National College of Ireland

# Sector Based Stock Market Prediction In USA

MSc Research Project
Data Analytics


Muhammad Nizam Uddin
Student ID: x14127032

School of Computing
National College of Ireland


Supervisor: Dr. Catherine Mulwa

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | Muhammad Nizam Uddin |
| **Student ID:** | x14127032 |
| **Programme:** | Data Analytics |
| **Year:** | 2021 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Dr. Catherine Mulwa |
| **Submission Due Date:** | 16/08/2021 |
| **Project Title:** | Sector Based Stock Market Prediction In USA |
| **Word Count:** | 6335 |
| **Page Count:** | 22 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Muhammad Nizam Uddin |
| **Date:** | 16th August 2021 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Sector Based Stock Market Prediction In USA

Muhammad Nizam Uddin
x14127032

16th of August 2021

**Abstract**

At a time when stock market investments have seen a surge and market unpredictability has hit new heights with global recessions due to events such as the covid 19 pandemic of 2020, this research offers a fresh outlook to lure in new investors and provide a safety net for seasoned investors by implementing machine learning models to predict stock market behaviour. This has been accomplished by building an Sector based index to segment out the companies and aggregating the closing prices based on that to determine how the Sector would perform over time using RNN, LSTM and Time Series ARIMA.Among the model LSTM obtained the average highest- RMSE value 6.64 with model accuracy 93.65%. This has been attained by establishing indices not just for companies but also paving the path for sector based segregation to enhance the research and make these indices more accurate over time. The research provides a crisp comparison between the existing models and highlights the pros and cons of implementing a new model as proposed for the research.

## 1 Introduction

Investing in stocks and trading has become more popular with the use of mobile applications and easy trading features which have taken E-Trading to a whole new level, more and more amateur investors have begun investing in the stock market over the past decade. With major companies such as Amazon, Google and Facebook projecting regular growth, several amateur investors have entered the market in hope for growth in investments. This trend of basic analysis and somewhat safe investment techniques is what motivates this research. The ideal motive being to be able to predict that to what extent a stock value could go up and what could be considered a 'safe investment'.

State of the art methods for stock prediction involve the use of a very high number of variables which provides an intrinsic high level of accuracy in the data domain but has an extrinsic nature when it comes to actual real-world implementation. This is due to the fact that most companies stock values are not subjected to the restricted number of variables and the consumption world changes rapidly. Some companies continue to adapt, for instance, Amazon started off with home deliveries in 1994 and capitalised on the market requirement at the time. In 2011 Robischon (2017) Amazon adapted to the culture of online video streaming to generate a market disruption and against the streaming giant Netflix. All this while Amazon Web Services was taking over the cloud computing world and the company grew four folds over the years with their latest ventures now being the

aviation industry. For an amateur investor, the outlook would be to invest in a company like Amazon and they would consider the company as a safe investment. The naïve investor would overlook aspects such as the sector involved for stock value prediction, the effects of market monopoly which giants like Amazon create on the investors mind and they might end up overlooking the value sector based investment has over company-based investments. The proposed research looks at this aspect of stock behaviour where sector based investments and market monopolies in sector are considered as a major checkpoint for stock behaviour analysis across various domains.The idea here was to implement a sector based index among the sector to determine which companies are following the sector index trend .The model provided users a platform for users to enter the data from the companies of their choice and compare it with the sector standards. Due to the time constraints presented by the research at this junction, the research has been based around the list of companies in the Fortune 500 for 2018, with a clear pathway to expand the research based on the results presented in the conclusion and future works section of the research.

## 1.1 Research Question and Objectives

RQ: *"How accurately can sector based index vs company performance be implemented to predict the stock behaviour?"*

Sub RQ: *"To what extent can the stock behaviour be linked to the sector types of the companies whose primary and secondary sources of income by sector domain are provided?"*

To solve the above research questions following objectives laid down and implemented which are discussed below:

Obj 1: Critique the state of the art methods for stock market related works (2012 - 2021)

Obj 2: Sector based segregation for companies and determine indices

Obj 3: Feature extraction in the form of stock opening and closing prices overtime and variable analysis

Obj 4: Aggregate stock price behaviour with sector based data

Obj 5 (a) Implementation, results and evaluation for RNN for sector based as well as for individual company

Obj 5 (b) Implementation, results and evaluation for LSTM for sector based as well as for individual company

Obj 5 (c) Implementation, results and evaluation for Time Series ARIMA for sector based as well as for individual company

Obj 6: Compare the results of sector based prediction indices with company stock behaviour individually to acquire accuracy of the model

**Major Contribution** This research made a major contribution by bringing a new model based on sector variable in the stock market prediction domain. Using price prediction of sector index and as well as individual company parallel, investor will have an extra safety net for their investment on which they can capitalise more for their investment.

Rest of the research will be as follows : Chapter 2 presented an investigation of existing state of the arts methods in stock markets.This is included methods such as RNN, LSTM and Time Series Analysis of stock trends. In Chapter 3 Research Methodology, which was focused on combining the two fields of sector based segregation and stock behaviour

prediction as well as design process. This was to set the theme for implementing the same machine learning models for sector based analysis which were used for normal stock price prediction. Another key factor in this research is that since we are using data collected from Fortune 500 companies, highlighting market monopolies and understanding their impact on stock values on other companies in the same field would also become feasible and lead to lesser skewed metrics in the output. Chapter 4 described the technicalities of implementing the research and showcasing the results based on how well the sector classification models work when clubbed with the stock prediction models. To conclude the research, this is followed up by the conclusions section where we answer the questions proposed in this section of the research and lay the ground works for the expansion of the research in a trickled down effect with more and more companies involved and to make the model more scale-able.

# 2 Critiques of stock market related works (2012 - 2021)

State of the art methods for stock behaviour prediction rely heavily on gathering information based on reporting and various media platforms. This is not restricted to information available online and expands to the domain of expert advice coming in from major Stock Market Prediction publishers such as Bloomberg Shah et al. (2018). With advancements in analysis techniques, data science stepped into the picture providing concrete analysis and proves of stock behaviour being predictable to a certain extent with Time Series analysis and Neural Networks taking the top spots guarantying up to 90 percent accuracy in stock prediction Borovkova and Tsiamas (2019) using various parameters such as closing prices, opening price, hourly variation in parent and partner companies, etc. While this sounds like a result that could drive an investor towards an easy fortune, if the accuracy of these models went up to 90%, anyone who is mediocre at data analysis would be a millionaire. The reason for this not happening is that the models were accurate based on the type of variables they were implementing, but the research including all parameters necessary for the output to be accurate was not assured. This research, however, adds in a variable that has not yet been analysed in the domain and focuses on an sector based stock prices in account. This section of the research will focus on highlighting the best methods for this segregation process and then on highlighting the best market practises for stock prediction, which will be followed up by an implementation strategy in the sections to follow.

## 2.1 A Critical Review of Market Segmentation such as sector Based Analysis

For an analysis such as the one implied, the first step is to initiate sector based segmentation process for the companies involved. To do this, the current system proposes a simple data download of the various sector names with a focus on their domains. These fortune 500 companies data publicly available [1]. The idea is provided in a simple form to be able to classify companies based on their dominating sources of income and per sector analysis. This type of analysis will focuses on designing ensemble techniques Seker et al.

[1] https://www.kaggle.com/agailloty/fortune1000?select=fortune1000.csv

(2013) specifically to understand the correlation among sector types to club with and company stock values, hence providing a concrete base to the research. State of the art methods for market segmentation analysis include the implementation of Decision trees Barak et al. (2017), Random Forest classifiers Khaidem et al. (2016) and correlation vs causation matrices Wu et al. (2020) to highlight the interdependence among variables such as stock price and sector domains. To the convenience of this research, the data available has been through this step of implementing the aforementioned methods and hence provided a boost to the research in the form of one step being implemented as a pre-requisite to the research. One of the biggest lags in modern day predictive methods is the lack of external variables considered when it comes to the companies whose stocks are in question. However, for this research, the idea is to eradicate the need of considering external variables as instead of companies, the sector will be analysed for price prediction. For this section of the research, the sector names have been taken from the data source provided and a simple classification method of using decision trees Kamiński et al. (2018) has been implemented by the author of the data and made available publicly. With this step out of the way, the final output is company names available with sector based classifications.

