PREDICTING THE GROWTH OF RESTAURANTS USING WEB DATA

Yiea-Funk Te
ETH Zurich, Switzerland
fte@ethz.ch

Daniel Müller
ETH Zurich, Switzerland
danielmueller@ethz.ch

Sebastian Wyder
Die Mobiliar, Switzerland
sebastian.wyder@mobi.ch

Dwian Pramono
Die Mobiliar, Switzerland
dwian.pramono@mobi.ch

ABSTRACT
The gastronomy industry plays an important role in 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. However, restaurants are facing tough challenges because of the recent economic turmoil. Despite their importance for the economic growth, limited attention has been paid to predicting restaurant growth. In this study, we propose the use of web mining techniques for restaurant growth prediction as a novel approach. Web mining enables automatic and large-scale collection and analysis of potentially valuable data from various online platforms, thus bearing a great potential for growth prediction. First, a systematic literature review on growth factors is conducted, which serves as a base to collect growth-relevant information from the web. Next, web mining methods are applied to automatically collect and extract growth factors from various web data sources. Finally, 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. Results show that random forests on web data outperform both logistic regressions and artificial neural networks and therefore are recommended for further investigations on predictive modeling of restaurant growth.

Keywords: *growth prediction, supervised machine learning, swiss restaurant firms, web mining*

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. 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 research field of SME risk and growth, researchers recently turned their focus on applying data mining techniques for restaurant failure and bankruptcy prediction (Kim and Upnejy, 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 Upnejy, 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 decision tree 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.

While web mining has shown to be very useful for e-commerce, where any information related to consumer behavior are extremely valuable to anticipate and increase the sales performance (Patel et al., 2011), 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 this study, we propose a novel approach for growth modeling which has not been considered so far. We explore the use of web mining techniques for restaurant growth prediction. First, a systematic literature review on growth factors is conducted, which will serve as a base to collect growth-relevant information from the web. Next, web mining methods are applied to automatically collect and extract growth factors from various web data sources. Finally, 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.

The remainder of this paper is structured as follows. First, we provide an overview of the previous work in related research areas. This contains an overview of the hospitality research including restaurant growth modeling and growth factors, followed by a survey of data mining studies in the domain of hospitality research. Next, we provide an overview of the

applied methodology. Consequently, we present the results and discuss our findings. This paper concludes with a summary and an outlook on future research.

2 RELATED WORK

2.1. Definition of growth

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 et al., 2010).

2.2. Survey of factors influencing restaurant growth

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 search process as transparent as possible we followed the guideline for systematic reviews provided by Okoli and Schabram (2010). We first included 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). Next, 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". Our search resulted in 174 papers. In the next step, 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 ended up with 107 articles meeting our criteria.

To summarize this part of work, we identified 49 factors influencing the growth of restaurants, which can be roughly divided into firm-internal and external factors (see Appendix). Firm-internal factors can be further divided into two groups: (1) the

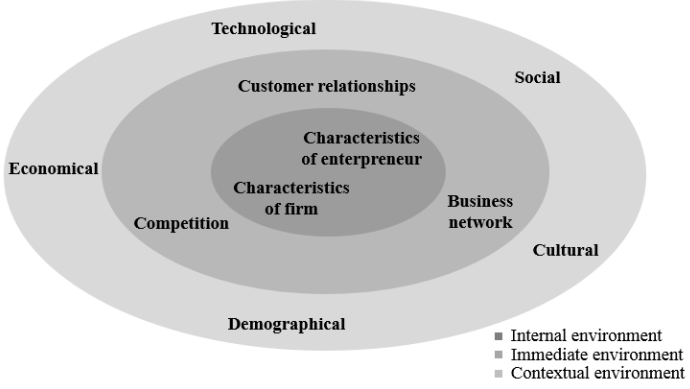

Figure 1: Factors influencing the growth of restaurants.

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 1 gives an overview of the factors influencing the growth of restaurants.

2.3. 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). Thus, we provide an 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 Upnejy, 2014). ANN models have been proposed as an attractive alternative because they are robust to some of these assumptions and do not require a priori specification of the functional relationship between the variables (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, there have been no previous studies that employed ANNs to predict the growth of restaurants.

