Detecting Misstatements in Financial Statements

Vincent Chiu
School of Computing Science
Simon Fraser University
vlc4@sfu.ca
vchiuwork@gmail.com
301095537

Vishal Shukla
School of Computing Science
Simon Fraser University
vshukla@sfu.ca
301337060

Kanika Sanduja
School of Computing Science
Simon Fraser University
ksanduja@sfu.ca
301347347

1 Motivation and Background

Financial statements are reports which corporations use to convey their current financial situation to the public including investors and shareholders. A misstatement occurs when a corporation misrepresents their financial situation in their financial report. Financial statement manipulation is an ongoing problem in corporate America. In a world driven by performance targets and high share prices, no one wants to be left behind. This pressure often leads corporations to manipulating their statements to portray a better but false financial picture. We are motivated to help auditors target companies that are more likely to make misstatements and enable investors to consider the misstatement risks before investing. We have created a program that can detect with high accuracy whether a given financial statement is misstated or not. We hope that our data product will help society invest with confidence and that corporations will become more responsible.

2 Problem Statement

Our primary goal is to build a program that classifies financial statements as a misstatement or not a misstatement.

We wish to answer the following questions:

- Is it possible to detect whether or not a given financial report has been misstated?
- Which industry has the most misstatements?
- Is it possible to use unsupervised learning to identify corporations that are outliers?
- Is there a correlation between a corporation being an outlier and submitting misstated financial reports?
- How correlated are the fields of a financial statement?
- What are the most common reasons for misstatements?

We had the following challenges:

- Financial dataset is large with many fields and requires specialized knowledge.
- There were multiple data sources that we had to integrate.
- There were few examples of actual misstated financial statements.
- The majority of features contain mostly null values.

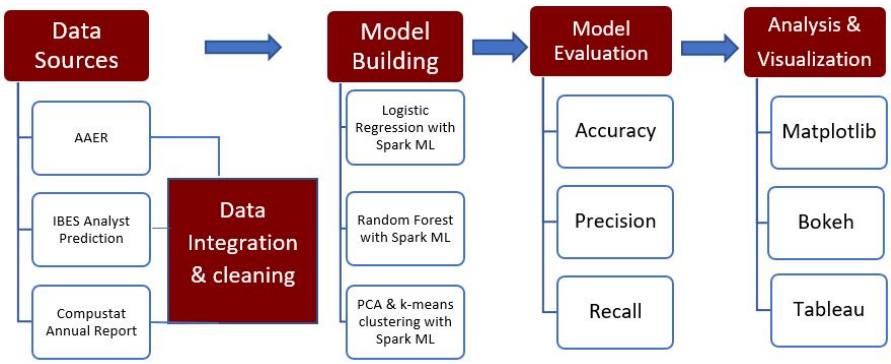

Figure 1: An overview of our project in a flow chart.

3 Data Science Pipeline

3.1 Datasets

We used three main datasets. Firstly, we used Compustat annual report data which consists of over 500,000 financial statements with over 1000 corporations included. Secondly, we used Accounting and Auditing Enforcement Releases (AAER) data. This data represents companies found guilty of filing misstatements. We used this to prepare our training data's class label. We also used IBES analyst earnings per share prediction data.

3.2 Data Integration

We performed data integration by performing a left join between the annual report data and the IBES prediction data on stock ticker and report year. To create the class label for misstatements, we wrote a user defined function for Spark and assigned a class label of misstatement (value of 1) to a given financial report if there was a corresponding AAER record with the same stock ticker and year. We assigned a class label of non-misstatement (value of 0) otherwise. Initially, we did not have the AAER dataset and we had to assign the class label of misstatement to records in which the retained earnings adjustment (REA) was not equal to zero. However, we could not achieve good results with our model with this label due to it being noisy. This is because there are other reasons other than misstatements that could result in retained earnings adjustment being non-zero.

