DISS. ETH NO. 25358

PREDICTING THE FINANCIAL GROWTH OF SMALL AND MEDIUM-SIZED ENTERPRISES USING WEB MINING

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

Presented by

Yiea-Funk Te

M. sc. University of Zurich

Born on 04.04.1985

Citizen of Switzerland

Accepted on the recommendation of

Prof. Dr. Elgar Fleisch

Prof. Dr. Florian von Wangenheim

2018

Abstract

Small and medium enterprises (SMEs) play an important role in the economy of many countries. When the overall world economy is considered, SMEs represent 95% of all businesses in the world, accounting for 66% of the total employment. Existing studies show that the current business environment is characterized as being highly turbulent and strongly influenced by modern information and communication technologies, thus forcing SMEs to experience more and more severe challenges in maintaining their existence and expanding their business. To support SMEs at improving their competitiveness, researchers turned their focus on applying data mining techniques to build growth prediction models. However, current prediction models only include few data types such as financial or operational data and thus cannot explain the whole and complex context of SME growth. Moreover, data used to construct these models is primarily obtained via questionnaires, which is very laborious and time-consuming, or is provided by financial institutes, thus not publicly available and highly sensitive to privacy issues. Recently, web mining has emerged as a new approach towards obtaining valuable insights in the business world. Web mining enables an automated and large scale collection and analysis of potentially valuable data from the web, a popular and interactive medium with immense amount of data freely available for users to access. While web mining methods have been frequently studied to anticipate growth of sales volume for e-commerce businesses, it remains unclear how web mining can be applied to leverage

SMEs growth prediction. In investigating this question, the present thesis analyses the use of web mining for SMEs growth prediction.

In a case study, we demonstrate the use of publicly available web data for growth prediction in the gastronomy industry. First, a comprehensive overview of factors influencing the growth of restaurants is provided through a systematic literature review. In total, 49 factors influencing the growth of restaurants are identified, serving as a knowledge base to develop a growth model for restaurants. Next, the usability of various web data sources is manually inspected with respect to the identified growth factors. Web mining techniques are applied for large-scale collection and preprocessing of unstructured web data. Finally, based on data from 403 Swiss restaurants, we build and compare different binary classification models using supervised machine learning algorithms. More specifically, the developed models classify a restaurant either in a non-growing or growing restaurant. The algorithms for predictive modeling include logistic regressions, random forests and artificial neural networks.

The present thesis makes a significant contribution to the body of literature at the intersection of SMEs growth research, web mining, and applied machine learning. To summarize, our findings suggest that web mining is a feasible approach to leverage growth prediction modelling for SMEs. By means of web mining, valuable business insights can be extracted from the web, which then can be further used for predictive modelling by applying machine learning techniques. Moreover, to the best of our knowledge, our case study is the first to apply web mining combined with supervised machine learning techniques to model the growth of restaurants based on publicly accessible web data.

This study contains both theoretical and practical implications. It contributes to the existing literature of SMEs growth research by confirming previous findings in a data-driven and model-based manner through machine learning. Furthermore, the proposed approach can be used to identify new growth factors based on the feature importance measure of the applied machine learning algorithms and thus, extend the empirical body of knowledge. As a practical application, the findings of the present thesis can be used to build an information system which allows an automated collection and analysis of publicly available web data in large scale with the objective of predicting future growth opportunities of SMEs.

Kurzfassung

Kleine und mittlere Unternehmen (KMUs) haben für die Wirtschaft in vielen Ländern eine wichtige Rolle. Betrachtet man die globale Weltwirtschaft, repräsentieren KMUs 95% aller Unternehmen weltweit, was 66% aller Arbeitsstellen ausmacht. Bestehende Studien zeigen, dass das gegenwärtige Geschäftsumfeld als äusserst turbulent und stark von modernen Informations- und Kommunikationstechnologien geprägt ist, was KMUs vor größere Herausforderungen bei der Erhaltung ihrer Existenz und dem Ausbau ihres Geschäfts stellt. Um KMU bei der Verbesserung ihrer Wettbewerbsfähigkeit zu unterstützen, konzentrieren sich Forscher auf die Anwendung von Data-Mining-Techniken zur Erstellung von Wachstumsprognosen. Aktuelle Prognosemodelle enthalten jedoch nur wenige Datentypen wie Finanz- oder Betriebsdaten, und können daher nicht den gesamten und komplexen Kontext des KMU-Wachstums erklären. Darüber hinaus werden die für die Erstellung dieser Modelle verwendeten Daten in erster Linie über Fragebögen erhoben, die sehr aufwendig und zeitraubend sind, oder von Finanzinstituten bereitgestellt werden und daher nicht öffentlich zugänglich und sehr sensibel für Datenschutzfragen sind. In jüngster Zeit hat sich Web Mining als neuer Ansatz zur Gewinnung wertvoller Einblicke in die Geschäftswelt herausgestellt. Web Mining ermöglicht eine automatisierte und umfangreiche Sammlung und Analyse potenziell wertvoller Daten aus dem Web, einem weit verbreiteten und interaktiven Medium mit Unmengen an Daten, auf die die Benutzer frei zugreifen können. Während Web Mining Methoden häufig untersucht wurden, um das Wachstum von E-Commerce-Unternehmen zu ermitteln, es ist nach wie vor unklar, wie Web Mining eingesetzt werden kann, um die Wachstumsprognose von KMUs zu optimieren. Bei der Untersuchung dieser Frage analysiert die vorliegende Arbeit den Einsatz von Web Mining für die Wachstumsprognose von KMUs.

In einer Fallstudie untersuchen wir die Nutzung öffentlich zugänglicher Web Daten für die Wachstumsprognose in der Gastronomie. Zunächst wird ein umfassender Überblick über die Faktoren, die das Wachstum von Restaurants beeinflussen, durch eine systematische Literaturrecherche gegeben. Insgesamt konnten dadurch 49 Wachstumsfaktoren identifiziert werden, die als Wissensbasis für die Entwicklung eines Wachstumsmodells für Restaurants dienen. Anschließend wird die Nutzbarkeit verschiedener Web Datenquellen im Hinblick auf die identifizierten Wachstumsfaktoren manuell überprüft. Web Mining Techniken werden zur großflächigen Erfassung und Aufarbeitung von unstrukturierten Web Daten eingesetzt. Basierend auf den Daten von 403 Schweizer Restaurants erstellen und vergleichen wir verschiedene binäre Klassifikationsmodelle mit Hilfe von maschinellen Lernalgorithmen. Konkret klassifizieren die entwickelten Modelle ein Restaurant entweder in ein nicht wachsendes oder in ein wachsendes Restaurant. Die Algorithmen zur prädiktiven Modellierung umfassen logistische Regressionen, Random Forests und künstliche neuronale Netzwerke.

Die vorliegende Arbeit leistet einen wesentlichen Beitrag zur Literatur an der Schnittstelle von KMU Wachstumsforschung, Web Mining und angewandtem maschinellen Lernen. Zusammenfassend weisen unsere Ergebnisse darauf hin, dass Web Mining ein praktikabler Ansatz ist, um Wachstumsprognosen für KMUs zu entwickeln. Mit Hilfe von Web Mining lassen sich aus dem Web wertvolle Geschäftsinformationen gewinnen, die dann mit Hilfe von maschinellen Lernverfahren zur prädiktiven Modellierung weiterverwendet werden können. Darüber hinaus ist unsere Fallstudie nach unserem besten Wissen die erste, die Web Mining in Kombination mit maschinellen Lernen einsetzt, um das Wachstum von Restaurants auf Basis öffentlich zugänglicher Web Daten zu modellieren.

Diese Studie enthält sowohl theoretische als auch praktische Folgerungen. Sie trägt zur bestehenden Literatur der KMU Wachstumsforschung bei, indem sie bisherige Erkenntnisse datengetrieben und modellbasiert durch maschinelles Lernen bestätigt. Darüber hinaus kann der vorgeschlagene Ansatz verwendet werden, um neue Wachstumsfaktoren zu identifizieren basierend auf maschinellem Lernen. Als praktische Anwendung können die Ergebnisse der vorliegenden Arbeit genutzt werden, um ein Informationssystem zu entwickeln, das eine automatisierte Sammlung und Analyse von öffentlich zugänglichen Web Daten in großem Stil ermöglicht - mit dem Ziel, zukünftige Wachstumsmöglichkeiten von KMUs zu prognostizieren.

Disclaimer

This dissertation contains parts of working papers and previous publications by the author. Please also refer to the following contributions when building upon the results of this thesis:

Te Y.-F., Müller D., Pletikosa Cvijikj I. (2016). Design of a Small and Medium Enterprise Growth Prediction Model Based on Web Mining (extended abstract). In: *18th International Conference on Business Information Systems and Information Management*. Seoul

Te Y.-F., Pletikosa Cvijikj I. (2017). Design of a Small and Medium Enterprise Growth Prediction Model Based on Web Mining. In: *17th International Conference on Web Engineering.* Rome

Te Y.-F., Müller D., Wyder S. (2018a). Analysis of Web Data for the Purpose of Predicting SMEs Financial Growth. In: *International Conference on Economic Modeling.* Venice

Te Y.-F., Müller D., Wyder S., Pramono D. (2018b). Predicting the Growth of Restaurants using Web data. In: *28th International Scientific Conference on Economic and Social Development.* Paris

The author also contributed to the following publications, which are not part of this dissertation:

Pletikosa Cvijikj, I., Kadar, C., Ivan, B., **Te Y.-F.** (2015). Towards a Crowdsourcing Approach for Crime Prevention. In: *International Joint Conference on Pervasive and Ubiquitous Computing*. Osaka

Pletikosa Cvijikj, I., Kadar, C., Ivan, B., **Te Y.-F.** (2015). Prevention or Panic: Design and Evaluation of a Crime Prevention IS. In: *International Conference on Information Systems.* Texas

Te, Y.-F., Kadar, C., Róses Brüngger R., Pletikosa Cvijikj, I. (2016). Human versus Technology: Comparing the Effect of Private Security Patrol and Crime Prevention Information System over the Crime Level and Safety Perception. In: *European Conference on Information Systems.* Istanbul

Kadar C, **Te Y.-F.,** Róses Brüngger R, Pletikosa Cvijikj I. (2016). Digital Neighborhood Watch: To Share or Not to Share?. In: *International Conference on Human-Computer Interaction.* Toronto

Müller D, **Te Y.-F.,** Flavien Meyer, Pletikosa Cvijikj I. (2016). Towards data driven decision support for financial institutions: Predicting small companies business volume in Switzerland. In: *7th International Conference on Computer Science & Information Technology.* Amman-Jordan

Müller D, **Te Y.-F.,** Pletikosa Cvijikj I. (2016). An e-government service as PaaS application to serve Switzerland's municipalities. In: *13th IEEE International Conference on Services Computing.* San Francisco

Müller D, **Te Y.-F.,** Pratiksha J. (2017) Predicting business performance through patent applications. In: *IEEE International Conference on Big Data.* Boston

Declaration of Co-Authorships

The individual contributions of the authors to the publications, which are primarily contained in the present dissertation, are summarized below:

Publication	**Co-authors**	**Contribution**
Te Y.-F., Müller D., Pletikosa Cvijikj I. (2016). Design of a Small and Medium Enterprise Growth Prediction Model Based on Web Mining (extended abstract). In: 18th International Conference on Business Information Systems and Information Management. Seoul	Te Y.-F.	Conception and design of the work; drafting the article; data collection; data analysis and interpretation; critical revision of the article
	Müller D.	Data collection (ground truth data)
	Pletikosa Cvijikj I.	Critical revision of the article; final approval of the version to be published
Te Y.-F., Pletikosa Cvijikj I. (2017). Design of a Small and Medium Enterprise Growth Prediction Model Based on Web Mining. In: 17th International Conference on Web Engineering. Rome	Te Y.-F.	Conception and design of the work; drafting the article; data collection; data analysis and interpretation; critical revision of the article
	Pletikosa Cvijikj I.	Critical revision of the article; final approval of the version to be published
Te Y.-F., Müller D., Wyder S. (2018a). Analysis of Web Data for the Purpose of Predicting SMEs Financial Growth. In: International Conference on Economic Modeling. Venice	Te Y.-F.	Conception and design of the work; drafting the article; data collection; data analysis and interpretation; critical revision of the article; final approval of the version to be published
	Müller D.	Data collection (ground truth data)
	Wyder S.	Data storage; technical support
Te Y.-F., Müller D., Wyder S., Pramono D. (2018b). Predicting the Growth of Restaurants using Web data. In: 28th International Scientific Conference on Economic and Social Development. Paris	Te Y.-F.	Conception and design of the work; drafting the article; data collection; data analysis and interpretation; critical revision of the article; final approval of the version to be published
	Müller D.	Data collection (ground truth data)
	Wyder S.	Data storage; technical support
	Pramono D.	Data collection (ground truth data)

Acknowledgments

This dissertation is the result of my work at the Institute of Information Management at ETH Zurich in the period from 2015 to 2018, where I was part of Mobiliar Lab for Analytics, a joint initiative of the ETH Zurich and partners from the insurance industry. The cross-institutional setting provided an interdisciplinary experience. Close collaboration with industry partners created a challenging and highly rewarding environment in which relevant business problems laid the foundation for rigorous research. As the surrounding conditions were an excellent premise, it was the people who supported me during the last three years who made this thesis possible. For that, I would like to express my deepest gratitude. First and foremost, I would like to thank my supervisor Prof. Dr. Elgar Fleisch who created a multi-faceted and stimulating work environment. His professional and personal guidance helped me develop as a researcher and as a person. Further, I would like to thank Prof. Dr. Florian Von Wangenheim for his willingness to co-supervise my thesis. I am grateful for his time and the valuable and constructive feedback he contributed.

Especially, I would like to thank Dr. Irena Pletikosa Cvijikj for her academic support and persistent engagement. Dr. Pletikosa Cvijikj introduced me to academic publishing and provided structural guidance throughout the different stages of my research. Further, I would like to thank Dr. Gundula Heinatz who led the Mobiliar Lab for Analytics during most of my time at ETH Zurich. She triggered and supervised the industry cooperations in which I had the opportunity to contribute and gather valuable insights for my research. I would like to thank Dr. Erika Meins and Dr. Andrea Ferrario for guiding and supporting me through the last months of my PhD. Monica Heinz from ETH Zurich and Elisabeth Vetsch-Keller of University of St. Gallen deserve a special shout out for their outstanding organizational support.

I am very grateful for the chance to work with excellent industry partners during my time as a PhD student. A big thank you goes to Sebastian Wyder and Dwian Pramono from Die Mobiliar for the exciting collaboration and fruitful discussions. Further, I would like to thank all my colleagues at ETH Zurich and University of St. Gallen, a group of highly talented, creative researchers and entrepreneurs. The many interesting discussions including a fruitful exchange of ideas were fantastic and had tremendous impact on my development. It was always a great pleasure to work with you and I hope to stay in touch with many of you. My special gratitude goes to my friend Dr. Fabian Wahle, who always advised and supported me in overcoming numerous obstacles I have been facing through my research.

Finally, I want to express my deepest gratitude towards my family, my parents, Pheap and Vuy-Kong, and brothers, Yiea-Zhung and Yiea-Wey. Thank you for always putting me on the right track, for your support through all the years and your unconditional love. Finally and most especially, I would like to thank my wonderful girlfriend Yeongji for her support during my PhD and for always bringing a smile back on my face.

Zurich, July 2018

Yiea-Funk Te

Contents

List of Figures

List of Tables

1 Introduction

In this chapter, the general motivation and objectives of this thesis are outlined. Further it provides an overview of the methodological approaches and closes with the remainder of this thesis.

1.1 Motivation

Small and medium–sized enterprises (SMEs) are recognized worldwide for their contribution to economic stability and development, new job creation and employment, and social cohesion and growth (OECD, 2004). When the overall world economy is considered, SMEs represent 95% of all business in the world, accounting for 66% of total employment and 55% of total production (OECD, 2004). Moreover, according to the 6th Annual Report of the European Small Business Observatory (Lukács, 2005), there are 19.3 million of the enterprises in the European Union, with over 99% of them defined as SMEs and employing approximately 75 million people. Especially, SMEs in Switzerland play a pivotal role in the development of the country. The importance of SMEs is evidenced by their high presence in the economic structure of the country. According to a study conducted by Fueglistaller (2017), 99.8% of all Swiss industrial firms are SMEs and account for over 55% of production and 68% of all jobs and thus, acting as the countries backbone for economic growth.

However, studies reveal that the current business environment is characterized as highly turbulent, influenced by modern information and communication technologies, globalization, short innovation cycles and employee mobility (Antlová, 2009; Post, 1997). Additionally, the growing number of SMEs caused competition to become increasingly intensive, forcing SMEs to experience more severe challenges in maintaining their existence and expanding their business.

Given the significant importance of SMEs to the economic growth, policy makers throughout the world have initiated support for SMEs at their various stages of development. Furthermore, in an attempt to reduce the global phenomena of unemployment and poverty, worldwide organizations such as the International Labour Organization (Ashton and Sung, 2002) and United Nations Industrial Development Organization (Klarer et al.) have shown a high level of interest in supporting SMEs. Furthermore, in order to support SMEs at improving their competitiveness, researchers and academics have been analyzing factors influencing the success of SMEs for many decades, e.g. Altman (1968), Ohlson (1980), and Henebry (1996), etc. Moreover, with the emergence of big data, researchers turned their focus on applying data mining techniques to build risk and growth prediction models for SMEs e.g. Kim and Sohn (2010), Duman et al. (2012), and Kruppa et al. (2013), etc.

However, current prediction models only include few data types such as financial or operational data and thus cannot explain the whole and complex context of SMEs growth (Patel et al., 2011). Moreover, conventional data collection is primarily conducted via questionnaire studies, which is very laborious and time-consuming, or provided by financial institutes, thus not publicly available and highly sensitive to privacy issues. In addition, data mining techniques such as artificial neural network and decision trees are extensively studied with a strong focus on risk assessment and bankruptcy forecasting for SMEs rather than growth prediction. Although numerous studies on SMEs growth factors and growth modelling exist, studies reporting data mining based SMEs growth prediction are scarce.

Recently, web mining has emerged as an important field of study for both practitioners and researchers towards obtaining valuable business insights from the web, reflecting the magnitude and impact of data-related problems to be solved in contemporary business organizations (Kosala and Blockeel, 2000). Web mining denotes the use of data mining techniques to automatically discover Web documents, extract information from Web resources and uncover general patterns on the Web. Web mining research overlaps with other areas such as artificial intelligence along with machine learning techniques, data mining, informational retrieval, text mining and Web retrieval. It thus enables an automated and large-scale collection and analysis of potentially valuable data from the web.

In particular, web mining has shown to be very useful for e-commerce. In the increasingly fierce competition in the e-commerce, any information related to consumer behavior are extremely valuable (Patel et al., 2011). A major challenge of e-commerce is to understand customers' needs and value orientation as much as possible, in order to ensure competitiveness in the E-commerce era. Therefore, web mining is used to gather data which have potential value from the website of e-commerce companies, for example to increase customer attraction and retention.

In the technology- and information-driven world, the web has become a popular and interactive medium not only for e-commerce businesses but for SMEs as well. Zooming in on Switzerland for instance, the proportion of SMEs with an own website increased sharply from 9% in 1998 to 40% in 2002 (Sieber, 2002).

Following this trend, one can assume that a large amount of potential valuable information is stored on the web, which theoretically can be used to gain better insights about the growth mechanism of SMEs. While WM methods has been well researched and used in the field of e-commerce research to increase the sales volume, it has barely been applied for SME growth prediction modelling. Antlová et al. (2011) is one of the first and few studies that demonstrated the power of web mining for SME growth prediction. In their paper, they studied the relationship between long-term growth of SMEs, Information and Communication Technology competencies and a web presentation by using web mining methods. They applied web mining techniques to automatically extract potential valuable information for growth prediction. Their study showed that a long-term growing company could be recognized from the web presentation with high accuracy. Another recent study conducted by Li et al. (2016) explored micro-level characteristics and impacts of external relationships such as government or university relations on the SME growth by extracting business-relevant indicators from websites through web mining, demonstrating the potential of web mining for SME risk and growth research.

However, these studies only focus on the information available in company websites and thus, restrict the amount and spectrum of growth factors to the information typically given in company websites. Moreover, given the immense amount of web data sources publicly available, studies applying web mining techniques for predictive modelling can be greatly enhanced by using and combining multiple data sources to derive useful growth-related information for SMEs growth prediction. Hence, further research exploiting the full potential of web mining for SMEs growth prediction is required.

This research aims at investigating the potential of using web mining for SME growth prediction, hence contributing to the research field of web mining, applied machine learning and SME growth research. Web mining methods will be explored with the goal to automatically extract valuable growth-related information stored in the web, whereas machine learning methods will be studied in order to develop SME growth prediction models with high performance and applicability. To address the aforementioned research gap, research questions are formulated and elaborated in the following section.

1.2 Research Questions

As pointed out in the previous subsection, several research gaps are identified in the domain of data mining and web mining based SMEs growth prediction modelling. First, most of the prediction models for SMEs focus on the anticipation of credit risk or bankruptcy. Although numerous studies on SMEs growth models exist, research reporting data mining based SMEs growth modelling are rare. Moreover, these models only include a few number of growth-influencing factors, which cannot capture the whole mechanism of SMEs growth. Second, conventional data collection to assess the growth factors is primarily conducted via questionnaire studies, which is very laborious and time-consuming. Furthermore, questionnaire studies suffer from well-known pitfalls such as low response rates and response bias (Lussier and Halabi, 2010). Third, studies using web data for SMEs growth prediction modelling are very limited. In addition, most web mining studies for SMEs only focus on the information contained in company websites, thereby limiting the number of factors integrated into the models.

In order to address these issues, in the present thesis we further investigate the use of publicly available web data for SMEs growth prediction. In particular, we aim at understanding how the web mining process can be used to systematically generate business-relevant knowledge from the informational richness of the web to predict the growth prediction model for SMEs. Thus, the first research question is stated as follows:

> ***RQ1. How can web mining be operationalized to study SME growth prediction?***

The growth of SMEs is an extremely complex mechanism which is characterized by a large amount of firm-internal and external factors. Moreover, the underlying factors for growth differ depending on the type of business (Scott and Bruce, 1987). Therefore, it is very important that growth models are developed for specific industries. Moreover, web mining is not equally useful for all industries. It can be stated that web mining unfolds its full potential if used for analyzing large volumes of web data, which presumes that business information are sufficiently covered on the web. Thus, business areas which benefit the most from web mining are those with a strong web presence (Gök et al., 2015).

Thus, to evaluate the feasibility of web mining for growth prediction, we investigate the use of web mining to forecast the growth of the gastronomy industry. The gastronomy business is in particular interest due to its importance to the economy of many countries. This is especially true for Switzerland, where the gastronomy industry accounts for a large share of all jobs in small and medium enterprises. More specifically, we believe that restaurants are a good choice to conduct the case study, as large volume of restaurant information are stored in the web due to their distinct marketing efforts for customer acquisition (Murphy et al., 1996). Moreover, although numerous studies has attempted to explain the growth of restaurants, studies reporting web mining based restaurant growth models cannot be identified. Hence, the second research question is formulated as follows:

RQ2. To which extend can we develop a web data based growth prediction model for the restaurant industry?

1.3 Approach

Relating to the research questions, the present thesis is designed to achieve a balance between theory and practice. Thus, the proposed research approach reflects an identical premise that combines literature analysis with practical investigations.

In order to address the first research question, a thorough literature survey is necessary in order to recapitulate the developments in the relevant fields. First, to understand the underlying mechanism of SMEs growth, we survey the key determinants of firm growth. Next, we review SMEs growth prediction studies applying data mining techniques and then review the literature yielding ideas for application scenarios of web mining. Finally, we derive a framework which includes all conceptual and technical aspects of web mining for SMEs growth prediction modelling. For an in-depth description of the applied methods please refer to Chapter 3.

To answer the second research question, we design a case study, where we apply and test the proposed framework resulting from the findings of RQ1 on the gastronomy industry. A systematic literature review is conducted to determine the factors influencing the growth of the restaurant business. Thereby, we followed the guideline for systematic reviews provided by Okoli and Schabram (2010). Next, web mining are applied to extract growth relevant information from the web, which were previously identified through our literature review on growth factors. For this purpose, we first inspect the usability of various web data sources with regard to growth-related information richness. Finally, several techniques from the field of Machine Learning are deployed in order to distinguish non-growing from growing restaurants based on the publicly available web data. For an in-depth description please refer to Chapter 4.

1.4 Thesis Outline

The remainder of this thesis is structured as follows: The next Chapter 2, provides further information about the research context of the present work. This contains a detailed explanation of web mining and its fields of application, an overview of SMEs growth factors, followed by a survey of SMEs growth prediction studies and web mining based SMEs growth prediction studies. Chapter 3 guides through the methodology used to accomplish the thesis. This is followed by Chapter 4, which presents the study in the context of gastronomy growth prediction where the described methodology are applied in order to address the aforementioned research questions. The study begins with a specific introduction and overview of the theoretical background of the gastronomy business. Then, a comprehensive explanation of the used web data sources and data collection is provided, followed by discussions of the data analysis and results. Chapter 5 concludes this thesis with a general discussion of the key findings and contributions to theory and practice. Finally, we summarize the limitations of our research and suggest future research directions.

2 Research Background

The present thesis is situated at the interface of three intensively investigated domains: web mining, data mining and SME growth research. To provide a better understanding of the core literature that will be considered, this chapter aims to provide an overview of the literature spanning specific topics from these domains. First, an overview of web mining and its applications is introduced. Further, a survey of factors influencing the growth of SMEs is provided, followed by an overview of SMEs growth prediction studies. Finally, this chapter concludes with an overview of web mining based growth prediction studies.

2.1 Overview of Web Mining and its Applications

2.1.1 Web Mining Taxonomy

The term Web Mining (WM) broadly covers an emerging field of research which has witnessed an enormous increase in the interest of researchers and scientific publications over the last ten years. Today, WM is a multidisciplinary pool of concepts and overlaps with other research fields such as artificial intelligence, machine learning techniques, data mining, informational retrieval, text mining and Web retrieval (Liu, 2007).

Therefore, it is difficult to find a generally accepted definition of web mining, since it varies depending on the application and task of web mining (Cooley et al., 1997;

Kosala and Blockeel, 2000; Stumme et al., 2006; Liu, 2007). For example, some researchers associate web mining with information retrieval from the web, while others consider it a tool for analyzing web usage patterns (Kosala and Blockeel, 2000). A rather comprehensive definition is proposed by Kosala and Blockeel (2000): Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Thus, web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data.

In the scope of the present thesis, web mining shall be seen as

…a set of techniques and technologies for the automated extraction and analysis of information, to generate knowledge and useful insights from web content.

It is clearly rooted on and strongly related to data mining and knowledge discovery in databases. However, as being remarked by Liu (2007), it is not sufficient to regard web mining as a sub-discipline of data mining, since it has to cope with unique issues such as detecting and extracting pieces of information from the web (see explanation in Chapter 2.3) .

Web mining can classified on the basis of two aspects: the retrieval and the mining. The retrieval focuses on retrieving relevant information from large repository whereas mining research focuses on extracting new information (Sharda and Chawla). In general, WM tasks can be categorized into three categories, as shown in Figure 1 (Kosala and Blockeel, 2000): **Web structure mining, web usage mining**

and web content mining. Furthermore, the categories contain the following tasks (Kosala & Blockeel, 2000): (1) retrieving intended web documents, (2) automatically selecting and pre-processing specific information from the retrieved web resources, (3) automatically discovering general patterns at individual websites and across multiple sites and (4) validation and interpretation of the mined patterns. In a first step, all relevant documents are retrieved using information retrieval, then relevant facts are extracted out of these relevant documents using information extraction. The next step is the use of machine learning techniques and data mining techniques to generalize this data and in the last step analysis is being made of these new mined patterns (Kosala & Blockeel, 2000). For the purpose of completeness, the three aforementioned categories will be elaborated, although the focus of this research lies on web content mining, which has shown to be very useful in the business world, particularly in e-commerce (Saini and Pandey, 2015).

Web structure mining is the process of finding structure information from the Web. Useful information are hyperlinks and other structural elements of Web documents such as HTML tags or metadata. Particularly, web structure mining aspires to obtain insights on how pages are structured and linked among each other. Web Structure Mining can be further differentiated into intra-page structure mining, which focuses with the structure of individual pages, and inter-page structure mining, which aims at the references and relationships between pages (Cooley and Srivastava, 2000; Svristava et al., 2005). Web structure mining is conceptually linked to the analysis of social networks and serves as a key technology for search engines (Liu, 2007).

Web usage mining aims to uncover patterns of usage patterns that track how visitors navigate a website (Kosala and Blockeel, 2000). The primary data source is the user's click-stream data located in the web server log file. Web usage mining provides information on how a user interacts with a website, how much time a user spends on a page or how users can be grouped and classified according to common criteria. Linder (2005) further differentiates between Web Log Mining and integrated Web Usage Mining. The former only considers log files and protocols, while the latter also evaluates additional data such as user profile data or sales data. The applications generated from web usage mining can be divided into personalization, system improvement, site modification, business intelligence and usage characterization (Srivastava et al., 2000).

Web content mining pursues the discovery and extraction of information and knowledge by directly processing the contents of a website (Kosala and Blockeel, 2000). Content manifests itself in numerous forms on the web and is usually provided in text documents, semi-structured documents or multimedia documents such as images or videos. Web content mining fulfills a number of tasks, ranging from the identification and classification of the page's content of a page to the collection of opinions and sentiments between the lines. It is considered as the most complex and technically demanding field in web mining and is currently the focus of interest of researchers. In this thesis we focus on the application of web mining to text information. Some of them are semi-structured such as HTML documents or more structured as data in the tables or database generated HTML pages, but most of the data is unstructured text data (Saini and Pandey, 2015). Different techniques need to be applied in all three types of data. Further details are provided in chapter 3.3.