## 2.2 A Critiques of Stock Market Prediction Methods

The current market analysis methods include different variations of Neural Networks Qiu and Song (2016) and Time Series Analysis Grigoryan et al. (2015) for stock prediction using different variables. Some of the techniques which hold an impact on the proposed research are discussed below.

### 2.2.1 Recurrent Neural Networks

Selvin et al. (2017) proposed the use of recurrent units of variable behaviour from stock data into an RNN model. While the accuracy of the research done by Selvin et al. (2017) was above 90%, its enhancement into forecasting was not available. The reason for using this model is that stock prediction would require understanding and implementation of long term dependencies among variables. Despite its pros, vanishing gradient problem might occur, which iterates that with more layers of information in the neural net, previous information is left out of the analysis. To solve this problem, the use of LSTM was proposed Pawar et al. (2019).

### 2.2.2 Long Short-Term Memory

LSTM provides an extra layer of memory security by implementing Long Short-Term memory cells instead of the traditional cells. As proposed by Pawar et al. (2019) , this is an advancement to the existing RNN model and provides better results than deep neural networks (DNN) in some cases. Towards the end of the research, Pawar et al. (2019) concluded that a high layered model with a double recurrent RNN model set up using LSTM as building blocks obtained the highest accuracy in stock prediction. However, the research conducted was limited to a small data set, as also highlighted by the author as the idea for that research was not fixed on stock prediction, but to provide a general 'ratings prediction scale using multivariate data'.

### 2.2.3 Convolutional Neural Networks

Hoseinzade and Haratizadeh (2018) proposed the use of CNN model to assure that a high number of variables with missing data are included and aggregated to ensure the integrity of the results provided. Due to the financial market behaviour showing a high relativity on various factors, this research focused on including as many variables as possible and filling out most of the blank data using aggregations. CNN methods are superseded by RNN and LSTM based on the fact that they have internal memory units, which provide better accuracies in historic data analysis such as the one being proposed here. However, a research conducted by Hinton et al. (2012) shows that in generic analysis such as this where the variables are high in number, CNN performed better (up to 96% accuracy as compared to $> 90\%$ for other methods), hence, an ensemble method where some of the features of CNN would benefit the research.

### 2.2.4 Time Series Analysis

Another classic example of an ensemble research comes in the form of the research done by Mehtab and Sen (2020) , where the various neural network outputs were analysed by defining the predictive analysis as a result of a solution consisting of first establishing a correlation matrix among the variables, then using Neural networks for price estimation and then using time series analysis to predict the future pricing based on historic trends Weng et al. (2018) . An update on this would be to use the ARIMA (Auto-regressive Integrated Moving Average) model, which focuses on analysing the correlations among variables overtime. The use of this model can provide a reason to leave out CNN and focus on an ensemble technique where the idea will be to segregate the market based on sector, predict current prices using a simple division of data and then use ARIMA to predict the future prices of the stocks.

Idrees et al. (2019) discussed the volatility of stock market being a major factor towards the failing outputs of most deep learning models towards predicting the prices. This approach breaks down the prediction into 3 aspects: long term, short term and medium term. On any change that is observed in the short term, the long-term model is adjusted, and it adapts to the research. ARIMA model adds accuracy to the forecast by highlighting univariate data points in the research, establishing relationships among variables which leads to an aggregated output and hence, for this research, as compared to using both CNN and RNN, we can leave out CNN from the analysis altogether, as post predictions, the forecast model will aggregate the values. If CNN is used as well, it would add a layer of aggregations and the output risks over fitting.

### 2.2.5 Identified Gaps in Literature and Conclusion

To get a result which is market ready and accurate enough to be stated as a reliable one, using an ensemble pack of techniques would be a must as suggested in various research critiqued in this section Pawar et al. (2019) ,Hoseinzade and Haratizadeh (2018) ,Mehtab and Sen (2020) ,Weng et al. (2018). The approach which researched in this research is derived from the research on ensemble techniques used by Mehtab and Sen (2020) hybridised with the idea of sector based segregation as compared to direct company based stock analysis. Another change being implemented to their research is the use of RNN (with LSTM cells) instead of CNN as the final phase of the research would implement ARIMA as compared to the basic time series analysis which would aggregate

the variables if missing to provide the nearest possible predictions. A small comparison between the techniques proposed is shown in the table 1.

Table 1: Pros and cons of various Techniques

| **Models** | **Pros** | **Contribution to this Research** | **Accuracy** |
|---|---|---|---|
| RNN | Recurrent units of repeated behavior can be analysed | Data has repeated trends, hence RNN is recommended Selvin et al. (2017) | Up to 82% accurate |
| LSTM | Adds to the RNN performance by providing building blocks which implement short term memory storage | Should enhance the performance of the predictive model as resaearched by Pawar et al. (2019) in multiple research methods | Up to 84% accurate |
| CNN | Implies a simple aggregation technique allowing the researchers to analyse a high number of variables Hoseinzade and Haratizadeh (2018) | The idea of aggregation can be used as a high number of variables are to be analysed using RNN | 80% accurate |
| Time Series | Allows forecasting using multivariate data but is constrained by data quality | Provides a base for forecasting in the domain of the research Mehtab and Sen (2020) | 84% accuracy |
| ARIMA | Aggregates the interdependent variables in a long,short,medium term forecasting model Idrees et al. (2019) | Proposed implementation for this to follow the results presented by RNN and LSTM Models | 96% accuracy (Limited data sources) |

To conclude this section of the research, a basic understanding of the thought process of researchers in the domain was established, which has further been implemented in this research. This step would attain the completion status of Objective 1.

# 3 Stock Market Methodology Approach

## 3.1 Introduction

For this research, the best approach is to implement a CRISP-DM data mining model as compared to the KDD. The reason for this approach is that CRISP-DM has the capability to invoke multiple layers of implementations in case of business understanding changes, which would be a more feasible approach in stock market scenarios. However, to imply

CRISP-DM, it has been modified to include certain elements of the KDD approach as well due to the use of a knowledge base being constructed in the form of sector being classified for the companies whose stocks are being analysed. The Sector based index will provide an additional layer of confirmation of the future trend for overall sector which will provide a extra safety net.

## 3.2 Data Sources, Usability and GDPR Compliance

The project implements data taken from three different sources [2] [3] [4]. All the data sources that have been used in the research are CCO Certified for public use and the research is hence GDPR compliant. Certificate can be viewable [5]. The data has an average usability index of 7.4 with a validity average of 9.2, pertaining to a high accuracy score of the outputs generated in the research. Data source 1 is a simple reference table for the industry standard abbreviations used by the NYSE (New york Stock exchange) for company references, which can be linked to data source 2 which contains the stock price values for all companies which are included in the data. Data source 3 contains information on the industries and sectors these companies are involved in. The reason data source 1 is required is to join the abbreviations in data source 2 to the full company names provided in data source 3.

## 3.3 Underlying Methodology

Figure 1 shows the approach this research has incorporated to predict the stock behaviour based on the sector types of the companies involved.

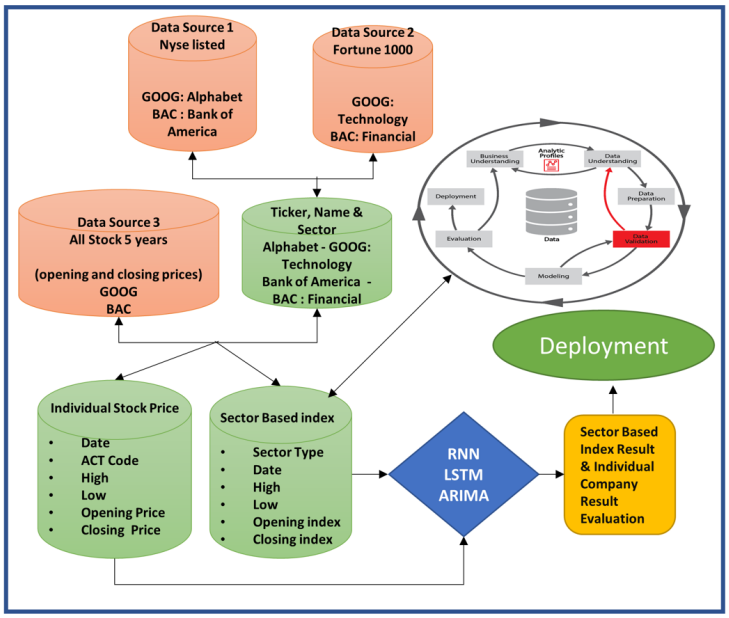


Figure 1: Methodology of Sector based stock Price Prediction

[2] https://datahub.io/core/nyse-other-listings#resource-nyse-listed
[3] https://www.kaggle.com/camnugent/sandp500
[4] https://www.kaggle.com/agailloty/fortune10,00?select=fortune1000.csv
[5] https://creativecommons.org/publicdomain/zero/1.0/

The process involves understanding the changing market, which is a desirable approach in the ever-evolving world of stock pricing models, followed up by the standard CRISP-DM methodology.The models implemented for the research are LSTM, RNN and Time Series Analysis (ARIMA) whereas the models for sector based segregation (a prerequisite within the available dataset) is Random Forest. The CRISP-DM data mining model has been implemented on the final data set obtained as a result of the process enunciated in figure 2.

