Predicting Stock Movements Using News Data

Capstone 2 Project
Springboard Data Science Career Track
By: Ivy Huong Nguyen, Ph.D.

Outlines

- Introduction, Data Description and Data Processing
- Exploratory Data Analysis
- Model Selection: Logistic Regression versus LightGBM
- Conclusions

INTRODUCTION

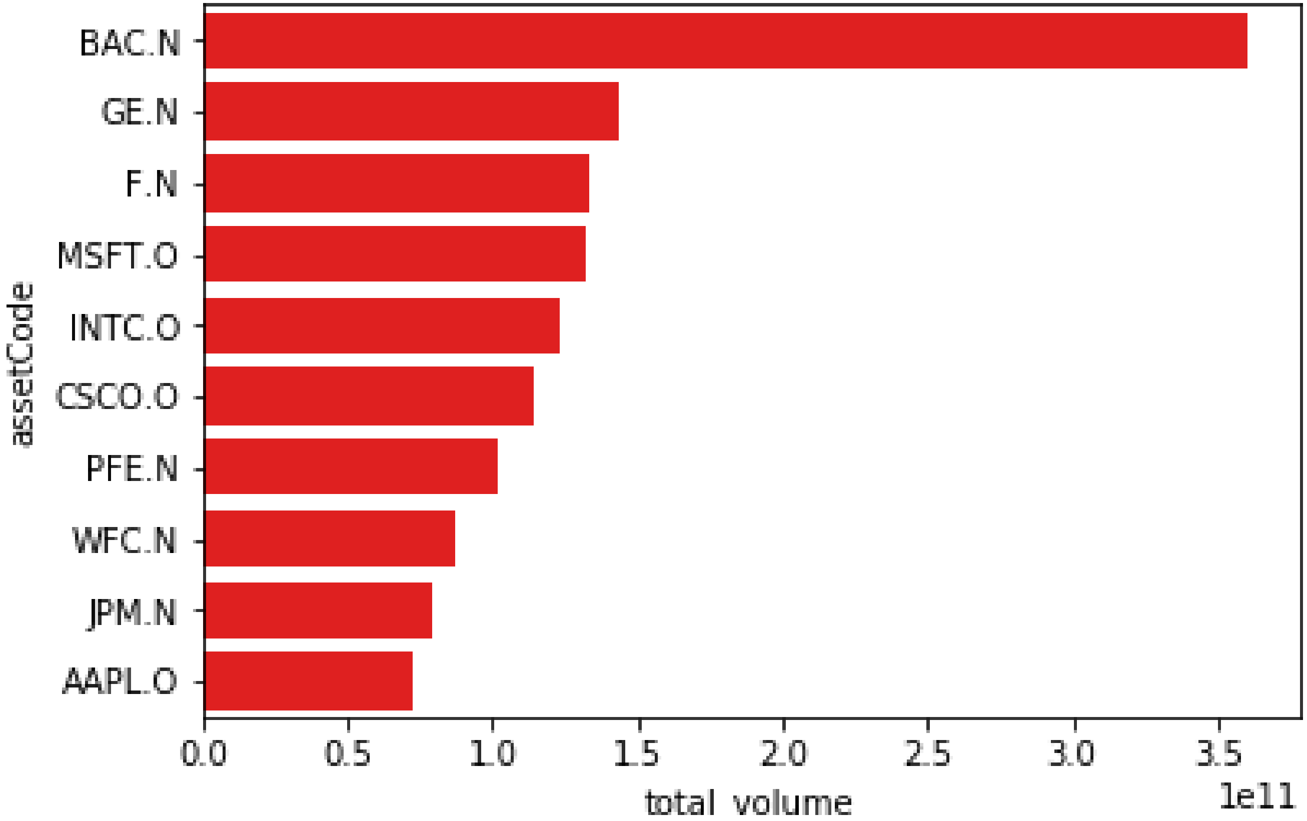
- This project is inspired by a Kaggle competition (https://www.kaggle.com/c/two-sigma-financial-news) hosted by Two Sigma, a company that is passionate about applying technology and data science to financial forecasts for over 17 years.
- There are two datasets being used in this project:
  - The news dataset
  - The market dataset
- Goal: Using features available in the news dataset to predict the target variable, returnsOpenNextMktres10 in the market dataset. Potentially we could also combine using the features available in the market dataset for the model selection as well.

EXPLORATORY DATA ANALYSIS

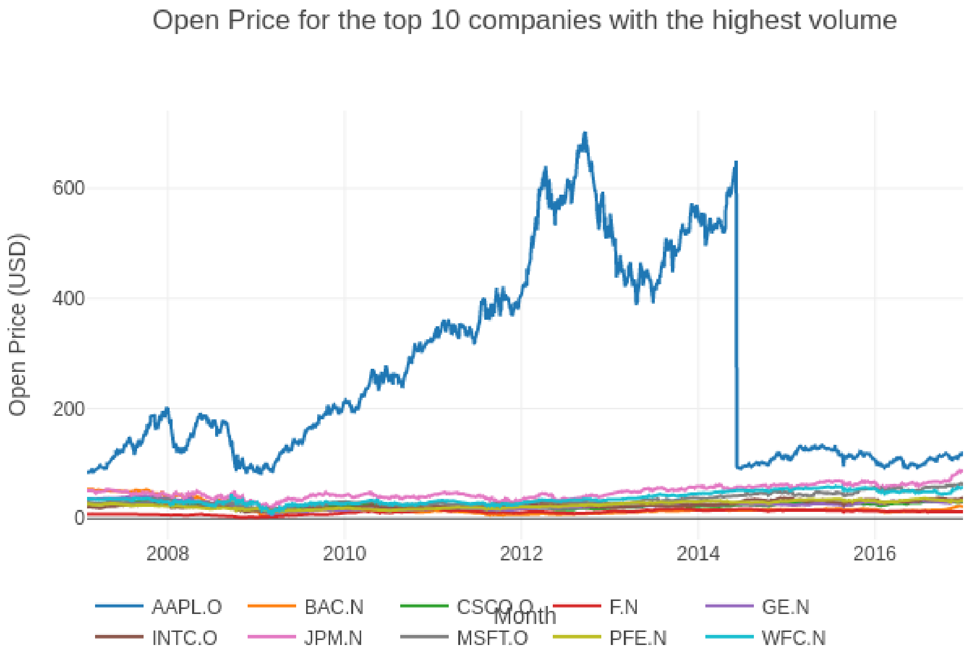
- EDA on the market dataset
- EDA on the news dataset

