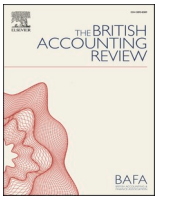
The British Accounting Review 54 (2022) 101096

Contents lists available at ScienceDirect

The British Accounting Review

journal homepage: www.elsevier.com/locate/bar



Predicting industry sectors from financial statements: An illustration of machine learning in accounting research

Hans van der Heijden

University of Sussex, Brighton, BN1 9RH, United Kingdom

A B S T R A C T

The main aim and contribution of this study is to outline and demonstrate the usefulness of a machine learning approach to address prediction-based research problems in accounting research, and to contrast this approach with a more conventional explanation-based approach familiar to most accounting scholars. To illustrate the approach, the study applies machine learning to predict a firm's industry sector using the firm's publicly available financial statement data. The results show that an algorithm can predict an industry sector with just this data to a high degree of accuracy, especially if a non-linear classifier is used instead of a linear classifier. Additionally, the algorithms were able to carry out an industry-firm pairing exercise taken from introductory accounting text books and MBA cases, with predicted answers showing a high degree of accuracy in carrying out this exercise. The study shows how machine learning approaches and algorithms can be valuable to a range of accounting domains where prediction rather than explanation of the dependent variable is the main area of concern.

1. Introduction

The main aim and contribution of this study is to outline and demonstrate the usefulness of a machine learning approach to address specific research problems in accounting research, and to contrast this approach with a more conventional explanation-based approach familiar to most accounting scholars. To illustrate the approach, the study sets out to predict a firm's industry sector, as specified by the North American Industry Classification System (NAICS), using the firm's publicly available financial statement data. The results show that an algorithm can predict an industry sector with just this data to a high degree of accuracy, especially if a non-linear classifier is used instead of a linear classifier.

The main difference between a machine learning approach and a conventional approach is that a machine learning approach is prediction-orientated whereas the conventional approach is explanation-orientated. In other words, a machine learning approach focuses primarily on the out-of-sample prediction of the dependent variable rather than the explanation of the dependent variable within-sample (Bao, Ke, Li, Yu, & Zhang, 2020). Prediction is not necessarily the same as explanation (Shmueli, 2010), and the machine learning approach is of value to a range of applications where prediction of a dependent variable is the main, and perhaps only, concern. Such applications are common in business and economics research (Kleinberg, Ludwig, Mullainathan, Nber, & Obermeyer, 2015). The measurement of success in prediction-orientated approaches is out-of-sample prediction accuracy rather than within-sample significance levels (p-values), and the theoretical specification of the conceptual model is, to a degree, determined by the algorithm, rather than a priori by the researcher.

The specific strand of machine learning of relevance to this study is the development of algorithms for 'supervised' or deep learning, where an algorithm is trained on a training set so it learns the connections between independent variables and dependent variables (Geron, 2019). The algorithm then uses this knowledge to predict dependent variables from independent variables on which it was not trained. The now classic example of supervised AI is the computerised recognition of hand-writing, where the algorithm learns the

E-mail address: h.vanderheijden@sussex.ac.uk.

https://doi.org/10.1016/j.bar.2022.101096
Received 20 April 2020; Received in revised form 6 April 2022; Accepted 9 April 2022
Available online 5 May 2022

connection between digits and the images of their hand-written equivalents (Lecun, Bottou, Bengio, & Haffner, 1998). Once it has learned the connections, the algorithm can then apply this knowledge to other images of hand-written digits for which the actual value is unknown. The specific algorithms used by supervised AI are the subject of intense research, and the volume of papers making contributions to this area is vast and growing.

Any categorical variable of interest to accounting researchers and practictioners can be a candidate application for supervised machine learning. Sample studies could cover the prediction of categorical variables extracted from the management discussion of the annual report, such as the high-level prospects of the firm, its environmental commitments, the nature of its operations, its business strategy, and its long-term outlook. Applications also include the prediction of market-based reactions such as investor sentiment and media response. For practising accountants, applications include the prediction of auditable categories that are either financially material, historically subject to manipulation, or politically sensitive. Examples include sales tax (VAT) categories and trade export status.

A small, not exhaustive, list of examples in the recent accounting literature may reinforce the broad applicability of machine learning. One recent study used machine learning to convert transcripts of conference calls into emotional traits, and link them to performance outcomes (Hrazdil, Novak, Rogo, Wiedman, & Zhang, 2019). Another recent study examined financial statement data to classify the likelihood of a firm committing accounting fraud (Bao et al., 2020). A stand-out domain is the prediction of future financial distress from publicly available financial statement data, given the benefits of accurate prediction such as the opportunity for timely remediation. This area of research is traditionally fertile, and has existed long before machine learning became popular. It has produced a variety of popular metrics, including the popular Altman Z-score (Altman, 1968). A number of reviews have been published that may serve as useful pointers (Alaka et al., 2018;Ravi Kumar & Ravi, 2007; Sun, Li, Huang, & He, 2014). Reviews also address some of the methodological challenges inherent to statistical methods and financial statement data (Balcaen & Ooghe, 2006). Early research used multiple discriminant analysis (Altman, 1968). This was then followed by logistic regression (Ohlson, 1980), and subsequently a range of other methods. In recent years, machine learning approaches have been used, such as neural networks (Khashman, 2010) and ensemble classifiers (Nanni & Lumini, 2009). Barboza et al. employ a range of machine learning models, including the random forest algorithm (Barboza, Kimura, & Altman, 2017). Studies suggest that the decision-tree based classifiers perform better than others (Gepp & Kumar, 2015).

In this paper we demonstrate the benefits of the machine learning approach by focusing on the prediction of a firm's industry sector from its public financial statement data. In selecting this prediction problem we drew inspiration from an exercise well known to accounting educators: to "connect the dots" between a set of industry sectors and a set of anonymous companies for which only certain financial ratios are revealed. Typically, financial accounting textbooks (for example, Libby, Libby, & Hodge, 2017) provide a select number of financial statements and a number of industries. Students are then asked to match the firm with the corresponding industry. They arrive at the solution using deductive reasoning and accumulated background knowledge. For example, manufacturers and hotels have higher fixed assets, service firms have no cost of goods sold, luxury firms have higher profit margins, and so on. This exercise is common in accounting classes for MBA students too. Over the years, Harvard Business School has published a number of cases in which students are asked to match the industry with a set of financial statement items (Crane & Reinbergs, 2000). A number of variations on this theme have been developed since (Bruner & Opitz, 1988; Bujaki & Durocher, 2012; Demers, 2011).