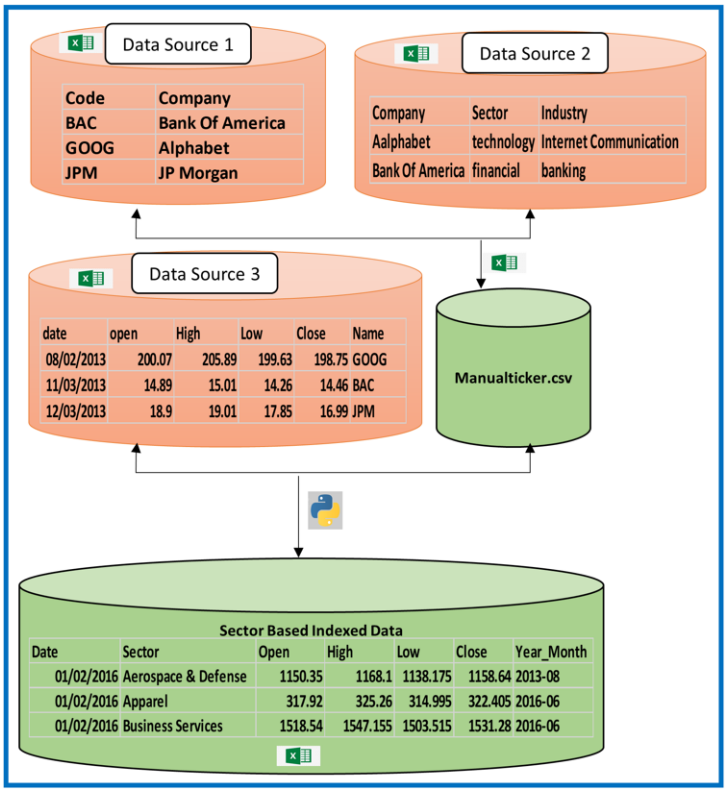


Figure 2: Data Processing

A brief elaboration of the methodology would be that the data is joined from all three sources and the final output is a file which has the sector types with the opening and closing values of all stock prices for the companies within a time frame of 2013 - 2018.

Once this was done, we moved on to implementing our models. The research follows a simple pattern of implementing the RNN, LSTM and Time Series Arima to the final output data using sectors to define indices of sector, this would allow the stake holders to view the companies they want to invest in on a scale of average performance of the sector under which their chosen companies fall.To establish an index as proposed, the steps taken to combine data from all three sources played a vital role in the research. Due to time constraint sector based index created with aggregating the closing prices rather creating price weighted or market cap weighted index. Again for individual companies data is readily available to run all three same models.

The process in figure 2 above shows how the three sources can be combined and what would form the index as proposed in the research. This step would attain the completion status of Objective 2.

## 3.4 Project Design Process Flow

The design flow of the research is shown in figure 3 and consists of a three-tier design which implements a data tier, a business logic tier and a client (presentation) tier. The client tier represents the segregation of companies based on their sector types and the classification process that goes into it. This layer also displays the final outputs of stock behaviour for the traditional company analysis as compared to sector based analysis as proposed in this research. The business logic tier consists of feature extraction of the data, modelling parameters and pre-processing methods for the data involved as shown in the figure.

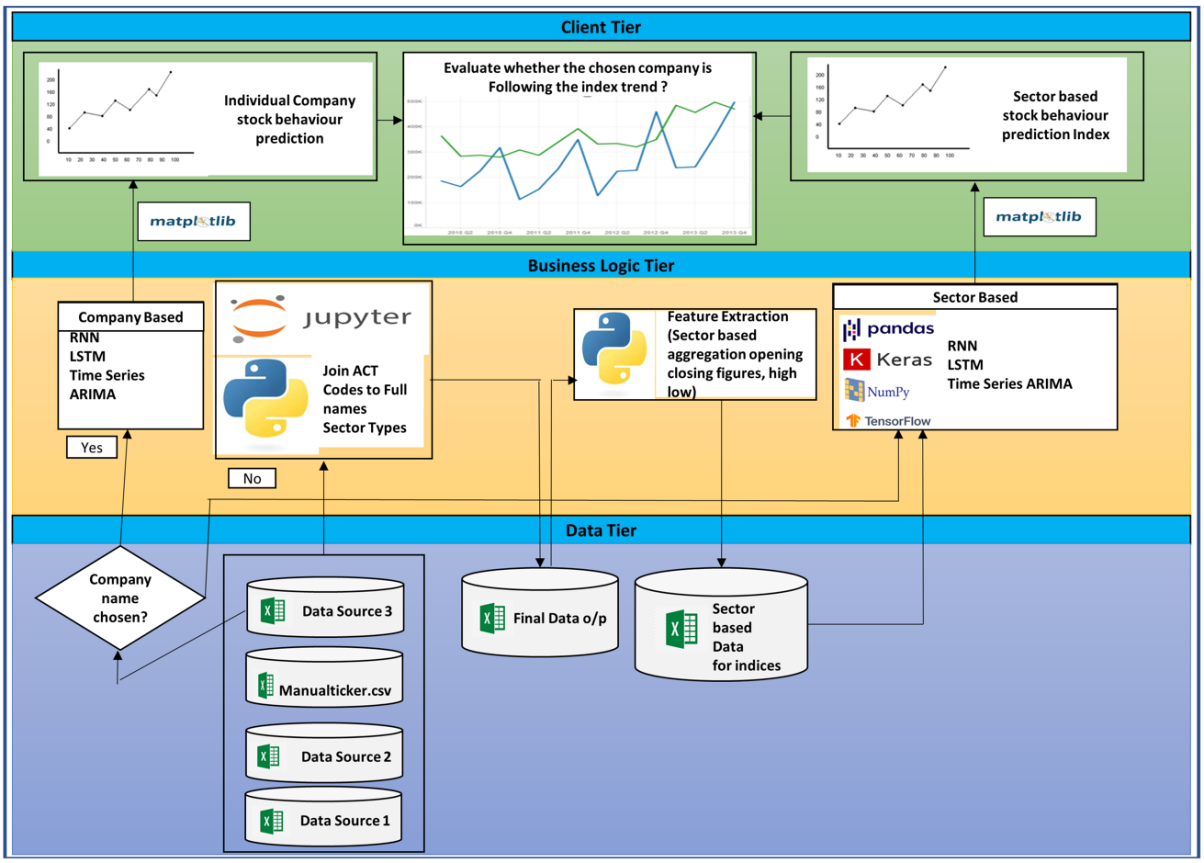


Figure 3: Project Design process of sector based stock Price Prediction

The first step would be to combine the data sources based on the ACT codes of the companies involved, then obtaining the final data as shown above in figure 2. This would then be uploaded back into Jupyter notebook and sector based data files would separately be created in a csv format. The next step is to simply implement RNN, LSTM and Time Series Arima on the data set to obtain a prediction index for the sector in question. On the other side, the clients can choose whichever company they want to compare with the index of the sector their chosen companies belong to. This design flow also allows users to observe indices until they can choose a company as per their desire with a decision choice provided in the data tier. This process is now able to detect if the companies involved show any different trends from the index of the sector they are derived from.

## 3.5 Conclusion

This research implemented the aforementioned techniques for data collection and establishing the indices as proposed previously. The model applies CRISP-DM methodology clubbed with a 3 tier design architecture to obtain the best results on a data set that has been certified for public use and has a high usability index. With the conclusion of this section, we are now set up to implement the research and evaluate how the indices perform as compared to individual company research using the same methods in the domain.

