Detecting Misstatements in Financial Statements

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

Vishal Shukla
School of Computing Science
Simon Fraser University
vshukla@sfu.ca
vishal_shukla@outlook.com

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

1 Motivation and Background

Financial statement manipulation is an ongoing problem in corporate America. In an industry 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 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 data product that classifies financial statements as a misstatement or not a misstatement. We wanted to answer the following questions:

- Is it possible to detect whether or not a given financial report has been misstated?
- How correlated are the fields of a financial statement?
- Which industry has the most misstatements?
- What are the most common reasons for 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?

We had the following challenges:

- Financial dataset is large with many fields and requires specialized knowledge.
- Multiple data sources
- Few examples of actual misstated financial statements.
- The majority of features contain mostly null values

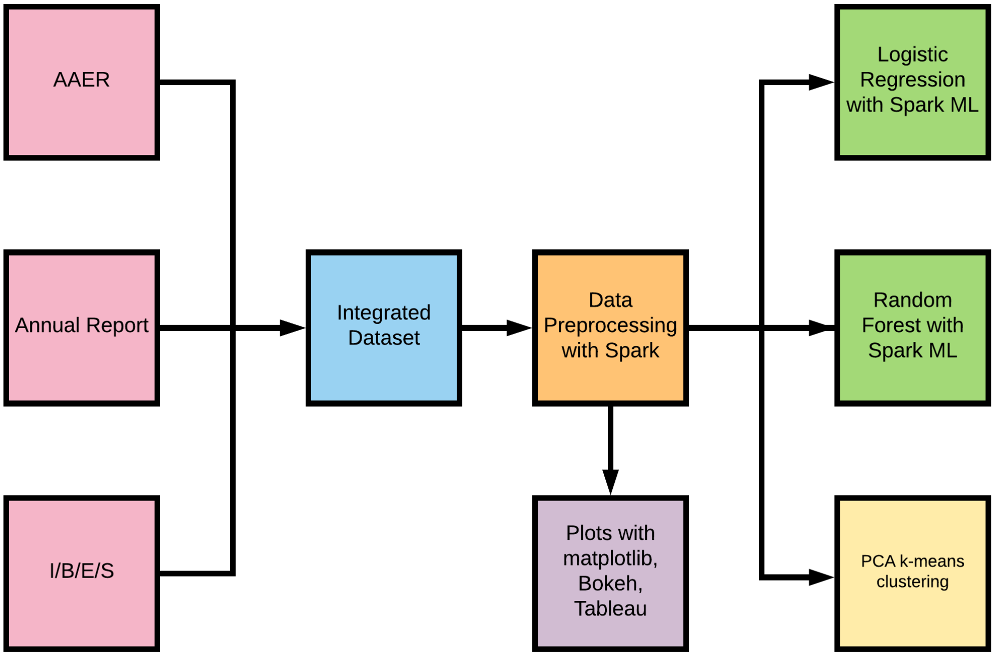
AAER
Annual Report
I/B/E/S
Integrated Dataset
Data Preprocessing with Spark
Plots with matplotlib, Bokeh, Tableau
Logistic Regression with Spark ML
Random Forest with Spark ML
PCA k-means clustering

Figure 1: The data flow of a data science pipeline in our project.

3 Data Science Pipeline

CompuStat annual report data consisting of over 500,000 financial statements of about 1000s of corporations. Accounting and Auditing Enforcement Releases (AAER) data: data representing companies found guilty of filing misstatements. We used this to prepare our training data's target variable. IBES analyst earnings per share prediction data. We performed data integration by taking the report data and performing a left join IBES prediction data on stock ticker and report year. To create the class label for misstatements, we wrote a custom user defined function for Spark and assigned a label of 1.0 to a given financial report if there was corresponding AAER record with the same stock ticker and year. We assigned a label of 0 otherwise.

To handle the class imbalance problem, we did try out different techniques such as assigning higher weights to minority class, down-sampling the majority class observations.

In terms of feature reduction technique used, we applied Principal Component Analysis (PCA) on the dataset.

4 Methodology

4.1 Preprocessing

We took the integrated dataset and imputed zeros in place of the null values. We consulted our subject matter expert and obtained a list with about 250 features that she believed to be the most important for classifying misstatements. Then we took only the columns that were in this list from our dataset. We only used the numerical variables as most of the string variables were not helpful in classifying misstatements, the non-numerical columns were mostly 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.

4.2 Exploratory Data Analysis

We initially explored the data by making several plots.

4.3 Supervised Learning

We used random forest and logistic regression. We used logistic regression model because we saw Dechow's paper [1] and wanted to see if we could replicate or exceed the performance of Dechow's logistic regression model. Then 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, and we could analyze the decision trees with the greatest weights.

We began by extracting the approximately 1500 misstatements. Due to the large class imbalance, we down sampled the non-misstatements and randomly sampled 1500 non-misstatements. Therefore, we had around 3000 total samples. We did a train test split. We used the training set to train each of our models and we used the test set for measuring our accuracy, precision and recall. Please see Figure 1 for a flow chart of how the data science pipeline and models fit together. 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.

4.4 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 4, 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 be further investigating the outliers or the unusual observations that are seen in the plot.

4.5 Tools

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

5 Evaluation and Results

We have a high accuracy and a fairly good precision and recall value for our models. Note that recall = sensitivity = true positive rate (TPR) and precision = positive predictive value

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

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

model name	random forest	logistic regression
accuracy (%)	81.818	70.503
misstatement precision (%)	78.198	64.454
misstatement recall (%)	86.627	87.226
non-misstatement precision (%)	86.013	82.022
non-misstatement recall (%)	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.