Moving away from accounting education, there are also realistic scenarios in practice where an AI generated industry code is beneficial to analysts. First, the industry code may be missing altogether because the firm is new or unknown, and a peer group cannot immediately be identified. Second, the actual industry code may be suspect, for example when the firm has diversified into other industries without updating its original industry code. It may also be that the firm is hiding its 'true' industry code if it fears that disclosing its true code will attract unwanted attention. Finally, the firm may be classified as "unclassified", which means technically the firm has an industry code, but in reality it does not.[1]

Being able to generate a missing industry code, or potentially replace a suspect industry code with an AI industry code can also have advantages for accounting scholars. Reviews show that researchers use industry codes such as the Standard Industry Classification (SIC) or the North American Industry Classification System (NAICS) extensively (Bhojraj, Lee, & Oler, 2003; Kahle & Walkling, 1996; Phillips & Ormsby, 2016; Weiner, 2011). Common uses are to provide descriptive statistics, to derive performance benchmarks, to exclude firms from study, to identify peer groups, and to predict market-based performance (examples include Abad, Ferreras, & Robles, 2020; Kou & Hussain, 2007). Researchers can impute synthetic codes into a dataset with missing industry codes to avoid firms being excluded from analysis. Analysts can study potential anomalies when synthetic codes predict entirely different industries from the actual code. And finally, analysts can use AI-generated codes to gain some additional insight into those firms with a suspect or "unclassified" industry code.

In summary, the purpose of this study is to demonstrate how machine learning and artificial intelligence can accurately predict the correct industry code for a company by just looking at the financial statements of that company. We use a sample of 80,790 firm-year combinations from the Compustat database, spanning a 10-year period from 2010 to 2019. The algorithm is trained on 64,632 firm-year combinations, and then tested on 16,158 different firm-year combinations. The algorithms developed in this paper generate AI industry codes for each firm. The performance of the predictions is assessed and analysed using the machine learning approach, where out-of-sample prediction is the critical measure of performance: we measure how accurately the algorithms avoid false positives (known as precision) and how accurately they avoid false negatives (known as recall).

[1] For example, at the time of writing, the SIC code of Google UK is 82990: "Other business support service activities not elsewhere classified".

The structure of the paper is as follows. We first illustrate the machine learning approach in more detail for the benefit of accounting scholars who are perhaps more used to a conventional approach. In this section we also describe a machine learning technique called random forest which is different from linear regression. We then briefly discuss previous literature that is relevant to the prediction of industry codes from financial statement data. This helps identifying the independent variables that we need for our study. The next section describes the data set collected for this study, and the subsequent section describes the results of the machine learning approach, and the prediction results. A special section of supplementary analysis returns to the original "connect-the-dots" Harvard Business School case exercise that we drew inspiration form, and shows how AI can perform this exercise, and potentially outperform human students. A discussion of results concludes the paper.

2. Machine learning approach

Fig. 1 depicts the machine learning approach used in this study and contrasts it with the conventional approach more familiar to most accounting scholars.

In this explanation-based approach, hypotheses are formulated that link independent variables to dependent variables (X represents x_1, x_2, …x_n independent variables, and y represents the dependent variable). A regression analysis is then conducted on the entire dataset. The regression is usually a linear combination of the independent variables that best fits the actual value of the dependent variable. Goodness of fit measures are obtained to examine the performance of the regression, and to test the hypotheses, p-values of the regression coefficients are examined to check whether the impact of the independent variable of interest is significantly different from zero. Interaction effects can be included to examine any non-linear effects.

In the prediction-based approach used in this study, the dataset is first split into a training and a test set. The training set is used to train the algorithm, i.e., to tell the algorithm which values of y need to be associated with which values of X. The test set is put aside to make sure it will not be affected by any further analysis. Similar to the conventional approach, a function is extracted from the training set that best maps the independent variables to the dependent variables. This could be a regression function (ie., a linear combination of the independent variables), but it could also be a non-linear function. Examples include decision trees (mapping X to y using if/then statements) and neural networks (mapping X to y using weightings and layers of nodes).

The function that is the output of the training set is then used to predict y values of the X values of the test set. The resulting predicted values are then compared with the actual y values of the test set. False positives and false negatives are then analysed, and performance measures known as precision and recall are used to expressed the performance *out-of-sample*. It is often the case that the

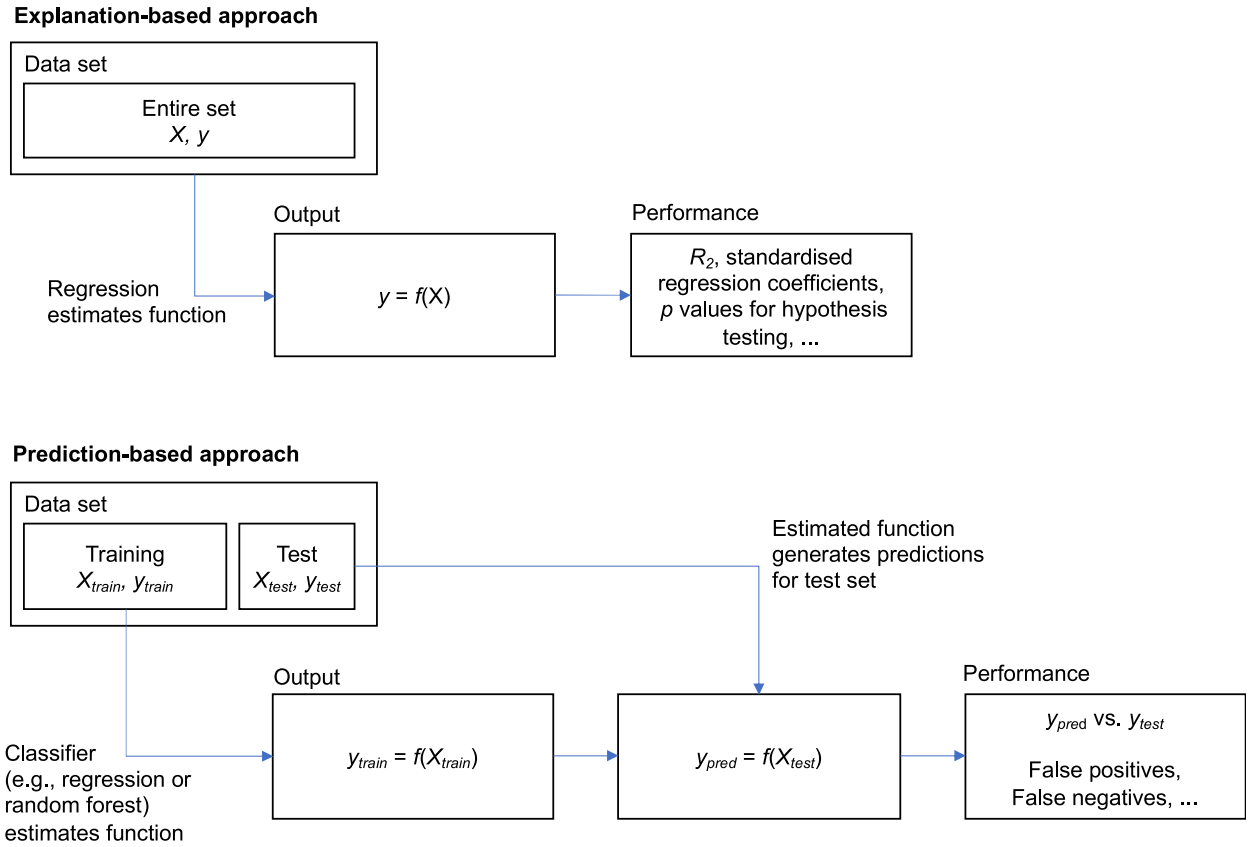


Fig. 1. Explanation-based approach vs prediction-based approach.

function does very well at predicting the *y* values of the training set (because this is what the algorithm is supposed to optimise), but much less so at predicting the values of the test set.