4 Methodology

4.1 Preprocessing

4.1.1 Feature Engineering & Data Cleaning

We consulted our subject matter expert, Dr. Kim Trottier and obtained a list with about 250 features that she believed to be the most important for classifying misstatements. We extracted the aforementioned columns in our integrated dataset. We performed Principal Component Analysis and discovered that many of the features are highly correlated. Please see figure 3. This justifies only retaining the features suggested by our subject matter expert as we would not be throwing away that much variability in our data. We only used the numerical variables as most of the string variables were not helpful in classifying misstatements. The columns with the string data type were mostly unique identifiers such as the business name or address. Therefore, we decided that it would be appropriate to drop the variables with the string data type. We then took this dataset and imputed zeros in place of the null values.

4.1.2 Class Balancing

We began by extracting the approximately 1500 misstatements. Due to the large class imbalance, we down-sampled the non-misstatements by randomly sampling 1500 non-misstatements. Therefore, we

had around 3000 total samples. 72% of the data was used for training and 18% was used for validation. 10% of the data was used for testing. We used the validation set to tune our hyper-parameters. We used the training set to train each of our models and we used the test set for measuring our accuracy, precision and recall.

4.2 Exploratory Data Analysis

We initially explored the data by making several plots. In the annual financial report dataset, there were many null values, this is because some industries have their own specific columns unique to them. Our initial strategy was to build a model that was specific to one industry as our minimum viable product. Therefore, we decided to make a visualization that will help us find the industry with the most misstatements so that we would have the most training examples. Please see Figure 4, for the visualization that we made. In the first iteration of our models, we only looked at corporations in the Prepackaged Software Services industry. In the subsequent iterations, we extended our analysis to other industrial segments as well as the complete dataset. As part of our EDA, we also experimented with performing k-means clustering on PCA components. However, the clusters that formed did not correspond to misstatements. We may investigate if there is any deeper meaning behind these clusters in the future.

4.3 Supervised Learning

4.3.1 Logistic Regression

We used random forest and logistic regression. We used the logistic regression model because we analyzed Dechow's [1] paper and wanted to see if we could replicate or exceed the performance of Dechow's logistic regression model. For implementation, we used the LogisticRegression class from the pyspark.ml.classification package for PySpark. [2]

4.3.2 Random Forest Model

Next, we decided to use a random forest model. We believe that random forest is a good choice because the results are more interpretable by accountants compared to neural networks. Each decision tree can be independently analyzed. Random Forest models are also very flexible and are capable of capturing non-linear boundaries between classes.

Please see Figure 1 for a flow chart of how the data science pipeline and models fit together.

4.3.3 Implementation

We used Spark.ML packages for both our random forest and logistic regression models. This is because the distributed nature of Spark allows us to harness the power of the cluster to train our models quickly.

4.3.4 Parameter Tuning using Validation Set

Spark offers TrainValidationSplit for hyper-parameter tuning [3]. For each of the supervised learning models, we performed a grid search for the best hyper-parameters. The best hyper-parameters for random forest when optimizing for accuracy was 200 trees with a maximum depth of 16. For logistic regression model the best hyper-paramters were a regularization parameter of 0.1 and a threshold of 0.6.

4.4 Unsupervised Learning

We attempted to segregate observations into two clusters hoping that each cluster would correspond to each class label. However, the actual class labels does not correspond to the clusters formed. As seen in the left plot of Figure 2, the misstated records are dispersed across all of the observations. Whereas, the output of the k-means groups all observations with high values of Principal Components into one cluster and the rest into another. In the future, we will consider investigating the outliers in the plot to see what makes the corporations outliers.

