**Stock Prediction**

**Rishabh Kasat, Kritika Karamchandani, Siddharth Singh**

**Business Understanding:**

The stock market is known as a place where people can make a fortune if they can crack the mantra to successfully predict stock prices.

The current stock price almost immediately reflects the newly revealed information about the company. A stock price may depend on several factors operating in the current world and stock market. We will try to take into account a combination of mainly two factors:

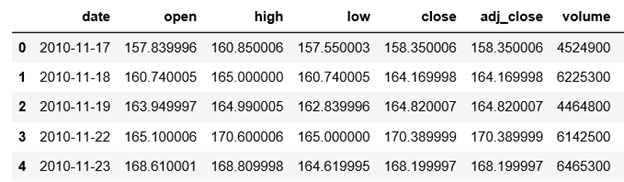
1. The past performances and records of the target company

2. How the tech industry, mainly Amazon, Google and Microsoft perform with respect to S&P 500

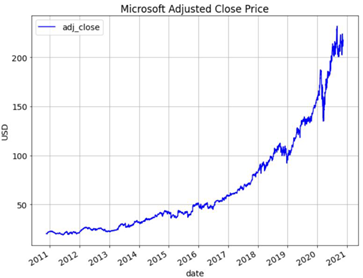
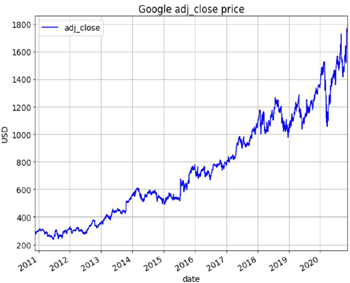
**Data Understanding:**

The Data was extracted from Yahoo Finance.

10 years of data for each of the companies was recorded to build machine learning algorithms for future price prediction.







**Data Preparation:**

Feature scaling is very important for the model to work properly. Our first model did not implement any scaling at all, and predictions on the validation set was way out of target.What happened here is that the model trained on adjusted closing price values between 89 to 125, and so the model can only output predictions within this range. When the model is trying to predict the validation set and it sees values out of this range, it is not able to generalize well. We used feature engineering to create new features which we thought could give better results. The new features our model had was :3 lags of difference between high and low price , 3 lags of difference between open and close price,3 lags for volume

We scaled the train set to have mean 0 and variance 1, and used this to train the model. Subsequently, when we are doing predictions on the validation set, for each feature group of each sample, we will scale them to have mean 0 and variance 1. For example, if we are doing predictions on day T, we will take the adjusted closing prices of the last N days (days T-N to T-1) and scale them to have mean 0 and variance 1. These scaled features were used in our model to predict the next days value

**Modelling:**

We are using Regression and Tree-based Regressor models. We used Regression as our base model and compared it to the other advanced Tree-based Regression models such as Random Forest and XgBoost.

Linear Regression:

Linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. We used linear regression here to fit a linear regression model to the previous N values, and use this model to predict the value on the current day.

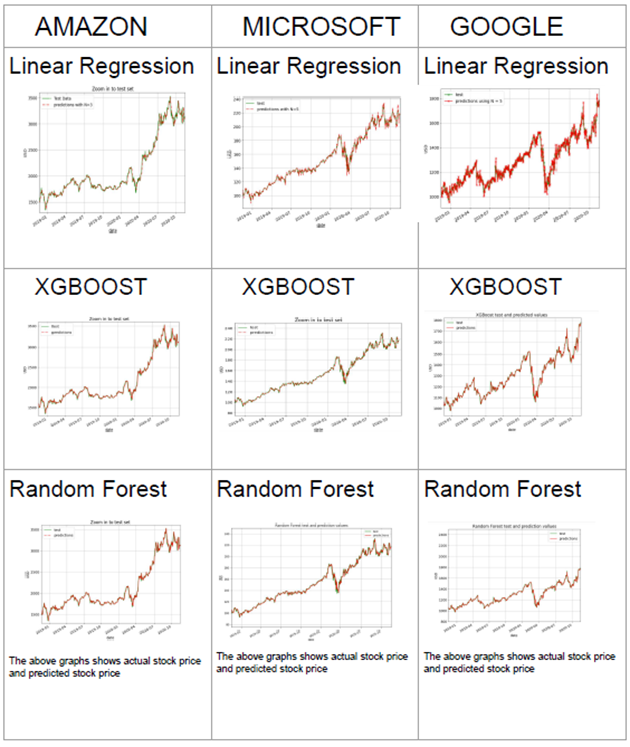
Tree-Based Models:

