Capstone 2: Project Proposal

Forecasting Stock Movement Using News

This capstone project is inspired by the Kaggle competition hosted by Two Sigma with the datasets are provided by Intrinio and Thomson Reuters.

Problem statement:
The ability to predict the movement of stock price in the financial industry is always an intriguing but challenging task; however, we might be able to find a solution to that challenge via numerous tools available to the field of data analytics. In this study, we would like to combine different techniques that machine learning offers us and news data as a feature to build an efficient predictive model. We are hoping to learn more about the predictive power of news data and thus explore its significant economic impact on the financial industry.

Clients: Financial-product companies and comsumers who want to invest in stocks.

Data description:
There are two sources of data in this study: the market data the news data. These two datasets are available on the Kaggle page of this competition and were collected from 2007 to present. Howeverm they are not guaranteed to be clean or have no errors. The market dataset includes financial market information such as opening price, closing price, trading volume, calculated returns, etc...whereas the news dataset contains information about news articles/alerts published about assets such as article details, sentiment, and other commentary. The description of each variable from each dataset is as follows (source: kaggle competition):

Outline to Solution:

- EDA process on the two datasets to see if there are any data errors -- > clean and merge the datasets after cleaning
- Feature engineering if needed
- Model Selection: comparing at least two different models
- Choose the best model based on AUC metrics since class imblance is potentially an issue.

Deliverables: a complete code of the project and a full write-up will be included.