The algorithm that produces the function is known as the 'classifier'. This study will use the linear discriminant analysis classifier and the random forest classifier. Standard regression cannot be used as the dependent variable of interest is categorical, but the linear discriminant analysis is a close alternative to regression. Optimising the choice of classifier, and configuring the parameters of a classifier is known as 'hypertuning'.

A random forest classifier uses a number of decision tree classifiers, each working on a different random subset of the training set (Breiman, 2001; Ho, 1995). Each decision tree classifier will then cast a vote for a particular class based on its results, and the target class that gets the most votes is the one selected (Geron, 2019).

Fig. 2 shows an analogy with a marble run to further illustrate the concept of the random forest algorithm. In this picture, the marble falls down, hits the spikes along the way, and depending on the spike pattern, it arrives at any particular bucket (Panel A). We could picture the marble as the values of the independent variables, x_1, x_2, …, x_n, and the buckets as the values of the dependent variable *y*. When random forest builds a decision tree, it can be said to build the spike pattern of the marble run. Panel B depicts a stylised example of such a decision tree, with the *x* values needing to meet specific criteria, to fall-through to the next node (spike).

The machine learning approach has two key differences with the conventional approach to be aware of. First, there is much less emphasis on the parameters of the function derived from the training set (explained variance, *p*-values, etc.). In the conventional approach, these parameters are of crucial importance because they often provide the verdict on whether a hypothesis is accepted or rejected. The machine learning approach does not have hypotheses in the traditional sense (as it focuses on out-of-sample prediction), so these parameters have less relevance.

The second difference is that there is much more emphasis on out-of-sample prediction accuracy. In the conventional approach, the total sample is considered to be taken from the entire population. Whether the function derived from the set is able to predict other cases not from the set is less important than the inferences that can be made within sample. However, in the machine learning approach, prediction accuracy of new cases the algorithm has not previously seen is the ultimate success measure.

3. Industry codes

Two popular industry classification schemes are the Standard Industry Classification (SIC), and the North American Industry Classification System (NAICS). The history and background of these schemes is documented elsewhere (e.g., Krishnan & Press, 2003; Pagell & Weaver, 1997). SIC and NAICS schemes are both 'production-technology-based' classifications. The difference is that SIC uses a mix of production processes and product-output categories, whereas NAICS uses only production processes. NAICS also includes emerging industries and more fine-grained types of services (Krishnan & Press, 2003; Walker & Murphy, 2001).

There is a body of literature that has examined how good the standard industry classification schemes are at combining similar firms and separating dissimilar firms. Several studies report that these schemes may not be sufficiently discriminative for the purposes of accounting research (Bhojraj et al., 2003; Clarke, 1989; Krishnan & Press, 2003). Given the historic origins of some schemes, notably SIC, there may not be sufficient coverage of the variety of activities in the high-tech services sector (Kile & Phillips, 2009). The industry classifications also do not group together firms that are tightly coupled in supply chains (Fan & Lang, 2000). And finally, conglomerates are ill-suited for the classification scheme as only one main industry class is reported (Amit & Livnat, 1990). In response, new industry classifications have been proposed based on input-output tables (Fan & Lang, 2000) and system-based features (Dalziel, 2007). More recently, innovative industry classification schemes have been proposed based on text analysis of business descriptions (Fang, Dutta, & Datta, 2013), product descriptions (Hoberg & Phillips, 2016), and financial statement patterns (Chong & Zhu, 2012; Yang, Liu, Zhu, & Yen, 2019).

These new developments notwithstanding, the SIC and NAICS schemes remain in widespread use in the accounting and finance literature. Reviews suggest a number of typical use cases. These include the identification of peer groups for a firm, the screening of firms in order to limit the scope of investigation, the development of industry-wide performance benchmarks, and the use of industry codes to provide descriptive statistics of the sample under study (Bhojraj et al., 2003; Kahle & Walkling, 1996).

4. Data

This section is split into three parts. The first part will discuss specifics of the data set. The second part will discuss the dependent variable, which the machine learning literature refers to as "targets". The next part will discuss the independent variables, called "features" (Chollet, 2018; Geron, 2019).

We selected firm-year data from North American companies from 2010 to 2019. Data was downloaded in March 2020 from Compustat through WRDS. The "North America – Daily" collection of datasets was selected, and within this collection, the "Fundamentals Annual" dataset. The date variable was "Data Date" and set to "2010–01" and "2019–12". Screening variables were kept at default settings. All 974 data types were selected in the raw download.

The study uses the NAICS code as its main target, or dependent variable. We use a specification of two-digit codes for the NAICS scheme. This ensures that the number of outputs is not too small to be meaningless, but not too large to be intractable. Although each code is made up of digits, they are nominal variables and not numeric, because the digits carry no ordinal or metric value.

Each industry code is mapped to a "target class code", to make sure that the scale of the variable is continuous, a requirement for the classifiers used in this study. To be sure, the classifiers still treat this target class variable as a nominal variable, but they will transform the continuous value into a dummy representation through a process called 'one-hot-encoding'.

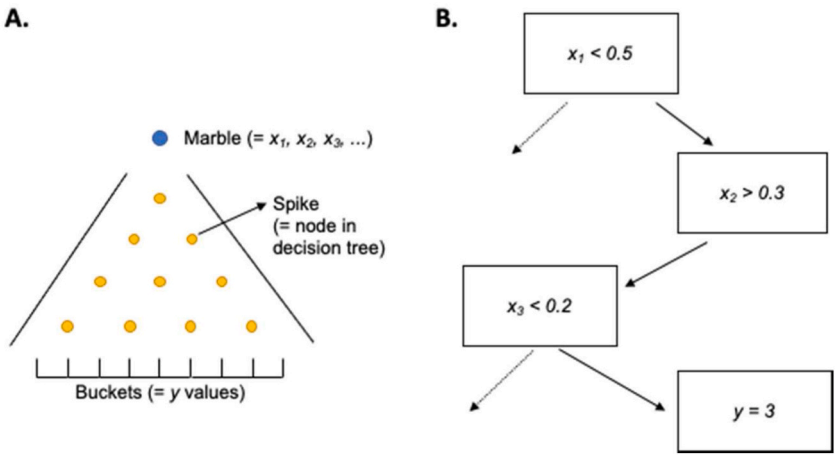


Fig. 2. Diagram visualising decision trees generated by the random forest algorithm. Panel A depicts the analogy with the marble run, and Panel B depicts a stylised version of a tree representing the if-then statement: if $x_1 < 0.5$ and $x_2 > 0.3$ and $x_3 < 0.2$ then $y = 3$.

The mapping of a target class code to the two-digit NAICS sector is displayed in Table 1. The Compustat data used in this study did not contain firms with NAICS code 55 (Management of Companies and Enterprises) and 92 (Public Administration).