EDA of the Market Dataset

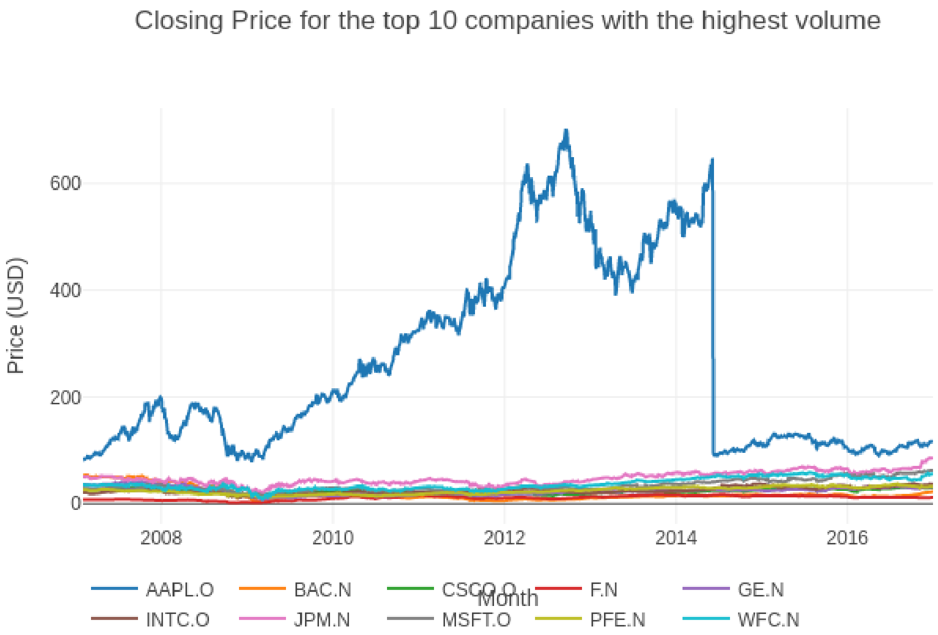
Bank of America corporation has the largest amount of trading in comparison to the others whereas Apple has the lowest amount in the list.



Close and open price of the top 10 Companies with the highest volume



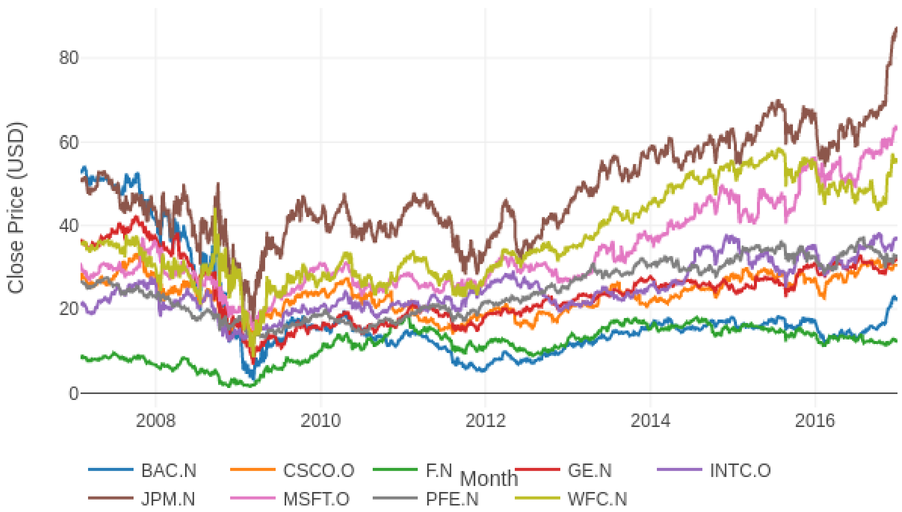
Open Price for the top 10 companies with the highest volume



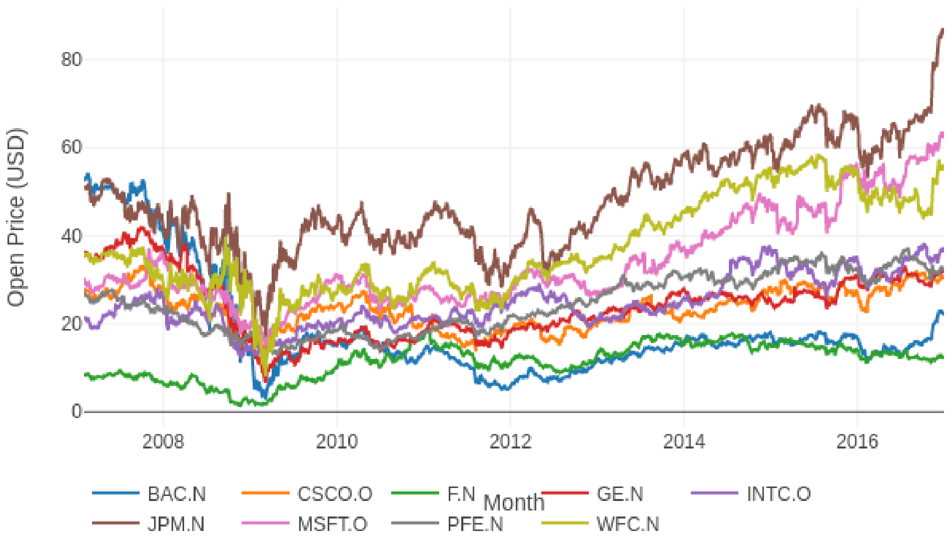
Closing Price for the top 10 companies with the highest volume

- It appears that there is a large gap in the close price and the open price of the APPL.O asset between May of 2014 and June of 2014.
- After researching, this gap was found to match with the 7-for-1 split that Apple initiated on June 9 of 2014.

Excluding APPL.O



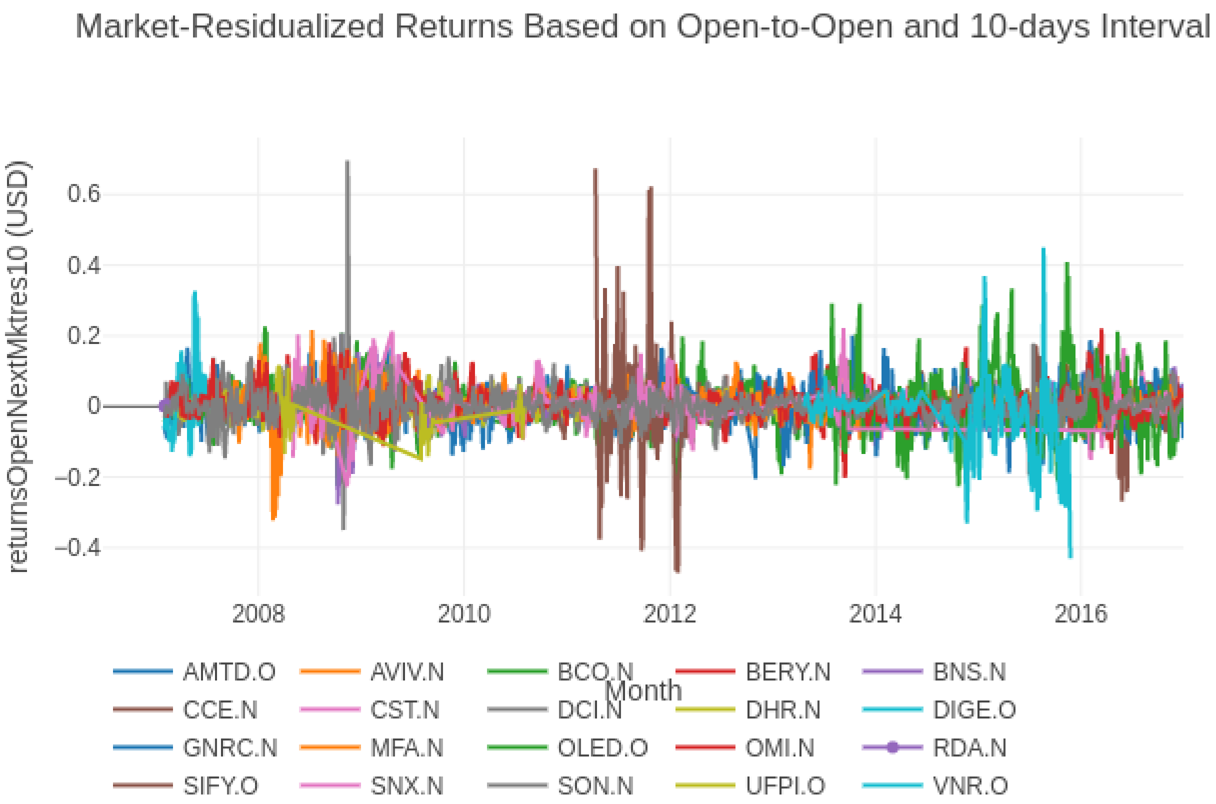
Open Price of top 10 highest volume companies without APPL.O



