Share Price Prediction for Increasing Market Efficiency using Random Forest

---

by

Tshinanne Angel Mbedzi
Student No: 11618267

submitted in partial fulfilment of the requirements for Master of Science Degree in e-Science

in the Department of Mathematical and Computational Sciences
Faculty of Science, Engineering and Agriculture
University of Venda
Thohoyandou, Limpopo Province
South Africa

Supervisor: Dr W Chagwiza
Co-Supervisor: Prof W Garira

25 June 2022

Abstract

The price of a single share of a collection of sell-able shares, options, or other financial assets, shall be the price of a share price. The share price is unpredictable since it primarily depends on buyers' and sellers' expectations. Share is a primary and secondary market equity security. In this study we will use machine learning techniques to predict the share price for increasing market efficiency. In addition, it is important for us to build a models to create appropriate features to improve the performance of the models. The random forest and the recurrent neural network will be used to achieve this. To fix class imbalance, we analyse preprocessing of the data set, like the selection of the features using filter and wrapper methods and selected oversampling techniques. The model's performance will be evaluated using Mean absolute error (MAE), Mean square error (MSE), Root mean square error (RMSE), Relative MAE (rMAE), and Relative RMSE (rRMSE). The performance of the RNN and Rf algorithms was compared for the prediction of the closing price. The Rf model was found to be the best model for predicting the stock price (closing price). This research project together with its findings will have an impact in increasing market efficiency. This will also promote potential economic growth.

Keywords: Lasso, Market efficiency, Prediction, Random forest, Share price.

Declaration

I, **Tshinanne Angel Mbedzi**, **Student No: 11618267**, declare that this mini-dissertation is my original work and has not been submitted for any degree at any other university or institution. The mini-dissertation does not contain other persons' writing unless specifically acknowledged and referenced accordingly.

Tshinanne Angel Mbedzi

25 June 2022

Dedication

This research paper is dedicated to my dear father, who has been nicely my supporter until my research was fully finished, and my beloved mother who, for months past, has encouraged me attentively with her fullest and truest attention to accomplish my work with truthful self-confidence.

Acknowledgements

I would like to thank my supervisors Prof W Garrira and Dr W Chagwiza for their patience and guidance and the support they provided throughout my research. I would also like to thank DST-CSIR National e-Science Postgraduate Teaching and Training Platform for providing financial support. I would also like to thank the University of Venda and the University of the Witswatersrand, Johannesburg for granting me the opportunity to be part of this program. Finally, I would like to thank Ms. C Sparks and my fellow classmates for the support.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
ARMA	**A**utoregressive **M**oving **A**verage
AUC	**A**rea **U**nder the ROC **C**urve
CDPA	**C**ommon **D**istinctive **P**attern **A**nalysis
CNN	**C**onvolutiona **N**eural **N**etwork
1D-CNN	**1D**imensional- **C**onvolutional **N**eural **N**etwork
COVID-19	Coronavirus disease of 2019
DT	**D**ecission **T**ree
ECDF	Emperical **D**istribution **F**unction
GLmnet	Lasso and elastic-net regularized generalized linear models
HMM	**H**idden **M**arkov **M**odel
K-S	**K**olmogorov **S**mirnov
Lars	Least Angle Regression
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election**O**perator
LSTM	**L**ong **S**hort **T**erm **M**emory
MAE	**M**ean **A**bsolute **E**rror
ML	**M**achine **L**earning
MSE	**M**ean **S**quare **E**rror
NN	**N**eural **N**etwork
RF	**R**andom**F**orest
RNN	**R**ecurrent **N**ueral **N**etwork
RST	**R**eal-time **S**ubstructure **T**esting
rMAE	**r**elative **M**ean **A**bsolute **E**rror
RMSE	**R**oot **M**ean **S**quare **E**rror
rRMSE	**r**elative **R**oot **M**ean **S**quare **E**rror
SVM	**S**upport **V**ector **M**achine

Chapter 1

Introduction

1.1 Background

1.1.1 Share Price

The price of a single share of a collection of sell-able shares, options, or other financial assets, shall be the price of a share price. Publicly traded corporations' share prices depend on market demand and market supply. The share price is unpredictable, since it primarily depends on buyers' and sellers' expectations [25]. We then have the stock price, which is in the simplest terms is, the largest amount, or the smallest amount that anyone is willing to pay for the stock.

The maximization of shareholder capital commonly structured is a financial priority of the company, as expressed in the market value of shares of the company. In the end, the income of a company is the key determining factor in its share price. Earnings provide a calculation of the increase in the value over a period of time between the company and common shareholders [7]. Share is a primary and secondary market equity security [7].

1.1.2 Stock Market

A stock market is an open market for a company to list its securities and to gain financial resources by selling its stock at the agreed amount. The stock holder is entitled in exchange to an annual profit dividend or bonus. Stock holders can also exchange the stock at an agreed price on the stock market if they wish to gain from

the difference in the price of buying-and-selling operations [37].

Predictions on the stock market are very hard. This is partly due to the complexities associated with the business activity. Several variables, including political developments, economic conditions and aspirations of traders, interfere in the stock market. Therefore, it is very difficult to forecast stock price fluctuations [5]. The efficient market hypothesis suggests all available information is completely expressed in the current market price. The stock prices are expected to follow a random trend and current value is the best bet for the next price, since news is random unpredictable in nature and is currently unknown [30].

The ability to predict the course and not necessarily the value of potential stock prices is the secret to making financial predictions of capital. Each investor must know that the anticipated course of the stock is the buying or selling decisions. Studies also found that value-based forecasts would yield higher profits [5].

Throughout finance research, one of two explanatory variables, fundamental analysis and technical analysis, usually associated with the estimation of stock prices. The forecasters often employ strategies from both groups to boost prediction efficiency at the same time. In addition, there are various time series foreseeing mathematical models using fundamental and technical research variables suggested by academics [14].

Figure 1.1 illustrates a diagrammatic representation of the processes and players on the stock market. The diagram shows what processes are involved in buying and selling, who is involved, current technologies, triggers to buy/ sell the instruments, the final product i.e shares and issuing of share certificates.

- **Institutions ("Buy Side" Fund Managers)**
  Institutions consist of fund managers, retail investors and institutional investors. These investment managers offer capital to companies which need the money to develop and run their companies. Corporations in exchange for their money give their debts or equities in the form of bonds and shares to the institutions respectively.The two major player cycle in capital markets is completed with the exchange of capital and debt or equity.

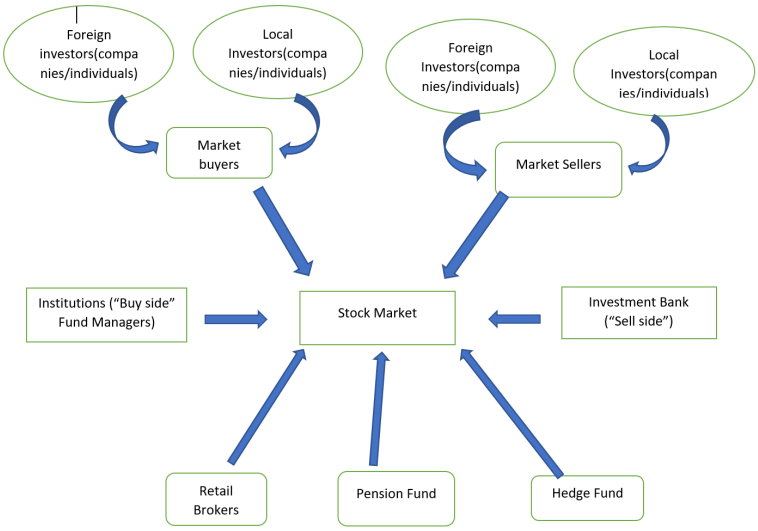

FIGURE 1.1: Diagrammatic representation of the stock market.

- **Investment Banks ("Sell Side")**
  Investment banks are contracted as an intermediary to promote transactions between companies and institutions. Investment banks are responsible for linking institutional investors with companies based on perceptions of risk and return and investment types. Investment banking careers include rigorous financial modeling and valuation research.

- **Hedge Fund & Pension Fund/Mutual Fund**
  The Hedge Fund Pension Fund/Mutual Fund, are major players, paying high fees for the self-financing operations of investment banks.Both hedge funds and mutual funds together generate 90% of the order flow from all exchanges around the world. Mutual funds / mutual funds; their primary revenue is generated from management fees, Spread & commission and financing activities.

- **Stock Market Brokers(Retail Brokers)**
  Stock market brokers are part and parcel of a country's investment scenario. They not only make it possible for individual or institutional investors to buy

and sell stocks in the stock markets, but also give customers valuable investment advisory service.

- **Investors**
  An investor is an individual or another entity (for example a corporation or mutual fund) that commits capital in order to receive financial returns. Investors rely on various financial instruments to obtain a return rate and meet essential financial goals, such as building retirement accounts, financing college education or simply accumulating more assets over time.

