Detecting Misstatements in Financial Statements

Vincent Chiu, Vishal Shukla, and Kanika Sanduja

School of Computing Science, Simon Fraser University

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?

What are the challenges?

- Financial dataset is large with many fields and requires specialized knowledge.
- Multiple data sources
- Few examples of actual misstated financial statements.

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 by joining annual report data with AAER and IBES prediction data on 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 used include:

- 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.
- k-Means: using unsupervised learning to cluster observations

Tools:

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

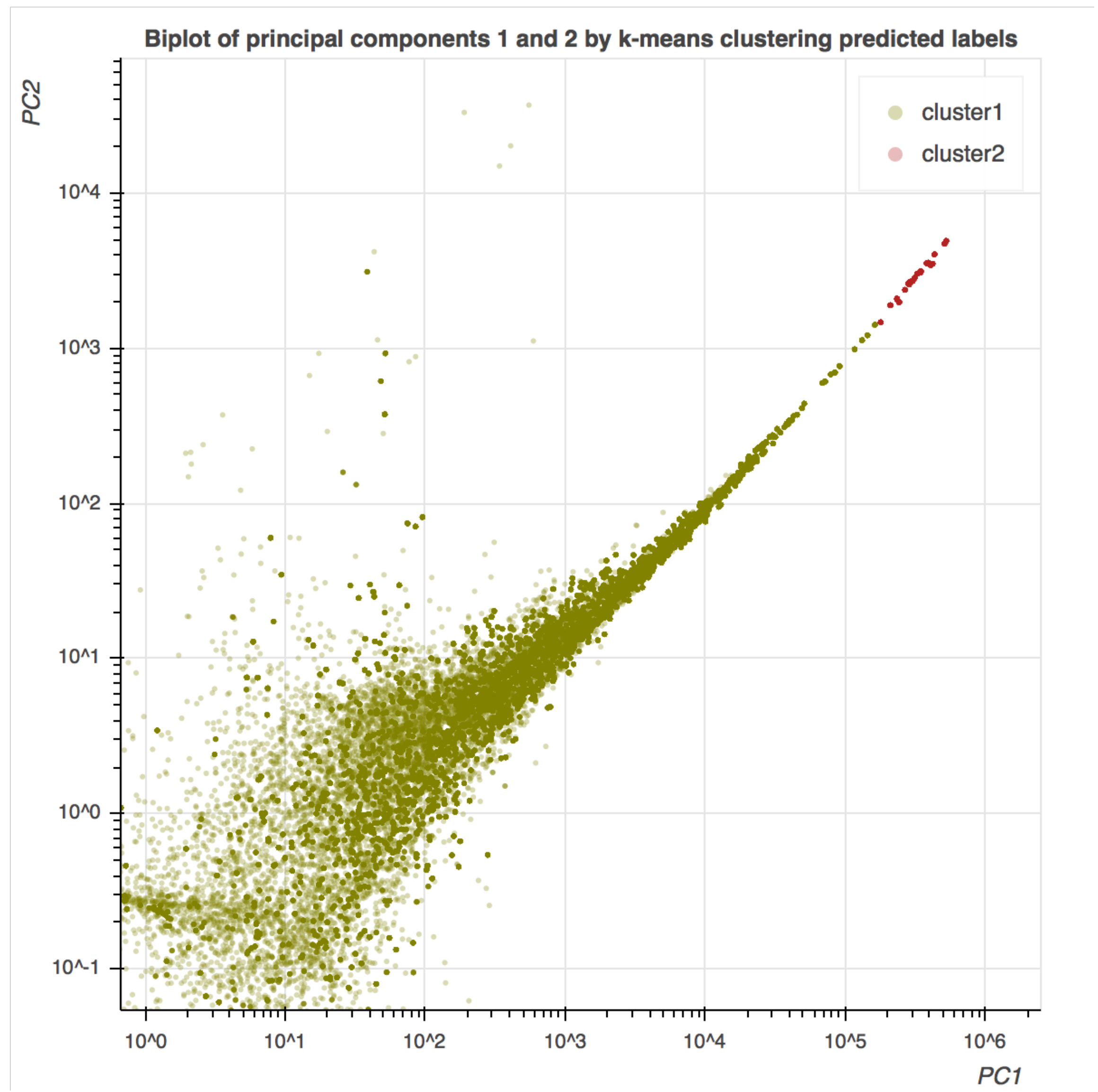

Figure 1: PCA plot with k-means clustering, however clusters did not correspond to misstatements