# 4 Implementation, Evaluation and Results of Sector Based Index Models

## 4.1 Introduction

This chapter of the report is dedicated towards highlighting the steps taken to implementing the research. The section is divided into a step-by-step methodology to show how the results are derived for the sector based indexing and how this research can be used to enunciate a better investment model for new investors. The idea here is to generate an evaluation metric system for performance evaluation of the sector based model as compared to an individual company analysis model using Neural Networks (RNN and LSTM) and Time Series.

The evaluation metrics used for this research are simple, the first step is to implement the models for RNN and LSTM for individual companies than sector based index which will generate a Root Mean Square Error (RMSE) value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Here N is the number observation . For stock Market prediction RMSE evaluation is widely used Waqar et al. (2017). For time series, the data was first broken down into three day intervals and analysed for 1500 days. The idea being to understand fluctuations in prediction vs price for the differences in opening and closing prices as compared to simply predicting the prices. Also it was observed if the the data is stationery or not. For the ARIMA model R- squared value used which is commonly used for time series ARIMA evaluation Ponnam et al. (2016). RSS is stand for sum squared regression error and TSS

$$R^2 = 1 - \frac{RSS}{TSS}$$

is stand for sum squared total error.

## 4.2 Data Extraction and Pre-processing: Generating the Final Data Output

To get a match on the ACT codes from Data Source 1 and matching it to company names in Data Source 2 (company names and sector), the first method of implementation was to use Levenshtein distance Lubis et al. (2018) using a simple ratio check of the uncertainties as shown in figure 4.

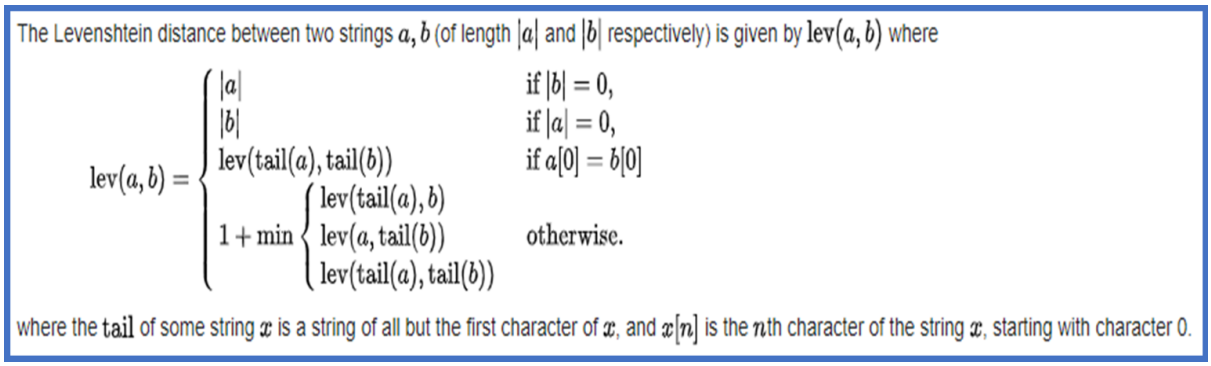


Figure 4: Levenshtein Distance

This Levenshtein distance calculates the average gap between all strings of a particular words and matches that to the nearest ones. It has scope for upper and lower limits ignorance but is incapable of reducing accuracy in scenarios where extra characters appear in one of the comparison data sets. The idea here was to get an overall ratio of over 90% and then do a manual evaluation of the results. But because of company names being extremely different in some cases the output ratio of this was a mere 11% match. There were only 505 companies in our target data set. so due to time constraint this part was done manually in excel file and save it to manualticker.csv. There were a few nulls, which were due to data source 1 having over 3299 companies while data source 2 had only 505. Hence, the next step was to remove all nulls and extract the relevant data only which led to a final list of 319 companies.

## 4.3 Exploratory Data Analysis

There was 505 individual companies stock prices available in the data. For this research Alphabet , Apple , JP Morgan and Bank of America randomly selected to analysed and for the sector indexes Technology and financial sector randomly chosen. The exploratory data analysis have been set up in a format that would take care of objectives 3 and 4 simultaneously, by combining all the data and extracting relevant features such as closing price. This step has been implemented in the research post the combination of all the three data sources after ensuring that the data wanted to use for the research is primarily usable. Some of the steps are implemented for a quick overall analysis of the data are discussed below. First, conducted a quick check on any nulls in the data for some random companies from data source 2, the reason for this granular check was to ensure that the data imported is not overtly impacted by nulls as that would not be an ideal scenario when the data sources are combined. Attached is the output of the null check, similar tests were run on industry-based data as well. The result of this was that no nulls were in the data as shown in figure 5.

```
# General info apple
BAC.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   date    1259 non-null   object
 1   open    1259 non-null   float64
 2   high    1259 non-null   float64
 3   low     1259 non-null   float64
 4   close   1259 non-null   float64
 5   volume  1259 non-null   int64
 6   Name    1259 non-null   object
dtypes: float64(4), int64(1), object(2)
memory usage: 69.0+ KB
```

```
# General info apple
JPM.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   date    1259 non-null   object
 1   open    1259 non-null   float64
 2   high    1259 non-null   float64
 3   low     1259 non-null   float64
 4   close   1259 non-null   float64
 5   volume  1259 non-null   int64
 6   Name    1259 non-null   object
dtypes: float64(4), int64(1), object(2)
memory usage: 69.0+ KB
```

Figure 5: Check Null

This was an important step because as seen in section 4.2, there are a few nulls which could skew Sector based information. Next EDA was a check on the high and low values

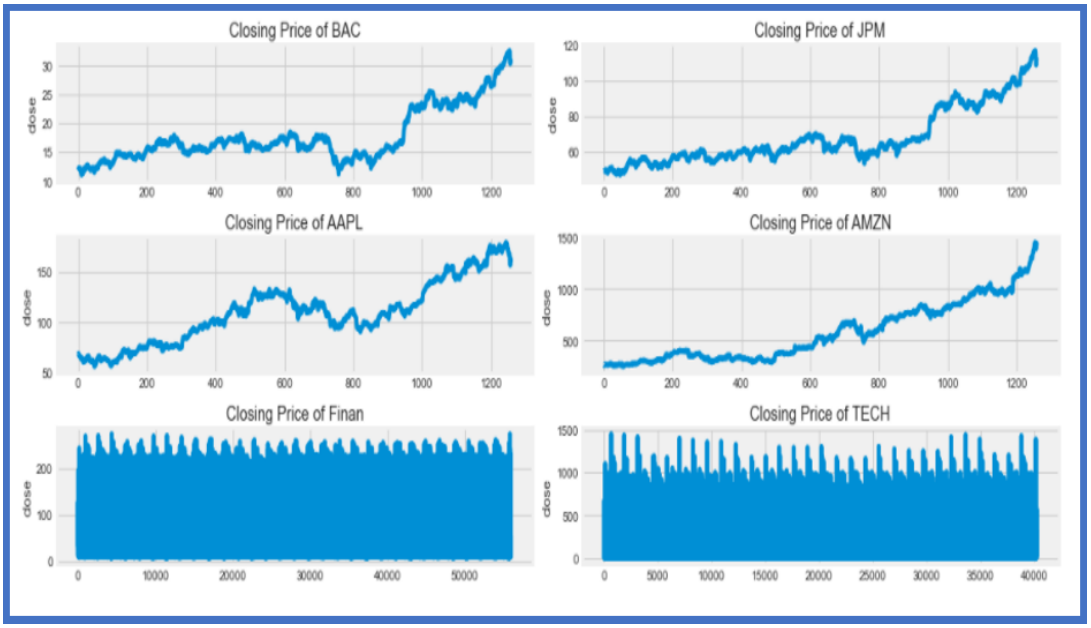


Figure 6: Check Outliers

along with the closing values of stock prices. ideally, there should not be any outliers as these values should be very close on a timeline. The output in figure 6 confirms there is no outliers exist in the data.

Another interesting metric was to observe the performance of the top 8 industries simply in terms of number of companies based on the difference of the share prices per unit time. While there were no discrepancies, but Tech sector index price action have noticeable uptrend. The Tech sector was scaling since mid-2017 shown in figure 7.