The data points were reduced to a useable set for the classifiers as follows. All firm-year observations where total assets was not positive or not available were discarded first. We then looked at the relevant balance sheet items (discussed next), and discarded every firm-year where one of the balance sheet items was negative or not less than total assets. This ensures that all common size percentages calculations had an expected value between 0 and 1. We then looked at all data points that had no valid NAICS code, or where the NAICS code was 'unclassified'. Table 2 summarises the results of this process.

In developing the feature set, or independent variables, for the classification, guidance was provided by the accounting education cases on pairing the firm's industry with the firm's financial statement (Crane & Reinbergs, 2000). These cases specify a combination of common size percentages for each balance sheet item, topped up with a select set of financial ratios covering the income statement. Examples include the cash balance sheet item scaled by total assets, and the R&D expenses scaled by sales revenue. As mentioned before, examples of these exercises are plentiful in the accounting education literature so the use of common-size financial statements appears a good place to start.

Our final selection contains 15 common size percentages and 12 ratios provided in the Crane and Reinbergs (2000) case. This case includes more than 15 common size percentages, but because common size percentages are subject to degrees of freedom, a common size percentage was not included if it could be readily calculated from other common size percentages. For example, total current assets was included but total non-current assets was not included because it can be readily derived as (1 – total current assets).

The Appendix to this paper shows how each common size percentage and ratio was calculated using data items from the Compustat database.

Table 1
Mapping of 18 target classes to two-digit NAICS industry codes (United States Census Bureau, 2017).

Target Class Code	NAICS Code starts with	Description	Acronym	Notes
0	11	Agriculture, Forestry, Fishing and Hunting	AGR	
1	21	Mining	MIN	
2	22	Utilities	UTL	
3	23	Construction	CON	
4	31–33	Manufacturing	MNF	
5	42	Wholesale trade	WHO	
6	44–45	Retail Trade	RET	
7	48–49	Transportation & warehousing	TRA	
8	51	Information	INF	
9	52	Finance & Insurance	FIN	
10	53	Real-Estate Rental & Leasing	EST	
11	54	Professional, Scientific, and Technical Services	PRO	
	55	*Management of Companies and Enterprises*		*Not in database*
12	56	Administrative and Support and Waste Management and Remediation Services	ADM	
13	61	Educational Services	EDU	
14	62	Health Care and Social Assistance	HLC	
15	71	Arts, Entertainment, and Recreation	ART	
16	72	Accommodation and Food Services	ACC	
17	81	Other Services (except Public Administration)	OTH	
	92	*Public Administration*		*Not in database*

Table 2
This table shows the process of reducing the sample of firm-year observations to a set with valid independent and dependent variables.

	Firm-years discarded	Number of firm-years
Available from Compustat		122,303
Rows where total assets not positive, or NaN (not a number)	28,505	93,798
Rows where any balance sheet items is $\geq$ total assets.	11,397	82,401
Rows where any balance sheet item is < 0	855	81,546
Rows with NAICS code absent or 'unclassified (99+)'	756	80,790
Final sample with valid NAICS codes only		**80,790**

Once data screening was complete, all common size percentages and financial ratios were computed as outlined in the Appendix. Data was then split in a training set and test set. Following guidelines by Geron (2019), the test set is a random set of 20% of the total set (16,158 rows). The test set was stratified to ensure it had proportional representation of each of the target classes. The training set has 64,632 rows. Rows were also randomly shuffled as part of this process.

Following on from the split into training and test set, the training set was scaled (standardised) using the RobustScaler from the SciKit-Learn library.[2] The test set was then scaled using the scaling algorithm of the training set scaler, ensuring that the scaling is not affected by the test set.

In this study we use two classifiers, guided by prior research on classification of financial statement data. The first classifier is linear discriminant analysis. This is a standard classifier that implements a method similar to multiple discriminant analysis. The linear discriminant analysis (LDA) classifier was implemented using the SciKit-Learn (Pedregosa et al., 2011) library.[3] Standard settings were used, with the solver set to the default singular value decomposition solver. The random forest classifier was also implemented using the SciKit-Learn (Pedregosa et al., 2011) library.[4] Default settings were used, which means 100 decision tree classifiers were generated, and to establish the quality of a split the Gini measure was employed.

5. Results

The median values of the features for the target NAICS codes are displayed in Table 3. These descriptives are of interest because they provide insight into the potential predictive value of the feature. For example, some features are zero or near-zero for most firms: A6 Investments, L3 Unearned Revenue, and E2: Preferred Stock. It is unlikely these features will provide much predictive value in separating out firms into industry sectors.

Correlations between the features are not tabulated for brevity, but it is apparent that some features are, virtually by definition, strongly correlated. For example, L5 Total Long-term Debt and R12 Long Term Debt/Capital. Given their high correlation, these features could serve as potential candidates for exclusion in future analysis.

It is common in accounting literature to measure the performance of classifiers using the Receiver Operating Curve (ROC) (see e.g. Jackson & Wood, 2013). However it is not possible to do this here because the ROC is confined to binary classifiers, and is not appropriate for multi-label classification. Instead, we follow the approach common in the machine learning literature (Geron, 2019), and work with *confusion matrices*, *precision*, *recall*, and the F_1 *score*.

Confusion matrices look at actual values and predicted values of the test set. All actual values are represented as rows, and all predicted values are represented as columns. Confusion matrices form a visual companion to the most common performance measures in classification machine learning, precision and recall. Precision is the accuracy of positive predictions, defined as $TP/(TP + FP)$, where TP is the number of true positives, and FP is the number of false positives. Recall is the accuracy of true positives, defined as $TP/(TP + FN)$, where TP is again the number of true positives, and FN is the number of false negatives (Geron, 2019).

There is a trade-off in precision and recall: it is possible to improve precision by sacrificing recall, and likewise, it is possible to improve recall by sacrificing precision. For this reason, another performance measure has been developed, the F_1 score, which is the harmonic mean of precision and recall (Geron, 2019). The F_1 score gives precision and recall equal weighting, and will be used in this study as the deciding performance measure.

Table 4 provides our final results, for each classifier. The results show much higher scores for the Random Forest classifier than linear discriminant analysis. The precision score, indicating absence of false positives, is 89.52%, which means that for every 100 firms where the algorithm predicted the industry code to be some value, it was correct 89 times out of 100. The recall score, indicating absence of false negatives, is 89.33%, which means that for every 100 firms where the actual industry code was X, the algorithm has these correct 89 times out of 100.

We now present the two confusion matrices for the two classifiers in Table 5. The diagonal (shaded gray in the tables) shows correct values, meaning those firm-year data points where the predicted industry code was equal to the actual industry code. All cells not on the diagonal are incorrect values, meaning the predicted industry code was not the actual code. A perfect, 100% score of precision, recall and F_1 would have all values on the diagonal and zeroes in all remaining cells not on the diagonal.

