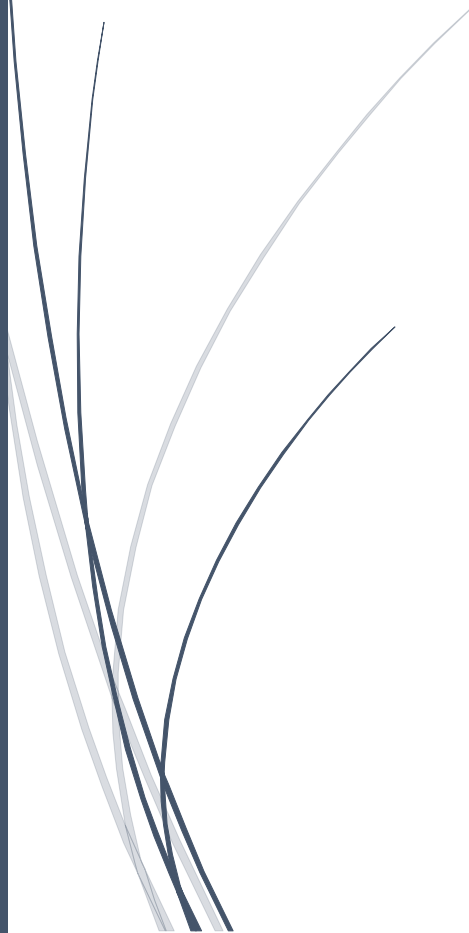


A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

12/8/2018

# Stock Price Prediction Using Financial News

Several thin, curved lines in shades of blue and grey originate from the bottom left and sweep upwards and to the right.

Author: Heyang Huang hh2720  
Lin Shi ls3311  
Yining Chen yc3566

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

# Stock Price Prediction Using Financial News

## 1. Overview

The ubiquity of data today enables investors to somehow predict stock price. However, a model that is resistant to financial crisis and efficiently utilizes both market and financial news data, has not yet been created. By ingesting and interpreting the data with regressions, data visualization, text mining and linear and ensemble machine learning methods, we develop this model to find signals in this sea of information and predict the stock price in a ten-day window. Our model stands out because it renders a text mining solution to the unexpected price volatility, foresees major economic downturns by monitoring volatility of past stock movement and implements a stacking algorithm involving eight machine learning models with a 0.1 Root-Mean-Square-Error in back-testing. Fed with large enough datasets, the model is indeed predictive and potentially a ‘money printer’.

## 2. Market Data

The market data contains a variety of returns calculated over different timespans. After the exploratory data analysis, we have a list of findings especially beneficial to the factor identification and data selection in machine learning model:

- ( **Figure 1** ) An increasing gap between blue-chip stocks and underperformed stocks.
- ( **Figure 2** ) Price volatility is an indicator of financial crisis. (2008-2009 Global Financial Crisis; 2015 Japan Crisis; 2016 Oil Price Drop; 2018 Trade War)
- ( **Figure 3** ) The Global Crisis in 2008-2009 is far more volatile than the rest of dataset and should be removed for a better prediction in smooth global economy.



Figure 1: Market Close Price by Quantile

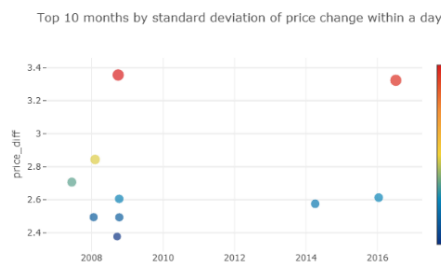


Figure 2: Top 10 Months by STD of Price within a Day

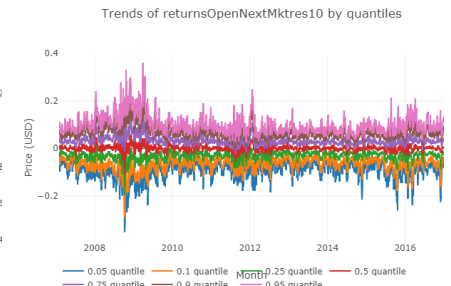
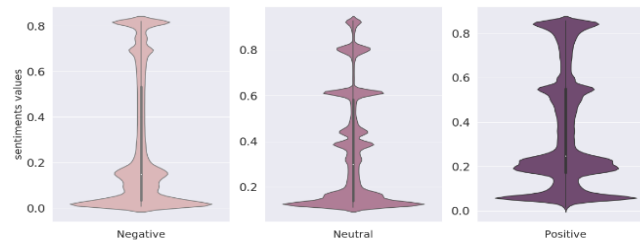


