Towards data driven decision support for financial institutions: Predicting small companies business volume in Switzerland

Daniel Müller, Funk Te, Flavien Meyer, Irena Pletikosa Cvijikj
Department of Management, Technology and Economics
ETH Zurich
Zurich, Switzerland
danielmueller@ethz.ch, fte@ethz.ch, flavien.meyer@student.unisg.ch, ipletikosa@ethz.ch

***Abstract*—**In Switzerland small and medium-sized enterprises represent more than 99% of all businesses. Therefore, prediction of their micro- and macroeconomic business development is of importance. In this paper, we propose a novel approach for predicting business volume using company characteristics and characteristics of the county the company operates in. We investigate which data sources can be combined to achieve this goal for small and midsized enterprises in Switzerland, building a model, irrespective of industry. We build our model based on the dataset obtained from an insurance company and combined the dataset with census data. We present two quantitative models, which allow to predict business volume in Swiss franks (CHF) and classify customers by size. Our results show that operational data from financial institutions (FI) customer relationship management (CRM) systems linked with census data are valuable to predict customer business volume.

Keywords: ***small and midsize enterprises (SME), business volume, random forests, customer classification, business risk management, open data***

I. INTRODUCTION

Small and medium-sized enterprises (SMEs) play an important role in the economy of many countries [1]. This is especially true for Switzerland, where SMEs represent more than 99% of all business and about 68% of all jobs in the country [2]. SMEs act as the countries backbone for growths but contribute unequally. Out of one hundred randomly drawn SMES companies, the fastest growing four firms will create 50% of the jobs in the group over a decade [2].

Financial institutions (FI) such as banks, brokers and insurance companies have robust relationships with many of the business owners and see themselves as important long-term partner in the value creation process and profit equally if the SMEs flourish [3]. Hence, it is in the FI's and in society's common interest [2] to understand what characterizes and drives the business of SMEs, given technical and economic feasibility.

For all FIs, operationalizing such a holistic customer perspective up to the individual level is a non-trivial task as it requires the collaboration of several departments, such as underwriting, IT, operations, strategy and marketing. However, the benefits for the FIs are substantial. Next to a better portfolio structure, this allows FIs to prioritize the customers based on business volume potential. A ranking of the most valuable customers, given the limited resources of customer relationship/sales personnel, can improve the efficiency of FI. Further, manual administrative tasks may cease to exist, given an automated computation of a holistic customer view [3]. This relieves the relationship manager [3], generally the ones in close contact with the SMEs and the ones responsible to document client characteristics (know your client [3], KYC). These managers may further benefit from automatically generated leads or prioritization tier changes, as variations in a customers' attributes can be made available to all hierarchies of the FI. However, these benefits mentioned are only possible given the non-manual collection and processing of a wide range of information to gain further insights about an FI's customer [4].

An Intelligent Decision Support System (IDSS) provides these benefits and enables better decision making [6]. An IDSS is a combination of traditional Decision Support Systems, which provide aggregated information for underspecified problems, and artificial intelligence, which learns from the collected data [6], [5].

To achieve such an IDSS, FIs need to automatically collect and process all available data sources, which relate to their customers. In previous decades, this was hardly possible due to lacking data sources and information processing limitations. However lately, even smaller companies in the financial technology sector (FinTech) have demonstrated the capability to assess the risk of individuals or companies in real time using various interchangeable sources of data [4]. The understanding of state-of-the-art methods to leverage modern information technology to build a more holistic picture of a business client, is essential for any FI [5], despite its size or maturity. An application of such methods resulting in fundamental information which can serve as the basis for a decision support system for FIs will be discussed in the next sections.

The rest of this paper is structured as follows. Section II will give an overview of the related literature we found to be guiding in explaining the context as well as the methodology applied in our experiments. In section III we will elaborate how we structured our data set, which we tested with two different models, as presented in section IV. In the V. section we discuss our results and conclude in section VI. The final section will summarize our findings and will give an outlook on our future work.

II. Related work

This research is based upon previous findings in four conceptual areas. It builds upon business growth models and integrates risk management research streams. It also refers to micro factors that help to explain both, risk and growth. Finally, it applies data science methodologies which were previously used in risk and growth research streams to predict a future customer state.

A. Business Growth Management

The literature on Business Growth Management dates back to 1967 and has proliferated since then into different streams addressing specific industries and business sizes. The most relevant growth models are Industry Growth [2], [3] [6], Large Business Growth [5], [7], Small Business Growth [8], [9] and General Growth [4], [12]. In combination, the authors' concepts for growth can be considered as the basis for emerging growth management approaches [10]. We will refer to small business growth [8] when examining our prediction results.

