

ICT FINAL REPORT for

Project 20 - “MAchine learnig & algoritms”

By Aakib, Manjeet, Andy & Calvin

**Sponsored by Dr Shuangzhe Liu & Dr Dat Tran**

Table of Contents

[ITS Capstone Project Sponsor signoff 3](#_Toc54970841)

[1) INTRODUCTION: 4](#_Toc54970842)

[2) ABOUT MACHINE LEARNING METHOD AND DEEP LEARNING METHOD: 5](#_Toc54970843)

[A. Machine Learning: 5](#_Toc54970844)

[B. Deep Learning 5](#_Toc54970845)

[3) SCOPE OF THE PROJECT: 6](#_Toc54970846)

[In-Scope: 6](#_Toc54970847)

[4) LIST OF RESOURCES UTILIZED: 7](#_Toc54970848)

[5) DELIVERABLES: 8](#_Toc54970849)

[A. Data Visualization: 9](#_Toc54970850)

[B. Data Pre-processing Techniques: 11](#_Toc54970851)

[C. Simple Linear Regression: 12](#_Toc54970852)

[A) What is simple linear regression? 12](#_Toc54970853)

[B) Why is Regression Model Important? 12](#_Toc54970854)

[C) 5 EASY STEPS HOW WE BUILD OUR REGRESSION MODEL: 12](#_Toc54970855)

[D) Findings of Linear Regrssion: 13](#_Toc54970856)

[D. LONG SHORT-TERM MEMORY(LSTM): 14](#_Toc54970857)

[A) What is LSTM model? 14](#_Toc54970858)

[B) Why LSTM? 15](#_Toc54970859)

[C) 5 easy steps how we build our lstm model: 15](#_Toc54970860)

[D) Findings of lstm model: 15](#_Toc54970861)

[E. ARIMA MODEL: 16](#_Toc54970862)

[A) What is ARIMA model? 16](#_Toc54970863)

[B) WHY ARE WE CHOOSING ARIMA MODEL? 17](#_Toc54970864)

[C) 5 easy steps how we build our lstm model: 17](#_Toc54970865)

[D) FINDINGS OF ARIMA MODEL: 18](#_Toc54970866)

[F. DECISION TREE: 19](#_Toc54970867)

[A) What is DECISION TREE model? 19](#_Toc54970868)

[B) WHY DID WE CHOOSE DECISION TREE: 20](#_Toc54970869)

[C) 5 easy steps how we build our lstm model: 20](#_Toc54970870)

[D) Findings of decision Tree: 21](#_Toc54970871)

[G. SUMMARY OF THE RESULTS: 22](#_Toc54970872)

[H. FORECAST: 23](#_Toc54970873)

[I. CONCLUSION: 24](#_Toc54970874)

[6) OUTSTANDING ISSUES: 25](#_Toc54970875)

[7) RISK MANAGEMENT: 26](#_Toc54970876)

[8) LESSONS LEARNED: 27](#_Toc54970877)

[9) HANDOVER THE MATERIALS TO THE SUPERVISOR: 28](#_Toc54970878)

[10) RECOMMEDATIONS TO THE SPONSERS / FUTURE DEVELOPMENT: 28](#_Toc54970879)

[Bibliography 29](#_Toc54970880)



# ITS Capstone Project Sponsor signoff

**Project name:** Machine learning techniques and applications

**Project id:** 20S2-20

**Team members:** Aakib Shah Sayed (U3188008), Calvin Le (U3167003), Manjeet Dahiya (U3148284), Xieyuan Huang (U3190995)

**Project Sponsor**: Dr Dat Tran & Dr Shuangzhe Liu

I  **D. Tran S. Liu**  as the sponsor of the project certify that:

The objectives of the project

☒ are fully met

☐ are partially met

☐ are not met

**Optional comments on team performance**

The team members have reviewed the literature, studied the models and analysed the data to produce good results. They have made significant progress with dedication and efforts.

A picture containing letter  Description automatically generated 22/10/2020

**Signature**  **Date**

# INTRODUCTION:



The stock market is the equity market or share market is the aggregation of buyers and sellers of stocks, which represent ownership claims on businesses. (Wikipedia, 2020)

We all know that stock is one of the interesting things that people spend on these days but due to some uncertainty some people lose money where some people gain money. The main reason behind this problem is people estimating stock value themselves. So, to come up with a solution we are going to apply the available machine learning techniques to have a data driven approach which can help people which they can rely on and have a much better estimation.

