NC STATE UNIVERSITY

US Stock Performance Exploration and Prediction

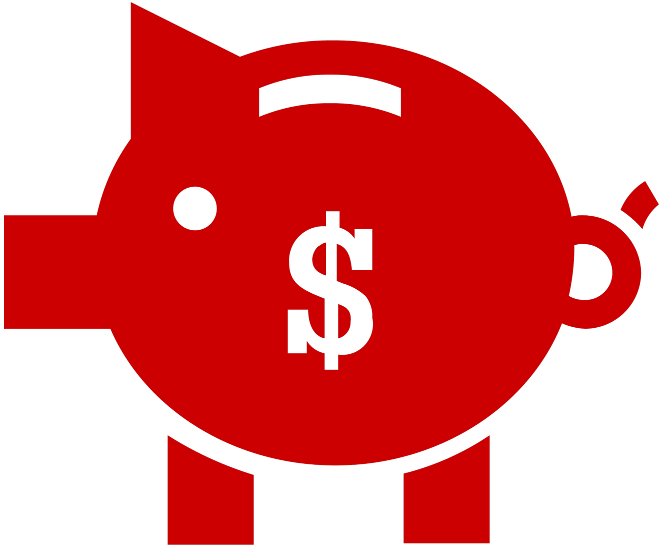
Machine Learning driven approach to understand the factors affecting US Stock Markets

By:
Yash Mishra, Anirudh Pandey,
Mithil Ghinaiya, Vraj Chokshi

1_NOV17_9_P22

Introduction

- Stock Markets are one of the most exciting and lucrative fields to study.
- We try to understand this dynamic and unpredictable chaos using machine learning models.
- We also explore if public opinion affects stocks.



Problem Statement

- Current year’s financial information is used to predict. We later explore the affect of public sentiments.

- Predicting Class has two values: 0 and 1. ‘**1**’ identifies stocks that one should BUY at the start of the year and sell at the end of the year for a profit.

- Correct prediction will help people allocate their finances better and earn more profits.

Data Set

- Kaggle-“**Financial Indicators of US stocks (2014-2018)**”
- Publicly traded company’s yearly 10-K filings
- Five CSV files with 225 columns each. Total 22,077 rows

