A

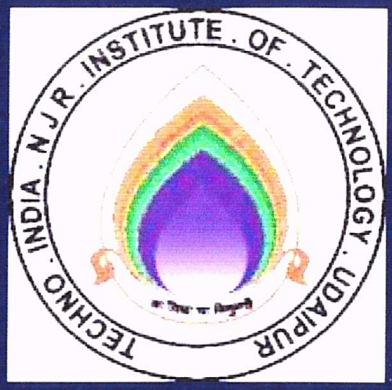
*MAJOR PROJECT REPORT*

*on*

**Stock Market prediction using LSTM**

*Submitted in partial fulfilment of the requirements of the degree of*

**BACHELOR OF TECHNOLOGY**




Under Guidance of

Mr. Yogendra Solanki
Ass. Professor
Electronics & Communication Engineering
TINJRIT, Udaipur

Submitted by

Jayesh Trivedi, 17ETCEC009
Anupreet Dube, 17ETCEC004
Preet Jhota, 17ETCEC014
Yash Jasnani, 17ETCEC023

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**

**TECHNO INDIA NJR INSTITUTE OF TECHNOLOGY, UDAIPUR**

**Sep - 2021**

Department of Electronics and Communication Engineering
Techno India NJR Institute of Technology, Udaipur

# Certificate

This is to certify that this Major Project report entitled **Stock Market Prediction Using LSTM** by **Jayesh Trivedi, Anupreet Dube, Preet Jhota, Yash Jasnani** have completed the work under my supervision and guidance, hence approved for submission in partial fulfilment for the award of degree of Bachelor of Technology in Electronics and Communication to the Department of Electronics and Communication Engineering, Techno India NJR Institute of Technology, Udaipur during academic session 2017-2021.

Mr. Yogendra Solanki
Ass. Professor
Dept. of E.C.E, TINJRIT, Udaipur
Date.....

Mr. Pradeep Chhawcharia
Head of Department
Dept. of E.C.E TINJRIT, Udaipur
Date.....

Department of Electronics and Communication Engineering
Techno India NJR Institute of Technology, Udaipur

## Examiner certificate

This is to certify that the following students

**Jayesh Trivedi**

**Anupreet Dube**

**Preet Jhota**

**Yash Jasnani**

of final year B.Tech. (Electronics & Communication Engineering), were examined for the project work entitled

***Stock Market Prediction Using LSTM***

during the academic year 2017 – 2021 at Techno India NJR Institute of Technology, Udaipur

**Remarks:**

**Date:**

Signature
(**Internal Examiner**)

Signature
(**External Examiner**)

# PREFACE

Predicting the Stock Market has been the bane and goal of investors since its existence. Everyday billions of dollars are traded on the exchange, and behind each dollar is an investor hoping to profit inone way or another. Despite its prevalence, Stock Market prediction remains a secretive and empirical art. Few people, ifany, are willing to share what successful strategies they have. Achieve goal of this project is to add to the academic understanding of stock market prediction. The hope is that with a greater understanding ofhow the market moves, investors will be better equipped to prevent another financial crisis. The project will evaluate some existing strategies from a rigorous scientific perspective and provide aquantitative evaluation of new strategies.

# ACKNOWLEDGMENT

We take this opportunity to record our sincere thanks to all who helped us to successfully complete this work. Firstly, we are grateful to our **supervisor Mr. Yogendra Singh Solanki** for his invaluable guidance and constant encouragement, support and most importantly for giving us the opportunity to carry out this work.

We would like to express our deepest sense of gratitude and humble regards to our

**Head of Department Mr. Pradeep Chhawcharia** for giving invariable encouragement in our endeavours and providing necessary facility for the same. Also, a sincere thanks to all faculty members of ECE, TINJRIT for their help in the project directly or indirectly.

Finally, we would like to thank my friends for their support and discussions that have proved very valuable for us. We are indebted to our parents for providing constant support, love and encouragement. We thank them for the sacrifices they made so that we could grow up in a learning environment. They have always stood by us in everything we have done, providing constant support, encouragement and love

**Jayesh Trivedi, 17ETCEC009**

**Anupreet Dube, 17ETCEC004**

**Preet Jhota, 17ETCEC014**

**Yash Jasnani, 17ETCEC023**

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
TECHNO INDIA NJR INSTITUTE OF TECHNOLOGY, UDAIPUR

# TABLE OF CONTENTS

# LIST OF FIGURES

## LIST OF TABLES



## LIST OF ACRONYMS

| Serial Number | ACRONYM | FULL FORM |
|---|---|---|
| 1 | SVM | Support Vector Machine |
| 2 | KNN | K Nearest Neighbor |
| 3 | LSTM | Long Short-Term Memory |
| 4 | RNN | Recurrent Neural Network |
| 5 | NSE | National Stock Exchange |
| 6 | BSE | Bombay Stock Exchange |
| 7 | ML | Machine Learning |
| 8 | ANN | Artificial Neural Network |

# ABSTRACT

In this report we analyze existing and new methods of stock market prediction. We take three different approaches at the problem: Fundamental analysis, Technical Analysis, and the application of Machine Learning. We find evidence in support of the weak form of the Efficient Market Hypothesis, that the historic price does not contain useful information but out of sample data may be predictive. We show that Fundamental Analysis and Machine Learning could be used to guide an investor's decisions. We demonstrate a common flaw in Technical Analysis methodology and show that it produces limited useful information. Based on our findings, algorithmic trading programs are developed and simulated using RNN and LSTM.

Predicting how the stock market will perform is one of the most difficult things to do. There are so many factors involved in the prediction – physical factors vs. physiological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

Stock market analysis is divided into two parts – Fundamental Analysis and Technical Analysis.

- Fundamental Analysis involves analyzing the company's future profitability on the basis of its current business environment and financial performance.
- Technical Analysis, on the other hand, includes reading the charts and using statistical figures to identify the trends in the stock market.

In this Project we have used LSTM, which stands for Long short term. LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:

- **The input gate:** The input gate adds information to the cell state
- **The forget gate:** It removes the information that is no longer required by the model
- **The output gate:** Output Gate at LSTM selects the information to be shown as output

**Key Words:** RNN, LSTM, Stock Market

# CHAPTER 1

# INTRODUCTION

Predicting the Stock Market has been the bane and goal of investors since its existence. Everydaybillions of dollars are traded on the exchange, and behind each dollar is an investor hoping to profit inone way or another. Entire companies rise and fall daily based on the behavior of the market. Should an investor be able to accurately predict market movements, it offers a tantalizing promise of wealth andinfluence. It is no wonder then that the Stock Market and its associated challenges find their way intothe public imagination every time it misbehaves. The 2008 financial crisis was no different, as evidenced by the flood of films and documentaries based on the crash. If there was a common themeamong those productions, it was that few people knew how the market worked or reacted. Perhaps abetter understanding of stock market prediction might help in the case of similar events in the future.

## 1.1Project Goals and Scope

Despite its prevalence, Stock Market prediction remains a secretive and empirical art. Few people, ifany, are willing to share what successful strategies they have. Achieve goal of this project is to add to the academic understanding of stock market prediction. The hope is that with a greater understanding ofhow the market moves, investors will be better equipped to prevent another financial crisis. The project will evaluate some existing strategies from a rigorous scientific perspective and provide aquantitative evaluation of new strategies.

It is important hereto define the scope of the project. Although vital to any investor operating in thereal world, no attempt is made in this project at portfolio management. Portfolio management is largely an extra step done after an investor has made a prediction on which direction any particular stockwill move.

The investor may choose to allocate funds across a range of stocks in such a way to minimize his or herrisk. For instance, the investor may choose not to invest all of their funds into a single company unless that company takes unexpected turn. A more common approach would be for an investor to invest across abroad range of stocks based on some criteria he has decided on before. This project will focus exclusively on predicting the daily trend (price movement) of individual stocks. The project will make no attempt to deciding how much money to allocate to each prediction. More so, the project will analyze the accuracies of these predictions.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 : REVIEW PROCESS ADOPTED

### 2.2 PAPER1

By Prof.:-**David M. Q. Nelson,**

Predictions on stock markets have been object of studies for many decades, but given its innate complexity, dynamism and chaoticness, it has proven to be a very difficult task. The number of variables and sources of information considered are immense and the signal-to-noise ratio insignificant. That makes the task of predicting stock market prices behavior in the future a very hard one. For many decades, there's been discussions in Science regarding the possibility of such a feat andit's notable in the related literature that most prediction models fail to provide precise prediction in a general sense. Nonetheless, there is huge amount of studies from various disciplines seeking to take on that challenge, presenting a large variety of approaches to reach that goal. One common approach is touse Machine Learning algorithms to learn from price historic data to predict future prices. This articlegoes in that direction but studying an specific method using recurrent neural networks. Such networkshave a short-term memory capability and the hypothesis to explore here is that this feature can present gains in terms of results when compared to others more traditional approaches in Machine Learningfield. The algorithm of choice here is the LSTM (Long-Short term memory) network. It's a type ofrecurrent network that has proved very successful on a number of problems given its capability todistinguish between recent and early examples by giving different weights for each while forgettingmemory it considers irrelevant to predict the next output. In that way, it is more capable to handle longsequences of input when compared to other recurrent neural networks that are only able to memorize short sequences.

**PAPER2**

By: -**Sreelekshmy Selvin, Vinayakumar R, Gopalakrishnan E.A,**

Forecasting can be defined as the prediction of some future event or events by analyzing the historical data. It spans many areas including business and industry, economics, environmental science andfinance. Deep learning algorithms are capable of identifying hidden patterns and underlying dynamics in the data through a self-learning process. In the case of stock market, the data generated isenormous and is highly non-linear. To model such kind of dynamical data we need models that canalyze the hidden patterns and underlying dynamics. Deep learning algorithms are capable of identifying and exploiting the interactions and patterns existing in a data through a self-learning process. Unlike other algorithms, deep learning models can effectively model these type of data andcan give a good prediction by analyzing the interactions and hidden patterns within the data. In , we can see the application of various deep learning models for multivariate time series analysis. The firstattempt to model a financial time series using a neural network model was introduced. This work made an attempt to model a neural network model for decoding the nonlinear regularities in asset price movements for IBM. However, the scope of the work was limited, but it helped in establishing evidences against EMH. Researches in the area of financial time series analysis using NN models used different input variables for predicting the stock return. In some works, data from a single time serieswere used as input. Certain works considered the inclusion of heterogeneous market information and macroeconomic variables. In a combination of financial time series analysis and NLP have been introduced. In deep learning architectures have been used for the modelling of multivariate financialtime series. In a NN model using technical analysis variables have been implemented for the prediction of Shanghai stock market. The work compared the performance of two learning algorithms and two weight initialization methods. The results shown that efficiency of back propagation can beincreased by conjugate gradient learning with multiple linear regression weight initialization.

