Large-scale machine learning for business sector prediction

Angenent, M.N.; Pereira Barata, A.P.; Takes, F.W.

Large-Scale Machine Learning for Business Sector Prediction

Mitch N. Angenent
Leiden Institute of Advanced
Computer Science, Leiden University
Leiden, the Netherlands
m.n.angenent@umail.leidenuniv.nl

António Pereira Barata
Leiden Institute of Advanced
Computer Science, Leiden University
Leiden, the Netherlands
a.p.pereira.barata@liacs.leidenuniv.nl

Frank W. Takes
Leiden Institute of Advanced
Computer Science, Leiden University
Leiden, the Netherlands
f.w.takes@liacs.leidenuniv.nl

ABSTRACT

In this study we use machine learning to perform explainable business sector prediction from financial statements. Financial statements are a valuable source of information on the financial state and performance of firms. Recently, large-scale data on financial statements has become available in the form of open data sets. Previous work on such data mainly focused on predicting fraud and bankruptcy. In this paper we devise a model for business sector prediction, which has several valuable applications, including automated error and fraud detection. In addition, such a predictive model may help in completing similar datasets with missing sector information. The proposed method employs a supervised learning approach based on random forests that addresses business sector prediction as a classification task. Using a dataset from the Netherlands Chamber of Commerce, containing over 1.5 million financial statements from Dutch companies, we created an adequately-performing model for business sector prediction. By assessing which features are instrumental in the final classification model, we found that a small number of attributes is crucial for predicting the majority of business sectors. Interestingly, in some cases the presence or absence of a feature was more important than the value itself. The resulting insights may also prove useful in accounting, where the relation between financial statements and characteristics of the company is a frequently studied topic.

CCS CONCEPTS

• **Applied computing** → **Economics**; • **Computing methodologies** → *Supervised learning by classification*;

KEYWORDS

business sector prediction, explainable machine learning, financial statements, data mining

ACM Reference Format:
Mitch N. Angenent, António Pereira Barata, and Frank W. Takes. 2020. Large-Scale Machine Learning for Business Sector Prediction. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3341105.3374084

1 INTRODUCTION

Financial statements form the backbone of accounting. They play a pivotal role in business by providing relevant financial information to company stakeholders. Companies generate annual reports containing these financial statements which, in turn, are comprised of attribute-value pairs. Although numerous in variables, only a small subset of attributes in financial statements are traditionally used by analysts for comparison of companies [2]. Amongst others, the business sector to which a company pertains is one such relevant feature. In fact, for business professionals, it is a paramount attribute for analysis. However, many a company fail to have their corresponding sector described. Predicting its value when absent would prove, thus, invaluable. Ultimately, the dependency of sector information availability, as well as the low cardinality and prevalence of the set of commonly used variables both act as a constraint upon conventional analysts and their practices.

From here on, we refer to *prediction* as establishing a predictive model to gain insights in how to perform such task. We are particularly interested in sector prediction by applying machine learning techniques on financial statements. Our focus on a predictive model is motivated three-fold: firstly, by predicting the sector of companies without a sector label it is possible to perform analysis on a larger proportion of a sector or market; following, a predictive model can aid government institutions in checking filed statements on their correctness by automatically detecting potential errors or fraud, as some sectors are subject to stricter regulations than others. Lastly, by merging the concepts of prediction and explainable machine learning, it is possible to further aid domain experts by providing them with new insights and tools (e.g., attributing relevance to a previously neglected set of features).

This study contributes towards current literature by not only assessing the suitability of machine learning algorithms with respect to sector prediction, but also through the use of a dataset of unprecedented scale (over 1.5 million instances) which originates from the Netherlands Chamber of Commerce [14]. In summary, our research question is: *Can machine learning be used to predict the business sector of a company based on their financial statement?* Complementarily, we are also interested which attributes of financial statements are most relevant for business sector prediction. We address our questions through an explainable *data driven* approach, by modelling business sectors as targets within a classification problem framework.

The structure of this paper is as follows. Section 2 discusses previous work and our contribution. Section 3 describes the characteristics of our data, and in Section 4 our methods are outlined. Section 5 presents the setup and results of our experiments. Section 6 concludes and offers recommendations for future work.