It is important to note that all categories aim to generate knowledge from web documents. The information obtained through Web Structure Mining is mainly of a technical nature and refers to the inherent structure of Web documents. Web usage mining provides behavioral, social or contextual information about the visitors of web documents. Web content mining enhances the information contained in the content of a Web document. Powerful applications are created when all three categories are used simultaneously, as suggested by Cooley and Srivastava (2000).

2.1.2 Web Mining Applications for Business

One result of the fascination with the web in recent years has been that Web applications have been developed at a much faster rate in the industry than research in Web related technologies. The aim of this chapter is to describe the applicability of Web Mining for business activities such as trend monitoring (Zaiane and Han,

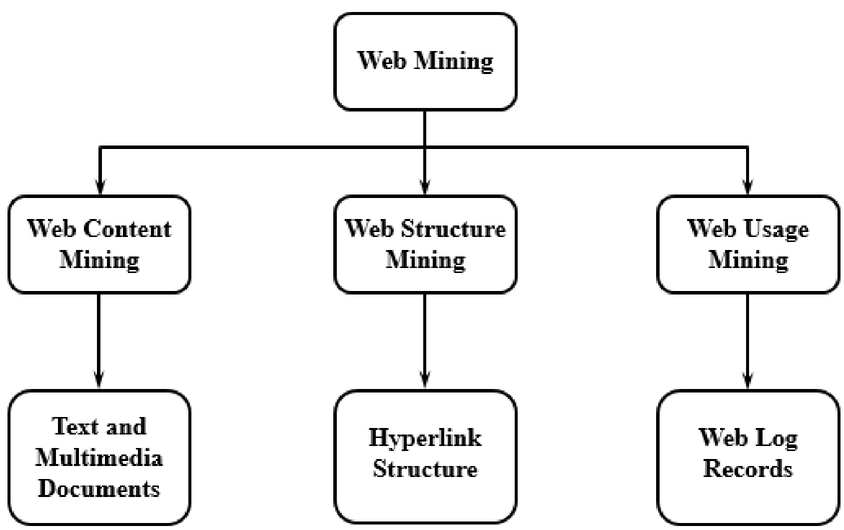


Figure 1: Overview of web mining categories and units of information which are examined (Saini and Pandey, 2015).

1998) and market research (Spangler and Chen, 2008) by presenting exemplary application scenarios, which are obtained from surveying web mining literature. Although the examples described are not limited to specific sectors, some are very well suited to the information demands of a single sector, such as e-commerce or retail. Following the focus of this work, this chapter focuses on the use of Web Mining for business applications.

Web mining covers a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. For research and development operations, businesses can harness web mining as source of inspiration for the generation of ideas and solutions to be infused in product or service innovations. Web mining sustains and enriches the innovation process through delivering first-hand information about current developments in technology. For instance, web mining can be applied to scan web documents efficiently, to identify and extract those with content of interest and to generate information about current innovation trends, subjects of research and directions of technological developments (Gu and Huang, 2008). Further, in the context of trend scouting, web mining can be applied to detect topics, themes and fashions of recent interest, which may impact the business environment in the near future. Especially the trends which affect public life and shape consumer behavior usually pervade the social web at a rather early stage (Schultze and Postler, 2008). Therefore, mining and analyzing user-generated content of online discussions may yield knowledge about the themes of public interest (Bandaru et al., 2011). Web mining can be applied to monitor the key topics, which are currently subject of discussion, or to identify topics of potential future interest (Liu and Chen-Chuan-Chang, 2004).

For marketing and sales activities, web mining supports for effective planning, optimization and evaluation of campaigns or other promotional activities. For instance, web mining can be applied to acquire knowledge about the contexts, in which products or brands are mentioned or discussed (Spangler and Chen, 2008). Moreover, web mining is also feasible to investigate the images of competitors' products or brands for comparison purposes (Xu et al., 2011). To evaluate and optimize online campaigns, web mining can be applied to analyze the speed by which campaigns diffuses on the web (Koran, 2010). Another interesting application of web mining in the field of marketing and sales is the detection of online communities, which is a vital component for conceptualizing campaigns of online promotions. Online communities establish through interaction over a longer period and find their foundations in common interests and affiliations and may offer the perfect target audience for a company's promotions (Java, 2008). Such communities might manifest as groups in social networks or as networks of affiliated companies, which can be seen as ideal ground for the placement of promotional measures. Thus, web mining facilitates the identification of suitable communities for marketing and sales purposes.

In the area of customer service management, web mining is a useful instrument for the analysis of online customer feedback and the provision of product recommendations. Online customer feedback provides valuable insights on the level of customer satisfaction, market adoption and improvement for a company (Thorleuchter et al., 2010). Thus, web mining can be used to identify and extract relevant insights from these reviews (Miner et al., 2012). In the case of product recommendations, recommender systems such as Amazon are especially useful for e-commerce, to support customers at finding articles of interest more quickly and to augment cross-sales significantly (Srivastava et al., 2005).

In the domain of public relation management, web mining comes to play essentially for the purpose of tracking the image and reputation a company faces in the media. For example, web mining can be used to observe online media and refined the flood of publications for articles of relevancy for a company (Brauckmann, 2010). Articles may be of relevancy when they refer to the company, to its executives, to competitors or other issues impacting the business environment. Especially user-generated media has proven to be a seeding place for rumors (Schultze and Postler, 2008). Due to the rapid rate of diffusion, such rumors often attract broad attention and pervade media long before the company acquired knowledge about the subject. Therefore, applying web mining for early identification of such issues enables the launch of respective counter-measures to prevent reputational damage for a company (Brauckmann, 2010).

Another domain of application for web mining is the detection of legal violations on the web. Common objectives include the identification of copyright infringement on intellectual property such as software, music or videos. Examining pages such as ware-sites or torrent-tracker-sites yield a good estimate on the scope and extent of the violation (Srivastava et al., 2005).

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. In this section, we described a number of prominent applications of web mining for the business context. However, the use of web mining is not limited to business applications, but it can also be applied to study the growth of SMEs. Thus, the remainder of the

Chapter 2 focuses on SMEs growth research, and prediction studies for SMEs applying data mining and web mining techniques.

2.2 Small and Medium Enterprises Growth Research

2.2.1 Definition of Small and Medium Enterprises

The definition of SMEs varies quite widely from country to country and even within single countries, depending on the business sector concerned. The World Business Council for Sustainable Development report (2007) stressed that there is no universally agreed definition of SMEs. Generally, scholars view small enterprises to be businesses employing between one and nine employees, and medium enterprises as those with between ten and ninety nine employees, although both types of SMEs have to be privately owned (Wijst, 1989). Other definitions state that SMEs are enterprises, which employ less than 100 people, and report an annual turnover of less than 10 million Euros. Another alternative is that SMEs are seen as firms which have a minimal share in the market, and which are not formally structured but are managed by personalized owners or part-owners that do not form a part of a large enterprise or firm (Storey, 1994).

In contrast, some other scholars define small businesses purely around their employee figure alone, and state that an enterprise with a workforce lower than 200 is small (Michaelas et al., 1999). In addition, different scholars have analyzed sales details, and defined a business as small when its annual sales fall between USD 0.5 to 2.5 million, and USD 2.5 to 16 million are deemed the sales margin for medium businesses (Lopez-Gracia and Aybar-Arias, 2000). Further, other scholars have stated that SME definitions are ambiguous, as basing the description on the size factor alone can be misleading because being small in one sector is not necessarily small in another. Likewise, it is perceived that a definition of SMEs cannot be universally agreed, as the nature and circumstances of their operations can alter from country to country (Mutula and Van Brakel, 2006).

In Switzerland, no official definition of SMEs exists (The World Business Council for Sustainable Development, 2007). The State Secretariat for Economic Affairs of Switzerland (SECO) applies a single criterion on the definition of SMEs: the number of employees. Each enterprise, irrespective of its legal form and activity, is regarded as an SME if it employs fewer than 250 people, i.e. between 1 and 249 employees. In addition, SMEs can be divided in three groups according to the number of employees (Fueglistaller, 2017): (1) micro enterprises with less than 10 employees, (2) small enterprises with number of employees between 10 and 50, and (3) medium enterprises with number of employees between 51 and 249. This definition is aligned with the definition provided by the commission of the European Union (European Union).

With regard to the Swiss business landscape of 2014, 99.8% of all 578'000 companies are SMEs which account for 68% of all 4'370'000 jobs (Fueglistaller, 2017). Thereby, micro enterprises account for 92.4% of business in Switzerland, followed by small enterprises and medium enterprises with 6.2% and 1.2% respectively. Moreover, 26.9% of all jobs are created by micro enterprises, 20.7% by small enterprises and 20.3% by medium enterprises. Further, SMEs can be broadly divided into three categories. The Federal Statistical Office distinguishes three sectors (Fueglistaller, 2017): the first, second and third sector. The first sector comprises agricultural and forestry enterprises, the second sector covers industrial and construction enterprises and the third sector (also called tertiary sector) concerns service enterprises. The Swiss SMEs landscape is dominated by the third sector, which accounts for 74.8% of all SMEs and 69% of all jobs in SMEs. The second sector accounts for 15.7% of all SMEs and 25.7% of all jobs in SMEs, followed by the third sector which constitutes 9.5% of all SMEs and 5.3% of all jobs in SMEs.

2.2.2 Definition of Growth

Growth is considered to be one of the key benchmarks of business success by practitioners. However, there is no consistency in the dimension of growth which theorists have used as the object of analysis. Different definitions have been used in the studies that attempted to explain the growth of SMEs. Some researchers advocated the strict use of financial indicators, while others emphasized the relevance of non-financial aspects of business success such as personal satisfaction and achievement (Buttner and Moore, 1997; Simpson et al., 2004; Walker and Brown, 2004). Financial growth measures include growth of revenues and profits (Cho et al., 2006). Researchers argued that for organizations to be considered successful, it is important for them to generate income and increases in profit, and to demonstrate some level of growth, as indicated in their sales revenue and income (Perren, 2000). Non-financial growth measures include growth of employment, customer satisfaction and loyalty (Brown and Mitchell, 1993). Jennings & Beaver (1997) argued that the attainment of personal objectives such as the desire for personal involvement, responsibility and the independent lifestyle, rather than financial outcome, is the best principal criterion of success for many business owners. In the present thesis, the adopted definition of growth of SMEs is the growth of annual revenue, due to its importance to the economy (Lev and Radhakrishan, 2010). Moreover, the growth in annual revenue is an objective measure which can be based on the accomplishment of the exact business objectives. It is considered to be a quantifiable measurement method with the ability to examine the quantity and quality of productivity of a business, such as sales or profit (Chong, 2008).

2.2.3 Survey of Factors influencing SME Growth

This sector provides an overview of the factors influencing the growth of SMEs. It is important to note that we do not limit the survey of growth factors to financial growth, as we aim to provide a broad overview of the factors influencing the growth of SMEs. Thus, this overview is a collection of factors that have been proven to be influential for the growth of SMEs in the general sense. In addition, this overview contains the survey of both qualitative and quantitative studies.

The current business environment is characterized as complex and fast-changing, influenced by a variety of firm internal and external factors. Beck and Demirguc-Kunt (2006) argued that for new SMEs to grow, it is important to strengthen not only the internal business environment but also the external environment. Literature on the success of SMEs usually identifies several factors with regard to the internal and external environment of the firm, as illustrated in Figure 2 (Worthington and Britton, 2009). In the following, the growth factors are briefly discussed.

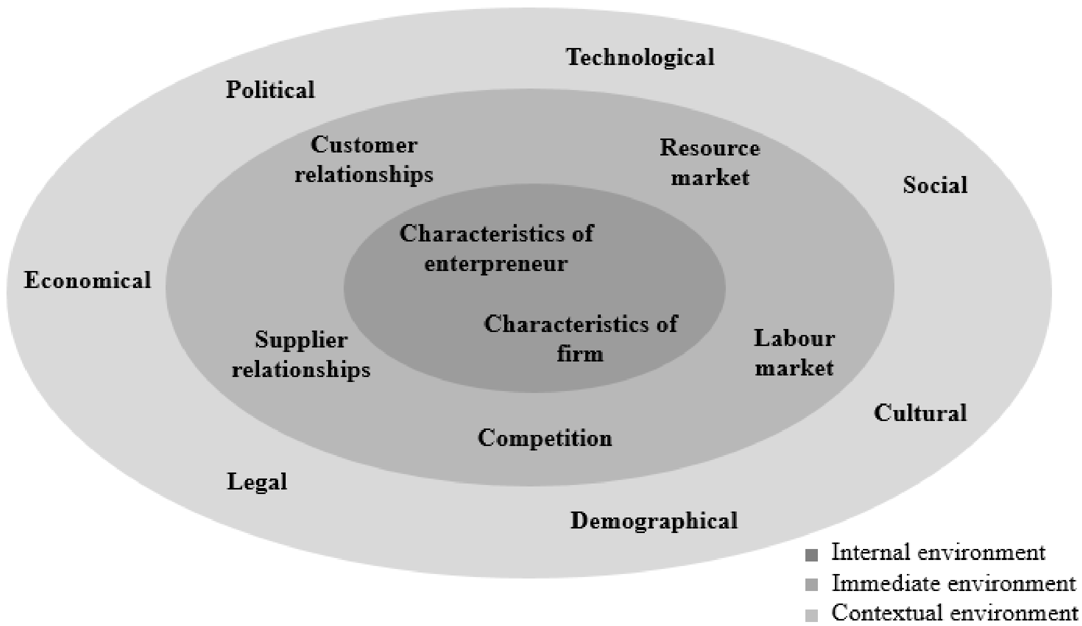


Figure 2: Overview of factors influencing the growth of SMEs.

Firm-internal factors

Firm-internal factors - denoted as internal environment in Figure 2, includes all firm-specific factors that are influenced by specific firm action, including the availability of resources, personal skills and abilities for pursuing entrepreneurial functions and the effective use of resources inside the firm (Chittithaworn et al., 2011). Thus, it can be argued that the internal business environment significantly influences the success of a business (Ligthelm and Cant, 2002; Dockel and Ligthelm, 2005). Furthermore, researchers have argued that characteristics of SMEs, characteristics of the entrepreneur, strategies and organizational structure of the firm are among the internal factors that influence SMEs success and growth (Storey, 1994). Therefore, the understanding and focus on the internal factors may improve business success

(Naffziger, 1995). Firm-internal factors can be roughly divided into two groups: (1) the characteristics of the firm such as firm attributes and firm strategies, and (2) the characteristics of the entrepreneur such as socio-demographic characteristics, and the personality and the competences of the entrepreneur.

Several studies have attempted to explain the link between the characteristics of firm and SMEs growth (Bates and Nucci, 1989; Storey, 1994; Baum and Locke, 2004). In general, the characteristics of firm can be grouped into 4 categories: Firm attributes, firm strategies and resources and organizational structure. For instance, Storey (1994) identified characteristics of the SMEs among the key components that are important in analyzing the growth of SMEs. Firm attributes that affect the growth of SMEs have been identified as age, size, and location of business (Kraut and Grambsch, 1987; Kallerberg and Leicht, 1991). Furthermore, numerous researchers have argued that the growth of a firm strongly depends on the firm strategy it adopts (Storey, 1994; Olson and Bokor, 1995; Pearce and Robinson, 2009). One key element of firm strategy is innovation, which effect on firm growth has been recognized in several studies. For instance, Ashton and Sung (2006) argued that innovation in product and services are essential to sustaining competitive advantage for firms offering differentiated good and services. Other firm characteristics well-studied by researchers and proved to be influencing the growth of SMEs include firm resources such as financial resources and human capital (Beck et al., 2006), and the organizational structure such as work specialization, centralization of work and the firm legal firm (Olson et al., 2005).

For many years, researchers have shown great interest in understanding the characteristics of entrepreneurs for many decades (Altman, 1968). Numerous studies analyzed the characteristics associated with entrepreneurship in order to distinguish the properties between entrepreneurs and non-entrepreneurs (Gartner, 1989). Characteristics of the entrepreneur such as specific traits and attitudes, which are defined heuristically, are often cited as the most influential factors related to the growth of SMEs and their competitiveness Man et al., 2002; Simpson et al., 2004; Gürol and Atsan, 2006). In this literature review, growth factors related to the characteristics of the entrepreneur are grouped into three categories: the socio-demographic characteristics of the entrepreneur, his personality characteristics and competences.

Numerous studies demonstrate that demographic characteristics, such as age and gender, and individual background influence the growth of SMEs. For instance, Reynolds et al. (2000) found that individuals aged 25-44 years were the most entrepreneurially active. This is supported by another study conducted by Woldie et al. (2008), which reported that middle-age and older owner-manager tend to run more growth oriented firms. In consideration of gender, a considerable amount of literature has been published on the effect of gender on the performance of SMEs (Johnsen and McMahon, 2005). However, these studies produce mixed and inconclusive results. For instance, in a quantitative and a qualitative study assessing the gender-related differences among 32 micro rural enterprises in Sweden, Sandberg (2003) concluded that there were few differences. Moreover, in examining whether gender has an impact on firm growth, Elizabeth and Baines (1998) conducted a study on a sample of 104 micro businesses in business services in two different locations in the UK. Their study found no effect of gender on firm growth. However, other studies report that male business owners were characterized with higher tendency of survival than their female counterparts (Boden and Nucci, 2000; Watson, 2003). Furthermore, several studies found that the individual background of the entrepreneur, such as education, previous work experience and family background, had an impact on the growth of SMEs (Richard, 2000; Brush, 2001; Gray et al., 2006). Moreover, personal qualities and traits such as high need for achievement, locus of control and propensity for risk-taking have often been associated with successful entrepreneurship (Begley and Boyd, 1987; Mueller and Thomas, 2001; Stewart et al., 2003). In addition, a large body of research highlight the importance of managerial and entrepreneurial competences for the growth of SMEs (Ibrahim and Goodwin, 1986; Walker and Brown, 2004).

Firm-external Factors

Firm-external factors - denoted as immediate and contextual environment in Figure 2, have been found to have a significant impact on the growth of SMEs. A study conducted by Hannan and Freeman (1977) suggests that organizations are constrained by the external environment they operate in. Consequently, the firm's growth is determined largely by these external factors. Davidsson et al. (2005) argued that growth is to a large extent a question of ambitions and abilities, but the fundamental drivers and barriers in the environment cannot be underestimated. The growth effects of a dynamic environment have been demonstrated in the literature. Numerous studies showed fast growing firms are more often found in industries and regions that are more dynamic (Jovanovich, 1982; Carroll and Hannan, 1989). Dahlqvist et al. (2000) highlighted that external factors offer opportunities and risks that can affect all entrepreneurs in their environment, regardless of their background, education or business concept. Further, Mazzarol et al. (1999) pointed out that these factors cannot be controlled and the success of SMEs often depends on the managerial ability to deal with them. Firm-external factors can be roughly divided into 2 groups: factors reflecting (1) the immediate and (2) the contextual environment.

According to Worthington and Britton (2009), the immediate environment includes supplier and customer relationship, competition, labor market and resource market. In general, shortcomings in the immediate business environment are the main obstacles to SMEs growth (Worthington and Britton, 2009). A large volume of studies describe customer relationship management as a key factor for the growth of SMEs (Dwyer et al., 1987; Morgan and Hunt, 1994). For instance, Temtime and Pansiri (2004) found in a survey of 203 SMEs that customer relationship was rated highly by the respondents in its impact on the performance of their firms, highlighting the strong business competition characterizing today's business environment. Therefore, focusing on how to find and retain profitable customers is a key factor for SMEs to survive in the global markets and in the increasing competitive environment (Kalakota and Robinson, 2001). Another crucial factor influencing the growth of SMEs is the understanding of the competition. SMEs operate in a global environment characterized by increased competition and unknown competitors (Ligthelm and Cant, 2002). The concentration of competition, and the market actions and strategies of competitors have an impact on the business process (Baron, 2004). Therefore, an analysis of the role of competitors and their behavior is crucial for the growth of an SME (Ligthelm and Cant, 2002). Further, several studies identified the importance of supplier relationships for the growth of

SMEs (Morrissey and Pittaway, 2006; Gelinas and Bigras, 2004). Suppliers directly influence production costs, quality and schedules of delivery of goods and services. Therefore, it is crucial for SMEs to have an established supply chain function to be successful in a complex business environment (Gelinas and Bigras, 2004).

The contextual environment comprises macro-environmental factors such as economic, political, socio-cultural, technological and legal influences on the growth of businesses which can emanate not only from local and national sources but also from international developments (Worthington and Britton, 2009). Economic factors include forces that regulate exchange of materials, money, energy and information. Several studies demonstrated that the general state of an economy, in which a firm competes, influences the performance of a business (Boddy, 2002; Ligthelm and Cant, 2002; Baron, 2004; Gurol and Atsan, 2006). Political-legal factors include forces that allocate power and provide constraining and protecting laws and regulations. Political and legal systems affect the way business is conducted, by defining what firms can and cannot do at a particular point in time (Boddy, 2002). Technological factors include forces that generate problem-solving inventions which can affect all aspects of a business, from its overall strategic position to how it manages marketing, design, production, and distribution (Boddy, 2002). The technological factors covered in this review are the accessibility to information and infrastructure (Swierczek and Ha, 2003; Bottasso and Conti, 2010). The socio-cultural factors involve the social and cultural aspects of the environment. These consist of customs, lifestyles, and values that characterize the society in which firms operate. Several studies demonstrated that these factors have a major impact on business growth (Wasilczuk, 2000; Boddy, 2002; Gürol and Atsan, 2006).

2.3 Survey of SMEs Growth Prediction Studies

This chapter provides an overview of the data mining and web mining studies related to SMEs growth. Data mining is commonly defined as the process of discovering useful patterns or knowledge from data sources, such as databases, texts and images (Witten et al., 2016). It is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization. Web mining refers to the process of discovering useful patterns or knowledge from the web. Given the complexity of the web and its unique characteristics, mining useful information and knowledge from the web is a challenging task (Liu, 2007). Data mining and web mining are strongly related to each other. Although web mining is strongly associated with data mining, web mining cannot be regarded as a sub-discipline of data mining, since web mining has to cope with unique challenges from the web (Liu, 2007).

2.3.1 Data Mining Studies on SMEs: Literature review

The literature on business growth models dates back to 1967 and has proliferated since then into different streams addressing specific industries and business sizes. Lippitt and Schmidt (1967) developed a general growth model for all sizes of businesses by examining how personality development theories influences the creation, growth and maturation of businesses in general. A few years later, Steinmetz (1969) qualitatively analyzed the growth of SMEs by partitioning the growth curve of SMEs into different stages and assessing the characteristic attributes of each stage. A qualitative study conducted by Scott and Bruce (1987) suggested a model for small business growth supporting managers to plan for future growth. The proposed model isolates five growth stages characterized by a unique combination of firm attributes. As a small company goes through different growth stages, attributes such as the management style, organizational structure and the use of technology changes. Although stage models are widely accepted among researchers and practitioners, stage models are criticized on some counts (O'Farrell and Hitchens, 1988). First, some of them seem more than heuristic classification schemes rather than a conceptualization of the processes underlying growth. Second, they implicitly assume that a small business will either grow and pass through all stages or fail in the attempt. Empirical evidence does not justify such an assumption. Third, the models only include firm internal characteristics such as management style and organizational structure and do not incorporate environmental influences on the business growth. Furthermore, empirical research is conducted on small sample sizes and specific types of businesses via questionnaires studies and thus, could threaten the validity of stage models (Farouk and Saleh, 2011).

With the emergence of big data, data mining techniques have been extensively studied in the domain of SME research. However, these studies mostly focused on the prediction of SME risk evaluation and bankruptcy rather than SMEs growth modelling. An early study indicated that backpropagation neural network were the most popular machine learning techniques among researchers in the finance and business domain during the 1990’s (Wong et al., 2000). For instance, Zhang et al. (1999) provide a comprehensive review of Artificial Neural Network (ANN) applications for bankruptcy prediction. Their findings indicated that ANN perform significantly better than logistic regression models. West (2000) investigated the credit scoring accuracy of various ANN models in comparison with traditional methods such as logistic regression and discriminant analysis model. Consistent with the findings of Zhang et al. (1999), ANN models perform slightly better than traditional methods.

Other techniques widely applied in the domain of risk evaluation and bankruptcy prediction includes decision trees (DT) and their ensemble variations such as random forests (RF). For instance, Fantazzini and Figini (2009) developed a model based on RF for SME credit risk measurement and compared its performance with the traditional logistic regression approach. They came to the conclusion that both models provided similar results in terms of performance, highlighting the potential of RF for credit risk modelling. Another recent and more application-oriented study conducted by Ozgulbas and Koyuncugil (2012) proposed an early warning system based on DT-algorithms for SMEs to detect risk profiles. The proposed system uses financial data to identify risk indicators and early warning signs, and create risk profiles for the classification of SMEs into different risk levels.

In summary, data mining techniques such as ANN, DT and RF are extensively applied for SME risk evaluation and bankruptcy prediction. However, even though numerous studies on SMEs growth factors and models exist, research reporting data mining based SMEs growth prediction are very limited. Furthermore, current prediction models usually include one type of data sources and thus cannot explain the whole and complex context of SMEs growth (Ozgulbas and Koyuncugil, 2012).

2.3.2 Web Mining Studies on SMEs: Literature review

The web is a popular and interactive medium with intense amount of data freely available for users to access. It is a collection of documents, text files, audios, videos and other multimedia data (Malarvizhi and Saraswathi, 2013). With billions of web pages available on the web, it is a rapidly growing key source of information, presenting an opportunity for businesses and researchers to derive useful knowledge out of it.

While WM methods has been well researched and applied in a wide range of the field, as described in Chapter 2.1.2, it has barely been used for SME growth research. Antlová et al. (2011) is one of the first and few studies that demonstrated the power of WM for SME growth prediction. In their paper, they studied the relationship between long-term growth of SMEs, Information and Communication Technology (ICT) competencies and a Web presentation by using WM methods. They applied web content mining techniques to automatically extract potential valuable information for growth prediction. Their study showed that a long-term growing company could be recognized from its Web presentation with high accuracy. Another notable study recently conducted by Li et al. (2016) explored micro-level characteristics and impacts of external relationships such as government or university relations on the SME growth by extracting business-relevant indicators from websites through web mining, demonstrating the potential of web mining for SME growth research. Finally, Thorleuchter and Van Den Poel (2012) analyzed the impact of textual information from e-commerce companies' websites on their commercial success by extracting web content data from the most successful top 500 worldwide companies. The authors demonstrated how text and web mining methods can be applied to extract e-commerce success factors from the company websites for predictive modelling.

However, these studies only focus on the information available in company websites and thus, restrict the amount and spectrum of growth factors to the information typically given in company websites (Gök et al., 2015). Hence, further research exploiting the full potential of web data for SMEs growth prediction is required.

3 Methodology

The development of the web has brought us enormous and constantly growing amounts of data and information. With the vast amount of data provided on the web, it has become an important resource for all kinds of research. However, due to the semi-structured or even unstructured nature of the web, traditional data extraction and data mining techniques cannot be applied. Web pages are hypertext documents that contain both text and hyperlinks to other documents. Moreover, the web data is heterogeneous and dynamic. Designing and implementing a web data mining based research framework has therefore become a challenge for researchers to use useful information from the web.

To this end, the purpose of this chapter is to introduce a framework on how web mining is applied for the purpose of building a growth prediction model for SMEs. The proposed framework is based on the concepts for knowledge discovery in databases proposed by Fayyad et al. (Fayyad et al., 1996), modified such that the framework can be used as an overall guideline for web mining based SMEs growth studies. This framework is composed of several elements. Features of each element are explored and implementation techniques are presented.

3.1 Research Framework

In order to explore web data for the purpose of SMEs growth prediction, we construct a research framework consisting of 10 elements, as shown in Figure 3. As mentioned in Chapter 2.3, web mining is closely related to data mining. Thus, the proposed web mining framework also contains elements which also apply to data mining. Elements specific to web mining only are marked by an asterisk. Furthermore, it is important to note that the framework is designed for research and not for web monitoring. Thus, the web data collection is conducted once at the beginning, followed by an extensive data exploration from the collected raw data. However, the proposed web mining framework can be adapted with minor changes such that it becomes suitable as a guideline for the creation of a web monitoring system.

In the following, the individual steps are explained:

1. In an initial step, the aims of a research project is defined. Typically, the research objectives are set by determining a set of research questions, which guide the research design (Thabane et al., 2009). The research questions of the present study are presented in Chapter 1.2.
2. Next, a systematic literature search will be conducted to build expertise on the growth mechanism of SMEs. This step is particularly critical to prevent collecting data which are redundant to achieve the stated research objectives. Furthermore, given the vast amount of web data, it is important know which data sources are to be collected, in order to prevent information overload (Petticrew and Roberts, 2006). Chapter 3.2 presents the methodology applied for a systematic literature review.
3. Next, appropriate web data sources should be selected according to the research needs and based on the findings of the literature review. Once the usability of the web data sources are determined, web data are either

downloaded via API (if available) or web crawling is applied for retrieving publicly accessible web documents in large volumes. In the context of this research project, it is important to note that web data are collected to derive the growth-indicating factors, which serve as input features to train a growth prediction model. Chapter 3.3 elaborates the web mining techniques used for collecting data from web data sources.