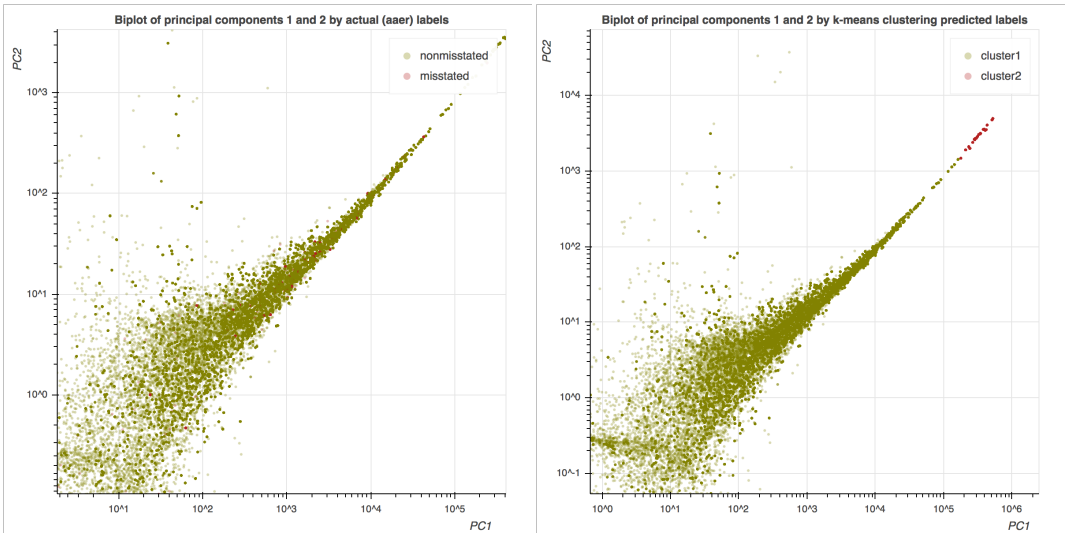

Figure 2: Left: misstatements are dispersed throughout the data points. Right: PCA plot with k-means clustering, however clusters did not correspond to class labels

4.5 Tools

- For Data Processing and Machine Learning: Spark, Spark ML
- For Visualization: matplotlib, Bokeh, Tableau.

5 Evaluation and Results

The results were evaluated with a test set with balanced classes. The evaluation metrics used include accuracy, precision and recall.The size of the test set was about 10 percent of the size of the training set. We attained high accuracies and fairly good precision and recall values for both our random forest and logistic regression models. For our project, we believe that recall is more important than precision. We expect auditors to be the primary user of our product. Auditors have to check all the financial reports anyway, so it is important that we alert auditors to as many likely misstatements as possible. In our models, both recall and precision are high which is desirable.
Note that recall = sensitivity = true positive rate (TPR)
precision = positive predictive value

$$\text{recall} = \text{sensitivity} = \frac{\#\text{ true positive}}{\#\text{ true positive} + \#\text{ false negative}}$$

$$\text{precision} = \text{positive predictive value} = \frac{\#\text{ true positive}}{\#\text{ true positive} + \#\text{ false positive}}$$

Evaluation Metrics Below are the evaluation metrics, tuned random forest refers to the random forest model's performance after its hyper-parameters have been optimized for accuracy.

model name	tuned random forest	untuned rf	logistic reg. tuned for f-measure
accuracy (%)	82.738	81.818	70.503
misstatement precision (%)	83.636	78.198	64.454
misstatement recall (%)	81.657	86.627	87.226
non-misstatement precision (%)	81.871	86.013	82.022
non-misstatement recall (%)	83.832	77.298	54.784

6 Data Product

Our data product is a program that takes in a set of financial statements as input and produces output labels as to whether these financial statements have been misstated or not.

7 Lessons Learned

- We learned how to apply machine learning models to real-time financial data with a large class imbalance.
- In terms of descriptive statistics, we discovered that the most common industries in which manipulations occurred are computers and computer services, retail, and general services.
- Using time series analysis, we discovered that there may be some relationship between organizations having misstatements and having the highest variation in the difference between the actual Earnings per Share and the Analyst Predicted Earnings per Share.
- From our logistic regression model, the features with the largest absolute weights include: Director's Emoluments, Auditors' remuneration, Dividends per share, Auditor's rank, gain or loss on sale of property and auditors' opinion.

feature	feature name	logistic_reg_coefficient
rmum	Auditors' Remuneration	-95.99
emol	Directors' Emoluments	-29.05
dvpsp_f	Dividends per Share - Pay Date - Fiscal	-0.30
auopic	Auditor Opinion	0.19
sret	Gain/Loss on Sale of Property	0.32
rank	Rank - Auditor	0.69

8 Summary

We created a program using the random forest model that is able to detect misstatements with 82.7% accuracy. Other metrics include 83.6% misstatement precision and 81.7% misstatement recall. We discovered many interesting insights which include:

- Some of the features with greatest weight for logistic regression include director's emoluments and amortization of goodwill. Director's emoluments are negatively correlated with committing misstatements since if directors are well paid, the organization will work efficiently and have lower chances of fraud. Another feature is amortization of goodwill. When a firm pays above the fair market price to acquire another firm, this overpayment is recorded as goodwill. This goodwill is amortized over the life of the net assets. Amortization of goodwill is positively correlated with committing a misstatement because the more a firm overpays for the acquisition of other firms, the more likely they are in financial distress.
- An attempt to partition the statements into two clusters: misstatements and non-misstatements and to identify corporations with outliers, was made using unsupervised models.

This model can be beneficial to financial institutions for three main reasons:
First, it is easy to use for personnel without programming backgrounds, such as auditors or investors, making the model highly accessible.
Second, it utilizes a wide range of data sources (three diverse datasets) which provides a balanced view of the financial status of an organization, making the model robust.
Third, we developed the model on Spark, which can scale up to even larger datasets having many features and records, making the model scalable.

9 Future Work

Some future ideas we wish to work on include:

- We will consider predicting future misstatements.
- We would also like to improve on our unsupervised clustering results to better understand the outliers and be able to correctly distinguish between the two classes.
- We will explore using LSTM and RNN models to improve our classification accuracy.

10 Acknowledgements

Acknowledgements: Thank you to Dr. Kim Trottier, Dr. Steven Bergner, Dr. Jiannan Wang, Hiral Patwa, and Simranjit Singh Bhatia

References

[1] Patricia M Dechow, Weili Ge, Chad R Larson, and Richard G Sloan. Predicting material accounting misstatements. *Contemporary accounting research*, 28(1):17–82, 2011.

[2] Apache Software Foundation. Classification and regression - spark 2.3.0 documentation, 2018.

[3] Apache Software Foundation. Ml tuning: model selection and hyperparameter tuning, 2018.

11 Appendix

This is version 30.