2.4. Survey of Web Mining Based Restaurant Growth Prediction Studies

The web is a popular and interactive medium with intense amount of data freely available for users to access. With billions of web pages available in 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. However, automatically extracting targeted or potential valuable information from the web is a challenging task because of many factors such as size of the web, its unstructured and dynamic content as well as its multilingual nature. Therefore, WM research has emerged for knowledge discovery from the web.

Web Mining (WM) denotes the use of data mining techniques to automatically discover Web documents, extract information from Web resources and uncover general patterns on the Web (Etzioni, 1996). WM research overlaps with other areas such as artificial intelligence along with machine learning techniques, data mining, informational retrieval, text mining and Web retrieval. WM research is 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 already existing data (Sharda and Chawla). In general, WM tasks can be classified into three categories (Kosala and Blockeel, 2000): Web Content Mining, Web structure mining and Web usage mining. WM has been proved very useful in the business world, especially in e-commerce where any information related to consumer behavior are extremely valuable. A major challenge of e-commerce is to understand customers' needs as much as possible, in order to ensure competitiveness in the e-commerce. Thus, WM can be used to find data which have potential value from the website of e-commerce companies. For instance, Morinaga et al. (2002) presented a system for finding the reputation of products from the internet to support marketing and customer relationship management in order to increase the sales growth. The proposed system automatically collects people's opinions about certain target product of webpages and uses different techniques for text mining to get the reputation of those products. 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 WM and text mining can be applied to extract e-commerce growth factors from the websites.

In the technology- and information-driven world, the web has become a popular and interactive medium not only for e-commerce but for restaurants as well. While WM methods has been well researched and used in the field of e-commerce research to increase the sales growth, it has not been applied for restaurant growth research to the best of our knowledge. Furthermore, these studies only focus on the information available in company websites and thus, restrict the amount and spectrum of information typically given in company websites (Gök et al., 2015). Hence, further research investigating the full potential of web data for growth prediction is required.

3 METHODOLOGY

3.1. Data collection

Based on our literature review on the factors influencing the growth of restaurants, we collect information from two different types of data: (1) not publicly available data provided by a large Swiss insurer, and (2) publicly available web data. We first elaborate the data provided by the Swiss insurer, which contain basic information about restaurants and mainly serve as a ground-truth, i.e. labelled data for supervised machine learning. Next, we describe various web data sources collected to derive the growth-indicating factors, which serve as input features to train a growth prediction models.

3.1.1 Insurer data

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 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 2000 Swiss restaurants are collected from the insurer for the purpose of this study.

3.1.2 Web data

Web data related to the set of Swiss restaurants with known revenues (i.e. ground truth) are collected and factors influencing growth are extracted by means of web mining techniques (Mitchell, 2015). First, the usability of various web data sources is manually inspected with respect to the identified growth factors, as summarized in the appendix. In this study, six web data sources are examined.

Central Business Names Index (CBNI): CBNI provides free access to basic firm information and links through to internet excerpts from the individual canton commercial registry databases (für Justiz, 2016). 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.

TripAdvisor.com (TripAdvisor): 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). Our dataset includes records of 20429 Swiss restaurant, which covers most restaurant businesses of Switzerland. The dataset consists 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. Furthermore, since we collected the data of all Swiss restaurants, we geocoded the locations of all restaurants to conduct a competition analysis. Competitive restaurants in the surroundings within a radius between 50m and 300m of our ground truth data are collected. Thereby, restaurants with same cuisine, better overall ratings, lower price category and more reviews are considered as competition, as illustrated in Figure 2 (right).

Open Street Map (OSM): OSM is a free-to-access web-based mapping system for location-based services and general information (OSM, 2016). In this study, two types of datasets are downloaded from the OSM database of Switzerland: (1) the Point of Interest (POI) dataset and (2) the Roads dataset. 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, 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 6 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 of the ground truth data collected from the CBNI 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 illustrated in Figure 2 (left).