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03.
https://doi.org/10.1145/3341105.3374084

2 RELATED WORK

The application of data mining techniques to financial statements generally focuses on fraud prediction as supervised learning problem [5, 9–12, 19]. Other supervised learning applications have been reported, such as revenue and bankruptcy prediction [8, 16, 20, 22]. A broad range of algorithms and methods have been applied and thoroughly compared in the aforementioned studies. One recurring classifier is the decision tree [6, 11, 12, 16, 20, 22], mostly as part of an ensemble [6, 12, 20, 22]. Although neural network classifiers have been reported to be of high performance, few insights can be attained (i.e., *black box* model) [11, 16, 19]. In contrast, decision trees are easily interpreted, with little to no performance loss [11, 16]. To note, the largest datasets used in these studies barely surpass 10,000 instances and 65 features. Ultimately, no literature currently exists regarding business sector prediction.

The contributions of this research consists foremost of establishing the adequacy of financial statements towards predicting business sectors. Additionally, we take advantage of the concept of explainable machine learning within our framework to produce new information about which financial statement attributes are relevant for the task of classification, improving on currently used methods by analysts. By achieving these goals, we further provide meaningful insights which may result in new applications within the economics domain, including novel frameworks for automated error and fraud detection. The emphasis of this work will be on assessing the ability to perform sector prediction by applying machine learning on financial statements.

3 DATA

The open dataset from the Netherlands Chamber of Commerce we use contains 1,517,400 anonymous financial statements, distributed over the years 2015-2018 [15]. We use only the financial statements with a sector code (SBI). A total of 923 unique SBI codes occur in the dataset, with non-uniform distribution. This imbalance motivates us to use a different hierarchical level of the SBI coding system. Instead of using the activity level, a higher level is used (Figure 1). By aggregating different sub-classes, 23 classes are established. A total of 593,090 instances and 151 attributes are extracted, with an average of 12 non-missing attribute values per instance. To address missing values, the missing-indicator approach is used. In this manner, a value of 0 is placed where values are missing ('NaN') and an additional column is added representing the missingness of each attribute (Figure 2). Given classifier robustness, the results obtained in performance should not alter significantly by using other imputation methods [18].

x_i	SBI	x_i
…	18129	…
…	49393	…
…	64191	…
…	66192	…
…	16239	…

$\Longrightarrow$

x_i	SBI	x_i
…	18	…
…	49	…
…	64	…
…	66	…
…	16	…

$\Longrightarrow$

x_i	SBI	x_i
…	C	…
…	H	…
…	K1	…
…	K4	…
…	C	…

Figure 1: Class reduction.

x_1	x_2
123	NaN
NaN	148
NaN	NaN
984	457

$\Longrightarrow$

x_1	x_2	M_1	M_2
123	0	1	0
0	148	0	1
0	0	0	0
984	457	1	1

Figure 2: Encoding missingness.

4 METHODS

The methods needed for the supervised classification problem that we are dealing with are described in this section. Different approaches to tackle class imbalance are discussed in Section 4.1. Section 4.2 consists of a consideration of the classification algorithm to apply. Finally, Section 4.3 reports our evaluation process.

4.1 Handling class imbalance

An excerpt of classes and their relative frequency reflects the problem of class imbalance (Table 1). Our approaches to tackle the class imbalance problem comes in two variations: 'cost sensitive learning' and 'sampling approaches'. The first adds weights to instances, with a higher weight for the instances of the minority class so that they contribute more into the total error. The sampling approach removes or adds samples to the train sets to obtain a more equal distribution of the classes. This study applies and compares four class imbalance approaches:

1. **Random undersampling (RUS)**: a sampling approach that randomly removes instances of the majority class.
2. **Random oversampling (ROS)**: a sampling approach that randomly adds extra instances of the minority class.
3. **Synthetic Minority Over-sampling (SMOTE)**: new instances are synthesized by computing feature values as slight variations of instances that are similar between each other. This leads to better generalization by decision trees [4].
4. **Weighted classes (CSL)**: the weight of a class is inversely proportional to its frequency (cost-sensitive learning).

Table 1: Excerpt of classes and their relative frequency.