4. In the context of machine learning, ground truth data denotes labeled data used for model training and evaluation. In the context of the present thesis, ground truth data (i.e. the growth labels of Swiss SMEs) are retrieved from the data provided by a Swiss insurance company, because financial measures such as sales growth of SMEs are not publicly available on the web. Therefore, Chapter 3.4 describes the insurer data used to construct the ground truth data for predictive modelling.
5. In web mining, it is a common practice to collect web data of different types and from different sources. Therefore, a proper data storage is important for the efficient use of data mining across various data sources. In the present thesis, we briefly distinguish between geographical and non-geographical data. Non-geographical data are further divided into structured (such as CSV format) and non-structured data (such as HTML documents). Chapter 3.5 elaborates the data storage methods used in this study.
6. In the next step, data from multiple sources are combined by matching records from different data sources representing the same real-world entity. In order to ensure a high quality of data matching, we adopt the entity identification approaches described by Denk (2009) to combine data provided by the Swiss

insurer with data of various web data sources. Chapter 3.6 provides further insights into the methodology of data sources linkage.

7. This step covers the preparation of the retrieved data for oncoming data mining tasks. Structured information are in the form of numerical data and thus, require minimal data preparation. However, unstructured data such as web documents in the form of HTML files must be first transferred into a unified and structured representation in order to gain useful insights for data mining and predictive modeling. In Chapter 3.7, the focus lies on the preprocessing methods of web documents, which include: preprocessing of HTML files, extraction of structured data, data cleaning and the application of text mining techniques.
8. In this step, the preprocessed information from the previous stage is converted to numerical and categorical values as input for machine learning algorithms. Moreover, because web data are often imperfect, the generated input features for machine learning are incomplete. Thus, additional steps such as feature imputation are conducted to handle missing values and to optimize the model training. Chapter 3.8 elaborates common practices in feature engineering.
9. The next step comprises operation related to machine learning, such as selecting machine learning algorithms and appropriate model parameters for pattern analysis. A wide variety of data mining algorithms exists and therefore, it is crucial to choose algorithms in accordance with the research objectives. The present research focuses on the prediction of SMEs growth using classification algorithms. Thus, Chapter 3.9 elaborates classification algorithms, which are widely used in various research domains: logistic regressions, random forests and artificial neural networks.

10. In the last step in the web mining process, discovered knowledge is consolidated. This involves the interpretation and visualization of the extracted patterns/models, answering the previously stated research questions and outlining the implications and future work. Chapter 3.10 describes the most common methods for visualization and evaluation of prediction models.

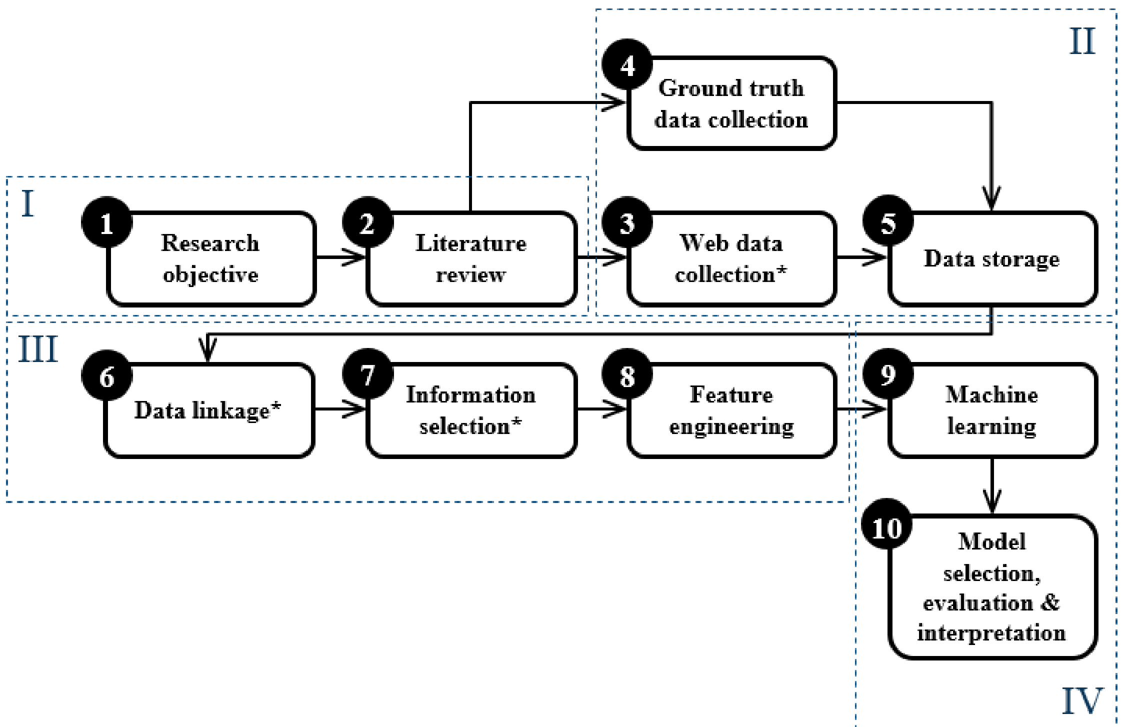


Figure 3: Web mining framework for SMEs growth prediction. Elements specific to web mining only are marked by an asterisk. The framework elements can be broadly grouped into categories: (I) Goal setting, (II) data collection, (III) data preparation and (IV) modeling and implication.

3.2 Systematic Literature Review Methodology

According to Fink (2005), a rigorous literature review must systematically pursue a methodological approach that (1) explicitly explains the procedures by which it was conducted, (2) comprehensively includes all relevant materials in its scope, and (3) is reproducible by others who would follow the same approach in reviewing the topic. When a literature review is conducted using a systematic, rigorous standard, it is called a systematic literature review.

Systematic literature reviews are conducted for a variety of purposes. They include providing a theoretical background for subsequent research; learning the breadth of research on a topic of interest; or answering practical questions by understanding what existing research has to say on the matter. Given the vast amount of publicly available data on the web, conducting a proper research review on the subject of interest before collecting web data is crucial to increase the web mining efficiency and to prevent information overload.

Different kinds of systematic literature review exist (Bero et al., 1998). Here, we follow the guidelines proposed by Okoli and Schabram (2010), modified to suit the needs of the present research. Figure 4 presents the procedures for building domain knowledge as a preparation for the oncoming web data collection.

3.2.1 Information Sources

In the first step, sources of information needs to be identified. Traditional information sources include books, journal articles, and already published literature reviews. Traditionally, these were accessed mainly by lengthy visits to libraries, but today these sources are widely available on the Internet via electronic databases, which are now the predominant source of literature collection (Okoli and Schabram, 2010). Open access databases such as Google Scholar and specific subject databases (such as ProQuest, Scopus, EBSCO, IEEE Xplore and the ACM Digital Library) offer electronic access to most published literature (Norris et al., 2008). Consistent with the work by Okoli and Schrabam (2010), we propose including the top ten most important journals in the research field of interest, by utilizing bibliometric indicators such as Scimago Journal & Country Rank (SJR).

3.2.2 Search Strategy

Next, a set of key terms describing the research of interest are defined to conduct the literature search, in order to assure that the results obtained are comprehensive and reproducible. The keywords should be defined in such a way that they leave the reviewers a large but manageable number of articles for further examination. Typically, key terms are defined separately for the title and abstract search (Okoli and Schabram, 2010). Thereby, it is crucial to understand the correct use of Boolean operators to take particular advantage of these databases (Fink, 2005). Furthermore,

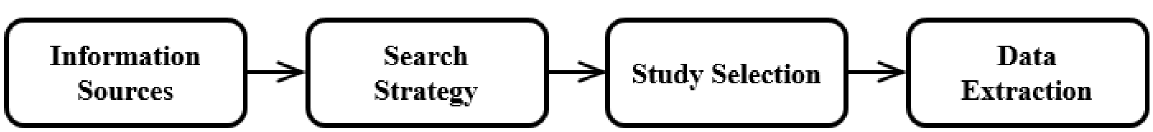


Figure 4: Procedures of a systematic literature review.

search criteria may include the article language (e.g. only English articles), type of publications (e.g. peer-reviewed conference proceedings or only journals) and date range (e.g. only articles after a certain date).

3.2.3 Study Selection

In this step, stricter criteria are defined for articles to be further considered in the review process. As suggested by Okoli and Schrabam (2010), the eligibility of studies are assessed by reviewing the abstracts of the articles identified by the search strategy. Full texts are additionally screened when necessary. As each systematic review varies depending on the review objective, a definitive guide to conducting the eligibility check does not exist. Typically, a standard form should be developed to employ in assessing each article. For instance, Fink (2005) proposes a form in which eligibility criterion is phrased as a yes or no answer. If an article does not meet one of the predefined criteria, then the assessment is finished and the article is excluded.

3.2.4 Data Extraction

This step represents a crucial phase in the systematic review procedure. At this point, the reviewers are left with a complete list of articles that will comprise the material for the final systematic review. Information are manually and systematically extracted from each article based on the objective of the literature review, such as the identification of key determinants of SMEs growth. Finally, the extracted information is summarized, which serves as knowledge base for the oncoming web mining tasks.

3.3 Web Data Collection

The web can be broadly subdivided into web documents and web applications. Web documents annotate a virtual unit located on the web which carries information, whereas web applications are interfaces designed to let people perform activities, tasks and requests online and do not explicitly carry information (Lewandowski, 2005). Hence, web applications are not of value for web mining purposes and the rest of the section focuses on web documents for web data collection. First, the selection of web data sources is discussed, followed by a brief explanation of the accessibility to web documents. Subsequently, we elaborate the principle of web crawling which is the primary data collection method applied in this work.

3.3.1 Selection of Web Data Sources

The most critical decision to be taken concerns the selection of sources which shall be mined. Common categories of Web sources are blogs, newsgroups, message boards, forums, news websites, business websites, portal sites, governmental sites and social platforms. After the domain knowledge is developed by conducting a literature review as explained in Chapter 3.2, the usability of various web data sources is manually inspected with respect to the following selection criteria:

- Information coverage, i.e. how much of the information to be gathered is covered by the source?
- Information completeness, i.e. how complete are the data provided by the source?
- Ease of web data collection, i.e. does the source provide an API for data collection, or facilitate the use of web crawlers?
- Ease of information extraction from web documents, i.e. is there a clear and comprehensive structure in the presentation of the information?
- Structural stability of the web data source, i.e. does the structure of the web data source frequently changing in time?
- Legal aspects of the collection of web documents, i.e. does the web data source explicitly prohibit the use of crawling techniques for data collection?

The last criterion is particularly important as web mining may pose a threat to important legal and ethical values (Van Wel and Royakkers, 2004; Velásquez, 2013). The present thesis focuses on the technological possibilities of web mining and the collection of publicly accessible web data for SMEs growth modeling without further consideration of regulatory and ethical aspects. However, a general discussion on the regulatory and ethical aspects can be found in Chapter 5.3.

Further, the selection can be restricted by only considering sources of a certain language or of a specific top-level domain. Another important consideration is, whether a closed- or an open set of sources shall be mined. Examples for closed-sets are the set of all pages belonging to a single website or a set consisting of a pre-defined list of URLs. Open Sets are not bound to the number of documents and may cover the entire Internet. While closed sets can leave out important sources, open sets are usually quite expensive in computing power and time.

3.3.2 Accessibility of web documents

Not all documents on the web are publicly and directly accessible. This must be of concern since web mining retrieves documents by automated means and might encounter other problems than human surfers do. In general, three categories classifying the accessibility of web documents are described. First, publicly accessible documents are available for everybody. The client sends a response to the server and obtains the documents, no matter if the client is a human or a program. Second, documents can be accessible through web forms which require inputs from a web interface. Generally speaking, the document is dynamically generated on demand according to input values. Third, documents accessible through application programming interfaces (API) are especially designed for automated retrieval. Such websites make their content available for automated querying, omitting the laborious task of using web forms. The results are generally returned in a structured form of XML-format (Web, 2007).

Here, the focus lies on the collection of publicly available and accessible web data which are either collected using web crawling or through web form submission (see Chapter 3.3.3). Web data collection via APIs is not considered due to the many limitations that often come with it. Many APIs require authentication, or have a restricted number of requests to be submitted per day (Mayr and Tosques, 2005). Nevertheless, it is important to note, that whenever it is feasible and the selection criteria for web data sources are fulfilled (Chapter 3.3.1), collecting web data using API should be the preferred method since the effort for data processing and transformation can be significantly reduced due to the structured format of the retrieved data.

3.3.3 Web Data Collection Methods

Web data collection methods need to match the different categories for accessing web content as outlined in Chapter 3.3.2. Further, they have to align with the selection of web data sources as elaborated in Chapter 3.3.1. In general, there are three methods for the acquisition of web data: (1) web crawling for retrieving publicly accessible web documents, (2) web interface submissions for retrieving web database content and (3) API integration for obtaining web data. The present work focuses on the use of the first two methods due to the limitations of APIs (see Chapter 3.3.2) and thus, this section elaborates web crawling and web form submission for web data retrieval. The explanation of APIs for data gathering can be found elsewhere, e.g. Lomborg and Bechmann (2014). Further, it is important to note that these methods yield unstructured results in the form of web documents (i.e. HTML-files). Thus, text mining must be applied for information extraction (see Chapter 3.7).

Web crawling: A web crawler must fulfill the function of automatically downloading web documents. The basic functionality of a crawler includes requesting, fetching and storing of web documents as well as automated redirection to the documents next to be retrieved (Liu, 2007). In general, a crawler requires three specifications (Liu, 2007). First, an initial set of URLs (known as seed pages) has to be elaborated as starting point for the crawling job. Second, conditions need to be provided whether the crawler should download a page or not. Third, the crawler requires instructions on how to continue, meaning which pages are lined in queue for the next visit. Different types of web crawling exist depending on the crawling modes. If a closed set of web documents shall be retrieved (see Chapter 3.3.1), the tasks is best performed by using a web scraper. If the documents to be obtained can not be specified in advance but can be described by common properties, a preferential crawler is the best solution. Finally, a universal crawler serves whenever it is not possible to make any specification regarding the source selection. In the following, all three types are briefly elaborated:

- A web scraper is a crawler in its most basic form. The URLs of all Web documents to be retrieved will be provided as initial set. The entire set is then sequentially processed by fetching and storing each document in the queue. Frequently, the URL-structure of pages from the same website is highly similar and varies only by certain characters. In this case it should suffice to define a pattern for URLs encompassing all documents to be retrieved.
- A universal crawler is commonly employed by search-engines for creating exhaustive, topic-independent indices of the web (Liu, 2007). The crawling process starts from a series of seed URLs and continues by tracking all links extracted from the retrieved documents (Markov and Larose, 2007). Consequently, the crawling job will progress arbitrarily in all directions and is able to capture the entire Internet to its maximum extent (Chakrabarti et al., 2002). Due to the limitation of computing resources and storage capacity, criteria for terminating the crawl must be defined, such as the specification of

a maximum number of page views, a maximum storage space or a maximum depth of linked pages to which the retrieval can descend (Liu, 2007). To increase scalability and reduce bottlenecks, crawling tasks can be executed simultaneously (Liu, 2007).

- Preferential crawlers are similar to universal crawlers, but the documents to be retrieved are preselected according to certain criteria. There are two approaches: Either the crawler can be modified in a certain direction by refining the selection of the links to be followed, or the relevance of the content of the documents is assessed before retrieval. For the first approach, crawling is driven by heuristics, such as the assumption that pages linked to each other are more similar in content or subject matter, or that the text surrounding a link can specify its subject, or that the text surrounding a link can specify its subject matter (Pant and Menczer, 2002). The second approach builds on the classification for deciding whether document should be retrieved or not (Liu, 2007).

Automated web interface submissions: Web crawlers are incapable of capturing dynamic web documents, which are produced on demand in relation to certain input values (Markov and Larose, 2007). Examples are result pages from search engines or content from web data repositories, such as archives, patent databases and digital libraries. These documents can only be accessed by a request via web form. The retrieval of this web data requires that an agent makes requests automatically. This can be realized through identification and replication of the essential interface elements. The essential elements of an interface are the input fields for data and the submission method, which both can be extracted from the HTML-code (Schrenk, 2012).

3.4 Ground Truth Data Collection

Ground truth data is a term used in various fields to refer to information provided by direct observation. In data mining, it is used to train and validate supervised machine learning techniques. Given the broad range of ground truth data types, providing a uniform guideline on how ground truth data is collected is not possible. Thus, in the following, we elaborate the collection of the ground truth data related to the primary interest of the present work, which is the growth prediction of SMEs. As discussed in Chapter 2.2.2, the adopted definition of growth is the growth of revenue due to its importance to the economy (Lev and Radhakrishan, 2010). Thus, revenue data from SMEs are collected as a ground truth for growth modeling.

Financial data of SMEs are highly sensitive to privacy issues and thus, are usually not publicly available in the web (Ozgulbas and Koyuncugil, 2012). In this thesis, ground truth data are retrieved from the data provided by a large Swiss insurer, which consists of SMEs' firm name, business type and the annual revenue in the period from 2010-2017. Furthermore, the data contain other basic information such as the business type, address, location and a unique business identification number (UID), which is assigned by the Swiss Federal Statistical Office to facilitate the corporation between the government and firms. This information is particularly useful to link corporate data with web data, as explained in Chapter 3.5. It is important to note that once the data linkage is completed, the data will be completely anonymized for further analysis in order to protect data privacy.

3.5 Data Storage

Data collected in the present thesis can briefly grouped into geographical and non-geographical data. Non-geographical data are further divided into structured (such as CSV format) and non-structured data (such as HTML documents). Therefore, two types of databases are used to securely store the collected web data: (1) PostgreSQL for geographical data such as data downloaded from Openstreetmap (PostgreSQL), and (2) ElasticSearch for non-geographical data including structured and unstructured web documents (ElasticSearch).

3.5.1 PostgreSQL

PostgreSQL is a powerful, free and open-source object-relational database system that uses and extends the SQL language combined with many features to safely store and scale the most complicated data workloads, capable of handling terabytes of data. The origins of PostgreSQL date back to 1986 as part of the POSTGRES project at the University of California at Berkeley and has more than 30 years of active development on the core platform. PostgreSQL runs on all major operating systems and has powerful add-ons such as the popular PostGIS geospatial database extender to facilitate spatial analysis (PostgreSQL). Therefore, PostgreSQL has become the open source relational database of choice for many researchers and practitioners. In this work, PostgreSQL is not only used as a tool to securely store geographical data in large scale, but also to facilitate spatial analysis, as further elaborated in the case study (Section 4).

3.5.2 ElasticSearch

ElasticSearch is a real time distributed analytics tool mainly designed to store and organize unstructured data in order to make it easily accessible (ElasticSearch). It is a distributed document store with strong full-text search capabilities, which stores all objects in JSON documents. These documents are indexed by default and are schema free, so that fields don't need to be defined for data types before adding data (Gupta and Rani, 2016), facilitating the storage of data from multiple data sources of prior unknown data structure. In this work, ElasticSearch is primarily used to store collected structured and unstructured web documents in large volumes.

3.6 Data Linkage

Data quality management is a crucial challenge in database management aiming at an improved usability and reliability of the data. Entity identification is defined as the detection and merging of two or more records representing the same real-world identity across multiple data sets, which is relevant in duplicate detection and elimination as well as data integration. Apart from data cleaning, data integration and data warehousing, entity identification is closely related to information retrieval, pattern recognition and data mining as well, thus, making use of ideas from several research areas (e.g. Bilenko et al., 2003). With the tremendous growth of web data sources, entity identification became an important issue in data warehousing (Aizawa and Oyama 2005).

A variety of data linking methods are available (Winkler, 2006). In the present work, we adopt and modify the data linkage method described by Denk (2009) to combine

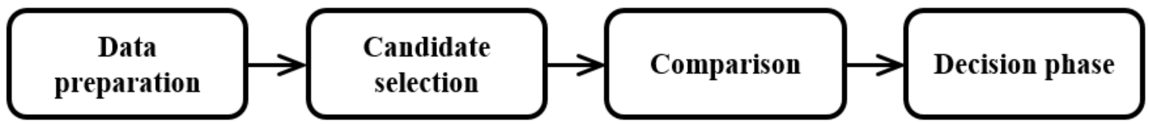


Figure 5: Data linkage procedures.

data provided by the Swiss insurer with data from various web data sources. The proposed data linkage method is a semi-automated rule- and knowledge-based method, which offers a high degree of flexibility and tuning possibilities, resulting in good performance for entity matching (Denk, 2009). Figure 5 illustrates the entity linkage procedure applied in this work.

3.6.1 Data Preparation

The first step of entity identification is the data preparation phase, encompassing different transformations of common variables to obtain comparable variables suitable for usage in the further identification process (Denk, 2009). In particular, string variables, such as names and addresses have to be pre-processed to be comparable among data sets, but also simple calculations can be necessary to derive matching variables, for example age determined from date of birth. Typically, standardization and parsing are required in case of string variables.

- Standardization is synonymous with the conversion of values into a uniform format, which includes the conversion of characters to lower- (or upper-) case, expansion of abbreviations, and the removal of language accents, punctuations and common words.

- Parsing is the process of splitting a string variable into a common set of components that are more comparable, such as dividing an address into zip code, city, street, and number. Examples of parsing criteria are spaces and hyphens.

3.6.2 Candidate Selection

This step includes a method for fast and computationally favorable filtering of data set pairs with negligible probability of containing data sets representing the same entity (Denk, 2009). In general, a detailed comparison with respect to all available matching variables is very time-consuming. Especially for large data sets, the selection of candidate record pairs with higher likelihood of belonging to the set of true matches is necessary to reduce the number of pairs that undergo the subsequent comparison of matching variables.

Blocking is a common approach to reduce the number of data pairs. Thereby, the set of all possible record pairs is subdivided into blocks agreeing on a specified blocking key. Only record pairs within these blocks are further analyzed, whereas the (usually larger) residual set of pairs are discarded. The best blocking variables have a high number of categories, high reliability and low error rates. Variables often used for blocking are regional classifications such as zip code (Fellegi and Sunter, 1969).

3.6.3 Comparison

In this step, similarity measures are used to assess the degree of similarity of the candidate pairs (Denk, 2009). Similarity measures are provided for different types of variables. For numerical variables, binary outcomes discerning agreement and disagreement, or tolerance limits (e.g. age difference of plus or minus one year) can be used to identify possible matches. For string variables, string comparator is a common approach to assess the similarity between two entities. String comparators are mappings from a pair of strings to the interval [0, 1], which measure the degree of similarity of the compared strings (Winkler, 1990). An example of string comparator is the edit distance (Marzal and Vidal, 1993). Its basic idea is that any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and other operations. The smallest number of such operations required to change one string into another is a measure of the difference between them.

Typically, string comparators are applied for each matching variable, such as firm name, address and location. Subsequently, the similarity measures are averaged and the candidate pairs are ranked according to their degree of similarity for the decision phase.

3.6.4 Decision Phase

In the last step, the candidate pairs are manually inspected to ensure a high data quality. Starting with the pair with highest similarity ranking, the final decision on entity matching is taken based on our knowledge and expertise (Denk, 2009). Thereby, manually looking into additional information about the entity on the web (e.g. social media, mentioning in news article) can be helpful at identifying true matches. Note, that only one among multiple candidate pairs are chosen or all pairs will be discarded to ensure a high data quality for model building.

3.7 Information Selection

This chapter elaborates the techniques used to extract useful information from the unstructured web documents. Due to the HTML syntax of web documents, pre-processing techniques particularly to web documents need to be applied before conventional text mining methods can be employed. Thus, we first elaborate the characteristics of web documents, followed by common methods used to preprocess web documents, which are the web document preprocessing (Chapter 3.7.2) and structured data extraction (Chapter 3.7.3). Subsequently, we describe conventional text mining techniques applied to prepare the data for oncoming feature engineering and machine learning tasks.

3.7.1 Characteristics of Web Documents

Web documents are very heterogeneous and differ in various ways: the format in which the document is displayed, the type of information it contains, the extent to which it is structured and whether it contains metadata. In general, a Web document can be divided into three levels. These layers are content, structure and layout (Balzert, 2007). The content refers to the actual information provided by the document, in the form of textual, numerical or visual data. Structure corresponds to the organization of the document and includes links, paragraphs, headings and elements of visual communication such as lists or tables. Layout describes the style

```
<!DOCTYPE html>
<html>
<head>
  <title>Page Title</title>
</head>

<body>
  <h1>This is a Heading</h1>
  <p> This is a paragraph </p>
  <a href="URL">This is a link</a>
  <img src="image.jpg" alt="This is an image">

  <ul>
    <li>List item 1</li>
    <li>List item 2</li>
    <li>List item 3</li>
  </ul>

  <table>
    <tr>
      <th>Table element 1</th>
      <td>Table element 2</td>
    </tr>
  </table>

<body>
</html>
```

Figure 6: Example of a HTML code with tags. Tags descriptions are in the code lines.

and visual representation of the document and determines the size, position, color and font of the structural part. Web documents optionally contain metadata that provides information about the document itself. Figure 6 shows an example of a HTML code. The description of the tags can be found in the code lines.

In the past, documents were individual HTML files whose content, structure and layout were inseparably linked. Today, there are many intelligent solutions to separate the three dimensions that facilitate the creation and revision of web documents (Balzert, 2007). The most common approach is to specify content and structure as HTML files, while defining the layout separately as CSS files. A complete separation is achieved by specifying the content as an XML file, the structure as an XSL file and the layout as a CSS file. Table 1 gives an overview of the most commonly used web document formats today and their compatibility with the dimensions mentioned above (marked with a cross):

3.7.2 Web Document Preprocessing

Preprocessing on the web document level needs to be applied in order to facilitate the extraction of relevant textual content. Typical tasks to be performed are HTML normalization and HTML tag removal, which are briefly elaborated in the following.

HTML Normalization: Although HTML syntax is standardized (W3C), many browsers interpret HTML code tolerantly and are resistant to syntax errors (Liu, 2007). Therefore, HTML files are often not coded according to standards and often contain inconsistencies, misspelled tags, incorrectly nested tags or missing relevant tags (Liu, 2007). These syntactical errors increase the identification of relevant entities and must be eliminated. Therefore, normalization is applied to recognize such errors and converts the HTML code into the standardized form.

Format	**Description**	**Content**	**Structure**	**Layout**
HTML	HyperText Markup Language	X	X	X
XML	Extensible Markup Language	X		
CSS	Cascading Stylesheet			X
XSL	Extensible Stylesheet Language		X	
Microsoft Office	.doc, .ppt, .xls, etc.	X	X	X
Multimedia	.jpg, .gif, .png, .mp3, etc.	X		

Table 1: Most commonly used web document formats. The crosses mark the compatibility.

HTML Tag Removal: The parts of an HTML file not containing any relevant information can be removed beforehand. They can be identified by the tags surrounding them (see Figure 6). Examples of embedded non-HTML code, such as scripting snippets (<script> tag), Flash objects (<object> tag) or internal CSS elements (<layout> tag) Balzert, 2007). Other classes of tags can be removed if required, e.g. the (<img> tag) if images are not relevant.

3.7.3 Structured Data Extraction

In this step, the relevant text information is extracted from the parsed Web documents. The aim is to distinguish the relevant content blocks from the auxiliary parts such as menus, navigation elements, disclaimers or embedded advertising spots. Examples of relevant content blocks are blog entries, comments, forum posts, data tables or product descriptions. Extraction can be performed by using extraction patterns, also known as wrappers (Liu, 2007).

Wrappers use the HTML structure as an orientation for identifying the interesting content blocks (Sarawagi, 2008). A wrapper must be specifically adapted to the structure of a document and the content of the blocks to be extracted in order provide useful results. Typically, the wrapper requires a predefined set of extraction rules that use the HTML tags as markers (see Chapter 3.7.2). Information is extracted whenever these markers are encountered. The development of the rules and regulations requires explicit knowledge of the document structure. However, it should be noted that rule sets can only be reused for documents with a similar structure. This is usually the case for web documents that belong to the same web data source, since these documents are usually generated by a Content Management

System with similar HTML templates (Jablonski and Meiler, 2002). Nevertheless, separate wrappers have to be developed when information is collected from different web data sources.

3.7.4 Natural Language Text Preprocessing

After the data extraction process described in the previous section, the data now consists mainly of textual and/or numerical entries. While numerical entries are usually presented in a well-structured way (e.g. in the form of tables) and thus require little preprocessing, preprocessing of text entries to generate useful information can be very demanding. Therefore, we briefly elaborate two common text mining techniques, which are applied to text data as preparation for the upcoming feature engineering tasks for machine learning: stopword removal and tokenization.

Stopwords such as articles, prepositions, conjunctions and pronouns are too common to be useful for text analysis (Baeza-Yates and Ribeiro-Neto, 1999). Therefore, elimination is reasonable and can lead to a significant reduction in the size of the text, which only contains essential information for further analysis. In addition, the data needs to be stripped from all other remaining HTML-tags (Sarawagi, 2008).

The tokenization task further splits the text instances into tokens (Feldman and Sanger, 2007). Tokens are obtained by splitting the text along certain separators, such as spaces, commas, quotation, marks or full stops (Sarawagi, 2008).

3.8 Feature Engineering

This chapter discusses the commonly used approaches in feature engineering, which is an important task to optimize the performance of machine learning algorithms. There are a large number of feature engineering methods that cannot be fully covered in this thesis. Thus, we elaborate the methods specifically applied in this work. First, this chapter describes how to deal with missing values in web data. Next, methods to eliminate redundant features are described.

3.8.1 Missing Values in Web Data

The Web is highly unstructured and often very chaotic. Therefore, web data are frequently incomplete, which leads to missing values in web data sets. Despite the frequent occurrence and relevance of the missing data problem, many machine learning algorithms deal with missing data rather naively. However, missing data processing should be treated carefully, as otherwise bias can be introduced into the induced knowledge (Batista and Monard, 2003). In the following, we elaborate five approaches to missing attribute values:

Discarding examples with missing attribute values: This method is the most basic, which consists of discarding all samples that have at least one unknown attribute value (Grzymala-Busse and Hu, 2000).

Discarding attributes with high level of missing values: This method consists of determining the extent of missing data on each attribute, and deleting the attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the upcoming analysis. Relevant attributes should be kept even with a high degree of missing values (Batista and Monard, 2003).