B. Business Risk Management

As organizations grow, they become more complex and risk needs to be managed [6]. [11] defines risk as "the potential of losing something of value, weighed against the potential to gain something of value". For FI, we consider the risk of losing business volume as risk. In an economic context, this restates how closely risk is connected with growth or negative growth of business volume. The desire to influence risk (and besides business volume growth) is illustrated by the many risk management research streams [12] that shaped in recent years. Within the scope of this research, risks are in the areas of strategy, market, human and operational issues, which may lead to a reduction in business volume. This choice is derived from the frequency of occurrence of generalizable risk types in the SME context [2].

C. Micro factors influencing business growth and risk

On a micro level, Porter (1960), Joyce et. al. (1990) showed that the choice of market segment had an explaining character for company growth. Further, company specific financials measures to predict bankrupt and healthy companies, are variables such profit, growth and employee efficiency ratios [3]. Further micro factors which impact the business of SMEs can be derived from the people's attitudes that influence the course of business. Entrepreneurial orientations, expressed through entrepreneurs' desire to be one's own boss and found strong correlation to company performance [13].

D. Combining data sources to predict business variation

Precursor authors have contributed to the business and risk prediction knowledge by combining qualitative inputs from financial data sources. Their works extend the possibilities to measure risk and provides deeper insight beyond well known financial ratios [3]. These authors have indicated that a wide range of industries that have been researching subjects in risk management (RM) and business growth literature. This same holds for the data science methodology literature listings about predicting SME risk/growths using non-financial data. This suggests that the majority of (reviewed) papers are industry/sector independent. Further, B2C e-commerce application aiming for cross selling and churn prediction are based on combining different data sources. These applications are matured concepts and can be considered industry standard [18]. However, model applications are not universal and can not be transferred into a B2B context, without model evaluation [18].

E. Literature gap

As [6] identified, current understanding of risk of SMEs is still in its infancy, wherefore the combination and application of the four research areas, may allow bringing forward new insights. The literature review [12] on managing risk in SMEs, came to the conclusion that "knowledge of the issues is inadequate at this early stage, and practical and academic studies are still very limited". The authors further emphasize that "many useful implications are expected in the future from this emerging stream" [12]. Widely discussed literature explaining how to predict business growth and risk with micro and other external data sources, which would be applicable for SMEs could not be identified yet. Neither has anybody to our best knowledge yet addressed the relevance for FI where fee revenue potential may correlate with business potential. Finally, the vast majority of research has been done with respect to US businesses which may not be representative of the Swiss business environment.

III. Methodology

Our chosen approach to predict business volume follows the Knowledge Discovery in Databases (KDD) process. It is commonly referred to as the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [13]. The process itself consist of combining related disciplines, notably Databases, Artificial Intelligence, Statistics, Scientific Discovery, and Visualization [14]. With the wealth of data sources available to the KDD process, our process was carefully guided by the domain experts, in our case insurance managers from the collaborating FI. A characterizing model of the process involving four primary stages, can be found below (from [14]).

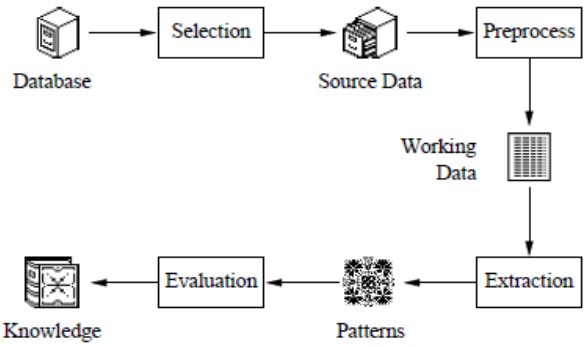

Figure 1. KDD process

Our process followed the ETL procedure, where we collected a set of features from available data sources [2], extracted, loaded and transformed them into a machine readable format, and then estimated the business volume of FI's customers from which some characteristics are known.

