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

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

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.

As increasingly larger volumes of open-sourced, structured, and standardized financial data are made available — mostly as a consequence of the introduction of the eXtensible Business Reporting Language (XBRL) [27] — so too is augmented the applicability of data mining techniques to that data. This promotes the potential to retrieve new relevant relations between attributes of financial statements. In other words, this enables analysis on a higher-level, including applications such as sector prediction and anomaly detection. We are particularly interested in sector prediction by applying machine learning techniques on financial statements. From here on, we refer to prediction as establishing a predictive model to gain insights in how to perform such task.

Our focus in 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).

Albeit previous work has been able to merge the fields of business and machine learning, its focus is generally the applicability of different classifiers within the task of fraud or bankruptcy detection within small datasets (100 to 1,000 instances) [24]. To the best of our knowledge, no research has been done in business sector prediction nor has any work reported the use of large-volume financial datasets. 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 [19].

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 provides background information about financial statements as well as sector categorization. Section 3 discusses previous work related to ours and how we further contribute towards current literature. Section 4 describes the characteristics of our data, and in Section 5 our methods are outlined. Section 6 presents the setup and results of our experiments. Section 7 concludes and offers recommendations for future work.

2 BACKGROUND

This section provides background information about the economic topics that are embedded in this research. The first section introduces the concept of financial statements, whereas the second section provides details of the sector categorization in the Netherlands.

2.1 Financial Statements

A financial statement must contain at least three elements: an income statement, a statement of cash flows, and a balance sheet. An income statement holds information about revenues and expenses of the company over a specific period, usually a calendar year. All single expenses from that period are summed up on ledger cards; these ledgers are based on accounting standards and are usually tailored for a specific company. The statement of cash flows shows the incoming and outgoing amounts of cash over the same specific period. Where the income statement represent the profitability of a company, the statement of case flows represents its liquidity. Lasty, the balance sheet indicates the state of a organization at specific time, usually the end of a financial year. This state is represented with the balance of standardized ledgers, such as: 'Accounts Receivables', 'Accounts Payable', 'Inventory' and 'Property', to name a few. These detailed ledgers are aggregated into the overarching categories 'Assets', 'Liabilities' and 'Equity'.

To enforce the standardization of financial statements, XBRL was created. XBRL is based on the XML language and allows reporting terms to be authoritatively defined. Central authorities, such as tax authorities and chambers of commerce, leverage the standardization for easier comparison [27]. An example is the Netherlands Chamber of Commerce, which obliges certain companies to deposit their financial statement in the XBRL standard. This led to faster depositions, smaller files, easier financial comparisons, and higher quality of financial statements for medium-sized Dutch companies in 2017 [7]. Not all Dutch companies have already made the transition to the XBRL standard. As of financial year 2016, micro and small are mandatory to deposit their annual reports in the new format. From financial year 2017, medium sized companies followed. Large companies (> €40 million in revenue, >250 employees) follow the new standard from financial year 2019.

Table 1: Number of distinct categories for the most commonly used levels of SBI coding. SBI 2008, version 2018 [14].

SBI coding	Type	Distinct categories
One character	Sections	21
Two digits	Departments	95
Three to five digits	Activities	1348

2.2 Sector categorization

Since this research focuses on sector prediction of Dutch companies, it is relevant to understand the Dutch sector coding. The Dutch organization for statistics, Statistics Netherlands, defined the standard industrial classification (SBI) for the Netherlands. The SBI is a hierarchical mapping based on economic activities to classify a business in terms of their primary business activity [17]. The longer the SBI code (represented by the number of characters), the more information it details about a company's activity. Moreover, as the hierarchical standing decreases, the number of distinct codes per hierarchical level increases as illustrated in Table 1. The SBI coding system distinguishes five hierarchical levels which reference other classification systems according to code length [16]:

(1) The section, represented by one character.
(2) The department, represented by two digits. These digits match the notation of the both the international (ISIC) and European (NACE) categorizations.
(3) The activity, represented by three digits.
(4) The activity, represented by four digits. This matches the NACE notation.
(5) The activity, represented by five digits. It is based on the four digit NACE notation, with small adjustments for the Netherlands.

