Detecting Misstatements in Financial Statements

Kanika Sanduja, Vincent Chiu, Vishal Shukla

School of Computing Science, Simon Fraser University

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.

Problem Statement

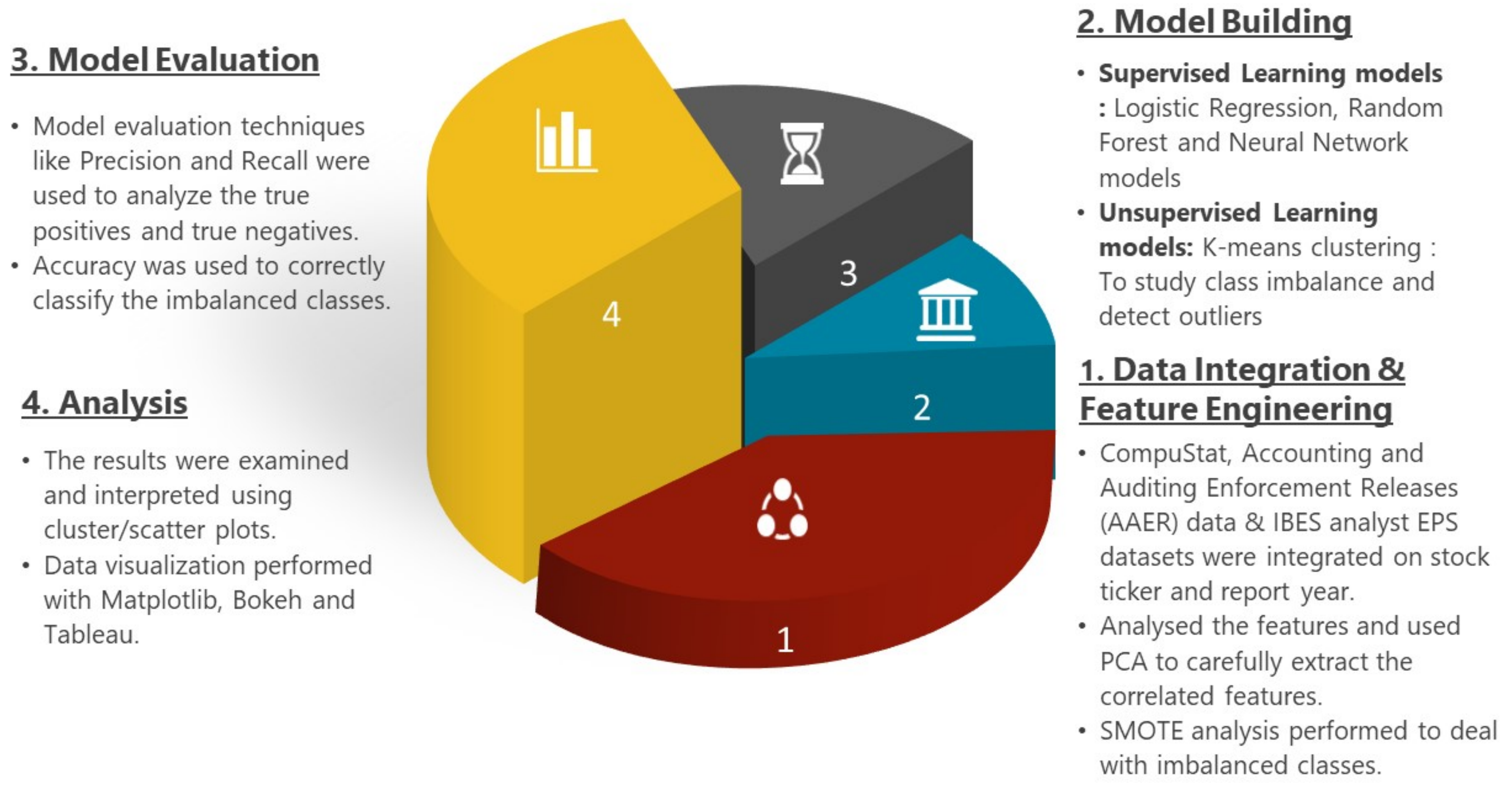
Primary Goal: build a data product that classifies financial statements as a misstatement or not a misstatement

- 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



3. Model Evaluation

- Model evaluation techniques like Precision and Recall were used to analyze the true positives and true negatives.
- Accuracy was used to correctly classify the imbalanced classes.