A. Sample description

The results of the KDD steps *Selection*, *Preprocess* and *Extraction* is a combined data source (N=7'064) consisting of FI customer related data combined with external information. These data sets are used for predictive modeling of SME's business volume with a FI. Company specific data has been provided by a Swiss insurer. It covers a 6-year period from 2010-2015 and consists of the company's name and address, the amount of contracts, and the length of the company's business relationship with the FI as well as the corporate form of the company. Further, we use data from the customer relationship management (CRM) system, which allows us to describe the company by size (number of employees) and type of business. The external data was used to enhance the internal client data by linking it to other information which are not gathered by FI yet. To leverage the core data source that the research partner has provided, we use public external data sources. Such a data source is provided by the *Bundesamt für Statistik (BFS)*, which allowed us to enrich the companies' characteristics. Using the companies' zip codes (domiciles), we linked the company with the county it is located in. Each company record was enriched with additional information obtained from census data. The mapping was done using the ZIP code of the company, thus linking the following variables to the data: Population density, the counties area, share of forest, arable and unproductive land within this county, sector distribution of the workforce as well as information about housing development - a driver for many business activities.

B. Data prepocressing and superviced learning

The sample we used consists of several continuous, ordinal and categorical variables. In order to prepare our data set for algorithmic evaluation, we binary coded the categorical variables. These were the variables describing the SME corporate form/type (AG, GmbH, Sarl, Co. KG, Cooperative, Association, School, Society) and the type of business they do (Noga Code).

The distribution of the business revenues we collected follows a power law distribution with many observations having low count values, i.e. strongly positively skewed (see Figure 2). Thus, applying an ordinary linear regression model is not suitable [18].

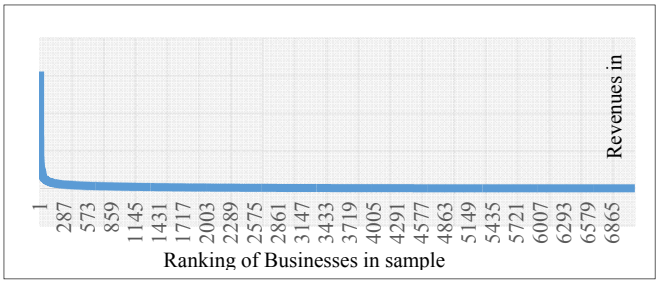

Figure 2. Revenue distribution ranked by volume in CHF

a) Prediction Experiment

To mitigate the positive skewed distribution, we log-transformed the dependent variable *y*, before processing the dataset any further. Our model assumes that the expected value $E(x) = \mu$ can be modeled as a linear combination of the independent variables

$$\boldsymbol{x} \in \mathbb{R}^p : \log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \qquad (1)$$

The combined data sources have been tested with a multiple linear regression, modeling the relationship between a scalar dependent variable y, in our case – SME's business revenue with a FI. All other variables were used as explanatory variables, reflecting the companies characteristics augmented with census data.

b) Classification Experiment

In a second experiment, we did a binary classification of small and large volume companies based on the proposed set of characteristics. To account for non-linear interactions between these features, we use the tree based ensemble model Random Forest (RF) [19]. RF provides a unique combination of prediction accuracy and model interpretability. The random sampling and ensemble strategies utilized in RF enable it to achieve accurate predictions while maintaining generalization. The advantage of this method, based on a bagging scheme, is the built-in and largely unbiased out-of-bag performance estimate [18]. In order to apply this method, we split our data sample in two groups and trained the model. The first group is made of customers with a business volume of more than CHF 10'000, while the second group consists of customers with equal or less than CHF 10'000.

IV. Results

Applying the multiple linear regression to our combined data sets results in the following model to forecast business revenue of SMEs with a FI. The results are summarized in table 1.

TABLE I. MODEL SUMMARY

R	*R-squared*	*Adj. R-squared*	*Standard Error*
.622	0.386	0.384	1.38

The model resulted in bivariate correlation coefficient of 62.2% and a Coefficient of determination of 38.6%, indicating model value in predicting business volume. From our ANOVA, we can derive the overall model significance with 29 degrees of freedom and a F score of 153, having a regression square of 8388.62, resulting in a total variance of 21'717. The significant effect our models yield, implies that the means differ more than would be expected by chance alone.