- **Foreigners**
  Foreigners play the roles of both buying and selling of stocks. Within the capital markets of developing host countries, foreign investors played an increasingly important role. While foreign investors carry large amounts of cash, they only investment on those stocks, rather than on the market as a whole, thanks to asymmetric knowledge and home distortion. The advantages of foreign investment may therefore be restricted and there may be dispersal of stock prices between favored foreign stocks and unfavoured foreign stocks [13].

1.2 Research Aims and Objectives

1.2.1 Research Aims

The aim of the research project is to use random forest and recurrent neural networks to predict share price in order to increase market efficiency.

1.2.2 Objectives

The above aim(s) of this research project will be achieved through the following objectives (goals):

- Build a model/s in order to create appropriate features to improve the performance of the model(random forest and a neural network).

- To develop appropriate feature selection strategy in order to obtain accurate predictions.

- Develop accurate models and compare their performances.

1.3 Limitations

At this given time various limitations maybe en-counted. Considering the COVID-19 pandemic, certain limitations and delays arose. There were delays in the execution of the project. Since the COVID-19 outbreak, schools promoted online learning. It was not easy to carry out the project at home with no physical meetings with supervisors and other external assistance. With limited access to resources to help carry out the project, there where delays in the completion of the project as a whole. Another limitation where having to test and compare the two models in the time that was given. Nonetheless we completed the project in the given time frame and carried it out accordingly.

1.4 Overview

The outline of the project is structured in the following way:

In Chapter 2, We explored previous literature on principles of machine learning and also the review of studies using methods similar to those suggested. A theoretical framework of the machine learning methods being used in this project is provided in Chapter 3. Chapter 4 provides the overview of the project. The conclusion and recommendation of the project are provided in Chapter 5.

Chapter 2

Literature Review

Machine Learning (ML)

Machine learning(ML) is an Artificial Intelligence (AI) sub-field. The goal of machine learning is to understand the data structure [34] and fit that data into models that humans can comprehend and use. ML is a field aimed at enabling machines to intelligently learn and carry out duties. In order to achieve this objective, a number of algorithms have been and remain in place over the years. These algorithms extract information insight into the modeling of a feature that can be used to forecast fresh information outputs [29].

Although in computer science, machine learning is a field, it differs from traditional approaches to computing. Algorithms in traditional computing are sets of specifically programmed orders that computers used to calculate or resolve problems. Instead of machine learning algorithms, computers can train on data inputs and then use statistical analysis to deliver values in a specific range [34].

Machine learning algorithms are usually divided into 3 types, namely supervised learning, where the form inputs and outputs are described by analysis. The main aim is to establish a connection that can estimate the relation between variables of input and output. Supervised learning is mostly used for classification of discrete class labels and regression for continuous outcomes. The second type of learning is unsupervised, in which the main aim is to examine the secret structure of the data without prior knowledge of the group/data, in light of unlabelled information. The third type is reinforcement learning, which is to build a system (agent), based on

interactions with the environment, that improves its efficiency.

Predicting stock movements is a fascinating subject and researchers in various fields are studied extensively. Machine learning has been widely studied for its potential for financial market forecasting, a well developed algorithm in a wide variety of applications [32].

2.1 Related work

By allowing for larger volumes of data to be processed by incredibly more smart algorithms not used before, since the speed of processing them has been not, for the time being, fast enough to deliver results for decision-making, but now it is sufficiently easy to see the big data revolution better by looking at its impact on the stock markets [20].

Predicting stock prices is incredibly difficult and frustrating, since prices just adjust like a random walk and time varies.Stocks pursue a random and unexpected route, according to random walk, making all methods of stock price prediction worthless in the long run. Various analysts have used sophisticated stock market approaches and techniques in trade decisions in recent years [31]. Many techniques for forecasting stock movements have been developed over the years. At first, standard methods of regression were used for forecasting stock patterns [28]. Artificial neural network models are the most important nonlinear models that have been commonly used in financial markets in recent years and have shown desired results[12]. In most cases, artificial neural networks outperform regression approaches. Hill et al. results on stock price prediction, for example, showed that this approach outperformed other models [12].

Kumar et al. investigated the efficacy of the ARMA, RF, SVM, and artificial neural network methods in predicting the SP CNX NIFTY index rate, and compared the functions of the three nonlinear models and one linear model. The results suggested that the SVM may be able to provide more accurate results [23]. Zhang used an SVM to forecast the Shanghai stock exchange price pattern, concluding that

the SVM has a high prediction capability and that combining an SVM with intelligent models produces even better outcomes than the model [44]. Kara et al. used fuzzy neural network models and SVM to predict the course of Istanbul's stock price movement. The fuzzy neural network's prediction score was 74.5 %, while the SVM's was 52.71%, indicating that the fuzzy neural network performed better than the SVM [18].

Artificial Networks (ANN) and Support Vector Machine (SVM) are two algorithms that are used most frequently to predict the transfer of stock and market price index. ANN emulates how the brain functions to create a neural network. Whereby a SVM is a highly different type of learning algorithm that is defined by decision-making capability control, kernel function usage and solution shortages [28]. Random forest surpasses the other three prediction models on overall performance for the first approach of input data where ten technical parameters are represented as continuous values, according to the testing results. Experiments also reveal that when these technical factors are provided as trend deterministic data, the performance of all prediction models improves.

Kannan et al. used data mining to identify hidden trends in its investment decisions from historic data that possibly predict. The forecasting of stocks is a daunting job for the forecasts of financial time series [17]. On half of the stocks, the algorithm performed well, while on the other half, it performed poorly. The algorithm predicted both increases and decreases, but they didn't happen very often. This algorithm could be used as a buying or selling signal, or to back up a trader's stock price predictions.

Sohangir et al. looked at how deep learning techniques like LSTM and CNN could boost market prediction accuracy using public sentiments. Deep learning (CNN) outperformed machine learning (ML) algorithms such as logistic regression and Doc2vec, according to the findings of the report. As compared to auto-regressive integrated moving average and generalised auto-regressive conditional heteroscedasticity, the simulation results showed that their proposed ensemble approach is more appealing [35]. According to the findings, CNNs have significantly better accuracy

than the other models. We can utilize CNN to extract the emotion of authors towards stocks from their words based on our findings. Some members of the financial social network have the ability to correctly predict the stock market. They can predict future market movement by utilizing CNN to determine their emotion. Abe et al. on the other hand, used a deep neural network approach to forecast stock price and found that it was more effective than shallow neural networks [1]. Deep neural networks outperform shallow neural networks in general, and the top networks exceed typical machine learning models, according to their findings. These findings suggest that deep learning has the potential to be a useful machine learning method for predicting cross-sectional stock returns.

To forecast the stock market, researchers compared single, ensemble, and integrated ensemble ML techniques. The AdaBoost-LSTM ensemble learning methodology outperforms several other single forecasting models and ensemble learning approaches, according to the empirical findings. Boosting ensemble classifiers outperformed bagged classifiers, according to the study [45]. According to the results of the trials, integrated ensemble machine learning (IEML, i.e., RS–boosting and multi-boosting) approaches outperform individual machine learning (IML, i.e., decision tree) and ensemble machine learning methods [EML, i.e., bagging, boosting, and random subspace (RS)]. RS–boosting is the best method for predicting credit risk in small and medium-sized businesses (SMEs) among the six techniques. For stock market prediction, Sun et al. proposed an ensemble LSTM using AdaBoost. The AdaBoost-LSTM ensemble outperformed several other single forecasting models, according to their findings. For forecasting stock price movement, a homogeneous ensemble of time-series models like SVM, logistic regression, Lasso regression, polynomial regression, Naive forecast, and others were proposed [39]. Similarly, Yang et al. used voting techniques to combine SVM, RF, and AdaBoost to predict whether to buy or sell stocks on an intraday, weekly, or monthly basis. In terms of precision, the analysis found that the ensemble technique outperformed a single classifier. As compared to single feedforward neural networks, Gan et al. proposed an ensemble of feedforward neural networks for predicting stock closing prices, and he showed a higher prediction accuracy. The empirical result shows that a homogeneous ensemble ANN predicts stock market price better than a single ANN [9].

Gao (2016) used a recurrent neural network (RNN) to predict the stock market using long short-term memory (LSTM). The aim of this analysis was to see whether LSTM could be used to predict stock market movements. According to the findings, the LSTM model's average precision and accuracy in predicting six stocks was 54.83%, with the highest and lowest precisions being 59.5% and 49.75%, respectively. Bao et al. demonstrated a new deep learning architecture for stock price prediction that combined wavelet transforms, stacked auto encoders, and LSTM methods. Six market indices and their associated future indices were chosen to test the proposed model's results. In comparison to other similar models, the findings showed that the proposed model is more precise and profitable [2].