4. Analysis

- The results were examined and interpreted using cluster/scatter plots.
- Data visualization performed with Matplotlib, Bokeh and Tableau.

2. Model Building

- **Supervised Learning models** : Logistic Regression, Random Forest and Neural Network models
- **Unsupervised Learning models:** K-means clustering : To study class imbalance and detect outliers

1. Data Integration & Feature Engineering

- CompuStat, Accounting and Auditing Enforcement Releases (AAER) data & IBES analyst EPS datasets were integrated on stock ticker and report year.
- Analysed the features and used PCA to carefully extract the correlated features.
- SMOTE analysis performed to deal with imbalanced classes.

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

- 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.

Evaluation and Results

We have a high accuracy and a fairly good precision and recall value for our models. $\text{recall} = \text{sensitivity} = \frac{\#\text{ true positive}}{\#\text{ true positive} + \#\text{ false negative}}$

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

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

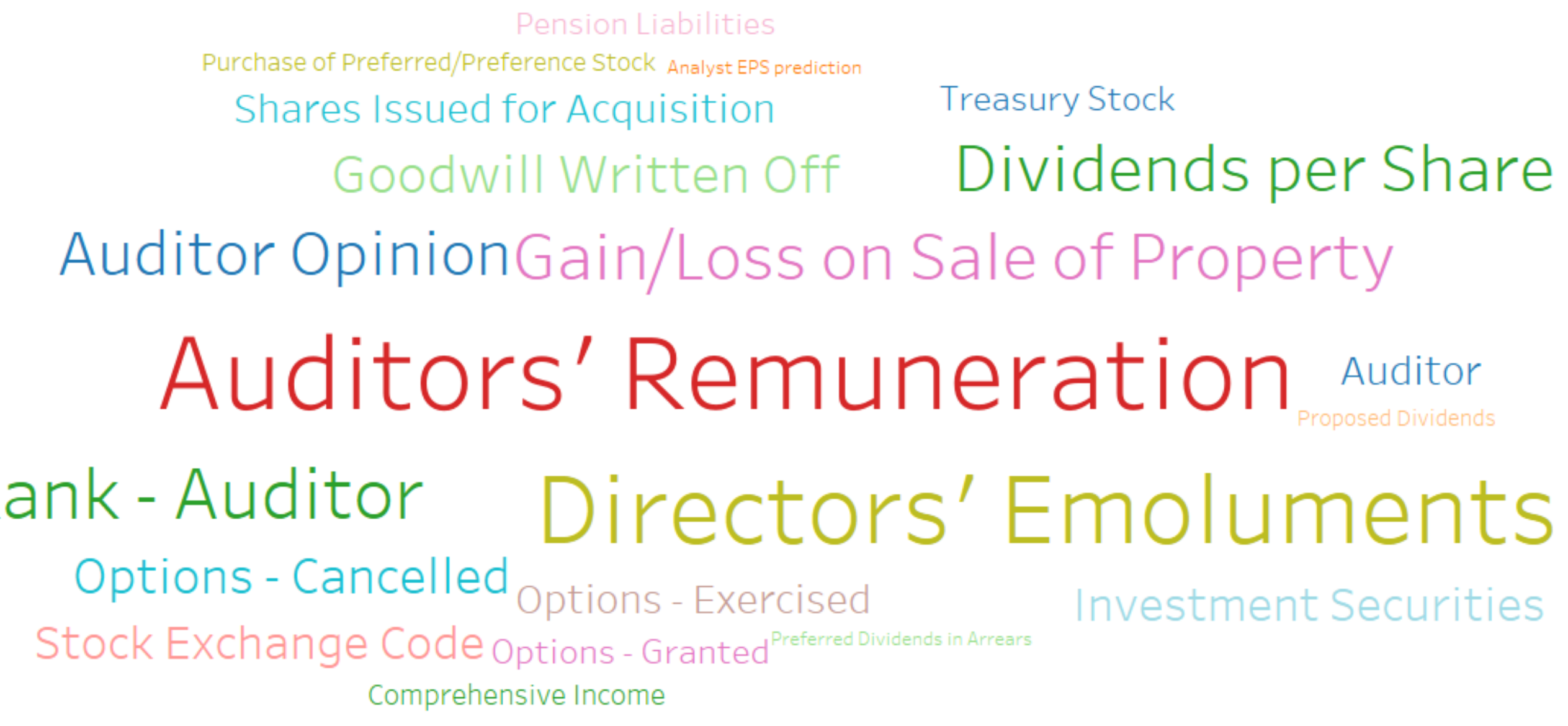


Figure 2: Word cloud of features with greatest weights in our logistic regression model