3 RELATED WORK

An abundance of research has been done in regard to the application of data mining techniques to financial statements, generally focused on fraud prediction. For example, [5, 10–13, 23] successfully modeled the problem whether or not a filed financial statement is a fraudulent financial statement as a supervised learning task. An overview of the field is provided by Sharma et al. [24]. They put effort in the categorization of 35 papers that researched fraudulent financial statement and other types of accounting fraud from the period between 1995 and 2012. The conclusion of the study is that neural networks have a good performance on classification problems, although they lack interpretability. Secondly, the researchers contributed with a framework for financial accounting fraud detection, emphasizing the use of sources other than financial statements alone. Another regularly recurring problem is predicting future revenue and ultimately bankruptcy of a company [9, 20, 25, 28]. Whether a company would go bankrupt in the following five years was modeled as a classification problem [28]. Accordingly, this problem was addressed by using an ensemble

of boosted trees (Extreme Gradient Boosting) and the addition of new features, by performing arithmetic operations on all possible combinations of features.

A broad range of algorithms and methods have been applied and thoroughly compared by the previously mentioned studies. One of the recurring classifiers is decision trees [6, 12, 13, 20, 25, 28], mostly as part of an ensemble [6, 13, 25, 28]. In the three studies that compared neural networks with other classifiers [12, 20, 23], the neural networks were the best in two studies [20, 23]. The main disadvantage of neural networks is that they act as a black box and therefore lack explainability. In contrast, decision trees are easily interpreted, without compromising on performance [12, 20].

The aforementioned researchers mainly used small datasets, shown here in four categories. Kirkos et al. [12] used the smallest dataset (< 100). Between 100 and 1,000 instances were used in five studies [5, 10, 13, 23, 25] and 1,000 - 10,000 instances were used by four [6, 9, 11, 20]. One research stood out: Sharma et al. [24] used more than 10,000 records. The number of features used differed from 18 [20] to 65 [9].

Based on the aforementioned studies, we can conclude that machine learning techniques have proven to be successful in several classification problems, such as fraud and bankruptcy. In addition, the performance of available algorithms has been widely investigated, mainly on small datasets. The contributions of this research consists foremost of establishing the adequacy of making use of financial statements towards predicting business sectors. Additionally, we utilize 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 therefore not be on the algorithmic part of machine learning, but rather focus on assessing the ability to perform sector prediction by applying machine learning on financial statements.

4 DATA

The open dataset from the Netherlands Chamber of Commerce we use [19] contains 1,517,400 anonymous financial statements, distributed over the years 2015-2018. We use only the financial statements with a SBI code (39%), leaving 593,090 instances. This segment of the dataset contains 154 different attributes. Two attributes with a descriptive function are dropped, and the target is extracted. This leaves 151 attributes for testing the model (see Section 5), with on average 12 attributes per financial statement. This means there are missing values. Attributes originating from balance sheets occur more frequent than attributes from income statements. This can be explained by the fact that less companies are obliged to deposit their income statement as part of their annual report [18]. Table 2 presents the 30 most frequent features where the column 'Frequency' denotes relative frequency.

There are a total of 923 unique SBI codes in the dataset. The distribution of the codes is not uniform. This imbalance motivates us to use a different hierarchical level of the SBI coding system. Instead of using the activity level (SBI code from three to five characters), a higher level such as the department or section will be used. Table 1 shows that this leads to a significant reduction of the number of distinct codes, which minimizes the level of detail within activities. This process is part of our efforts to address class imbalance. In other words, by aggregating different sub-classes into the same category, the overall number of instances per class should increase. Concretely, samples that would not be utilizable given their low class frequency can now be used within this study.