A 2-phase ensemble method for forecasting stock prices was proposed in another review, which included several non-classical disintegration models, such as ensemble empirical mode decomposition, empirical mode decomposition, and total ensemble empirical mode decomposition with adaptive noise, as well as ML models, such as SVM and NN. The use of RF robustness in stock selection strategy was implemented and evaluated [16]. They concluded that fundamental and long-term technical features are essential to long-term benefit in sound stock investments using the fundamental and technical dataset. For forecasting the stock market, Mehta et al. proposed a weighted ensemble model based on weighted SVM, LSTM, and multiple regression. Their findings show that when it comes to stock estimation, the ensemble learning methodology achieves the highest precision while reducing variance [24].

Jing et al. for classification , prediction and recognition used artificial neural networks. Neural network outputs or trading techniques are an art. The authors speak about a 7-step model design approach for a neural network prediction [16].

For forecasting stock price movement, Assis et al. suggested a NN ensemble. For stock market prediction, a deep NN ensemble with bagging was suggested. According to the findings, putting together multiple neural networks to predict stock price movement is more reliable than using a single deep neural network [40]. Jiang

et al. used a variety of cutting-edge machine learning techniques, including a tree-based and LSTM ensemble using stacking combination technique, to forecast stock price movement based on both macroeconomic data and historical transaction data. On average, the authors reported accuracy of 60–70% [15]. Kohli et al. investigated the success of various machine learning algorithms (SVM, RF, Gradient Boosting, and AdaBoost) in stock market price prediction. AdaBoost outperformed Gradient Boosting in terms of forecasting accuracy, according to the report [27].

An ensemble classifier for NN using bagging is presented in [22]. Their findings showed that an ensemble of NN classifiers outperforms a single NN classifier. Wang et al. also suggested an RNN ensemble system for effectively detecting stock-price manipulation activities, which combines trade-based features derived from historical trading records with characteristic features of the list companies[38]. Their findings show that the proposed RNN ensemble outperforms state-of-the-art approaches by an average of 29.8% in terms of AUC value in detecting stock price manipulation. Ensemble classifiers and regressors have been shown to have a higher predicting accuracy than single classifiers and regressors in previous studies [38].

Tsai et al. did a research that they tried, through ensemble training composed of decision-making trees and artificial neural networks, to forecast stock prices. They provided data from Taiwanese stock market data taking basic indexes, technical indexes and macroeconomic indexes into account. The Taiwan stock exchange data output of Decision Tree + Artificial Neural Network showed 77% F-score of results. F-score output shows up to 67% with single algorithms [36].

Tiffany et al. used the neural network for their handling capabilities nonlinear interactions and the introduction of a new fuzzy time series model in order to boost prediction. The fuzzy link is used to estimate the inventory index of Taiwan. Within the neural network fuzzy time series, observations are used for preparation and model observations for prediction as in-sample observations. The inconvenience of taking all membership levels for training and prediction will impact the neural network performance. They made the difference between results in order to prevent that. That lowered the discourse universe size [41].

Md. Rafiul et al. used Hidden Markov Model(HMM) method in the estimation of stock prices for the relevant markets. Due to its proven suitability for dynamic system modeling, HMM was used for pattern identification and classification problems. It can manage new data in a reliable and effective way for creating and comparing related trends [11].

Kim et al. employed a genetic algorithm to convert constants into discrete values. They introduced a new evolutionary calculation approach called the genetic quantum algorithm, which was used for reducing the complexity of the space of the characteristics. Genetic Quantum Algorithm is based on quantity measurement concepts and principles including qubits and overlays of states. The Genetic Quantum Algorithm may be represented by a linear superposition of solutions, because of its probabilistic representation, instead of binary, numeric or symbolic representations by using bit chromosomes. Quantity gates are used as genetic operators to find the right solution [19].

Ching-Hseue et al. introduced a multi-technical hybrid forecasting model to predict stock price patterns. In the cases in which four procedures listed such as selecting the critical technical indicators, CDPA is used to minimize entropy principle by common indicators based on a correlation matrix. The RST algorithms are also used to extract the linguistic rules and use genetic algorithms to improve the rules that were extracted to improve predictability and stock returns. The benefit was found that rules and forecasts on objective inventory data are more accurate and understandable than subjective human judgments [4].

Fazel et al. used the expert system based on a type-2 fuzzy rule to analyze stock price. The type-2 fuzzy model used appeared as input variables to the technological and fundamental indexes. The model used to forecast an automobile manufacturer's stock price in Asia. The output value of the input was projected onto the input field in order to produce next input variable membership values tuned with genetic algorithms. For inference and increase the strength of the system, the Type 1 technique was used. This technique has been used to reduce robustness, durability and error. It is used to predict more profitable stock trading [42] .

Many resources and applications are available to forecast stock markets, share amounts and share prices for any specific financial organization. Most say that they are almost 100% accurate in forecasting the stock market, however, users' opinions differ.

2.2 Summary

Many different techniques have been used to predict stocks in the past which have been discussed in the literature review, Very limited studies show the use of random forest to predict stocks. Hence in this project we seek to investigate the application of RF to predict stock and compare it's efficiency.

2.3 Significance and Motivation

The prediction of stock markets is an important topic for financial investors in determining which shares to buy and sell. Comparative output of various methodologies and data sets has the benefit of updating the techniques of existing methodologies to forecast share price efficiently. This project helps to explain and enhance efficiency of various machine learning approaches. In addition, the project will assist with the introduction of the stock market forecast economic growth model.

Chapter 3

Research Methodology

3.1 Introduction

Some of the possible techniques used in predicting share will be described in this chapter. Furthermore the chapter will entail some methods of machine learning that could be used to carry out this study.

3.1.1 Random forest

Random forest is a classification and regression ensemble-learning technique which structures forests by various tree modeling. A great number of individual unpruned decision trees are used for random forests that are generated by randomizing the split on each node of the decision tree. The accuracy of the prediction for each tree is typically less than the accurate split-size tree. Nevertheless, more robust accuracy obtain to that of a single tree with precise divisions by adding a couple of approximated trees in the ensemble [33].

The fundamental notion of ensemble learning is not enough of a single regression tree to evaluate the expected value of a dependent variable [28]. Random forests build n classified trees using replacement samples and estimate the class based on the predictions of the majority of the trees . After n trees have been created, the average predicted value is taken into account by each tree in the set when the test data is being used [28]. The Figure 3.1 is an example of a conceptual illustration of a random forest.

Decision tree learning is one of the most common classification techniques. The classification precision is highly efficient and superior to other classification methods. It is very reliable. The tree that is widely referred to as the decision tree reflects the classification model that these techniques include. Decision tree may also be learned as a regression tree and seems to be highly successful as well and has low error rates [28].

As it is shown in Figure 3.1, when the target information is an input into each tree model, the tree models generate their results independently, depending on their nature [33]. A majority judgment is sequentially determined as the forest production, or an arithmetic average of its findings. In this estimate, the random forest yield is a consensus judgment in the situation of the random forest built by classified trees, and the random forest built by regressive trees is an arithmetical average [33].

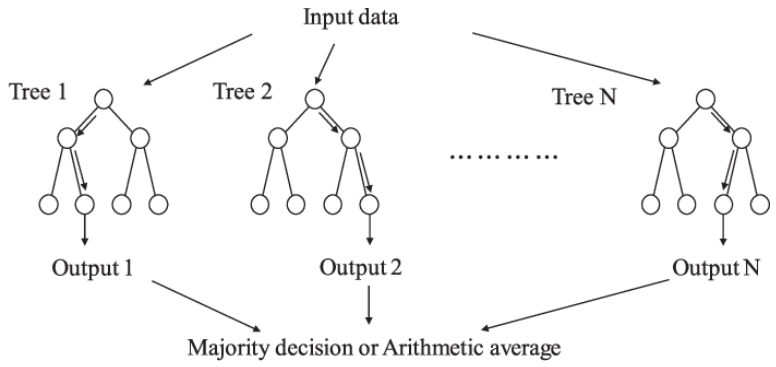

FIGURE 3.1: Conceptual illustration of random forest.

Random forests have advantages and disadvantages as well as every other technique. The advantages of random forest are:

- They are one of the most precise available learning algorithms. They create a very accurate classifier for many data sets.
- Effectively run in large databases, thousands of input variables can be managed without the variable deletion.
- They offer estimates for which variables in the classification are significant.
- They produce an unbiased internal estimation of the error of generalization with the forest construction.

- The reliability and precision of the estimation process is retained while a significant proportion of the data is missing.
- Methods to balance error are provided in class population unbalanced data sets.

The disadvantages of random forest would be:

- Several data sets with noisy tasks for classification and regression have been found to overfit the random forests.
- Random forests support certain attributes with more levels for data with categorical variables of different levels. Therefore, random forest variable scores for this type of data are not accurate.

Basic principles

The term "random forests" is vague. It's a generic expression for certain writers for combining random decision trees regardless of how trees are obtained. Whilst for other authors, it refers to the original algorithm of Breiman [3].