[2] See: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html.
[3] See: https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html.
[4] See: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Table 3
Median values of features by NAICS target code (n = 80,790 firm-years). See Table 1 for NAICS acronyms (e.g., AGR = Agriculture). Feature legend: A1 = Cash, A2 = Receivables, A3 = Inventory, A4 = Total Current Assets, A5 = Net Plant & Equipment, A6 = Investments, A7 = Goodwill & Intangibles. L1 = Accounts Payable, L2 = Total Debt in Current Liabilities, L3 = Unearned Revenue, L4 = Total Current Liabilities, L5 = Total Long-term Debt, L6 = Total Liabilities. E1 = Preferred Stock, E2 = Common Stock. R1 = Gross Margin, R2 = R&D/Sales, R3 = Net Income/Sales, R4 = Days of Receivables, R5 = Inventory Turnover, R6 = Fixed Asset Turnover, R7 = Total Asset Turnover, R8 = Net Income/Assets, R9 = Net Income/Equity, R10 = Assets/Equity, R11 = Debt/Equity, R12 = Long Term Debt/Capital.

	AGR	MIN	UTL	CON	MNF	WHO	RET	TRA	INF	FIN	EST	PRO	ADM	EDU	HLC	ART	ACC	OTH
A1	0.05	0.08	0.01	0.09	0.18	0.04	0.09	0.04	0.19	0.02	0.01	0.14	0.09	0.35	0.06	0.08	0.06	0.03
A2	0.05	0.02	0.04	0.12	0.12	0.21	0.04	0.04	0.10	0.02	0.01	0.21	0.15	0.07	0.13	0.02	0.02	0.05
A3	0.08	0.00	0.02	0.13	0.10	0.22	0.24	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03
A4	0.36	0.17	0.09	0.35	0.56	0.62	0.52	0.12	0.41	0.00	0.00	0.51	0.41	0.53	0.28	0.15	0.14	0.22
A5	0.37	0.76	0.72	0.10	0.13	0.11	0.26	0.70	0.07	0.01	0.00	0.05	0.09	0.14	0.15	0.47	0.53	0.19
A6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A7	0.01	0.00	0.01	0.04	0.08	0.12	0.05	0.03	0.28	0.01	0.01	0.27	0.21	0.11	0.31	0.21	0.09	0.30
L1	0.03	0.04	0.03	0.07	0.06	0.13	0.12	0.02	0.03	0.01	0.01	0.05	0.05	0.03	0.03	0.02	0.03	0.03
L2	0.01	0.00	0.03	0.02	0.01	0.01	0.01	0.03	0.01	0.02	0.02	0.01	0.01	0.00	0.01	0.01	0.01	0.01
L3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L4	0.11	0.08	0.10	0.18	0.19	0.27	0.28	0.11	0.22	0.00	0.00	0.27	0.25	0.27	0.17	0.12	0.14	0.14
L5	0.10	0.00	0.29	0.17	0.08	0.15	0.13	0.32	0.07	0.04	0.43	0.05	0.12	0.00	0.21	0.32	0.26	0.19
L6	0.31	0.25	0.68	0.54	0.43	0.58	0.56	0.57	0.49	0.88	0.54	0.47	0.56	0.42	0.55	0.57	0.57	0.64
E1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E2	0.67	0.73	0.30	0.45	0.55	0.41	0.43	0.41	0.50	0.12	0.41	0.52	0.44	0.57	0.42	0.41	0.40	0.35
R1	0.24	0.00	0.31	0.18	0.33	0.20	0.33	0.32	0.57	0.12	0.29	0.36	0.30	0.55	0.30	0.35	0.25	0.27
R2	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R3	0.05	0.00	0.09	0.02	0.02	0.02	0.02	0.06	0.03	0.00	0.01	0.03	0.02	0.06	0.02	0.02	0.04	0.04
R4	38.54	12.95	51.67	58.97	54.18	44.24	7.35	37.72	61.64	41.44	18.32	78.45	56.85	27.10	52.86	13.02	10.34	30.60
R5	4.05	0.00	10.58	1.41	3.39	6.77	4.96	17.78	0.00	0.00	0.00	0.00	0.00	0.00	2.65	22.61	47.47	10.22
R6	1.60	0.06	0.39	7.37	4.88	16.89	7.47	0.50	7.95	0.00	0.00	15.76	11.52	6.04	5.19	1.09	2.06	3.99
R7	0.44	0.04	0.28	0.83	0.74	1.77	1.87	0.36	0.57	0.04	0.10	0.87	1.06	0.89	0.87	0.52	0.82	0.99
R8	0.02	−0.07	0.03	0.02	0.02	0.04	0.04	0.03	0.02	0.01	0.02	0.03	0.03	0.05	0.02	0.02	0.03	0.02
R9	0.05	−0.10	0.09	0.06	0.04	0.09	0.10	0.08	0.04	0.08	0.05	0.06	0.07	0.08	0.06	0.04	0.09	0.07
R10	1.48	1.36	3.30	2.21	1.80	2.44	2.31	2.41	2.00	8.58	2.43	1.93	2.29	1.76	2.36	2.44	2.49	2.88
R11	0.20	0.05	1.07	0.54	0.25	0.56	0.43	0.90	0.26	0.64	1.13	0.23	0.42	0.06	0.64	0.90	0.72	0.74
R12	0.13	0.00	0.48	0.28	0.12	0.27	0.23	0.44	0.12	0.23	0.50	0.08	0.20	0.00	0.34	0.44	0.39	0.36

Table 4
Precision, Recall and F_1 scores for two classifiers, comparing predicted versus actual NAICS codes for test set (n = 16,154 Firm-years).

	Classification Accuracy Metric
1. Linear Discriminant Analysis	
Precision	67.19%
Recall	67.09%
F_1	66.38%
2. Random Forest	
Precision	89.52%
Recall	89.33%
F_1	89.01%

Table 5
Confusion matrices for linear discriminant analysis and random forest classifiers, classifying NAICS industry codes (n = 16,158 Firm-Years). Rows represent actual values, columns represent predicted values. For example, cell AGR/MIN (value 14) means that 14 firm-years with NAICS code AGR where (incorrectly) classified as Mining.

Confusion matrix for LDA classifier

	AGR	MIN	UTL	CON	MNF	WHO	RET	TRA	INF	FIN	EST	PRO	ADM	EDU	HLC	ART	ACC	OTH
AGR	0	9	2	0	28	0	0	5	1	0	1	1	0	0	0	0	0	0
MIN	14	1536	74	0	278	1	5	70	17	3	26	4	0	2	1	0	2	0
UTL	4	101	361	0	19	0	0	37	15	2	7	3	0	0	3	0	0	0
CON	0	22	2	57	58	6	2	2	12	1	1	30	1	0	0	0	0	0
MNF	6	197	26	20	3504	92	127	36	304	19	24	27	5	4	0	1	4	0
WHO	2	11	5	0	146	92	33	14	10	1	2	13	0	0	2	0	3	0
RET	0	2	0	0	150	18	238	3	17	3	4	0	0	0	0	0	5	0
TRA	4	184	53	0	33	6	0	103	15	0	2	0	11	0	1	2	2	0
INF	4	57	17	0	422	2	0	18	590	14	17	62	4	9	7	6	2	0
FIN	6	15	3	53	126	2	2	6	132	3480	453	22	7	1	0	1	1	2
EST	1	69	42	3	53	2	4	39	53	30	756	4	0	0	3	4	3	2
PRO	2	11	1	0	154	4	0	1	151	1	1	64	6	0	0	1	0	0
ADM	0	13	3	0	35	13	2	6	60	1	2	35	27	0	4	0	0	0
EDU	0	2	0	0	33	0	0	0	22	0	0	7	0	3	0	0	2	0
HLC	0	6	5	0	54	1	0	12	60	0	0	8	5	0	3	1	1	0
ART	0	22	4	0	13	0	0	17	23	1	3	4	0	0	0	1	0	0
ACC	0	39	9	13	19	1	1	39	36	3	6	0	0	2	0	0	27	0
OTH	2	2	0	0	7	1	2	0	7	8	0	0	0	0	0	0	0	0