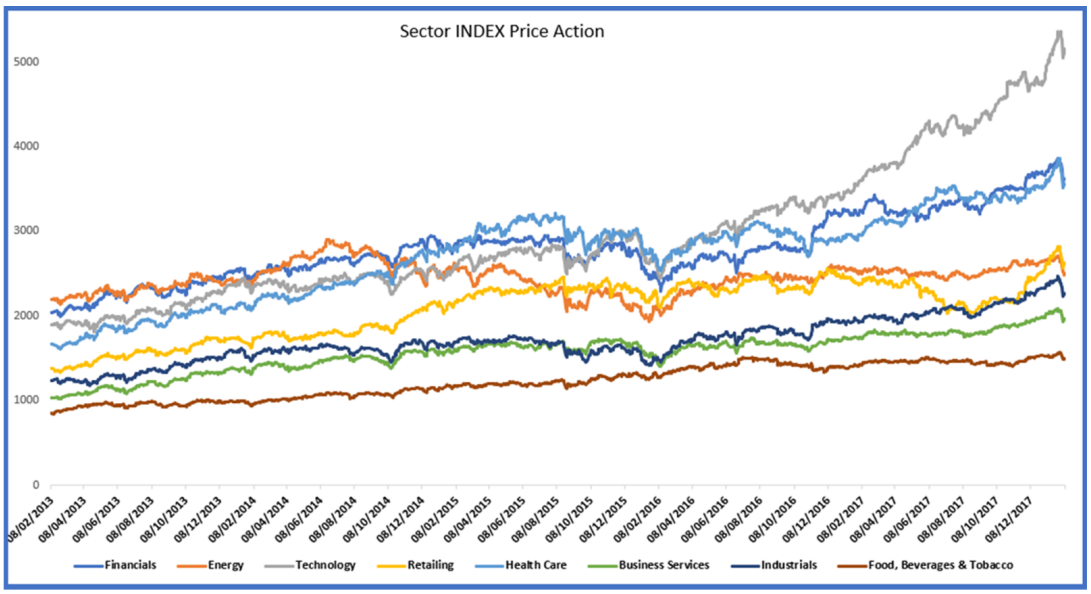


Figure 7: Sector INDEX price Action

Overall Technology, Financial and energy have more than 30% of the market share of the overall NYSE Stock Market.Below figure 8 shows the sector wise market share in the overall NYSE stock market.

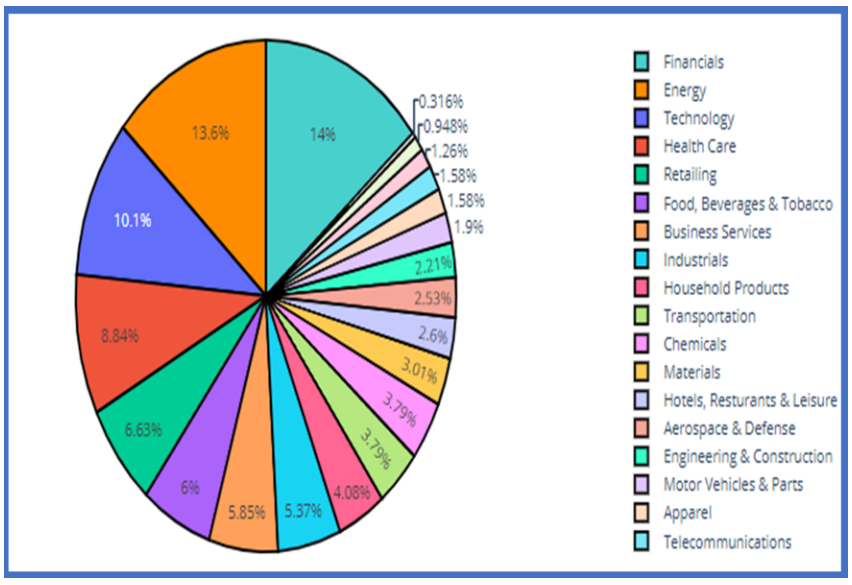


Figure 8: Sector Based Market Share

An interesting pattern detected in the correlation matrix that companies are in the same sector highly and positively correlated but on the other hand Sectors have very low correlation between them shown in figure 9.

Once the exploratory analysis was concluded, we had a basic idea on how sector or Companies performed around and what time duration's did sudden changes occurred.

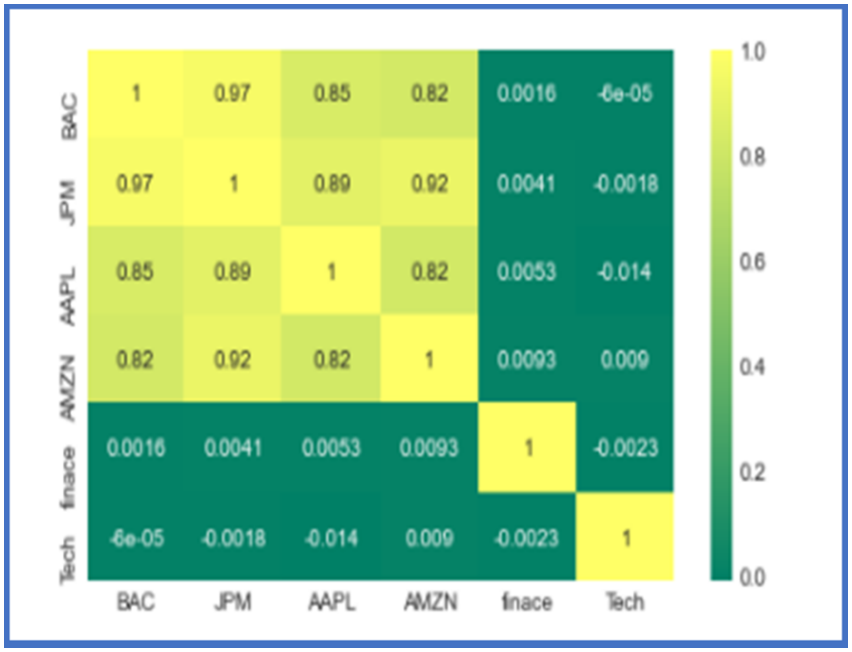


Figure 9: Correlation Among the Sector Based Companies

## 4.4 Implementation, Results and Evaluation of RNN

To implement the RNN model, the first step taken was to select the Sector types we initially wanted to start off the research with. Hence, from an EDA performed on the data, a quick view showed that the highest number of companies in the data were 'Financial','Technology' and 'Energy', all of them were presenting over 100 companies. Technology and Financial sector was chosen to initiate the implementation process to get a better understanding on how the sector based index would look like. To initialise the

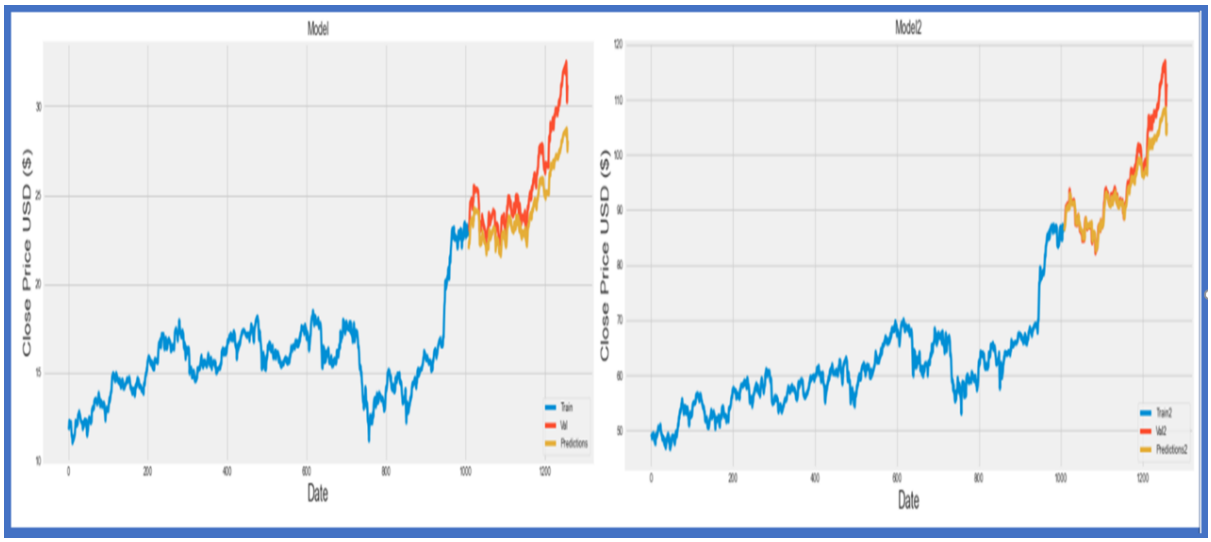


Figure 10: RNN Actual Vs Predicted for BAC and JPM

process, RNN layers were set up and an 80-20 train to test ratio was established for the data on 50 time steps. The reason samples were not chosen was to ensure scalability of

the model in case further elements were to be added. To avoid model over fit, dropout mechanism for regularisation has been implemented along with root mean squared error to evaluate how close the values foretasted are as compared to the actual values. While attempts to increase the epochs to 200 were made, this was not possible due to the overfit in terms of the loss values estimated and the process had to be dropped on an sector level, however, at the company level, this was scalable to 500 epochs successfully. Now, the next step is to compare the predicted stock behaviour to the actual values show in figure 10 above for Bank of America and JP Morgan chase investment bank.