The forest mechanism is, as already stated, sufficiently versatile for the supervised classification and regression tasks. However, in this introduction we focus on regression analysis and discuss the classification case only briefly, in order to keep matters simple. A non-parametric regression approximation is the general context in which a random vector $X \in \chi \subset \mathbb{R}^p$ is observed, and the objective is to predict the integrable square random response $Y \in \mathbb{R}$ by estimating the return function $m(x) = \mathbb{E}[Y|X = x]$. Assuming a training sample $D_n = ((X_1, Y_1), ..., (X_n, Yn))$, is distributed as the independent random variables pair of prototypes (X,Y). The dataset D_n is used to construct $m_n : \chi \to \mathbb{R}$ of the function m. Therefore, m (mean squared error) regression function estimate is compatible if $\mathbb{E}[m_n(X) - m(X)]^2 = 0$ as $n \to \infty$ [3].

A random forest is the predictor of the random M regression trees set. The expected value at query point X for the j-th tree in the family is denoted by $m_n(X : \Theta_j, D_n)$, where $\Theta_1, ..., \Theta_M$is a separate random variable Θ distributed as a generic random

variable m and independent of D_n. Variable Θ is used to evaluate the training set before each trees grow and to choose the following instructions, more detailed definitions are given later. The $j-th$ tree estimate takes the mathematical form:

$$m_n(x;\Theta_j,D_n) = \sum_{i\in D_n{}^*(\Theta_j)} \frac{\mathbb{1}x_i \in A_n(x;\Theta_j D_n)Y_i}{N_n(x;\Theta_j,D_n)}. \tag{3.1}$$

Where;

- $D_n{}^* =$ the set of data points selected prior to the tree construction,
- $A_n(x,\Theta_j,D_n) =$ the cell containing x, and
- $N_n(x;\Theta,D_n) =$ the number of points that fall into $A_n(x,\Theta_j,D_n)$.

The trees are now combined to make the (finite) forest estimate as written;

$$m_{M,n}(x;\Theta_1,...,\Theta_M,D_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta_j,D_n). \tag{3.2}$$

Since M can be arbitrarily larger, modeling, it makes sense to allow M to be infinitely large [3] and to take into account instead of the forest estimates

$$m_{\infty,n}(x;D_n) = \mathbb{E}_\Theta[m_n(x;\Theta,D_n)]. \tag{3.3}$$

Where in equation 3.3;

- $\mathbb{E}_\Theta =$ the expectation with respect to the random parameter Θ, condition on D_n

Operation "$M \to \infty$" is justified by large numbers, which almost definitely asserts that it is subject to D_n,

$$\lim_{M\to\infty} m_{M,n}(x;\Theta_1,...,\Theta_M,D_n) = m_{\infty,n}(x;D_n). \tag{3.4}$$

3.1.2 Recurrent Neural Networks

Recurrent neural networks are a class of neural networks that can be used in feed-back loop concepts. After processing the data with one of the hidden layers in

the network, the layer along with the new input is sent back and forth processed. Depending on how they are organized, various kinds of recurrent networks exist. Recurrent neural networks are used when data, like time series, is sequential in nature. A RNN can be used if you assume the data have some form of autoregressive structure. Some other way of looking at RNNs is by providing a database that records information about what was measured until now [6].

RNNs are fuzzy because they do not make forecasting using exact models in the training data but use their internal representation to perform a high-interpolation among training examples— like other neural networks. In addition, fuzzy predictions are not influenced by the dimensionality curse, and thus the simulation of real or multivariate information is much superior to exact matches [10].

In reality, the open content reach of a standard RNN structure is very small. The vanishing gradient problem is caused by the fact that the effect of a given input on the hidden layer, and thus on the network output, decreases exponentially and vanishes. RNNs are called recurrent since the performance of each layer is dependent on the calculations of the layers before it; in other words, these networks have memory that stores data-related information. These networks are actually multiple copies of ordinary neural networks organized next to one another, each transmitting a message to the next [26]. It should be noted that RNNs can arbitrarily process long data sequences, but only look back for some steps in practice [6].

There are four phases to neural network training.

1. Training data preparation: the more data there are, the easier the training. Data preprocessing is one of the stages in this stage, with the aim of removing missing data and unwanted variations, as well as data simulation. Normalization (e.g. transmission to a zero–one range) is an example of data preprocessing.

2. Choosing a network architecture that is right for you. The number of neurons, layers, and network form will be calculated at this stage; the number of layers and neurons will be determined by trial and error.

3. Networking training. The training cycle is iterated at this point until the desired results are achieved [26].

There are several different ways in which RNNs can be used. Below are some examples of other RNNs being used:

i) Language modelling and generating text

ii) Machine translation

iii) Speech recognition

iv) Gathering image description

v) Video tagging

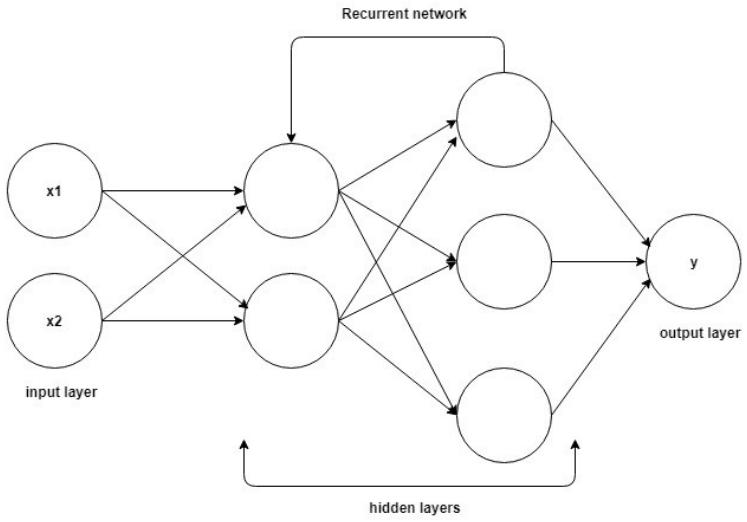

FIGURE 3.2: Simple Recurrent Neural Network structure.

The Figure 3.2 illustrates the architecture of a simple recurrent neural network structure. In the Figure 3.2 it is shown that we have x_1 and x_2 as the input variables in the input layer. They are then followed by the hidden layer which consists of the recurrent network and then followed by the output layer y.

RNN Architecture

We introduce a time notion to the model in recurrent neural networks. It can also be called neural feedforward networks because they are generated by combining borders that reach adjacent time steps [21]. Edges that link adjacent time steps are called recurrent edges. They form cycles in the network. This includes self connection cycles from a node to itself across time in the network. The cycles can be seen in Figure 3.3 below.

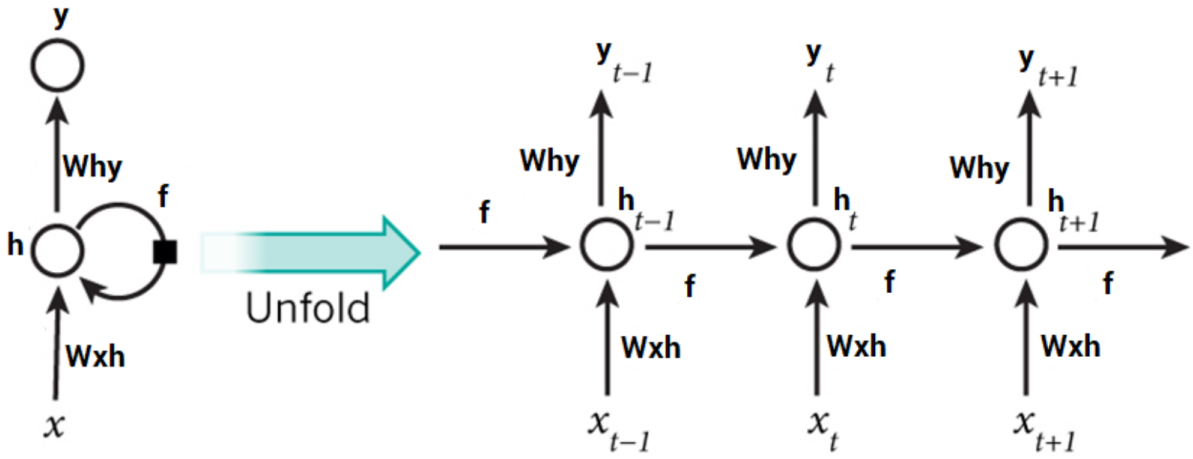

FIGURE 3.3: A folded simple Recurrent Neural Network [21]

Suppose x_t was a sequence of inputs in order to understand formulas that control recurrent neural networks. Nodes with recurrent edges receive input from the current input point and also from hidden state values h_{t-1} in the network's previous state. Every neuron in the hidden nodes performs a linear matrix operation with input data x_t, and according to the form of prediction, the output of this operation is fed into an activation function. The output $\hat{y}_t$ is determined by the hidden values h_t at each time t. Input x_{t-1} at time $t-1$ influences the output $\hat{y}_t$ at time t and later by the way of the recurrent connections. The results of linear matrix operation of these parameters gives two equations that are necessary for computation at each time step on the forward pass in a folded simple recurrent neural network shown

in Figure 3.3. The two equations are given as:

$$h_t = \sigma(W^{xh}x_t + W^{hh}h_{t-1} + b_h), \tag{3.5}$$

$$\hat{y}_t = \sigma(W^{hy}h_t + b_y). \tag{3.6}$$

Equation 3.5 is used to calculate the output of the hidden layer each time and equation 3.6 is used to calculate the output of $\hat{y}_t$. Here W^{hh} is the matrix of the weight that connects the hidden layers, W^{xh} is the matrix of weight between the input layer and the hidden layers and W^{hy} the matrix of weight between the hidden and output layers. The bias parameters are represented by vectors b_h and b_y. The weights matrices here in recurrent neural networks are shared parameters as like in feedfoward neural networks, they are just used in different time steps. If the dimensions of the input layer, output layer and hidden layer are D_x,D_y and D_h respectively. Then the dimensions of weights W^{hh}, W^{xh} and W^{hy} are $D_h \times D_h$, $D_h \times D_x$ and $D_y \times D_h$ respectively.