**PAPER3**

By:-**Osman Hegazy 1, Omar S. Soliman 2 and Mustafa Abdul Salam**

STOCK price prediction has been at focus for years since t can yield significant profits. Predicting the stock market is not a simple task, mainly as a consequence of the close to random-walk behavior of a stock time series. Fundamental and technical analyses were the first two methods used to forecast stock prices. Artificial Neural networks (ANNs) i s the most commonly used technique. In most cases ANNs suffer fromover-fitting problem due to the large number of parameters to fix, and the little prior user knowledge about the relevance of the inputs n the analyzed problem. Also, Support vector machines (SVMs) had been developed as an alternative that avoids such limitations. Theirpractical successes can be attributed to solid theoretical foundations based on VC-theory. SVM compute globally optimal solutions, unlike those obtained with ANNs, which tend to fall into local minima. Least squares –support vector machines (LS-SVM) method was presented in, which was reformulated the traditional SVM algorithm. LS-SVM uses a regularized least squares function with equality constraints, leading to a linear system which meets the Karush-Kuhn-Tucker (KKT) conditions for obtaining an optimal solution. Although LS-SVM simplifies the SVM procedure, the regularization parameter and the kernel parameters play an important role in the regression system. Therefore, t s necessary to establish a methodology for properly selecting the LS-SVM free parameters, n such a way that the regression obtained by LS-SVM must be robust against noisy conditions, and it does not need priori user knowledge about the influence of the free parameters values n the problem studied. to develop a machine learning model that hybrids the PSO and LS-SVM model.The performance of LS-SVM s based on the selection of free parameters C (cost penalty), $\epsilon$ (insensitive-loss function) and $\gamma$ (kernel parameter). PSO will be used to find the best parameter combination for LS-SVM.

**PAPER4**

**By:-Shen et al.**

The article makes a case for the use of machine learning to predict large American stock indices, including the Dow Jones Industrial Average. The article boasts a 77.6% accuracy rate for the DowJones specifically. The team uses a set of 16 financial products and and uses their movements to predict movements in American stock exchanges. Many of these products will be used later in thisreport. Some of the financial products used are listed below.

- FTSEindexprice
- DAX indexprice
- Oilprice
- EURO/USDexchangerate

The article makes good use of explorative methods in their data preparation stage. They show usinggraphs there is a some features may have predictive power because of their correlation to theNASDAQ index. They then go on to perform feature selection based on the predictive power of each feature on its own. The results presented in this section, the predictive power of single features, arevery similar to results that we will show later in this report. After they have selected their top 4 features, they compare a Support vector machine model to a Multiple Additive Regression Trees model forpredicting the daily NASDAQ trend. Their winning model is the SVM with 74.4% accuracy.

While the results presented by Shen et al. in this article appear to be very much in line with results thatwill be later presented in this report, it is quite vague in terms of the methodology that was used. Forinstance, there is no record of what model was used in the feature selection step or how cross- validation was carried out to calculate any of their results. However, their results will be supported byresults in this report, so it is safe to assume that no critical errors were made in these steps.

While training and testing the models, both the article and this report ignore the overlapping

of tradinghours. However, this should certainly not be ignored when estimating real world trading performance.

## PAPER 5

By:-**Kara et al.**

The article uses technical analysis indicators to predict the direction of the ISE National 100 Index, anindex traded on the Istanbul Stock Exchange. The article claims impressive results, up to 75.74%accuracy.

Technical analysis is a method that attempts to exploit recurring patterns and trends within the price of the stock itself. It goes directly against all forms of the Efficient Market Hypothesis. As described previously, even the weak form of the Hypothesis rules out prediction using historic price data alone.The team uses a set of 10 technical analysis indicators which are listed below.

- Simple 10-daymovingaverage
- Weighted 10-daymovingaverage
- Momentum
- Stochastic K%
- Stochastic D%
- Relative Strength Index
- Moving Average Convergence Divergence
- Larry Williams R%
- Accumulation Distribution Oscillator
- Commodity Channel Index

Thedailyvaluesfortheindicatorsarecalculatedandcoupledwiththedailypricemovement direction, the dependent variable. Two types of model are tested; a support vector machine and a neural network. Results are cross- validated using a single-holdout method. This method of cross-validation is known to be inferior when compared too the

techniques such as k-fold cross- validation[12], but it is unlikely that this would have a drastic effect on the results presented in the article. The team's neural networkmodel had an accuracy of 75.74% and their support vector machine had an accuracy of 71.52%.

## PAPER 6

*[Elsevier B.V 2016]:* Stock market price data is generated in huge volume and it changes every second. Stock market is a complex and challenging system where people will either gain money or lose their entire life savings. In this work, an attempt is made for prediction of stock market trend. Two models are built one for daily prediction and the other one is for monthly prediction. Supervised machine learning algorithms are used to build the models. As part of the daily prediction model, historical prices are combined with sentiments. Up to 70% of accuracy is observed using supervised machine learning algorithms on daily prediction model. Monthly prediction model tries to evaluate whether there is any similarity between any two months trend. Evaluation proves that trend of one month is least correlated with the trend of another month

## PAPER 7

[ *E. F. Fama , 1995 ] :* In order to obtain stable returns, this paper aims to establish a quantitative model with higher prediction accuracy. According to the time series characteristicsof financial data, the prediction model with financial time series data is constructed by using the time-memory sequence model LSTM, which is applied to the representative SSE 50 series stocks. And based on this, the LSTM model with Encoder-Decoder mode and a hybridized framework of LSTM with CNN are built to improve the original model. Feature extraction is performed on the input data by using CNN, and then as an input to the LSTM, the extracted features are used for sequence prediction with the LSTM model.

## PAPER 8

*[S. A. R. Nai-Fu Chen and Richard Roll, 1986 ] :* Prediction model can be applied on the historical data to get future trend. As researchers have discussed in S. J. Grossmara and R. J. Shiller3 and L. Andrew and M. A. Craig4, as and when new information comes in the market stock market value varies. Technical analysis and semi strong form of efficient market hypothesis are followed, to build prediction model in the proposed work.

The goal of this research work is to build a model which predicts stock trend movement (trend will be up or down) using historical data and social media data. Two models are built as part of research work.Both models use supervised machine learning algorithm. First model is daily prediction model, considers both sentiment and historical data.

## PAPER 9

***[S. J. Grossman and R. J. Shiller , 1980] :*** This model predicts the future trend for the next day. Sentiment of the company has been computed by using twitter data and news of the company. Outcome of sentiment analysis is considered along with open price, close price of stock with extracted statistical parameters to build model. Second model is monthly prediction model, considers only historical data and predicts the trend for next one month. Proposed work 3 investigates whether the outcome of model is inline with the actual trend movement. The rest of this paper is organized is as follows. Section 2 introduces some previous research work on sentiment analysis for stock market prediction and stock trend movement using historical price. Section 3 describes proposed method. Section 4 shows the dataset used and evaluates the results of the experiments. Finally, Section 5 concludes the contribution of this research work.

## PAPER 10

***[ Engle, Robert F. , 1982]:*** Although stocks are highly real-time and unstable, the large amount of historical data, as an objective reaction of the stock market, necessarily indicates the future trend of the stock market to a certain extent. Experts, scholars and investment institutions aim to obtain a certain degree of prediction on the stock market, and get higher benefits by analyzing these data. To this end, numerous experts and scholars at home and abroad have proposed a variety of methods to predict the changing trend of the stock market and build a stockforecasting model. Statistical and econometric models such as multiple regression and exponential smoothing [1]. However, the extremely complex nonlinear characteristics of the stock market make the limitations of the traditional statistical model in dealing with nonlinear problems, and the prediction effect is not satisfactory.

## PAPER 11

[Dou Wei] (2019 International Conference on Artificial Intelligence and Advanced

Manufacturing (AIAM)) This study, based on the demand for stock price prediction and the practical problems it faces, compared and analyzed a variety of neural network predictionmethods, and finally chose LSTM (Long Short-Term Memory, LSTM) neural network. Then, through in-depth study on how to predict the stock price by the LSTM neural network optimized by MBGD algorithm, the feasibility of the method and the applicability of the model areanalyzed, and finally the conclusion is drawn. It is found that historical information is very important to investors as the basis of investment decisions. Past studies have used opening and closing prices as key new predicators of financial markets, but extreme maxima and minima mayprovide additional information about future price behavior. Therefore, the index of three representative stocks in China's stock market are selected as the research objects, and the key data collected from them include the opening price, closing price, lowest price, highest price,date and daily trading volume. The results show that although LSTM neural network model has some limitations, such as the time lag of prediction, but with attention layer, it can predict stock prices. Its main principle is to discover the role of time series through analyzing the historical information of the stock market, and to deeply explore its internal rules through the selective memory advanced deep learning function of LSTM neural network model, so as to achieve the prediction of stock price trend. **INSPEC Accession Number: 19264745**

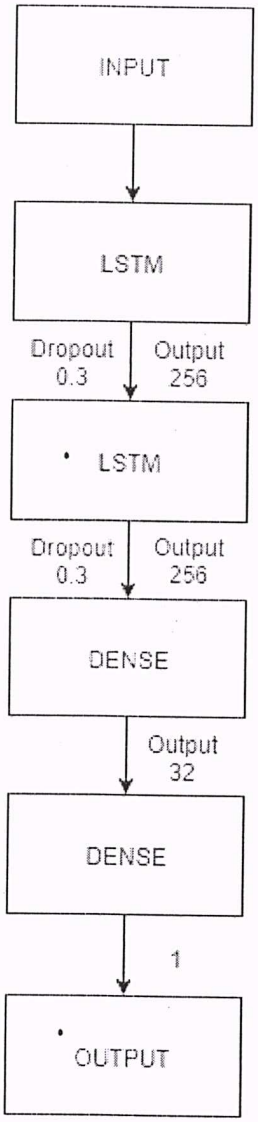
## PAPER 12

**[Ishita Parmar]** (2018 First International Conference on Secure Cyber Computing and Communication(ICSCCC)) Stock market prediction seems a complex problem because there are many factors that have yet to be addressed and it doesn't seem statistical at first. But by proper use of machine learning techniques, one can relate previous data to the current data and train the machine to learn from it and make appropriate assumptions. The dataset being utilized for analysis was picked up from Yahoo Finance. The dataset consisted of approximately 9 lakh records of the required stock prices and other relevant values. The data reflected the stock prices at certain time intervals for each day of the year. It consisted of various sections namely date, symbol, open, close, low, high and volume. For the purpose of simulation and analysis, the data for only one company was considered. All the data was available in a file of *csv* format which was first read and transformed into a data-frame using the *Pandas* library in Python. From this,the data for one particular company was extracted by segregating data on the basis of the symbol field. Following this normalization of the data was performed through usage of the *sklearn* library in Python and the data was divided into training and testing sets. The test set was kept as 20% of the available dataset. Although machine learning as such has many models but this paperfocuses on two of the most important amongst them and made the

predictions using these.