Replacing with most frequent or mean value: These are one of the simplest methods to deal with missing attribute values. The mean value or most frequently occurring value of the attribute is selected to be the value for all the unknown values of the attribute. (Grzymala-Busse and Hu, 2000)

Treating Missing Attribute Values as Special Values: This method uses the unknown attribute values in a completely different approach. Instead of trying to find some known attribute value as its value, the "unknown" itself is treated as a new value for the attributes that contain missing values and treat it in the same way as other values. (Grzymala-Busse and Hu, 2000)

Using prediction models: Prediction models are sophisticated methods for handling missing data. These methods consist of creating a prediction model to estimate values that replace the missing data. The attribute with missing data is used as the class attribute and the remaining attributes as input for the prediction model. An important argument for this approach is that attributes often have relationships (correlations) to each other. In this way, these correlations could be used to create a predictive model for the classification or regression of qualitative and quantitative attributes with missing data (Batista and Monard, 2003). Well known examples of prediction models for handling missing data are CN2 (Clark and Niblett, 1989), C4.5 (Quinlan, 2014) and kNN (Song et al., 2008).

3.8.2 Multicollinearity of Features

Collinearity is the technical term for the situation in which a pair of variables have a significant correlation to each other. Multicollinearity refers to the presence of strong relationships between several variables simultaneously (Kuhn, 2013). The presence of multicollinearity may be due to the scarcity of data samples or is inherent in the investigated problem (Mansfield, 1982). Prediction models derived from such data without a check on multicollinearity may lead to reduced model performance, erroneous analysis and adverse model interpretation (Garg and Tai, 2013). Therefore, the removal of multicollinearity before the application of machine learning algorithms is an important feature engineering task. A variety of approaches exist to tackle multicollinearity, such as principal component analysis or factor analysis (Manly and Alberto, 2016). In this thesis, we follow a heuristic approach proposed by Kuhn (2013), in which the minimum number of predictors is removed to ensure that all pairwise correlations are below a certain threshold. While this method only identify collinearities in two dimensions, it can have a significantly positive effect on model performance (Kuhn, 2013). The algorithm is as follows. First the correlation matrix of all variables are calculated. Next, the two variables associated with the largest absolute pairwise correlation is determined. Subsequently, the average correlation of the two variables in conjunction with the other variables are calculated. Finally, the variable (i.e. one of the two variables) possessing a larger average correlation is removed. These steps are repeated until no absolute correlations are above a given threshold. It is important to note, that a best approach to tackle multicollinearity does not exist in general. The suitability of the methods strongly depends on the size and structure of the data and on the investigated problem. Thus, testing and comparing different approaches is needed to identify the most suitable technique for the investigated problem.

3.9 Supervised Machine Learning for Predictive Modelling

This chapter provides an introduction to a subset of supervised machine learning algorithms which are mainly used in predictive modeling. First, we provide a basic explanation of supervised learning (Bishop, 2006). Next, we elaborate a subclass of machine-learning methods, which are widely used in various business-related research fields: logistic regressions, random forests and artificial neural networks. Finally, we discuss commonly used methods for model optimization.

3.9.1 Supervised Machine Learning

A supervised learning algorithm is a method of creating a mathematical model or function that generates a specific output on a given input. The algorithm derives this model from a training data set, which is a collection of data points (also called "samples" or "examples") consisting of example entries paired with their corresponding outputs or "labels". The process of creating a model from the training set is called training. After this model has been created, it can calculate new output values for new inputs, even for inputs that are not available in the training set. In the remainder of the present thesis, we only focus on classification algorithms, where the output of the model is one of a finite set of discrete labels, also referred to as the set of classes. A more detailed explanation can be found in Bishop (2006).

3.9.2 Logistic Regression

Logistic regression is a widely used statistical modeling technique in which the probability of an outcome is related to a set of independent variable, given by an equation of the form

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

where p denotes the probability of the outcome, β_0 is an intercept term, β_1,..., β_i are the coefficients associated with each independent variable X_1, ... X_i. Index *i* represents a unique subscript denoting each variable. The usual assumption is that the independent variables are related in a linear manner to the log odds ($\log\left[\frac{p}{1-p}\right]$) of the outcome. Logistic regression models use the method of maximum likelihood as their convergence criterion (Oates, 2015). The β coefficients can easily be converted into the corresponding odds ratios by raising the exponential function to the coefficient if variables are represented by a single linear term, resulting in a simple interpretation of the importance of the independent variables (Oates, 2015).

3.9.3 Random Forest

Random Forest classifiers are derived from decision tree models. Decision Trees are non-parametric supervised learning algorithms used for classification and regression. A decision tree is a flowchart-like tree structure consisting of nodes, branches and leafs built to make a classification decision based on a set of feature characteristics. Each node represents a "test" on a feature (e.g. feature: firm age, test: >20 years?); each branch represents the outcome of the test and each leaf node represents a final decision assigning a class label. The paths from the start node (root node) to the leaf nodes represent the classification rules. A decision tree is built from training data by selecting appropriate tests in each of the test nodes. Most training strategies train a tree from top to bottom, first selecting tests that maximize the information gain about the classification. Therefore, the algorithm attempts to find the split that creates subgroups that best distinguish the samples in terms of different class labels. A number of algorithms exist for the construction of decision trees, such as ID3, C4.5, C5.0 and CART (Hastie et al., 2008). However, the simplicity of decision trees has some drawbacks. A single decision tree cannot model complex nonlinear decision patterns because the decision boundaries generated by the test nodes are always parallel to the axes in a feature space representation. Furthermore, decision trees have shown to be unstable when exposed to noise in the data (Patil and Bichkar, 2012). The Random Forest approach improves the stability and accuracy of decision trees by integrating a large number of decision trees into an ensemble classifier. For instance, a Random Forest might contain 500 decision trees, where every decision tree is trained on a so called bootstrapped subsample from the training set. A prediction is obtained by taking the average of predictions from the individual trees (Hastie et al., 2008). The application of this method to decision trees is called bagging, which stands for bootstrap aggregation. It can lead to higher stability and better accuracy. Random Forest further improves bagging by "de-correlating" the trees. This is achieved by taking into account only a small and random subset of characteristics in each tree split. If there are many functions in the data set, this restriction ensures that the individual decision trees are very different from each other. Further, each split within each tree is created based on a random subset of features. The algorithm for creating a random forest is implemented as follows. A predefined number of decision trees is trained. A bootstrap sample is taken from the training set for each tree. This tree is then trained on the bootstrap sample, where only a fixed number of randomly selected features are selected for each split. Predictions can be made from the random forest by feeding a new test observation into all individual decision trees and then averaging their predictions or making a majority decision.

3.9.4 Artificial Neural Network

An artificial neural network is a classifier modeled after how the structure of the human brain functions (Tu, 1996). A human brain contains a huge amount of nerve cells and neurons. Each of these cells is connected to many other cells, creating a very complex network of signal transmission. Each cell collects inputs from all the other nerve cells to which it is connected, and when it reaches a certain threshold, it signals to all cells to which it is connected.

When creating an artificial neural network, this is imitated by using a "perceptron" as the basic unit instead of the neuron. The perceptron can receive and combine multiple weighted inputs. If the combined input exceeds a threshold, the perceptron is activated and an output is sent. Which output it sends is determined by the activation function and is often chosen to be between 0 and 1 or -1 and 1. The equation for a perceptron can be written as

$$y = \Phi \left(\sum_{i=1}^{n} w_i x_i + b \right),$$

where y is the output signal, Φ is the activation function, n is the number of connections to the perceptron, w_i is the weight of the ith connection and x_i is the value of the ith connection. The quantity b represents the threshold (in the scalar case).

The strength of an artificial neural network can be shown by combining several perceptrons and working together. Perceptrons are often organized in layers, with each layer taking inputs from the previous one, applying weights and then, if necessary, sending signals to the next layer. The learning process of an artificial neural network is achieved as follows. First, the weights associated with the connections between the layers are updated. Thereby, several ways exist, and most involve initializing the weights and fed the network a training sample. The error of the network at the output is then calculated and fed back by a process called "back propagation" (Buscema, 1998). This process is then used to update the weights, and by repeatedly using this process, the network can learn to distinguish between several different classes. Figure 7 depicts a graphical representation of an artificial neural network.

3.10 Model Selection, Evaluation and Interpretation

This chapter elaborates the methods used for model selection, evaluation and interpretation. Figure 8 provides a general overview of the complete procedure, which consists of four parts. (1) We address the generalization performance by discussing train/test split and the need of performing multiple runs of modeling which is denoted as n repeats in Figure 8 (Chapter 3.10.1). (2) We discuss a hyper-parameter optimization technique, which includes the methods k-fold cross-validation for model selection and random search for hyper-parameter selection (Chapter 3.10.2). (3) We elaborate how the model performance is optimized by leveraging the receiver operating characteristics (ROC) curve and Youden's index (Chapter 3.10.3). Finally, (4) we elaborate the model interpretation, which includes the explanation of the performance measures used in the present thesis (Chapter 3.10.4). It is important to note, that this section is limited to the subclass of models described in Chapter 3.9. Furthermore, we restrict this section to the binary classification task, which is the primary objective of the present thesis.

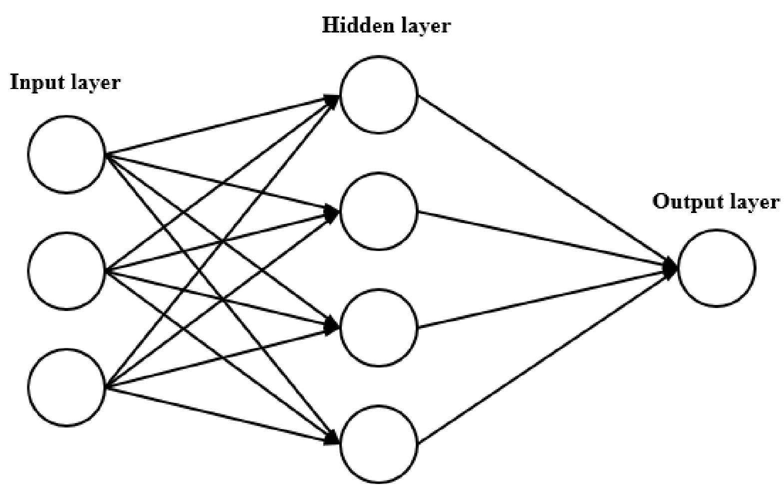


Figure 7: A graphical representation of an artificial neural network with one hidden layer.

3.10.1 Generalization Performance

The common objective of any supervised machine learning algorithm is to model the relationship between inputs and outputs in the training set in a way that allows generalization, or to generate meaningful results for new inputs not included in the training data, also called generalization performance (Bishop, 2006). The main focus on generalization performance is to use standard metrics to evaluate the suitability of an algorithm for modeling a particular dataset. Unless one knows the identity of all future inputs to a model and their correct labels, the generalization performance must be estimated from the available data. For example, using the accuracy of the training dataset of the model is a poor estimate because it can assign a too favorable rating to a model which is "overfitted" and poor at generalizing. Therefore, in the initial step, the dataset is split into a training and test set. The training set are used for hyper-parameter tuning and model training, while the test data set is used to report models' performance. Furthermore, in order to reduce the variance due to the

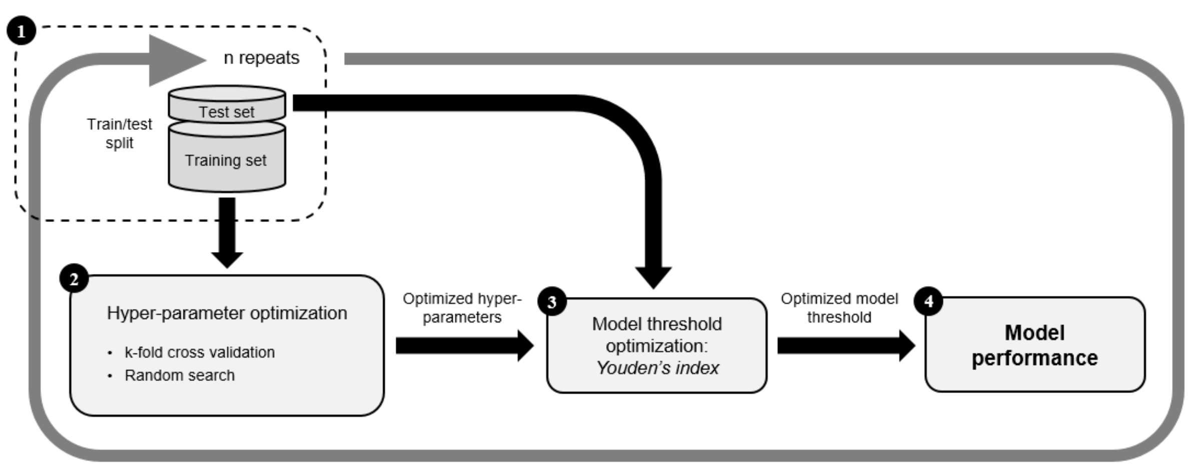


Figure 8: An overview of the model selection, evaluation and interpretation procedures.

training-test split, and to obtain reliable performance estimation for model comparison, we repeated the aforementioned procedure multiple times (denoted as n repeats in Figure 8). Therefore, the dataset is successively split into training and test set, and the proposed procedure is executed multiple times (Kim, 2009). Thereby, the dataset is reshuffled before each round, and the average performance of the models is reported.

3.10.2 Hyper-parameter Optimization Method

To optimize models' hyper-parameters, a random search is conducted to find the optimal value for the hyper-parameters (Bergstra and Bengio, 2012). Thereby, the random search approach queries a given amount of combinations of hyper-parameters at random, where each hyper-parameter consists of a continuous, uniform distribution with pre-defined lower and upper limits (Bergstra and Bengio, 2012). Several hyper-parameter tuning methods exist, with grid search being the most widely used method. In the present thesis, Random search was chosen over the standard grid search method due to the reduced computational time while producing comparative results (Bergstra and Bengio, 2012).

Furthermore, in order to validate the optimized classifiers to the training set, a k-fold cross-validation procedure is applied for model selection. In a k-fold cross-validation (CV), the original sample is partitioned into 10 subsamples while maintaining the ratio of the classes in the target variable. Of the k subsamples, a single subsample is retained as the validation data for testing the model, while the remaining 9 subsamples are used as training data. The CV process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds is averaged to produce a single performance estimation on the training set for model selection (Kohavi, 1995). Finally, the performance of the final model is reported on the test set.

3.10.3 Model Performance Optimization using ROC-Curve

This section elaborates the use of receiver operating characteristic (ROC) curve to optimize the performance measures for machine learning algorithms covered in Section 3.9. The ROC curve is a popular graphical method of displaying the discriminatory accuracy of a binary classifier for distinguishing between two classes. It is widely used in many research domains (Bradley, 1997).

The ROC curve is a plot of the "sensitivity" versus "1-specificity" over all possible threshold values of the classifier. Figure 9 depicts exemplary two ROC curves. A perfect classifier has an ROC curve starting from the origin, going straight to (0,1), then turning right at ninety degrees and ending at (1,1). However that is the ideal case, which often cannot be achieved, as illustrated by the sub-optimal ROC curve. The curve clearly demonstrates the trade-off relationship between sensitivity and specificity, where each point on the ROC curve denotes a cut-off point for the classifier. Thus, choosing a wise cut-off point is crucial to reporting optimized performance measures of a classifier.

Several methods have been proposed to choose the optimal cut-off points (Steinhauser, 2016). Here, we focus on the Youden Index method, which is widely used in many research areas (Fluss et al., 2005). The Youden index is defined as follows (Youden, 1950):

$$J = \max_{c}\{sensitivity(c) + specificity(c) - 1\}$$

$$= sensitivity\ (c_0) + specificity(c_0) - 1,$$

where J ranges between 0 and 1. J = 0 indicates that the classifier has no discriminating ability and J = 1 indicates a perfect classifier (Fluss et al., 2005). c_0 is the optimal cut-off point. From a graphical perspective, Youden's Index is the maximum vertical distance between the ROC curve and the imaginary diagonal chance line from (0,0) to (1,1). To summarize, it is important to report model performance measures such as accuracy, sensitivity and specificity at the optimized cut-off point.

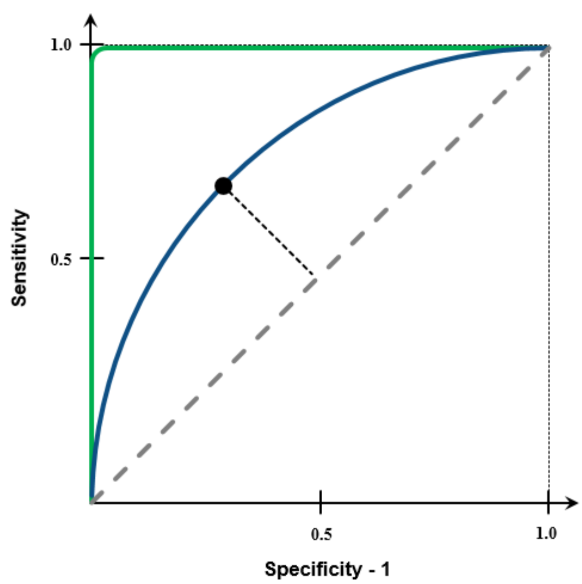


Figure 9: ROC curve of a perfect classifier (green) and sub-optimal classifier (blue). The grey dashed line denotes the random line. The black dot represents the values of sensitivity and specificity at the optimal cut-off point.

3.10.4 Performance Measures

The interpretation of the models represents the most important final step in the presented web mining framework. Here, we discuss the performance measures used in the present thesis.

When validating the performance of a classifier on the test set, the predicted outcome produced by the classifier are counts of the correct and incorrect classifications from each class. This information is commonly displayed in a confusion matrix. A confusion matrix is a form of contingency table indicating the differences between the true and predicted classes for a series of labelled samples, as shown in Figure 10 for a binary classification case. In Figure 10, *TP* and *TN* are the number of true positives and true negatives respectively, whereas *FP* and *FN* are the numbers of false positives and false negatives respectively. The row totals, (*TP+FP*) and (*FN+TN*), are the number of predicted negative and positive examples, whereas the column totals, (*TP+FN*) and (*FP+TN*), are the number of truly negative and positive

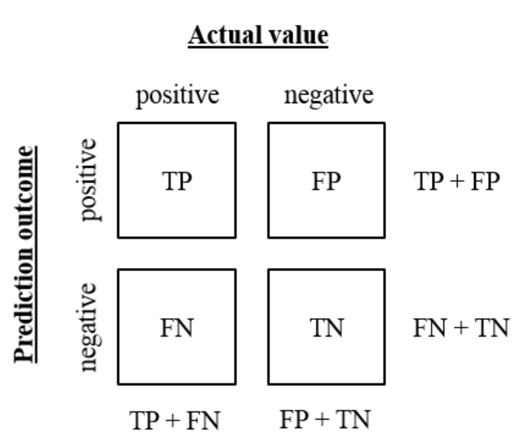


Figure 10: A confusion matrix.

examples. The confusion matrix shows all of the information about the classifier's performance. However, more meaningful performance measures can be derived from the confusion matrix, which are introduced in the following:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

Accuracy refers to the overall percentage correctly classified. Sensitivity refers to the probability that a true positive sample is correctly classified into true positive group, whereas the specificity refers to the probability that a true negative sample is correctly grouped into the true negative group. It is important to note, that these performance measures are functions of one particular cut-off point of the classifier, which in the optimal case is estimated by maximizing the Youden index (see Chapter 3.10.3).

In addition, we make use of the area under the ROC curve (AUC) measure, which is obtained by summarizing the information of the ROC curve into a single global value using the trapezoidal integration (Purves, 1992). AUC has been shown to exhibit a number of desirable properties as a classification measure when compared to the overall accuracy, such as increased sensitivity in analysis of variance (ANOVA) tests, decision threshold independent, and invariant to a priori class probabilities (O'brien, 1979). Therefore, AUC is the preferred choice for comparing the performance of different classification schemes (Bradley, 1997).

4 Case Study: Predicting the Growth of Restaurants using Web Mining

This chapter presents a practical case study investigating the application of the proposed web mining framework to complement the theoretical elaboration explained in Chapter 3. Furthermore, the proposed method can be considered as a novel approach for the restaurant growth research which has not been considered so far.

The remainder of this chapter employs the research framework outlined in Chapter 3 and is structured as follows. Chapter 4.1 provides a short introduction to the present case study. Chapter 4.2 provide an overview of the previous studies in restaurant growth modeling. In Chapter 4.3, a systematic literature review on growth factors is conducted, which will serve as a groundwork to collect growth-relevant information from the web. Chapter 4.4 elaborates the ground truth data used in this case study. In Chapter 4.5, web data collection methods are applied to automatically collect and store data from various web data sources, and information extraction methods are applied to extract growth factors from various web data sources. Thereby, web data collection, data storage and information extraction are grouped by web data source for better readability. Subsequently, Chapter 4.6 explains how the different data sources are interlinked. Chapter 4.7 discusses the label creation and feature engineering task. In Chapter 4.8, we build and compare different binary classification models using supervised machine learning algorithms. More specifically, the developed models classify a restaurant either in a non-growing or growing restaurant. The algorithms which have been considered include logistic regressions, random forests and artificial neural networks, which share a predominant role in a range of research domains. Further, the results are presented and the findings are discussed. Finally, Chapter 4.10 concludes with a summary and an outlook on future research in the field of restaurant growth modeling using web mining.

4.1 Introduction

The gastronomy industry play an important role in the economy of many countries. Especially in Switzerland the gastronomy industry is particularly relevant, as 10% of all jobs in small and medium enterprises are created by gastronomy, acting as the countries backbone for growth (Gastrosuisse, 2017). However, existing studies show that the gastronomy industry is facing many tough challenges because of the recent economic turmoil (Neuman, 2009; Gastrosuisse, 2017). Only one-third of all Swiss restaurants generate an appropriate income in order to maintain their existence and expanding their business. Moreover, the study conducted by GastroSuisse (2017) revealed that the sales performance of the gastronomy industry has been dropping continuously over the past eight years, highlighting the urgent need to counteract the negative trend.

Given the importance of the gastronomy industry to the Swiss economy, researchers and academics have been analyzing factors influencing the risk and growth of restaurants, and developing models to anticipate restaurant failure and bankruptcy for many decades (Dimitras et al., 1996).

With the emergence of data mining in the gastronomy research field, researchers recently turned their focus on applying data mining techniques for restaurant failure and bankruptcy prediction (Kim and Upneja, 2014). However, these prediction models only include few data types such as financial or operational data and thus cannot explain the whole and complex context of restaurant growth (Kim and Upneja, 2014). Moreover, conventional data collection is primarily conducted via questionnaire studies, which is very laborious and time-consuming, or provided by financial institutes, thus highly sensitive to privacy issues. Furthermore, data mining techniques such as artificial neural network and random forest are extensively studied with a strong focus on the prediction of bankruptcy rather than growth of restaurants. Although numerous studies has attempted to explain the growth of restaurants, studies reporting data mining based restaurant growth models cannot be identified.

Simultaneously, web mining has emerged as an important approach to obtain valuable business insights from the web, as enterprises post increasing information about their business activities on websites. In particular, restaurants post their publicly-viewable information on their website and online platforms for various reasons, including promoting their food, presenting their facility and expanding their customer base, with the goal to outperform their competition and increase the sales performance. Furthermore, the web also contain valuable information about the firm's location, specifications of products and services offered, key personnel, and strategies and relationships with other firms. Thus, the web can be viewed as a huge and ever-growing database containing valuable business-related information, which is readily and publicly available, cost-effective to obtain, and extensive in terms of coverage and the amount of data contained

Web mining has shown to be very useful for a wide range of business-oriented applications (see Chapter 2.1.2). In particular, web mining has proven to be very valuable for e-commerce, where any information related to consumer behavior are extremely valuable to anticipate and increase the sales performance (Patel et al., 2011). However, web mining has been barely used in the research of the hospitality industry (Kong et al.). Considering the vast and increasing amount of data freely available online, web mining bears a great potential in revealing valuable information hidden in web, which can be further used to study the growth of restaurants. In the present case study, we present how web mining can be utilized to leverage growth modeling for restaurants by following the web mining framework presented in Chapter 3.

4.2 Related Work in Restaurant Growth Prediction

4.2.1 Definition of Growth in Restaurant Growth Studies

Growth is considered to be one of the key benchmarks of success by practitioners in the restaurant industry. However, there is no consistency in the dimension of growth which theorists have used as the object of analysis. Different definitions have been used in the studies that attempted to explain the growth of restaurants. Non-financial growth measures include growth of employment, customer satisfaction and loyalty (Brown and Mitchell, 1993). Financial growth measures include growth of revenues and profits (Cho et al., 2006). In this study, the adopted definition of growth is the growth of revenue, due to its importance to the economy (Lev and Radhakrishan, 2010).

4.2.2 Survey of Prediction studies for Restaurants

For the gastronomy industry, there is not much documented bankruptcy prediction research, and even less for growth prediction (Kim and Gu, 2006). More surprisingly, we were not able to identify restaurant growth studies utilizing to web mining to the best of our knowledge. Thus, we provide a general overview of bankruptcy prediction studies in the gastronomy industry. Olsen et al. (1983) first attempted to predict business failure in the restaurant industry. In their study, 7 failed restaurant firms were compared with 12 non-failed, using a graph analysis of financial ratios rather than sophisticated models. Later, Multivariate Discriminant Analysis (MDA) and logit analyses have become popular tools for financial distress prediction (Dimitras et al., 1996). Using logistic regression analysis, Cho (1994) extensively investigated business failure in the hospitality industry. Defining failure as a firm with 3 or more years of consecutive negative net income, he developed logistic regression models for predicting restaurant and hotel failures, respectively. Gu and Gao (2000) predicted business failure of hospitality firms by using financial ratios and multivariate discriminant analysis (MDA). They developed a failure prediction model for hospitality firms using a combined sample of hotels and restaurants that went bankrupt between 1987 and 1996. However, these methods suffer from the disadvantages associated with parametric and distribution-dependent approaches (Dragos et al., 2008). Drawbacks to MDA are the assumptions of normally distributed independent variables Balcaen and Ooghe, 2006), whereas the shortcomings of logit analysis are the assumptions of the variation homogeneity of data (Lee et al., 2006) and the sensitivity to multicollinearity (Doumpos and Zopounidis, 1999). It is well known that these assumptions are incompatible with the complex nature of business growth (Lacher et al, 1995).

Consequently, with the emergence of data mining, machine learning algorithms such as random forests (RF) and artificial neural networks (ANN) have been used in an attempt to overcome the above mentions limitations in MDA and logit (Kim and Upneja, 2014). ANN models have been proposed as an attractive alternative because they are robust to some of these assumptions (Jain and Nag, 1997). Various studies report that ANNs models achieve better prediction results than traditional statistical techniques (Lacher et al., 1995; Etheridge et al., 2000; Bloom, 2004). For instance, Zhang et al. (1999) provide a comprehensive review of ANN applications for bankruptcy prediction. However, although many of previous studies report that ANNs models can produce better prediction results than logistic regressions, ANNs do not always result in superior predictive performance, leading to inconclusive outcomes when comparing these two models (Boritz et al., 1995). Thus further studies in the direction of model comparison is needed.

Another technique widely applied in various business-related research fields includes decision trees (DT) and their ensemble variations such as random forests (RF). For instance, Gepp et al. (2010) assessed the performance of the DT model for business failure prediction. They compared the prediction accuracy between the DT model and MDA based on Frydman et al.'s (1985) cross-sectional dataset during the period from 1971 to 1981 and included 20 financial variables to ensure the validity of comparisons with their research. They concluded that DT models show better predictive power than MDA. Li et al. (2010) demonstrated the applicability of the DT model in the area of business failure prediction and compared the predictive performance with four other classification methods including MDA, logit, kNN, and SVM. They predicted short-term business failure of Chinese listed companies on Shanghai Stock Exchanges. They used 135 pairs of companies in failure and healthy conditions and concluded that the predictive performance of DT models outperformed the other models for short-term business failure prediction. Another recent and more application-oriented study conducted by Ozgulbas and Koyuncugil (2012) proposed an early warning system based on DT-algorithms for SMEs to detect risk profiles. The proposed system uses financial data to identify risk indicators and early warning signs, and create risk profiles for the classification of SMEs into different risk levels.

In summary, despite the wide use of ANN, DT and RF in various research fields and industries for predictive modelling, the use of these models in the hospitality research is very scarce. Moreover, to the best of our knowledge, there have been no previous studies that employed web mining to predict the growth of restaurants.

4.3 Systematic Review of Restaurant Growth Factors

The restaurant business environment is complex and covered by a variety of firm-internal and external factors. To discover the factors influencing the growth of restaurants, we conducted a systematic literature review. To make the review process as transparent as possible we followed the guideline for systematic reviews as outlined in Chapter 3.2.

In the first step, information source are identified to be the top ten journals for hospitality research, which are Journal of Travel Research, Tourism Management, Annals of Tourism Research, Cornell Hospitality Quarterly, International Journal of Hospitality Management, Journal of Service Management, International Journal of Contemporary hospitality Management, Journal of Sustainable Tourism, Journal of Hospitality Marketing and Management and Journal of Hospitality and Tourism Research (Scientific Journal Ranking). As a search strategy, we developed a set of keywords describing the review work on the factors influencing restaurant business. The title was restricted to at least one of the following keywords: "restaurant", "gastronomy" and "food service industry". The abstract had to include at least one of the following keywords: "growth", "success", "key determinant", "bankruptcy" and "failure". The search resulted in 174 papers. In the study selection phase, we validated the relevancy of the 174 articles based on title, abstract, keywords and the full text. Studies not directly related to the performance of restaurants or determinants of growth are excluded from the review, such as “service failure and recovery strategies” or “menu engineering”. Finally, we found 107 articles that meet our criteria for data extraction, from which we manually extract information on restaurant growth factors.