# ABOUT MACHINE LEARNING METHOD AND DEEP LEARNING METHOD:

## Machine Learning:

It is a technique to break up data, learn from the data and use it for certain type of prediction. It is a type of data mining technique that learns from the provided data and helps to make an informed decision. These days big companies like Amazon or eBay uses these techniques to better suggest user suggestion or experience while shopping. The main drawbacks of machine learning are the user must extract the required data for the available datasets and the feed into the model, if any error occur in prediction, the programmer fix it. (Yadav, 2018)

## Deep Learning

It is a process which involves feeding a computer with lots of data which it can use to make a certain decision about other data. Here the word deep means it having number of layers in the neural network which is popularly known as hidden layers. As a result, deep learning requires high computer power like GPU, and have a higher accuracy. (Ahmad, 2019)

LSTM

Decision Tree

ARIMA

Linear Regression

# SCOPE OF THE PROJECT:

*Our purpose in this project is to create different machine learning models and compare their results and predict the future using the best available model.*

And our scope of the project is as given below:

## In-Scope:

1. Collect Dataset from Yahoo Finance for about 10 years’ time of different organization.
2. Visualize the Data and find different co-relation among the above-mentioned datasets.
3. Apply 4 machine learning techniques to predict the stock of different organizations such as Qantas, CBA and Telstra.

The machine learning techniques we are going to use are:

1. LSTM (Long-Short term memory):
2. ARIMA (Auto Regressive Integrated Moving Average):
3. Decision Tree and Random Forest:
4. Linear Regression
5. Predict the future stock prices of above-mentioned companies.
6. Comparison of datasets, findings and model outcomes with possible applications.
7. Testing and debugging different models.
8. A written report where all findings, comparisons, and different model techniques are explained.
9. A python running file or a collection of Jupiter Notebook of different models created.

# LIST OF RESOURCES UTILIZED:

The list of resources utilised in our project is as listed below:

1. Laptop with GPU power.
2. Use of stock data of 15 years (2005-2020) for companies including Biogen, Quanta’s, Sydney, Telstra, Commonwealth Bank
3. Use of Anaconda software and the list of default packages available.
4. Use of Jupyter notebook.
5. Use of some of the popular libraries including:
   1. NumPy for working with arrays
   2. Pandas for data manipulation and analysis
   3. TensorFlow
   4. Seaborn for visualisation
   5. Matplotlib for visualization
   6. Keras for LSTM model
   7. Sklearn for pre-processing
   8. Os for saving model content
   9. Yfinance for collecting data
   10. Stats for ARIMA model

# DELIVERABLES:

A list of briefs about our deliverables are:

1. Collect the dataset of 5 different organization included in different fields like Biogen in medical, Telstra in medical and others in banking and transportation
2. Interpret the dataset with some graph and some latest news data.
3. Pre-process the dataset.
4. Build all 4 models for all 5 datasets.
5. Get the results and find the error matrix for each model.
6. Compare the error matrix and find some of the best model.
7. Interpret the result.
8. Come up with a conclusion.

Now, we will dive deeper into the deliverables, where we will show you some of our best model results and summarize them up for the purpose of simplicity in this report.

## Data Visualization:

**Line Graph:**

Background pattern

Description automatically generated**Figure: Closing Stock Price of different datasets applied in our models.**

The main point to understand from this graph is that BIIB the top line graph has its value ranging from A$ 33 to A$ 475 while others have it less than A$ 10 except CBA which means higher variability in data which machine needs to learn.

**Biogen Inc:**

The Biogen Inc stocks starts at A$ 64.7 in 1st Jan 2005 and ends at A$ 274.69 on 31st July 2020. We can see there is a high gap between the starting point and the end we have collected. The main reason behind the company hype in stock price is they recruited a new CEO George Scangos, who turns the company in his two years’ time by reducing the costs and job and redirecting the company into smarter marketing in cancer research and development which helped increased the share price by 212%. The highest stock price for Biogen is A$ 475.98 during this period. (Herper, 2012)

**Commonwealth Bank:**

The Commonwealth Bank stock started at A$ 31.9 in 1st Jan 2005 and ended at A$ 72.48 on 31st July 2020. There is a slight rise in the stock market because it has been able to expand its market to customer across different nations as well as acquired different partnerships and earned the interest of investors. The highest stock price for this Bank is A$ 95.8 during this period. (Newman, 2015)

**Quanta’s Airline:**

The Quanta’s Airline stock started at A$ 3.9 in 1st Jan 2005 and ended at A$ 3.2 on 31st July 2020. As from the graph the Quanta’s Airline had a loss in 2012 and 2014 which was mainly due to rise of fuel prices, intense competition, and industrial disputes. In 2013 we can see a slight rise as they alliance with Emirates to offer carrier flights to Dubai. Later, in 2015 it sold its lease at Sydney Airport for A$535 million. (Quantas, 2020)

**Sydney Airport:**

The Sydney Airport stock started at A$ 3.125 in 3rd Jan 2005 and ended at A$ 5.23 on 31st July 2020. The highest stock price for Sydney Airport was A$9.2 within this time period during late 2019 and because of COVID-19 its stocks has fallen below but due to continuous renovation and expansion and alliance it was able to reach at that level. (Sydney, 2020)

**Telstra:**

Telstra stock shared started at A$ 4.184 in 3rd Jan 2020 and ended at A$ 3.35 on 31st July 2020. The highest stock price for Telstra was A$ 6.67. The reason behind Telstra climbing the stock market from 2012 to 2016 because it started addressing customer problem and started solving their issues. Later again in 2018 investors sold the shares which declined its dividends and caused a fall in market. Finally, in 2020 as of COVID-19 virus its market value again raised by 1.8%. (Mickleboro, 2019)

## Data Pre-processing Techniques:

There are different sets of pre-processing techniques available for different sets of data applied. It can be clearly understood the different models need different pre-processing techniques.