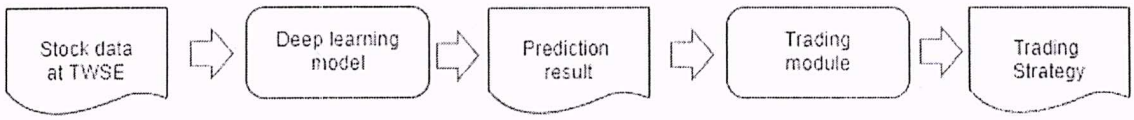
## PAPER 13

**[Samuel Olusegun Ojo]** A stacked LSTM model is trained on the data, and evaluated by measuring the Mean Squared Error (MSE) and Mean Absolute Diviation (MAD) of the model. 90% of the data is used as training while the remaining 10% is used as testing and validation.



## PAPER 14

[Li-Chen Cheng] ( 2018 IEEE International Conference on Big Data (Big Data)) Stock marketprediction is traditionally one of the most challenging time series prediction tasks, as stock datais noisy and non-stationary. Stock return predictions are broadly classified into linear and non- linear models: linear autoregressive integrated moving average models, exponential smoothing models, and the generalized autoregressive conditional heteroscedasticity model. Non-linearmodels include SVMs, genetic algorithms, and state-of-the-art deep learning. Akita build an LSTM model using textual and numerical information to predict ten company's closing stock prices . Ding propose a deep convolutional neural network using event embeddings whichcombines the influence of long-term events and short-term events to predict stock prices . Nelsonpropose a model based on LSTM using five historic price measures (open, close, low, high, volume) and 175 technical indicators to predict stock price movement . Fischer apply an LSTM model to a large-scale financial market prediction task on S&P 500 data from December 1992 until October 2015. They show that the LSTM model outperforms the standard deep net and traditional machine learning methods INSPEC Accession Number: 18412109




## PAPER 15

[CHUN YUAN LAI]

Predicting stock price has been a challenging project for many researchers, investors, and analysts. Most of them are interested in knowing the stock price trend in the future. To get a precise and winning model is the wish of them. Recently, Neural Network has been a prevalent means for stock prediction. However, there are many ways and different predicting models such as Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). In this paper, we propose a novel idea that average previous five days stock market

information (open, high, low, volume, close) as a new value then use this value to predict, and use the predicted value as the average of the stock price information for the next five days. Moreover, we utilize Technical Analysis Indicators to consider whether to buy stocks or continue to hold stocks or sell stocks. We use Foxconn company data collected from Taiwan Stock Exchange for testing with the Neural Network Long Short-Term Memory (LSTM).

**INSPEC Accession Number:** 19247354

## 2.3 : Gaps

| **Author Name** | **Gaps** |
|---|---|
| Osman Hegazy 1, Omar S.Soliman 2 and Mustafa Abdul Salam | ANN |
| Sreelekshmy Selvin, Vinayakumar R, Gopalakrishnan E.A, | Hidden nputs |
| Shen et al. | Accuracy |

## 2.4 :Problem and Solution of Paper

| Author Name | Problem | Solution |
|---|---|---|
| David M. Q. Nelson, | How can we improve the LSTMtime series prediction? | After forecasting through LSTM, this difference has to be added back to get the original series. |
| Sreelekshmy Selvin ,VinayakumarR, Gopalakrishnan E.A, | Non- linear data prediction in hidden inputs in self learning | Deep learning algorithms analysing the interaction and hidden pattern within data. |
| Osman Hegazy 1, Omar S. Soliman 2 and Mustafa Abdul Salam | In most cases ANNs suffer fromover-fitting problem due to the large number of parameters to fix. | (SVMs) had been developed as an alternative that avoids such limitations |
| Shenetal | Accuracy in single . featurepreductive power by support vector machine i.e., 71.52% only. | K-fold cross validation withneural network had an accuracy of 75.74%. |
| Karaetal | Is time series analysis usefulfor LSTM stock prediction? | Time series analysis refers to the analysis of the changen the trend of the data over a period of time. Time series analysis has a variety of applications. |

## 2.5 : Strength and Weakness

| Author Name | Strengths | Weakness |
|---|---|---|
| David M. Q. Nelson, | Lstm asuumes that there areinput values (time series) which are to be used to predictan output value. | Layers require a large amount of memory bandwidth to be computed, because the system has not enough memory bandwidth to feed the computational units. |
| ,Sreelekshmy,Selvin , VinayakumarR, GopalakrishnanE. A, | Multivariable dat apredictionin single algorithm. | Scope is less against Neural Network Model. |
| Kara et al. | Lstm are very powerful in sequence prediction problems because they are able to store past information | They are an unsupervised learning method, although techniocally they are trained using supervised learning methods. |

# CHAPTER 3

# THEORETICAL ASPECTS

## Gathering the Datasets

A primary dataset will be used throughout the project. The dataset will contain the daily percentage change in stock price for all 30 components of the Dow Jones. Luckily, daily stock price data is easy to come by. Google and Yahoo both operate websites which offer a facility to download CSV files containing a price history. These are useful for looking at individual companies but cumber some when accessing large amounts of data cross many stocks.

For this reason, Quandl was used to gather the data instead of using Google and Yahoo directly. Quandl is a free to use website that hosts and maintains vast amounts of numerical datasets with a focus specifically on economic datasets, including stock market data which is backed by Google and Yahoo. Quandl also provides a small python library that is useful for accessing the database programmatically. The library provides a simple method for calculating the daily percentage change daily in prices. This calculation is defined in equation 4.1 where $p_t$ is the closing price on day $d$, and $\delta p_d$ is the resulting percentage change.

$$= \frac{p_d - p_{d-1}}{} \qquad (4.1$$

To build the primary dataset, a simple program was built to query the Quandl API for each of the 30 Dow Jones components using the transformation outlined in equation 4.1. Before finally saving the data, the gathered data was augmented with an additional column containing the classification of the daily percentage change. This augmentation is defined in equation 4.2 where $trend_d$ is the price movement direction on day $d$ and $\delta p_d$ as defined

in equation 4.1.

The final step is shifting all of the $\delta p_d$ and $trend_d$ data points backwards by one day. The data we have gathered for any day should be used to predict the trend tomorrow, not the same day. By shifting $\delta p_d$ and $trend_d$ backwards, data gathered in later sections will be paired with the correct dependent variable. For instance, data we gather for a Monday will be matched with, and try to predict, Tuesday's trend.

This dataset was then saved in CSV format for simple retrial as needed throughout the project. This dataset containing the daily trends of companies will serve as the core dataset that will be used in most experiments later in the report. When we want to use the dataset later with the extra data we collect for each experiment, we only need to do a simple database join operation on the company and date. The dataset contains 122,121 rows covering all 30 companies daily since January 1st 2000. Table 4.1 shows an extract of a number of rows from this dataset.

With the primary dataset prepared, we will combine it as needed with additional datasets to carry out individual experiments.

## 3.1 Limitations of the Data

Although the data gathered in the previous section is certainly a good start, it is admittedly far behind what any serious investor more than likely has access to.

One obvious piece of missing data that is the intraday prices, i.e. the prices minute by minute. It is possible that this data could be used to guide and investors decisions on the interday level. However, intraday prices are not as freely available as interday prices and are considered a commodity in themselves. To get hold of such a dataset would incur a large cost, one that is not within the budget of a project such as this. Later in the project we will evaluate a strategy in which this limitation becomes significant.

Another important piece of missing data is the order book. The order book is a record of live buy and sell orders for a particular stock. It consists of the amount of stock each trader is willing to buy or sell, as well as their price. Successful orders are matched off against the orderbook by the exchange. The price of a stock is usually considered to be half way between the highest buying price and the lowest selling price. It is easy to imagine that the order book contains useful data. For instance, the weighted average of orders might be

predictive of the price. However access to this data is extremely costly and far beyond what most casual investors can afford, let alone the budget for this project. With no way around these limitations, we use the data provided by Quandl.

## 3.2 Tools

*Python and associated packages*

Python was the language of choice for this project. This was an easy decision for the multiple reasons.

1. Python as a language has an enormous community behind it. Any problems that might be encountered can be easily solved with a trip to Stack Overflow. Python is among the most popularlanguages on the site which makes it very likely there will be a direct answer to any query [17].

2. Python has an abundance of powerful tools ready for scientific computing. Packages such as NumPy, Pandas, and SciPy are freely available, performant, and well documented. Packages such as these candramatically reduce, and simplify the code needed to write a given program. This makes iterationquick.

3. Python as a language is forgiving and allows for programs that look like pseudo code. This is usefulwhen pseudo code given in academic papers needs to be implemented and tested. Using Python, thisstep is usually reasonably trivial.
   However, Python is not without its flaws. The language is dynamically typed and packages are notorious for Duck Typing. This can be frustrating when a package method returns something that, for example, looks like an array rather than being an actual array. Coupled with the fact that standard Python documentation does not explicitly state the return type of a method, this can lead to a lot of trialand error testing that would not otherwise happen in a strongly typed language. This is an issue that makes learning to use a new Python package or library more difficult than it otherwise.

# CHAPTER 4

## TECHNICAL DESCRIPTION

The first approach we take to solving the problem of market prediction is to use Fundamental Analysis. This approach tries to find the true value of a company, and thus determine how much oneshare of that company should really be worth. The assumption then is that given enough time, themarket will generally agree with your prediction and move to correct its error. If you determine themarket has undervalued a company, then the market price should rise to correct this inefficiency, andconversely fall to correct the price of an overvalued company.