Swiss Federal Statistical Office (SFSO): SFSO is the national service provider and competence center for statistical observations in areas of national, social, economic and

Figure 2: 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 (left), and its competition, i.e. restaurants with same cuisine, better ratings, lower price category and more reviews (right, denoted in different colors).

environmental importance (Chen and Tsai, 2016). The FSO 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 (Swiss Federal Statistical Office 2016): population density, population change, foreign nationals, age pyramid (young, adult, and old 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). All data is aggregated on the level of municipalities - the lowest administrative unit on which Swiss census data is publicly available.

Swiss Federal Tax Administration (SFTA): 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.

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 al., 2002). Thus, in order to evaluate the location quality of our ground truth data, we inspect their proximity to chain branches. Therefore, we collected the geocoded location of all Swiss

Table1: Data sources and extracted growth factors. A detailed list of all growth factors is given in the appendix.

Data sources	**Factor type**	**Growth factor**
Insurer	Firm attributes	Type of restaurant
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

branches of the best-known fast-food chains, which include McDonald's, Subway, Starbucks and Burger King. The geocodes of the branches are downloaded from the Google maps on each chain's website (McDonald's Switzerland, 2017; Subway Switzerland, 2017; Starbucks Switzerland, 2017; Burger King Switzerland, 2017). In total, 1783 branches are recorded. In line with the collection of the above mentioned location-based information, the number of branches within a radius between 50m and 300m of our ground truth data are counted as a measure for the location quality.
The data from OSM, SFSO and SFTA are downloaded in CSV and PDF format from the respective website. However, the information in CBNI, TripAdvisor and fast-food chains are only visible. Thus, the information given in the fast-food chains' website are manually downloaded, whereas the information given in CBNI and TripAdvisor are automatically collected in the form of HTML files by applying web scraping techniques (Mitchell, 2015). Therefore, a self-developed scraper based on Python is deployed. Next, text mining methods are used to extract information of our interest and to store them in a structure format. Therefore, we use the python library BeautifulSoup (Richardson, 2007). In total, 27 out of the 49 identified growth factors are extracted from the above mentioned web data sources. The web data sources along with the extracted growth factors are summarized in Table 1.

3.2. Data sources 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). In this study, we adopt the data linkage method described by Denk (2009) to combine data provided by the Swiss insurer with data of various web data sources. As shown in Figure 3, 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. 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. Since the officially registered legal firm name in CBNI may differ from the actual restaurant name given in TripAdvisor, we used the following

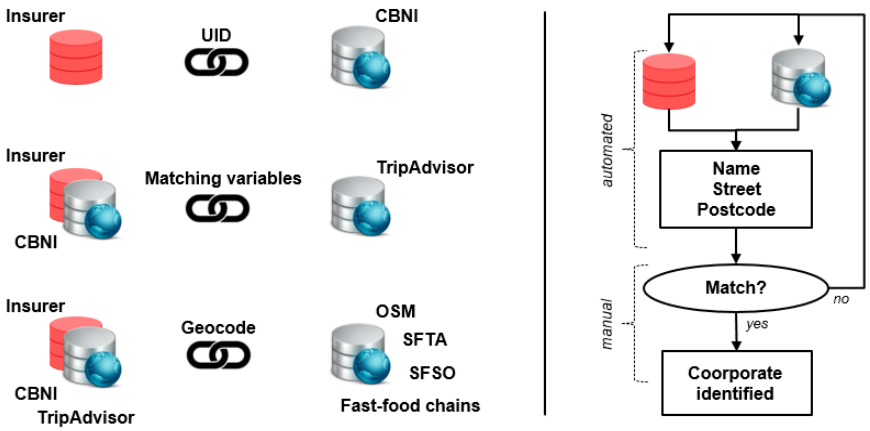

Figure 3: Linking corporate data with web data.

matching criteria for this matching step: name, postcode 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. Standardization refers to the conversion of values to a consistent format. Parsing deals with the decomposition of a string variable into a common set of components that are better comparable, as for instance splitting up a general address variable into postcode, city, street address and number. Then, location-based web data sources (OSM, SFSO, SFTA, fast-food chains) are matched with the geocoded address of our database. The processes are conducted automatically and the potential matches are returned for each ground truth sample. In the last step, to ensure a high data quality, the potential matches are inspected manually. Based on our knowledge and expertise, the final decision is taken. 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, 516 restaurants of the initial 2000 restaurants could be successfully identified and matched with web data sources.