So, some of the popular method we applied for the dataset we collected are:

1. Taking things in consideration we have applied “dropna ()” function which helps to eliminate the null values from the Data Frame.
2. Use of Min Max scaler which helps to reduce the values in range between 0 and 1 which is necessary in regards with LSTM model.
3. Use of difference between datasets in ARIMA model to stationary the dataset.
4. We are shifting one row of the close price in the decision tree to create the true value of the target.
5. Use of reshape function to fit into the model.
6. Using round function to minimize the error rate.
7. Selecting only the required close columns.
8. Data split of 70 % train and 30 % test applied and 80% train and 20% test.

## Simple Linear Regression:

### What is simple linear regression?

Simple Linear Regression is used to find the relationship between two quantitative variables. It allows us to estimate how a change in independent (x) variable can bring a change in dependent variable (Y). There are various types of regression as per the requirement like multi-variable regression or logistic regression but here we will be applying a simple linear regression which is indicated by a straight line.

The formula for simple linear regression is as:

where y is the predicted value of the dependent variable x.

*B0 is the intercept*

*B1 is the regression coefficient*

*X is the independent variable influencing y*.

It is observed that the linear the Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE. (Bevans, 2020)

### Why is Regression Model Important?

Our objective of choosing the regression model was simple. For our group we understood that Regression model is one of the easiest models to apply for and it helps to identity if a dependent variable (x) has any affect in the independent variable (Y).

It is important in a business or any other datasets to get a valuable insight, actionable insight which can help the business. (Foley, 2018)

### 5 EASY STEPS HOW WE BUILD OUR REGRESSION MODEL:

The five easy steps how we builded our Regression model is as follows:

1. Choose the relevant section of data which is Closing price.
2. Import the necessary package from python known as “Sci-kit” learn.
3. Apply the Linear Regression.
4. Fit it into the model.
5. Get Results.

### Findings of Linear Regrssion:



Figure: A Linear Regression being fit into our Telstra dataset.

Error matrix for different dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| Company | MAE | MSE | R2 |
| BIIB | 42.53 | 2990.85 | 0.78 |
| CBA | 7.46 | 86.66 | 0.68 |
| QAN | 1.33 | 2.49 | 0.02 |
| SYD | 0.95 | 1.12 | 0.70 |
| TLS | 0.81 | 0.88 | 0.019 |

In the above table we calculated the MAE, MSE, and R2 for all the 5 different datasets. The MAE represents the absolute value difference between the forecasted value and the actual value, MSE represents the measure of average square of errors which is difference between estimated values and actual value Here, the R2value also indicates the goodness of fit into our model, the higher the R2 value the better but as R2 has certain limitations as a good model can have a low R2 value while a bad model can have a good R2 value.

According to the table above it can be clearly seen that BIIB has a higher error matrix than any other dataset, the main cause behind this could be as it has higher variance in data which starts from A$ 33 and ends at A$ 475. It also states that the difference between the actual price and the predicted price is lot higher which can vary A$ 42 or more while forecasting, and the average square of error between them can range from 2990 but the R2 is higher means the dataset of BIIB is more similar between actual value and predicted value.

Likewise, having a closure look at TLS dataset which has the lowest error matrix as its closing price variance A$ 2.5 to A$ 6.67. It MAE is only about 0.82 which means there is not much actual difference between the original price and the forecasted price and also the MSE and RMSE is also low which are 0.88 and 0.94 where the average error could be of 0.88 and the dataset is almost near the best fit line with a R2 of 0.011 which means the Y variable is not dependent of the X variable.

Furthermore, we applied a special normalized case for SYD data which gave a rise to error matrix which at moment is around 1.11 is rise to 5.1 so we also found that normalization is not a step to attempt for linear regression models so we did not applied it to any other models.

Also, from the above table it can be clearly seen that, linear regression is not suitable for different varieties of data like BIIB as it has higher number of MSE cause the value of closing price for BIIB goes from A$ 33.34 to A$ 475.98 but if the dataset values does not have high variance than the linear regression model fits better with MSE ranging from 2.53 to 0.88 which can be difference range between the predicted price and the actual price.

Here, if MAE & MSE less than 1 than it is best and if R2 ~ 1 to then it is better.

## LONG SHORT-TERM MEMORY(LSTM):

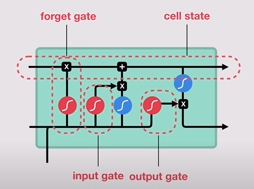
### What is LSTM model?