Closing Price of top 10 highest volume companies without APPL.O

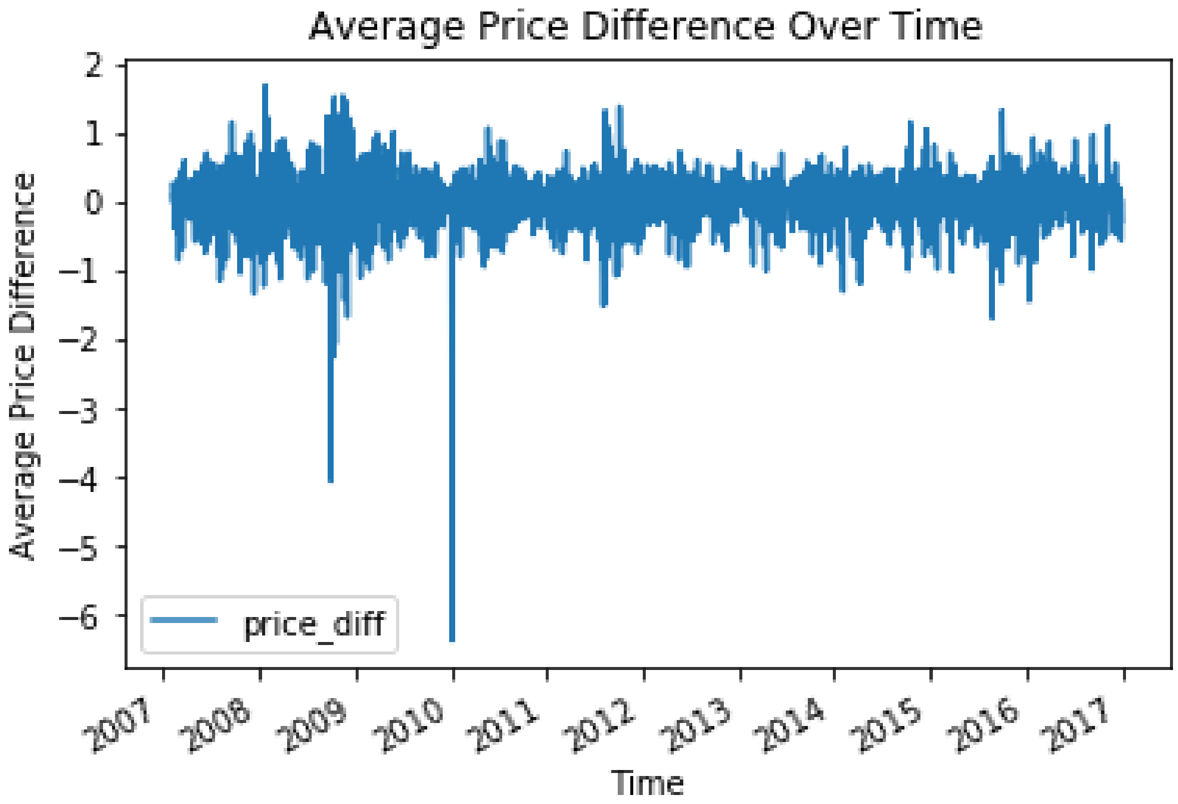
- After excluding APPL.O, we can clearly see the collapse of Lehman Brothers during the year of 2008 in both close price and open graphs.
- Another significant event that we can also point out in these graphs is towards the end of the year 2011, which marks the Black Monday of 2011 in the finance and investing industry.

Target variable close-up for 20 random companies



- The fluctuation of the market-residualized returns based on open-to-open price and within a 10-day interval of the 20 randomly selected assets is represented
- We can somewhat see the collapse of the Lehman Brothers back in 2008 and the Black Monday of 2011 → looked at the price difference from open to close