Graham et al laid the ground work for the field with the book *Security Analysis*. He encouraged would-be investors to estimate the intrinsic value of a stock before buying or selling based on trends, a novelidea at the time. It stands as testament to his approach that his only A+ student was Warren Buffet, who methodically applied the strategy and has enjoyed renowned success since. This gives us some hope,but we should be cautious and remember that the market might behave differently today than it did before.

It should be noted that Fundamental Analysis is compatible with the weak form of the Efficient Market Hypothesis. As explained earlier, the weak form does not rule out prediction from data sources external to the price, which is what we will use to determine our fair market price.

We will look at two of the most common metrics used in Fundamental Analysis, Price to EarningsRatio, and Price to Book Ratio to predict long term price movements on a year to year basis. This is the typical prediction range for Fundamental Analysis.

Due to the prediction range of Fundamental Analysis however, this chapter will not focus only on thecomponent companies of the Dow Jones. We would not be able to generate a sufficiently large dataset by looking at recent year to year changes of only 30 companies. Instead, in this chapter we will takeour sample companies from the S&P 500 index. This will provide a sufficiently large dataset to do theanalysis required in this chapter.

## 4.1 Price to Earnings Ratio

The first metric for the value of a company that we will look at is the Price to Earnings Ratio. The Price to Earnings Ratio calculation is defined in equation 5.1.

$$\text{P/E Ratio} \qquad \frac{\text{Share Price}}{\text{Earnings Per Share}}$$

Roughly speaking, what the P/E Ratio calculates is the price an investor is willing to pay for every $1 of company earnings. If this ratio is high, it might be a sign of high investor confidence. If investor confidence is high, that might mean they expect high returns in the following year. We should then expect to see a relationship between high P/E ratio and high returns in the following year.

To investigate this relationship, we plotted the P/E ratio of 456 companies on the 31st of December against the change in stock price for the following year. We gathered these data points from the year 2000 to 2014. Figure 5.1 plots this relationship.

| Date | Symbol | $\delta p_d$ | trend |
|---|---|---|---|
| 2000-01-05 | DD | 0.04230 | Gain |
| 2000-01-05 | DIS | 0.03748 | Gain |
| 2000-01-05 | GE | -0.00749 | Loss |
| 2000-01-05 | GS | -0.04248 | Loss |
| 2000-01-05 | HD | -0.00097 | Loss |
| 2000-01-05 | IBM | 0.03515 | Gain |

Figure 4.1: Data Extract of Primary Dataset

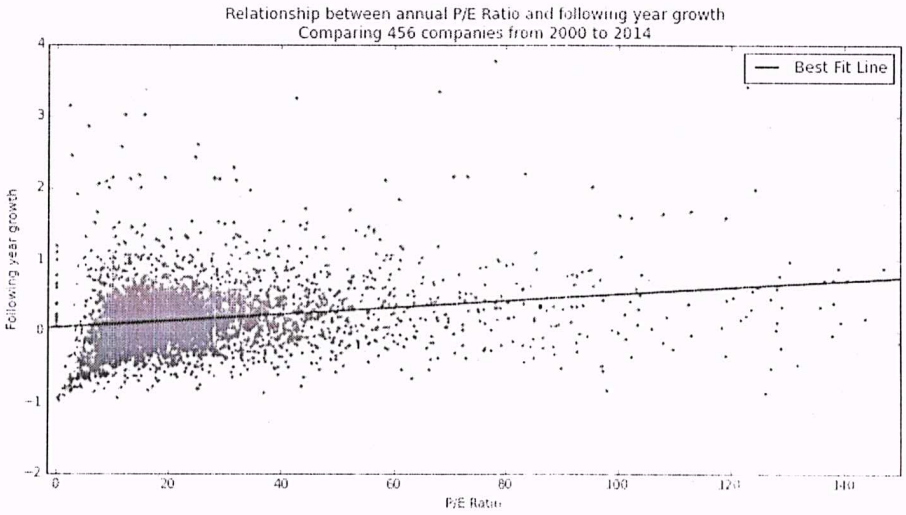



Figure 4.2: Relationship between P/E Ratio and following year growth

The best fit line was calculated using the standard Least Squares method. If the P/E ratio was indeedpredictive, we might have expected a better fitting line.

We can investigate the data further using a boxplot. Figure 5.2 divides companies into two categories. The first category, *titan Companies*, are companies whose share price increased in a given year, andthe second category, *Loss Companies*, are companies whose share price fell in a given year. The boxplot shows the distribution ofP/E Ratios foreach category.

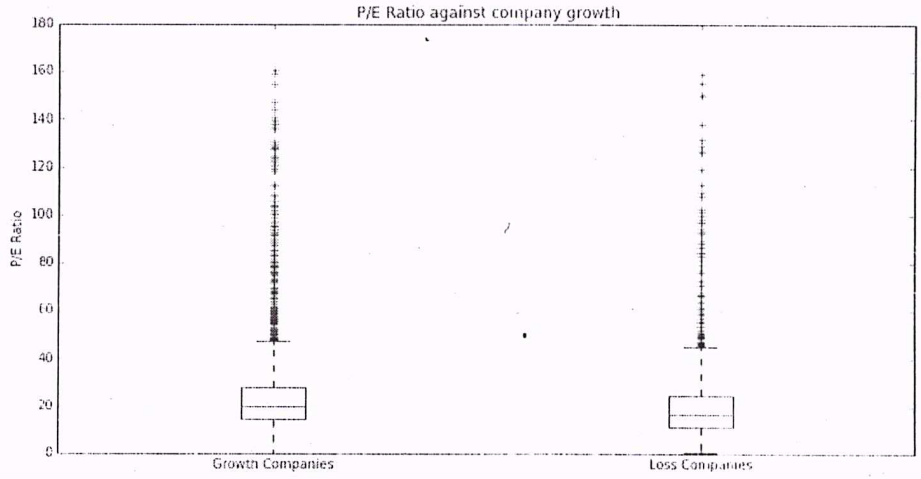



Figure 4.3: Investigation of P/E Ratio predictive value using Box plot

If the P/E Ratio was predictive, we would have expected a noticeable difference in the P/E Ratio distribution of companies whose share increased, and those whose share price decreased. However, this is not the case. The P/E Ratio distribution between these categories is almost identical. We canthere- fore conclude that the P/E Ratio has little predictive value when it comes to estimating company performance for the following year.

## 4.2 Price to Book Ratio

The second metric for the value of a company that we will look at is the Price to Book Ratio. The price to Book Ratio calculation is defined in equation 5.2.

$$\text{P/B Ratio} \quad \frac{\text{Share Price}}{\text{Book Value of Company}}$$

Informally, what the Price to Book Ratio calculates is the ratio between the value of a company according to the market and the value of the company on paper. ilf the ratio is high; this might be a signal that the market has overvalued a company and the price may fall over time. Conversely if the ratio islow, that may signal that the market has undervalued the company and the price may rise overtime. We should then expect to see a relationship between high P/B ratio and low returns in the following year.

To investigate this relationship, we plotted the P/B ratio for of 316 companies on the 31stof Decemberagainst the change in stock price for the following year. We gathered these data points from the year2000 to 2014. Figure 5.3 plots of this relationship.

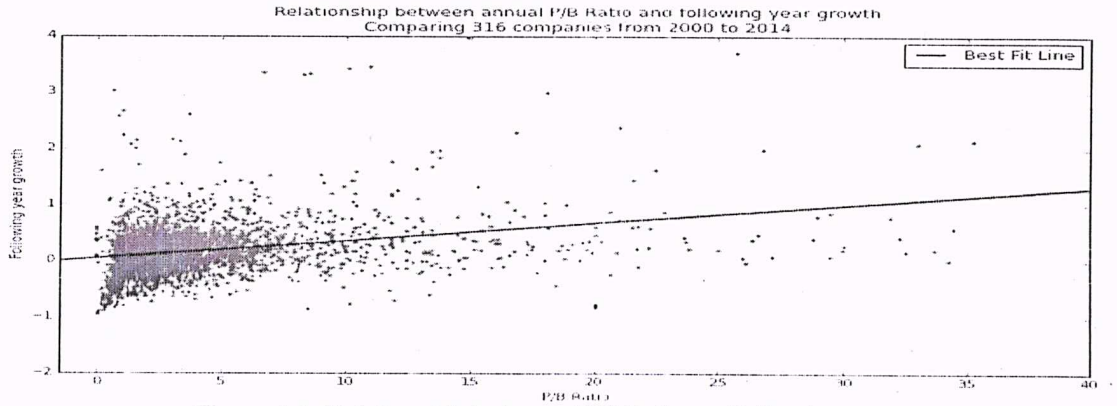



Figure 4.4: Relationship between P/B Ratio and following year growth

Just as in the P/B diagram, the best fit line was calculated using the standard Least Squares method. The slope of the best fit line here is greater than that of the P/E ratio which is the opposite of what we might have expected. The data suggests that a high P/B ratio is some what predictive of a high growth in the stock price. This is an unexpected result and is directly opposed to available literature on the subject [9]. This suggests that a high P/B ratio may act as a signal of investor confidence, a better signal than the P/E Ratio.

To better understand the predictive value of the P/B Ratio we can use a box plot. Figure 5.4 divides companies into two categories, exactly as in the earlier P/E Ratio example. The box plot shows the distribution of P/B Ratios for each category.

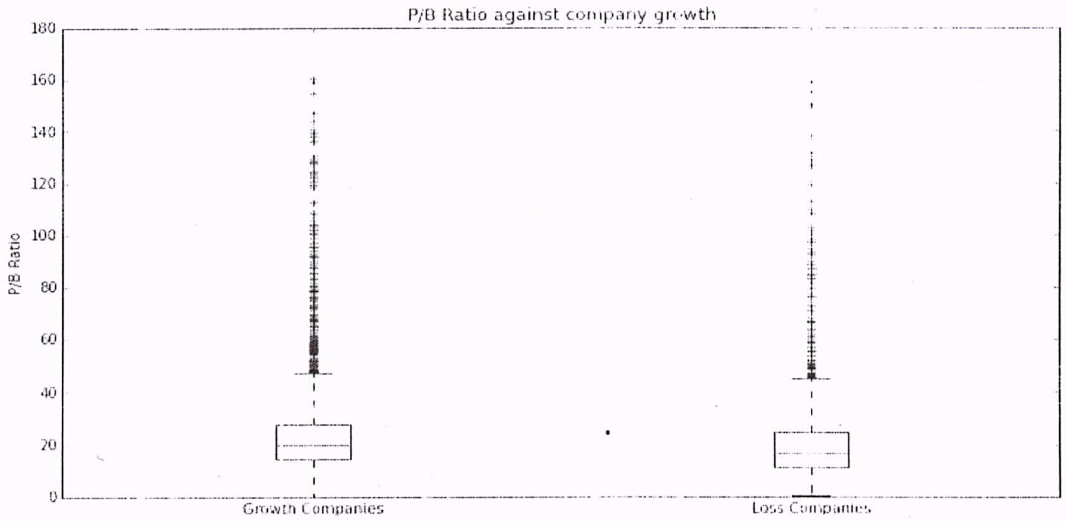



Figure 4.5: Investigation of P/B Ratio predictive value using Box plot

It is evident that figure 5.4 tells a very similar story to figure 5.2, the P/E Ratio box plot. We can see that companies that grew and companies that shrank had an almost identical distribution of P/B Ratios. If it were predictive, we would have expected different distributions for each category. We can therefore conclude that the P/B Ratio also has little predictive value when it comes to estimating company performance for the following year.