Long short-term memory (LSTM) is a special machine type of algorithm which is capable of learning long-term dependencies. LSTM networks are a type of Recurrent Neural Network (RNN) to predict problems in a sequence progressing. This network was introduced by Hochreiter and Schmidhuber in 1997 to work on a variety of problems. (Olah, 2015)

The logic behind LSTM is a memory cell known as a “cell state” which keeps its state over time.

LSTM is designed for application for example if the input is an ordered sequence then information from earlier in the sequence, may or may not be important to predict the outcome.

LSTM uses the output from a previous step as an input for the next step. (Yoshua Bengio, 1994)



**Figure: Showing one input and output gate of LSTM with its function.**

These gates are used to regulate the flow of information which is then used if the data is important to keep or throw away.  This is used to predict.

### Why LSTM?

It is said that LSTM is one of the best models for time series analysis and has a great capacity to train the model accurately. As it has the power of hidden layers which are available to attach which really helps to learn about the data fed into it and understand the links between the data.

As this model works on layers basis where different number of hidden layers can be added as per our requirement plus different types of gates which determines to proceed forward or not to another layers.

Also, LSTM is one of the new techniques in deep learning and has more control ability.

### 5 easy steps how we build our lstm model:

1. Choose the relevant section of data which is Closing price.
2. Import the necessary package from python known as “Keras and TensorFlow”.
3. Apply the different pre-processing technique.
4. Fit it into the model by adding different hidden layers.
5. Get Results.

### Findings of lstm model:



Figure: A section of Telstra dataset where our LSTM model is fitted between predicted value (blue) vs real stock value (red)

LSTM model test with condition:

80/20 data split and Epochs: 100, Layers: 4

|  |  |  |  |
| --- | --- | --- | --- |
| Company | MAE | MSE | R2 |
| BIIB | 9.97 | 218.48 | 0.86 |
| CBA | 1.52 | 4.39 | 0.89 |
| QAN | 0.17 | 0.06 | 0.94 |
| SYD | 0.19 | 0.082 | 0.92 |
| TLS | 0.06 | 0.006 | 0.94 |

Having a closer look at the above table it can be clearly seen that BIIB in have the highest number of Error matrix which is verified as it has the higher variance of data in amount of dollars whereas Telstra is performing better as it has lowest amount of data variance it has throughout the time period we have selected.

In conclusion, it can be seen that LSTM model is performing at its best where the actual value with the predicted value fit range from 80% or above for all the different varieties of dataset where its lowest for BIIB with MAE at 9.97 which is the price difference that can occur between actual value and predicted value, MSE 218.48 which is the square of errors it can have and R2 if 0.86 which is the goodness of fit between the actual testing dataset and the predicted dataset. So, this concludes that LSTM models are pretty good in time series if there is not much variance in the dataset.

## ARIMA MODEL:

### What is ARIMA model?

ARIMA stands for Autoregressive integrated moving average. ARIMA stands for Autoregressive integrated moving average. An ARIMA model constitute 3 parameters which are p (the order of AR model), d (the number of differencing to make time series stationary) and q (the order of MA model).

Auto Regressive means that the model uses the dependent relationship between an observation and some predefined number of lagged observations. (also known as “time lag” or “lag” of Yt).

Yt-1 is the first lag of the time series. is the coefficient of the first lag estimated by the model. is the constant value estimated by the model?

Moving Average means that the model exploits the relationship between the residual error and the observations.

€t is the white noise. Фi is the coefficient of the error in a specific lag estimated by the model.

In a simple word, Predicted Yt = Constant + Linear combination Lags of Y (up to i lags) + Linear Combination of Lagged forecast errors (up to i lags)

(ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+, 2020)

### WHY ARE WE CHOOSING ARIMA MODEL?

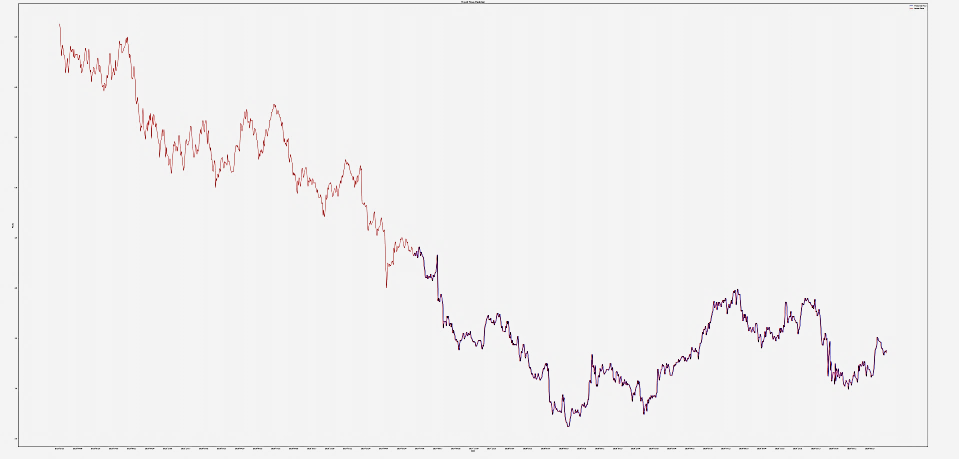
The list of reason we are choosing the ARIMA model are:

1. It is one of the best models in time-series analysis.
2. ARIMA models are a subset of linear regression models that attempt to use the past observations of the target variable to forecast its future values in time series analysis.
3. It is more flexible than other statistical model such as simple linear regression.
4. The data can be modified into stationary data which meets the requirement of ARIMA Model.

### 5 easy steps how we build our lstm model:

1. Choose the relevant section of data which is Closing price.
2. Import the necessary package from python known as “Stats”.
3. Apply the different pre-processing technique.
4. Fit it into the model by choosing the relevant p, d, q value.
5. Get Results.

### FINDINGS OF ARIMA MODEL:



**Figure: ARIMA model being fit into Telstra dataset where blue is the predicted value and the red is the actual value.**

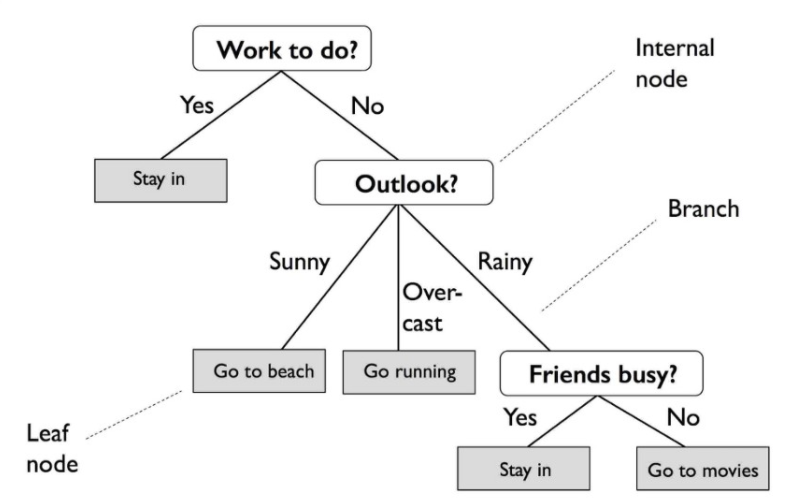
|  |  |  |  |
| --- | --- | --- | --- |
| Company | MAE | MSE | R2 |
| BIIB | 4.426 | 59.12 | 0.958 |
| CBA | 0.726 | 1.230 | 0.968 |
| QAN | 0.084 | 0.013 | 0.985 |
| SYD | 0.084 | 0.014 | 0.981 |
| TLS | 0.035 | 0.002 | 0.980 |

From the above table it can be found that ARIMA (1,1,0) is better model as it has the minimum value for the error matrix generated and higher value of goodness of fit which are less than zero for 4 dataset except for BIIB which is 4.426 MAE as it interprets as A$ 4.46 could be the difference between the actual value and the predicted value. Similarly, MSE is 59.12 with an R2 (goodness of fit) around 0.95. Likewise, Telstra on the dataset is performing at its best which is 0.996 R2 and the error matrix MAE 0.035 and MSE 0.002 between the actual and the predicted value. An important point to notice here is the less the value of MAE which is the actual difference between the real value of stock and the predicted value of the stock the lesser the MSE gets whereas if higher the value of MAE the higher the MSE gets. So, one of the best performing models is the ARIMA (1,1,0) in our dataset.

## DECISION TREE:

### What is DECISION TREE model?

Decision tree is a supervised machine learning model used to predict target by learning decision rules from features (Li, 2020). It has been widely used in classification problem such as the graph shown below. In classification problem, splitting the node by maximising the information gain or minimising the Gini impurity.



**Figure: (Li, 2020)**

In stock price prediction, the model is expected to predict a certain price of a stock in the future, which is not classification problem, but regression problem.

Three criterions on measuring the quality of a split in decision tree regression model are mean squared error, mean squared error with Friedman’s improvement and mean absolute error. To be more specific, mean squared error is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node. Friedman MSE uses mean squared error with Friedman’s improvement score for potential splits and mean absolute error minimizes the L1 loss using the median of each terminal node (sklearn.tree.DecisionTreeRegressor — Scikit-learn 0.23.2 documentation, 2020).

In this project, weighted mean squared error was selected as the criterion of the model.

Nt can be considered as the number of training samples at node t, Dt is the training subset at node t, y(i) is the true target value and is thee predicted value based on the features. In this project, we only introduced the close price of the stock in previous day as the feature so that we can compare with other models since they are introducing the same feature.