Confusion matrix for random forest classifier

	AGR	MIN	UTL	CON	MNF	WHO	RET	TRA	INF	FIN	EST	PRO	ADM	EDU	HLC	ART	ACC	OTH
AGR	10	7	1	0	24	0	0	3	1	1	0	0	0	0	0	0	0	0
MIN	0	1885	9	1	99	0	3	11	9	9	3	1	1	0	0	0	1	1
UTL	0	47	478	0	6	0	0	11	4	1	2	3	0	0	0	0	0	0
CON	0	14	0	126	35	1	0	0	4	3	2	7	2	0	0	0	0	0
MNF	0	95	2	4	4184	13	26	1	45	11	6	4	0	1	0	1	3	0
WHO	0	9	3	0	78	219	5	5	8	1	0	3	0	0	0	0	3	0
RET	0	1	0	0	54	4	373	0	6	0	1	0	0	1	0	0	0	0
TRA	0	52	6	0	7	1	0	334	6	2	4	1	0	0	0	1	2	0
INF	0	30	1	4	107	2	1	3	1043	18	3	10	2	0	3	0	4	0
FIN	0	23	2	1	32	0	3	4	41	4180	20	4	0	1	1	0	0	0
EST	0	33	3	1	24	0	1	3	31	56	903	8	0	2	1	1	1	0
PRO	0	7	1	1	90	4	0	1	64	7	3	215	2	0	1	0	1	0
ADM	0	11	0	0	21	1	1	1	26	2	1	12	122	0	3	0	0	0
EDU	0	1	0	0	0	0	1	0	16	2	0	2	0	47	0	0	0	0
HLC	0	8	1	0	25	1	0	1	20	3	0	10	3	1	83	0	0	0
ART	0	2	2	0	9	0	0	5	8	6	1	3	0	0	0	50	2	0
ACC	0	3	1	2	4	0	0	5	5	2	5	0	0	0	0	2	166	0
OTH	0	2	0	0	8	0	0	0	1	0	0	0	0	0	1	0	0	17

6. Supplementary analysis: Harvard Business School case

In this section we explore how accurate a machine learning model would be in carrying out a typical "pair the firm with the industry" exercise found in introductory financial accounting textbooks and MBA cases. To examine this we ran the algorithms on the same Harvard Business School industry-pairing exercise that was discussed in the introduction, and used to identify the features for classification (Crane & Reinbergs, 2000). The case lists 10 firms, identified as A to J, and 10 industries. In the case, the student is asked to pair the firm with the correct industry. More specifically, the training set remained the same, but the test set is now taken to be the 10 firms. We recognise the anachronistic setting of this analysis, in that the training data is from 2010 to 2019, and the test data from 2000.

Following on from the machine learning approach, we analysed the predictions of each classifier and compared them with the

correct industry code for each firm. Table 6 shows the detailed results for each classifier. The table shows that the random forest classifier is able to correctly predict the industry of 8 out of 10 firms, and the linear discriminant analysis predicts 6 out of 10. Both classifiers struggled with the online retailer (thinking it was a manufacturer) and the hotel (thinking it came from the transportation or real-estate industry).

The random forest algorithm also produces probabilities for each industry sector, indicating how likely it believes that a firm is of a particular industry. We have mapped these probabilities graphically for this exercise, and they can be seen in Fig. 3.

7. Discussion

The main finding of our research is that, using a machine learning approach and using machine learning classifiers, it is possible to predict the industry sector of a company from its financial statement data to a high level of accuracy (9 out of 10). The non-linear random forest classifier outperformed the linear discriminant analysis. In the supplementary analysis, the main finding is that the algorithms were able to carry out the industry-pairing exercise, with predicted answers showing good prediction accuracy.

The findings of the study demonstrate the usefulness of a machine learning approach, and its orientation towards prediction out-of-sample as opposed to explanation within-sample. In terms of the relative performance of classification methods, our study is in line with previous financial statement distress prediction, which suggests that the random forest (or decision-tree based ensemble) classifiers are the most predictive in this line of research (Barboza et al., 2017; Gepp & Kumar, 2015).

The implications of these findings are that we can meaningfully generate AI industry codes that will allows us to address some of the challenges associated with the use of standard industry classification schemes. If an industry code is missing or otherwise unavailable for a firm, and it is desirable to have one, then these algorithms can create a synthetic code instead. If an industry code is suspect for some reason, then these algorithms can generate a synthetic code and the appropriateness of the actual code can be reviewed and appraised. This study contributes to this literature by offering a solution for creating a "stand-in" industry code in these circumstances.

Returning to the domain of accounting education, we have illustrated how machine learning can be used to solve reasonably complex exercises such as the industry-pairing exercise. These exercises were perhaps traditionally seen as requiring human judgement, and not within the scope of algorithmic applications. Of course, a number of caveats can be made. First, the machine learning algorithms required additional, machine-readable, specification of the exercise (such as the decision to predict the NAICS code). The algorithms also did not "strike out" a sector from the universe of possible answers once the sector had been paired to a firm. Another caveat is that it is not possible for algorithms to discern firms at the more granular level of industry sector that they operated. For example, three firms in the exercise were manufacturers of specific items, and the algorithms would not be able to say which one was which.

This study has a number of limitations. We made a number of deliberate choices to reduce the scope of the analysis. The study confined itself to NAICS industry codes, and used North-American financial statement data. This implies that the findings do not necessarily carry over to other countries, especially those where other standard industry schemes are in use. Further research could examine the performance of these classifiers in a more international context. A further limitation is that we kept the industry codes at a high level, to keep the study tractable. Further studies could potentially extend the granularity of the codes, so that industries can be classified at a more fine-grained level.

There are other areas for future research. The first is related to improving the classifiers. It may be possible to exceed the 90% accuracy by using non-financial statement data, for example by incorporating natural language text from annual reports and other company filings. We also kept the parameters of these classifiers at their default settings, and it is possible that fine-tuning these parameters (hyper-tuning) might result in better performance.

A second area for further research is to investigate if we can reduce the number of financial statement data items. We previously identified that some data items were near-constant, or exhibited high multi-collinearity. Further research could explore the minimum number of financial statement data required to establish an acceptable accuracy.

