Classifying Misstatements in Financial Statements

Vincent Chiu, Vishal Shukla, Kanika Sanduja

CS Department, Simon Fraser University

SFU BIG DATA

Motivation and Background

- Enable auditors to focus on companies that are more likely to make a misstatement.
- Investors will be able take into account the misstatement risk in their decisions.

Problem Statement

- Is it possible to detect whether or not a given financial report has been misstated?
- Which industry has the most misstatements?

Why are they challenging?

- Very large data set with hundreds of thousands of financial statements
- integrating multiple data sets
- Very few examples of actual misstated financial statements.
- having to acquire accounting domain knowledge

Data Science Pipeline

- CompuStat annual report data: 1000s of corporations, around 500,000 financial statements
- Accounting and Auditing Enforcement Releases (AAER) data: data representing companies found guilty of misstatements. We used this as our training label.
- IBES analyst earnings per share prediction data:
- Data Integration, joining tables through stock ticker and report year.

Methodology

Preprocessing: We took the integrated dataset and imputed zeros in place of the null values. We only used the numerical variables. Models:

- Random Forest: ensemble learning with multiple decision trees
- Logistic Regression: estimating probabilities using a logistic function and applying a threshold to the probability to classify.

Tools:

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

Evaluation

We have a high accuracy and a fairly good sensitivity and recall value for our models.

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

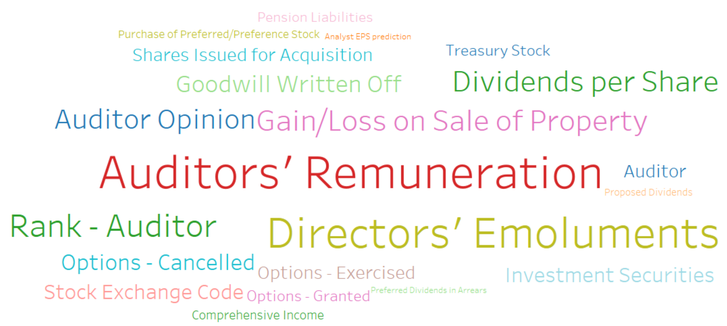

Figure 1: word cloud of features with greatest weight in our logistic regression model

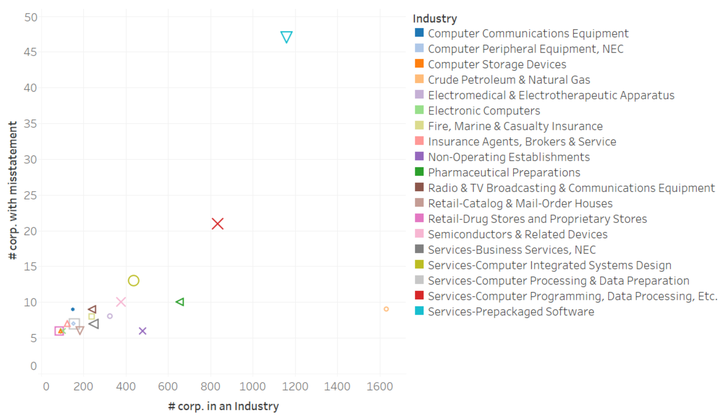

Figure 2: number of corporations with misstatements vs. total number of corporations in a given industry

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.

Lessons Learnt

- how to apply our machine learning and big data skills to financial datasets.
- Financial data is difficult to deal with. There are many fields and you need specialized knowledge.

Summary

We took annual financial report data, AAER data, analyst prediction data, and we integrated them into one parquet file. Then we applied multiple machine learning models to this integrated dataset to classify misstatements.

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.

Acknowledgements

- Kim Trottier, Associate Professor of Accounting, SFU
- Steven Bergner, Research Associate, SFU
- Jiannan Wang, Assistant Professor, SFU
- Hiral Patwa, TA
- Simranjit Singh Bhatia, TA