	Description	Freq.
A	Agriculture, forestry and fishing	0.01474
B	Mining and quarrying	0.00026
C	Manufacturing	0.03661
D	Electricity, gas, steam, air conditioning supply	0.00204
E	Water supply; sewerage, waste management	0.00190
F	Construction	0.04202
G	Wholesale and retail trade	0.11077
H	Transportation and storage	0.01987
I	Accommodation and food service activities	0.01592
J	Information and communication	0.03192
K1	Financial institutions - Other	0.03051
K2	Financial institutions - Financial holdings	0.36829
K3	Financial institutions - Investment funds	0.05301
K4	Financial institutions - Insurance and pension	0.00081

4.2 Classification algorithm

In this work, we make use of a random forest classifier: a bagging ensemble method in which weak classifiers (trees) are jointly created from random samples of the entire dataset [7]. Primarily, we chose this algorithm for its explainability, which translates into assessing feature importance, allowing us to extract insights from the obtained model. Besides, random forest competitive performance has been proven in previous work [11, 16]. Additionally, there are several other advantages to using this classifier. First, it requires little preprocessing of data (e.g., scaling, feature selection). Additionally, it requires little tuning of hyperparameters to produce adequate and usable results. Third, it offers appropriate scalability in both sample size and dimensionality. Lastly, it is mostly insensitive to outliers, and overall noisy data [3].

4.3 Evaluation

The goal of this study is to provide insights into business sector prediction. As these insights are only representative when obtained from a reliable model, we need an evaluation metric to assess our models performance. The performance metric of choice for classification tasks is the consensually used area under the curve (AUC) of the receiving operator characteristic curve (ROC). A *OneVsRest* classification problem approach is followed as to be able to produce such a metric, and obtain insights in the performance per class. For each class imbalance-handling approach, AUCs are yielded through stratified 10-fold cross validation with respect to each class. Combining the AUCs of all classes and computing their mean produces the final AUC of each specific approach. All classes weigh equally during all computations, independently of their frequency.

5 EXPERIMENTS

In this section, the experiments are outlined. Section 5.1 describes the experimental setup. The results are shown in Section 5.2 and discussed in Section 5.3.

5.1 Experimental Setup

The complete implementation of this experiment is performed in *Python*. Machine learning algorithms and measures were supplied by Scikit-learn [17]. The three sampling class imbalance approaches were implemented using the imbalance-learn module [13]. Classifiers were initialized with default parameters. This resulted in 100 trees used for one random forest model. For reproducibility, the random seed value was set to 42. These conditions apply to all objects initialized during the experiments.

The train sets and test sets are stratified, so that the class frequency in each of the train/test sets are a reflection of the complete dataset. Ten train sets and ten test sets are determined once and used over the complete course of the experiment. The AUC per class is computed by taking the mean of the AUC of each fold. Then, the mean of all classes is used to compute the performance of the complete method by means of AUC. Wilcoxon signed-rank tests [21] are applied to determine whether the class imbalance approaches are statistical significant compared to the regular approach. A p-value below 0.05 is considered to indicate a significant change in the distribution of performance.

The class imbalance approaches as described in Section 4.1 are implemented as follows. For RUS, the default sampling strategy implies that all classes except the minority class are undersampled during preprocessing each fold. For ROS and SMOTE, the default sampling strategy implies that all classes except the majority class are oversampled during the preprocessing for each fold. CSL is implemented by setting parameter 'class_weight' of RandomForestClassifier to 'balanced'. Ultimately, values of feature importance are retrieved from the best-performing approach. Feature importance is computed as the Gini variable importance measure [1]; higher values equate to higher relevance, with a cumulative sum of 1.

5.2 Results

The overall mean and standard deviation values of performance per class imbalance-handling approach can be regarded in Table 2. Wilcoxon signed-rank test p-values between *Regular* and every other approach are also denoted. Additionally, a baseline *Random* performance value (AUC value of 0.5) was added representing the score of an uninformed model; i.e., random guessing. Mean and standard deviation values of performance for each individual sector classifier yielded within the regular approach to handling class imbalance are listed in Table 3. Table 4 lists the frequency and mean feature importance for the 10 most important features across all classes where the prefix *M* denotes the missing-indicator. For example, feature x_{17} (the 17th most frequently valued attribute in the dataset) occurs as one of the 10 most important features for all 23 classes, with an average importance of 0.09.

Table 2: Mean and standard deviation values of performance (AUC) per approach.