The model depicted overall high RMSE value for training to testing data, meaning that despite of all the attempts made using dropout mechanism and layering the RNN multiple times, there was a high overfit in the sample data specially in the sector based index.The accuracy's of prediction have been calculated for the Technology and Financial sector shown in table 2.

Table 2: RNN - RMSE For INDEX

| **Sector** | **Companies** | **Train Loss** | **Test RMSE** | **Accuracy** | **Epoch** |
|---|---|---|---|---|---|
| Technology | 102 | Loss:0.0071 | 37.95 | 62.95% | 50 |
| Financial | 139 | Loss:0.0236 | 40.78 | 59.22% | 50 |

Below table 3 shows the RMSE value captured for the APPLE, Amazon, Bank of America and JP Morgan Chase.

Table 3: RNN - RMSE For Companies

| **Company** | **Train Loss** | **Test RMSE** | **Accuracy** | **Epoch** |
|---|---|---|---|---|
| Apple | Loss:3.0361e-04 | 10.58 | 89.15% | 200 |
| Amazon | Loss:9.6011e-05 | 19.09 | 80.91% | 100 |
| Bank Of America | Loss:1.9023e-04 | 1.648 | 98.35% | 500 |
| JP Morgan Chase | Loss:1.5443e-0 | 2.759 | 59.22% | 500 |

In comparison to the previous market research models for RNN which involved simply using stock price values on a company level and attained an accuracy of 81.91%, in compare to that this research got nearly similar result but for sector index obtained very poor accuracy. This section marked of the completion of objective 5a.

## 4.5 Implementation, Results and Evaluation of LSTM

Due to high dropout rates and low accuracies of RNN, LSTM has been implemented.The RMSE values obtained by the LSTM Model for companies and sector based is shown in table 4 and 5

Table 4: LSTM - RMSE For INDEX

| **Sector** | **Companies** | **Train Loss** | **Test RMSE** | **Accuracy** | **Epoch** |
|---|---|---|---|---|---|
| Technology | 102 | loss: 0.0238 | 40.915 | 59.084% | 80 |
| Financial | 139 | loss: 0.0071 | 23.632 | 76.367% | 60 |

Table 5: LSTM - RMSE For INDEX

| Company | Train Loss | Test RMSE | Accuracy | Epoch |
|---|---|---|---|---|
| Apple | loss: 2.2313e-04 | 2.679 | 97.320% | 509 |
| Amazon | loss: 1.5426e-04 | 20.80 | 79.199% | 100 |
| Bank of America | loss: 1.6794e-04 | 0.758 | 99.24% | 400 |
| JP Morgan | loss: 1.8923e-04 | 2.380 | 97.619% | 200 |

The output of the graph depicting predicted values vs actual values is shown in figure 11.

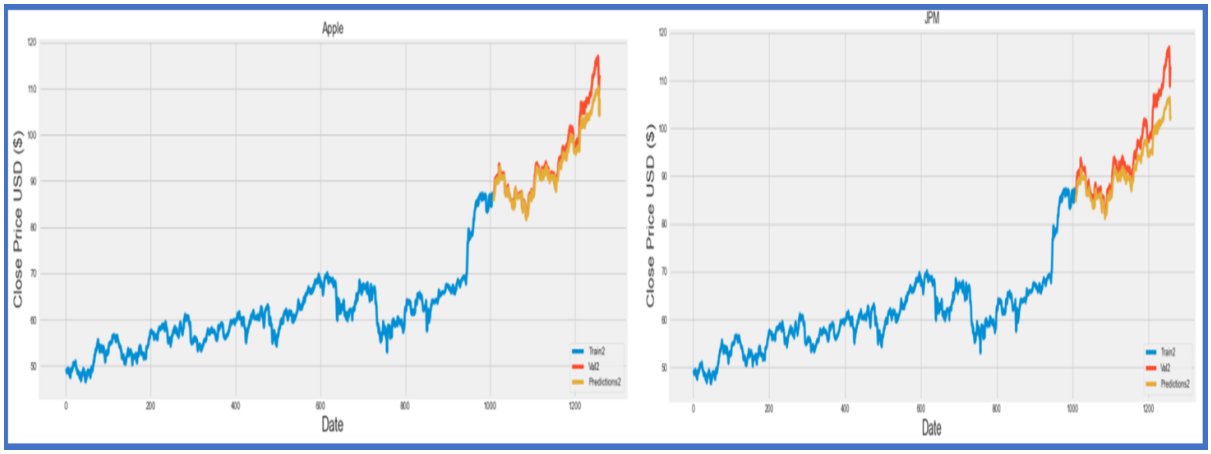


Figure 11: LSTM Actual Vs Predicted APPLE and JPM

This model has implied a significant increase to the RNN approach for sector index based models as as well individual companies. LSTM standards accuracy, which were previously attained at 84%, but in this research LSTM model Accuracy on average 93.34% . Though it was a small data set ,but it's worth mentioning this accuracy. There is a still scope of improvement the sector based model where accuracy achieved mere 67.7%. This section marked the completion of objective 5b.

## 4.6 Implementation, Results and Evaluation of ARIMA Time Series

To implement time series forecasting, a slightly different approach has been used. The first step here was to analyse the differences between opening and closing prices in a 3-day time interval for 4 years (2015- 2017). This has been shown below in figure 12.

Then, to implemented Time Series ARIMA (Auto regression integrated moving average) , regressor values based on previous closing values of the time duration's were selected (4 years). It is widely used time series and it makes use of lagged moving average to smooth the time series data. However he next step was to include the parameters of testing and training by splitting the data and implementing x and y tests to find out how close the data was to predicted values, followed initiating the training mechanism.

Table 6 shows the RMSE and $R\hat{2}$ attained by ARIMA time series for the individual companies as well as sector based index in industries involved.

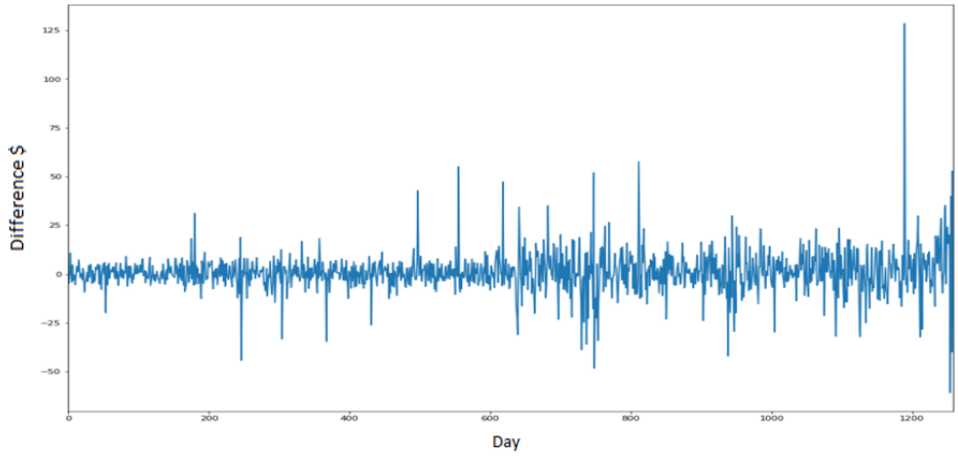


Figure 12: Opening  Closing Price Differences in 3 day Interval

Table 6: Arima Time Series evaluation Matrix for the Sector

| **Sector** | **Companies** | **R square Value** | **RMSE** | **Accuracy** |
|---|---|---|---|---|
| Technology | 102 | -1.495 | 6.215 | 93.75% |
| Financial | 139 | 0.297 | 5.26 | 94.74% |

Table 7 shows the same metrics for the Apple, Amazon , Bank of America and JP Morgan chase.