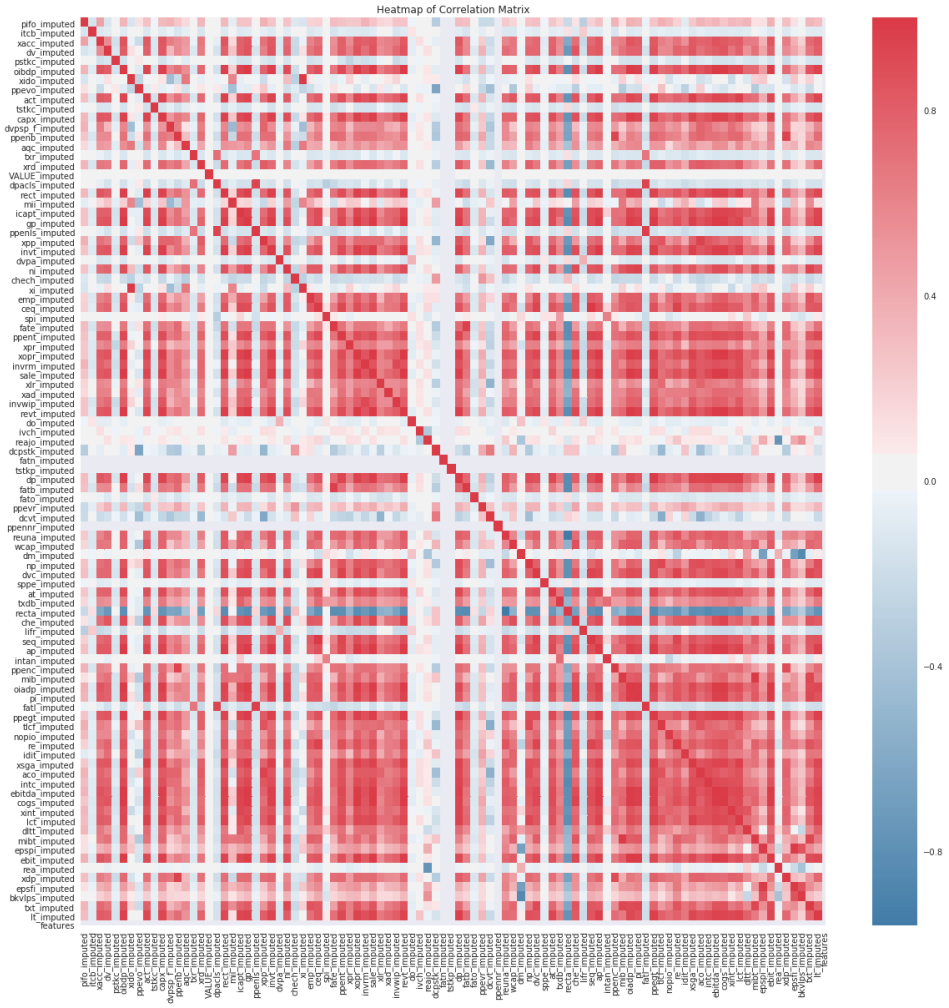

Figure 3: Correlation between Features in Financial Statements; Many features are highly correlated; red is highly positively correlated, blue is highly negatively correlated and white is no correlation

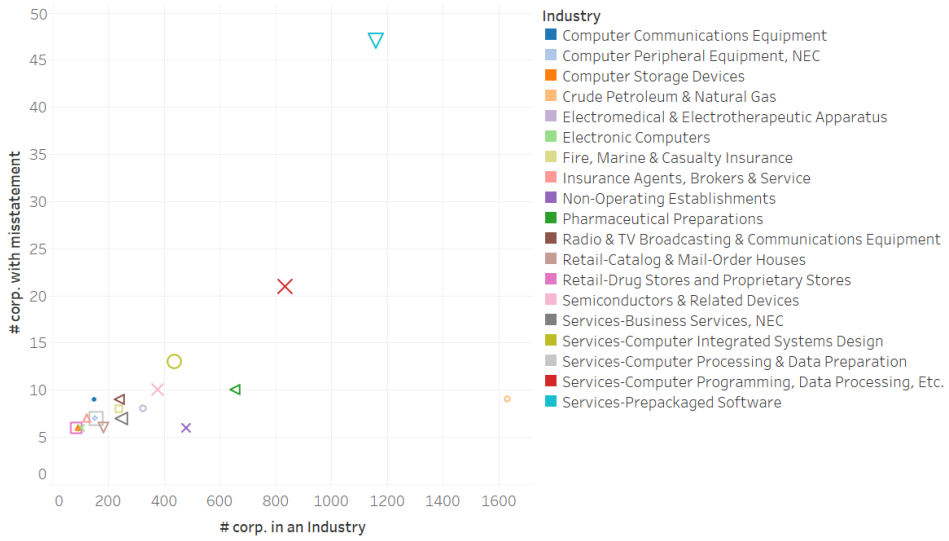

Figure 4: Number of corporations with at least one misstatement vs. total number of corporations in a given industry