3.3. Label creation & data preprocessing

3.3.1 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 to calculate the relative change of revenue over the corresponding timespan using linear regression (Montgomery et al., 2012).
Figure 4 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%).

3.3.2 Input feature creation

The input features for growth modelling are derived from the collected web data as described 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.
$The OSM and fast-food chains data are stored in PostgreSQL, a powerful, free and open-source database system typically used for geographical data (Stonebraker and Kemnitz, 1991). To derive the features reflecting the infrastructure and competition, the objects of interest within a radius between 50m and 300m are counted, e.g. counts of restaurants. Therefore, we make use of the function ST_DWithin() from the Python library pygresql,

which returns a Boolean value which is True, if the object of interest is within the defined distance (Cain, 2006).
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 (Batista and Monard, 2003). In our dataset, the range of missing data are between 0% and 52%, with TripAdvisor data containing the most missing data due to the incompleteness of information generated by TripAdvisor users. To address this issue, the following measures have been taken based on the business and structural characteristics of the features: 1) delete samples if only few 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 $r_{prs} \geq 0.95$) are removed (Hunt, 1986). In total, 85 input features are generated for the purpose of supervised machine learning, as summarized in Table 2. Note, that features denoted with a digit at the end are dummy variables derived from categorical features. In total, 85 input features are generated for the purpose of supervised machine learning, as summarized in Table 2. Note, that features denoted with a digit at the end are dummy variables derived from categorical features.

3.4. Construction of the growth models

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). Furthermore, logistic regression (LR) is chosen 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. From the ensemble of trees, the predicted class of an observation is calculated as the class with the majority vote (Breiman 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

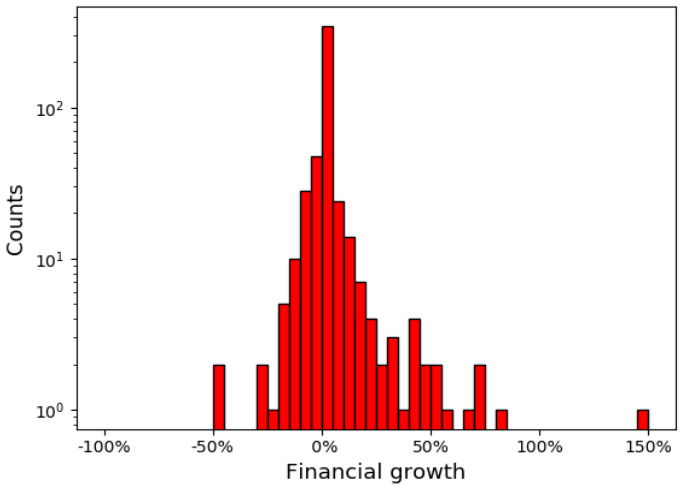

Figure 4: Distribution of the ground truth data (histogram bins = 50).

Table 2: Input features for supervised machine learning algorithms.

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
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		

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 (Haykin et al., 2009). 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 (Storey et al., 1990). 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).

In the first 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.

To optimize models' hyper-parameters, 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. Randomized grid 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 stratified 10-fold cross-validation procedure was applied for model selection. 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. Of the 10 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). 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 reported on the test set.

Furthermore, 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 (Kim, 2009). 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 study, the number of repeats is set to 10.

To compare and evaluate the classification performance of our classification algorithms, we make use of 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). In order to provide further insights on the classification performance, 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, are reported together with the AUC measure. 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.

4 RESULTS

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 3 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 5 depicts the mean feature importance plot of our RFCs (left) and LRs (right) only for the top 20 features due to the large amount of input features, which have been used to train the models. 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).

5 CONCLUSION

In this 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 literature review, as summarized in the Appendix. Next, a set of web data sources are examined with regards to the identified growth factors. Within the scope of this study, six

Table 3: Average performance of classifier families.

	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 %

web data sources containing information reflecting the business internal 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 scraping. 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 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 information related to customer relationship extracted from TripAdvisor are very useful to model the growth of restaurants. Moreover, external environmental factors such as the infrastructure, competition and demographics also play an crucial role, highlighting the importance of including a wide range of factors when modelling the growth of restaurants. 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.