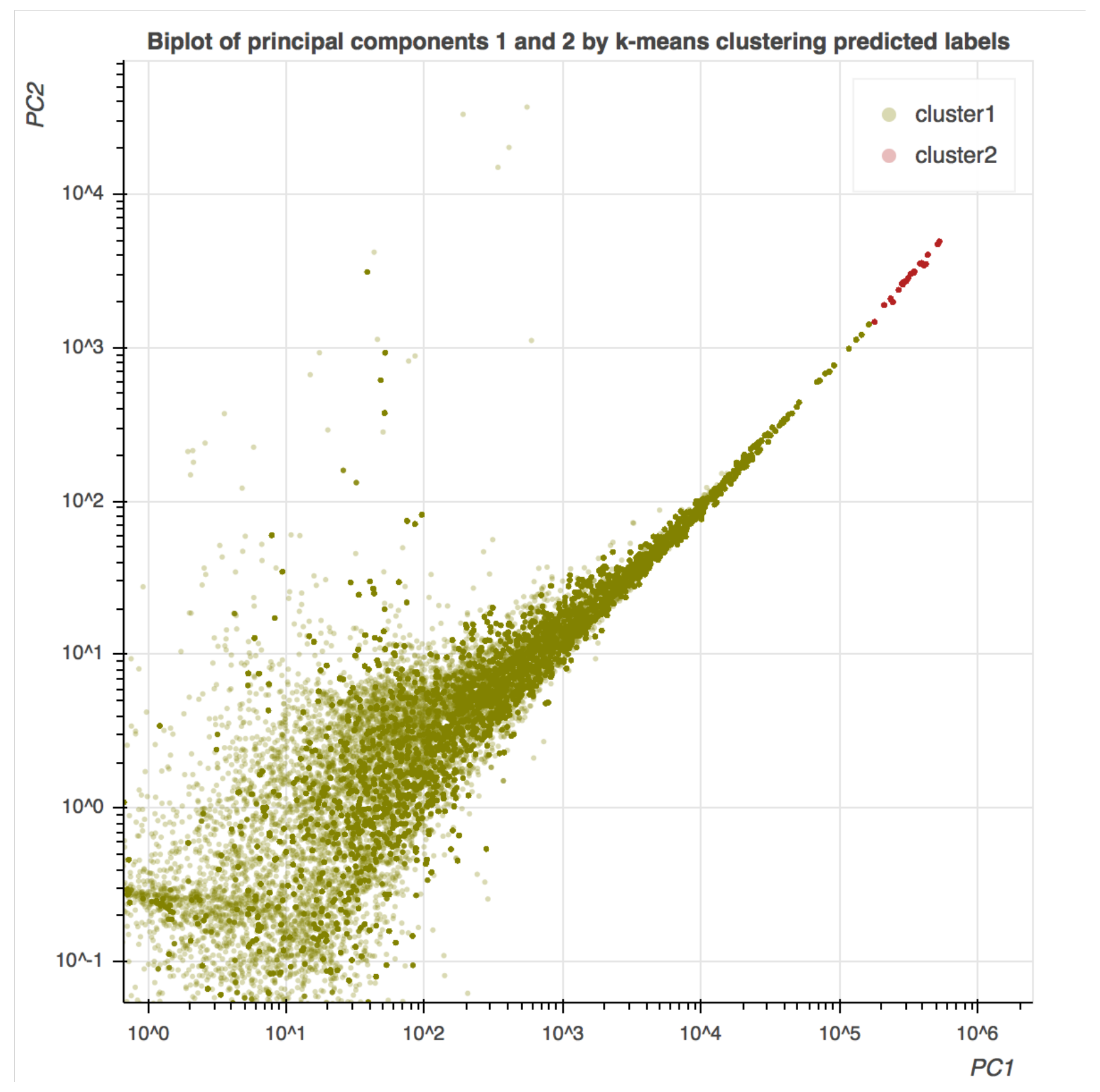


Figure 3: PCA plot with k-means clustering, however clusters did not correspond to class labels