Feature engineering, data error elimination process

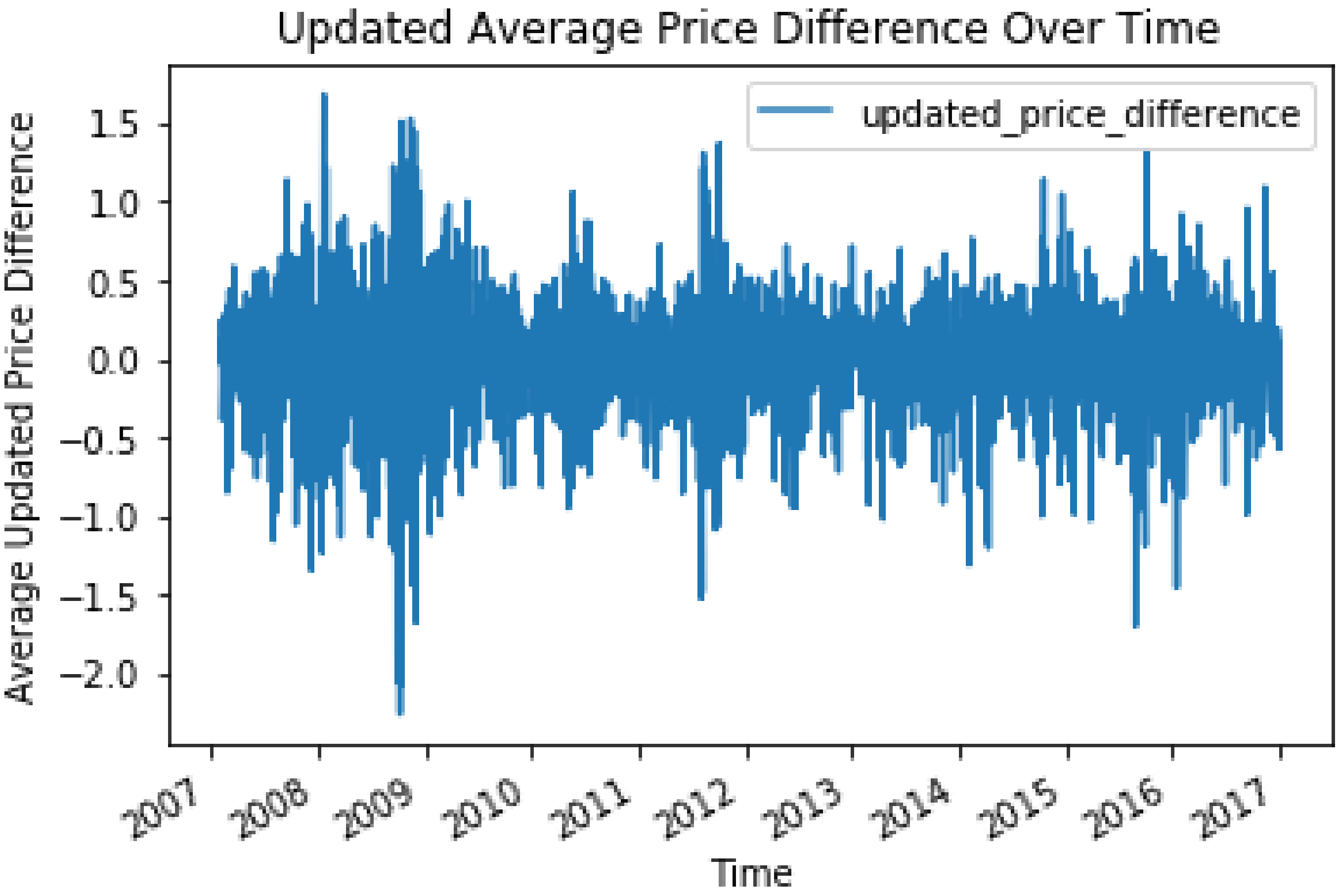


- Noticed errors in the time-frame of the significant events via price difference graph
  → Sorted the price difference column that was generated earlier by taking the difference between the open and the close price.
- noticed that the Towers Watson and Co has a huge price difference (almost $10,000 in difference) on the 4th of Jan in 2010

Fixed the errors by:

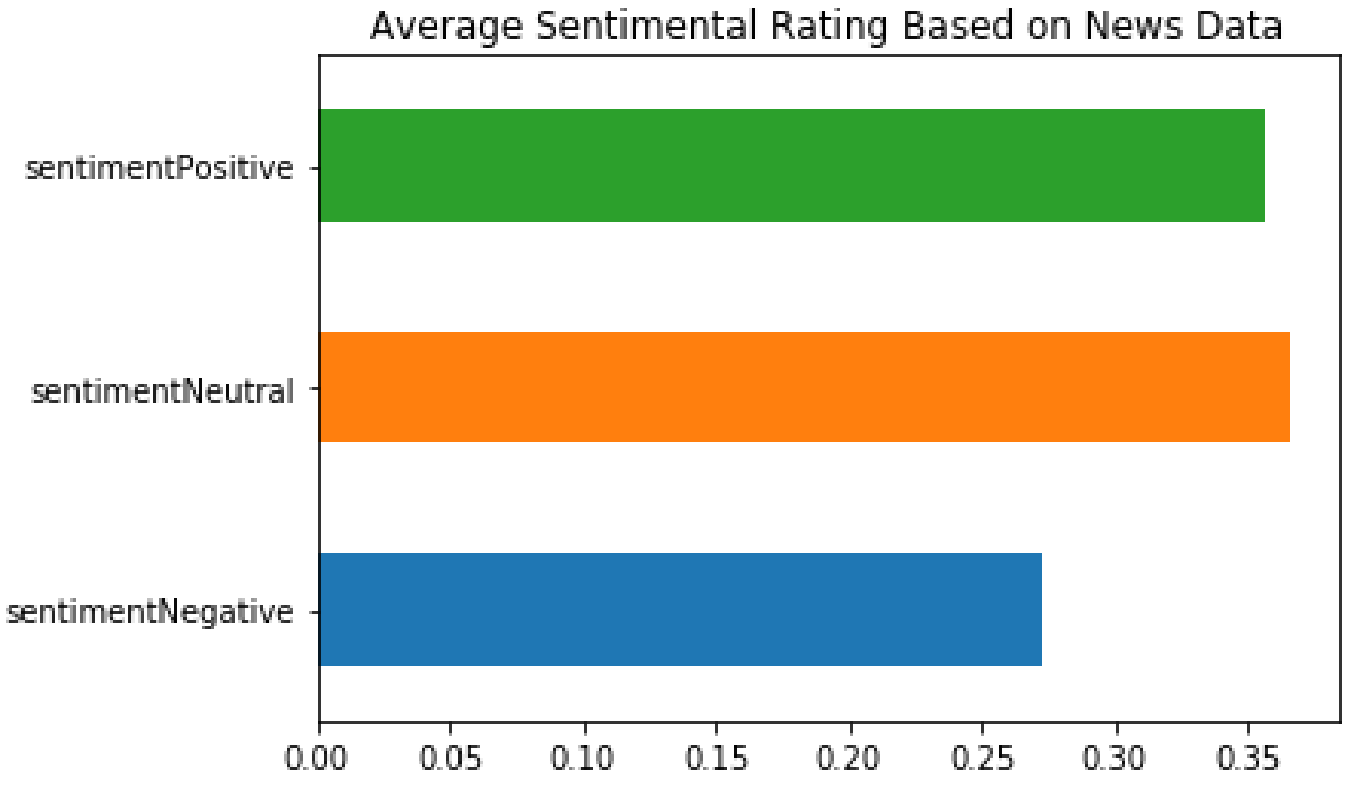
- Took the ratio between the close stock price and the open stock price of the same day for each asset.
- If the close price increases or decreases more than twice the open price then we classify that data point as an error and replace it with the median of the open price or the median of the close price.