## 4.3 Limitations of Fundamental Analysis

There is an obvious problem in Fundamental Analysis. We are trying to quantify the true value of acompany when almost every company has in someway or another some purely qualitative value.

Fundamental Analysis methods do not attempt to capture these qualitative values. How should onequantify the value of a brand, the size of its customer base, or a competitive advantage? Until thesevalues are quantified, it leaves a large gap in what an algorithmic style approach can achieve. What algorithm, for instance, would have valued WhatsApp at $22 billion while still making a year-on-yearloss?

## 4.4 Fundamental Analysis– Conclusion

We evaluated two Fundamental Analysis metrics and found no conclusive proof of their predictivevalue These predictions are also very long term, looking one year into the future. Predictions on thistime scale are not the focus of the project. Instead we will focus on predicting daily trends in the market. Due to these issues that we moved away from Fundamental Analysis.

# CHAPTER 5

# TECHNICAL ANALYSIS

These second approach we take at market predictions Technical Analysis. As described in there view of Kara et al this makes use of recurring patterns and trends within the price of a stock and goes directlyagainst even the weak form of the Efficient Market Hypothesis.

Nevertheless, Technical Analysis remains popular in online non-academic literature. This is likely because it is simple, intuitive, and what many casual investors imagine a professional trader might use. The quintessential image of the work of a professional trader is analyzing a historic price graphs usingsophisticated diagrams and charts to predict the future price. This is Technical Analysis.

## 5.1 Moving Average Crossover

Next, we move to Technical Analysis models that are sound in theory. These models work on a statistical basis rather than patterns and make explicit pre- dictions about the future. One of the simplest and most common models of this type is the Moving Average Cross over strategy.

The moving average crossover strategy relies on the interaction between two moving averages ofdifferent periods. One is a moving average over a short period, and the other is a moving average over a longer period. For example, the short moving average might be the mean price over a period of the last 10 days, and the long moving average might be the mean price over a period of the last 20 days. When the short moving average crosses under the long, this can be interpreted as a negative signal that themarket is trending downwards. Conversely if the short moving average crosses over the long, this canbe interpreted as a positive signal that the market is trending upwards. Accordingly, the points wherethese events happen are called the crossover points and can be categorized into negative and positivecrossovers point.

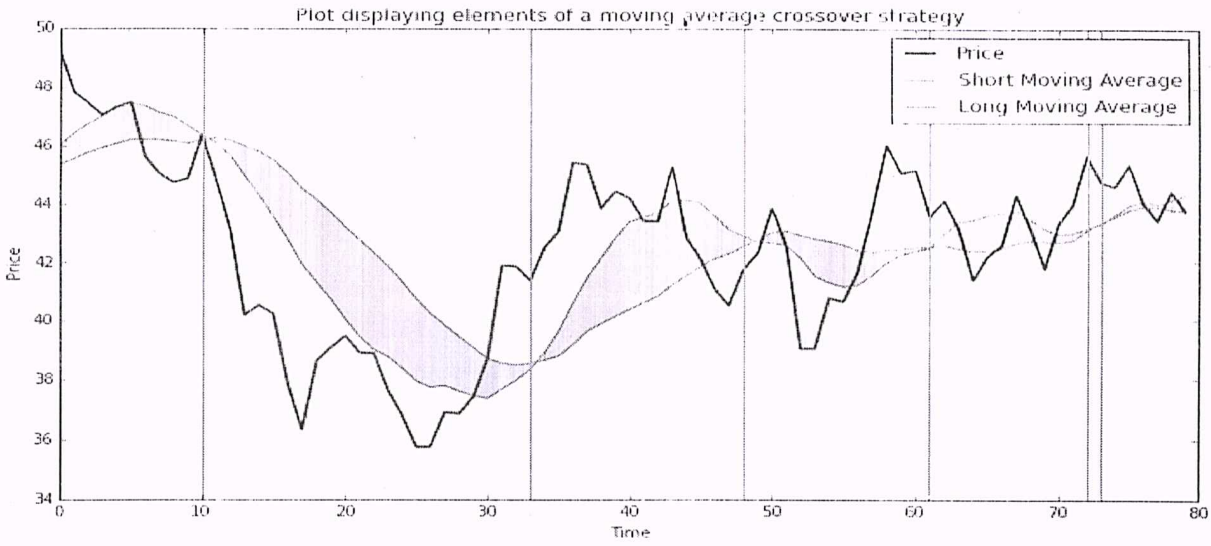



Figure 5.1 : Moving Average Crossover

In figure 6.2, the red areas are where the short moving average is below the long moving average andthe green areas are where the short moving average is above the long moving average. The diagram seems to give us hope for this strategy. The large green and red areas on the left of the diagram doindeed appear to be predictive of market upward and downward trends respectively.

However, while it is attractive to look at the crossover points on the left of the diagram, one should notignore the less significant crossover points on the right of the diagram. These are crossover points justas much as the ones on the left area but these are not as predictive for market trends. The goal is to tochoose long and short periods moving average to maximize the predictive value.

### 5.1.1 Evaluating the Moving Average Crossover Model

To evaluate the predictive value moving average crossover model, we attempted to build a predictorusing these crossover signals (positive and negative) as the input features and the market trend for thefollowing day as the dependent variable we are trying to predict. We performed a rigorous evaluationof long- and short-term pairs on a training set, and tested the winning long/short term pair against anindependent test set. To perform this analysis, the database previously assembled in the Data and Tools section was used. This

database was then joined with 50 new moving average columns. The first of these columns contained the 1-day moving average price, the second contained the 2-daymovingaverageprice, etc., up to the 50-day moving average price. This precomputation of divided based on company. We chose 20 random companies and used them as the training set. The remaining 10 companies were used as a test set. This is similar to the single-holdout method, whichunder normal circumstances is not considered to be statistically credible. However, there was sufficient data in this case for single-holdout to be viable. The training set contained 74,000 trading days between all 20 companies the test set contained 38,000 trading days between the 10 companies. The model itself was purposefully kept extremely simple so as to remain true to the intended usage ofthe moving average strategy. When a positive crossover occurred (short crosses over long), the modelpredicted the stock price would increase tomorrow, and when a negative crossover occurred itpredicted the stock price would fall tomorrow. Finding the optimal pair of long and short movingaverage periods is equivalent to finding the best hyperparameters for the model. This is modelselection.

We perform a grid search over all possible long and short period pairs. For each period pair, we find the crossover points between them. For each crossover point, we make a prediction based on its positivity or negativity, and compare tomorrow's predicted trend against the actual trend. We can then calculatethe accuracy of this long/short period pair and remember it if it is the best so far. When we have iterated over all possible long and short pairs, we will have found the best period pair for predicting tomorrow's trend in the training set. Table 6.3.1 displays the top 5 short and long period pairs and theirtest accuracy.

| Rank | Short Period | Long Period | Training Score |
|---|---|---|---|
| 1 | 8 | 35 | 0.5229 |
| 2 | 8 | 34 | 0.5219 |
| 3 | 8 | 33 | 0.5127 |
| 4 | 2 | 37 | 0.5073 |
| 5 | 2 | 36 | 0.5016 |

The winning pair after model selection was 8, and 35 for the short and long period respectively. Wethen cross validated this against our test set. Our accuracy on the test set using the 8, 35 pair was 0.5157. This is slightly lower than our training score, as should be expected.

However, both of these scores are still extremely poor. They perform only marginally better than acoin toss. We can put into context how well the model performs using the Kappa Statistic. The KappaStatistic compares the model's performance to a random version of itself. The Kappa Statistic is defined in equation 6.1 where $P(A)$ is the observed proportion of agreement and $P(E)$ is expectedproportion of agreement.

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

This model scores a kappa of 0.0427, which is far from statistical significance. The model performs no better than random in any significant way. We must conclude that the Moving Average Crossover is not predictive in any meaningful way, at least for predicting the movement of companies on the DowJones.

## 5.2 Additional Technical Analysis Models

Although we had no success using moving average crossovers, there are many other technical analysis indicators that are regularly used by traders. We will evaluate a further five of the most common indicators.These indicators are described below.

**Weighted Moving Average** A weighted moving average is the average where later data pointsare given more weight than earlier data points. In terms of this experiment, the data points will be theprices of individual days. The intuition in using the weighted moving average indicator follows an
idea of reverting to the mean, but in this case the mean is calculated using weighted prices. Equation

6.2 defines weighted moving average $wma^n_d$, on day $d$ using the previous $n$ days where $p_d$

Is the price on day $d$.

$$wma_n = \frac{n * p_d + (n-1) * p_{d-1} + \ldots + 1 * p_{d-n}}{n + (n-1) + \ldots + 1}$$

**Momentum** A momentum indicator calculates the relative change in price over the last $n$ days. This is based on a straight forward percentage change. Equation 6.3 defines $momentum^n$ which calculates momentum on day $d$ by looking at the price difference from $n$ days ago.

$$momentum^n = \frac{p_d - p_{d-n}}{d}$$

**Relative Strength Index** The Relative Strength Index (RSI) attempts to capture the trading strength of a stock, i.e will how likely a stock price is to continue rising or falling. Borrowing from Wong et al.
[20] for the intuition, Readings of 100 imply that there are pure upward price movements, while readings of 0 imply that there are pure downward price movements. Hence a reading close to 100 indicates an over bought market, while a reading around 0 indicates an over-sold market.

Much like our other indicators then, RSI should act as an indicator that the market price is about to revert back to a some kind of mean.