### WHY DID WE CHOOSE DECISION TREE:

The list of reason why we choose decision tree are:

1. It is a supervised machine learning model used to predict target by learning decision rules from features (Li, 2020).
2. It has been widely used in classification problem and regression problem.
3. It is easy to interpret, powerful and versatile model.

### 5 easy steps how we build our lstm model:

1. Choose the relevant section of data which is Closing price.
2. Import the necessary package from python known as “Sklearn.DecisionTree”.
3. Apply the different pre-processing technique.
4. Fit it into the model.
5. Get Results.

### Findings of decision Tree:



**Figure: A part of decision tree model being fitted into Telstra dataset where blue is the predicted value and the red is the real value.**

|  |  |  |  |
| --- | --- | --- | --- |
| Company | MAE | MSE | R2 |
| BIIB | 2.275 | 29.624 | 0.974 |
| CBA | 0.446 | 0.649 | 0.981 |
| QAN | 0.062 | 0.008 | 0.995 |
| SYD | 0.070 | 0.010 | 0.985 |
| TLS | 0.037 | 0.003 | 0.997 |

From the above table, it can be clearly seen that MAE for BIIB is 2.275 which means there is only 2.275 difference in value between the actual value and the predicted value. Likewise, MSE is 29.624 which shows the actual square of errors and R2 is 0.974 which is the goodness of fit between the actual and predicted value.

Similarly, for Telstra the MAE is very low as A$ 0.037 in actual stock price and the real price with MSE 0.003 and the line of best fit R2 is 0.997% which is 99.7% which is awesome.

## SUMMARY OF THE RESULTS:

We took the best performing models of each models with the best trials and now we are going to summarize the MAE and MSE with R2 for all different models and come up with the conclusion.

**MAE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Company\Models | Regression | LSTM | ARIMA | Decision Tree |
| BIIB | 42.53 | 9.97 | 4.426 | 1.609 |
| CBA | 7.46 | 1.52 | 0.726 | 0.351 |
| QAN | 1.33 | 0.17 | 0.084 | 0.068 |
| SYD | 0.95 | 0.19 | 0.084 | 0.069 |
| TLS | 0.81 | 0.06 | 0.035 | 0.035 |

**MSE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Company\Models | Regression | LSTM | ARIMA | Decision Tree |
| BIIB | 2990.85 | 218.48 | 59.12 | 23.199 |
| CBA | 86.66 | 4.39 | 1.230 | 0.453 |
| QAN | 2.49 | 0.06 | 0.013 | 0.001 |
| SYD | 1.12 | 0.082 | 0.014 | 0.010 |
| TLS | 0.88 | 0.006 | 0.002 | 0.003 |

**R2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Company\Models | Regression | LSTM | ARIMA | Decision Tree |
| BIIB | **0.78** | **0.86** | **0.958** | **0.984** |
| CBA | **0.68** | **0.89** | **0.968** | **0.988** |
| QAN | **0.02** | **0.94** | **0.985** | **0.989** |
| SYD | **0.70** | **0.92** | **0.981** | **0.987** |
| TLS | **0.019** | **0.94** | **0.980** | **0.982** |

In Conclusion, from the above table of Error Matrix it can be found that Decision Tree seems to be performing better, but the problem we found with the decision tree is that there are the possibility of data leakage between the training set and the testing set which in turn caused the model to be accurate because while training in the decision tree if we have certain range of data split then while testing and predicting it can only stick to that range of the dataset, so to avoid this limitation we have to put up certain range of testing data to help pick up the full range to forecast for the future.

Whereas in LSTM and ARIMA models there is no certain problem of data leakage and as per our available dataset and the split of 80 / 20 ARIMA is one of the best models performing at the moment as per shown in the Error matrix table. Also, ARIMA is better model if there are simple problems to tackle and if we want to understand the stock trends and we know our features of dataset well. In case of LSTM, for Deep learning it is one of the powerful models where it doesn’t not have any problem, but it needs powerful computer with GPU to execute this event which cannot be useful in every case if simple Machine Learning can solve the problem.

Finally, it depends on what our requirement are, and both the models are good with the results so one should understand what problem they are solving and select the appropriate model as per the timings and requirement satisfied.

## FORECAST:

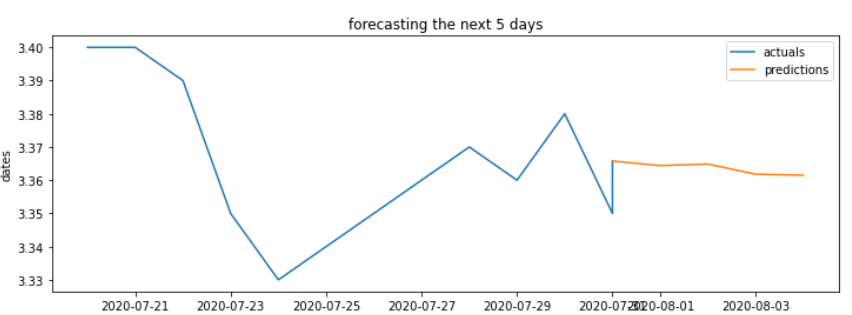


Figure: The Forecasting price of Telstra dataset for next 5 days.