Table 7: Arima Time Series evaluation Matrix for the companies

| **Companies** | **R squared Value** | **RMSE** | **Accuracy** |
|---|---|---|---|
| Apple | -0.049 | 7.86 | 92.14% |
| Amazon | -5.07 | 22.329 | 77.67% |
| Bank of America | -.083 | 1.253 | 98.74% |
| JP Morgan Chase | 1.05 | 4.6 | 95.40% |

While only this model provided the best performance on an sector based index as compared to RNN and LSTM, the existing threshold for comparison by Time Series ARIMA model was set at a high accuracy of 94.75%. while the technology index for this does average out to 94.25% for the first sector types.The company-based data has led to a significant decrease in the overall accuracies predicted for the model, averaging out to 90.5%. This section concludes objective 5c.

## 4.7 Comparison of Developed Models with existing Models

After all the model implementation we can conclude that LSTM gave the on average highest accuracy for individual companies, 93% as compare to previous researcher 84% Pawar et al. (2019). whereas Arima is the only model, it gave highest accuracy for sector based index in compare to other model in this research .Below is the detail summary in figure 13 and figure 14. Previous ARIMA model has obtained 96% accuracy Mehtab and Sen (2020) However , still there is a room for increase accuracy for sector based model.

| Sector | Test RMSE | Developed Model Accuracy | Existing Model Accuracy |
|---|---|---|---|
| Technology | 37.95 | 62.95% | Selvinet al. (2017)82% |
| Finacial | 40.78 | 59.22% | Selvinet al. (2017)82% |
| Avg | 39.365 | 61% | Selvinet al. (2017)82% |

| Company | Test RMSE | Developed Model Accuracy | Existing Model Accuracy |
|---|---|---|---|
| APPLE | 10.58 | 89.15% | Selvinet al. (2017)82% |
| Amazon | 19.09 | 80.91% | Selvinet al. (2017)82% |
| BAC | 1.648 | 98.35% | Selvinet al. (2017)82% |
| JPM | 2.759 | 59.22% | Selvinet al. (2017)82% |
| Avg | 8.51925 | 82% | Selvinet al. (2017)82% |

**LSTM**

| Sector | Test RMSE | Developed Model Accuracy | Existing Model Accuracy |
|---|---|---|---|
| Technology | 40.91 | 59.08% | Pawaret al. (2019)84% |
| Finacial | 23.632 | 76.37% | Pawaret al. (2019)84% |
| Avg | 32.271 | 68% | Pawaret al. (2019)84% |

| Sector | Test RMSE | Developed Model Accuracy | Existing Model Accuracy |
|---|---|---|---|
| APPLE | 2.679 | 97.32% | Pawaret al. (2019)84% |
| Amazon | 20.8 | 79.20% | Pawaret al. (2019)84% |
| BAC | 0.758 | 99.24% | Pawaret al. (2019)84% |
| JPM | 2.38 | 97.61% | Pawaret al. (2019)84% |
| Avg | 6.65425 | 93.34% | |

**ARIMA**

| Sector | Test RMSE | Developed Model Accuracy | Existing Model Accuracy |
|---|---|---|---|
| Technology | -1.495 | 93.75% | Idrees et al.(2019) 96% |
| Finacial | 0.297 | 94.74% | Idrees et al.(2019) 96% |
| Avg | -0.599 | 94% | Idrees et al.(2019) 96% |

| Sector | Test RMSE | Developed Model Accuracy | Existing Model Accuracy |
|---|---|---|---|
| APPLE | 7.86 | 92.24% | Idrees et al.(2019) 96% |
| Amazon | 22.32 | 77.67% | Idrees et al.(2019) 96% |
| BAC | 1.253 | 98.74 | Idrees et al.(2019) 96% |
| JPM | 4.6 | 95.4 | Idrees et al.(2019) 96% |
| Avg | 9.00825 | 91% | |

Figure 13: Summary Result

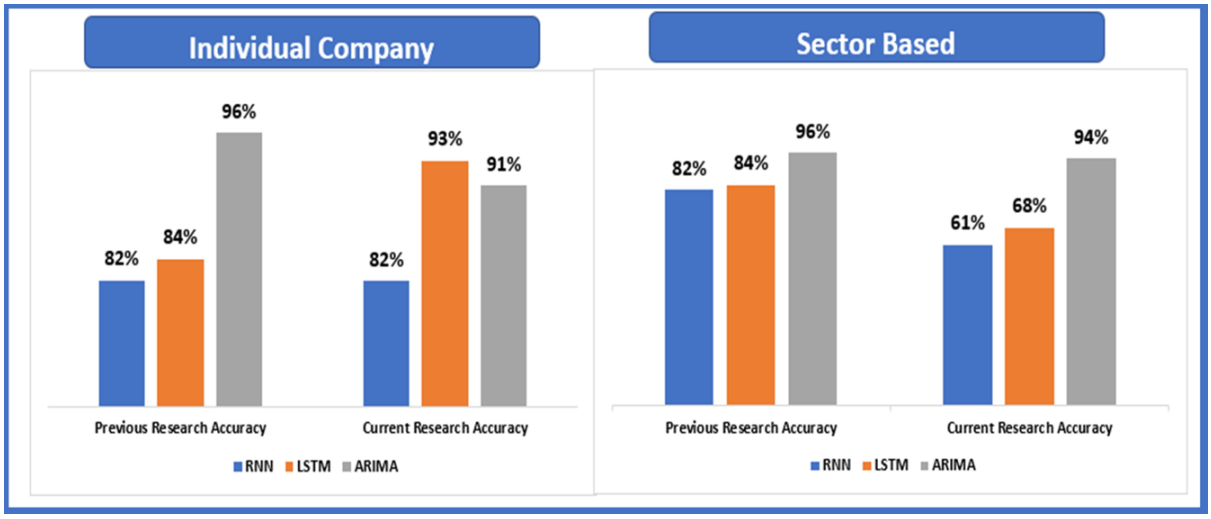


Figure 14: Summary of Average Accuracy

once sector based prediction improve it will give the investor another safety net for relatively safer investment.This means that company's alignment with the index would provide a security factor to the investors who choose to invest in these companies based on the performances of the sector. To conclude illustrate how this index performs in comparison to other companies in the same sector, the companies with top revenues in the domain were chosen and compared to the index. The output of individual company performances shown as well as the sector the company in figure 15. This is for the

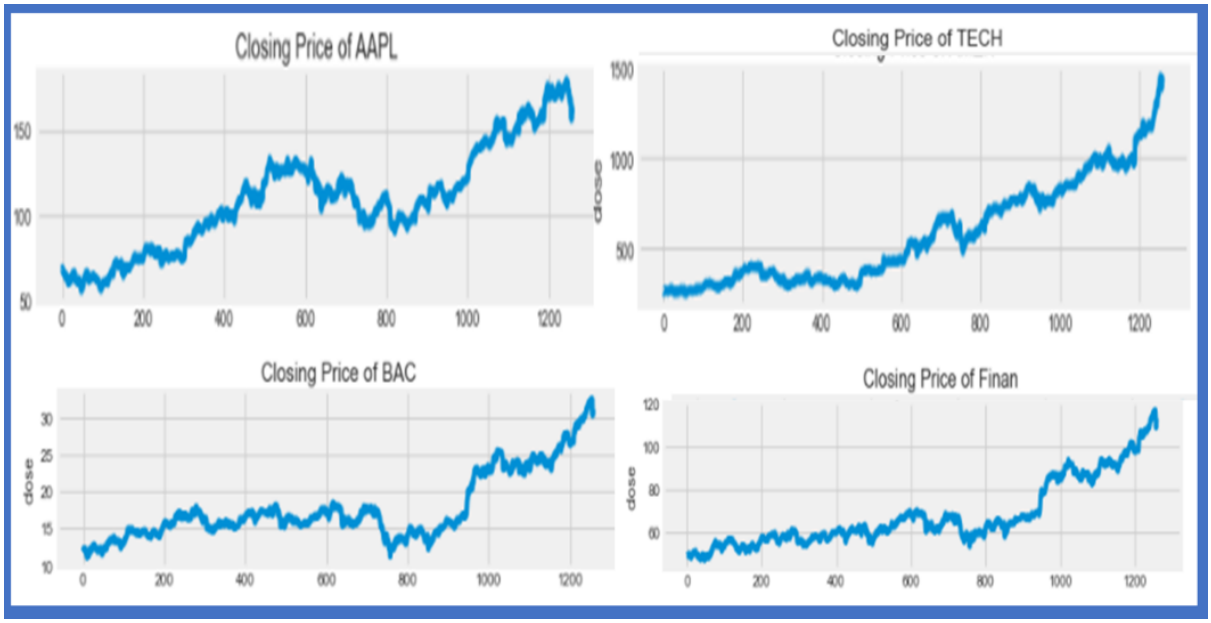


Figure 15: Comparison of Companies Closing Price with INDEX

users to understand how much the sector-based market is behaving differently from the individual companies in question. This process allows the clients to compare as many companies as they would prefer to the index and invest accordingly along with an accuracy measure of the industries chosen (limited to 8 for this research).with this,it has solved and implemented the research question.