There are three essential points to note regarding the represented companies in our dataset. First, not all Dutch companies are obliged to deposit their annual reports. This depends on their legal form: only companies with limited liability legal form, and companies that have public shares are obliged to deposit [18]. Second, not all companies that are obliged to deposit their annual reports are obligated to deposit them in XBRL and may therefore not be present in the dataset. Third, the sector code is not a mandatory field to fill in. Therefore, it is likely that the dataset mainly contains small to medium sized companies with a limited liability legal form. Nevertheless, the dataset contains a large amount of data from a wide range of sectors, and is therefore an adequate source for our experiments.

Table 2: Most frequent attributes.

$\mathbf{x}_i$	Feature	Frequency
x_1	EquityAndLiabilities	0.99578
x_2	BalanceSheetBeforeAfterAppropriation…	0.99512
x_3	Assets	0.99301
x_4	Equity	0.98992
x_5	AssetsCurrent	0.96979
x_6	AssetsNoncurrent	0.85330
x_7	ShareCapital	0.68637
x_8	LiabilitiesCurrent	0.67616
x_9	Receivables	0.64786
x_{10}	ReservesOther	0.63630
x_{11}	CashAndCashEquivalents	0.62448
x_{12}	PropertyPlantEquipment	0.42538
x_{13}	Provisions	0.41828
x_{14}	FinancialAssets	0.41599
x_{15}	AssetsCurrentOther	0.30076
x_{16}	Liabilities	0.29789
x_{17}	AssetsNoncurrentOther	0.25422
x_{18}	LiabilitiesNoncurrent	0.24889
x_{19}	Inventories	0.15787
x_{20}	SharePremium	0.14394
x_{21}	IntangibleAssets	0.08732
x_{22}	RetainedEarnings	0.07145
x_{23}	LegalReserves	0.04965
x_{24}	ResultForTheYear	0.03713
x_{25}	LegalStatutoryReserves	0.02481
x_{26}	SecuritiesCurrent	0.02148
x_{27}	RevaluationReserve	0.01919
x_{28}	ConstructionContractsAssets	0.01704
x_{29}	Securities	0.01145
x_{30}	ResultAfterTax	0.00650

x_i	SBI	x_i
...	18129	...
...	49393	...
...	64191	...
...	66192	...
...	16239	...

(a)

x_i	SBI	x_i
...	**18**	...
...	49	...
...	64	...
...	66	...
...	**16**	...

(b)

x_i	SBI	x_i
...	**C**	...
...	H	...
...	K1 *	...
...	K4 *	...
...	**C**	...

(c)

Figure 1: Class reduction.

x_1	x_2
123	NaN
NaN	148
NaN	NaN
984	457

(a)

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

(b)

Figure 2: Encoding missingness.

5 METHODS

The methods needed for the supervised classification problem that we are dealing with are described in this section. In Section 5.1 the steps undertaken during preprocessing are shown. Different approaches to tackle class imbalance are discussed in Section 5.2. Section 5.3 consists of a consideration of the classification algorithm to apply. Finally, Section 5.4 reports the cross validation methods and performance metric used.

5.1 Data preprocessing

The process of class reduction is demonstrated in Figure 1. As described in Section 4, for practical reasons we choose to use a SBI coding that is of a higher hierarchical level. The three to five digit codes (activity, Figure 1a) are transformed into two digit codes (departments, Figure 1b). Note that different departments can be in the same section (marked bold). Then, the two digit codes are transformed into one character code (sections, Figure 1c) except for the section with the highest frequency. This section makes distinctions between the departments and activities within it, by adding a postfix (marked $*$). The split of this section into four subgroups further reduces class imbalance. The final classes we use as target for our classification are listed in Table 3.