The unrolled recurrent neural is shown in Figure 3.3. The network is no longer seen in this picture as cyclically, but rather as a whole series that shares weight over time. The entire unrolling network can be trained over many time steps. RNN training is similar to the simple artificial neural network and also uses a minor difference back propagation process. Since the weights and biases of this network are shared by all time steps, the gradient in each output depends on previous steps calculated and not just on the current time steps calculations.

3.2 Feature Selection Strategies

The process of selecting predictive features and removing redundant and unpredictive features is known as feature selection. [8].The key reasons for using the feature selection are:

- to facilitate analysis of the model, remove the redundant variables and do not add any information;

- reduce the problem to allow algorithms to run more quickly, enabling high-dimensional data to be processed;
- reduce overfitting.

Popular algorithms for selecting features may be grouped in two types: (1) filter and (2) wrapper approaches [43]. In order to pick key features independently, the filter approaches use the general characteristics of the training data; that is, the input variables only are considered. Wrapper methods evaluate and assess the optimal subset of features using the prediction output of a specified learning algorithm. In this section, several common feature selection algorithms will be briefly reviewed.

3.2.1 LASSO (Least Absolute Shrinkage and Selection Operator)

In 1996, Robert Tibshirani first formulated LASSO. It is an efficient way of performing two key tasks: regularization and selection of features. This method places restriction on the sum of the model parameters' absolute values, the sum must be smaller than the value. The method applies a shrinking process (regularization) that penalizes the regression coefficients, which some of them shrink to zero. The variables which still have a non-zero coefficient after the process of shrinkage are selected to belong to the model during the selection process [8]. The aim of this method is to reduce the prediction error to a minimum.

In practice, the tuning parameter λ, which controls the penalty's power, is very important. When λ is large enough, the coefficients are required to be exactly equal to zero, resulting in a reduction in dimensionality. More coefficients are shrunk to zero as the parameter is increased. We have an OLS (Ordinary Least Sqaure) regression if $\lambda = 0$ on the other side [8].

The use of the LASSO method provides many advantages, but first and foremost a good precision because shrinkage and removal of the coefficients reduces variation without significantly raising bias, particularly if you have a few observations and a lot of features. When it comes to the tuning parameter λ, we know that as λ increases, bias increases and variance decreases, so a trade-off between bias

and variance must be sought. In addition, LASSO helps improve the model interpretability by removing unnecessary variables not related to the response variable, thereby minimizing overfitting [8].

3.2.2 DT (Decision Tree)

A DT consists of one root and different branches, nodes and leaves. Every feature is included in one node and only features that help to classify appear in the tree, while others are unable to choose good features. The DT is focused on the entropy principle, which selected as discriminating and explanatory features variables with the highest information gain [8]. First the information required for classification of a given sample is determined and after divising the sample in accordance with the selected variable X, this Information will be recomputed. The distinction between these two information is known as the variable X information gain.

3.3 Data

The data used to conduct this research is sourced from an online source, Kaggle. The dataset includes information on the largest 35 companies in South Africa by market value, some economic data that may be applicable to the prices and other measured indexes: a composite SA40 index and a SA40 "VIX" index for volatility measurements in the composite index. The price information is in the following format: Date, Close, Open, High, Low, Vol, Change, Beta, Alpha. Financial information is in the following format: Date, Revenue, Cost of sales, Gross profit, Operating profit, Net profit, Headline earnings, Shares in issue, Non-current assets, Current assets, Non-current liabilities, Current liabilities, Cash from operations, Cash from investing, Cash from financing, Change in cash, Full year, Half year. Economic data (including a variety of resources and potentially important global values) are typically in format: Date, Value.

3.3.1 Data Preprocessing

Prior to training the models, data must be preprocessed. It is a method of converting and cleaning data into a usable format in order to enhance the model's efficiency

and data quality. Since training models with raw financial data, which is noisy by design, results in overfitting and underfitting of the model in predicting the production(performance), the performance of predictive models is dependent on the stage of preprocessing.

3.3.2 Data Splitting

The dataset must be partitioned into two sections, the training set and the test set, before the models can be trained. This is done for the sake of performance. The two sets are used for different purposes: the Training set is used to build the model, while the Test set is used to evaluate the model. In practice, after preprocessing the data, the first 80% of the compiled data was used for preparation, 10% for validation, and 10% for evaluating the performance of trained models.

3.4 Variable selection

When restricting the number of variables in the method, we use function variables to evaluate the target variables. In terms of preventing overfitting, facilitating pattern analysis, and reducing computational time, the variable selection method plays an important role. There are several different methods for selecting variables, but we chose LASSO for this project.

3.5 Evaluation metrics

Various performance metrics are used to assess regression performance. We focused on the regression issue in this project. There are also several regression efficiency measures such as mean absolute error, mean square error, root mean square error etc. In this project we considered most of those.

3.5.1 Mean absolute error, Mean square error, Root mean square error, Relative MAE, Relative RMSE

The output of the models was evaluated using Mean absolute error (MAE), Mean square error (MSE), Root mean square error (RMSE), Relative MAE (rMAE), and Relative RMSE (rRMSE) in this project. We came up with the following definitions:

The mean of the absolute errors is called the mean absolute error. The MAE units are the same as the expected target, which helps determine if the error size is significant. The MAE is a measure of how well a model performs. The smaller the MAE, the better.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|, \tag{3.7}$$

$$rMAE = \frac{1}{n}\sum_{t=1}^{n}[\frac{\hat{y}_t - y_t}{y_t}]. \tag{3.8}$$

Equation 3.9 is the mathematical expression for mean square error(MSE). The MSE metric is used to calculate the average of the squares of errors or deviations. MSE removes any negative signs by squaring the gaps between the points and the regression line. MSE takes into account the predictor's variance as well as bias. Larger variations are also given more weight by MSE. The greater the error, the higher the penalty. The MSE is a measure of how well a model performs. The smaller it is, the better.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2. \tag{3.9}$$

The root Mean Squared Error(RMSE) metric measures a model's ability to predict a continuous value. The RMSE units are the same as the expected target, which helps determine if the error size is important. The model's output is better when the RMSE is low. The mathematical representation of the RMSE is expressed in 3.10.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)}{n}}, \tag{3.10}$$

$$rRMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{P}[\frac{y_k - \hat{y}_k}{y_k}]^2}. \tag{3.11}$$

where;

- n is the number of samples,
- y_t is the actual value,
- $\hat{y}_t$ is the predicted values,
- $\overline{y_t}$ is the mean value of y_t, $t = 1, 2, \cdots, n$, and
- k is the dummy variable time.

The lower the value error, the closer the approximate true values are.

3.6 Implementation

The RF, and RNN algorithms implemented using the Keras machine learning package and LASSO was implemented using R. All models were implemented using sklearn and tensor flow Python packages. Both Python and R used to execute this project.

3.6.1 LASSO R implementation

In order to implement LASSO, we used the R statistical software, which already has a number of built-in functions. Regularization approaches are performed by two major packages in particular [8].

- Glmnet (Lasso and elastic-net regularized generalized linear models)is a R package that suits linear and generalized linear models, penalizing the maximum likelihood with the LASSO method and Ridge Regression, as well as a combination of the two penalties. Glmnet uses cyclical coordinate descent to find the minimum.
- Lars (Least Angle Regression) is a modern model-selection approach based on conventional forward selection, in which we choose the explanatory variable with the highest absolute correlation with the response y from a set of possible explanatory variables. The LASSO method is implemented in the Lars package [8].

