**Business Problem**

Stock market analysis and prediction will bear huge financial payoffs as this will decide the profit-making for many individuals/firms. With the amount of data available on the internet, Stock market analysis and prediction have gained much momentum of late.

Analyzing data related to the stock market can be tricky because of the amount of data available, but it is equally interesting. Creating a proper prediction system will help in profit maximization.

The primary aim of this project is to build a prediction system using different models to see past trends and forecast if the value of a stock will rise or fall.

**Background/ History**

Stock market prediction attempts to determine the future value of the company's stock or other financial instruments traded on an exchange. Successful prediction of a stock's future price could help gain significant profits.

Stock market data is complicated, and predictions are not straightforward as they depend on various factors. The stock market prediction is not always accurate, but we need some prediction to be done to invest in stocks.

Prediction depends on the stock's previous performance, various factors like company revenue and particular sector performance, and multiple announcements made by organizations and the Government.

Stock prediction methodologies fall under the following categories: fundamental analysis, technical analysis, and technological methods.

**Datasets**

The datasets used in this project will be sourced from Kaggle, a public-domain dataset. I am planning to use data from two different sources and utilize the following data:

Date – Trade date.

Open – Open price of the day.

High – The highest price in a day.

Low – Lowest price in a day.

Close – Close price of the day.

Volume - Number of shares traded.

Name - Stock's ticker name

**Methods**

As part of this project, we use stock market data with the last five years data for each stock name. Stock market data is time series data.

Time-series forecasting models predict future values considering values previously observed. Time-series forecasting is used widely for non-stationary like temperature data for a period of time and stock prices over a period of time.

For this time series data, we will be using the ARIMA model.

I will try to implement any time forecasting model other than the ARIMA model.

**Analysis**

Initially, I was planning to use Linear Regression and Decision Tree. However, based on the feedback, I have chosen to use the Time Series Forecasting model – ARIMA.

Stock prices are not randomly generated; they can be treated as a discrete-time series model based on well-defined numerical data collected at successive points at regular intervals.

I have taken just a couple of stocks – AAL and UAL and split the data into training and test data sets, built the model and calculated the MSE (Mean Squared Error). I have got MSE of 0.748.

**Assumptions**

Share price depends on factors like Quarterly results and any announcements by the organization or Government. However, while doing the analysis here, we will not consider all these details except the history of the stock.

**Limitations**

As we are sourcing the data from public websites, which are not the latest, our analysis will not be able to provide future predictions from the date the paper is published.

**Ethical Considerations**

We don't foresee any privacy issues when sourcing this data from public sites. However, stock-related data is always posted on websites that are accessible to the public. Hence, there are no ethical issues.

The only issue is that the data is not the latest, which might impact the model predictions.

**Challenges**

I will be using two different datasets, and if the data size creates a problem with running on my local machine, I might have to delete a few records, which can impact the analysis/prediction.

To build a system, we need business expertise along with technical knowledge. As I am not a Subject matter expert, the model implementation might need to be revised.

While investing in stocks, people will look at data for the last five years from the current date; we will only have visualizations on the current date.

The models built here in a short period might not be accurate; we might have to tweak them to make them accurate.

**Implementation Plan**

Below are the steps which I will be following as part of the implementation plan:

* Data Reader
* Data Cleansing
* Select two stocks to analyze ( as we cannot implement all)
* Data Visualizations
* Build and fit ARIMA Model
* Calculate accuracy
* Visualize the predictions.

**Future Uses/Additional Applications**

To make this application easily accessible to traders and investors, we can implement the following:

1. Simple UI will allow the user to input the ticker symbol to get the predictions charts.
2. Provide a public API that will give data in JSON format and chart in png format as API response.

In addition, we need to find an API that will provide the latest data based on the ticker symbol so that the data is the latest to get accurate results.

**Recommendations**

Apart from the ARIMA forecasting model, we should also implement other time series forecasting models, compare their accuracy, and select whichever suits us better.