An average statement contains 12 attributes and therefore for each row in the tabular file, on average 139 columns are empty. In this step of preprocessing we deal with this missing data. The presence of an attribute (whether a company uses this attribute in financial statements) may be characteristic for a company and its legal form, therefore, a missing-indicator will be added to encode the missingness, a commonly used approach for encoding missing data. This process is illustrated in Figure 2. First, the columns are duplicated. In the first set of columns, missing values ('NaN') are replaced with zero values. In the second set of columns, binary attributes represent whether that an attribute was present (1) or not (0) in the corresponding record. We use the prefix $M\_$ for missing-indicators. The missing-indicator is used to encode missingness so that patterns in missing data may be used in conjunction with patterns in observed data. Adding missing-indicators enables the quantification of the missingness regarding feature importance, assuming that the distribution of missingness is related to the target class distribution. Given the robustness of random forest classifiers however, should this assumption not hold, the results obtained in classifier performances should not alter significantly from the ones obtained by using other imputation methods [22].

5.2 Handling class imbalance

Table 3 shows that there is an imbalance in the classes. 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)**: instead of duplicating instance i, a new instance is synthesized by computing its features as slight variations of the features of the instances that are similar to i, based on sharing a cluster i. This leads to better generalization by decision trees. [4]
(4) **Weighted classes (CSL)**: a cost sensitive learning approach where the weight of a class is inversely proportional to their frequency.

5.3 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 [8]. 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 [12, 20]. 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]. Thus, a random forest classifier was selected.

5.4 Evaluation

Recall from our introduction that the goal of this study is to provide insights in 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

Table 3: Classes and their relative frequency.

Class	Description	Frequency
A	Agriculture, forestry and fishing	0.01474
B	Mining and quarrying	0.00026
C	Manufacturing	0.03661
D	Electricity, gas, steam and air conditioning supply	0.00204
E	Water supply; sewerage, waste management and remediation activities	0.00190
F	Construction	0.04202
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	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 financial institutions	0.03051
K2	Financial institutions - Financial holdings	0.36829
K3	Financial institutions - Investment funds	0.05301
K4	Financial institutions - Insurance and pension funding	0.00081
L	Renting, buying and selling of real estate	0.04390
M	Consultancy, research and other specialised business services	0.18743
O	Public administration, public services and compulsory social security	0.00004
P	Education	0.00593
Q	Human health and social work activities	0.02002
R	Culture, sports and recreation	0.00899
S	Other service activities	0.00490
T	Activities of households as employers; goods and service- producing of households for own use	0.00001
U	Extraterritorial organisations and bodies	0.00002

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 distribution.

6 EXPERIMENTS

In this section, the experiments are outlined. Section 6.1 describes the experimental setup. The results are shown in Section 6.2 and discussed in Section 6.3.

6.1 Experimental Setup

The complete implementation of this experiment is performed in *Python*. Machine learning algorithms and measures were supplied by Scikit-learn [21]. The three sampling class imbalance approaches were implemented using the imbalance-learn module [15]. 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. Roughly 534,000 instances are part of each train set and 59,000 instances part of each test set. 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 [26] are applied to determine whether the class imbalance approaches are statistical significant compared to the regular approach. Hereby, we do not directly assume that an approach is better is there is a slight improvement in performance. We say that p-values below 0.05 are considered to indicate a significant change in the distribution of performance.

The class imbalance approaches as described in Section 5.2 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.

6.2 Results

The overall mean and standard deviation values of performance per class imbalance-handling approach can be regarded in Table 4. 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

Table 4: 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 5: 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

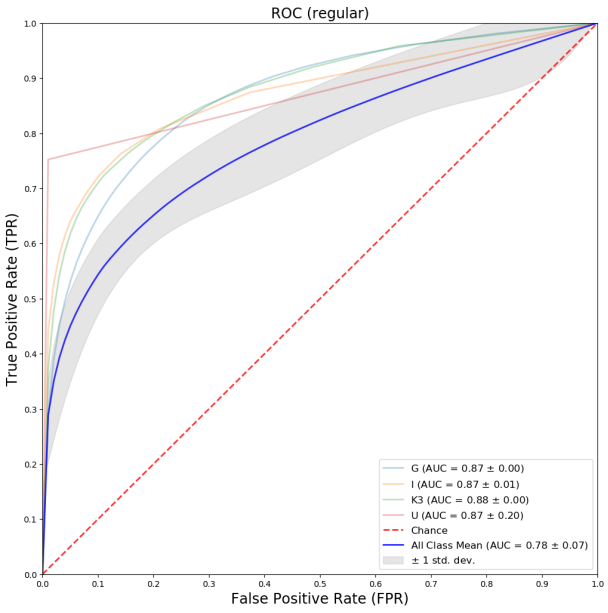

Figure 3: ROC curves and AUC values of overall mean classifier (blue) and the four top performing classes (*Regular*).