2014: 3808x225

2015: 4120x225

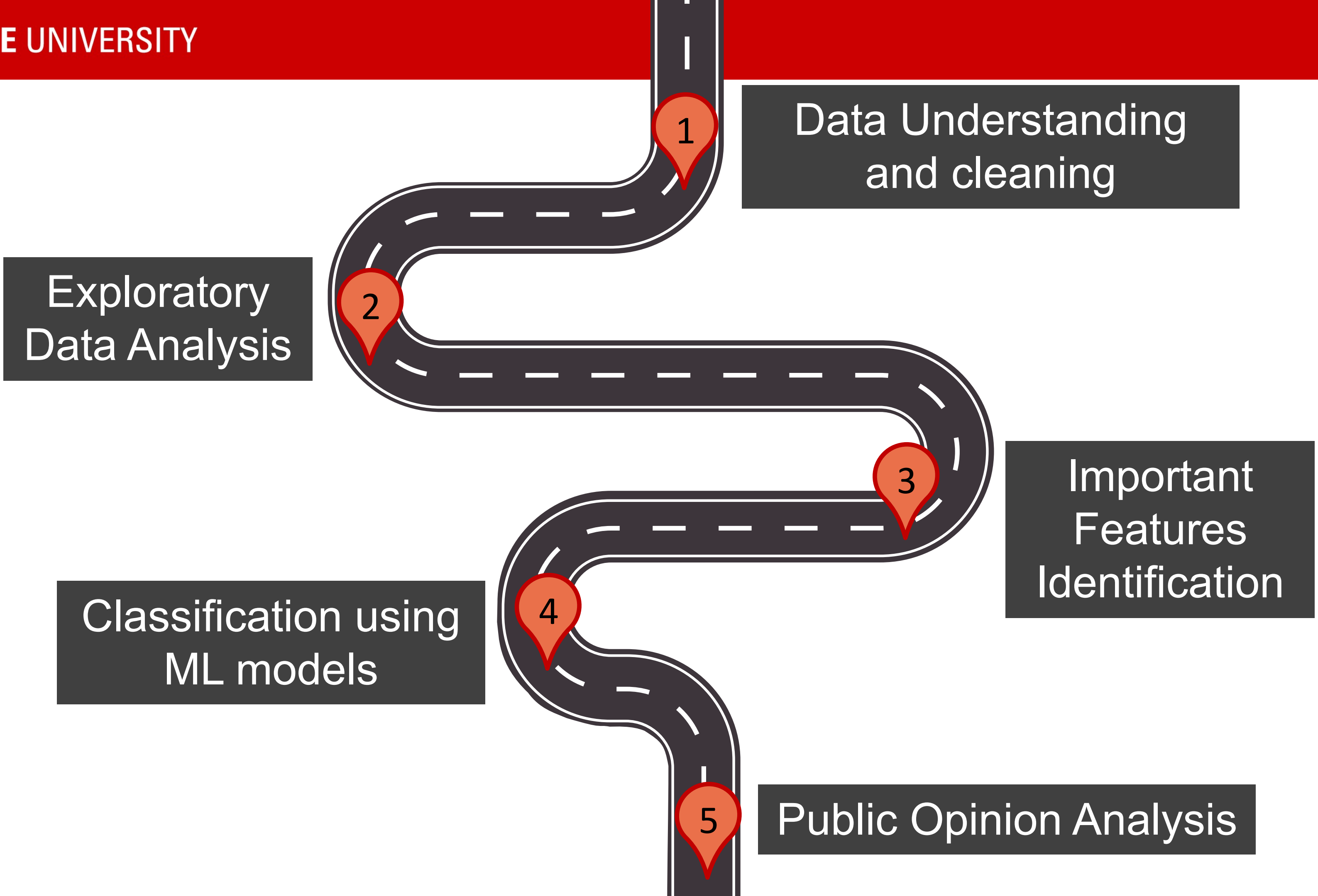
2016: 4797x225

2017: 4960x225

2018: 4392x225

Past Work

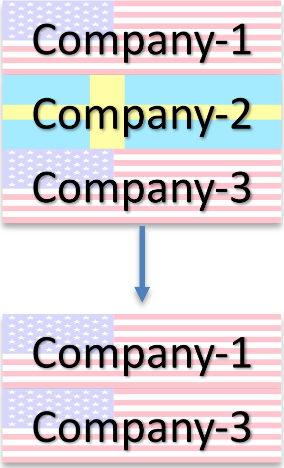
- Nguyen et al. (2015) got 54% accuracy with SVM.
- Attigeri et al. (2015) got 70% accuracy with LR.
- Dang, Duong (2016) got 73% accuracy with SVM.
- But these works focus on few companies only. While we handle 4116 unique companies and 5 years of data.



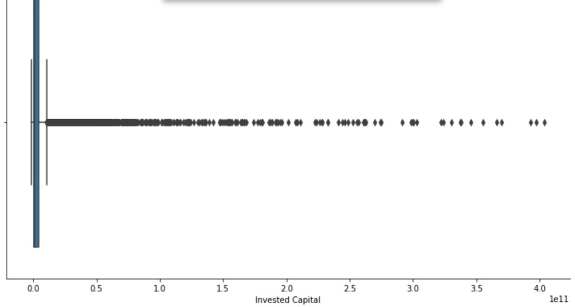
1. Data Understanding and cleaning
2. Exploratory Data Analysis
3. Important Features Identification
4. Classification using ML models
5. Public Opinion Analysis

1. Data Understanding and Cleaning

- Observed big outliers. Had to filter US companies using their ticker symbols.