This study has both theoretical and practical implications. It contributes to the existing literature of restaurant 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 measures of RFCs and LRs thus, extend the empirical body of knowledge. As a practical application, the proposed research method 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 restaurants.

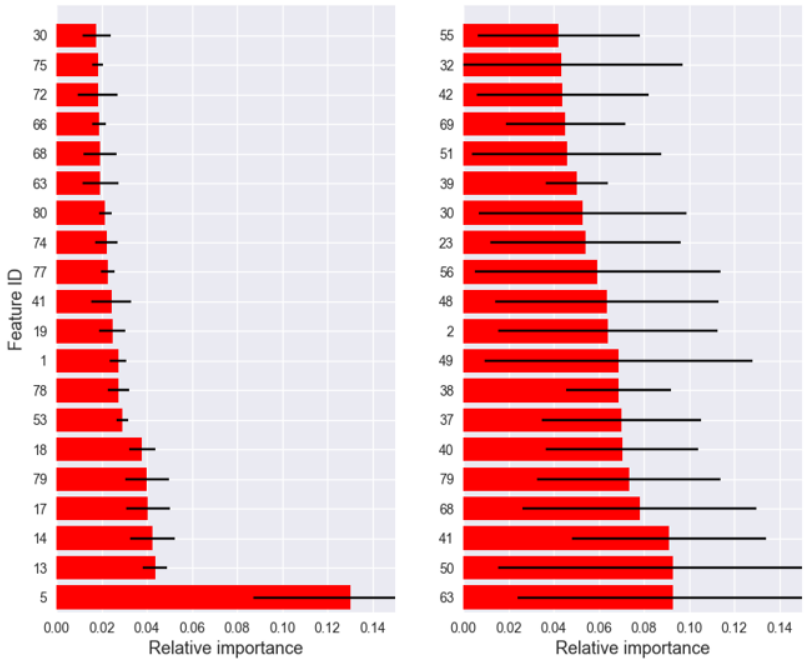

Figure 5: Feature importance plot including the top 20 features of RFCs (left) and LRs (right).

6 LIMITATIONS AND FUTURE WORK

This study is not without limitations and provides several opportunities for further research. First, our work is limited to Switzerland, 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, we plan to apply web mining techniques 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 model. Finally, we plan to test other machine learning methods such as stacking classifiers with to goal to optimize the performance restaurant growth prediction.

7 APPENDIX

Business environment		Factor type	Growth factor
Internal environment	Characteristics of firm	Firm attributes	Age of firm
			Size
			Location
			Reputation
			Service quality
			Physical environment
			Type of restaurant
			Kitchen & service operation
		Firm resources	Financial resources
			Human capital
		Firm strategies	Marketing / innovation
			Restaurant concept
			Service cycle optimization
			Business / menu planning
			HR management
		Food	Price
			Quality
			Type
			Variety of menu
	Characteristics of entrepreneur	Organization structure	Work specialization
			Centralization
			Legal form
		Socio-demographic	Age of entrepreneur
			Family background
			Education
			Experience
		Personality	Need for achievement
			Locus of control
			Attitude
		Competences	Managerial
			Entrepreneurial
External environment	Immediate environment	Customer relationships	Customer / market needs
			Customer acquisition
			Customer retention
			Customer satisfaction & feedback
		Network	Inter-organizational links
		Competition	Cluster of restaurants
			Food pricing
	Contextual environment	Technological	Infrastructure
		Socio-cultural	Tourism
			Social class
			Lifestyle
			Cultural diversity
		Economical	Taxation
		Demographical	Population size, growth & density
			Age & gender distribution
			Employment & income
			Education level
			Household size