**Visualizations**

**AAL:**

**Chart, scatter chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**UAL:**

**Chart, scatter chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

Chart, line chart

Description automatically generated

**References**

* Damarla, R. (n.d.). Stock Market Prediction using Decision Tree. Kaggle. Retrieved March 18, 2023, from <https://www.kaggle.com/datasets/camnugent/sandp500?resource=download>
* Mooney, P. (n.d.). Stock Market Data (NASDAQ, NYSE, S&P500). Kaggle. Retrieved March 18, 2023, from <https://www.kaggle.com/datasets/camnugent/sandp500?resource=download>

**Conclusion**

After evaluating the model, we have observed that this model has an accuracy of approx. 98%. This shows that this model is best fit for stock market data. We can use the stock market data APIs to get the latest predictions.

**Appendix**

1. NYSE: It stands for New York Stock Exchange. It is the biggest marketplace in the world for buying and selling shares.
2. NASDAQ: National Association of Securities Dealers Automated Quotations. It is an electronic stock exchange to buy and sell shares.
3. S&P 500: It stands for Standard and Poor's 500. It is a stock market index tracking the stock performance of the top 500 companies listed on the stock exchanges of the USA.
4. ARIMA Model: It stands for AutoRegressive Integrated Moving Average. It is a statistical analysis model which uses time series data to analyze the dataset or predict future values.
5. Stock symbol – Stock Symbol is also known as Ticker Symbol. It is a unique series of letters assigned to a security for trading purposes.

**Questions and Answers**

1. **What is time series data?**

Time Series Data is also called Time Stamped Data. Time Series data is recorded at consistent intervals of time.

1. **Is stock market data complicated? If so, why?**

Stock market data is complicated as it consists of various investors (small and large) making uncoordinated decisions about various investments. Stock values are not always linear as the value is not just time based but involves various factors.

1. **What are the time-series methods that have been used to predict stock prices?**

The following are the models used for time-series data prediction or forecasting:

* 1. Autoregression (AR)
  2. Moving Average (MA)
  3. Autoregressive Moving Average (ARMA)
  4. Autoregressive Integrated Moving Average (ARIMA)
  5. Seasonal Autoregressive Integrated Moving-Average (SARIMA).

1. **What is the ARIMA model?**

Autoregressive Integrated Moving Average (ARIMA) is a statistical mode that uses time-series data to predict or forecast future trends or better understand the data.

1. **Can we use any model other than the ARIMA model?**

As stock market data is time-series data, we can use any of the above-mentioned models (mentioned in question# 3)

1. **How to calculate the accuracy of time series models?**

As in other models, we cannot directly calculate the accuracy score. However, we can calculate Mean absolute percentage error (MAPE) and subtract that from 100 to get accuracy.

1. **Can we use this model implementation with datasets other than NYSE?**

We can use this model for any stock market data regardless of stock exchanges. However, it might need some tweaks to the implementation depending on the column names used.

1. **Do we have any APIs or sources which can provide the latest data instead of old data?**

There are many APIs available. Below are few APIs to name a few:

* 1. [YH Finance](https://rapidapi.com/blog/best-stock-api/%0ahttps:/rapidapi.com/belchiorarkad-FqvHs2EDOtP/api/yh-finance-complete)
  2. [Yahoo Finance](https://rapidapi.com/blog/best-stock-api/%0ahttps:/rapidapi.com/sparior/api/yahoo-finance15)

1. **Will these models withstand the vast datasets?**

To withstand vast datasets, we need to use hardware accordingly with higher configurations.

1. **What visualizations can be used, and how will they help?**

As this is time series data, per my knowledge, Line charts are the best.

1. **Can we implement an interface for the user to provide stock symbols and provide predictions?**

In the current implementation, I have provided an option to input the Stock symbol to get the prediction. We can implement this as an API and include the required details in response along with the URL for the prediction graph.