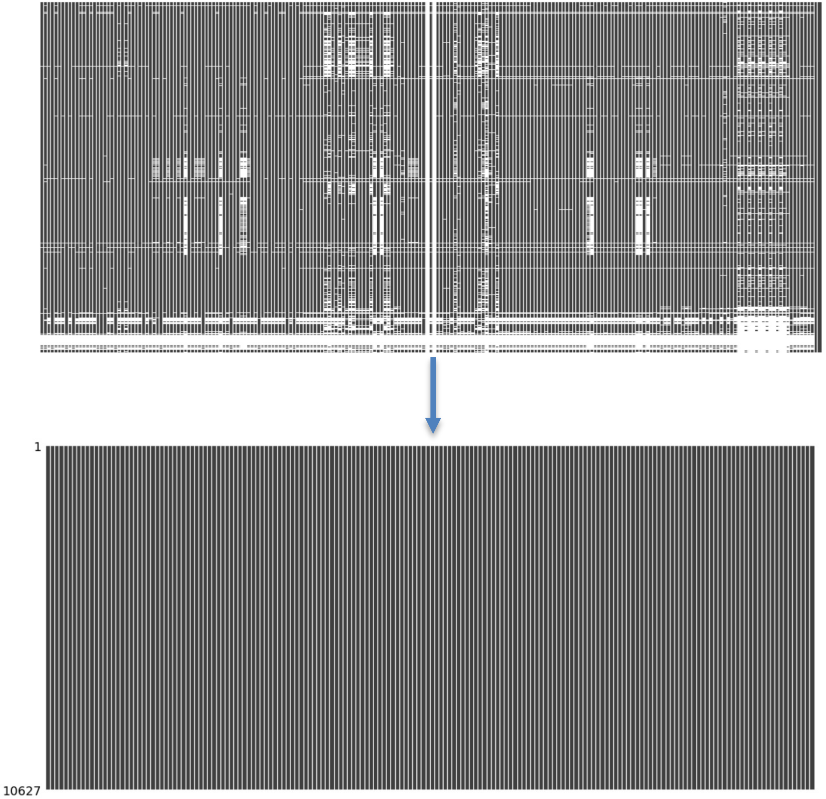


- Outlier were treated.
  - Poor Results with: Winsorize, Log Transformation, 75-25 percentile, 90-10 percentile, IQR outlier treatment.
  - Anything beyond 2.5 SD unit was discarded.



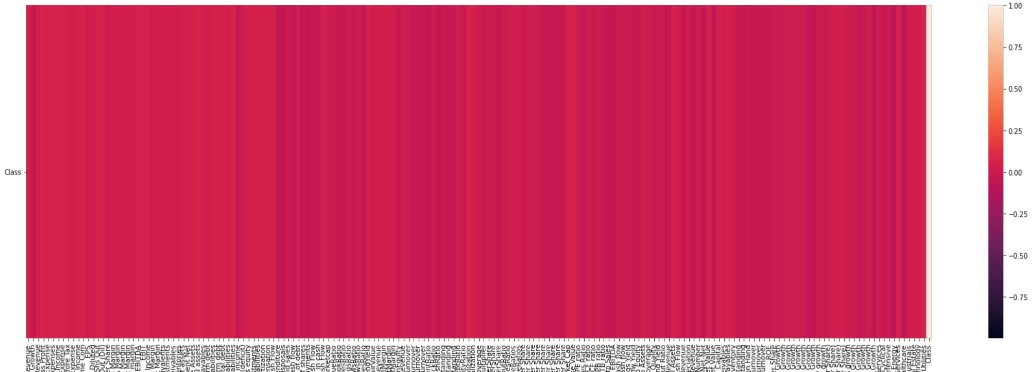
1. Data Understanding and Cleaning

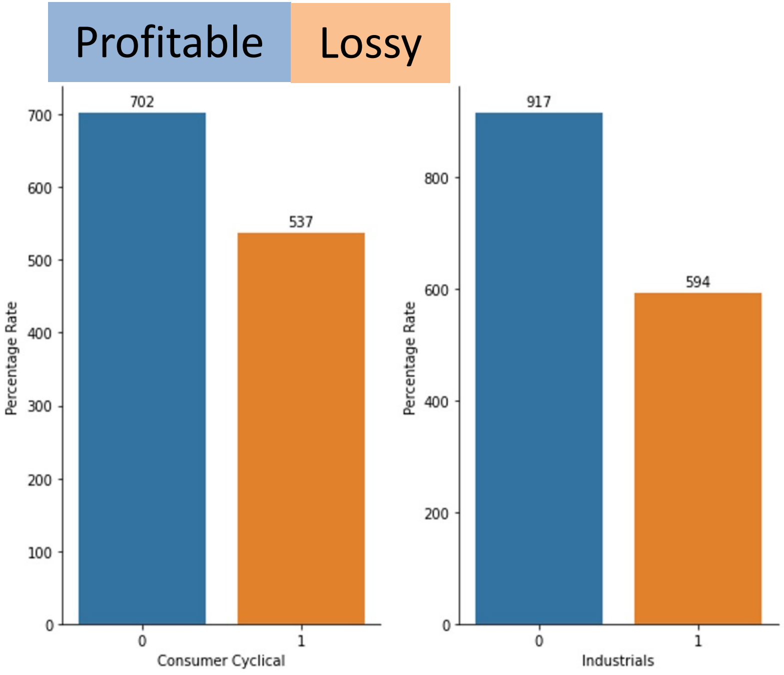
- Missing values had to be treated.
  - White space shows missing value.
- Columns with low variance were also removed