classifier yielded within the regular approach to handling class imbalance are listed in Table 5. Figure 3 shows the performance of the regular approach. The mean of the ROC curves per class is represented by the blue line, with the standard deviation in gray. The ten most important features per class are presented in Table 6. A feature's index is based on frequency ranking. The missing-indicator of feature x_i is represented as M_x_i. Table 7 lists the frequency (column '#') and mean importance of features in the previous table.

6.3 Discussion

Without applying a class imbalance approach, we obtained an AUC of 0.78 ± 0.07. This can be interpreted as a adequate result, given that a random prediction would yield an AUC value of 0.5. The results vary per class with a minimum AUC of 0.62 (class O: Public administration, public services and compulsory social security) and maximum of 0.88 (class K3: Financial institutions - Investment funds). Despite that we observe a relationship between the frequencies of classes (see Table 3) and their performances, we can not conclude that this is the only dependency. This is illustrated by classes U (relative frequency = 0.00002, AUC = 0.87), K2 (relative frequency = 0.36829, AUC = 0.84) and O (relative frequency = 0.00004, AUC = 0.62). One explanation could be that the characteristics of some classes are to more extend expressed by unique attributes in financial statements, yielding a better classification result. Despite not having sufficient sample size to conduct performance measurements for class T, feature importances are still computable.

There is no approach with a statistically significant improvement compared to the regular approach. Therefore, the regular approach is used for determining feature importance. Regarding Table 7, six features ('AssetsNonCurrentOther', 'Inventories', 'InterestReceivedClassifiedAsInvestingActivities', 'CalledUpShareCapital', 'CashFlowFromOperations', 'M_InvestmentProperties') were among the ten most important features for all classes. In addition to this list, three features ('PaymentReclaimingValueAddedTax', 'ProceedsSalesIntangibleAssets', 'ChangesValueFinancialAssetsSecurities') were present in the top 10 for a majority of the classes. Remarkably, only two of the twenty most frequent attributes are among the aforementioned lists ('AssetsNoncurrentOther' and 'Inventories'). Among the 20 unique features in the top ten, six missing-indicators are listed. For example, the presence of the attribute 'InventmentProperties' appear to hold considerable information for classification for all classes.

There are several points to note about the obtained results. First, all insights are obtained from one classification algorithm without tuning hyperparameters. The latter also applies to the three class imbalance approaches from the imbalance-learn package. Furthermore, all results are obtained from one dataset, and as stated in Section 4 this dataset is not a perfect representation of all Dutch companies. Given our dataset characteristics and limitations, we thus consider our yielded performance values and results to be adequately computed and therefore valid within our scope.

Table 6: Top 10 most important features per class (*Regular*).