# 5 Conclusion and Future Work

The research implemented established an index based on closing values throughout sector with the top eight chosen in terms of sample size of companies. This covers the objectives list stated in the beginning of the research. The outputs of the research set up a narrative for a future research where the index models need to be improved. With RNN indices scoring a low average accuracy of 81.91% for individual companies and the organisation level accuracies being relatively high for the small data sample (4 companies) evaluated, there could be a doubt on whether the index-based approach is preferable in these scenarios or not. Another confirmation on this came once the results for LSTM were established, the index improved significantly with an average accuracy of 67.7%, but this was further violated by an even higher accuracy attained for all the individual companies observed under the spectrum chosen for technology. Time series ARIMA, on the other hand, provided the best results in terms of accuracy for the index with an average

accuracy of 94.25%, which was actually higher than the time series prediction accuracies for some of the companies, leading to this approach proving more beneficial for the stake holders as compared to using neural networks. All the techniques implemented have a great accuracy in contrast with previous researches pertaining to the performance of the models used, specially LSTM Average accuracy 93% compare to previously achieved 84% Pawar et al. (2019). This adds to the current value of the research by providing a safety net to the investors in the form of sector index where they can now compare the performances of their chosen companies to see how they differ from the indices.

Another major difference between the time series approach as compared to RNN and LSTM is that this is not reliant on the closing values of the industry prices by dates, but on the differences between opening and closing values over certain time intervals (initially set to 3 days). To conclude the research semantically, the results, while accurate for time series, do not provide concrete evidence to the high functionality of the model for modern day implications. There are several aspects leading to this conclusion such as the lack of variables in the data with only opening and closing values being the measures of understanding. A short time span for the research led to it being impossible for expanding the results to include a larger data sample and get a more concrete view of the index performance as compared to individual company performances. The sector based segregation was only limited to primary income factors for companies such as Amazon, where other income sources also cover a major aspect of stock investments by stake holders. While this research is limited to sector based information, the dataset is capable of expanding further and incorporating industry-based information to include something like a correlation matrix to understand industry type information and superimpose that on company-based outputs. This research opens up horizons for the research of stock price prediction from a different aspect and with improvement in index establishment (be it industry based, sector based or any other new innovation based), the otherwise risky stock investment patterns could be made a little more predictable and less risky for the investors.

**Future Works** The research laid the groundwork towards a new parameter for stock investments by designing a benchmark in the form of sector based indices. This provides the new investors a basic understanding on what to expect from the companies which fall under their chosen indices and how the companies they have chosen to perform as compared to the index ratings. To continue researching in this domain, the researchers will have to focus on improving the index which is currently in a crude condition with the only understanding of segregation coming in the form of the sector chosen. This research can then further be granulated in the form of including industry based information along with industry based indices which would help understanding the process for improving the indices by adding in more data points to the research and the models. Another aspect for improvement here would be to incorporate stock price data of more companies, considering we already have the industry types and sector types coming in from over 1000 companies and the ACT codes which are already in a usable manner as well. The research has the scope of expanding into more countries if a directory of information and data is implemented and investors are allowed to add in companies, they would like to see be a part of the indices already in place. This would not only add more data points to the research but also allow the research a better understanding on what various aspects can affect the values of the indices in future. The research depends on an ever-improving knowledge based which relies on index improvement and hence, there would be a scope for improvement till the point company based results align with the index based results

and all the companies which are outliers can be classified and removed, to provide the users with a complete risk free investment approach.

# 6 Acknowledgement

I would like to dedicate this research to my beloved children as they missed so much playtime with me while i was busy with doing this research.I Am also indebted to my beloved parents and friends for their constant motivation and encouragement. I am grateful to my wife for her support . I also would like to extend my gratitude to Dr. Catherine Mulwa for her advise and guidance through out the Research.

# References

Barak, S., Arjmand, A. and Ortobelli, S. (2017). Fusion of multiple diverse predictors in stock market, *Information Fusion* **36**: 90–102.

Borovkova, S. and Tsiamas, I. (2019). An ensemble of lstm neural networks for high-frequency stock market classification, *Journal of Forecasting* **38**(6): 600–619.

Grigoryan, H. et al. (2015). Stock market prediction using artificial neural networks. case study of tal1t, nasdaq omx baltic stock, *Database Systems Journal* **6**(2): 14–23.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors, *arXiv preprint arXiv:1207.0580* .

Hoseinzade, E. and Haratizadeh, S. (2018). Cnnpred: Cnn-based stock market prediction using several data sources, *arXiv preprint arXiv:1810.08923* .

Idrees, S. M., Alam, M. A. and Agarwal, P. (2019). A prediction approach for stock market volatility based on time series data, *IEEE Access* **7**: 17287–17298.

Kamiński, B., Jakubczyk, M. and Szufel, P. (2018). A framework for sensitivity analysis of decision trees, *Central European journal of operations research* **26**(1): 135–159.

Khaidem, L., Saha, S. and Dey, S. R. (2016). Predicting the direction of stock market prices using random forest, *arXiv preprint arXiv:1605.00003* .

Lubis, A. H., Ikhwan, A. and Kan, P. L. E. (2018). Combination of levenshtein distance and rabin-karp to improve the accuracy of document equivalence level, *International Journal of Engineering & Technology* **7**(2.27): 17–21.

Mehtab, S. and Sen, J. (2020). A time series analysis-based stock price prediction using machine learning and deep learning models, *International Journal of Business Forecasting and Marketing Intelligence* **6**(4): 272–335.

Pawar, K., Jalem, R. S. and Tiwari, V. (2019). Stock market price prediction using lstm rnn, *Emerging Trends in Expert Applications and Security*, Springer, pp. 493–503.

Ponnam, L. T., Rao, V. S., Srinivas, K. and Raavi, V. (2016). A comparative study on techniques used for prediction of stock market, *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, IEEE, pp. 1–6.

Qiu, M. and Song, Y. (2016). Predicting the direction of stock market index movement using an optimized artificial neural network model, *PloS one* **11**(5): e0155133.

Robischon, N. (2017). Why amazon is the world's most innovative company of 2017, *Fast Company Magazine* **2**.

Seker, S. E., Mert, C., Al-Naami, K., Ayan, U. and Ozalp, N. (2013). Ensemble classification over stock market time series and economy news, *2013 IEEE International Conference on Intelligence and Security Informatics*, IEEE, pp. 272–273.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K. and Soman, K. (2017). Stock price prediction using lstm, rnn and cnn-sliding window model, *2017 international conference on advances in computing, communications and informatics (icacci)*, IEEE, pp. 1643–1647.

Shah, D., Isah, H. and Zulkernine, F. (2018). Predicting the effects of news sentiments on the stock market, *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 4705–4708.

Waqar, M., Dawood, H., Guo, P., Shahnawaz, M. B. and Ghazanfar, M. A. (2017). Prediction of stock market by principal component analysis, *2017 13th International Conference on Computational Intelligence and Security (CIS)*, IEEE, pp. 599–602.

Weng, B., Lu, L., Wang, X., Megahed, F. M. and Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources, *Expert Systems with Applications* **112**: 258–273.

Wu, F., Zhao, W.-L., Ji, Q. and Zhang, D. (2020). Dependency, centrality and dynamic networks for international commodity futures prices, *International Review of Economics & Finance* **67**: 118–132.