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

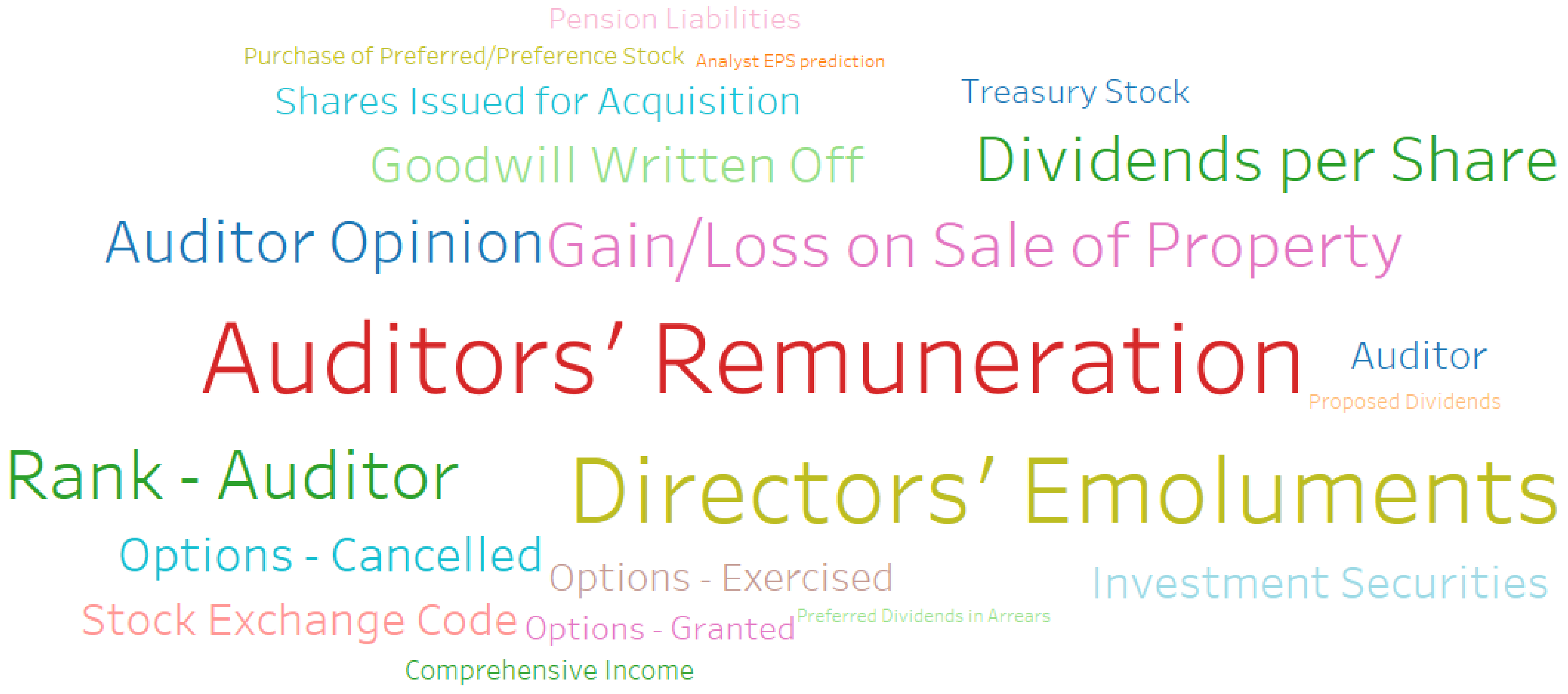

Figure 2: word cloud of features with greatest weights in our logistic regression model