To summarize the systematic literature review, we identified 49 factors influencing the growth of restaurants, which can be roughly divided into firm-internal and external factors (see Appendix A1-2). Firm-internal factors can be further divided into two groups: (1) the characteristics of the firm such as firm attributes (age, size, location), firm strategies (marketing, business concept) and food-related factors (price, quality and type of food), and (2) the characteristics of the entrepreneur such as socio-demographic characteristics (age, gender, family and educational background) and the personality of the entrepreneur (need for achievement, risk-taking propensity). Firm-external factors can be divided into 2 groups: factors reflecting (1) the immediate and (2) the contextual environment. The immediate environment includes customer relationship, competition and business network. In contrast, the contextual environment comprises macro-environmental factors such as economical, socio-cultural, technological and demographical determinants on the growth of restaurants. Figure 11 gives an overview of the factors influencing the growth of restaurants.

4.4 Ground Truth Data Collection

In the present case study, not publicly available data provided by a large Swiss insurer are used as a ground truth for growth model construction. The data provided by the Swiss insurer contain information of a set of Swiss restaurant, which consists of the restaurant's name, the annual revenue in the period from 2010-2017 and the type of restaurant, e.g. inn, snack-restaurant, hotel-restaurant etc. Furthermore, each restaurant contain a unique business identification number (UID) assigned by the Swiss Federal Statistical Office to facilitate the corporation between the government

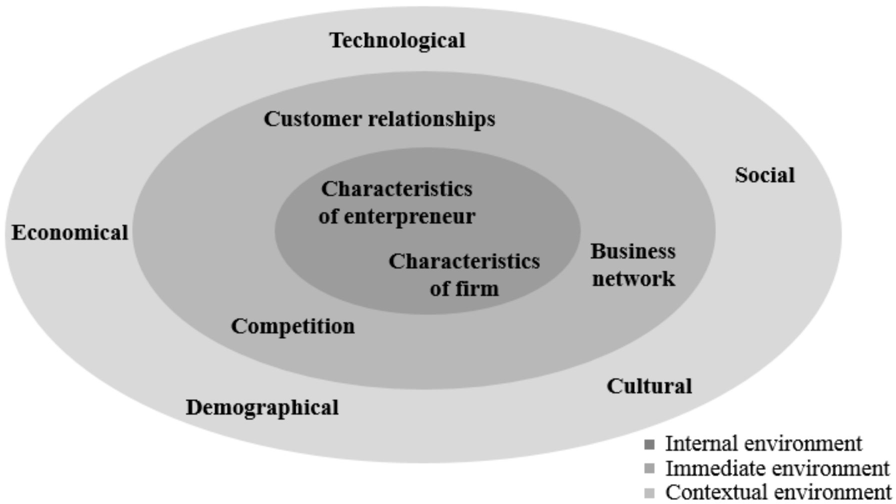


Figure 11: Restaurant business environment.

and firms. Thus, the data are used as following: (1) as a ground truth to train the growth model by constructing the growth label from the revenue data, (2) as a linkage to collect firm-related data from the web via UID, and (3) to construct input features for model training. In total, data of 2'014 Swiss restaurants are collected from the insurer for the purpose of this study.

4.5 Web Data Collection

Based on our literature review on the factors influencing the growth of restaurants, we collect information from various web data sources to cover a wide range of the aforementioned growth-indicating factors, which serve as input features to build growth prediction models for restaurants. For this purpose, web data related to the set of Swiss restaurants with known revenues (i.e. ground truth) are collected. In the first step, the usability of various web data sources is manually inspected with

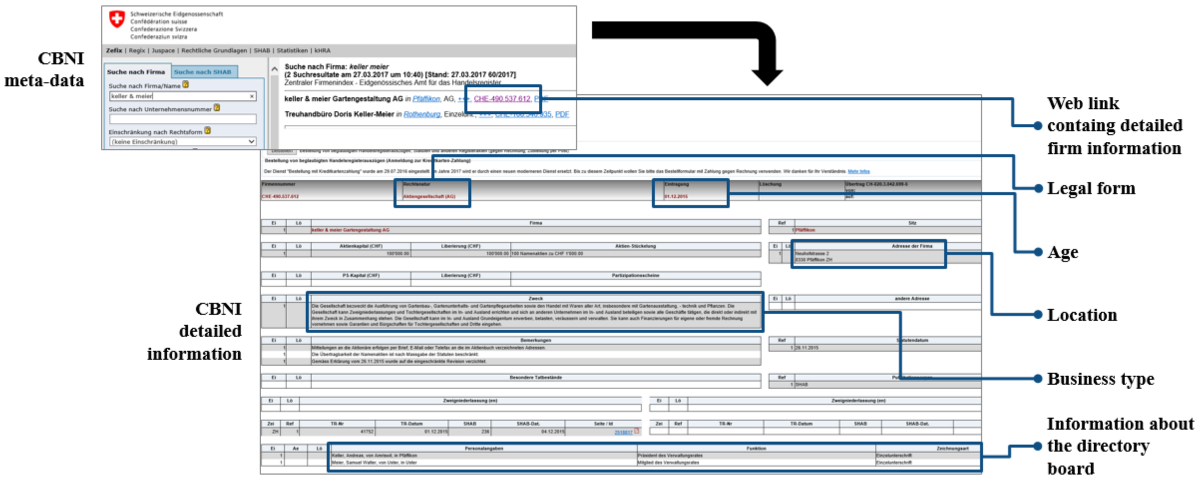


Figure 12: Exemplary excerpt from Central Business Names Index for demonstration purposes.

respect to the identified growth factors, as summarized in Appendix A1-2. Next, web data are collected and stored as described in Chapters 3.3 and 3.5 respectively. It is important to note, that each web data source has its individual structure and therefore, data collection methods must be developed for each web data source separately. Subsequently, in order to extract the information related to growth factors from the raw web data, text mining methods are utilized, as explained in Chapter 3.7. A total of six web data sources are considered, which are explained in more detail below.

4.5.1 Web Data Source “Central Business Names Index”

The Central Business Names Index (CBNI) provides free access to basic firm information and links through to internet excerpts from the individual canton commercial registry databases (für Justiz, 2001). The freely viewable information for each firm includes: UID, firm name, Swiss-wide identification number, registration date, legal form, address, purpose, status, and information about the members of the administrative board and their work function. Figure 12 shows an excerpt of the publicly accessible and viewable CBNI for an exemplary Swiss firm.

Collecting the data from CBNI is challenging and requires the application of both automated web interface submissions and preferential crawlers (see Chapter 3.3.3). In the first step, the meta-data of all Swiss firms are retrieved from the CBNI database using an automated web interface submission. The meta-data include basic information such as legal name and the unique identification number (UID) of a firm, and an additional web link that provides detailed information about the firms, as mentioned above. Thereby, Python’s library mechanize is used to implement the automated web interface submission (Lee, 2013). Next, a preferential crawler is deployed to collect detailed information about the firms by following the additional web links. Thereby, Python's library BeautifulSoup is used to extract the additional web links (Richardson, 2013), whereas Python's library urllib2 is used to access the detailed firm information (Lawson, 2015). The architecture of the data collection for CBNI is briefly explained in Figure 13.

In total, data of 577'540 Swiss firms are collected, covering the complete business population of Switzerland (Fueglistaller, 2017). The collected HTML raw data are stored in ElasticSearch for further processing (see Chapter 3.5.2).

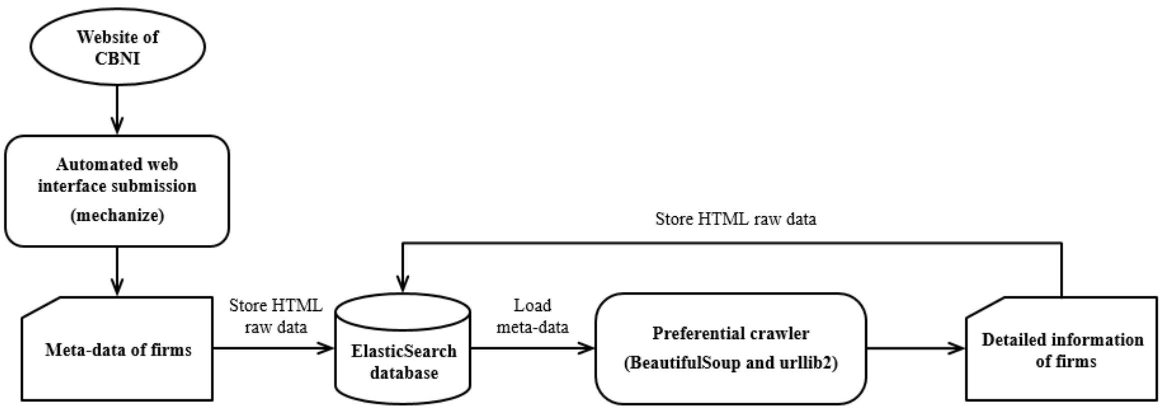


Figure 13: High-level view of the CBNI crawling architecture.

4.5.2 Web Data Source "TripAdvisor.com"

TripAdvisor.com (TripAdvisor) is one of the world's largest tourism communities (TripAdvisor, 2017). Founded in early 2000, it now covers restaurants in more than 190 countries, with over 200 million ratings and reviews autonomously generated by its users. Users can post reviews and opinions of travel-related content, such as hotels, restaurants and attractions. Furthermore, it is possible to add multimedia elements (photos and videos) or travel maps of previous trips or take part in discussion forums, web-based applications that allow users to post some material and discuss some specific topic. Moreover, TripAdvisor allows tourists to rate restaurants in a 5-star marking system from four separate aspects: food, service, value and atmosphere. These four criteria do have been proven to be able to influence consumers' restaurant decision-making (Heung, 2002).

To collect TripAdvisor data, we follow a similar approach as described for the web data collection in CBNI (see Chapter 4.5.1). In the first step, an automated web interface submission is applied to gather the meta-data of all Swiss restaurants using Python's library mechanize (Lee, 2013). The meta-data include basic information such as restaurant name and an additional web link, which provides detailed information about the restaurants. Next, a preferential crawler is deployed to collect detailed information about the restaurants by following the additional web links. We use the same Python libraries as mentioned in chapter 3.5.1, namely BeautifulSoup for extracting additional web links (Richardson, 2013) and urllib2 for accessing the detailed information of restaurants (Lawson, 2015). It is important to note that despite the similarity of the crawling architecture for CBNI and Tripadvisor, the automated web interface form submission and preferential crawler for Tripadvisor must be developed separately, since the website structure of TripAdvisor is significantly different from the one of CBNI.

The collected data are stored in ElasticSearch (see Chapter 3.5.2) and includes records of 20'429 Swiss restaurant, which covers most restaurant businesses of Switzerland. The collected data are HTML raw files and consist of information about the restaurant name and location, the cuisine type, price category, location-based ranking, number of reviews and review languages, the total ratings and ratings of the four criteria, i.e. food, service, value and atmosphere. These information are extracted using the methods described in Chapter 3.7. Figure 14 depicts an excerpt of a random exemplary restaurant on the publicly accessible and viewable TripAdvisor website.

Figure 14: Random example of a restaurant in TripAdvisor website.

4.5.3 Web Data Source: "Open Street Map"

Open Street Map (OSM) is a free-to-access web-based mapping system for location-based services and general information (OSM, 2016). In this case study, two types of datasets are directly downloaded from the OSM database of Switzerland and stored in PostgreSQL (see Chapter 3.5.1): (1) the Point of Interest (POI) dataset which contains 251'517 data points and (2) the Roads dataset which contains 295'819 data points. POIs are specific point locations on a map that are considered as useful or interesting for specific activities. They are described by the latitude and longitude or address of the location, type and name and contain six categories: public buildings (post, police, bank, school, university), healthcare (hospital, pharmacy, doctor), public transportations (bus, tram, taxi and train station), tourism (museum, attraction, gallery), entertainment (cinema, theatre, casino, arts center, nightclub), parking lots and residential area. The Roads dataset contains six types of roads: motorway, trunk roads, primary road, secondary road, tertiary road and unclassified roads, which are described by the latitude and longitude of the nodes spanned across the roads. These datasets are used to derive factors reflecting the infrastructure surrounding the restaurants, which are proven to be influential on restaurants growth (Park and Khan, 2006). Therefore, the restaurant address are geocoded, and the POIs and roads within a radius between 50m and 300m are extracted for each restaurant based in previous studies (Rammer et al., 2016; Chen and Tsai, 2016), as exemplary illustrated in Figure 15.

4.5.4 Web Data Source: "Swiss Federal Statistical Office"

Swiss Federal Statistical Office (SFSO) is the national service provider and competence center for statistical observations in areas of national, social, economic and environmental importance (Chen and Tsai, 2016). The SFSO is the main producer of statistics in the country and runs the Swiss Statistics data pool, providing information on all subject areas covered by official statistics. The dataset include socio-demographic, cultural and economic describing the Swiss population. Many of these factors are considered as significantly influencing the SMEs growth in past studies. The census data were derived from annual portraits provided by the SFSO and consists of 2'396 data points (Swiss Federal Statistical Office, 2016): population density, population change, foreign nationals, age pyramid (young, adult, and old

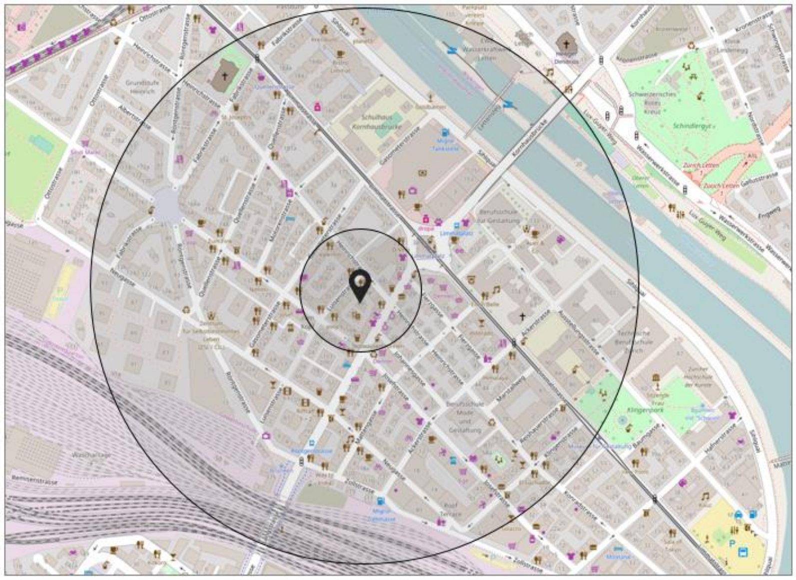


Figure 15: Exemplary illustration of POIs and roads within a radius between 50m and 500m as factors reflecting the infrastructure surrounding a restaurant located in Zurich city.

population ratios), area usage (settled and used for agriculture/forests/unused ratios), unemployment rate, residential density (persons per apartment room), and the number of businesses and residents employed in the different economy sectors (primary, secondary, and tertiary sector ratios). The data can be directly downloaded from SFSO in CSV-format. In addition, the data are provided as geographical data, which are aggregated on the level of municipalities - the lowest administrative unit on which Swiss census data is publicly available. Thus, the dataset is stored in PostgreSQL to facilitate location-based analysis.

4.5.5 Web Data Source:“Swiss Federal Tax Administration”

Swiss Federal Tax Administration (SFTA) is the Swiss administration for taxation, which manages the cantonal and municipal tax regulations (Swiss Federal Tax Administration, 2016). The Swiss taxation system is very complex, divided into many tax categories. In this study, we focus on the collection of the corporate taxation, which has proven to influence the restaurant growth (Borde, 1998). Therefore, we extracted two factors reflecting the corporate taxation: (1) the profit tax, based on the net profit as accounted for in the corporate income statement, and (2) the capital tax, which is levied on the ownership equity of companies. The tax data are provided on a cantonal level and consists of 52 data points. The data from SFTA are downloaded as CSV file and stored as geographical data (in cantonal units) in PostgreSQL (see Chapter 3.5.1).

4.5.6 Web Data Source: “Fast-Food Chains”

Fast-food chain giants such as McDonalds or Starbucks are well-known for conducting an extensive location assessment before a branch is opened (Morland et

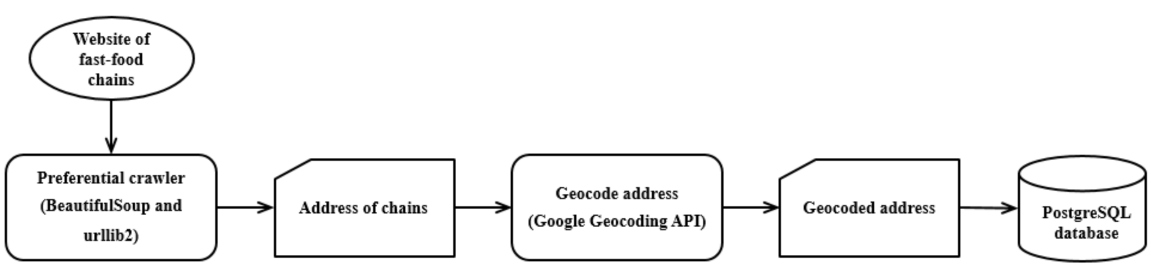


Figure 16: High-level crawling architecture and geocoding process for fast-food chains data.

al., 2002). Thus, in order to evaluate the location quality of restaurants, we inspect their proximity to chain branches. Therefore, we collected the address of all Swiss branches of the best-known fast-food chains, which include McDonald's, Subway, Starbucks and Burger King. The address of the branches are collected from each chain's website using a preferential crawler (McDonald's Switzerland, 2017; Subway Switzerland, 2017; Starbucks Switzerland, 2017; Burger King Switzerland, 2017). The preferential crawler is constructed as follows: First, Python's library urllib2 is used to retrieve each fast-food chains HTML raw file which contains a collection of the locations of the corresponding chains. Next, Python's library BeautifulSoup is used to extract the address of the chains. Finally, the addresses are geocoded using Google Geocoding API (Bernhard, 2013).

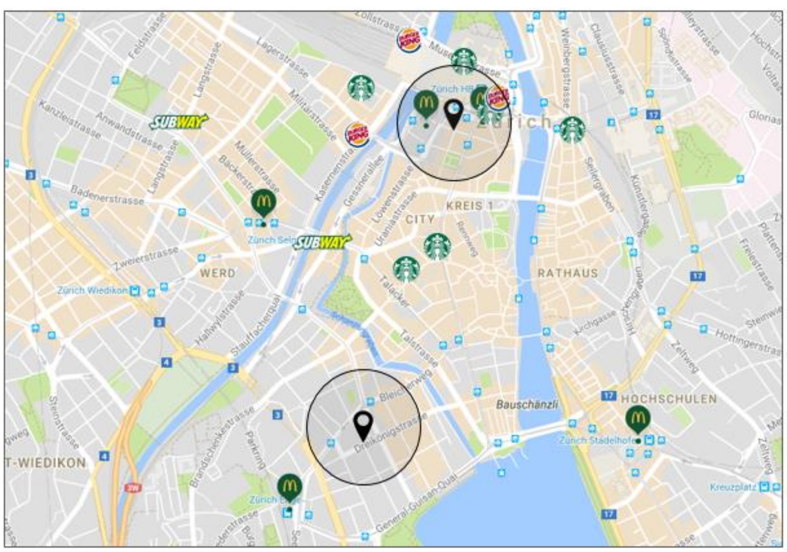


Figure 17: Exemplary illustration of two restaurants in the close proximity of fast-food chains (upper pin mark) and far away from fast-food chains (lower pin mark).

The data collection process for fast-food chains are illustrated in Figure 16. In total, 1'783 branches are collected and stored in PostgreSQL (see Chapter 3.5.1). In line with the collection of the above mentioned location-based information, the number of branches within a radius between 50m and 300m of restaurants are counted as a measure for the location quality for restaurants. Figure 17 exemplary illustrates a restaurant in the close proximity of fast-food chains.

4.5.7 Overview of Web Data Sources

To summarize the web data collection part of work, we have collected information related the growth of restaurants from six web data sources either by directly downloading the data or utilizing web crawling methods. The web data sources along with the data collection methods, data storage and number of records are summarized in Table 2.

4.6 Data Linkage

In this case study, we adopt the data linkage method described in Chapter 3.6 to combine data provided by the Swiss insurer with data of various web data sources. As shown in Figure 18 (right), our linkage approach is a semi-automated and rule- & knowledge-based method, which offers a high degree of flexibility and tuning possibilities, resulting in good data quality (Denk, 2009).

In the first step, insurer data are matched with the CBNI data source via UID, as the UID is unique for each firm (Figure 18: left, A). Next, a set of matching variables are defined to further match our newly created database (i.e. insurer data linked with CBNI data) with TripAdvisor data (Figure 18: left, B). Since the officially registered legal firm name in CBNI may differ from the actual restaurant name given in TripAdvisor, we define the following matching criteria for this matching step: name, zip code and street. String variables, such as names and addresses have to be pre-processed to be comparable among data sets. Therefore, standardization and parsing are required (see Chapter 3.6.1). Next, blocking method is applied to reduce the amount of data pairs for comparison (see Chapter 3.6.2). Thereby, zip code is used as a blocking variable. Subsequently, a string comparator is applied on the above mentioned matching variables in order assess the degree of similarity of the candidate pairs (see Chapter 3.6.3). Thereby, Python's library FuzzyWuzzy is utilized (Cohen, 2011). Finally, in the decision phase, the pairs of candidates are checked manually and the final decision on entity matching is made on the basis of our knowledge and experience (see chapter 3.6.4).

Further, location-based web data sources (OSM, SFSO, SFTA, fast-food chains) are matched with the geocoded address of our database (Figure 18: left, C). The

Web data source	**Data collection**	**Data storage**	**Number of records**
CBNI	Web crawling	ElasticSearch	577'540
TripAdvisor	Web crawling	ElasticSearch	20'429
OSM	Download	PostgreSQL	547'336
SFTA	Download	PostgreSQL	2'396
SFSO	Download	PostgreSQL	52
Fast-food chains	Web crawling	PostgreSQL	1'783

Table 2: Overview of the web data sources, data collection methods, data storage and number of records.

processes are conducted automatically and the potential matches are returned for each ground truth sample. As in the data linkage process described above, the potential matches are inspected manually to ensure a high data quality. Note, that only one among multiple matches are chosen or all matches will be discarded to ensure a high data quality for model building. In total, 403 restaurants of the initial 2000 restaurants could be successfully identified and matched with web data sources.

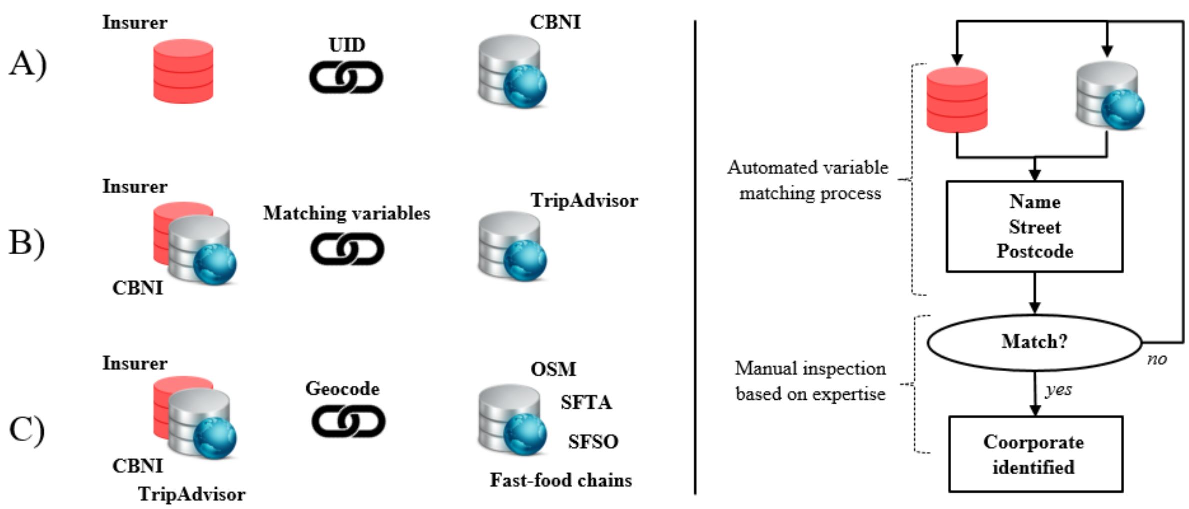


Figure 18: Linking corporate data with web data.

4.7 Label Creation & Data Preprocessing

4.7.1 Growth Label Creation

A crucial part of the data mining procedure is to define the proper label based on the business objective for data mining analysis. In this study, we test binary classification models for restaurant growth, i.e. separating restaurants into non-growing and growing ones. In the first step, we use the annual revenue between 2010 and 2016 of the ground truth data to calculate the relative change of revenue over the corresponding timespan using linear regression (Montgomery et al., 2012). It is important to note that the granularity of financial data provided by the Swiss insurer is limited to ten thousandths of Swiss francs, e.g. 150'000 CHF instead of the factual 154'350 CHF. Therefore, annual financial growth and shrinkage in the thousandth range are unlikely to be recorded.

Figure 19 shows the distribution of the ground truth data as a function of the relative revenue growth in percent. Out of 403 restaurants, 73 restaurants (18.11%) showed a negative revenue growth (relative_growth < 0), whereas 234 restaurants (58.06%) showed no signs of growth (relative_growth = 0), and 96 restaurants (23.83%) experienced a growth between 2010 and 2016 (relative_growth > 0). Since the primary interest of our study is to model the growth of restaurants, a cut off value of 0.0 is chosen to separate non-growing restaurants from the growing ones. To construct the binary labels, restaurants showing no signs of growth are assigned the value 0, whereas growing restaurants are assigned the value 1. Finally, the dataset consists of 307 samples with 0 as the majority class (76.18%) and 96 samples labelled with 1 as the minority class (23.82%).

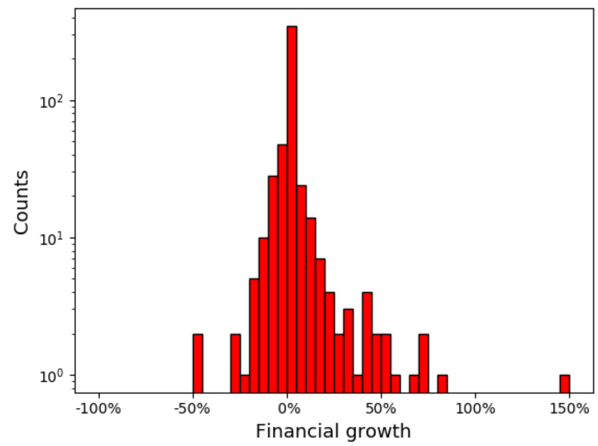


Figure 19: Distribution of the ground truth data (n = 403, histogram bins = 50).

4.7.2 Feature Engineering

Feature engineering: General procedures

The input features for growth modelling are derived from the collected web data as summarized in Table 2. The information from the Swiss insurer, SFSO and SFTA are provided in the form of structured numerical and categorical data and thus, require minimal data preprocessing. In contrast, the information extracted from CBNI and TripAdvisor are provided in the form of textual information, whereas data from OSM and fast-food chains are presented as geographical coordinates. First, the textual information are converted to a numeric representation. For instance in CBNI, registration date of firms are converted to a number of months to represent the age the of firm, work specialization are approximated by the number of distinct job functions, and the centralization of work are given in the form of a binary-valued variable by verifying the existence of sole signature authority within the firm.

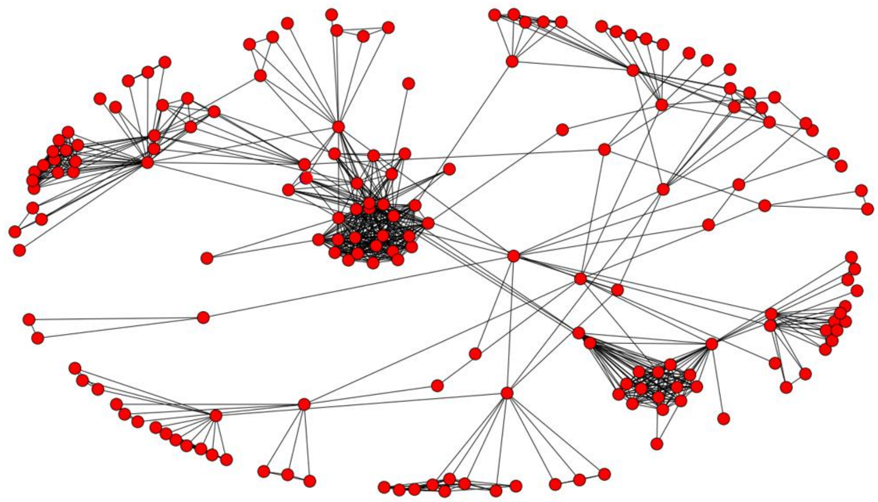


Figure 20: Example of a large and complex restaurant business network using Python's library networkx

Feature engineering: Business network

Furthermore, we use the information provided by the CBNI to analyze the business network of restaurants, which has shown to be very influential for restaurant growth (Hjalager, 2000). Business network analysis is a research field of network science which is a complete research domain in itself. Therefore, only the basics of network science used for feature engineering are described in this thesis. For more information on network science, please refer to Brandes et al. (2013).