Both the algorithm and the procedure for implementing the LASSO are identical. The key difference is that glmnet performs a cyclical coordinate descent on a grid of possible values for λ and finds a sequence of loss function-related models as an output. Lars, on the other hand, can calculate the exact value as to where a new variable is introduced into the model [8].

Glmnet, on the other hand, is the most widely used nowadays, and one of its advantages is that it can be used for generalized linear models, while Lars is limited to linear regression models. Furthermore, experiments have shown that glmnet performs better in a variety of circumstances. In conclusion, we conducted our research using glmnet [8].

3.7 Summary

This section described the techniques, methods and measures to be applied when forecasting stock price. The methods used to interpret the data include time series data techniques, RF, RNN and performance measures.

Chapter 4

Results and Discussion

4.1 Introduction

With the algorithms discussed in Chapter 3, the analysis of the data was presented in this chapter. Using Python and R, the performances of the algorithms(RF and RNN) were compared in this chapter, using the performance measures discussed in Chapter 3. We will compare the performance of the random forest with that of the recurrent neural network for the prediction of the closing price.

4.2 The Data

4.2.1 The Source of Data

The data used to conduct this research is sourced from an online source, Kaggle.

4.2.2 Data Exploration

The data here shows that the stock prices where only recorded during week days(Monday-Friday) from 2010-07-21 to 2019-01-04.

FIGURE 4.1: Graphical representation of the data.

The graph in Figure 4.1 is a visualisation of the closing price. The graph shows when the stocks increases and decreases. This will then show when it would have been a good time to buy and a good time to sell. Looking at the Figure 4.1, it is seen that the best time buy would have been between December 2011 and January 2012, and the best time to sell would have been in January 2019. Between 2014 and 2017, the stock prices have been stable.

4.2.3 Descriptive statistics

The table 4.1 shows the descriptive statistics for all the variables in the data set. The table outlines the minimum, maximum, median, mean and the standard deviation.

	Min	Max	Median	Mean	St.Div
Open	81.100	327.700	142.250	151.819	49.350
High	82.800	328.750	144.300	154,145	50,127
Low	80.000	325.650	140.425	149.361	48.597
Last	81.000	325.95	142.000	151.585	49.420
Close	80.950	325.750	141.950	151.562	49.402
Total trade quantity	39610.000	29191015.000	1768579.000	232764.392	2081347,579
Turnover (Lacs)	37,040	55755.080	2552.165	3919.237	4547.901

TABLE 4.1: The descriptive statistics of the dataset.

The minimum being the smallest data value and the maximum being the largest data value. The median is the 'middle number' of the sorted data values which is mathematically represented by;

$$median = Z_{50} \tag{4.1}$$

The average of the data values is known as the mean which is mathematically represented by;

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.2}$$

The data's standard deviation , denoted by s, is a popular measure of dispersion. It measures the average distance between a single observation and the mean.

$$\bar{x} = \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1}} \tag{4.3}$$

Equation 4.3 is the mathematical representation of the standard deviation. Which is equal to the square root of the sample variance.

Min	Max	Median	Mean	St.Div
80.950	325.750	141.950	151.562	49.402

TABLE 4.2: The descriptive statistics of the Close price.

Since our main interest for the project is to predict the close price, table 4.2 show the descriptive statistic for the close price. The table outlines the minimum value, maximum value, median, mean and the standard deviation.

4.3 Variable selection using LASSO

The LASSO was used to select variables in this project so that a useful variable could be used to predict stock price. LASSO is a regression method for selecting variables.

Variables	Coefficients
(Intercept)	4.906668×10^{-2}
Open	5.739528×10^{-2}
High	8.381753×10^{-2}
Low	8.259775×10^{-2}
Last	8.905957×10^{-1}
Total Trade Quantity	-1.013947×10^{-8}
Turnover (Lacs)	9.439260×10^{-7}

TABLE 4.3: Parametric coefficients.

From the data set evaluated by LASSO, Table 4.3 indicates parametric coefficients and the importance of the variables in stock price prediction. As a result, according to the LASSO, all variables are important in predicting stock price.

4.4 Machine learning models

Predictions of stock prices are extremely important. We could improve return on investment if we could come up with a forecasting method. The stock market is a multi-dimensional monstrosity full of contradictions and interdependencies that is affected by a variety of factors. What if this problem could be resolved using machine learning? So far, neural networks and random forests have been shown to be capable of resolving the existing problems.

Built-in python functions were used to help train the RNN and RF models. In simulation of the model, we used python 3.0 to develop the RNN model to predict the

closing price. This research made use of the Python packages Numpy, pandas, matplotlib, scikitlearn, and tensorflow. The data was pre-processed using Scikitlearn.

As in previous dissertations, performance measures including relative mean absolute error (rMAE) and root mean square error (RMSE) will be used to simplify the method of predicting the best predicting technique. The RNN's RMSE and rMAE are 2.51 and 1.38, respectively, while the RF's are 2.44 and 1.11. As a result of the rMAE results provided in Table 4.4, the RF was determined to be the best predicting technique.

	RF	RNN
MSE	5.96	6.30
RMSE	2.44	2.51
MAE	1.23	1.90
rMAE	1.11	1.38

TABLE 4.4: Assessment of models.

The Figure 4.2 shows how the model performed. As shown in the figure, the train set is represented by the blue line, the validation set by the orange line and the prediction set by the yellow line. The data had a 80:10:10 split. Through observation it is seen that the model did well in predicting.

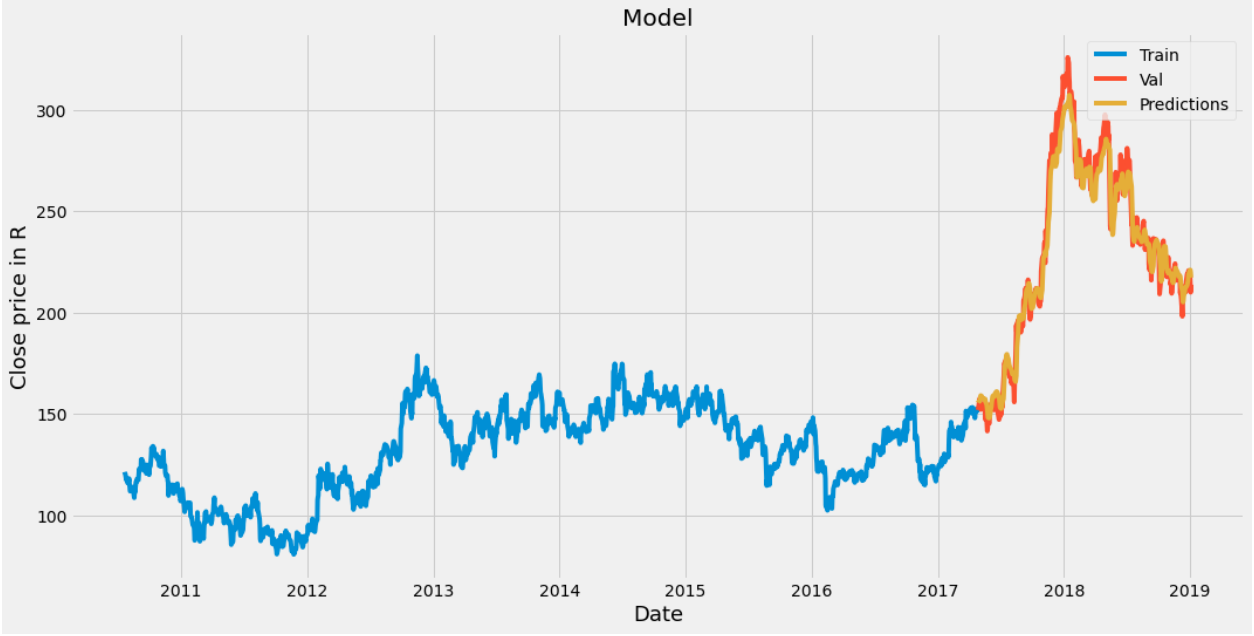

FIGURE 4.2: RNN model performance.

The Figure 4.3 show the comparison of the actual against the predicted. The actual being the blue line and the predicted represented by the orange line. Looking at these lines carefully it show that the model tried performed well.