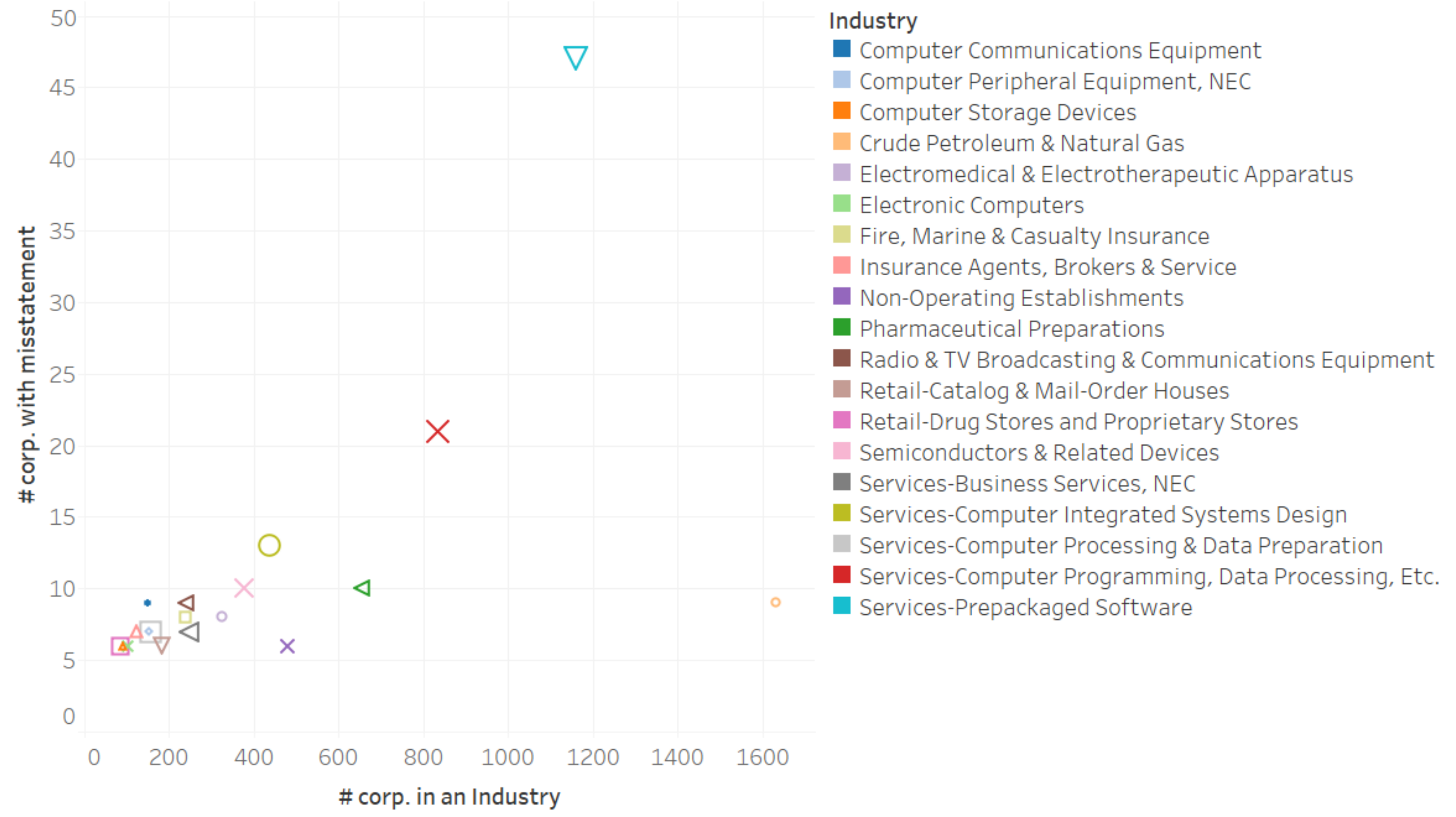

Figure 3: number of corporations with misstatements vs. total number of corporations in a given industry

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.

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 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.

References and Acknowledgements

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