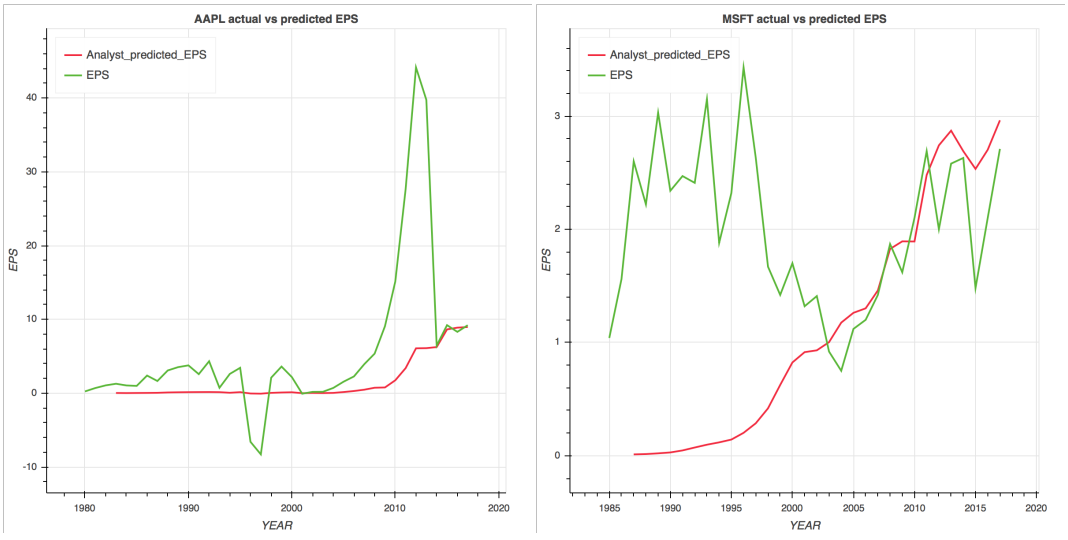

Figure 5: Red line is analyst predicted Earnings per Share (EPS), greenline is actual EPS, x-axis is year. Left is Apple, Right is Microsoft. The years of actual misstatements are: For Apple: 2001 and 2002. For Microsoft: 1995, 1996, 1997, 1998.

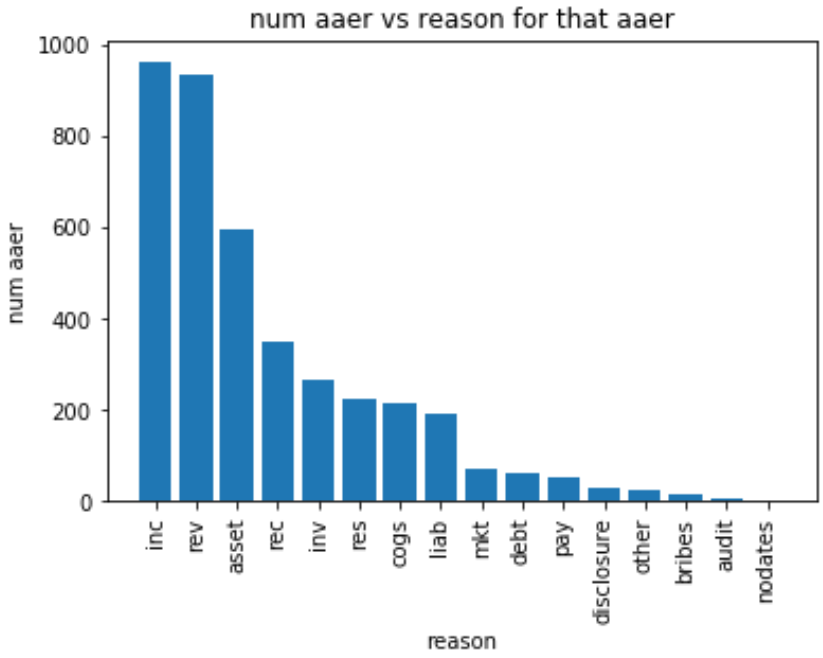

Figure 6: Number of AAER vs Reason for Misstatement

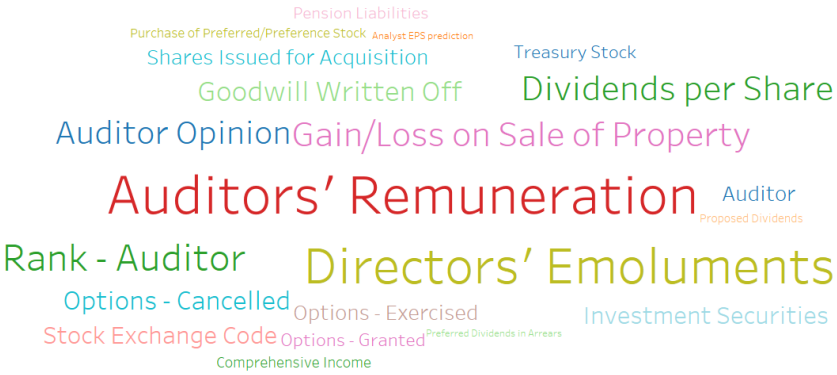

Figure 7: Word Cloud of Features with Greatest Weights in our Logistic Regression Model