2. Exploratory Data Analysis

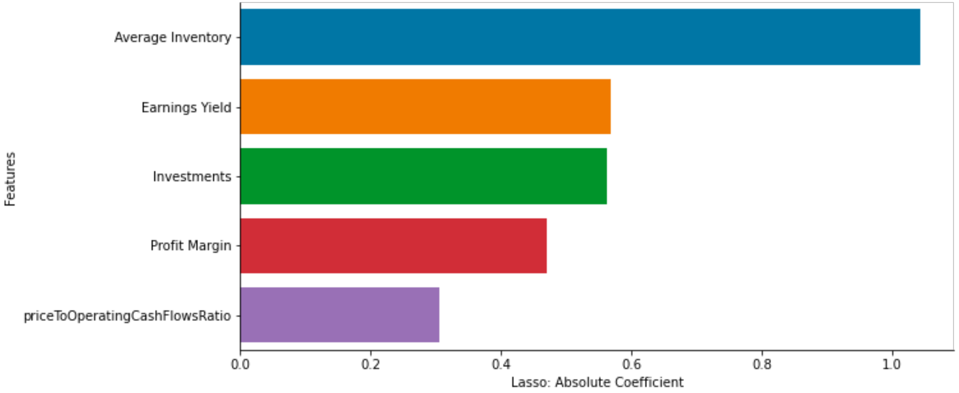
- Sector-wise analysis shows majority (59%) of the companies are profitable.
- Class variable is not strongly correlated to any columns.

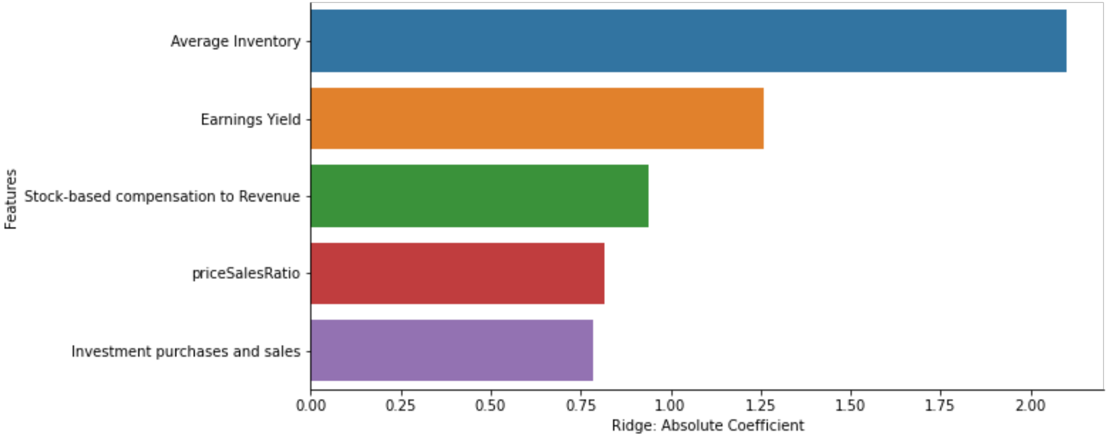




3. Important Features

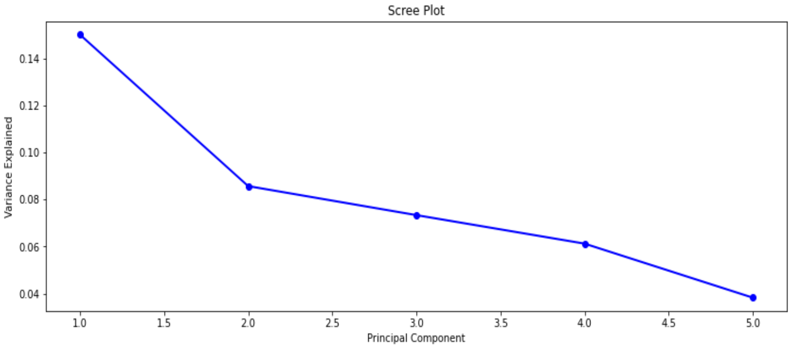
- “Average Inventory” and ”Earning Yields” identified as important by Lasso and Ridge Regression.
- Lasso and Ridge differ from the 3rd feature.





4. Data Preparation for Classification

- PCA done on normalized and standardized data.
  - Columns explain 95% variance
  - Classification done on both datasets.
- Train-Test split: 80-20
  - Validation data not used because outlier treatment removed many rows
  - 70-30, 75-25 split led to poor precision.