Tree-based ensemble models have multiple advantages compared to other machine learning algorithms: they are easy to understanding, their results are better than those of the underlying weak learners, they don’t have a normality distribution assumption and can handle mixed data types, they are robust against overfitting, outliers, noise, multicollinearity, and input and they are [computationally](https://www.researchgate.net/publication/275641579_COMPARISON_OF_MACHINE_LEARNING_ALGORITHMS_RANDOM_FOREST_ARTIFICIAL_NEURAL_NETWORK_AND_SUPPORT_VECTOR_MACHINE_TO_MAXIMUM_LIKELIHOOD_FOR_SUPERVISED_CROP_TYPE_CLASSIFICATION) inexpensive.

We trained the XgBoost and Random forest model on the train set, tuned its hyperparameters using the validation set, and finally applied the models on the test set and reported the results.

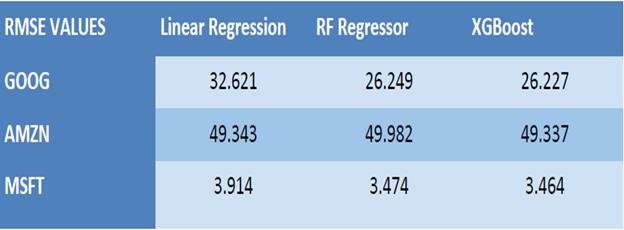
The model is robust to overfitting and outliers, and on evaluating the model on different data sets we found the maximum RSME was less than 50, so we expect our model to perform well on different data sets with similar number of data points.

**Evaluation:**

****

We plotted the actual stock prices of Amazon, Microsoft and Google from the test set which is denoted by the green lines and compared it to the predicted values of the test set which are denoted by the red dotted lines. We observed that our models were able to capture the overall positive trend for all three companies for the test set.

We used ‘RMSE’ as the performance metric for our models. Below is the ‘RMSE’ score for each of our models: -



As we can see from the above table, XgBoost model performance is the best for all three companies. We couldn’t observe much difference between the RMSE score for the tree-based models i.e., Random Forest and XgBoost but overall, the tree-based models performed better than Regression.

The results from our model will help us predict closing price as accurately as possible and better predictions will lead to higher clientele which in turn will give the company access to more data which can be used for training purposes and experimentation on using different variables which can add value to the models.

**Deployment**:

• Discuss how the result of the data mining will be deployed.

The main aim for the project is to predict the stock price for which the model should be trained, and then the predictions can be done on the test set. In the business set up this process should be a continuous process where a pipeline is to be created such that the model incorporates the new data points available for further training purposes.

Deployment will also include creation of reports, generating a user base which will receive the report, and how the report should be delivered to the users. Some important factors that will be considered are:

* Frequency of prediction generation
* Applications that can access the model

• Discuss any issues the firm should be aware of regarding deployment.

Deploying a machine learning model is not an end of the deployment, it will be an iterative and continuous process. Some issues that the company might face in the long run are:

Continuous Improvement: Over a period of time new features might be available that add more value to the model which will have to be integrated in the model. As new data is available the model will require hyperparameter tuning to avoid any kind of bias.

Continuous Monitoring: The company will need to have a dedicated team which monitors the model and its predictions.

• Are there important ethical considerations?

One of the most important ethical considerations is that any sensitive information should not be available to unauthorized people. Few measures that can be taken with respect to this are:

* Anonymization of data for development teams
* Restricting access to other users and applications
* Signing of NDA for the employees that have access to data

• Identify the risks associated with your proposed plan and how you would mitigate them.

One of the major and the most common risks that the company faces is the advancement of information and changes. The people working on the project can change over a course of time which creates a difference in the knowledge and experiences related to the application. To mitigate this every team should maintain proper technical and functional documentations which will help the next person in identifying the flow of the project and any existing issues.

**Appendix:**

<https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/> . This is a tutorial on Machine Learning methods used for predicting Stock Price

<https://www.sciencedirect.com/science/article/pii/S1877050920307924>. This is an article on using Machine Learning for stock prediction.

<https://randerson112358.medium.com/predict-stock-prices-using-machine-learning-python-f554b7167b36>. This is an article on using Machine Learning models for predicting closing precise for stocks

**Dataset**:

The data set was procured from Yahoo Finance. - https://finance.yahoo.com/