Following prior research in business network, we define a business network as a group of firms which are interconnected by the entrepreneurs involved in the firm (Provan et al., 2007). In other words: If an entrepreneur is involved in two distinct firms, these two firms form business network. According to this logic, business networks are created for restaurants in which restaurants can be connected directly

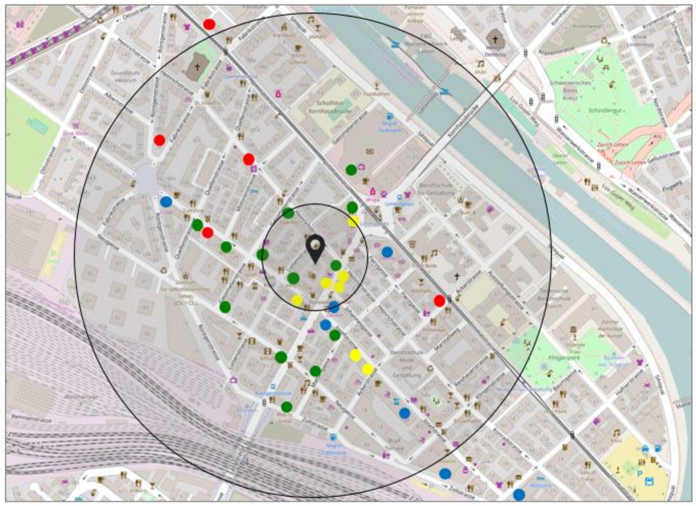


Figure 21: Random and exemplary illustration of competitive restaurants in Zurich within a radius between 50m and 300m. Competitive restaurants with same cuisine, better ratings, lower price category and more reviews are denoted in different colors.

or indirectly within a business network. Figure 20 depicts an example of a large and complex restaurant business network. Further, measures known from network science are utilized to assess the network characteristics, including network size - number of restaurants within a network, network density - the degree of cross-linking of restaurants within a network, and centrality measures – taking the importance of restaurant positions within a network into account, similarly to Google's PageRank which is used to rank websites in their search engine (Rogers, 2002). These network measures are calculated by utilizing Python's library networkx. For more information on the network measures, see Provan et al. (2007).

Feature engineering: Competition analysis

In addition, since we collected the data of all Swiss restaurants, we geocoded the locations of all restaurants to conduct a competition analysis using Google

Geocoding API (Bernhard, 2013). Thereby, the geocodes are loaded into PostgreSQL, which facilitates location-based analysis (see Chapter 3.5.1). Following prior research in analyzing business competition, competitive restaurants in the surroundings within a radius between 50m and 300m of our ground truth data are counted (Chen and Tsai, 2016; Rammer et al., 2016). The more competitive restaurants in the surroundings, the greater the competition. In the present case study, restaurants with same cuisine, better overall ratings, lower price category and more reviews are considered as competition, as exemplary illustrated in Figure 21.

Table 3 provides an overview of the growth factors which are derived through feature engineering from the six aforementioned web data sources. In total, 27 out of the initially 49 identified growth factors are covered. Note, that the specific input features for supervised machine learning are elaborated in Chapter 4.7.3.

4.7.3 Feature Preprocessing

Because web data are often incomplete, the generated features are incomplete. Missing data treatment should be carefully treated, otherwise bias might be introduced into the knowledge induced, as outlined in Chapter 3.8.1. In our dataset, the range of missing data are between 0% and 52%, as shown in Table 4. TripAdvisor data containing the most missing data due to the incompleteness of information generated by TripAdvisor users, such as information about the type of meal (i.e. breakfast, lunch, dinner) or the availability of parking lots. To address this issue, the following measures have been taken based on the business and structural characteristics of the features (see Chapter 3.8.1): 1) delete samples if only few

Data sources	Factor type	Growth factor
CBNI	Firm attributes	Age, size
	Firm resources	Human capital
	Organization structure	Work specialization, Centralization
	Network	Inter-organizational links
TripAdvisor	Firm attributes	Reputation, service quality, physical environment
	Food	Price, quality, type
	Customer relationships	customer satisfaction & feedback
	Competition	Clusters of restaurants, food pricing
OSM	Technological	Infrastructure, tourism
	Social-cultural	Lifestyle
SFTA	Economical	Taxation
SFSO	Social-cultural	Social class, cultural diversity
	Demographical	Population size, growth & density, Age & gender distribution, employment & income, household size
Fast-food chains	Firm attributes	Location

Table 3: Web data sources and growth factors extracted through feature engineering. A detailed list of all growth factors is given in Appendix A1-2.

samples are involved (missing data less than 10%), 2) delete features if imputation is not suitable, 3) impute missing numerical values with the mean value (Batista and Monard, 2003), and 4) impute missing categorical value with -1 which represents the absence of a particular information (Gryzmala-Busse and Hu, 2000).

Furthermore, features with zero variance and high correlation (Pearson correlation coefficient rprs ≥ 0.95) are removed, as explained in Chapter 3.8.2. In total, 85 input features are generated for the purpose of supervised machine learning, as summarized in Table 5. Note, that features denoted with a digit at the end are dummy variables derived from categorical features.

Features (grouped)	**Missing values**
CBNI	0%
TripAdvisor	0% - 52%
OSM	8% - 36%
SFTA	26%
SFSO	24%
Fast-food chains	0%

Table 4: The extent of missing values in our dataset in the web data sources.

Feature ID	Feature name	Feature ID	Feature name
1	Revenue level	44	Streets within 50m
2	Restaurant type 1	45	Pedestrian zones within 50m
3	Restaurant type 2	46	Parking lots within 50m
4	Restaurant type 3	47	Public transportation within 50m
5	Firm age	48	Public building within 50m
6	Management size	49	Residential within 50m
7	Centralization of work	50	Fast-food chains within 50m
8	Ratio management vs functions	51	Tourism within 300m
9	Legal form 1	52	Motorway within 300m
10	Legal form 2	53	Streets within 300m
11	Legal form 3	54	Pedestrian zones within 300m
12	Number of cuisine	55	Parking lots within 300m
13	Number of feedback	56	Public transportation within 300m
14	Ranking	57	Public building within 300m
15	Number of feedback languages	58	Healthcare within 300m
16	Rating overall	59	Entertainment within 300m
17	Rating best	60	Residential within 300m
18	Rating good	61	Number of restaurants within 50m
19	Rating satisfied	62	Number of restaurants with same cuisine within 50m
20	Rating insufficient	63	Number of restaurants with lower price within 50m
21	Rating bad	64	Number of restaurants with more review within 50m
22	Rating service	65	Number of restaurants with better feedback within 50m
23	Rating cuisine	66	Number of restaurants within 300m
24	Rating quality	67	Number of restaurants with same cuisine within 300m
25	Number of meal type	68	Number of restaurants with lower price within 300m
26	Meal type 1	69	Number of restaurants with more review within 300m
27	Meal type 2	70	Number of restaurants with better feedback within 300m
28	Meal type 3	71	Business network size: only direct partners
29	Meal type 4	72	Business network size: including indirect partners
30	Number of characteristics	73	Business network density
31	Characteristics 1	74	Population size
32	Characteristics 2	75	Population density
33	Characteristics 3	76	Foreigner
34	Characteristics 4	77	Population (0 to 19 years)
35	Characteristics 5	78	Population (20 to 64 years)
36	Characteristics 6	79	Population (over 64 years)
37	Number of occasions	80	Housing ownership rate
38	Occasion 1	81	Empty flat rate
39	Occasion 2	82	Rating atmosphere 1
40	Occasion 3	83	Rating atmosphere 2
41	Price 1	84	Rating atmosphere 3
42	Price 2	85	Rating atmosphere 4
43	Tourism within 50m		

Table 5: Input features for supervised machine learning algorithms.

4.8 Supervised Machine Learning for Growth Modeling

4.8.1 Model Selection and Evaluation Methodology

Restaurant growth is a highly complex mechanism, thus predicting the growth of restaurants requires machine learning algorithms which are capable to handle a high level of complexity. Therefore, we use the Random Forest Classifier (RFC) and Multi-layer Perception (MLP) neural network, a subclass of ANN, which are able to model complex interactions between the input variables and thus, share a predominant role in a range of research domains (Cutler et al. 2007). In addition, comparable models from other studies (i.e. growth models based on web mining) cannot be identified as a baseline to the best of our knowledge (see Chapter 4.2.2). Therefore, we utilize linear regression (LR) as a benchmark due to its wide use for economic modelling in the past (Youn and Gu, 2010).

RFC is a non-parametric non-linear classification algorithm that fits an ensemble of decision trees to a dataset, and then combines the predictions from all the trees (see Chapter 3.9.2). From the ensemble of trees, the predicted class of an observation is calculated as the class with the majority vote (Chen et al., 2004). Furthermore, a by-product of the random forest algorithm is the measure of feature importance, which allows a data-based evaluation of the relative importance of the growth factors.

MLP neural network is powerful machine learning algorithm for pattern recognition and classification due to the non-linear, non-parametric adaptive learning properties and thus, is capable of modelling highly non-linear relationships (see Chapter 3.9.3). MLPs are typically composed of at least three layers of nodes: the input layer, at least one hidden layer and the output layer. The network architecture is characterized a large set of parameters, such as the number of layers, the number of nodes in each layer and how the nodes are inter-connected. The input layer consists of input features, whereas the output layer produces the model outcome. In between, there are one or more hidden layers which aims at model the complex relationship between the input layer and the output layer. One drawback of MLPs, when compared to RFC, is their limited explanatory power due to the "black-box" nature of MLPs.

LR is another machine learning algorithm estimates the relationship between the dependent variable and a set of features using a logistic function (see Chapter 3.9.1). Furthermore, the relative contribution of each feature on the actual classification can be determined, which is a key advantage in contrast to the MLPs (Neophytou and Molinero, 2004).

The model selection, evaluation and interpretation follow the methods outlined in Chapter 3.10. In the initial step, our dataset is split into a training and test set following a 90/10 ratio. The training set are used for hyper-parameter tuning and model training, while the test data set is used to report models' performance.

In this study, we use Python’s sklearn implementation of the above mentioned machine learning algorithms (Pedregosa et al., 2011). The models' hyper-parameters to be optimized are summarized in Appendix A3 for each model class. Therefore, we conducted a randomized grid search to find the optimal value for the parameters for each classifier with 500 iterations, i.e. 500 combinations of hyper-parameters are tested for each classifier (see Chapter 3.10.2). Randomized grid search was chosen over the standard grid search method due to the reduced computational time while producing comparative results. Furthermore, in order to validate the optimized classifiers to the training set, a stratified 10-fold cross-validation procedure was applied for model selection (see Chapter 3.10.3). In a stratified 10-fold cross-validation (CV), the original sample is partitioned into 10 subsamples while maintaining the ratio of the classes in the target variable (McKay et al., 1979). Thereby, we make use the function RandomizedSearchCV() of the Python library sklearn, which combines both of the aforementioned methods (Pedregosa et al., 2011). Finally, the performance of the final model is optimized and reported on the test set, as outlined in Chapter 3.10.3.

In addition, in order to reduce the variance due to the training-test split, and to obtain reliable performance estimation for model comparison, we repeated the aforementioned procedure multiple times (see Chapter 3.10.1). Therefore, we successively split our dataset into training and test set, and execute the proposed procedure multiple times. Thereby, the dataset is reshuffled and re-stratified before each round. Finally, we then report the average performances of the classifier families, i.e. RFCs, MLPs and LRs. In this case study, the number of repeats is set to 10. The performance of each of the ten modeling can be found in Appendix A4.

To compare and evaluate the classification performance of our classification algorithms, we make use of the performance measures described in Chapter 3.10.4, which includes the area under the receiver operating characteristic curve (AUC) measure - a commonly used measure for model comparison and effective evaluation of the accuracy measure (Bradley, 1997), accuracy - the overall percentage correctly classified, sensitivity - the fraction of samples correctly classified as growing restaurants, and specificity - the percentage of samples correctly classified as non-growing restaurants. Note, that the performance measures are determined for each repeat, and finally averaged and reported as the mean performance of the classification method along with the standard deviation (Yamane, 1973).

4.8.2 Model Results and Interpretation

We first evaluate the models based on the performance measures mentioned above. Subsequently, we elaborate the explanatory power of the input features by reporting the mean feature importance across the RFCs, which is an inherent measure of the random forest algorithm. In addition, we report the relative contribution of each feature from LRs as a mean feature importance measure by following the study concept of Grömping (2009). Finally we discuss and compare the factors influencing the growth of restaurants of our RFCs and LRs.

Table 6 shows the average classification performance of RFCs, MLPs and LRs with respect to a binary classification of samples into non-growing and growing restaurants. Based on the AUC and accuracy, RFCs yield the best results among the tested models, with mean AUC and accuracy of 68.1% and 68.0% respectively. LRs reports slightly lower mean AUC and accuracy of 65.8% and 66.3% respectively, which clearly outperform the MLPs with mean AUC and accuracy of 62.0% and 57.7% respectively. Furthermore, our results suggest that both RFCs and MLPs favored specificity over sensitivity, while LRs favored sensitivity over specificity.

Figure 22 depicts the mean feature importance plot of our RFCs (left) and LRs (right) only for the 20 most important features due to the large amount of input features, which have been used to train the models. The complete feature importance plot of all input features can be found in Appendix A5-6. In addition, the numeration of the

	Roc_auc	Accuracy	Sensitivity	Specificity
Random forests	68.1 ± 5.0 %	68.0 ± 7.0 %	65.6 ± 17.0 %	68.8 ± 15.0 %
Multi-layer perceptrons	62.0 ± 6.0 %	57.7 ± 6.0 %	72.2 ± 11.0 %	52.7 ± 10.0 %
Logistic regressions	65.8 ± 8.0 %	66.3 ± 6.0 %	60.0 ± 15.0 %	68.5 ± 10.0 %

Table 6: Average performance and standard deviation of classifier families.

features refers to the feature ID of Table 5. Despite the different ranking of the RFCs' and LRs' features, we can observe five common features among the top 20 features, namely features related to the price of food (feature 41), competition (feature 63 and 68), firm characteristics (feature 30) and demographical factor (feature 79). The feature importance of RFCs shows, that "firm age" (feature 5) is clearly the most predictive feature with a with a substantially larger importance value than all other predictors, followed by the number of feedbacks given in TripAdvisor (feature 13), the overall ranking of the restaurant in TripAdvisor (feature 14), and the rating "best" (feature 17). The subsequent features are characterized by a mixture of features reflecting factors mainly related to the demographics, customer relationship and competition. The top 20 features of LRs are characterized by a set of factors with a flat distribution of the relative importance. In line with the feature importance of RFCs, factors reflecting the competition play in important role for LRs as well (feature 63, 68 and 69). However in contrast to RFCs, the top 20 features of RFCs are governed by factors reflecting the infrastructure, such as the proximity to public transportation, building, parking lots and fast-food chains (feature 48 - 50, 55 - 56).

4.9 Case Study Conclusions

4.9.1 Case Study Summary

In the present case study, we analyze the use of web data for the purpose of predicting the financial growth of restaurants. First, 49 factors influencing the growth of restaurants are identified through an extensive systematic literature review, as summarized in Appendix A1-2. Next, a set of web data sources are examined with regards to the identified growth factors. Within the scope of this study, six web data sources containing information reflecting the business internal

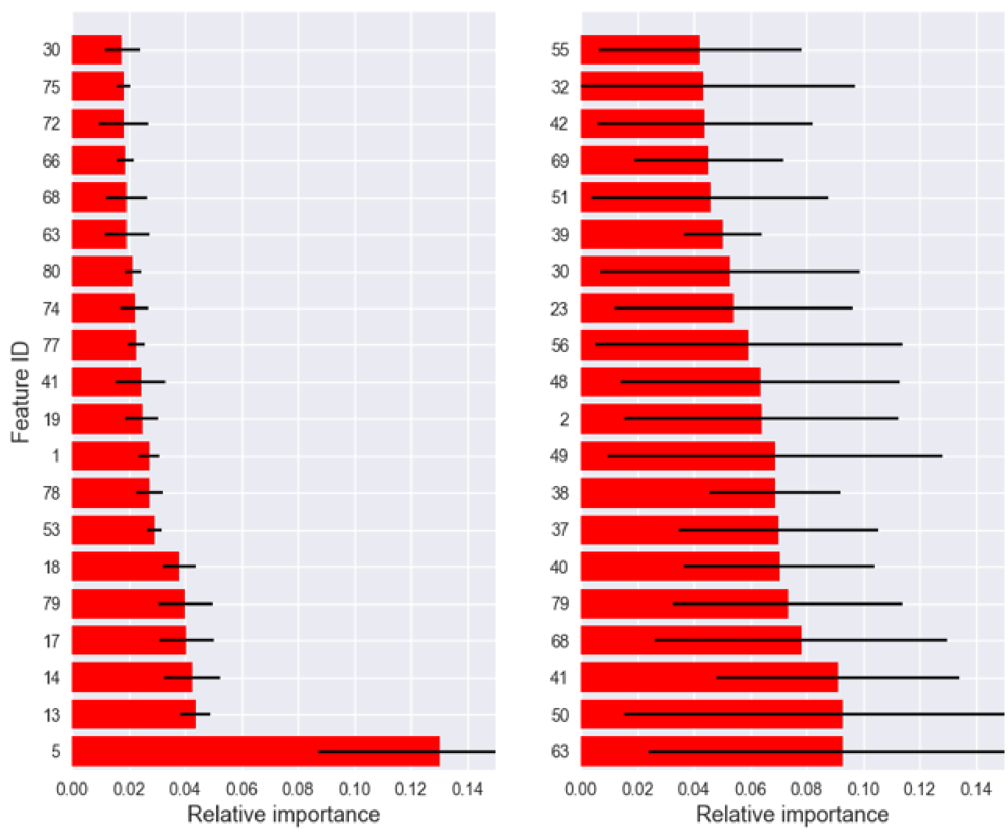


Figure 22: Feature importance plot including the top 20 features of RFCs (left) and LRs (right). The numeration of features refers to feature ID used in Table 5.

and external environment of restaurants are identified: Central Business Names Index, TripAdvisor, OpenStreetMap, Swiss Federal Statistical Office, Swiss Federal Tax Administration and fast-food chains data. The data are either downloaded from the websites or collected by means of web crawling. Text mining methods are applied to extract the growth factors from textual information and to construct the input features for predictive modelling. Therefore, RFCs, MLPs and LRs are tested and compared with the goal to predict a binary outcome, i.e. non-growing versus growing restaurants.

Our results suggest, that RFCs with a mean accuracy of 68% outperform MLPs and our in-house built benchmark LRs. Furthermore, our study shows that the LRs is not inferior to MLPs in terms of growth prediction accuracy for restaurants, as opposed to many studies reporting MLPs' better prediction accuracy when compared to LCRs. The feature importance measure of our RFCs and LRs suggest that wide selection of factors are important for the construction of the growth model. Especially, information related to customer relationship (number of feedback and ranking) extracted from TripAdvisor are very useful to model the growth of restaurants. Moreover, external environmental factors such as the infrastructure (number of streets within 300m), competition (e.g. number of restaurants with lower price within 300m) demographics (population size, density and age distribution) also play a crucial role in modelling the growth of restaurants. The complete list of input factors ranked by their feature importance is provided in Appendix A5.

To the best of our knowledge, this study is the first to apply WM techniques combined with supervised machine learning techniques to model the growth of restaurants. Our result demonstrates the potential of building growth prediction models for restaurants based on publicly accessible web data.

4.9.2 Case Study Limitations

This study is not without limitations and provides several opportunities for further research. First, our work is limited to Swiss restaurants, thus the obtained results might differ in different geographical regions. Second, the revenue data of restaurants are provided by an insurer, which might differ from the actual revenue. Third, important growth factors completing the firm-internal environment, such as the characteristics of the entrepreneur (appendix) are not included in our model because they are not available in the examined web data sources. To address this issue, the proposed web mining research can be applied to collect and preprocess textual information given in company websites and social platforms like Xing, with the goal to enlarge the input feature space of our growth models. Finally, other machine learning methods such as stacking classifiers could be tested with to goal to optimize the performance restaurant growth prediction.

5 General Discussion and Implications

This chapter presents the discussion of the research results of this thesis. It begins with a summary and discussion of principal findings, followed by a reflection on theoretical and practical implications. Furthermore, the limitations of this thesis and the future prospects for the application of web mining for SME growth research are discussed. At the end of this chapter a final conclusion of the overall work is presented.

5.1 Key Findings

Given the importance of SMEs for the economy and society, decision-makers and researchers have made efforts to promote SME growth and to improve overall economic performance (Carter and Auken, 2006). Therefore, analyzing and predicting the growth of SMEs has become an important area of research. However, the potential of web mining for the growth prediction of SMEs has not yet been thoroughly evaluated, despite its wide use in many business-relevant applications. In addition, the use of web data offers many advantages. There is a huge and growing amount of easily and publicly accessible web data that can be obtained cost-effectively and in large quantities, which can be used for SME growth modelling (Gök and Shapira, 2015; Saini and Pandey, 2015).

Therefore, the central objective of research which underlies this thesis is to investigate, which potential web mining bears for the SMEs growth prediction. A corresponding research question is formulated in the introduction, where we investigate the applicability of web mining for SME growth research. In the course of the investigation, a web mining framework is constructed and the framework elements are elaborated in detail. Further, a case study is conducted to assess to applicability of the proposed web mining framework. In the scope this thesis, web mining is regarded as a method for the automated identification, retrieval, extraction and analysis of web content with the goal to predict the growth of SMEs. Hence it must not be confused with the discipline of web usage mining - focusing on the analysis of user data, and web structure mining - analyzing topographical aspects of the internet.

In general it can be stated that web mining unfolds its true potential if used for analyzing large volumes of web content. Further, web mining is not equally useful for all industries. Business areas benefiting the most are those which are sufficiently covered on the web. For instance, this comprises consumer-centric businesses which are interested in social information like sentiments and consumer behavior, and businesses which are strongly influenced by geospatial information as demonstrated in our case study. Therefore, the development of a web mining-based growth model for individual business sectors must be carefully evaluated on the web presentation. In this regard, conducting a systematic literature search is a suitable approach, in which growth factors for individual business segments and the potentially useful web data sources are identified.

Further, web data sources are by nature not designed to be easily linked to each other, even less to be linked with company-internal databases. Therefore, a high level of caution is necessary when conducting data linkage, as this may tremendously influence the data quality for growth modeling. For this purpose, we apply a semi-automated rule- and knowledge-based method adopted from Denk (2009), which yields good performance for data matching (see Chapter 3.6).

In general, working with web data is challenging task due to their unstructured and chaotic nature resulting in high quantity of missing values in web data sets. Thus, dealing with missing values in web data are highly sensitive as wrong treatment may induce erroneous bias into the knowledge. The optimal solution to this issue would be to collect data from additional web data sources to fill the unavailable information. However, if this is not possible or affordable, more sophisticated methods such as imputation techniques must be applied. It is recommended to first analyze the business and structural characteristics of the information before applying any methods (see Chapter 3.8).

Further, the feasibility of the proposed research framework strongly depends on the availability and quality of the ground truth data. In the present thesis, we aim at predicting the financial growth of firms. In order to develop the financial growth measure, data on the financial situation of companies must be collected which are highly sensitive and most likely not publicly available on the web. Commonly, this data are obtained through financial institutions or questionnaire studies. In our case study, the ground truth data are provided by a Swiss large insurer, which contains information about the annual revenue of Swiss restaurants over a large period of time (see Chapter 4.5).

Finally, our case study demonstrates that the application of web mining for SMEs growth prediction is a very promising approach. Based on six web data sources we are able to predict the growth of restaurants with an overall accuracy of 68%. Given that the growth mechanism of restaurants is highly complex and that the constructed growth model is based on web mining, we consider the results to be encouraging both for further research and commercial implementation. In consideration of the novelty of the proposed approach in the field of gastronomy research, we were not able to identify a benchmark model for comparison within this specific research area. Furthermore, we were not able to compare our model with the performance of experts (e.g insurance agents) in growth prediction for restaurants. Nevertheless, our results are comparable to the latest research findings from related research areas, where SMEs bankruptcy are predicted with an accuracy of 68% using Random Forest Classifier (RFC) (Sigrist and Hirnschall, 2018).

5.2 Contributions to Theory and Practice

The present work has both theoretical and practical implications. It contributes to the existing literature of SMEs growth research by confirming previous findings in a data-driven and model-based manner through supervised machine learning. Furthermore, the proposed approach can be used to identify new growth factors, for instance based on the feature importance measure of the RFC and thus, extend the empirical body of knowledge.

Besides of the theoretical aspects, this study has a number of important practical implications. The research findings can be used to build an information system for SMEs which allows an automated collection and analysis of publicly available web data in large scale with the objective of predicting future growth opportunities of SMEs. For the Swiss SME organizations, the insights generated in this thesis may support the Swiss SME organizations at understanding the growth of SMEs, thus strengthen their supportive role for SMEs. For investment companies, the proposed information system can be used to monitor the development of SMEs by mining the changes in the internal and external business environment from web data, serving as an “early recognition system” for future opportunities of growth. Finally, for SMEs, the information system can be used to evaluate the characteristics of firms based on the information given in the web. The absence of important key success factors can be pointed out to firms, thus serving as a consulting program. In particular for the insurance industry, there are many areas of application. The proposed information system allows an automated collection of business-relevant information, followed by a structured representation of web data. These information may facilitate electronic sales support for insurance consultants and improve the quality of advisory work. For instance, insurance consultants receive suggestions for discussions with customers through aggregated data information. Further, the information system can be used to monitor changes in the directory board of firms, which is considered one the most frequent reasons of contract cancellation. In addition, the information system allows the identification of business relationships among SMEs, which is a very valuable information for customer acquisition.

5.3 Limitations and Future Directions

In the following, a range of limitations present in this thesis work and the potential further outlook are discussed. First, the proposed web mining framework for SMEs growth modeling has only be tested in one case study. Further case studies considering other business industries should be conducted to validate the generalization and applicability of the web mining framework. In particular, it would be interesting to understand whether the proposed framework can be applied to predict the growth of emerging high-tech businesses where growth modeling has not been studied yet such as blockchain startups or firms dealing with VR and AR.

Second, the ground truth data to train and validate our models are provided by a large Swiss insurer. In our case study, the dataset includes information about restaurant's name, the annual revenue in the period from 2010-2017 and the type of restaurant, e.g. inn, snack-restaurant, hotel-restaurant etc. However, the revenue data suffer of approximations - including the lack of proper reporting/updating, and therefore the predictive modeling is impacted by this phenomenon. In addition, the granularity of financial data provided by the Swiss insurer is limited to ten thousandths of Swiss francs and thus, small annual financial growth and shrinkage in the thousandth range are unlikely to be recorded. This fact may lead to the deviation of the predictive power of our models from the actual one. Therefore, multiple or more reliable sources as a ground truth are desirable to validate the proposed web mining framework. For instance, data from tax authorities or questionnaires for assessing growth measures directly from SMEs are recommended.

Third, as mentioned in Chapter 4.9.2, this thesis did not consider factors reflecting the characteristics of the entrepreneur, which has shown to be influencing the growth of SMEs. This is due to the fact that these information are very difficult to obtain from publicly available web data. In addition, this information is traditionally evaluated using questionnaire studies, which is not our research focus. Unfortunately, we were not able to identify suitable web data sources which broadly covers the characteristics of entrepreneurs. Initially, Linkedin was considered to be a suitable web data source for the analysis of entrepreneurial characteristics (Prodromou, 2012). However, after Linkedin have severely restricted its policy for scientific applications, we have not further considered this web data source (Linkedin, 2017). Future research directions could assess the availability of information about entrepreneurs on the publicly accessible web, such as social media platforms or entrepreneurship platforms accessible to researchers.

Fourth, this thesis did not include information given in company websites. Various research confirmed that firms having a website enable to reach wider geographical markets and increase customers because more people were able to access information about the business, thus improving the business effectiveness (Lunati, 2000; Pages, 2002). A study have shown that over 70% of all companies in the EU own a website (Eurostat, 2015). Considering this fact, one can assume that a large amount of potential valuable information is hidden on the websites of SMEs, which theoretically can be used to gain better insights about the growth mechanism of SMEs. However, studying the use of information in company website are the subject of the latest research efforts, where many caveats and restrictions in the interpretation of corporate website information have been identified (Gök and Shapira, 2015). Moreover, using website data needs particular technical skills, including skills which are different from those used in the handling structured or semi-structured web data sources such as TripAdvisor or ZEFIX. Therefore, future studies should focus on the use of company websites as a novel data source for the SMEs growth prediction.

Finally, this thesis presents the technological possibilities arising through the use of web mining for SMEs growth prediction without further considerations of the legal and ethical issues of web mining. Despite the potential of web mining, web mining does pose a threat to important legal and ethical values which should be respected and protected by web mining practitioners and web users (Van Wel and Royakkers, 2004; Velásquez, 2013). Moreover, the linkage of various web data sources with firm-internal customer data as presented in this thesis must be viewed critically from a data privacy perspective. With the General Data Protection Regulation (GDPR Regulation (EU) 2016/679) that came into effect in the EU in May 2018, the topic is of high actuality as well for Switzerland. Moreover, the linkage of distinct sources of customer data (i.e. SMEs data) within this thesis must be viewed critically from a data privacy perspective. Especially, the use of customer data by insurers is highly regulated and restricted by law in the United States (NCSL, 2016). However, due to the actuality of the new regulation by the General Data Protection Regulation (GDPR Regulation (EU) 2016/679), studies or general guidelines dealing with these new regulatory aspects in conjunction with web mining cannot be identified. Therefore, the technological possibilities of web mining and the legal and ethical framework requirements must be carefully coordinated for use in business operations to prevent violations of compliance, existing regulations and minimize conduct/reputational risk.

Bibliography

Aizawa, A., & Oyama, K. (2005, April). A fast linkage detection scheme for multi-source information integration. In *Web Information Retrieval and Integration, 2005. WIRI'05. Proceedings. International Workshop on Challenges in* (pp. 30-39). IEEE.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589-609.

Antlová, K. (2009). Motivation and barriers of ICT adoption in small and medium-sized enterprises. *E+ M Ekonomie a management*, (2), 140.

Antlová, K., Popelínsky, L., & Tandler, J. (2011). Long term growth of SME from the view of ICT competencies and web presentations. *E + M Ekonomie a management*, (4), 125.

Ashton, D. N., & Sung, J. (2002). *Supporting workplace learning for high performance working*. International Labour Organization.