Variance explained after Standardization

Stand. Train Size	3581 x 69
Stand. Test Size	896 x 69
Norm. Train Size	3581 x 37
Norm. Test Size	896 x 37

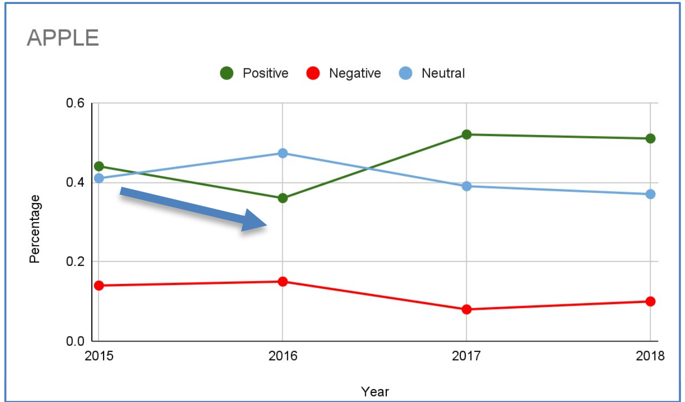
4. Classification

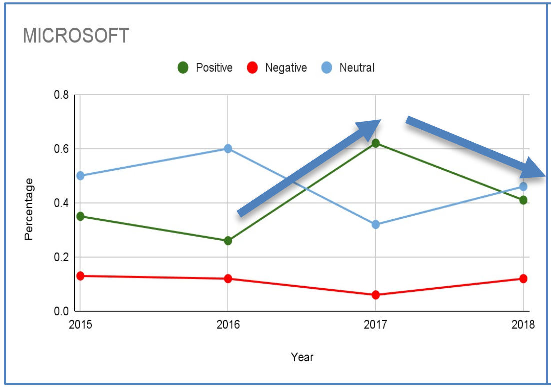
- Metric: PRECISION: TP/(TP+FP)
  - It’s okay if you don’t get RICH.
  - Losing money is not acceptable.
  - Minimize FALSE-POSITIVES
- 10-fold CV done on training data for hyper parameter tuning.
- Standardization showed better results than normalization.

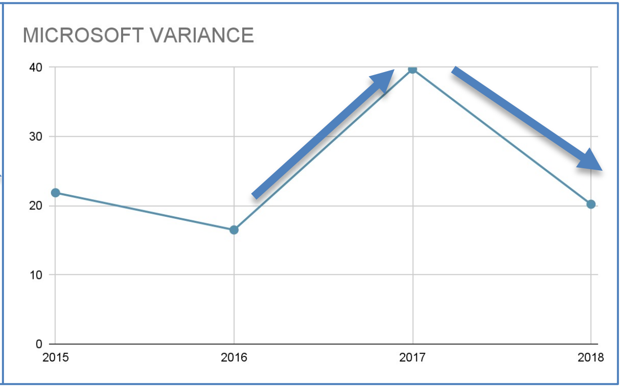
Models	Precision
Decision Tree	0.786
Random Forest	0.750
Logistic Regression	0.720
Gradient Boosting	0.665
XG-Boost	0.645
SVM (RBF-kernel)	0.626
Naïve Bayes	0.601

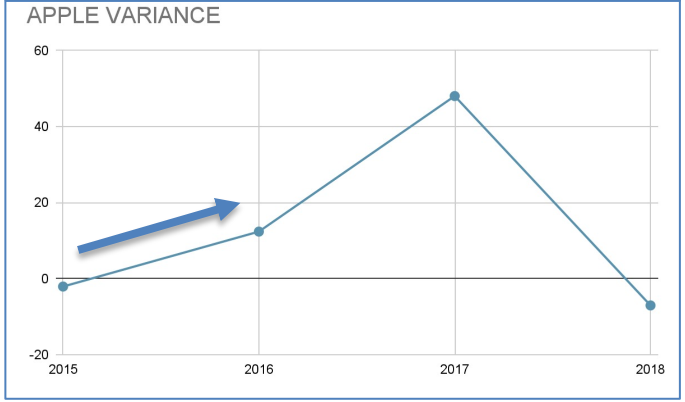
5. Public Opinion Analysis

- We scrapped Tweets from 2014 to 2018 to find relation between public opinion and company's performance in the stock market.


APPLE


MICROSOFT


MICROSOFT VARIANCE


APPLE VARIANCE

Results

- Average Inventory and Earning Yields are important indicators of company’s performance.

- Decision Tree model has the highest precision (0.786). But only 62% accuracy.

- Public Sentiment is a good but inconsistent indicator of company’s stock performance.

FINAL THOUGHTS

- There is a reason why there is only one Warren Buffet.
- No ML model can always guarantee you profit.
- One needs to be wise about his investments.