Approach	Performance	p-value
Random	0.5	-
Regular	0.78 ± 0.07	-
RUS	0.59 ± 0.08	0.000
ROS	0.78 ± 0.06	0.768
SMOTE	0.79 ± 0.05	0.848
CSL	0.78 ± 0.07	0.357

Table 3: Mean and standard deviation values of performance (AUC) per class (*Regular*).

Class	Performance	Class	Performance
A	0.82 ± 0.01	K2	0.84 ± 0.00
B	0.68 ± 0.03	K3	0.88 ± 0.00
C	0.84 ± 0.00	K4	0.74 ± 0.02
D	0.76 ± 0.02	L	0.81 ± 0.00
E	0.72 ± 0.02	M	0.72 ± 0.00
F	0.82 ± 0.00	O	0.62 ± 0.15
G	0.87 ± 0.00	P	0.73 ± 0.01
H	0.82 ± 0.00	Q	0.79 ± 0.01
I	0.87 ± 0.01	R	0.77 ± 0.01
J	0.79 ± 0.01	S	0.76 ± 0.01
K1	0.72 ± 0.01	U	0.87 ± 0.20

Table 4: Top occurring features and importance (*Regular*).

x_i	Attribute	#	FI
x_{17}	AssetsNoncurrentOther	23	0.090
x_{19}	Inventories	23	0.089
x_{128}	InterestReceived…InvestingActivities	23	0.085
x_{55}	CalledUpShareCapital	23	0.083
x_{45}	CashFlowFromOperations	23	0.066
M_x_{69}	InvestmentProperties	23	0.051
x_{125}	PaymentsReclaimingValueAddedTax	22	0.049
x_{99}	ProceedsSalesIntangibleAssets	22	0.048
x_{90}	ChangesValueFinancialAssetsSecurities	18	0.045
x_{47}	CashAndCashEquivalentsCashFlow	6	0.034
M_x_{20}	SharePremium	6	0.033
M_x_{59}	InterestReceived…OperatingActivities	4	0.033
M_x_{70}	Incr.Decr.PayablesCreditInstitutions	4	0.023
M_x_{35}	SumOfExpenses	3	0.030
x_{141}	ResultBeforeTaxOrdinaryActivities	2	0.033
x_{43}	CashFlowOperatingActivities	1	0.025
x_{111}	LineItems…NotOperatingActivities	1	0.025
x_{106}	RevaluationReserveRelease	1	0.013
M_x_{61}	IncreaseDecreaseProvisions	1	0.008
x_{85}	CashFlowsOperatingActivitiesOther	1	0.002

5.3 Discussion

Since no approach provided significant performance improvement compared to the regular approach, this approach will be analysed. Results vary per class with a minimum AUC of 0.62 (class O) and maximum of 0.88 (class K3), with a mean performance of 0.78 ± 0.07. Despite observing a relationship between class frequency and model performance, we do not conclude a linear dependency; e.g., classes U and K2 have similar performance with distinct frequencies. Hence, the characteristics of some classes are well encoded by financial statements alone. From Table 4, 6 features were among the 10 most important features for all classes, while 6 missing-indicators were listed for at least one class. Thus, it is deducible that business sectors are characterisable by a small subset of attributes, while the presence or absence of attributes holds considerable information for classification as well.

6 CONCLUSIONS AND FUTURE WORK

In this study we have performed explainable business sector prediction by applying machine learning techniques to financial statements. We provide insights into which features relate the most to business sectors. We conclude that a small subset of all features from both balance sheets and income statements are the top features for the majority of the classes. Additionally, the presence or absence of an attribute on a financial statement can be as important as the value itself. This enables future applications such as the detection of mislabeled statements and potential fraud, while overall augmenting data accuracy and aiding domain experts in their work. In summary, we conclude that machine learning can be used for accurate business sector prediction. Future work can entail using other classification algorithms, optimized hyperparameters, and different data sources such as written annual reports (text mining).