Baek, S. H., Ham, S., & Yang, I. S. (2006). A cross-cultural comparison of fast food restaurant selection criteria between Korean and Filipino college students. *International Journal of Hospitality Management*, *25*(4), 683-698.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, *38*(1), 63-93.

Balzert, H. (2007). *Basiswissen Web-Programmierung: XHTML, CSS, JavaScript, XML, PHP, JSP, ASP. NET, Ajax*. W3l GmbH.

Bandaru, N., Moyer, E. D., & Radhakrishna, S. (2011). *U.S. Patent No. 7,930,302*. Washington, DC: U.S. Patent and Trademark Office.

Baron, R. A. (2004). The cognitive perspective: a valuable tool for answering entrepreneurship's basic “why” questions. *Journal of business venturing*, *19*(2), 221-239.

Bates, T., & Nucci, A. (1989). An analysis of small business size and rate of discontinuance. *Journal of Small Business Management*, *27*(4), 1.

Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, *17*(5-6), 519-533.

Baum, J. R., & Locke, E. A. (2004). The relationship of entrepreneurial traits, skill, and motivation to subsequent venture growth. *Journal of applied psychology*, *89*(4), 587.

Beck, T., & Demirguc-Kunt, A. (2006). Small and medium-size enterprises: Access to finance as a growth constraint. *Journal of Banking & finance*, *30*(11), 2931-2943.

Beck, T., Demirgüç-Kunt, A., & Maksimovic, V. (2006). The influence of financial and legal institutions on firm size. *Journal of Banking & Finance*, *30*(11), 2995-3015.

Begley, T. M., & Boyd, D. P. (1987). Psychological characteristics associated with performence in entrepreneurial firms and smaller businesses. *Journal of business venturing*, *2*(1), 79-93.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281-305.

Bernhard, S. (2013). GEOCODE3: Stata module to retrieve coordinates or addresses from Google Geocoding API Version 3.

Bero, L. A., Grilli, R., Grimshaw, J. M., Harvey, E., Oxman, A. D., & Thomson, M. A. (1998). Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. *BMJ (Clinical research ed.)*, *317*(7156), 465-468.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, *18*(5), 16-23.

Bishop, C. M. (2006). Pattern recognition and machine learning (information science and statistics), Springer-Verlag New York. *Inc. Secaucus, NJ, USA.*

Bloom, J. Z. (2004). Tourist market segmentation with linear and non-linear techniques. *Tourism Management*, *25*(6), 723-733.

Boddy, D. (2009). *Management*. Pearson Education.

Boden Jr, R. J., & Nucci, A. R. (2000). On the survival prospects of men's and women's new business ventures. *Journal of business venturing*, *15*(4), 347-362.

Borde, S. F. (1998). Risk diversity across restaurants: An empirical analysis. *Cornell Hotel and Restaurant Administration Quarterly*, *39*(2), 64-69.

Boritz, J. E., Kennedy, D. B., & Albuquerque, A. D. M. E. (1995). Predicting corporate failure using a neural network approach. *Intelligent Systems in Accounting, Finance and Management*, *4*(2), 95-111.

Bottasso, A., & Conti, M. (2010). The productive effect of transport infrastructures: does road transport liberalization matter?. *Journal of regulatory economics*, *38*(1), 27-48.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159.

Brauckmann, P. (Ed.). (2010). *Web-Monitoring: Gewinnung und Analyse von Daten über das Kommunikationsverhalten im Internet*. Uvk Verlags GmbH.

Brown, K. A., & Mitchell, T. R. (1993). Organizational obstacles: Links with financial performance, customer satisfaction, and job satisfaction in a service environment. *Human Relations*, *46*(6), 725-757.

Brush, C. G., Greene, P. G., & Hart, M. M. (2001). From initial idea to unique advantage: The entrepreneurial challenge of constructing a resource base. *The academy of management executive*, *15*(1), 64-78.

Burger King Switzerland (2017). https://de.burger-king.ch/

Buscema, M. (1998). Back propagation neural networks. *Substance use & misuse*, *33*(2), 233-270.

Buttner, E. H., & Moore, D. P. (1997). Women's organizational exodus to entrepreneurship: self-reported motivations and correlates with success. *Journal of small business management*, *35*(1), 34.

Carroll, G. R., & Hannan, M. T. (1989). Density delay in the evolution of organizational populations: A model and five empirical tests. *Administrative Science Quarterly*, 411-430.

Carter, R., & Auken, H. V. (2006). Small firm bankruptcy. *Journal of Small Business Management*, *44*(4), 493-512.

Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002, May). The structure of broad topics on the web. In *Proceedings of the 11th international conference on World Wide Web* (pp. 251-262). ACM.

Chang, K. C. (2013). How reputation creates loyalty in the restaurant sector. *International Journal of Contemporary Hospitality Management*, *25*(4), 536-557.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*, 1-12.

Chen, L. F., & Tsai, C. T. (2016). Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management*, *53*, 197-206.

Chittithaworn, C., Islam, M. A., Keawchana, T., & Yusuf, D. H. M. (2011). Factors affecting business success of small & medium enterprises (SMEs) in Thailand. *Asian Social Science*, *7*(5), 180.

Cho, M. H. (1994). *Predicting business failure in the hospitality industry: An application of logit model* (Doctoral dissertation, Virginia Tech).

Cho, S., Woods, R. H., Jang, S. S., & Erdem, M. (2006). Measuring the impact of human resource management practices on hospitality firms' performances. *International Journal of Hospitality Management*, *25*(2), 262-277.

Chong, H. G. (2008). Measuring performance of small-and-medium sized enterprises: the grounded theory approach. *Journal of Business and Public affairs*, *2*(1), 1-10.

Cichy, R. F., Sciarini, M. P., & Patton, M. E. (1992). Food-Servce Leadership: Could Attila Run a Restaurant?. *Cornell Hotel and Restaurant Administration Quarterly*, *33*(1), 46-55.

Cichy, R. F., Aoki, T. T., Patton, M. E., & Hwang, K. Y. (1993). Shidō-sei: Leadership in Japan's Commercial Food-Service Industry. *Cornell Hotel and Restaurant Administration Quarterly*, *34*(1), 88-95.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine learning*, *3*(4), 261-283.

Clark, M. A., & Wood, R. C. (1998). Consumer loyalty in the restaurant industry-a preliminary exploration of the issues. *International Journal of Contemporary Hospitality Management*, *10*(4), 139-144.

Cohen, A. (2011). FuzzyWuzzy: Fuzzy string matching in python.

Cooley, R., Mobasher, B., & Srivastava, J. (1997, November). Web mining: Information and pattern discovery on the world wide web. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on* (pp. 558-567). IEEE.

Cooley, R. W., & Srivastava, J. (2000). *Web usage mining: discovery and application of interesting patterns from web data*. Minneapolis, MN: University of Minnesota.

Costello, T. A., & Small, R. W. (1988). Restaurant entrepreneurs revisited: Systematic vs. intuitive entrepreneurs and the restaurants they operate. *Hospitality Education and Research Journal*, *12*(2), 377-389.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

Dahlqvist, J., Davidsson, P., & Wiklund, J. (2000). Initial conditions as predictors of new venture performance: A replication and extension of the Cooper et al. study. *Enterprise and innovation management studies*, *1*(1), 1-17.

Davidsson, P., Achtenhagen, L., & Naldi, L. (2005). Research on small firm growth: A review.

Denk, M. (2009). A framework for statistical entity identification to enhance data quality. *Insights on Data Integration Methodologies*, 89.

Dienhart, J. R., & Gregoire, M. B. (1993). Job satisfaction, job involvement, job security, and customer focus of quick-service restaurant employees. *Hospitality research journal*, *16*(2), 29-43.

Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, *90*(3), 487-513.

DiPietro, R. B., Parsa, H. G., & Gregory, A. (2011). Restaurant QSC inspections and financial performance: an empirical investigation. *International Journal of Contemporary Hospitality Management*, *23*(7), 982-999.

Dockel, J. A., & Ligthelm, A. A. (2005). Factors responsible for the growth of small businesses: management. *South African Journal of Economic and Management Sciences*, *8*(1), 54-62.

Doumpos, M., & Zopounidis, C. (1999). A multicriteria discrimination method for the prediction of financial distress: The case of Greece. *Multinational Finance Journal*, *3*(2), 71.

Dragos, C., Dragos, S., & Dumitru, A. (2008). Financial scoring: a literature review and experimental study. *Economic and Business Review for Central and South-Eastern Europe*, *10*(1), 53.

Duarte Alonso, A., O'neill, M., Liu, Y., & O'shea, M. (2013). Factors driving consumer restaurant choice: An exploratory study from the Southeastern United States. *Journal of Hospitality Marketing & Management*, *22*(5), 547-567.

Dube, L., & Renaghan, L. M. (1994). Measuring Customer Satisfaction for Strategic Management: For financial success, a restaurant's management must make the connection between service attributes and return patronage. Here's a way to establish that connection. *Cornell Hotel and Restaurant Administration Quarterly*, *35*(1), 39-47.

Duman, E., Ekinci, Y., & Tanrıverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications*, *39*(1), 48-53.

Dwyer, F. R., Schurr, P. H., & Oh, S. (1987). Developing buyer-seller relationships. *The Journal of marketing*, 11-27.

Edwards, J. S. A., & Meiselman, H. L. (2005). The influence of positive and negative cues on restaurant food choice and food acceptance. *International Journal of Contemporary Hospitality Management*, *17*(4), 332-344.

ElasticSearch. https://www.elastic.co/

Elizabeth, C., & Baines, S. (1998). Does gender affect business 'performance'? A study of microbusinesses in business services in the UK. *Entrepreneurship & Regional Development*, *10*(2), 117-135.

Emenheiser, D. A., Clay, J. M., & Palakurthi, R. (1998). Profiles of successful restaurant managers for recruitment and selection in the US. *International Journal of Contemporary Hospitality Management*, *10*(2), 54-62.

English, W. (1996). Restaurant attrition: a longitudinal analysis of restaurant failures. *International Journal of Contemporary Hospitality Management*, *8*(2), 17-20.

Etheridge, H. L., Sriram, R. S., & Hsu, H. Y. (2000). A comparison of selected artificial neural networks that help auditors evaluate client financial viability. *Decision Sciences*, *31*(2), 531-550.

European Union. https://europa.eu/european-union/about-eu/institutions-bodies/european-commission_en

Eurostat (2015). http://ec.europa.eu/eurostat/statistics-explained/index.php/E-business_integration

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. *Methodology and computing in applied probability*, *11*(1), 29-45.

Farouk, A., & Saleh, M. (2011). An explanatory framework for the growth of small and medium enterprises. In *International Conference of System Dynamics Society*.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

Fink, A. (2005). *Conducting research literature reviews: From the internet to paper*. Sage.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical journal*, *47*(4), 458-472.

Frydman, H., Altman, E. I., & KAO, D. L. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. *The Journal of Finance*, *40*(1), 269-291.

Fueglistaller, U., Fust, A., & Brunner, C. (2017). Schweizer KMU. Eine Analyse der aktuellsten Zahlen–Ausgabe 2017.

für Justiz, B. (2001). Zefix - Der zentrale Firmenindex auf Internet. *Reden*, *2000*, 1999.

Garg, A., & Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *International Journal of Modelling, Identification and Control*, *18*(4), 295-312.

Gartner, W. B. (1989). Some suggestions for research on entrepreneurial traits and characteristics. *Entrepreneurship theory and practice*, *14*(1), 27-38.

Gastrosuisse (2017). *Branchenspiegel 2017*. Gastrosuisse, Verband für Hotellerie und Restauration.

Gelinas, R., & Bigras, Y. (2004). The characteristics and features of SMEs: favorable or unfavorable to logistics integration?. *Journal of Small Business Management*, *42*(3), 263-278.

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of forecasting*, *29*(6), 536-555.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653-671.

Goldman, K. (1993). Concept selection for independent restaurants. *Cornell Hotel and Restaurant Administration Quarterly*, *34*(6), 59-72.

Gordon L. Lippitt and Warren H. Schmidt. 1967. Crises in a developing organization. Harvard Business Review.

Gray, K. R., Foster, H., & Howard, M. (2006). Motivations of Moroccans to be entrepreneurs. *Journal of Developmental Entrepreneurship*, *11*(04), 297-318.

Gregory, S. R., Smith, K. D., & Lenk, M. M. (1998). Factors contributing to internal customer satisfaction and commitment in quick service restaurants. *Journal of Restaurant & Foodservice Marketing*, *2*(4), 21-47.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308-319.

Grzymala-Busse, J. W. and Hu, M. (2000). A comparison of several approaches to missing attribute values in data mining. In: International Conference on Rough Sets and Current Trends in Computing pp. 378-385.

Gu, C., & Huang, L. (2008, December). Web Mining in Technology Management. In *Business and Information Management, 2008. ISBIM'08. International Seminar on* (Vol. 2, pp. 88-91). IEEE.

Gu, Z., & Gao, L. (2000). A multivariate model for predicting business failures of hospitality firms. *Tourism and Hospitality Research*, *2*(1), 37-49.

Gupta, S., & Rani, R. G. (2016). *Data Transformation and Query Analysis of Elasticsearch and CouchDB Document Oriented Databases* (Doctoral dissertation).

Gürol, Y., & Atsan, N. (2006). Entrepreneurial characteristics amongst university students: Some insights for entrepreneurship education and training in Turkey. *Education+ Training*, *48*(1), 25-38.

Ha, J., & Jang, S. (2013). Attributes, consequences, and consumer values: A means-end chain approach across restaurant segments. *International Journal of Contemporary Hospitality Management*, *25*(3), 383-409.

Han, H., Back, K. J., & Barrett, B. (2009). Influencing factors on restaurant customers' revisit intention: The roles of emotions and switching barriers. *International Journal of Hospitality Management*, *28*(4), 563-572.

Hannan, M. T., & Freeman, J. (1988). Density dependence in the growth of organizational populations. *Ecological models of organizations*, *7*, 31.

Hastie, T., & Tibshirani, R. & Friedman, J.(2008). The Elements of Statistical Learning; Data Mining, Inference and Prediction.

Henebry, K. L. (1996). Do cash flow variables improve the predictive accuracy of a Cox proportional hazards model for bank failure?. *The quarterly review of economics and finance*, *36*(3), 395-409.

Heo, C. Y. (2016). Exploring group-buying platforms for restaurant revenue management. *International Journal of Hospitality Management*, *52*, 154-159.

Heung, V. C. (2002). American theme restaurants: A study of consumer's perceptions of the important attributes in restaurant selection. *Asia Pacific Journal of Tourism Research*, *7*(1), 19-28.

Heung, V. C., & Gu, T. (2012). Influence of restaurant atmospherics on patron satisfaction and behavioral intentions. *International Journal of Hospitality Management*, *31*(4), 1167-1177.

Hiemstra, S. J., & Kosiba, S. T. (1994). Recession and tax impacts on the US restaurant industry. *Hospitality Research Journal*, *17*(2), 17-23.

Hjalager, A. M. (2000). Organisational ecology in the Danish restaurant sector. *Tourism Management*, *21*(3), 271-280.

Hudson, B. T. (1995). Venture capital in the restaurant industry. *Cornell Hotel and Restaurant Administration Quarterly*, *36*(3), 50-61.

Hwang, J., Yoon, Y. S., & Park, N. H. (2011). Structural effects of cognitive and affective reponses to web advertisements, website and brand attitudes, and purchase intentions: The case of casual-dining restaurants. *International Journal of Hospitality Management*, *30*(4), 897-907.

Hyun, S. S. (2009). Creating a model of customer equity for chain restaurant brand formation. *International Journal of Hospitality Management*, *28*(4), 529-539.

Hyun, S. S., & Perdue, R. R. (2017). Understanding the dimensions of customer relationships in the hotel and restaurant industries. *International Journal of Hospitality Management*, *64*, 73-84.

Ibrahim, A. B., & Goodwin, J. R. (1986). Perceived causes of success in small business. *American journal of small business*, *11*(2), 41-50.

Jablonski, S., & Meiler, C. (2002). Web-Content-Managementsysteme. *Informatik-Spektrum*, *25*(2), 101-119.

Jack Kivela, J. (1997). Restaurant marketing: selection and segmentation in Hong Kong. *International Journal of Contemporary Hospitality Management*, *9*(3), 116-123.

Jacob, C., Guéguen, N., & Boulbry, G. (2014). Using verbal attention to enhance restaurant customer satisfaction and behavior. *International Journal of Hospitality Management*, *39*, 50-52.

Jain, B. A., & Nag, B. N. (1997). Performance evaluation of neural network decision models. *Journal of Management Information Systems*, *14*(2), 201-216.

Jang, S. S., Ha, J., & Park, K. (2012). Effects of ethnic authenticity: Investigating Korean restaurant customers in the US. *International Journal of Hospitality Management*, *31*(3), 990-1003.

Jani, D., & Han, H. (2011). Investigating the key factors affecting behavioral intentions: Evidence from a full-service restaurant setting. *International Journal of Contemporary Hospitality Management*, *23*(7), 1000-1018.

Java, A. (2008). *Mining social media communities and content* (Doctoral dissertation, University of Maryland, Baltimore County).

Jennings, P., & Beaver, G. (1997). The performance and competitive advantage of small firms: a management perspective. *International small business journal*, *15*(2), 63-75.

Johnsen, G. J., & McMahon, R. G. (2005). Owner-manager gender, financial performance and business growth amongst SMEs from Australia's business longitudinal survey. *International Small Business Journal*, *23*(2), 115-142.

Josiam, B. M., & Monteiro, P. A. (2004). Tandoori tastes: Perceptions of Indian restaurants in America. *International Journal of Contemporary Hospitality Management*, *16*(1), 18-26.

Jovanovic, B. (1982). Selection and the Evolution of Industry. *Econometrica: Journal of the Econometric Society*, 649-670.

Kalakota, R., & Robinson, M. (2001). e-Business 2.0: Roadmap forSuccess. *Massachusetts: Addsion Wesley Longman Inc*.

Kalleberg, A. L., & Leicht, K. T. (1991). Gender and organizational performance: Determinants of small business survival and success. *Academy of management journal*, *34*(1), 136-161.

Kang, J., Tang, L., & Fiore, A. M. (2015). Restaurant brand pages on Facebook: do active member participation and monetary sales promotions matter?. *International Journal of Contemporary Hospitality Management*, *27*(7), 1662-1684.

Kara, A., Kaynak, E., & Kucukemiroglu, O. (1995). Marketing strategies for fast-food restaurants: a customer view. *International Journal of Contemporary Hospitality Management*, *7*(4), 16-22.

Kim, H. B., & Kim, W. G. (2005). The relationship between brand equity and firms' performance in luxury hotels and chain restaurants. *Tourism management*, *26*(4), 549-560.

Kim, H. J., Park, J., Kim, M. J., & Ryu, K. (2013). Does perceived restaurant food healthiness matter? Its influence on value, satisfaction and revisit intentions in restaurant operations in South Korea. *International Journal of Hospitality Management*, *33*, 397-405.

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, *201*(3), 838-846.

Kim, H., & Gu, Z. (2006). A logistic regression analysis for predicting bankruptcy in the hospitality industry. *The Journal of Hospitality Financial Management*, *14*(1), 17-34.

Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, *53*(11), 3735-3745.

Kim, J. H., & Jang, S. (2016). Determinants of authentic experiences: An extended Gilmore and Pine model for ethnic restaurants. *International Journal of Contemporary Hospitality Management*, *28*(10), 2247-2266.

Kim, J. H., Youn, H., & Rao, Y. (2017). Customer responses to food-related attributes in ethnic restaurants. *International Journal of Hospitality Management*, *61*, 129-139.

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, *36*, 354-362.

Kim, W. G., & Kim, H. B. (2004). Measuring customer-based restaurant brand equity. *Cornell Hotel and Restaurant Administration Quarterly*, *45*(2), 115-131.

Kim, W. G., & Moon, Y. J. (2009). Customers' cognitive, emotional, and actionable response to the servicescape: A test of the moderating effect of the restaurant type. *International journal of hospitality management*, *28*(1), 144-156.

Kimes, S. E. (2004). Restaurant revenue management: implementation at Chevys Arrowhead. *Cornell Hotel and Restaurant Administration Quarterly*, *45*(1), 52-67.

Klarer, M. J., Tran, M., Chi, M., & Marchich, M. United Nations Industrial Development Organization.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, *14*(2), 1137-1145.

Kong, A., Nguyen, V., & Xu, C. Predicting International Restaurant Success with Yelp.

Koran, J. M. (2010). *U.S. Patent No. 7,844,605*. Washington, DC: U.S. Patent and Trademark Office.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, *2*(1), 1-15.

Kraut, R. E., & Grambsch, P. (1987). Home-based white collar employment: Lessons from the 1980 Census. *Social Forces*, *66*(2), 410-426.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, *40*(13), 5125-5131.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York: Springer.

Lacher, R. C., Coats, P. K., Sharma, S. C., & Fant, L. F. (1995). A neural network for classifying the financial health of a firm. *European Journal of Operational Research*, *85*(1), 53-65.

Lawson, R. (2015). *Web scraping with Python*. Packt Publishing Ltd.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113-1130.

Lee, J. J. (2013). Mechanize: Stateful programmatic web browsing in Python.

Lev, B., Petrovits, C., & Radhakrishnan, S. (2010). Is doing good good for you? How corporate chariTable contributions enhance revenue growth. *Strategic Management Journal*, *31*(2), 182-200.

Lewandowski, D. (2005). *Web Information Retrieval: Technologien zur Informationssuche im Internet*. DGI.

Lewis, R. C. (1985). Predicting hotel choice: The factors underlying perception. *Cornell Hotel and Restaurant Administration Quarterly*, *25*(4), 82-96.

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*, *37*(8), 5895-5904.

Ligthelm, A. A., & Cant, M. C. (2002). *Business success factors of SMEs in Gauteng: A proactive entrepeneurial approach*. Bureau of Market Research, University of South Africa.

Linder, A. (2005). Web Mining-Die Fallstudie Swarovski. 1. *Auflage. Wiesbaden: Deutscher Universitäts-Verlag*.

Linkedin policy (2017). https://www.linkedin.com/legal/privacy-policy

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Liu, B., & Chen-Chuan-Chang, K. (2004). Special issue on web content mining. *Acm Sigkdd explorations newsletter*, *6*(2), 1-4.

Liu, Y., & Jang, S. S. (2009). Perceptions of Chinese restaurants in the US: what affects customer satisfaction and behavioral intentions?. *International Journal of Hospitality Management*, *28*(3), 338-348.

Lombardi, D. (1996). Trends and directions in the chain-restaurant industry. *Cornell Hotel and Restaurant Administration Quarterly*, *37*(3), 14-17.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256-265.

Lopez-Gracia, J., & Aybar-Arias, C. (2000). An empirical approach to the financial behaviour of small and medium sized companies. *Small Business Economics*, *14*(1), 55-63.

Lukács, E. (2005). The economic role of SMEs in world economy, especially in Europe. *European integration studies*, *4*(1), 3-12.

Lunati, M. (2000). SMEs and electronic commerce: An overview.

Lussier, R. N., & Halabi, C. E. (2010). A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management*, *48*(3), 360-377.

Magnini, V. P., Garcia, C., & Honeycutt Jr, E. D. (2010). Identifying the attributes of an effective restaurant chain endorser. *Cornell Hospitality Quarterly*, *51*(2), 238-250.

Malarvizhi, R., & Saraswathi, K. (2013). Web Content Mining Techniques Tools & Algorithms–A Comprehensive Study. *International Journal of Computer Trends and Technology (IJCTT)*, *4*(8), 2940-2945.

Man, T. W., Lau, T., & Chan, K. F. (2002). The competitiveness of small and medium enterprises: A conceptualization with focus on entrepreneurial competencies. *Journal of business venturing*, *17*(2), 123-142.

Manly, B. F., & Alberto, J. A. N. (2016). *Multivariate statistical methods: a primer*. CRC Press.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician*, *36*(3a), 158-160.

Markov, Z., & Larose, D. T. (2007). *Data mining the Web: uncovering patterns in Web content, structure, and usage*. John Wiley & Sons.

Marzal, A., & Vidal, E. (1993). Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence*, *15*(9), 926-932.

Mathe-Soulek, K., Slevitch, L., & Dallinger, I. (2015). Applying mixed methods to identify what drives quick service restaurant's customer satisfaction at the unit-level. *International Journal of Hospitality Management*, *50*, 46-54.

Mayr, P., & Tosques, F. (2005). Webometrische Analysen mit Hilfe der Google Web APIs. *Information Wissenschaft und Praxis*, *56*(1), 41-48.

Mazzarol, T., Volery, T., Doss, N., & Thein, V. (1999). Factors influencing small business start-ups: a comparison with previous research. *International Journal of Entrepreneurial Behavior & Research*, *5*(2), 48-63.

McCleary, K. W. (1978). The corporate-meetings market: components of success in attracting group business. *Cornell Hotel and Restaurant Administration Quarterly*, *19*(2), 30-35.

McDonald's Switzerland (2017). https://www.mcdonalds.ch/de/restaurants/.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239-245.

Michaelas, N., Chittenden, F., & Poutziouris, P. (1999). Financial policy and capital structure choice in UK SMEs: Empirical evidence from company panel data. *Small business economics*, *12*(2), 113-130.

Miner, G., Elder IV, J., & Hill, T. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

Morgan, M. S. (1995). Assessing Chain-Restaurant Impact - Using Linear Regression. *Cornell Hotel and Restaurant Administration Quarterly*, *36*(3), 30-33.

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *The journal of marketing*, 20-38.

Morland, K., Wing, S., Roux, A. D., & Poole, C. (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine*, *22*(1), 23-29.

Morrissey, W. J., & Pittaway, L. (2006). Buyer-supplier relationships in small firms: the use of social factors to manage relationships. *International Small Business Journal*, *24*(3), 272-298.

Mueller, S. L., & Thomas, A. S. (2001). Culture and entrepreneurial potential: A nine country study of locus of control and innovativeness. *Journal of business venturing*, *16*(1), 51-75.

Muller, C. C. (1999). A simple measure of restaurant efficiency. *Cornell Hotel and Restaurant Administration Quarterly*, *40*(3), 31-37.

Muller, C., & Inman, C. (1996). Characteristics and behavior of top chain-restaurant CEOs. *Cornell Hotel and Restaurant Administration Quarterly*, *37*(3), 64-69.

Mun, S. G., & Jang, S. S. (2015). Working capital, cash holding, and profitability of restaurant firms. *International Journal of Hospitality Management*, *48*, 1-11.

Murphy, J., Forrest, E., & Wotring, C. E. (1996). Restaurant marketing on the worldwide web. *Cornell Hotel and Restaurant Administration Quarterly*, *37*(1), 61-71.

Mutula, S. M., & van Brakel, P. (2006). E-readiness of SMEs in the ICT sector in Botswana with respect to information access. *The electronic library*, *24*(3), 402-417.

Naffziger, D. (1995). Entrepreneurship: A person based theory approach. *Advances in entrepreneurship, firm emergence, and growth*, *2*, 21-50.

Nam, J. H., & Lee, T. J. (2011). Foreign travelers' satisfaction with traditional Korean restaurants. *International Journal of Hospitality Management*, *30*(4), 982-989.

Namkung, Y., & Jang, S. (2008). Are highly satisfied restaurant customers really different? A quality perception perspective. *International Journal of Contemporary Hospitality Management*, *20*(2), 142-155.

Namkung, Y., & Jang, S. (2010). Service failures in restaurants: Which stage of service failure is the most critical?. *Cornell Hospitality Quarterly*, *51*(3), 323-343.

Nandola, K., Koshal, M., & Koshal, R. K. (1982). Forecasting restaurant food sales. *Cornell Hotel and Restaurant Administration Quarterly*, *23*(2), 92-96.

NCSL (2016). http://www.ncsl.org/research/financial-services-and-commerce/use -of-credit-information-in-insurance-2016-legislation.aspx

Neophytou, E., & Molinero, C. M. (2004). Predicting corporate failure in the UK: a multidimensional scaling approach. *Journal of Business Finance & Accounting*, *31*(5-6), 677-710.

Neuman, W. (2009). Discounts have restaurants eating own lunch.

Noone, B. M., Kimes, S. E., Mattila, A. S., & Wirtz, J. (2009). Perceived service encounter pace and customer satisfaction: An empirical study of restaurant experiences. *Journal of Service Management*, *20*(4), 380-403.

Norris, M., Oppenheim, C., & Rowland, F. (2008, November). Open Access Citation Rates and Developing Countries. In *ELPUB* (pp. 335-342).

Oates, K. S. (2015). *A logistic regression analysis of score sending and college matching among high school students*. The University of Iowa.

O'brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, *74*(368), 877-880.

O'Farrell, P. N., & Hitchens, D. M. (1988). Alternative theories of small-firm growth: a critical review. *Environment and Planning A*, *20*(10), 1365-1383.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

Ok, C., Back, K. J., & Shanklin, C. W. (2006). Service recovery paradox: Implications from an experimental study in a restaurant setting. *Journal of Hospitality & Leisure Marketing*, *14*(3), 17-33.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research.

Olsen, M., Bellas, C., & Kish, L. V. (1983). Improving the prediction of restaurant failure through ratio analysis. *International Journal of Hospitality Management*, *2*(4), 187-193.

Olson, P. D., & Bokor, D. W. (1995). Strategy process-content interaction: Effects on growth performance in small, start-up firms. *Journal of small business management*, *33*(1), 34.

Olson, E. M., Slater, S. F., & Hult, G. T. M. (2005). The performance implications of fit among business strategy, marketing organization structure, and strategic behavior. *Journal of marketing*, *69*(3), 49-65.

OSM Data for Switzerland (2016). http://planet.osm.ch/

Ozgulbas, N., & Koyuncugil, A. S. (2012). Risk Classification of SMEs by Early Warning Model Based on Data Mining. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, *6*(10), 2649-2660.