Updated average price difference Over Time of the randomly-selected 20 companies



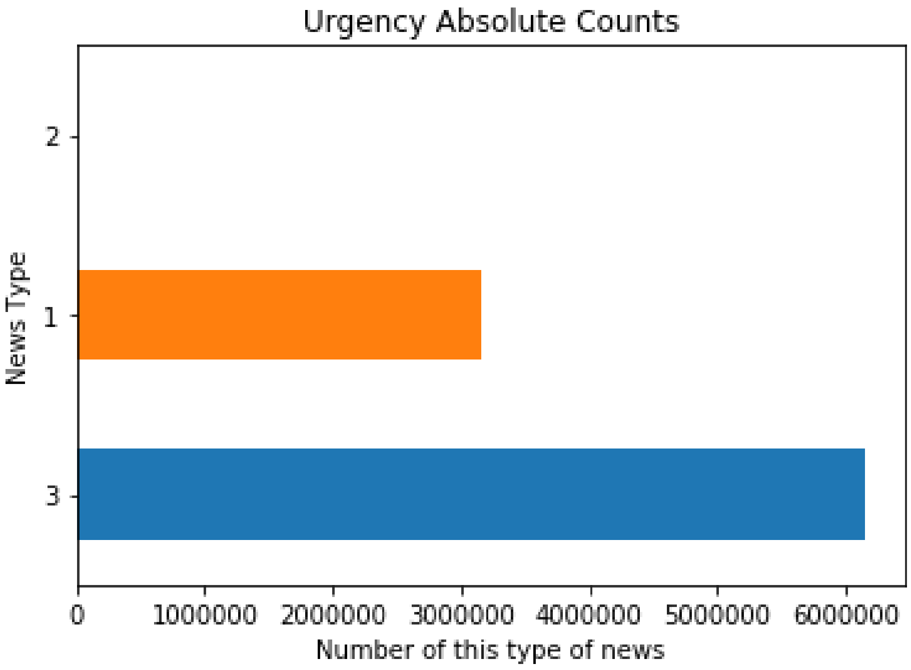
EDA of the News Dataset

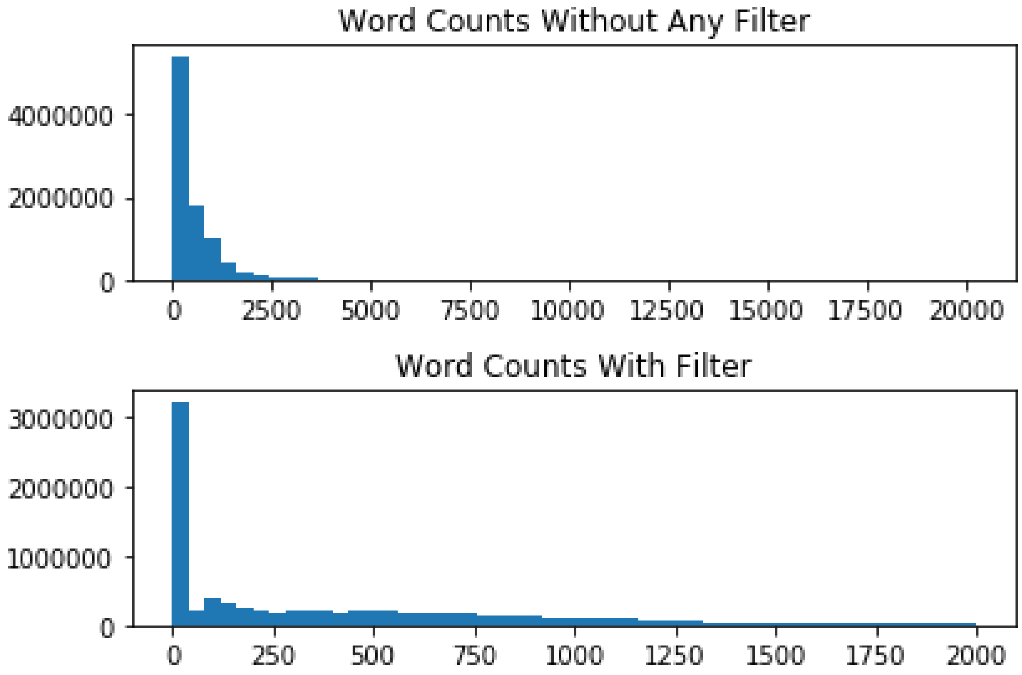
a) Sentimental Rating



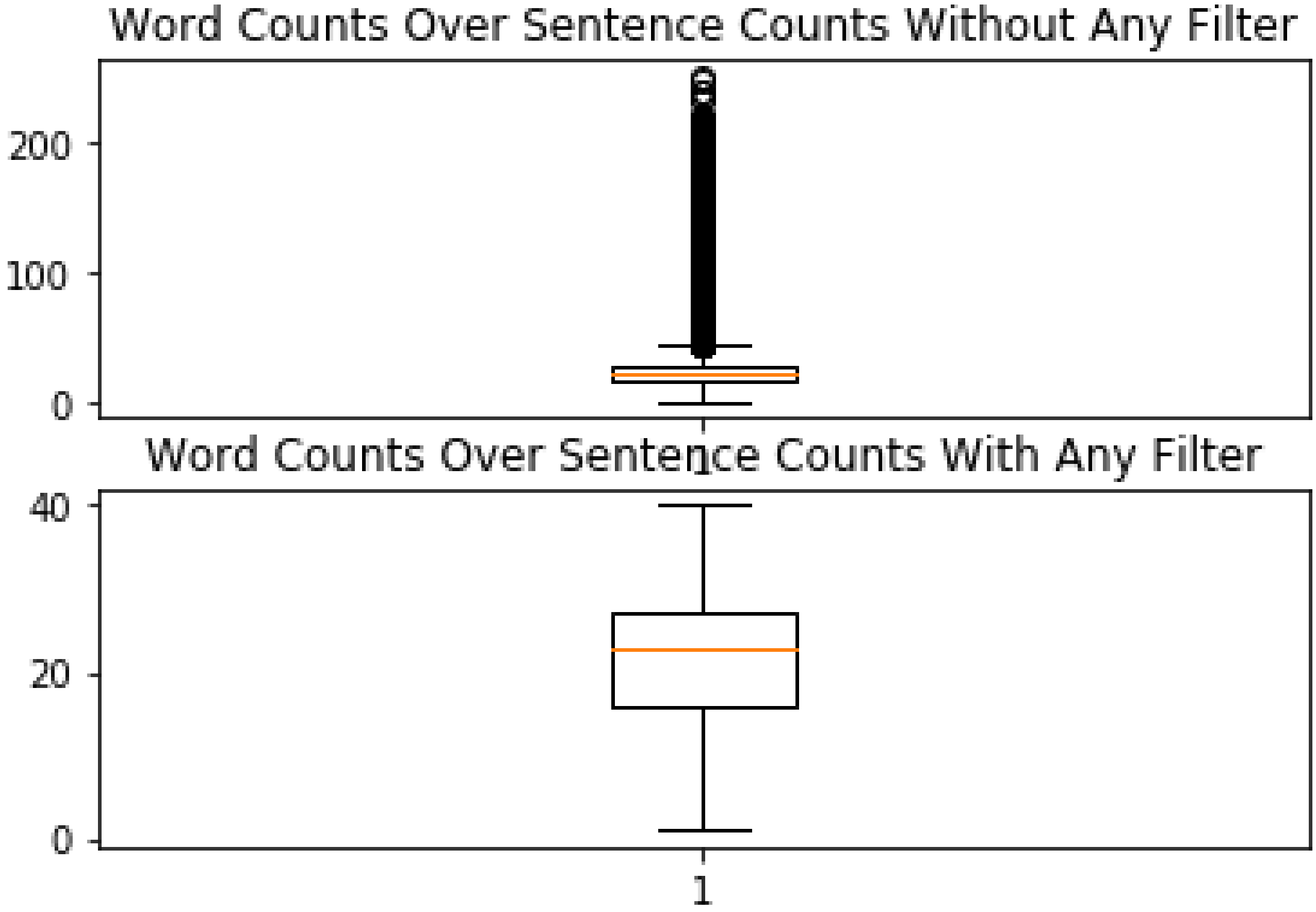
EDA of the News Dataset (cont.)