Calculation of RSI is done in two parts. First $RS^n$ is calculated. $RS^n$ is, informally, the average price increase over the last $n$ days divided by the average price decrease over the last $n$ days. $RS^n$ is then used to calculate $RSI^n$ which is the relative strength index on day $d$ using the preceding $n$ days.

## 5.3 Data Preparation

Before we begin training the models, we must prepare the data to feed them. Because no thresholds had to be defined, this step was relatively straight forward.
The primary price database originally prepared in the Data and Tools chapter was used

for this experiment. For each company, its price history was iterated over and the value of each indicator wascalculated for each day. The results of these calculations were stored in a temporary collection.

When iteration was completed, the data for each company was individually standardized so that bothhigh and low priced companies could be used in training and testing the same model. Standardization shifts and scales the data so that it has a mean of 0 and a standard deviation of 1. It should be noted that standardizing in this manner is not correct methodology. Ideally, the training data should be standardized before the test data, and the test data should be standardized using calculationsbased on the training data. This is done to avoid leakage, i.e. when the test data influences the training data. However, this would be difficult to do in this situation as each company needed to be standardized separately. Since this error is more likely to artificially increase the model's performance, and our models will fail to perform anyway, this error is ignored.

With each of the indicators for each company standardized, the data is stored for later use in erroreslimation. There are 95,407 examples in the database.

## 5.4 Error Estimation

In this step, we determine which class of model should be used on this data. We will compare threetypes of model, ak-Nearest Neighbors model, a Logistic Regression model, and a Naive Bayes basedmodel. We also need to search a range of hyperparameters for each model. For instance, what value ofk do we use in k-NN? To get an error estimation of each model type, we do a grid search over allhyperparameter values and test using nested kFold cross validation. In this case, we split the data into10 outer folds and 10 inner folds. The inner folds determine the winning hyperparameter which is crossvalidated by the outer fold. In total, for each hyperparameter there are 100 models trained. Table 6.1details the models tested, the hyperparameters searched, and their estimated accuracy.

| Model Name | Model Specific Hyperparameters | Estimated Accuracy |
|---|---|---|
| Logistic Regression | Norm penalization: l1 and l2 | 0.5143 |

| | | |
|---|---|---|
| K Neighbors Classifier | k: 5 to 15, weights: uniform and distance | 0.5034 |
| Gaussian NB | - | 0.5131 |

Table 5.2: Error Estimation Scores

It is evident from table 6.1 that none of these models were terribly successful. This is a strong sign that the input data, the indicator values, were not predictive of the following day's price trend.

This does not bode well for technical analysis; all technical analysis methods that have been tested have failed. However it would seem that our work is supportive of most serious academic literature on the subject. Prominent economist Burton Malkiel summarises the academic attitude towards technical analysis quite well[14].

Whether deserving of it or not, it would be right to analyze what went wrong. Does the fault indeed lie with lack of predictive power in the technical analysis indicators? Or, perhaps it is that our models were not complex enough to capture their signals. This question will be answered in the following chapter where we will question whether there is any useful information in the price history at all. Ultimately, we will provide strong evidence that there is in fact no useful information in price history. It follows logically then that any method that tries to leverage price history is doomed to fail. This includes technical analysis.

### 5.4.1 Common Problems with Technical Analysis

For a casual investor, navigating online literature in this area poses a significant challenge. An extremely common theme in this literature is the poor method- ology applied to evaluating trading patterns. We have seen two examples of confirmation bias when we looked at the Head and Shoulders pattern and when we looked at Moving Average crossover points. In the former, patterns that didn't fit the narrative were simply ignored and in the latter people focused too heavily on the instances where it did work.

Even when there is no confirmation bias present, there is very rarely any proper separation of training and test set. Correct methodology would separate these examples so that one could accurately estimate how the model would perform given unseen examples, like it would have to do in reality.

This Problem is prevalent when looking for long and short periods in moving average crossover. Many analysts find the best long/short period pair for their given time period and expect that to be just as predictive in future periods. This is incorrect methodology. It is easy to overfit a model to performwell on a single piece of data, but this may not carry over to unseen examples.

Above, we applied the correct methodology. In our experiments, the data was split into test and training sets and hyperparameters were determined through cross-validation.
This gives us a true estimate of how our best estimator carries over to future data. This proper methodology is not common in online literature.

### 5.4.2 Technical Analysis-Conclusion

It might have been expected that, given the popularity of Technical Analysis for stock market trading, there would have been a more positive result. However, somewhat surprisingly, the data shows thatthere is little predictive value to be found in Technical Analysis.

# CHAPTER 6

## ANALYZING THE PROBLEM

Our final approach to attacking the problem of stock market prediction is to look at machine learning. With machine learning, we will be building models that learn to exploit relationships within the data that we might have otherwise missed using Fundamental Analysis or Technical Analysis.

### Preceding 5-day prices

In Technical Analysis, we attempted to find patterns and trends in the data that we could use to predict the price movement, the trend, for the following day. Ultimately Technical Analysis failed to produce any notable results, but perhaps it was because the models are not complex enough to capture any hypothetical pattern that might exist in market data.
Similarly to Technical Analysis, in this section we will try to apply machine learning techniques to the price of the stock itself. Again, the Efficient Market Hypothesis says that it should not be possible to gain any predictive value from the price alone.

The data that we will be using is the percentage change in closing price of the stock of the preceding 5 days. The dependent variable will be the trend of the 6th day i.e. whether the stock price fell or rose. This dataset is built using the dataset assembled in the Data and Tools section. Table 7.1 is an extract of the first 4 rows in the data set where *trend* is the encoding of the 6th day according to equation 4.2.
Columns labelled day1 to day5 are the percentage change in closing price of the preceding 5 days. We will use the 6th day trend as the dependent variable. The dataset gathered contained 206635 examples. It was gathered from the the daily closing price of companies in the Dow Jones from the year 2000 to 2014. Before feeding the data into the models as discussed later, the data rows were randomly permuted to remove any bias the model could learn from an ordered dataset.

| | day1 | day2 | day3 | day4 | day5 | 6th day trend |
|---|---|---|---|---|---|---|
| 0 | 0.0492 | -0.0029 | 0.0176 | 0.0115 | 0.0028 | Loss |
| 1 | -0.0029 | 0.0176 | 0.0115 | 0.0028 | -0.0142 | Loss |
| 2 | 0.0176 | 0.0115 | 0.0028 | -0.0142 | -0.0028 | Gain |
| 3 | 0.0115 | 0.0028 | -0.0142 | -0.0028 | 0.0115 | Loss |

Table 6.1:Data Extract

## 6.1 Error Estimation

After we have gathered the data, the next step in building our model is choosing which base model we should work with. This step is called Error Estimation. It involves training and testing the performance of various models on the dataset to see which one we should focus on optimizing.

We tested each model by doing a nested kFold test. In this case, we split the data into 10 outer folds and 10 inner folds. The inner folds determine the winning hyperparameter which is cross validated by the outerfold. In total, for each hyperparameter there are 100 model strained.

| Model Name | Hyperparameters Tested | Nested KFold Accuracy |
|---|---|---|
| Logistic Regression | Norm penalization: l1 and l2 | 0.5343 |
| K Neighbors Classifier | k: 1 to 10, weights: uniform and distance | 0.5172 |
| GaussianNB | - | 0.5292 |

Table 6.2: Error Estimation Scores

Table 6.2 shows the accuracy scores of each model that we tried on the given data. A support vectormachine based model was also tried, but it proved too slow to train with the computational resources at hand.

However in smaller trials the SVM model did not seem to be significantly better than any of the modeltypes presented above.

It is immediately clear that there is a problem here None of the models tested scored significantly above 0.5% accurate in our classification test. A coin toss predicting the outcome of the dependentvariable would have performed similarly to these models Although some models appear to be slightly better, we cannot place too much value in the actual value cross validations core in error estimation.

## 6.2 Analysis of Model Failure

The failure of all models there were tested gives us a strong indication that something is deeply wrongsomewhere. In this section, we will show that there is good evidence that the problem is in fact that there is no information to be found from the previous 5 day prices. This is entirely supportive of theEfficient Market Hypothesis.

### 6.2.1 Model Complexity

It is important here to clarify what is meant by model complexity. The complexity of a model is thesize of the set of hypotheses that it can produce. A hypothesis is a guess that the model can make aboutthe relationship between input features and the dependent variable. A linear model would guess, or hypothesize, that the relationship is a linear combination of the input features. A nearest neighbors' model would guess that the relationship is going to copy previously seen examples. A model that canproduce a larger number of hypotheses is more complex than one which can produce a smaller set. Itis also important to note that these hypothesis sets do not necessarily overlap. One model may be morecomplex than the other, but that is not to say that the best hypothesis does not belong to the model oflesser complexity.

It is possible that the models in the previous section failed because the models tested were not sufficiently complex, or that they did not cover the optimal set of hypotheses. However, the models and hyperparameters that were tested covered a large hypothesis base. If there were a good hypothesisto find, it is reasonable that one of the models would have performed better than a coin toss.

We can also inspect the effect of model complexity visually using graphs. For convenience we chooseto model the kNN model here. Analysis of the kNN model is simpler and no

less consequential than the other models.

The complexity of kNN can be varied by changing k, where k is the number of neighbors the modelwill consider to make its prediction. A low value of k means the model will look at only a few of theclosest neighbors to attempt to infer the value of the input, and conversely a high value of kmeans that the model will consider a larger number of neighbors. Recall that our measure of complexity is thesize of the hypothesis set. It follows that the complexity of kNN is inversely proportional to k. Given *n* examples, a uniformly weighted 1-NN model is able to produce *n* different predictions, but a *n N N*model can only produce one prediction.