In conclusion, the main aim and contribution of this study was to outline and demonstrate the usefulness of a machine learning approach to address specific research problems in accounting research, and to contrast this approach with a more conventional

Table 6
This table shows the results of the two classifiers in predicting the correct industry from an industry-financial statement pairing exercise (Crane & Reinbergs, 2000).

Firm	Industry	Correct NAICS code	Prediction Linear Discriminant Analysis	Prediction Random Forest
A	Online Retailer	RET	MNF	MNF
B	Supermarket	RET	RET – OK	RET – OK
C	Hotel	ACC	TRA	EST
D	Airline	TRA	UTL	TRA – OK
E	Consumer Products	MNF	MNF - OK	MNF – OK
F	Pharmaceuticals	MNF	MNF – OK	MNF – OK
G	Electronic Comms	MNF	MNF – OK	MNF – OK
H	Warehouse Club	RET	RET – OK	RET – OK
I	Staffing Agency	ADM	ADM – OK	ADM – OK
J	Software Developer	INF	MNF	INF–OK
TOTAL CORRECT			6	8

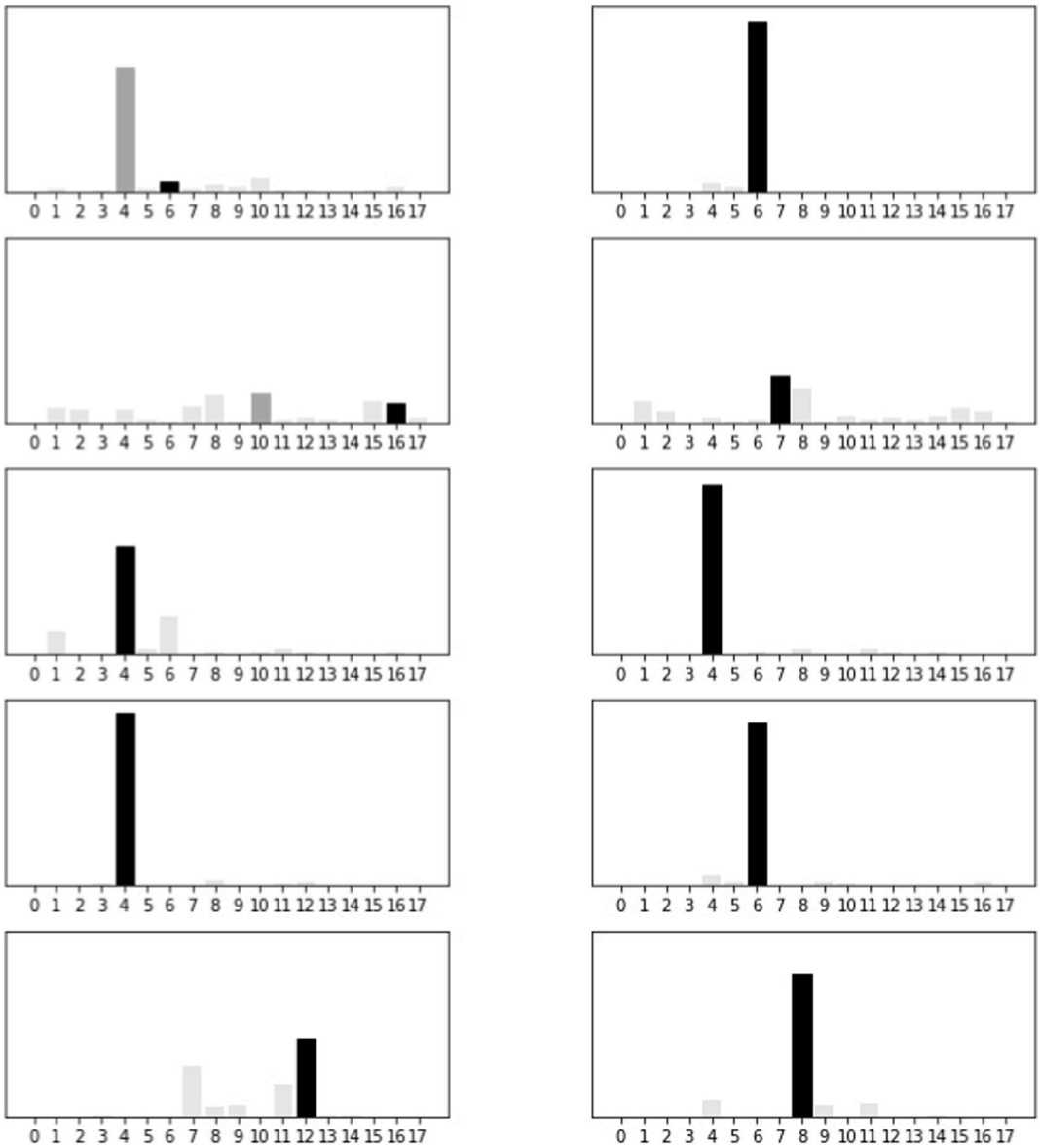


Fig. 3. Probabilities for each of the ten firms in the exercises, as produced by the random forest algorithm. The range 0–17 refers to the NAICS codes as explained in Table 1. Black bars indicate the actual industry code value of the firm. Dark gray bars indicate situations where the algorithm chose an incorrect sector. For example, the third firm was predicted to be value 10 (NAICS EST, or Real Estate) but was actually value 16 (NAICS ACC, or Accommodation & Food services).

regression-based approach familiar to most accounting scholars. We hope that this provides a good introduction in machine learning to accounting scholars who are less familiar with machine learning approaches and methods, and that it will lead to machine learning and artifical intelligence eventually being used more widely in accounting research and scholarship.

Appendix

Mapping of Compustat data types to features.

	Common size components		Compustat items used	Calculation	Notes
	ASSETS				
A1	Cash & Marketable Securities		CHE	CHE/AT	
A2	Receivables		RECT	RECT/AT	0 if RECT not available
A3	Inventories		INVT	INVT/AT	0 if INT not available
	Other current assets	Not included (= A4 – A3-A2 - A1)			
A4	Total current assets		ACT	ACT/AT	0 if ACT not available
A5	Net Plant & Equipment		PPENT	PPENT/AT	0 if PPENT not available
A6	Investments		IVAEQ	IVAEQ/AT	0 if IVAEQ not available
A7	Goodwill & Intangibles		INTAN	INTAN/AT	0 if INTAN not available
	Other non-current assets	Not included (= 1 – A4 – A5 – A6 – A7)			
	Total non-current assets	= 1 – A4			
	Total assets	= 1			Always 1
	LIABILITIES				
L1	Accounts Payable		AP	AP/AT	0 if AP not available
L2	Total debt in current liabilities		DLC	DLC/AT	0 if DLC not available
L3	Unearned Revenues		UI	UI/AT	0 if UI not available
	Other current liabilities	Not included = L4 – L3 – L2 – L1			
L4	Total current liabilities		LCT	LCT/AT	0 if LCT not available
L5	Total long-term debt		DLTT	DLTT/AT	0 if DLTT not available
	Other non-current Liabilities	Not included = L6 – L5 – L4			
L6	Total liabilities		LT	LT/AT	0 if LT not included
	EQUITY				
E1	Preferred stock		PSTK	PSTK/AT	0 if PSTK not included
E2	Common stock		CEQ	CEQ/AT	0 if CEQ not included
	Total stockholder equity	Not included = E1 + E2			
	Total Liabilities & Equity	= 1			Always 1
	RATIOS				
R1	Gross Margin		SALE, COGS	(SALE – COGS)/SALE	
R2	R&D/Sales		XRD, SALE	XRD/SALE	
R3	Net Income/Sales		IB, SALE	IB/SALE	
R4	Days of Receivables		RECT, SALE	RECT (SALE/365)	
R5	Inventory Turnover		COGS, INVT	COGS/INVT	
R6	Fixed Asset Turnover		SALE, PPENT	SALE/PPENT	
R7	Total Asset Turnover		SALE, AT	SALE/AT	
R8	Net Income/Assets		IB, AT	IB/AT	
R9	Net Income/Equity		IB, CEQ	IB/CEQ	
R10	Assets/Equity		AT, CEQ	AT/CEQ	
R11	Debt/Equity		DLC, DLTT, CEQ	(DLC + DLTT)/CEQ	
R12	L/T Debt/Total Capital		DLTT, PSTK, CEQ, DLTT	DLTT/(PSTK + CEQ + DLTT)	