Class	x # 1	x # 2	x # 3	x # 4	x # 5	x # 6	x # 7	x # 8	x # 9	x # 10
A	x_{19}	x_{17}	x_{128}	x_{55}	x_{45}	M_x_{69}	x_{99}	x_{125}	x_{90}	x_{141}
B	x_{17}	x_{19}	x_{55}	x_{128}	x_{45}	x_{99}	M_x_{69}	x_{125}	x_{90}	x_{47}
C	x_{128}	x_{17}	x_{19}	x_{55}	x_{45}	M_x_{70}	M_x_{69}	x_{125}	x_{99}	x_{90}
D	x_{17}	x_{19}	x_{128}	x_{55}	x_{45}	M_x_{69}	x_{99}	x_{125}	x_{90}	x_{47}
E	x_{19}	x_{17}	x_{55}	x_{128}	x_{45}	M_x_{69}	x_{99}	x_{125}	x_{90}	M_x_{20}
F	x_{128}	x_{19}	x_{17}	x_{55}	x_{45}	M_x_{69}	x_{125}	x_{99}	x_{90}	M_x_{70}
G	M_x_{70}	x_{128}	x_{55}	x_{17}	x_{19}	x_{45}	M_x_{61}	M_x_{69}	x_{125}	x_{99}
H	x_{17}	x_{19}	x_{128}	x_{55}	x_{45}	M_x_{69}	x_{125}	x_{99}	x_{90}	M_x_{20}
I	x_{55}	x_{17}	x_{19}	x_{128}	x_{45}	M_x_{70}	x_{125}	M_x_{20}	M_x_{69}	x_{90}
J	x_{19}	x_{17}	x_{128}	x_{55}	x_{45}	M_x_{69}	x_{125}	x_{99}	x_{90}	x_{47}
K1	x_{19}	x_{17}	x_{128}	x_{55}	x_{45}	M_x_{69}	x_{99}	x_{125}	x_{90}	x_{47}
K2	x_{128}	x_{45}	x_{17}	x_{19}	x_{55}	x_{111}	M_x_{69}	M_x_{35}	M_x_{59}	x_{99}
K3	x_{128}	M_x_{35}	x_{17}	x_{19}	x_{55}	x_{45}	M_x_{69}	M_x_{59}	x_{99}	x_{125}
K4	x_{19}	x_{17}	x_{55}	x_{128}	M_x_{35}	x_{45}	M_x_{69}	x_{99}	x_{125}	x_{90}
L	x_{128}	x_{19}	x_{17}	x_{55}	x_{45}	M_x_{20}	M_x_{69}	M_x_{59}	x_{99}	x_{125}
M	x_{128}	x_{17}	x_{19}	x_{55}	x_{45}	M_x_{69}	x_{125}	x_{99}	x_{90}	M_x_{59}
O	x_{19}	x_{17}	x_{55}	x_{128}	x_{45}	x_{125}	x_{99}	M_x_{69}	x_{90}	x_{43}
P	x_{17}	x_{19}	x_{55}	x_{128}	x_{45}	M_x_{69}	x_{125}	x_{90}	x_{99}	x_{47}
Q	x_{19}	x_{17}	x_{128}	x_{55}	x_{45}	M_x_{69}	x_{90}	x_{99}	x_{125}	x_{47}
R	x_{19}	x_{17}	x_{55}	x_{128}	x_{45}	M_x_{69}	x_{99}	x_{125}	x_{90}	M_x_{20}
S	x_{17}	x_{19}	x_{55}	x_{128}	x_{45}	M_x_{69}	x_{125}	x_{99}	x_{90}	M_x_{20}
T	x_{17}	x_{19}	x_{55}	x_{128}	x_{125}	M_x_{69}	x_{45}	x_{90}	x_{99}	x_{106}
U	x_{128}	x_{17}	x_{19}	x_{55}	x_{141}	x_{85}	x_{99}	M_x_{69}	x_{125}	x_{45}

Table 7: Top occurring feature count and importance (*Regular*).

x_i	Feature	#	Importance
x_{17}	AssetsNoncurrentOther	23	0.090
x_{19}	Inventories	23	0.089
x_{128}	InterestReceivedClassifiedAsInvestingActivities	23	0.085
x_{55}	CalledUpShareCapital	23	0.083
x_{45}	CashFlowFromOperations	23	0.066
M_x_{69}	M_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}	M_SharePremium	6	0.033
M_x_{59}	M_InterestReceivedClassifiedAsOperatingActivities	4	0.033
M_x_{70}	M_IncreaseDecreasePayablesCreditInstitutions	4	0.023
M_x_{35}	M_SumOfExpenses	3	0.030
x_{141}	ResultBeforeTaxOrdinaryActivities	2	0.033
x_{43}	CashFlowOperatingActivities	1	0.025
x_{111}	LineItemsOtherIncomeStatementReceiptsPaymentsNotConsideredOperatingActivities	1	0.025
x_{106}	RevaluationReserveRelease	1	0.013
M_x_{61}	M_IncreaseDecreaseProvisions	1	0.008
x_{85}	CashFlowsOperatingActivitiesOther	1	0.002