TABLE II. MODEL COEFFICIENTS RANKED BY SIGNIFICANCE

Model 1	Coefficients				
	Unstandardized Coefficients		*Standard Coefficients*		
	B	Std. Error [a]	Beta	T	Sig.
(Constant)	5.55	.275		2.16	.000
Lengths of relationship	.438	.012	.369	37.785	.000
AG	.643	.127	.133	5.063	.000
Societies	.988	.212	.054	4.652	.000
Number of employees	.237	.014	.169	17.072	.000
Business depth	.101	.003	.327	33.478	.000
Noga 6	-.800	.103	-.148	-7.734	.000
Company Age	.002	.001	.041	4.231	.000
Noga 3	-.286	.097	-.072	-2.946	.003
Welfare	.026	.010	.031	2.614	.009
Sector 3 workforce	.001	.000	.046	2.566	.010
Sector 2 workforce	.001	.000	-.050	-2.522	.012
Population Density	.000	.000	-.053	-2.347	.019
Empty Houses	.029	.014	.020	2.066	.039
Noga 7	-.260	.116	-.043	-2.247	.025
Forests	-.005	.003	-.038	-1.859	.063
Noga 4	-.189	.104	-.034	-1.823	.068
Foundations	.264	.159	.028	1.653	.098
Total area	-.001	.001	-.026	-1.654	.098
Noga 2	-.165	.101	-.034	-1.644	.100
GmbH & Sarl	.232	.146	.028	1.582	.114
Noga 5	-.156	.102	-.030	-1.526	.127
Noga 1	-.143	.106	-.023	-1.344	.179
Co KG	.361	.287	.013	1.259	.208
Arrable land	-.003	.003	-.033	-1.24	.215
Sector 1 workforce	.001	.000	.016	1.223	.221
Unproductive area	-.003	.003	-.021	-1.005	.315
Associations and Schools	-.120	.187	-.008	-.645	.519
Cooperative	-.023	.162	-.002	-.141	.888
Housing Dev	.001	.002	-.001	-.073	.941

a: dependent variable: log (revenue)

Our coefficient table indicates that several company characteristics are valuable contributors to predict business volume.

Our second approach classifying customers in *large* and *small* using RF yielded the following results.

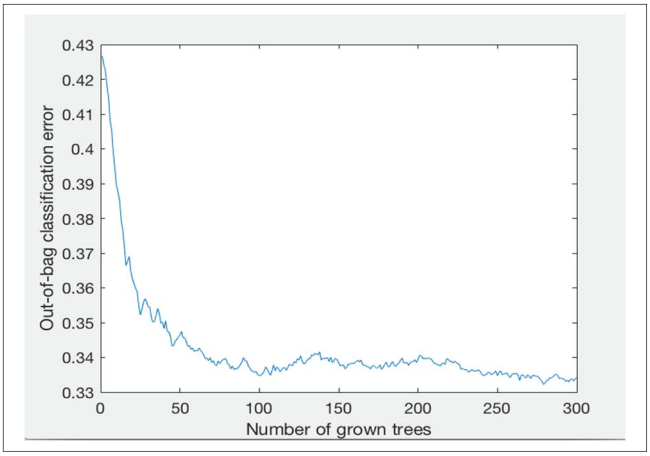

Figure 3. Random forest classification to predict group affiliation

Growing 250 trees allows the model to forecast the group affiliation (<> CHF 10'000 in business volume) with an accuracy of 66.4%. The models learning performance quickly decreased after the 50th tree and plateaued around 66%.

V. DISCUSSION

The chosen approaches demonstrated that SME business volume can be predicted by analyzing augmented company specific data, which many FI have collected throughout the years of operations. Suitable data sources for augmentation are e.g. publically available data sources such as the ones from *BFS* or the *Handelsregister (HReg)* provides the ability to calculate company age and describe the type of business. In our experiments, we narrowed our test sample to small companies in Switzerland and added data relating to the counties these SMEs are registered and in. We justify this due to the nature of SME business, which is locally oriented. Most SMEs conduct the majority of their business within a small radius of their domicile, hence we assumed the counties composition and historic growth influences local business revenue.

There was a significant effect for the form of the company such as AG, $t(29) = 5.06$, $p < .001$ or Societies, $t(29) = 4.65$, $p < .001$. Many of the social-demographic characteristics from the census data we found to be non-significant ($p > .100$) in predicting company business volume. Our model results were in line with our expectations, however we were surprised that the housing development/idle rate and welfare KPIs were not found to be significant, $t(29) = -.073$, $p = .941$.

VI. CONCLUSIONS

FIs which plan to build capacity in business analytics and support corporate decision making with machine learning approaches could benefit from applying the approach represented in this paper. Knowing which customers will generate more business volume (or have the potential) allows a better allocation of resources and may especially be important for FI, as only a few of their SME clients will eventually become large clients. Hence, early identification of

these high value clients is elementary to foster the FI's business.

VII. Summary and Future Work

In this paper we presented two models to estimate and classify business volume of FI's customer. The obtained results suggest that adding additional data to the model could improve the predictive power. These could be unstructured operational data which many FIs collect unintentionally, but may also be data collected independently via a survey.