References

Abad, P., Ferreras, R., & Robles, M. D. (2020). Intra-industry transfer effects of credit risk news: Rated versus unrated rivals. *The British Accounting Review, 52*(1). https://doi.org/10.1016/j.bar.2018.12.002

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., et al. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications, 94*, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. https://doi.org/10.2307/2325319

Amit, R., & Livnat, J. (1990). Grouping of conglomerates by their segments' economic attributes: Towards a more meaningful ratio analysis. *Journal of Business Finance & Accounting, 17*(1), 85–100. https://doi.org/10.1111/j.1468-5957.1990.tb00551.x

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

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. Firms using a machine learning approach. *Journal of Accounting Research, 58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006
Bhojraj, S., Lee, C. M. C., & Oler, D. K. (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research, 41*(5), 745–774. https://doi.org/10.1046/j.1475-679X.2003.00122.x
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
Bruner, R. F., & Opitz, C. S. (1988). The financial detective. *Darden Business Publishing Cases, 1*(1), 1–5. https://doi.org/10.1108/case.darden.2016.000300
Bujaki, M., & Durocher, S. (2012). Industry identification through ratio analysis. *Accounting Perspectives, 11*(4), 315–322. https://doi.org/10.1111/1911-3838.12003
Chollet, F. (2018). *Deep learning with phyton*. Shelter Island, NY: Manning.
Chong, D., & Zhu, H. (2012). Firm clustering based on financial statements. In *, Vol. 2012. 22nd workshop on information technologies and systems, WITS* (pp. 43–48).
Clarke, R. N. (1989). SICs as delineators of economic markets. *Journal of Business, 62*(1), 17–31.
Crane, D. B., & Reinbergs, I. (2000). Drivers of industry financial structure. In *Harvard business School exercise 201-039, september*.
Dalziel, M. (2007). A systems-based approach to industry classification. *Research Policy, 36*(10), 1559–1574. https://doi.org/10.1016/j.respol.2007.06.008
Demers, E. (2011). Balance sheet detective. *INSEAD Case, 167*(1), 1–3.
Fang, F., Dutta, K., & Datta, A. (2013). LDA-based industry classification. *International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design, 3*, 2500–2509.
Fan, J. P. H., & Lang, L. H. P. (2000). The measurement of relatedness: An application to corporate diversification. *Journal of Business, 73*(4), 629–660. https://doi.org/10.1086/209657
Gepp, A., & Kumar, K. (2015). Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science, 54*, 396–404. https://doi.org/10.1016/j.procs.2015.06.046
Geron, A. (2019). *Hands-on machine learning with SciKit-learn, keras & tensorflow* (2nd ed.). Sebastopol, CA: O'Reilly.
Ho, T. K. (1995). Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994
Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy, 124*(5), 1423–1465. https://doi.org/10.1086/688176
Hrazdil, K., Novak, J., Rogo, R., Wiedman, C., & Zhang, R. (2019). Measuring executive personality using machine-learning algorithms: A new approach and audit fee-based validation tests. *Journal of Business Finance & Accounting*, 1–26. https://doi.org/10.1111/jbfa.12406
Jackson, R. H. G., & Wood, A. (2013). The performance of insolvency prediction and credit risk models in the UK: A comparative study. *The British Accounting Review, 45*(3), 183–202. https://doi.org/10.1016/j.bar.2013.06.009
Kahle, K. M., & Walkling, R. A. (1996). The impact of industry classifications on financial research. In *, Vol. 31. Journal of financial and quantitative analysis*.
Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications, 37*(9), 6233–6239. https://doi.org/10.1016/j.eswa.2010.02.101
Kile, C. O., & Phillips, M. E. (2009). Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing and Finance, 24*(1), 35–58. https://doi.org/10.1177/0148558X0902400104
Kleinberg, J., Ludwig, J., Mullainathan, S., Nber, H., & Obermeyer, Z. (2015). Prediction policy problems. *The American Economic Review, 105*(5), 491–495.
Kou, W., & Hussain, S. (2007). Predictive gains to segmental disclosure matrices, geographic information and industry sector comparability. *The British Accounting Review, 39*(3), 183–195. https://doi.org/10.1016/j.bar.2007.05.002
Krishnan, J., & Press, E. (2003). the North American industry classification system and its implications for accounting research. *Contemporary Accounting Research, 20*(4), 685–717. https://doi.org/10.1506/N57L-0462-856V-7144
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791
Libby, R., Libby, P. A., & Hodge, F. (2017). *Financial accounting* (9th ed.). New York: McGraw-Hill.
Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*(2 PART 2), 3028–3033. https://doi.org/10.1016/j.eswa.2008.01.018
Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109. https://doi.org/10.2307/2490395
Pagell, R. A., & Weaver, P. J. S. (1997). NAICS: NAFTA's industrial classification system. *Business Information Review, 14*(1), 36–44. https://doi.org/10.1177/0266382974236192
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
Phillips, R. L., & Ormsby, R. (2016). Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship, 21*(1), 1–25. https://doi.org/10.1080/08963568.2015.1110229
Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. *European Journal of Operational Research, 180*(1), 1–28. https://doi.org/10.1016/j.ejor.2006.08.043
Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330
Sun, J., Li, H., Huang, Q. H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems, 57*, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006
United States Census Bureau. (2017). *North American industry classification system executive office of the president office of management and budget*.
Walker, J. A., & Murphy, J. B. (2001). Implementing the North American industry classification system at BLS. *Monthly Labor Review, 124*(12), 15–21.
Weiner, C. (2011). The impact of industry classification schemes on financial research. *SSRN Electronic Journal*, 1–48. https://doi.org/10.2139/ssrn.871173
Yang, S. Y., Liu, F. C., Zhu, X., & Yen, D. C. (2019). A graph mining approach to identify financial reporting patterns: An empirical examination of industry classifications. *Decision Sciences, 50*(4), 847–876. https://doi.org/10.1111/deci.12345