8 BIBLIOGRAPHY

1. Aizawa A. and Oyama K. (2005). A Fast Linkage Detection Scheme for Multi-Source Information Integration, in: *Proc. WIRI'05,* 30-39.
2. Balcaen, S. and Ooghe, H. (2006). Thirty-five years of studies on business failure: and overview of the classic statistical methodologies and their related problems. *Br. Account. Rev.* 38 (1), 63–93.
3. Batista, G. E. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence* 17(5-6), 519-533.
4. Batista, G. E. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence* 17(5-6), 519-533.
5. Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13: 281-305.
6. Bilenko M., Mooney R., Cohen W., Ravikumar P. and Fienberg S. (2003) Adaptive Name Matching in Information Integration. *IEEE Intelligent Systems* 18(5), 16-23.
7. Bloom, J. Z. (2004). Tourist market segmentation with linear and non-linear techniques. *Tourism Management* 25(6), 723-733.
8. Borde, S. F. (1998). Risk diversity across restaurants: An empirical analysis. *Cornell Hotel and Restaurant Administration Quarterly* 39(2), 64-69.
9. Boritz, J. E., Kennedy, D. B. and Albuquerque, A. (1995). Predicting corporate failure using a neural network approach. *Intelligent Systems in Accounting, Finance and Management* 4, 95-111 .
10. 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.
11. Breiman, L., Chen, C. and Liaw, A. (2004). Using random forest to learn imbalanced data. *Journal of Machine Learning Research* p. 666.
12. Brown, K. A. and 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.
13. Cain, D. J. M. (2006). PyGreSQL–PostgreSQL module for Python. URL: http://www. pygresql. org.
14. Chen, L. F., and 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.
15. Cho, Min-Ho (1994). *Predicting business failure in the hospitality industry: An application of logit model.* Doctoral dissertation, Virginia Polytechnic Institute and State University.
16. Cho, S., Woods, R. H., Jang, S. S. and Erdem, M. (2006). Measuring the impact of human resource management practices on hospitality firms' performances. *International Journal of Hospitality Management 25*(2), 262-277.
17. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology* (88:11), pp. 2783-2792.
18. Denk, M. (2009). A framework for statistical entity identification to enhance data quality. *Insights on Data Integration Methodologies* 89.
19. Dimitras, A. I., Zanakis, S. H., and 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.

20. Doumpos, M. and Zopounidis, C. (1999). A multicriteria discrimination method for the prediction of financial distress: the case of Greece. *Multinatl. Finance J.* 3 (2), 71–101.
21. Dragos, C., Dragos, S. and Emitru, A. (2008). Financial scoring: a literature review and experimental study. *Econ. Bus. Rev.* 10 (1), 53–66.
22. Etheridge, H . L ., Sriram, R. S. and Hsu, H. Y. K. (2000). A comparison of selected artifi cial neural networks that help auditors evaluate client financial viability. *Decision Sciences* 31(2), 531-550 .
23. Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine?. *Communications of the ACM* 39:11, pp. 65-68.
24. Frydman, H., Altman, E., Kao, D. (1985). Introducing recursive partitioning for financial classification: the case of financial distress. *J. Financ.* 40 (1), 269–291.
25. für Justiz, B. (2001). *Zefix-Der zentrale Firmenindex auf Internet*. Reden, 2000, 1999.
26. Gastrosuisse (2017). *Branchenspiegel 2017.* Gastrosuisse, Verband für Hotellerie und Restauration.
27. Gepp, A., Kumar, K., Bhattacharya, S. (2010). Business failure prediction using decision trees. *J. Forecast.* 2(6), 536–555.
28. Gök, A., Waterworth, A. and Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics* (102:1), pp. 653-671.
29. Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician* 63(4), 308-319.
30. 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.
31. Gu, Z., and Gao, L. (2000). A multivariate model for predicting business failures of hospitality firms. *Tourism and Hospitality Research: The Survey Quarterly Review* 2(1), 37-49.
32. Haykin, S. S., Haykin, S. S., Haykin, S. S. and Haykin, S. S. (2009). *Neural networks and learning machines (Vol. 3).* Upper Saddle River, NJ, USA:: Pearson.
33. 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, 19-28.
34. Jain , B . A . and Nag , B. N. (1997) Performance evaluation of neural network decision models. *Journal of Management Information Systems* 14(2), 201-216 .
35. Kim, H., and Gu, Z. (2006). A logistic regression analysis for predicting bankruptcy in the hospitality industry. *The Journal of Hospitality Financial Management* 14(1), 17-34.
36. Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis* 53(11), 3735-3745.
37. Kim, S. Y. and Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling* 36, 354-362.
38. Kong, A., Nguyen, V., and Xu, C. (2015). *Predicting International Restaurant Success with Yelp.*
39. Kosala, R. and Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD Explorations Newsletter* (2:1).
40. Lacher, R., Pamela, S., Sharma, L. and Fant, A. (1995). A neural network for classifying the financial health of a firm. *Eur. J. Oper. Res.* 85 (1), 53–63.

