

1. Goals:

This project on is making Stock Market Predictions based on Stock data of Dow Jones and the news headlines for the past 8 years. Stock prices of companies usually has a high impact depending on the things happening around the country. The idea is to use a company's stock prices and the corresponding news headlines for 8 years and build a model which predicts the stock prices.

2. Clients:

Listed Companies:

Currently the prediction model is only for one company. But the same model can further be used to predict stock prices for any of the companies listed on Nasdaq. The companies can use the model to predict the company's stock prices and make business decisions.

Stock Brokers and Buyers:

The other set of customers would be stock brokers and buyers who would want to predict stock prices and make decisions (buy or sell) accordingly. Insights on the seasonal trends and growth can be analyzed using the model.

3. Data:

The data is taken from kaggle It has information of the top 25 headlines for the last 8 years and also the stock price data for the same duration. As of now, the model will be built using the information. Further, twitter tweets and hashtags of CNBC is also planned to be scraped from the web and used in the model.

Kaggle dataset: <https://www.kaggle.com/aaron7sun/stocknews>

4. Brief Outline(Approach):

The data in kaggle has 25 headlines which are sentences. The sentences will be concatenated and parsed to figure out the polarity (positive, negative, neutral) and subjectivity and objectivity of the sentences(headlines) for all the rows. This data will further be used to perform analysis and build the model. The focus will be on creating a model for binary classification which would predict whether the stock prices would go up or down(0 or 1). It might further be used in the future to predict accurate stock prices.

5. Deliverables:

Sep 14 – Research on what effects stock prices and what data could be used for stock price prediction. Come up with the initial plan

Sep 21 – Research and figure out a way to find subjectivity and polarity of headlines

Sep 28 – Perform Exploratory Analysis on the data and implement code to plot graphs.
Upload all code on GitHub.

Oct 5 – Start working on the model and finish cleaning up the dataset.

Oct 12 – Finish the first iteration of building the model. Validate the model against all machine learning algorithms.

Oct 19 – Fine tune two of the best performing algorithms for higher accuracy.

Oct 26 – Buffer Week.

Nov 2 – Final Code Submission and presentation.