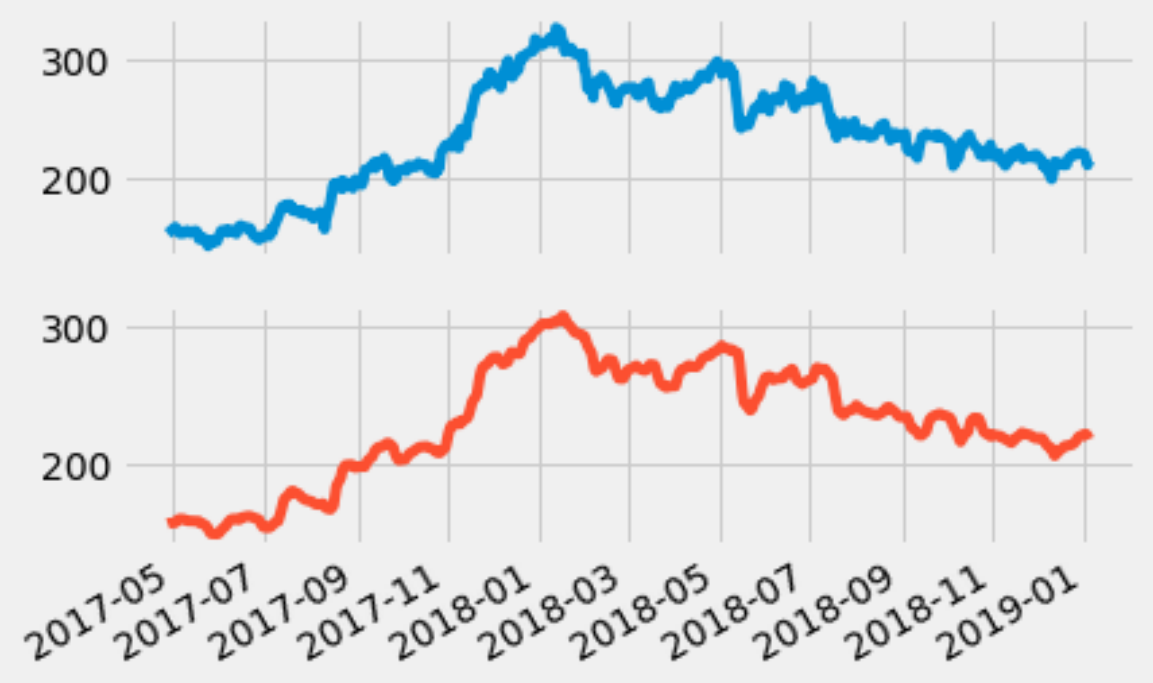

FIGURE 4.3: Comparing Actual Against Predicted in (RNN).

The Figure 4.4 represents the obtained regression predictions results of the Rf method. It show that the Rf model was able to predict the closing price almost accurately. As like in Figure 4.2 the train set is represented by the blue line, the validation set by the orange line and the prediction set by the yellow line.

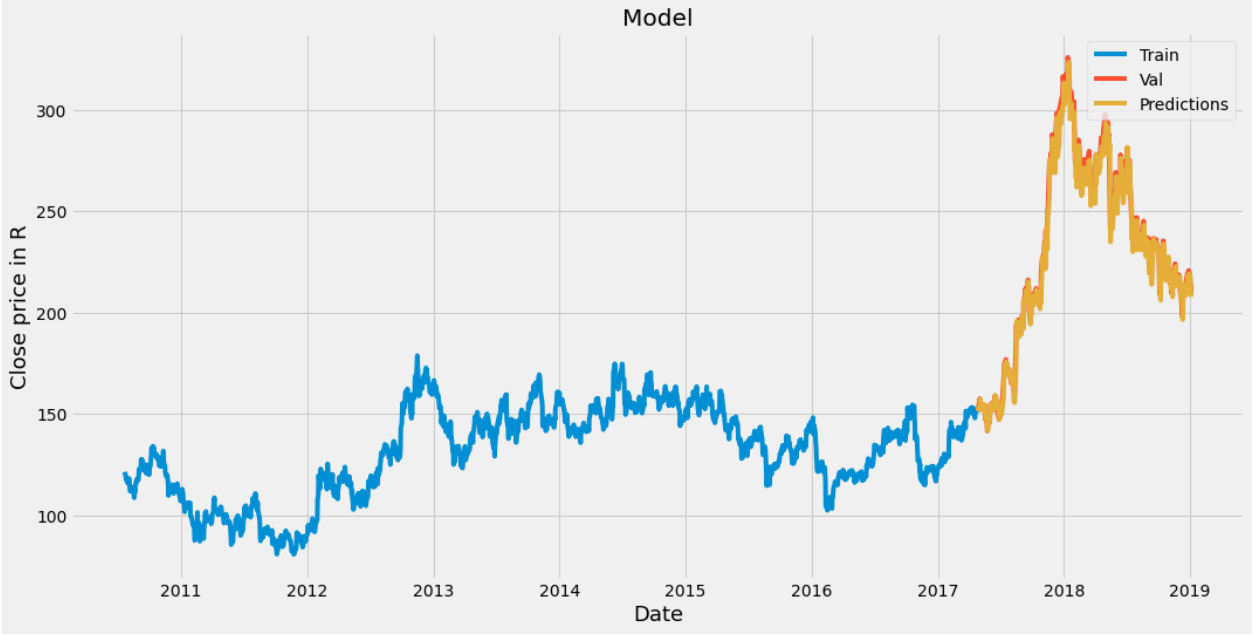

FIGURE 4.4: RF model performance.

The Figure 4.5 show the comparison of the actual against the predicted. The actual being the blue line and the predicted represented by the orange line. Looking at these lines carefully it show that the model tried performed well.

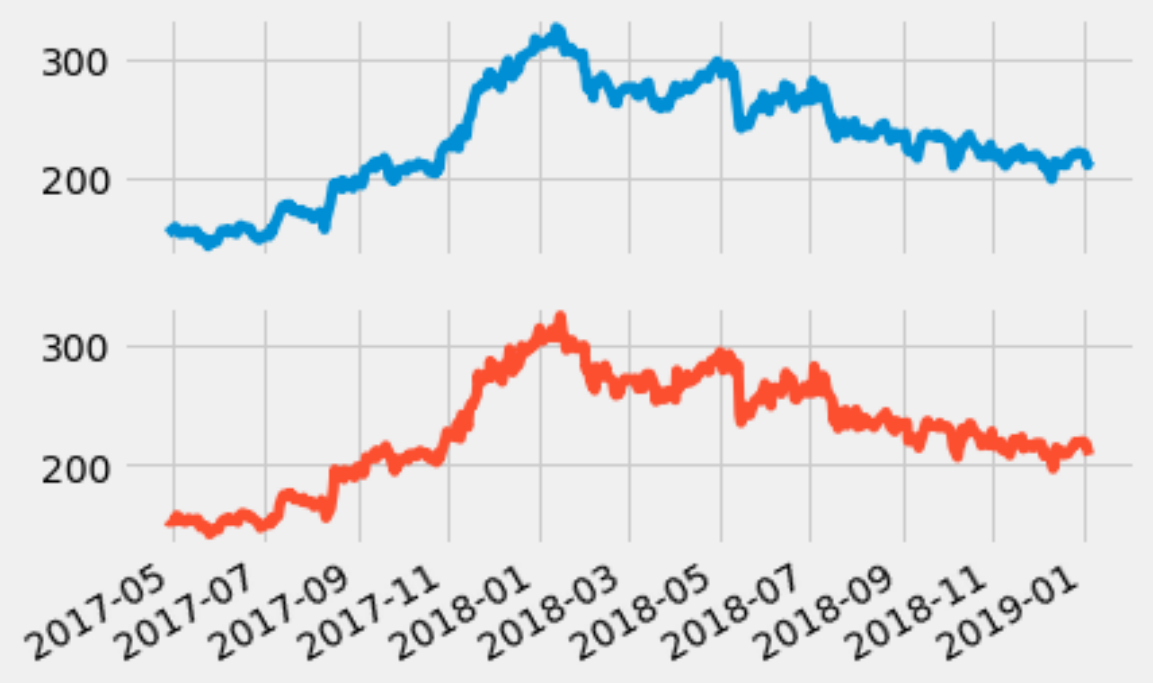

FIGURE 4.5: Comparing Actual Against Predicted in (RF).

In comparison with Figure 4.2 and 4.3 it is clearly seen that that the Rf model outperformed the RNN model.

4.5 Conclusion

This chapter presented the data analysis with the assistance of the algorithms discussed in the previous Chapter(3). The performance of the RNN and RF algorithms was compared for the prediction of the closing price. The Rf model was found to be the best model for predicting the closing price.

Chapter 5

Conclusions and Future Work

5.1 Introduction

This chapter summarizes the findings of the dissertation presented in Chapter 4 as well as some relevant concepts, such as potential research areas.

5.2 Findings

This research project focused on stock price prediction on data from an online source(Kaggle). The data is of from 21 July 2010 to 04 January 2019.

The goal of this study is to forecast the direction of stock price movement. The performance of two models, recurrent neural networks (RNN) and random forests (RF), in terms of prediction. In Chapter 4, we addressed stock price prediction using RNN and RF, as well as variable selection using the least absolute shrinkage and selection operator (LASSO).

In terms of relative root mean square square error (rRMSE) and relative mean square error (RMSE), the findings in Chapter 4 Section 4.4 showed that RF is the best predicting model as compared to the RNN. The Figures 4.2 and 4.4 shows how well the models performed. As well as the graphs in Figures 4.3 and 4.5 show the comparison between the actual and the predicted. It was concluded that the RF model outperformed the RNN model due to these graphs.