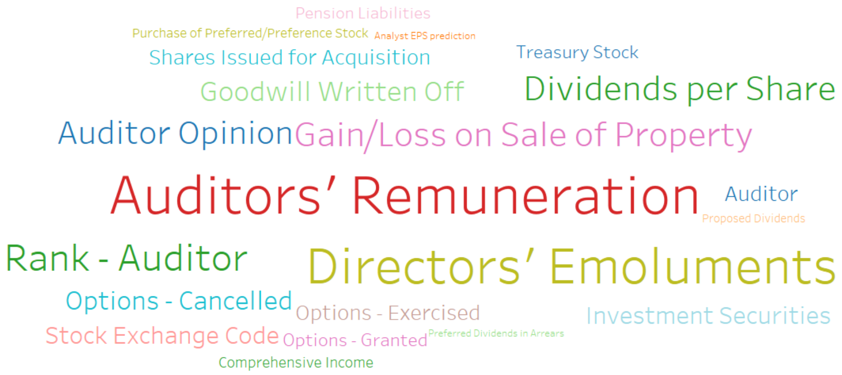

Figure 2: Word cloud of features with greatest weights in our logistic regression model

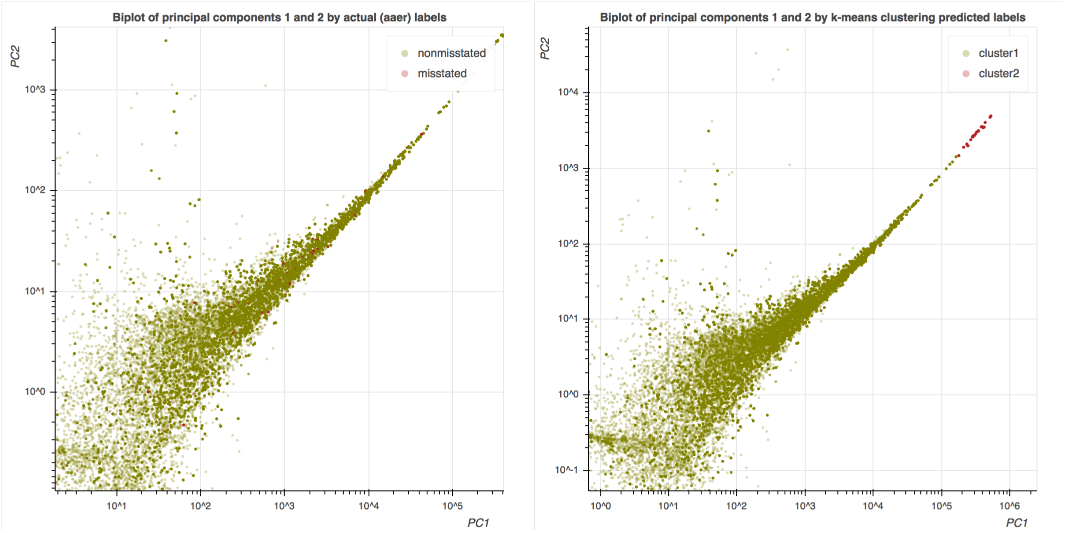

Figure 3: PCA plot with k-means clustering, however clusters did not correspond to class labels

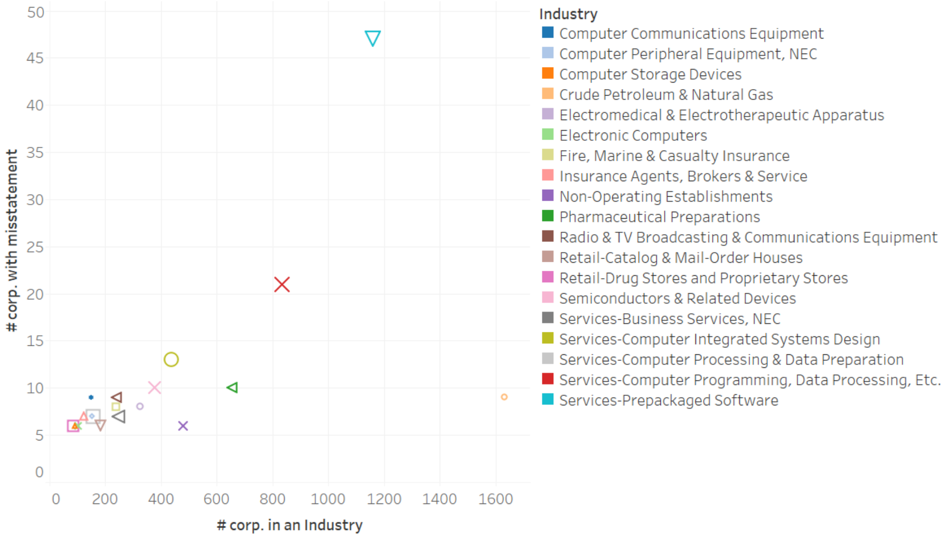

Figure 4: Number of corporations with at least one misstatement vs. total number of corporations in a given industry

7 Summary and Future Plans

- Detected misstatements correctly with accuracy of 82 % using supervised machine learning algorithms.
- Attempted using unsupervised learning algorithm to partition the statements in two clusters: misstatements and non-misstatements.
- The most common industries in which manipulations occurred are computers and computer services, retail, and general services
- The features with the most weight for logistic regression include Director's Emoluments, Auditors' remuneration, and Earnings per share.
- Conducted time series analysis of the difference between Actual EPS and Analyst predicted EPS for misstated corporations.
- Attempt to predict future misstatements

8 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.