|  |  |  |  |
| --- | --- | --- | --- |
| TLS | Predicted | Actual | Difference |
| 2020-07-31 | 3.365753 | 3.35 | -0.016 |
| 2020-08-03 | 3.364391 | 3.4 | 0.036 |
| 2020-08-04 | 3.364854 | 3.48 | 0.115 |
| 2020-08-05 | 3.361800 | 3.4 | 0.038 |
| 2020-08-06 | 3.361503 | 3.39 | 0.028 |

Figure: The table above is of forecasting for Telstra dataset using LSTM model.

We have used LSTM one of the deep learning method to apply into our forecast model to predict the price for the next 5 business days as shown in the table above. It can also be noticed that the difference between the predicted price and the actual price is pretty low, so it is one the good model to follow in deep learning.

## CONCLUSION:

Finally, it can be clearly seen that BIIB is performing worst because of larger variation in dataset where TLS is doing best.

To sum up, it depends on the available resources, which model is appropriate while solving a problem but by taking this project as an example there are many other different types of machine learning models that can be applied to solve various other real-life classification and regression problems which can be in Financial, Medical, Real-Estate or Astronomy sector which can have an impact in the future.

# OUTSTANDING ISSUES:

Our project does not have any outstanding issues as all the necessary steps to successfully complete this project as mentioned in the scope has been accomplished.

The steps include

1. Collecting the data from yahoo finance
2. Pre-process the data as per the requirement of the model
3. Analyze the datasets.
4. Build the model.
5. Compare the results of the model.
6. Finalize the conclusion.

# RISK MANAGEMENT:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Risk | Event description | Probability/Likelihood | Impact | Rating | Mitigation strategy | Residual Risk |
| Unclear Scope | Team cannot understand the purpose and aim of the project. | High | High | High | Communicate with the sponsor | Low |
| Not having enough data or bad data | In model building process we might not have enough data to train model properly with all verities | High | High | High | Shuffle the data so each time machine learns it learn in random way rather than in chronological order | Low |
| Scaling Problem | Once a certain model was fitted to a model same fit wont work perfectly for other dataset | Low | High | Medium | Change certain hyperparameters in the model | Low |
| Random Split problem | As we didn’t set the state of split each time running a code will produce in different split | High | High | High | Specify the random state in the model. | Low |
| Different results in each different run | As our split was random each time there was a different result from our models which made us hard to record the result for reports | High | High | High | Save it as pdf and fix the random state of the model | High |
| Crashing of computer during run | The possibility of computer crash during running of epochs in a model | High | High | High | Have a backup of code so that it can be run again after crash resolved. | High |
| Different epochs different results | While changing the number of epochs in our model each time different epoch would produce different problem. | High | High | High | Find one suitable epoch and stick with in for all different model | High |
| Overfitting of the model | This occurs when model learns a lot and produces null result | Low | Low | Low | Use of Dropout Layers to drop neurons to avoid overfitting | Low |
| Finding the adequate P, d, q value in ARIMA | There are lot of varieties of p, d, q in ARIMA model for a single dataset. | High | Medium | Low | Using the error matrix to determine which has the lowest one and following the respective one or two sets of digits | Low |
| Misinterpreting the result | Due to different error matrix having different importance different team members can interpret in different way | High | High | Low | Communicating with the team and bringing everyone in the same page | Low |

# LESSONS LEARNED:

There are many lessons we learned while accomplishing this project some of the popular ones are as mentioned below:

1. Understanding how to use Python and its codes.
2. Understood the application of machine learning.
3. Understood all the 4 model such as linear regression, LSTM, ARIMA and decision tree in detail.
4. Understood the different machine learning basic library such as Numpy, Pandas and Seaborn.
5. Understood the specific important library’s for machine learning such as s TensorFlow, Keras, Sci-kit learn.
7. Understood how to interpret results.
8. Understood the importance of communication in the team.
9. Understood the importance of pre-processing techniques and the importance of quality of data
10. Understood the importance of how visuals can help.
11. Understood how summary of results and findings can make life easier to the audience.

***Ethical Learning:***

1. Understood the sense of taking responsibility for the team and completing the task assigned by the team leaders.
2. Understanding the importance of working in the team and the importance of team.
3. Understood that being honest and communicating with the team is important.

# HANDOVER THE MATERIALS TO THE SUPERVISOR:

We handed all our environment files which contains all the necessary library to execute our code plus all the code for all 4 different models (LSTM, ARIMA, Linear Regression, Decision Tree) for all 5 different dataset (SYD, TLS, BIIB, QAN, CBA) we had to the supervisor.