Pages, Y. (2002). E-Business Report: the online experience of small and medium enterprises. *Melbourne: Pacific Access for Telstra Corporation Limited*.

Pant, G., & Menczer, F. (2002). MySpiders: Evolve your own intelligent Web crawlers. *Autonomous agents and multi-agent systems*, *5*(2), 221-229.

Park, C. (2004). Efficient or enjoyable? Consumer values of eating-out and fast food restaurant consumption in Korea. *International Journal of Hospitality Management*, *23*(1), 87-94.

Park, K., & Khan, M. A. (2006). An exploratory study to identify the site selection factors for US franchise restaurants. *Journal of Foodservice Business Research*, *8*(1), 97-114.

Park, K., & Jang, S. S. (2012). Duration of advertising effect: Considering franchising in the restaurant industry. *International Journal of Hospitality management*, *31*(1), 257-265.

Park, K., & Jang, S. (2015). The cyclical effect of advertising: Is reducing restaurant advertising appropriate in periods of economic contraction?. *International Journal of Contemporary Hospitality Management*, *27*(7), 1386-1408.

Parsa, H. G., & Kahn, M. A. (1991). Menu trends in the quick service restaurant industry during the various stages of the industry life cycle (1919-1988). *Hospitality Research Journal*, *15*(1), 93-109.

Parsa, H. G., & Khan, M. A. (1993). Quick-Service Restaurants of the 21 St Century: an Analytical Review of Macro Factors. *Hospitality Research Journal*, *17*(1), 161-173.

Parsa, H. G., Self, J. T., Njite, D., & King, T. (2005). Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, *46*(3), 304-322.

Parsa, H. G., van der Rest, J. P. I., Smith, S. R., Parsa, R. A., & Bujisic, M. (2015). Why restaurants fail? Part IV: The relationship between restaurant failures and demographic factors. *Cornell Hospitality Quarterly*, *56*(1), 80-90.

Patel, K. B., Chauhan, J. A., & Patel, J. D. (2011). Web mining in e-commerce: Pattern discovery, issues and applications. *International Journal of P2P Network Trends and Technology*, *1*(3), 40-45.

Patil, D. V., & Bichkar, R. S. (2012). Issues in optimization of decision tree learning: A survey. *International Journal of Applied Information Systems (IJAIS), New York, USA*, *3*(5).

Paul, R. N. (1994). Status and outlook of the chain-restaurant industry. *Cornell Hotel and Restaurant Administration Quarterly*, *35*(3), 23-26.

Pearce, J. A., Robinson, R. B., & Subramanian, R. (2000). *Strategic management: Formulation, implementation, and control*. Columbus, OH: Irwin/McGraw-Hill.

Pedraja, M., & Yagüe, J. (2001). What information do customers use when choosing a restaurant?. *International Journal of Contemporary Hospitality Management*, *13*(6), 316-318.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*, 2825-2830.

Perren, L. (2000). Factors in the growth of micro-enterprises (part 2): exploring the implications. *Journal of small business and enterprise development*, *7*(1), 58-68.

Petticrew, M., & Roberts, H. (2006). How to appraise the studies: an introduction to assessing study quality. *Systematic reviews in the social sciences: A practical guide*, 125-163.

Post, H. A. (1997). Building a strategy on competences. *Long Range Planning*, *30*(5), 733-740.

PostgreSQL. https://www.postgresql.org/

Poynter, J. T. (1992). Georgia's bed and breakfast inn industry: An exploratory study to identify success factors.

Prodromou, T. (2012). *Ultimate Guide to LinkedIn for Business: How To Get Connected with 130 Million Customers in 10 Minutes*. Entrepreneur Press.

Provan, K. G., Fish, A., & Sydow, J. (2007). Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of management*, *33*(3), 479-516.

Purves, R. D. (1992). Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). *Journal of pharmacokinetics and biopharmaceutics*, *20*(3), 211-226.

Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.

Rammer, C., Kinne, J., & Blind, K. (2016). Microgeography of innovation in the city: Location patterns of innovative firms in Berlin.

Regulation, G. D. P. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, *59*, 1-88.

Reynolds, D., & Biel, D. (2007). Incorporating satisfaction measures into a restaurant productivity index. *International Journal of Hospitality Management*, *26*(2), 352-361.

Reynolds, P. D., Hay, M., Bygrave, W. D., Camp, S. M., & Autio, E. (2000). Global entrepreneurship monitor: 2000 executive report.

Richard, O. C. (2000). Racial diversity, business strategy, and firm performance: A resource-based view. *Academy of management journal*, *43*(2), 164-177.

Richardson, L. (2013). Beautiful soup. *Crummy: The Site*.

Rogers, I. (2002). The Google Pagerank algorithm and how it works. *URL: http://www.iprcom.com/papers/pagerank/index.html*.

Ryu, K., Lee, H. R., & Gon Kim, W. (2012). The influence of the quality of the physical environment, food, and service on restaurant image, customer perceived value, customer satisfaction, and behavioral intentions. *International Journal of Contemporary Hospitality Management*, *24*(2), 200-223.

Saini, S., & Pandey, H. M. (2015). Review on web content mining techniques. *International Journal of Computer Applications*, *118*(18).

Sandberg, K. W. (2003). An exploratory study of women in micro enterprises: gender-related differences. *Journal of small business and enterprise development*, *10*(4), 408-417.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, *1*(3), 261-377.

Schrenk, M. (2012). *Webbots, spiders, and screen scrapers: A guide to developing Internet agents with PHP/CURL*. No Starch Press.

Schultze, M., & Postler, A. (2008). Online-Trend-Monitoring bei der EnBW: Mit dem Ohr am Kunden. *Kommunikation, Partizipation und Wirkungen im Social Web*, *2*, 370-382.

Scientific Journal Ranking. https://www.scimagojr.com/journalrank.php

Scott, M., & Bruce, R. (1987). Five stages of growth in small business. *Long range planning*, *20*(3), 45-52.

SECO. https://www.seco.admin.ch/seco/en/home.html

Seo, S., & Hwang, J. (2014). Does gender matter? Examining gender composition's relationships with meal duration and spending in restaurants. *International Journal of Hospitality Management*, *42*, 61-70.

Sharda, D., & Chawla, S. Web Content Mining Techniques-A Study. *International Journal of Innovative Research in Technology and Science (IJIRTS)*.

Shriber, M., Muller, C., & Inman, C. (1995). Population changes and restaurant success. *Cornell Hotel and Restaurant Administration Quarterly*, *36*(3), 43-49.

Sieber, P. (2002). Einsatz und Nutzung des Internet in kleinen und mittleren Unternehmen in der Schweiz 2000: Bern, Dr. Pascal Sieber & Partners AG. *URL: www.pascal-sieber. ch/Files/kmuinfo-2000.pdf*.

Sigrist, F., & Hirnschall, C. (2018). Gradient Tree-Boosted Tobit Models for Default Prediction. *arXiv preprint arXiv:1711.08695*.

Silber, I., Israeli, A., Bustin, A., & Zvi, O. B. (2009). Recovery strategies for service failures: The case of restaurants. *Journal of Hospitality Marketing & Management*, *18*(7), 730-740.

Simpson, M., Tuck, N., & Bellamy, S. (2004). Small business success factors: the role of education and training. *Education+ Training*, *46*(8/9), 481-491.

SJR. https://www.scimagojr.com/

Small, O. E. C. D. (2004). Medium-sized enterprises in Turkey: issues and policies.

Song, Q., Shepperd, M., Chen, X., & Liu, J. (2008). Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation. *Journal of Systems and software*, *81*(12), 2361-2370.

Spangler, S., Proctor, L., & Chen, Y. (2008, December). Multi-Taxonomy: Determining perceived brand characteristics from web data. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*(Vol. 1, pp. 258-264). IEEE.

Sparks, B., Bowen, J., & Klag, S. (2003). Restaurants and the tourist market. *International Journal of Contemporary Hospitality Management*, *15*(1), 6-13.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, *1*(2), 12-23.

Srivastava, T., Desikan, P., & Kumar, V. (2005). Web mining–concepts, applications and research directions. In *Foundations and advances in data mining* (pp. 275-307). Springer, Berlin, Heidelberg.

Starbucks Switzerland (2017). http://www.starbucks.ch/store-locator/search.

Steinhauser, S., Schumacher, M., & Rücker, G. (2016). *Determining optimal cut-offs in the meta-analysis of diagnostic test accuracy studies* (Doctoral dissertation).

Steinmetz, L. L. (1969). Critical stages of small business growth: When they occur and how to survive them. *Business horizons*, *12*(1), 29-36.

Stewart Jr, W. H., Carland, J. C., Carland, J. W., Watson, W. E., & Sweo, R. (2003). Entrepreneurial dispositions and goal orientations: A comparative exploration of United States and Russian entrepreneurs. *Journal of small business management*, *41*(1), 27-46.

Storey, D. J. (1994). Understanding the Small Business Sector Routledge London Google Scholar.

Stumme, G., Hotho, A., & Berendt, B. (2006). Semantic web mining: State of the art and future directions. *Web semantics: Science, services and agents on the world wide web*, *4*(2), 124-143.

Subway Switzerland (2017). http://www.subway-sandwiches.ch/restaurants.php.

Swierczek, F. W., & Ha, T. T. (2003). Entrepreneurial orientation, uncertainty avoidance and firm performance: an analysis of Thai and Vietnamese SMEs. *The International Journal of Entrepreneurship and Innovation*, *4*(1), 46-58.

Swiss Federal Statistical Office (2016). http://www.bfs.admin.ch/bfs/portal/en/index/infothek/onlinedb/stattab.html.

Swiss Federal Tax Administration (2016). https://www.estv.admin.ch/.

Temtime, Z. T., & Pansiri, J. (2004). Small business critical success/failure factors in developing countries: some evidences from Botswana.

Thabane, L., Thomas, T., Ye, C., & Paul, J. (2009). Posing the research question: not so simple. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, *56*(1), 71.

The World Business Council for Sustainable Development 2007. *A business guide to development actors.* Switzerland: Atar Roto Presse SA.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010). Mining ideas from textual information. *Expert Systems with Applications*, *37*(10), 7182-7188.

Thorleuchter, D., & Van Den Poel, D. (2012). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, *39*(17), 13026-13034.

TripAdvisor (2017). https://www.tripadvisor.ch/.

Tse, E. C. Y., & Olsen, M. D. (1988). The impact of strategy and structure on the organizational performance of restaurant firms. *Hospitality Education and Research Journal*, *12*(2), 265-276.

Tse, E., & Olsen, M. D. (1990). Business strategy and organisational structure: a case of US restaurant firms. *International Journal of Contemporary Hospitality Management*, *2*(3).

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225-1231.

Tzeng, G. H., Teng, M. H., Chen, J. J., & Opricovic, S. (2002). Multicriteria selection for a restaurant location in Taipei. *International journal of hospitality management*, *21*(2), 171-187.

Van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129-140.

Velásquez, J. D. (2013). Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments. *Expert Systems with Applications*, *40*(13), 5228-5239.

W3C. https://www.w3.org/standards/

Walker, E., & Brown, A. (2004). What success factors are important to small business owners?. *International small business journal*, *22*(6), 577-594.

Wasilczuk, J. (2000). Advantageous competence of owner/managers to grow the firm in Poland: Empirical evidence. *Journal of small business management*, *38*(2), 88.

Watson, J. (2003). Failure Rates for Female-Controlled Businesses: Are They Any Different?. *Journal of small business management*, *41*(3), 262-277.

Web, P. API Dashboard, 2007.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11-12), 1131-1152.

Wijst, D. (1989). *Financial structure in small business: Theory, tests and applications*. Springer Berlin Heidelberg.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

Winkler, W. E. (2006). Overview of record linkage and current research directions. In *Bureau of the Census*.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Woldie, A., Leighton, P., & Adesua, A. (2008). Factors influencing small and medium enterprises (SMEs): an exploratory study of owner/manager and firm characteristics.

Wong, B. K., Lai, V. S., & Lam, J. (2000). A bibliography of neural network business applications research: 1994–1998. *Computers & Operations Research*, *27*(11-12), 1045-1076.

Worthington, I., & Britton, C. (2009). *The business environment*. Pearson Education.

Wu, C. H. J., & Liang, R. D. (2009). Effect of experiential value on customer satisfaction with service encounters in luxury-hotel restaurants. *International Journal of Hospitality Management*, *28*(4), 586-593.

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision support systems*, *50*(4), 743-754.

Yamane, T. (1973). Statistics: An introductory analysis.

Yang, Y., Roehl, W. S., & Huang, J. H. (2017). Understanding and projecting the restaurantscape: The influence of neighborhood sociodemographic characteristics on restaurant location. *International Journal of Hospitality Management*, *67*, 33-45.

Yeh, S. S., & Huan, T. C. (2017). Assessing the impact of work environment factors on employee creative performance of fine-dining restaurants. *Tourism Management*, *58*, 119-131.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

Youn, H., & Gu, Z. (2010). Predict US restaurant firm failures: The artificial neural network model versus logistic regression model. *Tourism and Hospitality Research*, *10*(3), 171-187.

Zaiane, O. R., Xin, M., & Han, J. (1998, April). Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on* (pp. 19-29). IEEE.

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research*, *116*(1), 16-32.

Zhang, Z., Zhang, Z., & Law, R. (2013). Regional effects on customer satisfaction with restaurants. *International Journal of Contemporary Hospitality Management*, *25*(5), 705-722.

Appendix

Business environment		Factor type	Growth factor	Literature
Internal environment	Characteristics of firm	Firm attributes	Age of firm	Hjalager (2000); Namkung (2010)
			Size	Hjalager (2000); Muller (1999); Namkung (2010); Parsa (2005, 2015)
			Location	Chen (2016); Jack Kivela (1997); Mathe-Soulek (2015); Parsa (2015); Tzeng (2002)
			Reputation	Baek (2006); Chang (2013); Cichy (1992); Hwang (2011); Hyun (2009); Kang (2015); Kim (2004, 2005); Parsa (1993);
			Service quality	Baek (2006); DiPietro (2011); Duarte Alonso (2013); Jani (2011); Josiam (2004); Kim (2004, 2005); Mathe-Soulek (2015); Namkung (2008); Noone (2009); Ok (2006); Parsa (1993, 2005); Park (2004); Silber (2009); Wu (2009)
			Physical environment	Baek (2006); DiPietro (2011); Duarte Alonso (2013); Jack Kivela (1997); Jang (2012); Josiam (2004); Kara (1995); Liu (2009); Mathe-Soulek (2015); Namkung (2008); Park (2004); Pedraja (2001); Ryu (2012); Heung (2012); Wu (2009)
			Type of restaurant	Clark (1998); Goldman (1993); Ha (2013); Jack Kivela (1997); Kim (2009)
			Kitchen & service operation	Costello (1998); Kimes (2004)
		Firm resources	Financial resources	English (1996); Parsa (2005); Poynter (1992)
			Human capital	Gregory (1998); Muller (1996); Mun (2015); Poynter (1992); Yeh (2017)
		Firm strategies	Marketing / innovation	English (1996); Gregory (1998); Hwang (2011); Jack Kivela (1997); Magnini (2010); Kang (2015); Kara (1995); Park (2004, 2012, 2015); Poynter (1992); Zhang (2013)
			Restaurant concept	Parsa (2005)
			Service cycle optimization	Heo (2016); Kimes (2004)
			Business / menu planning	Hudson (1995)
			HR management	Borde (1998); Diehnhart (1993); Parsa (1993)
		Food	Price	Baek (2006); Hiemstra (1994); Jani (2011); Joasim (2004); Kara (1995); Kimes (2004); Mathe-Soulek (2015); Nam (2011); Nandola (1982); Park (2004); Parsa (1993); Pedraja (2001); Zhang (2013)
			Quality	Baek (2006); Duarte Alonso (2013); Jang (2012); Joasim (2004); Kim (2013, 2016, 2017); Liu (2009); Mathe-Soulek (2015); Namkung (2008); Park (2004)
			Type	Clark (1998); Ha (2013); Josiam (2004); Kim (2013)
			Variety of menu	Baek (2006); Josiam (2004); Kara (1995); Mathe-Soulek (2015)
		Organization structure	Work specialization	Hjalager (2000); Parsa (2015); Tse (1988)
			Centralization	Tse (1988)
			Legal form	Tse (1988)

Appendix A1: Systematic review on growth factors of restaurants (part 1). For simplification, only the first author of the reviewed studies is denoted.

Business environment		Factor type	Growth factor	Literature
Internal environment	**Characteristics of entrepreneur**	Socio-demographic	Age of entrepreneur	Emenheiser (1998)
			Family background	Emenheiser (1998)
			Education	Muller (1996)
			Experience	Hudson (1995); Kim (2016); Parsa (2005)
		Personality	Need for achievement	Cichy (1992);Lewis (1985); Parsa (2005); Poynter (1992)
			Locus of control	Kim (2013); Parsa (2005)
			Attitude	Parsa (2005)
		Competences	Managerial	Cichy (1992, 1993); Kimes (2004); Muller (1996); Mun (2015); Parsa (2005); Pyonter (1992); Tse (1990)
			Entrepreneurial	Hudson (1995)
External environment	**Immediate environment**	Customer relationships	Customer / market needs	Lewis (1985); McCleary (1978); Parsa (1991)
			Customer acquisition	Hyun (2017)
			Customer retention	Chang (2013); Clark (1998); Dubé (1994); Hyun (2017)
			Customer satisfaction & feedback	Dubé (1994); Chang (2013); Edwards (2005); Han (2009); Jacob (2014); Kim (2009); Reynolds (2007)
		Network	Inter-organizational links	Hjalager (2000)
		Competition	Cluster of restaurants	Hudson (1995); Morgan (1995); Parsa (1993, 2005); Paul (1994); Tzeng (2002)
			Food pricing	Hudson (1995); Morgan (1995); Parsa (1993, 2005); Paul (1994); Tzeng (2002)
	Contextual environment	Technological	Infrastructure	Chen (2016); Tzeng (2002)
		Socio-cultural	Tourism	Sparks (2003)
			Social class	Goldman (1993); Parsa (2015)
			Lifestyle	Goldman (1993)
			Cultural diversity	Muller (1996)
		Economical	Taxation	Borde (1998); Hiemstra (1994)
		Demographical	Population size, growth & density	Goldman (1993); Shriber (1995, 2016); Yang (2017); Zhang (2013)
			Age & gender distribution	Goldman (1993); Lombardi (1996); Nandola (1982); Parsa (2015); Seo (2014); Yang (2017)
			Employment & income	Hiemstra (1994); Nandola (1982); Paul (1994); Shriber (2016)
			Education level	Zhang (2013)
			Household size	Goldman (1993); Hiemstra (1994); Nandola (1982); Parsa (2015); Shriber (2016); Yang (2017)

Appendix A2: Systematic review on growth factors of restaurants (part 2). For simplification, only the first author of the reviewed studies is denoted.

Classification algorithm	**Hyper-parameter**
RFC	n_estimator criterion max_features max_depth min_samples_split min_samples_leaf class_weight bootstrap
MLP	hidden_layer_size activation solver alpha learning_rate learning_rate_init power_t momentum beta_1 beta_2
LR	C penalty solver max_iter fit_intercept tol class_weight

Appendix A3: Models' hyper-parameters to be optimized. For detailed explanation of the hyper-parameters, see Pedregosa et al. (2011).

RFC Performance

Modeling	Roc_auc	Accuracy	Sensitivity	Specificity
1	75.6%	62.9%	77.8%	57.7%
2	66.7%	77.1%	33.3%	92.3%
3	70.9%	60.0%	88.9%	50.0%
4	68.4%	68.6%	66.7%	69.2%
5	62.0%	77.1%	44.4%	88.5%
6	70.9%	68.6%	66.7%	69.2%
7	71.4%	68.6%	66.7%	69.2%
8	59.8%	65.7%	55.6%	69.2%
9	70.5%	77.1%	66.7%	80.8%
10	64.5%	54.3%	88.9%	42.3%
Mean	**68.1%**	**68.0%**	**65.6%**	**68.8%**
Std	**4.6%**	**7.3%**	**16.8%**	**15.0%**

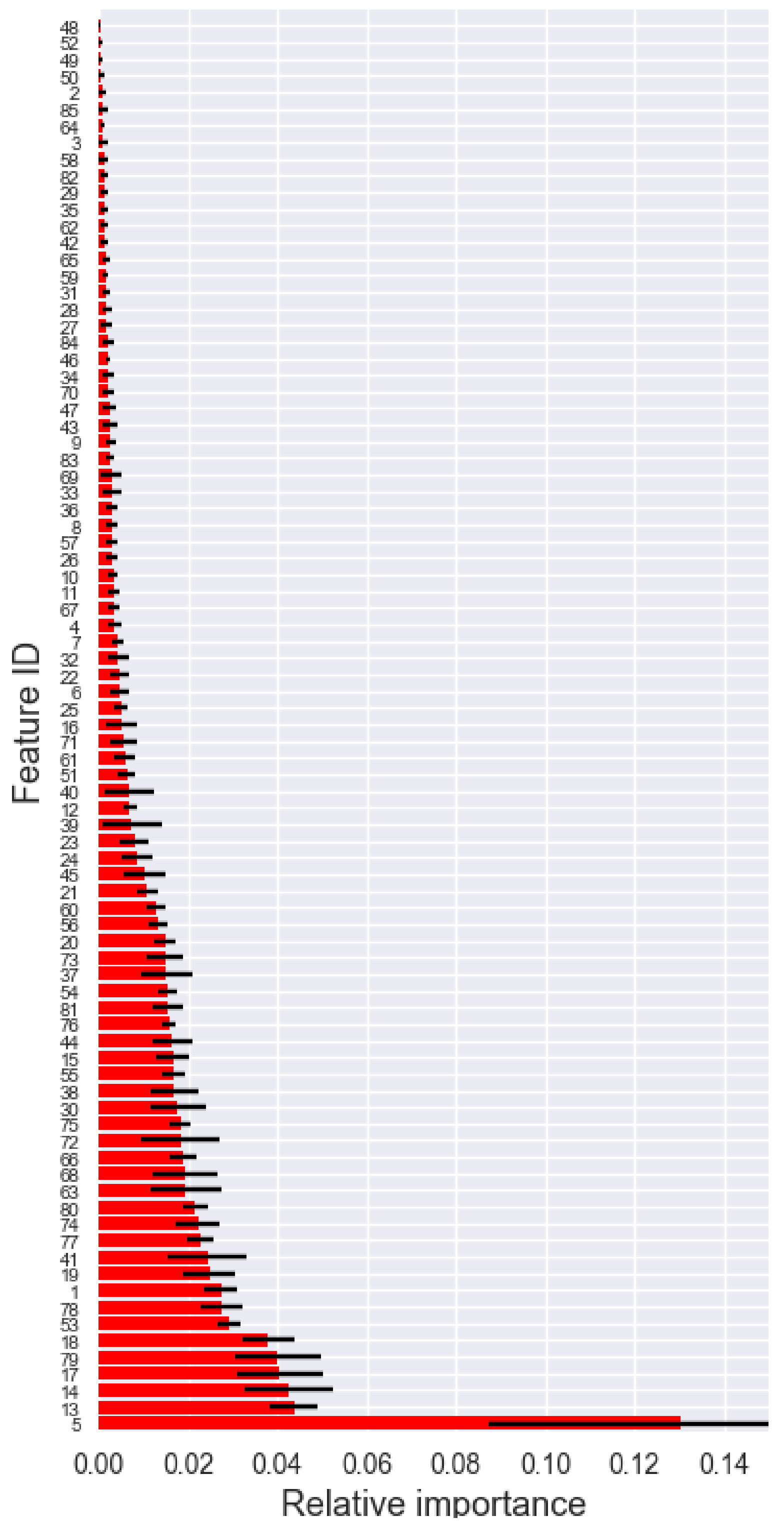
LR Performance

Modeling	Roc_auc	Accuracy	Sensitivity	Specificity
1	78.2%	71.4%	88.9%	65.4%
2	66.7%	65.7%	55.6%	69.2%
3	64.5%	57.1%	66.7%	53.8%
4	55.6%	62.9%	44.4%	69.2%
5	64.5%	68.6%	44.4%	76.9%
6	68.8%	65.7%	66.7%	65.4%
7	55.6%	54.3%	66.7%	50.0%
8	71.8%	68.6%	66.7%	69.2%
9	57.3%	71.4%	33.3%	84.6%
10	75.2%	77.1%	66.7%	80.8%
Mean	**65.8%**	**66.3%**	**60.0%**	**68.5%**
Std	**7.6%**	**6.5%**	**15.1%**	**10.3%**

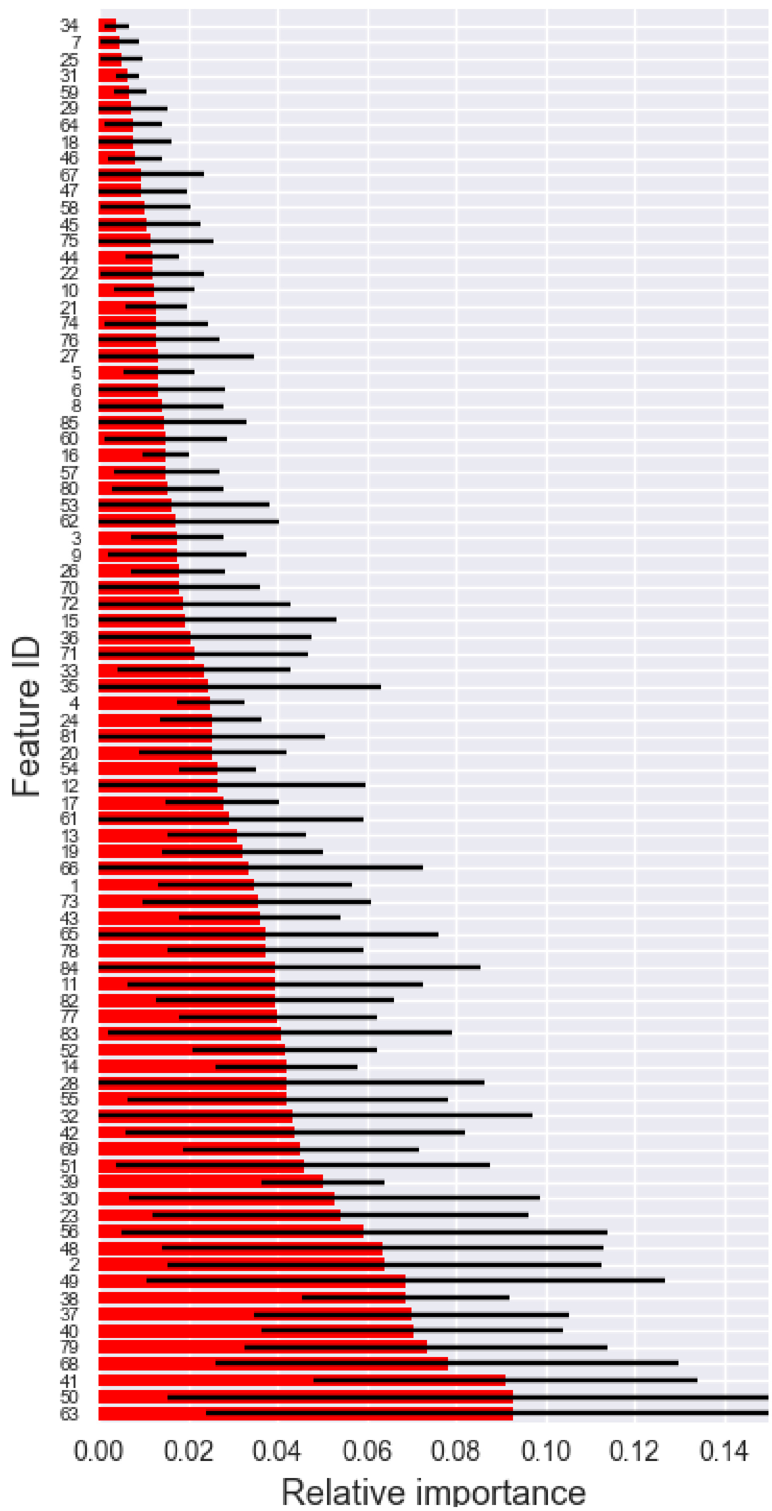
MLP Performance

Modeling	Roc_auc	Accuracy	Sensitivity	Specificity
1	72.2%	54.3%	88.9%	42.3%
2	64.5%	65.7%	55.6%	69.2%
3	59.4%	60.0%	55.6%	61.5%
4	51.3%	45.7%	77.8%	34.6%
5	60.7%	57.1%	66.7%	53.8%
6	53.4%	54.3%	66.7%	50.0%
7	71.4%	57.1%	88.9%	46.2%
8	58.5%	65.7%	77.8%	61.5%
9	63.2%	54.3%	77.8%	46.2%
10	65.0%	62.9%	66.7%	61.5%
Mean	**62.0%**	**57.7%**	**72.2%**	**52.7%**
Std	**6.5%**	**5.8%**	**11.4%**	**10.2%**

Appendix A4: Model performances in detail.



Appendix A5: Complete mean feature importance plot of RFCs.



Appendix A6: Complete mean feature importance plot of LRs.

Curriculum Vitae

Personal Information

Name	Yiea-Funk Te
Date of birth	04.04.1985
Place of birth	Emmen, Switzerland
Nationality	Switzerland
Contact	yfunk.te@gmx.ch

Education

01/2015 – 06/2018	ETH Zürich, The Department Management, Technology and Economics, Zurich, Switzerland (Doctoral studies)
05/2010 – 03/2012	University of Zurich, Switzerland, Master of Science in Physics
09/2005 – 05/2010	University of Zurich, Switzerland, Bachelor of Science in Physics

Professional Experience

07/2012 – 12/2013	Kistler Instrumente AG, Winterthur, Switzerland: Measurement Engineer