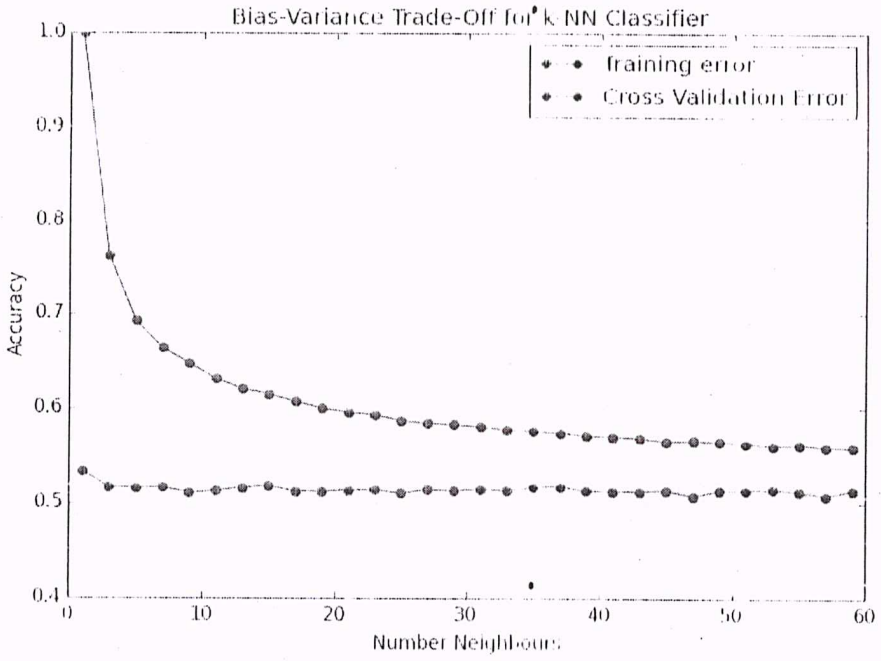



Figure 6.3: kNN Validation Curve

Figure 6.3 plots the validation curve for the kNN model. To generate this graph, we generated 60 kNNmodels with different k values. The training and test error for each value of k was then plotted. Since the complexity of themodel varies with k as explained earlier, we can interpret the graph as being a performance comparison of models of varying complexity.

On the left are models of high complexity and high variance (leading to high training score) and on the right are models of low complexity and high bias.

Normally in such a graph we would expect the training and test errors to become closer somewhere in the middle, indicating the optimum trade-off between bias and variance. However there is no such clear point in this graph. No matter what the complexity of the model, the Cross Validation Accuracy never rises significantly above 0.5%. We can then conclude that the problem is not due to the complexity of the model.

**Training Data**

If the problem does not lie with the complexity of our models, then maybe the problem is due to the amount of training data that we have. Perhaps if we had more data, we could build better models. As mentioned previously, the dataset we used for training and testing these models had over 200,000 examples. This seems like it should be enough, but maybe the stock markets are so complex they need more.

To properly diagnose this, we can plot the learning curve. In the validation curve, we varied the complexity of the model. In the learning curve, we will vary the number of examples presented to the model. Because of computational resources available, we were forced to constrain the number of examples in the learning curve to approximately 20,000. This is much smaller than the 200,000 examples in the dataset, but certainly not too small to ignore the findings. A 50-Nearest Neighbors model was used because there appears to be a slight upwards bump in the validation curve around 50NN.

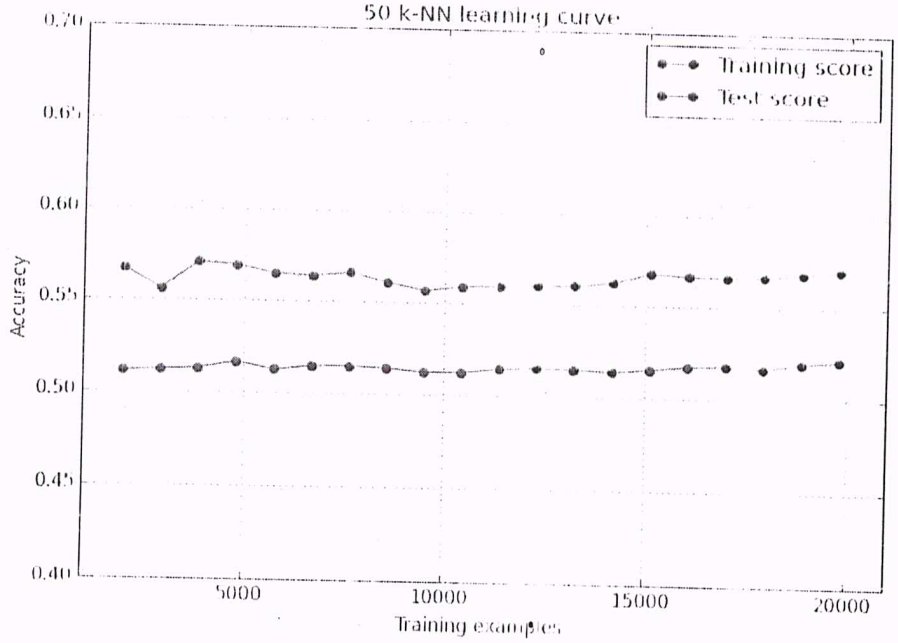



Figure 7.2: kNN Learning Curve

Figure 7.2 shows the effect of varying the number of examples on the model. Normally in a learning curve graph, you would expect the test and training errors to converge. The intuition is that as the model is presented with more data, it is better equipped to produce better hypotheses. Because these improved hypotheses are more general we expect the training accuracy to decrease (less overfitting), but the test accuracy to increase (less underfitting). However, we can see that this is not the case for our graph. The graph shows that for our model, the test and training accuracy remain apart and do not begin to converge no matter how much data they were given. We can then conclude that the problem is not due to lack of data for training and testing our model.

**Preceeding 5 day prices – Conclusion**

In this section, we attempted to train a model to predict the trend in the 6th day given the trends of the previous 5 days but all models were unsuccessful. We then analyzed why they might have failed and concluded that it was a problem with neither the complexity of the models nor the amount of data we had to train the models.

We must then conclude that the problem must lie in the data itself. It would seem that the preceding 5 day prices contain no information useful in predicting the following day trend. This supports our findings in the Technical Analysis approach and the Efficient Market Hypothesis. Clearly, we must begin to look at using external data.

## Related Assets

With the failure of using the price itself to predict stock movements, we turn to other sources of predictive value. Perhaps the most obvious source we should seek to use is the movement of assets related to the Dow Jones.

Intuitive, one might suppose that when the price of Oil rises, that is a good sign for the Dow Jones and we can expect it to rise too. Similarly, if the price of Oil falls we might expect the price of the Dow Jones to fall with it. In this section, we will search for features that rise and fall with the Dow Jones index.

## Data

Since in the last section we showed that yesterday's prices appear to have no influence on today's prices, it is unreasonable to expect yesterday's prices of related assets to influence today's price of the Dow Jones. Because of this, instead of predicting based on yesterday's prices, we will predict based on price movements in assets that are traded earlier in the same day. We will use the price movements of assets that are traded in Europe and Asia to predict the trend in the Dow Jones, which is traded in New York. If markets in Europe and Asia are trending heavily in one direction, it might follow that the Dow Jones will also trend that way when the markets open. What is important is that in the real world, we can observe the trends of our related assets before we need to make a trend prediction for the Dow Jones.

However, some market times overlap. For instance, the London Stock Ex- change and the New York Stock Exchange are both trading at the same time for approximately 4 hours daily. It would be ideal then to have price data which can be cleanly partitioned into intraday prices before and after the Dow Jones begins. However, this data is not easily available in the public domain. Intraday price data is a commodity and not something which is distributed as freely as interday price data. Quandl, which has been our source of much of the data up to this point, does not offer intraday price data.

We are forced to use subpar data which does not allow for proper preparation. We will continue to use the data provided by Quandl from which we can extract the daily closing price of each asset and index. It is conceivable that the trend in the New York Stock Exchange can affect closing prices on the London Stock Exchange, even though it only opens in the final 4 trading hours. This means that we have a dependent variable that potentially influences our independent variables. This is bad data preparation. Due to this issue we must be cautious and suspicious of positive results moving forward. It is unfortunate that later in the report, when we implement the trading algorithms in Quantopian, we will show that positive results mentioned in this section do indeed appear to be because of this error in the data.

## 6.3 Exploration of Feature Utility

The next step in creating a model to predict the daily trend of the Dow Jones index is to determine the predictive value of the features exploratively and qualitative manner, we first analyzed the predictive value of each feature on its own. Eight models were trained, one for every feature. For each feature a, kNN model was trained to predict the trend for the Dow Jones Index and the result was cross validated. Figure 7.3 shows the cross validation results for each feature.

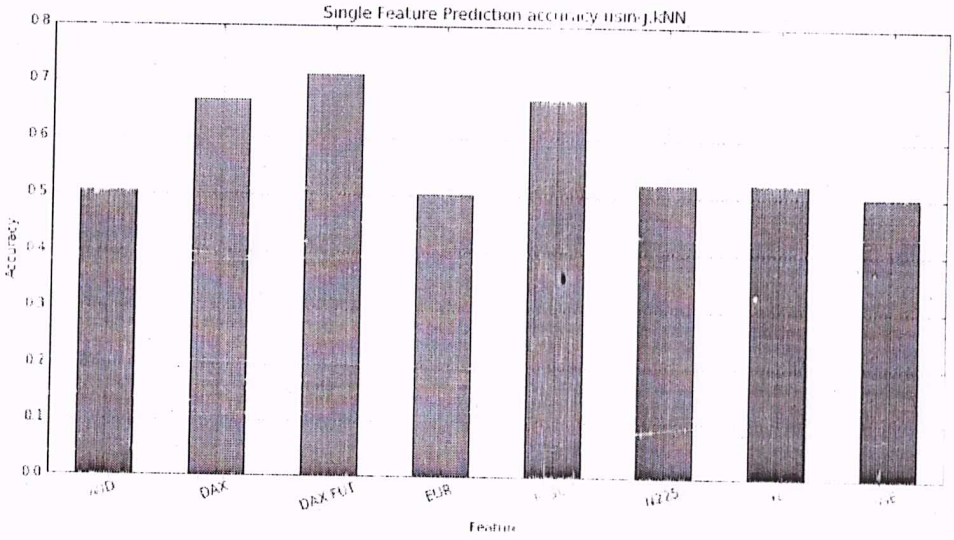



Figure 6.5: Single Feature Prediction Results

It is evident from figure 7.3 that there is indeed some predictive value in this dataset. There are three features that appear to be highly predictive. The German DAX, DAX futures, and the FTSE100 are all roughly 70% accurate without doing any work whatsoever. These results are quite similar to results presented by Shen et al. as discussed earlier.

## Modelling

Now that we know that there is at least some predictive value in our dataset we can begin to create our model.

### Error Estimation

Similar to what we attempt to do with the model based on the historical prices, the first step in creating our model is error estimation. In this step we estimate the performance of various classes of model.