We plan to further enhance our model by using company specific information such as the owner's personality and perception of risk to predict company revenues. An applicable framework is Entrepreneurial orientation (EO), reflecting on a CEO's quality to lead. To optimize the revenue prediction model, we plan to conduct a survey to analyze how the above mentioned measures provided by EO affect growth. If the measures indeed show a significant impact of the EO features towards a company's revenue, they will be implemented in the model. Alternative approaches capturing extraversion, emotional stability, agreeability, conscientiousness and openness of company managers, such as the Big Five Personality Theory [21], may also be considered, even though they may be harder to capture on scale, assuming they can not be derived from operational systems.

References

[1] A. Flynn and P. Davis, "The policy–practice divide and SME-friendly public procurement," *Environment and Planning,* vol. XX, pp. 1-20, 2015.

[2] D. J. Storey, "Understanding the Small Business Sector," in *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, Urbana, 1994.

[3] S. Auge-Dickhut, B. Koye and A. Liebetrau, Customer Value Generation in Banking - The Zurich Model of Customer-Centricity, Zurich: Springer International Publishing, 2016.

[4] C. W. Choo, Information management for the intelligent organization: The art of scanning the environment, Medford, NJ: Information Today, 2002.

[5] A. Soni and R. Duggal, "Reducing Risk in KYC (Know Your Customer) for large Indian banks using Big Data Analytics.," *International Journal of Computer Applications,* vol. 97, no. 9, pp. 49 - 60, 2014.

[6] M. Scott and R. Bruce, "Five Stages of Growth in Small Business," *Long Range Planning,* pp. 45-52, 1987.

[7] I. M. Fund, "The Special Data Dissemination System: Guide for subscribers and users.," Washington, D.C., IMF, 2013, pp. 1-113.

[8] F. Zhou, Y. Bingru, L. Li and Z. Chen, "Innovative Computing Information and Control," in *ICICIC '08, 3rd International Conference on. Date, 18-20 June 2008*, Piscataway, N.J., 2008.

[9] G. Lang, Decision support systems. Theory and application., Berlin: Springer, 2012.

[10] L. Ammore and V. Piotukh, Algorithmic Life: Calculative Devices in the Age of Big Data, Oxon: Routledge, 2016.

[11] R. V. Wright, "Strategy centers: A contemporary managing system.," Arthur D. Little Inc. , 1975.

[12] C. Verbano and K. Venturini, "Managing Risks in SMEs: A Literatrue Review and Research Agenda," *Journal of Technology Management and Innovation,* pp. 186-197, 2013.

[13] A. D. Little, "A system for managing diversity," Cambridge, Mass., 1974.

[14] M. E. Porter, Competitive strategy: Techniques for analyzing industries and competitors, New York: The Free Press, 1980.

[15] J. I. Channon, Business Strategy and Policy, New York: Harcourt Brace Jovanovich, 1968.

[16] M. Salter, Stages of Corporate Development: Implications for Management Control, Boston: Graduate School of Business Administration, George F. Baker Foundation, Harvard University, 1968.

[17] P. M. Kreiser, D. L. Marino, D. F. Kuratko and M. K. Weaver, "Disaggregating entrepreneurial orientation: the non-linear impact of innovativeness, proactiveness and risk-taking on SME performance," *Frontiers of Entrepreneurship Research,* vol. 40, p. 273–291, 2012.

[18] L. L. Steinmetz, "Critical stages of small business growth: When they occur and how to survive them," *Business horizons,* vol. 12, no. 1, pp. 29-36, 1969.

[19] L. B. Barnes and S. A. Hershon, "Transferring power in family business," *Harvard Business Review,* vol. 54, no. 4, pp. 105-114, 1976.

[20] W. Verbeke, K. Dejaeger, D. Martens, J. Hur and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach.," *Computational Intelligence and Information Management,* vol. 218, no. 1, p. 211–229, 2011.

[21] G. Shmueli, "To Explain or to Predict?," *Statistical Science,* vol. 25, no. 3, pp. 289-310, 2010.

[22] S. Rosenstein and J. G. Wyatt, "Outside directors, board independence, and shareholder wealth," *Journal of Financial Economics,* vol. 26, pp. 175-191, 1990.

[23] E. W. Frees, Regression Modeling with Actuarial and Financial Applications, Madison: Cambridge Press, 2010.

[24] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R news,* pp. 18-22, 12 2002.

[25] Y. Qi, "Random Forest for Bioinformatics," NEC Labs America, Cupertino, CA, 2008.

[26] M. A. Ciavarella, A. K. Buchholtz, C. M. Riordan, R. D. Gatewood and G. S. Stokes, "The Big Five and venture survival. Is there a linkage?," *Journal of Business Venturing,* vol. 4, no. 19, p. 465–483, 2004.