41. Lee, T., Chiu, C., Chou, Y. and Lu, C. (2006). Mining the customer credit using classification and regression tree andmultivariate adaptive regression splines. *Comput. Stat. Data Anal.* 50 (4), 1113–1130.
42. Lev, B., Petrovits, C. and Radhakrishnan, S. (2010). Is doing good good for you? How corporate charitable contributions enhance revenue growth. *Strategic Management Journal* 31(2), 182-200.
43. Li, H., Sun, J. and Wu, J. (2010). Predicting business failure using classification and regression tree: an empirical comparison with popular classical methods and top classification mining methods. *Expert Syst. Appl.* 37(8), 5895–5904.
44. McDonald's Switzerland (2017). https://www.mcdonalds.ch/de/restaurants/.
45. Mitchell, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc.
46. Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis (Vol. 821)*. John Wiley & Sons.
47. Morinaga, S., K. Yamanishi, K. Tateishi and Fukushima, T. (2002). Mining product reputations on the web. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM*, pp. 341-349.
48. Morland, K., Wing, S., Roux, A. D., and 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.
49. Neophytou , E. and Molinero, C. M. (2004). Predicting corporate failure in the UK: A multidimentional scaling approach. *Journal of Business Finance and Accounting* 31(5/6), 677 – 710.
50. Okoli, C., and Schabram, K. (2010). *A guide to conducting a systematic literature review of information systems research.*
51. Olsen, M., Bellas, C., and Kish, L. V. (1983). Improving the prediction of restaurant failure through ratio analysis. *International Journal of Hospitality Management* 2, 187-193.
52. OSM Data for Switzerland 2016. http://planet.osm.ch/
53. Ozgulbas, N. and 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), pp. 2649-2660.
54. Park, K., and 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.
55. Patel, K. B., Chauhan, J. A. and Patel, J. D. 2011. Web Mining in E-Commerce: Pattern Discovery, Issues and Applications. *International Journal of P2P Network Trends and Technology* (1:3), pp. 40-45.
56. 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.
57. Rammer, C., Kinne, J. and Blind, K. (2016). *Microgeography of innovation in the city: Location patterns of innovative firms in Berlin.*
58. Richardson, L. (2007). *Beautiful soup documentation.*
59. Ron Kohavi (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *International Joint Conference on Artificial Intelligence* pp. 1137-1145.

60. Ronald J. Hunt (1986). Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research* 65(2), 128-130.
61. Scientific Journal Ranking. http://www.scimagojr.com/journalrank.php?category=1409
62. Sharda, D. and Chawla, S.. Web Content Mining Techniques-A Study. *International Journal of In-novative Research in Technology and Science (IJIRTS).*
63. Starbucks Switzerland (2017). http://www.starbucks.ch/store-locator/search.
64. Stonebraker, M., and Kemnitz, G. (1991). The POSTGRES next generation database management system. *Communications of the ACM* 34(10), 78-92.
65. Storey, D., Keasey, K., Watson, R. and Wynarczyk, P. (1990). *The Performance of Small Firms: Profits, Jobs, and Failures.* London: Routledge Small Business Series.
66. Subway Switzerland (2017). http://www.subway-sandwiches.ch/restaurants.php.
67. Swiss Federal Statistical Office 2016. http://www.bfs.admin.ch/bfs/portal/en/index/infothek/onlinedb/stattab.html.
68. Swiss Federal Statistical Office: STAT-TAB 2016. http://www.bfs.admin.ch/bfs/portal/en/index/infothek/onlinedb/stattab.html.
69. Swiss Federal Tax Administration 2016. https://www.estv.admin.ch/.
70. Thorleuchter, D. and 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), pp. 13026-13034.
71. TripAdvisor (2017). https://www.tripadvisor.ch/.
72. 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.
73. Zhang, G., Hu, M. Y., Patuwo, B. E. and Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research* (116:1), pp. 16-32.