7 CONCLUSIONS AND FUTURE WORK

In this study we have performed explainable business sector prediction by applying machine learning on financial statements. This enables applications such as the detection of mislabeled company statements and potential cases of fraud, overall augmenting data accuracy. First, we determined the performance of random forest classifiers, in terms of a ROC curve for each class and mean of their AUCs. An AUC of 0.78 ± 0.07 demonstrates the suitability the approach. We determined feature importance and thereby give insight in the most important features for business sector prediction.

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, without a strong connection to the frequency of the feature. Missing-indicators appeared as important feature for a majority of the classes which indicated that the presence of an attribute on a financial statement is occasionally at least as important as the value itself. This is not surprising, as it is well known that certain assets are only present on balance sheets of specific companies. Interestingly, the aforementioned insights are discovered automatically by the machine learning algorithm by using a large number of attributes from balance sheets and income statements. This enables domain experts such accountants to select a small number of attributes characteristic for a sector, reducing a tremendous amount of manual work. In summary, we conclude that machine learning by means of supervised learning on financial statements can be used for accurate business sector prediction. Our conclusion can serve as a baseline for future studies.

We make five recommendations for future work. The first two imply a repetition of the experiments on a later moment in time: the combination of several versions of the dataset at different points in time, containing financial statements over different years, will yield a bigger dataset that would result in a better performance. Besides, an increased dataset can be used for time series analysis of the development of all companies or specific sectors. Another future direction could be the comparison between different classification algorithms for sector prediction. Due to global and European standards for sector categorization, contribution could be made by combining international datasets. At last, future work could deepen our knowledge of sector categorization by using other methods and data sources, such as text mining on written annual reports.

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] Enrico Evink and Leo van der Tas. 2018. Digitalisering van de jaarrekening: het gebruik van XBRL in gedeponeerde jaarrekeningen van middelgrote ondernemingen. *Maandblad voor Accountancy en Bedrijfseconomie* 92 (2018), 361–373. https://doi.org/10.5117/mab.92.30421

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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.

[14] P. Kruiskamp. 2019. Standaard Bedrijfs Indeling 2008. Retrieved May 5, 2019 from https://www.cbs.nl/-/media/_pdf/2018/01/sbi%202008%20versie%202018%20engels.pdf

[15] 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

[16] Statistics Netherlands. 2012. Relaties tussen (inter)nationale standaardclassificaties. Retrieved June 10, 2019 from https://www.cbs.nl/-/media/imported/onze%20diensten/methoden/classificaties/documents/2012/26/schemaclassificaties.pdf

[17] Statistics Netherlands. 2018. Standard Industrial Classifications (Dutch SBI 2008, NACE and ISIC). Retrieved June 10, 2019 from https://www.cbs.nl/en-gb/our-services/methods/classifications/activiteiten/standard-industrial-classifications--dutch-sbi-2008-nace-and-isic--

[18] Netherlands Chamber of Commerce. 2019. Handleiding jaarrekeningen maart 2019. Retrieved May 5, 2019 from https://www.kvk.nl/download/Handleiding_jaarrekening_tcm109-464530.pdf

[19] 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/

[20] 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

[21] 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.

[22] 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.

[23] 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

[24] Anuj Sharma and Prabin Panigrahi. 2013. A Review of Financial Accounting Fraud Detection Based On Data Mining Techniques. *International Journal of Computer Applications* 39 (2013), 37–47. https://doi.org/10.5120/4787-7016

[25] 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

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

[27] XBRL-International. 2014. An Introduction to XBRL. Retrieved June 9, 2019 from https://www.xbrl.org/the-standard/what/an-introduction-to-xbrl/

[28] 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.