Above all, the proposed approach's success, which is based on a human approach to investing, motivates the development of expert systems and the use of machine learning algorithms for challenges in a variety of different domains to replicate human ways to decision making.

5.3 Future Work

Deep learning models that evaluate financial news stories as well as financial data such as a closing price, trading volume, profit and loss statements, and other financial criteria could be developed in the future for potentially superior results. Furthermore, deep learning methods such as Long Short Term Memory(LSTM) and 1Dimensional-Convolutional Neural Network (1D-CNN) can be used to continue this work in the future.

5.4 Conclusion

Due to constantly changing stock values that are dependent on various parameters that produce complicated patterns, stock price prediction is a difficult process. The findings and future research directions were summarized in this chapter. Machine Learning algorithms were used after all of the results had been included. After analyzing the data, we concluded that while all of these algorithms are capable of accurately predicting stock prices, the Random Forest method was the better of the two. Traditional algorithms and systems may not be able to tackle difficulties connected with this massive amount of data quickly, causing systems to operate slowly and unable to produce the best and most accurate forecast results. However, using the Python environment, we can handle vast amounts of data quickly and efficiently without having to change the techniques in the current processes.
This research project together with its findings will have an impact in increasing market efficiency. This will also promote potential economic growth.

Bibliography

[1] Masaya Abe and Hideki Nakayama. "Deep learning for forecasting stock returns in the cross-section". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2018, pp. 273–284.

[2] Wei Bao, Jun Yue, and Yulei Rao. "A deep learning framework for financial time series using stacked autoencoders and long-short term memory". In: *PloS one* 12.7 (2017), e0180944.

[3] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *Test* 25.2 (2016), pp. 197–227.

[4] Ching-Hsue Cheng, Tai-Liang Chen, and Liang-Ying Wei. "A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting". In: *Information Sciences* 180.9 (2010), pp. 1610–1629.

[5] Rohit Choudhry and Kumkum Garg. "A hybrid machine learning system for stock market forecasting". In: *World Academy of Science, Engineering and Technology* 39.3 (2008), pp. 315–318.

[6] Luca Di Persio and Oleksandr Honchar. "Artificial neural networks architectures for stock price prediction: Comparisons and applications". In: *International journal of circuits, systems and signal processing* 10.2016 (2016), pp. 403–413.

[7] Sulochana Dissanayake and Mihiri Wickramasinghe. "Earnings Fluctuation on Share Price Volatility". In: *Account and Financial Management Journal* 1.5 (2016).

[8] Valeria Fonti and Eduard Belitser. "Feature selection using lasso". In: *VU Amsterdam Research Paper in Business Analytics* 30 (2017), pp. 1–25.

[9] Kim Soon Gan et al. "Homogeneous ensemble feedforward neural network in CIMB stock price forecasting". In: *2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. IEEE. 2018, pp. 1–6.

[10] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[11] Md Rafiul Hassan and Baikunth Nath. "Stock market forecasting using hidden Markov model: a new approach". In: *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*. IEEE. 2005, pp. 192–196.

[12] Tim Hill et al. "Artificial neural network models for forecasting and decision making". In: *International journal of forecasting* 10.1 (1994), pp. 5–15.

[13] Chun-Pin Hsu. "The influence of foreign portfolio investment on domestic stock returns: Evidence from Taiwan". In: *The International Journal of Business and Finance Research* 7.3 (2013), pp. 1–11.

[14] Ozgur Ican, Taha Bugra Celik, et al. "Stock market prediction performance of neural networks: A literature review". In: *International Journal of Economics and Finance* 9.11 (2017), pp. 100–108.

[15] Minqi Jiang et al. "An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms". In: *Physica A: Statistical Mechanics and its Applications* 541 (2020), p. 122272.

[16] YAO JingTao and TAN Chew Lim. "Guidelines for Financial Prediction with Artificial neural networks". In: (2009).

[17] K Senthamarai Kannan et al. "Financial stock market forecast using data mining techniques". In: *Proceedings of the International Multiconference of Engineers and computer scientists*. Vol. 1. 2010, p. 4.

[18] Yakup Kara, Melek Acar Boyacioglu, and Ömer Kaan Baykan. "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange". In: *Expert systems with Applications* 38.5 (2011), pp. 5311–5319.

[19] Kyoung-jae Kim, Ingoo Han, and Jonh S Chandler. "Extracting trading rules from the multiple classifiers and technical indicators in stock market". In: *The Korea Society of Management Information Systems' 98 International Conference on IS Paradigm reestablishment*. The Korea Society of Management Information Systems. 1998.

[20] Avery Leider. "Computer Science Predicts the Stock Market". In: ().

[21] Zachary C Lipton, John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning". In: *arXiv preprint arXiv:1506.00019* (2015).

[22] Shingo Mabu, Masanao Obayashi, and Takashi Kuremoto. "Ensemble learning of rule-based evolutionary algorithm using multi-layer perceptron for supporting decisions in stock trading problems". In: *Applied soft computing* 36 (2015), pp. 357–367.

[23] Kumar Manish and M Thenmozhi. "Forecasting stock index movement: A comparison of support vector machines and random forest". In: *Proceedings of ninth Indian institute of capital markets conference, Mumbai, India*. 2005.

[24] Shikha Mehta et al. "Ensemble learning approach for enhanced stock prediction". In: *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE. 2019, pp. 1–5.

[25] Placido M Menaje. "Impact of selected financial variables on share price of publicly listed firms in the Philippines". In: *American International Journal of Contemporary Research* 2.9 (2012), pp. 98–104.

[26] Mahla Nikou, Gholamreza Mansourfar, and Jamshid Bagherzadeh. "Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms". In: *Intelligent Systems in Accounting, Finance and Management* 26.4 (2019), pp. 164–174.

[27] Isaac Kofi Nti, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. "A comprehensive evaluation of ensemble learning for stock-market prediction". In: *Journal of Big Data* 7.1 (2020), pp. 1–40.

[28] Jigar Patel et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques". In: *Expert systems with applications* 42.1 (2015), pp. 259–268.

[29] Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017.

[30] Bo Qian and Khaled Rasheed. "Stock market prediction with multiple classifiers". In: *Applied Intelligence* 26.1 (2007), pp. 25–33.

[31] Ashish Sharma, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach". In: *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. Vol. 2. IEEE. 2017, pp. 506–509.

[32] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. "Stock market forecasting using machine learning algorithms". In: *Department of Electrical Engineering, Stanford University, Stanford, CA* (2012), pp. 1–5.

[33] Ryuhei Shiomi et al. "A study on operating lifetime estimation for electrical components in power grids on the basis of analysis of maintenance records". In: *Journal of International Council on Electrical Engineering* 9.1 (2019), pp. 45–52.

[34] Pinky Sodhi, Naman Awasthi, and Vishal Sharma. "Introduction to Machine Learning and Its Basic Application in Python". In: *Available at SSRN 3323796* (2019).

[35] Sahar Sohangir et al. "Big Data: Deep Learning for financial sentiment analysis". In: *Journal of Big Data* 5.1 (2018), pp. 1–25.

[36] Yu Tang et al. "Web-based fuzzy neural networks for stock prediction". In: *Proceedings of Second International Workshop on Intelligent Systems Design and Application*. 2002, pp. 169–174.

[37] Chang Sim Vui et al. "A review of stock market prediction with Artificial neural network (ANN)". In: *2013 IEEE international conference on control system, computing and engineering*. IEEE. 2013, pp. 477–482.

[38] Qili Wang et al. "Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning". In: *Neurocomputing* 347 (2019), pp. 46–58.

[39] Sourabh Yadav and Nonita Sharma. "Homogenous ensemble of time-series models for indian stock market". In: *International Conference on Big Data Analytics*. Springer. 2018, pp. 100–114.

[40] Bing Yang, Zi-Jia Gong, and Wenqi Yang. "Stock market index prediction using deep neural network ensemble". In: *2017 36th Chinese Control Conference (CCC)*. IEEE. 2017, pp. 3882–3887.

[41] Tiffany Hui-Kuang Yu and Kun-Huang Huarng. "A neural network-based fuzzy time series model to improve forecasting". In: *Expert Systems with Applications* 37.4 (2010), pp. 3366–3372.

[42] MH Fazel Zarandi et al. "A type-2 fuzzy rule-based expert system model for stock price analysis". In: *Expert Systems with Applications* 36.1 (2009), pp. 139–154.

[43] Xiangzhou Zhang et al. "A causal feature selection algorithm for stock prediction modeling". In: *Neurocomputing* 142 (2014), pp. 48–59.

[44] Zhi-yong Zhang et al. "Stock time series forecasting using support vector machines employing analyst recommendations". In: *International Symposium on Neural Networks*. Springer. 2006, pp. 452–457.

[45] You Zhu et al. "Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance". In: *Neural Computing and Applications* 28.1 (2017), pp. 41–50.