b) Urgency and word counts



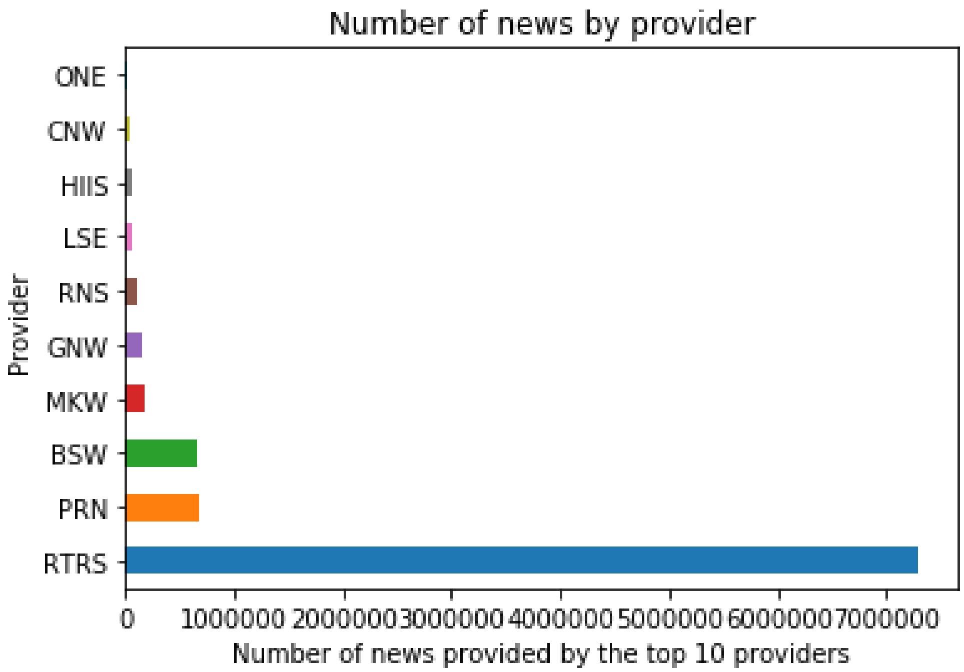


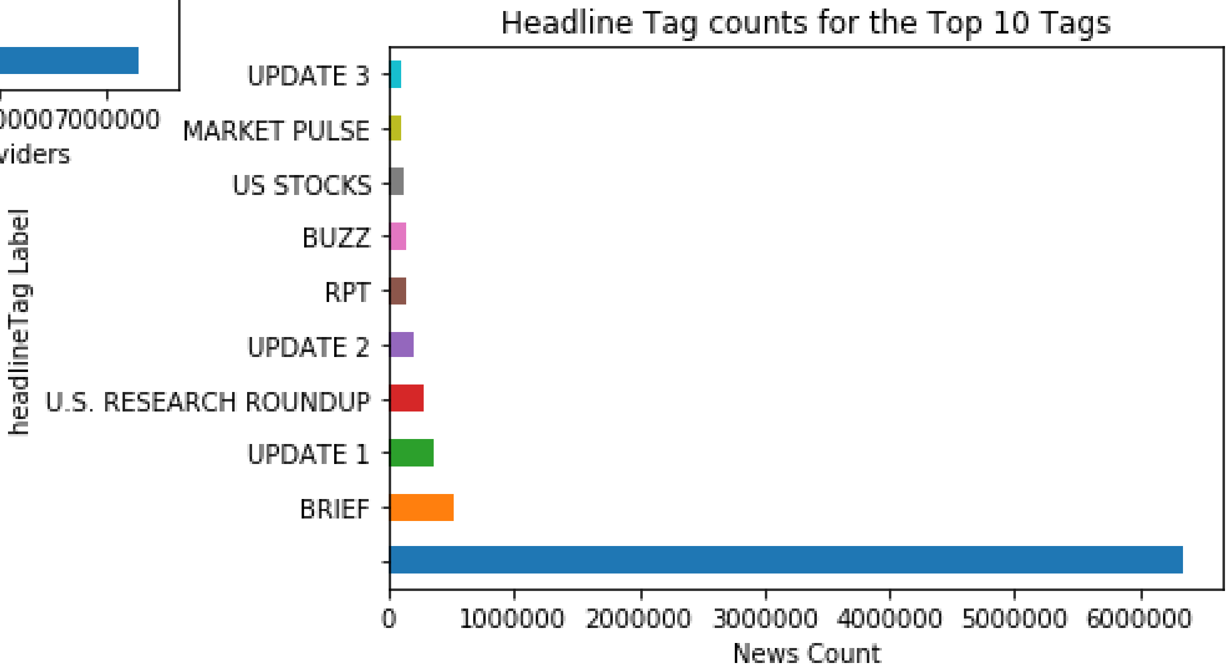
EDA of the News Dataset (cont.)



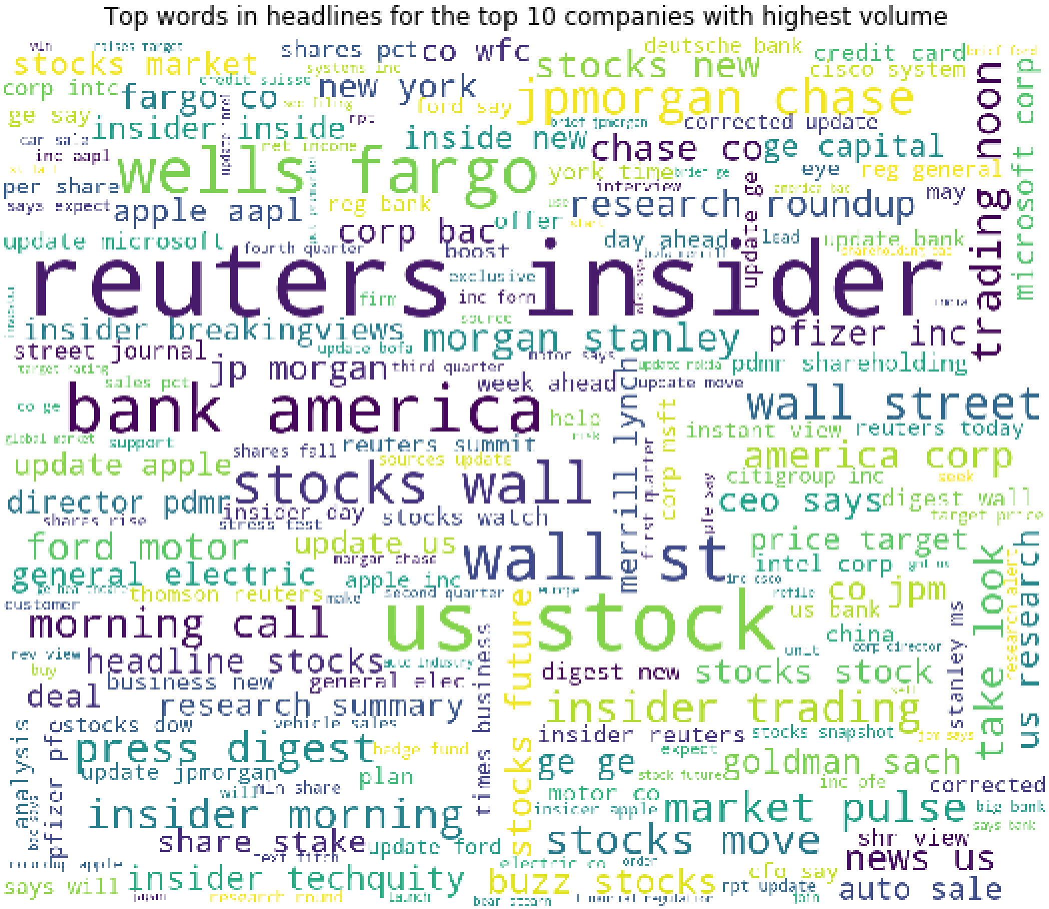
EDA of the News Dataset (cont.)

c) Providers and headline tags





EDA of the News Dataset (cont.)