REFERENCES

[1] Kellie J. Archer and Ryan V. Kimes. 2008. Empirical Characterization of Random Forest Variable Importance Measures. *Computational Statistics & Data Analysis* 52, 4 (2008), 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015

[2] Paul Barnes. 1987. The Analysis and Use of Financial Ratios: A Review Article. *Journal of Business Finance & Accounting* 14, 4 (1987), 449–461. https://doi.org/10.1111/j.1468-5957.1987.tb00106.x

[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32. Issue 1. https://doi.org/10.1023/A:1010933404324

[4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. http://dx.doi.org/10.1613/jair.953

[5] Hawariah Dalnial, Amrizah Kamaluddin, Zuraidah Mohd Sanusi, and Khairun Syafiza Khairuddin. 2014. Accountability in Financial Reporting: Detecting Fraudulent Firms. *Procedia - Social and Behavioral Science* 145 (2014), 61–69. https://doi.org/10.1016/j.sbspro.2014.06.011

[6] Giuseppe Dattilo, Sergio Greco, Elio Masciari, and Luigi Pontieri. 2000. A Hybrid Technique for Data Mining on Balance-Sheet Data. In *Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000)*. Springer-Verlag, London, UK, 419–424. http://dl.acm.org/citation.cfm?id=646109.679295

[7] Tim Kam Ho. 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (1998), 832–844. https://doi.org/10.1109/34.709601

[8] Ken Ishibashi, Takuya Iswaki, Shota Otomasa, and Katsutoshi Yada. 2016. Model Selection for Financial Statement Analysis: Variable Selection With Data Mining Technique. *Procedia Computer Science* 96 (2016), 1681–1690. https://doi.org/10.1016/j.procs.2016.08.216

[9] Rasa Kanapickiene and Zivele Gundiene. 2015. The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. *Procedia - Social and Behavioral Science* 213 (2015), 321–327. https://doi.org/10.1016/j.sbspro.2015.11.545

[10] Yeonkook J. Kim, Bok Baik, and Sungzoon Cho. 2016. Detecting Financial Misstatements With Fraud Intention Using Multi-class Cost-sensitive Learning. *Expert Systems with Applications* 62 (2016), 32–43. https://doi.org/10.1016/j.eswa.2016.06.016

[11] Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. 2007. Data Mining Techniques For the Detection of Fraudulent Financial Statements. *Expert Systems with Applications* 32 (2007), 995–1003. Issue 4. https://doi.org/10.1016/j.eswa.2006.02.016

[12] Sotiris Kotsiantis, E. Koumanakos, D. Tzelepis, and V. Tampakas. 2005. Forecasting Fraudulent Financial Statements using Data Mining. *International Journal of Computational Intelligence* 3 (2005), 104–110.

[13] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365.html

[14] Netherlands Chamber of Commerce. 2019. Home. Retrieved May 5, 2019 from https://www.kvk.nl/english/

[15] Netherlands Chamber of Commerce. 2019. Jaarrekeningen Open Data Set. Retrieved January 21, 2019 from https://www.kvk.nl/producten-bestellen/koppeling-handelsregister/kvk-jaarrekeningen-open-data-set/

[16] David L. Olson, Dursun Delen, and Yanyan Meng. 2012. Comparative Analysis of Data Mining Methods for Bankruptcy Prediction. *Decision Support Systems* 52 (2012), 464–473. Issue 2. https://doi.org/10.1016/j.dss.2011.10.007

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[18] António Pereira Barata, Frank W. Takes, H. Jaap van den Herik, and Cor J. Veenman. 2019. Imputation Methods Outperform Missing-Indicator for Data Missing Completely at Random. In *Proceedings of the 19th IEEE International Conference on Data Mining Workshops*. Institute of Electrical and Electronics Engineers, Beijing, China, 407–414.

[19] P. Ravisankar, V Ravi, G. Raghava Rao, and I. Bose. 2011. Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques. *Decision Support Systems* 50 (2011), 491–500. Issue 2. https://doi.org/10.1016/j.dss.2010.11.006

[20] Tae Kyung Sung, Namsik Chang, and Gunhee Lee. 1999. Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction. *Journal of Management Information Systems* 16 (1999), 63–86. https://doi.org/10.1080/07421222.1999.11518234

[21] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics* 1, 6 (1945), 80–83. https://doi.org/10.2307/3001968

[22] Maciej Zieba, Sebastian K. Tomczak, and Jakub M. Tomzak. 2016. Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. *Expert Systems with Applications* 58 (2016), 93–101. https://doi.org/10.1016/j.eswa.2016.04.001.