# RECOMMEDATIONS TO THE SPONSERS / FUTURE DEVELOPMENT:

1. We Can have more detail varieties of testing in models which might help to find the best models having even less error rate then now in each model.
2. Use of multiple variables to train / test the dataset.
3. Use of news data following sentiment analysis for more accurate prediction.

# Bibliography

Ahmad, S. H., 2019. *AI vs Machine Learning vs Deep Learning: What the differences?.* [Online]   
Available at: https://medium.com/@sofyan\_ahmad/ai-vs-machine-learning-vs-deep-learning-what-the-differences-7f0934054479  
[Accessed 01 Sept 2020].

Bevans, R., 2020. *An introduction to simple linear regression.* [Online]   
Available at: https://www.scribbr.com/statistics/simple-linear-regression/  
[Accessed 01 Sept 2020].

Foley, B., 2018. *What is Regression Analysis and Why Should I Use It?.* [Online]   
Available at: https://www.surveygizmo.com/resources/blog/regression-analysis/Ben Foley  
[Accessed 01 Sept 2020].

Goel, V., 2020. *Data Science Life Cycle in 5 Steps!.* [Online]   
Available at: https://medium.com/@vanshikagoel/data-science-life-cycle-in-5-steps-3fe720b2a2fe  
[Accessed 01 Sept 2020].

Herper, M., 2012. *Biotech's Comeback Kid.* [Online]   
Available at: https://www.forbes.com/sites/matthewherper/2012/12/04/biogen-idec-biotechs-comeback-kid/#18cd35a36c89  
[Accessed 1 Sept 2020].

Kenton, W., 2020. *Investopedia.* [Online]   
Available at: https://www.investopedia.com/terms/t/timeseries.asp  
[Accessed 30 Aug 2020].

Mickleboro, J., 2019. *The Telstra share price fell 23% in 2018.* [Online]   
Available at: https://www.fool.com.au/2019/01/02/the-telstra-share-price-fell-23-in-2018/   
[Accessed 01 Sept 2020].

Newman, R., 2015. *3 reasons why Commonwealth Bank of Australia’s share price jumped 5% in November.* [Online]   
Available at: https://www.fool.com.au/2015/11/30/3-reasons-why-commonwealth-bank-of-australias-share-price-jumped-5-in-november-2/  
[Accessed 01 Sept 2020].

Olah, C., 2015. *Understanding LSTM Networks.* [Online]   
Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/#:~:text=Long%20Short%20Term%20Memory%20networks,many%20people%20in%20following%20work.  
[Accessed 01 Oct 2020].

Pai, A., 2020. *CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/  
[Accessed 01 Sept 2020].

Phi, M., 2018. *Illustrated Guide to LSTM’s and GRU’s: A step by step explanation.* [Online]   
Available at: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21  
[Accessed 01 Oct 2020].

Quantas, 2020. *Qantas Airways Limited.* [Online]   
Available at: https://en.wikipedia.org/wiki/Qantas  
[Accessed 1 Sept 2020].

SuperDataScience, 2020. *Artificial Neural Network For Regression.* [Online]   
Available at: https://www.udemy.com/course/linear-regression-with-artificial-neural-network/learn/lecture/18888982#announcements  
[Accessed 20 Aug 2020].

Sydney, 2020. *Sydney Airport.* [Online]   
Available at: https://en.wikipedia.org/wiki/Sydney\_Airport  
[Accessed 01 Sept 2020].

Wiki, 2020. *Advantage LSTM.* [Online]   
Available at: https://www.google.com/search?safe=active&client=firefox-b-d&sxsrf=ALeKk000ny6VALcAlQHShp66GuqIwdDNHA%3A1602070098134&ei=UqZ9X-7gB5iG4-EPtKm2wAM&q=what+are+the+benefits+of+lstm+models&oq=what+are+the+benefits+of+LSTM+mode&gs\_lcp=CgZwc3ktYWIQAxgAMggIIRAWEB  
[Accessed 01 Oct 2020].

Yadav, P., 2018. *Responses (1).* [Online]   
Available at: https://medium.com/@mail2princeyadav/machine-learning-vs-deep-learning-b5c5a4fc5c  
[Accessed 01 Sept 2020].

Yoshua Bengio, P. S. P. F., 1994. *Learning LSTM with Gradient Descent is Difficult.* [Online]   
Available at: http://ai.dinfo.unifi.it/paolo//ps/tnn-94-gradient.pdf  
[Accessed 1 Oct 2020].

Li, L., 2020. Classification And Regression Analysis With Decision Trees. [online] Medium. Available at: https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054 [Accessed 8 October 2020].

Scikit-learn.org. 2020. Sklearn.Tree.Decisiontreeregressor — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html [Accessed 8 October 2020].

ML+. 2020. ARIMA Model - Complete Guide To Time Series Forecasting In Python | ML+. [online] Available at: <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/#:~:text=ARIMA%2C%20short%20for%20'Auto%20Regressive,used%20to%20forecast%20future%20values.> [Accessed 9 October 2020].