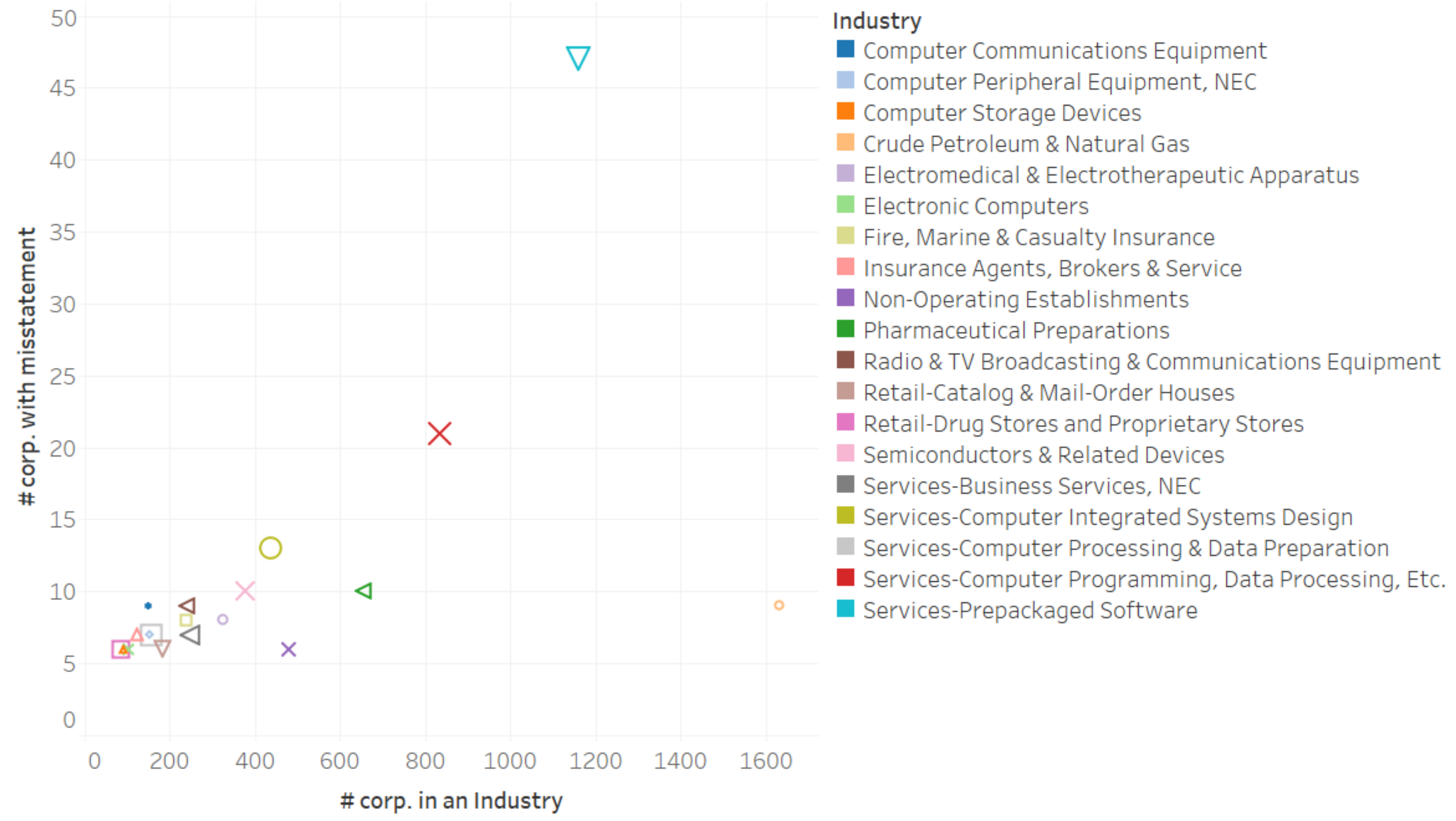


Figure 4: Number of corporations with at least one misstatement vs. total number of corporations in a given industry

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 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

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.