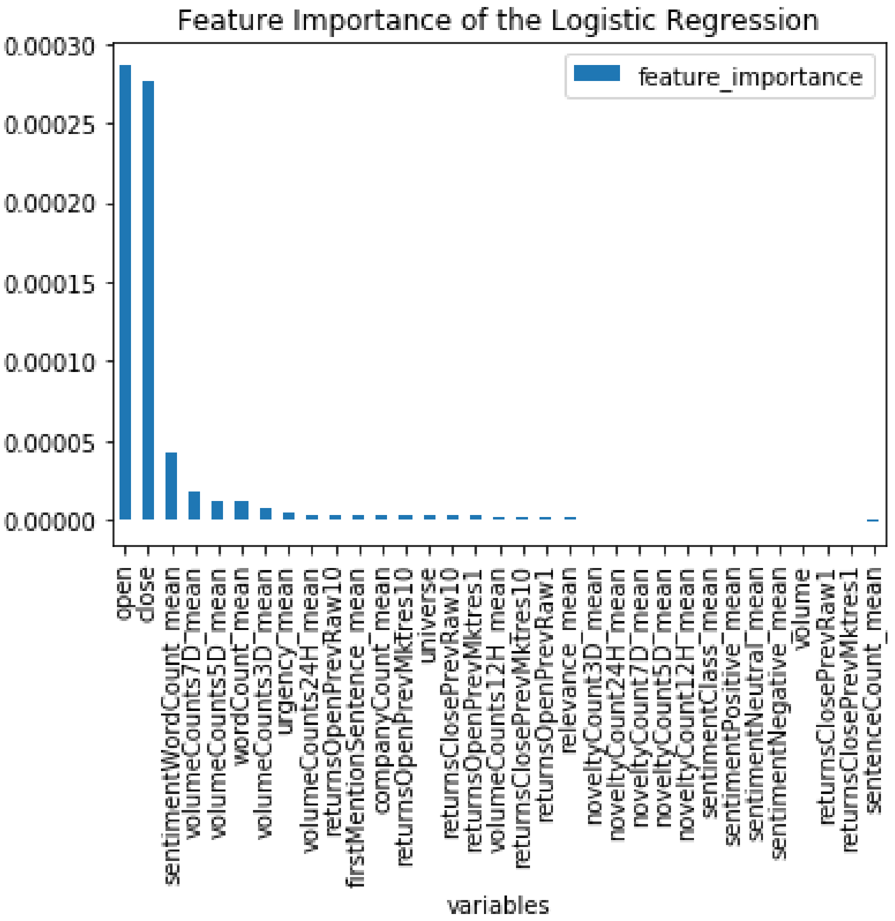
Top words in headlines for the top 10 companies with highest volume

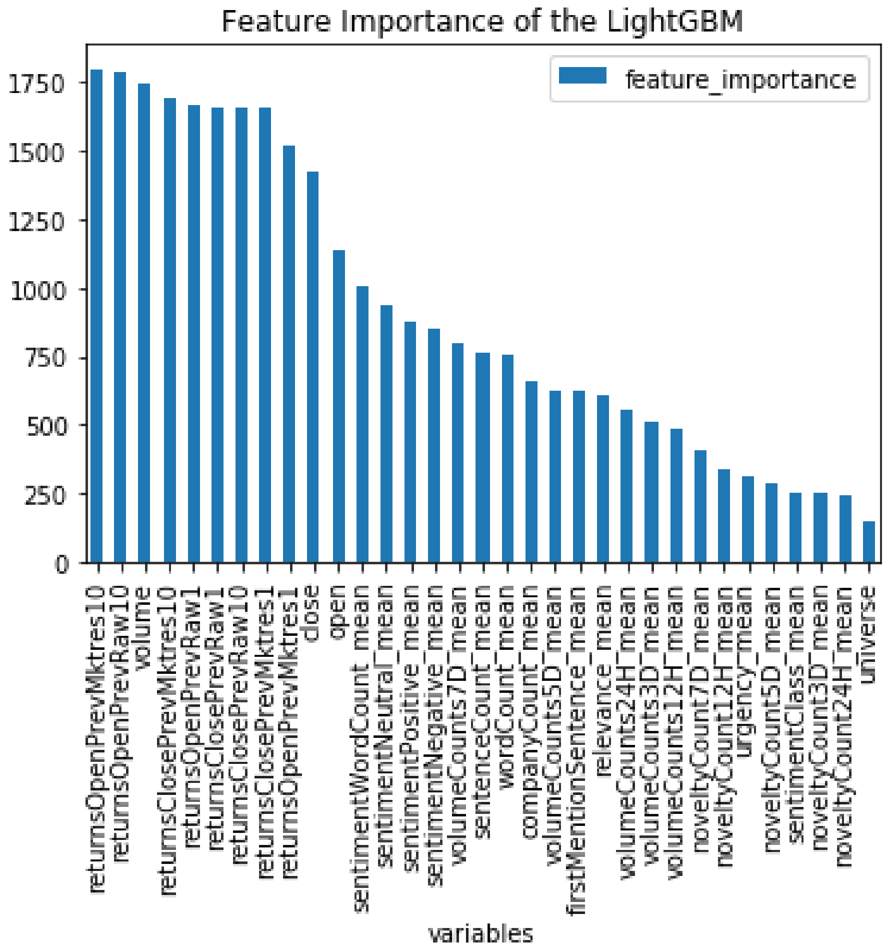
WordCloud generated image shows the impact of each word presented in the headlines of the top 10 companies that have highest trading volume

MODEL SELECTION

- Comparing Model Performance
  - Logistic regression model is the first choice when it comes to a binary classification problem → predict probability between 0 and 1
  - Second choice is the lightGBM. LightGBM model is a relatively new but efficient model that is based on tree-based learning algorithm
    - LightGBM is preferred as ‘light’ because of its high speed and its ability to handle large size dataset