Figure 3: Return in 10 Days by Quantile

## 3. News Data

The market dataset contains the news headlines, audience, provider and urgency since 2008. As aforementioned, we cut 2008-2009 from the dataset, remove low-frequency word aggregates, and converts null values to empty strings. The word clouds and sentiments distributions are shown below:





## 4. Machine Learning Modeling

### 4.1 Feature Selection

After merging market and news data and deleting highly correlated features as shown in **Figure 5**, we perform feature selection. Feature scaling is not needed since we plan to use lightgbm - a tree-based model, which do not require standardization. We tried using a regressor model, but a problem is that it gives close-to-0 values for most of prediction. Thus, we convert this problem into a classification problem: 0 for negative return and 1 for positive return. The Result is shown in **Figure 6**.

### 4.2 General Linear Model

Four 'regularised' regression models are implemented: Ridge, Lasso, Elastic Net, and Kernel Ridge. Regularisation is a form of regression that shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. This will be particularly helpful for the current dataset where the model needs to account for 80 features.

### 4.3 Boosting Ensemble Model

Four boosting models are implemented: gradient boosting, XGboost, LightGBM, and CATboost. Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. It employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors. Therefore, the observations have an unequal probability of appearing in subsequent models and ones with the highest error appear most. While boosting saves time, we have to choose the stopping criteria carefully or it could lead to overfitting on training data.

### 4.4 Rank of Models and Stacking Algorithm

We run eight models with a ranking shown in **Figure 7**. We can see from the above graph that the LASSO and ElasticNet are the best cross-validated models, scoring very closely to one another. Gradient boosting hasn't fared quite as well, however each algorithm still obtains a very respectable RMSE. We explore stacking as a means of achieving an even higher score. In a nutshell, stacking uses as a first-level (base) the predictions of a few basic classifiers and then uses another model at the second-level to predict the output from the earlier first-level predictions. Stacking can be beneficial as combining models allows the best elements of their predictive power on the given challenged to be pooled, thus smoothing over any gaps left from an individual model and increasing the likelihood of stronger overall model performance.

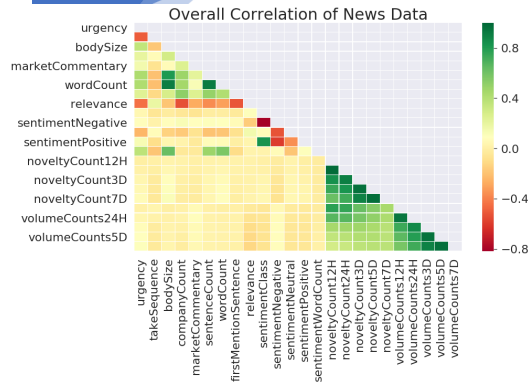


Figure 5: Correlation Heat map

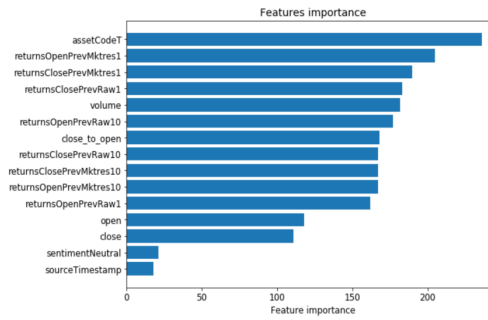


Figure 6: Feature Importance

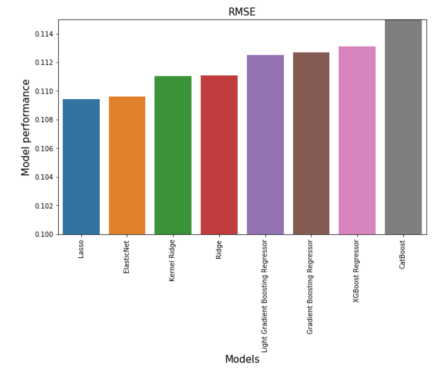


Figure 7: Model RSME Ranking

## 5. Conclusion and What We Did Above

Data preprocess and feature selection is the most important things for data analysis and prediction project. Finding a good feature is like digging gold and it will contribute a lot to the accuracy of prediction. The final testing RSME score of 0.1 indicates the success of our model. It demonstrates how we rigorously preprocess large dataset, remove outliers, gain inspirations from exploratory data analysis and build upon the machine learning model covered in class with self-studied advanced machine learning methods to enhance the predictive power. Most importantly, making accurate predictions that can beat the market is very useful for every marketing makers.