**Milestone Report**

**Problem:**

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.

**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

**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

Initially, I checked the datatypes of all the columns and the shape of the datasets. Then I researched online and found a research paper and an algorithm explaining the way to find out the subjectivity, objectivity and polarity of a give sentence or a paragraph. The algorithm was written in a way to read only a given sentence or a paragraph to find out the polarity. I modified the algorithm such that it accepts a dataframe and runs the algorithm across all the rows with 25 columns of headlines in conjunction. After running across the dataframe, it generates new columns with data for **subjectivity, objectivity, negative, positive, neutral** in terms of percentage for each row. At the end of the algorithm, a new dataframe is generated.

I will then **concatenate** the subjectivity, objectivity, negative, positive, neutral columns of one dataframe with another dataframe which has data about the stock prices for the past 8 years. After creating the new concatenated dataframe, I checked the shape and info of the dataframe and found some null values. I replaced the null values with the **mean** of the respective column the null values were present.

After this I used describe() to check the statistics of the dataframe and find out any outliers in the dataframe and luckily did not find any outliers in the dataframe. This newly created dataframe will be used for further evaluation and predictive analytics.

**Potential Datasets that could be used:**

I was thinking of using twitter data by CNBC and the company(for which the prediction is being made) itself to perform sentiment analysis and place side by side with the stock prices for better prediction.

**Findings**:

LDA seemed to be performing good with an accuracy of 94% and a precision of 0.94, but it seemed a little fishy and when checked the ROC AUC score, it turned out to be 0.5 making the model worthless. However, I still plan to simulate the model and use LDA to see the prediction and if it works well.

There are other models XGBoost and Random Forests, which are giving an accuracy of 58%, which can still be considered because it is giving a more than 50% probability of predicting the stock prices. These models will also be used in simulation for prediction.

**GitHub Repo and Code:**

<https://github.com/ShreyamsJain/Stock-Price-Prediction-Model>