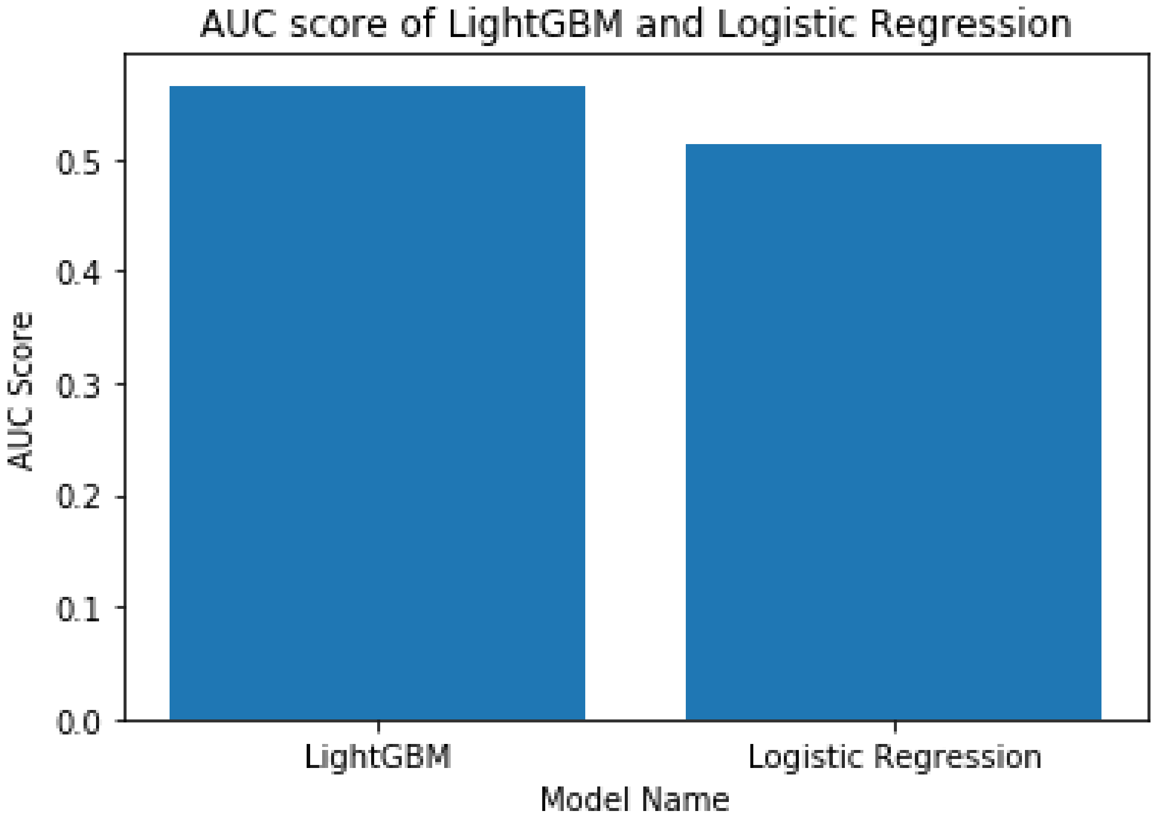
Feature Importance





- The top three features that have the highest contribution to the logistic regression model are ‘open’, ‘close’, and ‘ sentimentWordCount-mean’.
- The lightGBM model has a totally different distribution of the feature importance, in which the top three most important features are: returnsOpenPrevMktres10, returnsOpenPrevRaw10, and volume.

AUC Score Comparison

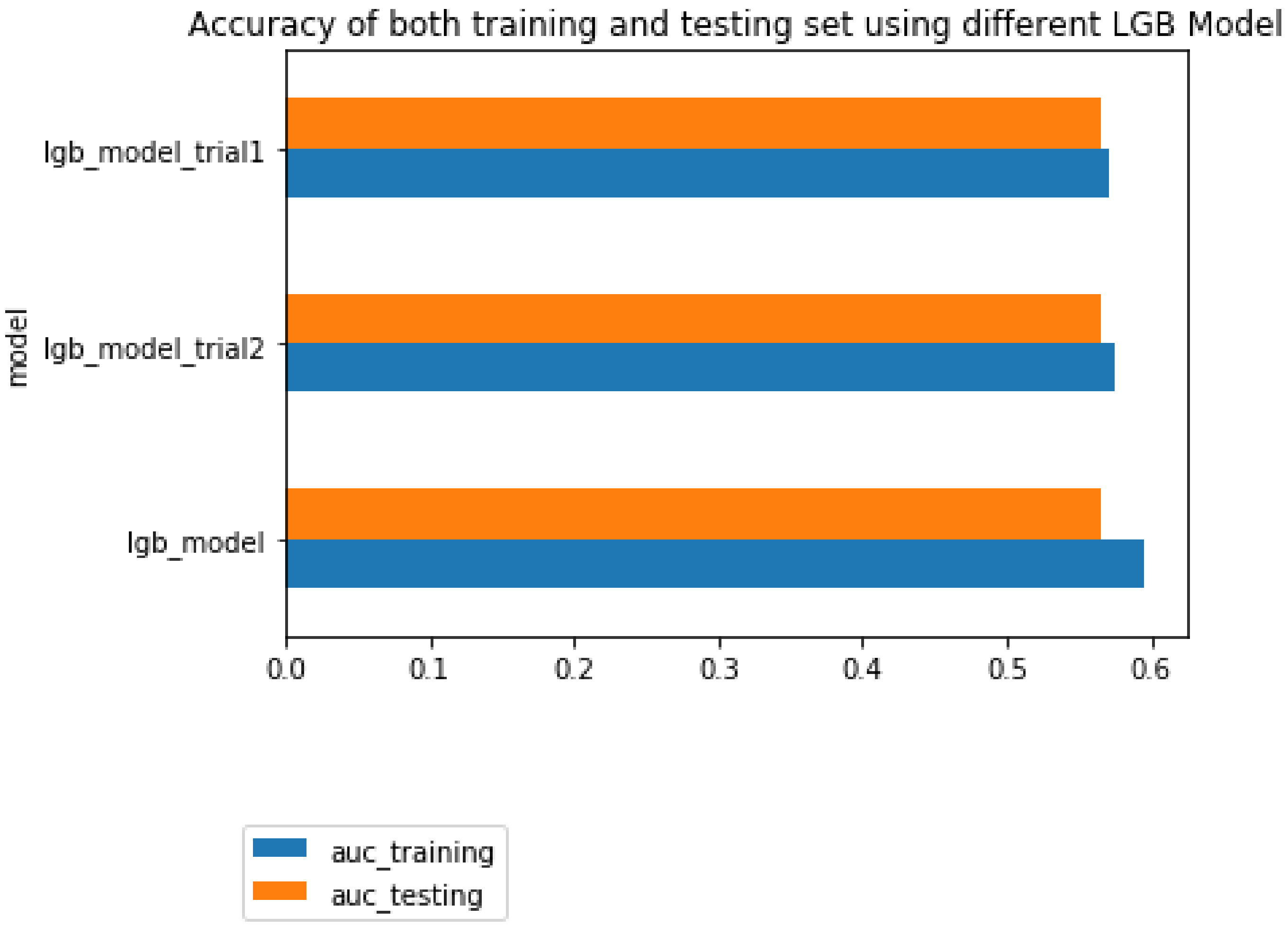


The Area Under the ROC Curve (AUC) score of the logistic regression is 51.41% whereas that of the lightGBM model is 56.62%

MODEL SELECTION

- Tuning LightGBM Model: To further optimize the accuracy of the LightGBM model, we can vary the following features
  - use large max_bin (slow down working process due to large dataset)
  - use small learning_rate in addition to increasing the num_boost_round
  - use large num_leaves (may cause overfitting)
  - use bigger training dataset (not applicable in our case)
  - try using 'dart' method
  - or try using categorical features directly (already eliminated)

MODEL SELECTION



CONCLUSIONS

- The prediction was made via the Kaggle kernel of this competition and was submitted; however the validation score is not available since this competition is still going.
- Overall, the lightGBM model gives us a better AUC score in comparison to the logistic regression model → LightGBM is recommended for this dataset with further tuning to avoid overfitting and improving the AUC score.