We used a nested kFold method to perform the error estimation. The inner kFold performed a GridSearch over a set of hyperparameters. The outer kFold was responsible for the cross validation of the set of hyperparameters found within the inner kFold. The Grid Search searched over two domains. The first domain was the hyperparameter determining how many of the $k$ best features to keep. For instance, the search might determine that it was best to keep the top 3 most predictive features. The second domain were the hyperparameters specific to the model being tested. Table 7.3 details the model class tested, the model specific hyperparamters searched over, and the cross validation score.

| Model Name | Model Speciftc Hyperparameters | Estimated Accuracy |
|---|---|---|
| Logistic Regression | Normpenalization: l1, l2 | 0.5717 |
| KNeighborsClassifier | $0 < k \leq 25$, weights: uniform, distance | 0.7251 |

| | | |
|---|---|---|
| GaussianNB | - | 0.7208 |

The scores from table 7.3 look hopeful. We can see that the kNN model has the best estimated accuracy, albeit not by much. This means that this is the model we should bring forward to Model Selection to further optimize it.

**Model Selection**

Now that we have our winning model class, kNN, we now need to determine the optimum hyperparameters for our model.

A kNN model has two important hyperparameters, $k$ and how to weigh examples in the neighborhood. $k$ dictates how many neighbors the model should examine to make a prediction. We can then weigh those neighbors either uniformly or by their distance to the input. We can search over the same value space as we did in Error Estimation for kNN, $0 < k < 25$ and weights *uniform* or *distance*. We will also be searching for the best number of features to keep in the model. This can be treated as another hyperparameter.

After performing a grid search over the value space and cross validating each combination of hyperparameters, we find that the best combination of hyperparameters gives us a cross validated

## Model Selection

Now that we have our winning model class, kNN, we now need to determine the optimum hyperparameters for our model. A kNN model has two important hyperparameters, $k$ and how to weigh examples in the neighborhood.

$k$ dictates how many neighbors the model should examine to make a prediction. We can then weigh those neighbors either uniformly or by their distance to the input. We can search over the same value space as we did in Error Estimation for kNN, $0 < k$ 25 and weights *uniform* or *distance*. We will also be searching for the best number of features to keep in the

model. This can be treated as anotherhyperparameter.

After performing a grid search over the value space and cross validating each combination of hyperparameters, we find that the best combination of hyperparameters gives us a cross validatedaccuracy of 74.63% at predicting Dow Jones Index daily trends. Table 7.4 shows the best combinationof hyperparameters found.

| Hyperparameter | Best value |
|---|---|
| Number of features to keep | 3 |
| Number of neighbors toexamine | 14 |
| Method of weighing neighbors | uniform |

Table 7.4 indicates that the search determined that keeping only three of the original 8 features gave us the best cross validation score. This is in line with what we might have expected from the explorationof the predictive value of the features where we pointed out three features that appeared to be highlypredictive already.

By training the model on the full dataset with the winning hyperparameters and inspecting the 3 features the feature selector decided to keep, we can see that they are indeed the three features weassumed would be predicted. The German DAX, DAX futures, and the FTSE100 are the featuresselected in the final model.

## Data Exploration

As we did in the previous section, to get a better sense of our data we will explore it a little further. The company with the most opinions in the gathered dataset is Intel (INTC) which had 326 individualanalyst opinions. To get a sense of whether the data might be useful,

we can plot these opinions inrelation to the INTC price. First we filtered the dataset so that we only consider INTC opinions.

We then aggregated the opinions that were issued on the same day by different research firms by summing the number of Upgrades and Downgrades. Finally the aggregated opinions were merged with the INTC stock price. This data was plotted in figure 7.4. Green upwards pointing arrows signalan upgrade and red downwards pointing arrows signal a downgrade. The size of the triangle represents the number of coinciding opinions across all research firms on a particular day.

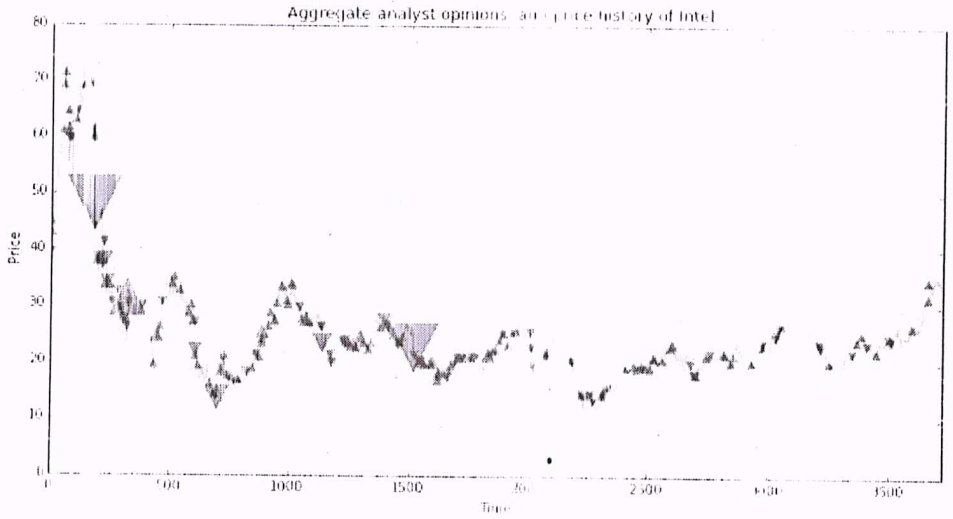



Figure 6.6: Visualisation of analyst opinions and INTC price

Figure 6.6 provides a visualization of the relationship between opinion sentiment (Upgrade or Downgrade) and price. The large red down pointing triangle ion the left hand side of the diagram is a point representing 9 separate research firms all downgrading their opinion of INTC on the same day. The continued sharp fall in price after this point is evidence that the opinions may be predictive.

From visual inspection it would appear there may be some value in these opinions, but we cannotbe certain until we attempt to build a model.

## Data Preparation

As evident from Table 7.5 all of the feature values in the dataset are labels and categories, not numerical data. This presents a problem. The machine learning library used in this project, sklearn, does not provide models which can handle non-numeric data. We must therefore transform our data into numeric form.

The correct way to transform a label type feature is to use One-hot Encoding. One-hot Encoding will transform a single feature with $n$ unique values into $n$ different features with binary values.

| | Location |
|---|---|
| 0 | London |
| 1 | Paris |
| 2 | New York |

(a)Before

| | London | Paris | New York |
|---|---|---|---|
| 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 |
| 2 | 0 | 0 | 1 |

(b) After

Figure 6.7: One-hot Encoding Example

Figure 6.7 demonstrates the concept of One-hot encoding. After One-hot encoding, the original feature is removed but no information is lost. To signify a value in the original column, a 1 is placed in the new corresponding column. For the opinions dataset, it was decided that the most important original columns were *Research Firm*, *Action*, and *To*. *Research Firm* might be important because an opinion from some firms may have more of an influence on the movement than others. The *Action* column is intuitively important because it summarizes the sentiment of the opinion into Upgrade or Downgrade. The *To* column is the recommendation of the Research Firm.

This is also intuitively important because it could be are commendation to Buy or Sell.

It was then necessary to One-hot encode each of these three features. This resulted in a total of 266 new feature columns, one new column for each unique value in the original three columns. Now that all the features have been one hot encoded, we can aggregate rows where more than one research firm issued an opinion for the same company on the

same date.

We will use a sum operation to aggregate the rows across all columns. For example, if both research firms upgraded a company on the same day then we will merge these rows and place a value of 2 in the upgrade column.

The entire data preparation stage can be done very concisely using the Pan- das python library. The code used for this step is shown in listing 7.3.3.

## Error Estimation

With our data prepared, we must decide which class of model to use.

We used a nested kFold method to perform the error estimation. The inner kFold performed a GridSearch over a set of hyperparameters. The outer kFold was responsible for the cross validation of the set of hyperparameters found within the inner kFold. The Grid Search searched over two domains. The first domain was the hyperparameter determining how many of the $k$ best features to keep. The second domain is the hyperparameters specific to the model being tested. Table 7.6 details the model class tested, the model specific hyperparameters searched over, and the cross validation score.

| Model Name | Model Specific Hyperparameters | Estimated Accuracy |
|---|---|---|
| LogisticRegression | Norm penalization: l1, l2 | 0.6627 |
| KNeighborsClassifier | $0 < k \leq 20$, weights: uniform, distance | 0.6504 |
| MultinomialNB | - | 0.6729 |

The scores from table 7.6 look positive. We can see that the MultinomialNB model has the best estimated accuracy. This means that this is the model we should bring forward to Model Selection.

### Model Selection

The MultinomialNB model is a naive Bayes classifier and only has one hyper- parameter. The hyperparameter, called alpha in sklearn, controls the additive smoothing in the model.

Additive Smoothing is a technique used to smooth categorical data.

At the same time as we are searching over the model hyperparameter, we will again be searching forthe best *k* features to keep. This is an important step in this case because we have a large number offeatures (266 in total) to begin with.

| Hyperparameter | Best value |
|---|---|
| Number of features to keep | 11 |
| MultinomialNBalpha | 0.7333 |

Table 7.7 shows the best set of hyperparameters found in the search. The winning model had a crossvalidation accuracy of 67.40%.

**Analyst Opinions-Conclusion**

In this section we gathered a list of analyst opinions and used them to build a model to predict the sameday price movement for companies in the NSE

Although concise in its final version, the data preparation for this section proved difficult to get right. The final model had an estimated accuracy of 67.40% which is an extremely positive result.

Although the data used in this section does not have the same problems the data in the previous section had, we will see similar results when it is simulated in a realistic trading environment. The model doesnot fair as well in the real world as our results may suggest.

**Disasters**

The final source of predictive data we will consider are natural disasters in the United States. It is easy to imagine why these might be predictive of changes in the stock price. A large storm could damage machinery or interfere with manufacture and logistics. In turn, this could have an impact a company's profit and therefore the stock price.

## Data Preparation

The first task was to gather a database of natural disasters in the United States. This data is provided byEM-DAT, The International Disaster Database [1]. Disasters in this dataset are guaranteed to be relatively large. The website states the qualifications necessary for a disaster to be entered into thedatabase

In order for a disaster to be entered into the database at least one of the following criteria has to befulfilled:

- 10 or more people reported killed
- 100 people reported affected
- A call for international assistance
- Declaration of a state of emergency

All disasters occurring in the United States from the year 2000 to 2014 were extracted. Before anyfurther preparation, there were 423 entries in this dataset. Disasters with missing or invalid dates were then removed. Following this, disasters with anomalous durations were removed. Figure 7.6 shows that there were three clear outliers with durations exceeding 100 days. These were removed from the dataset. After all preparation was completed on this dataset, there were 395 disaster events remaining.

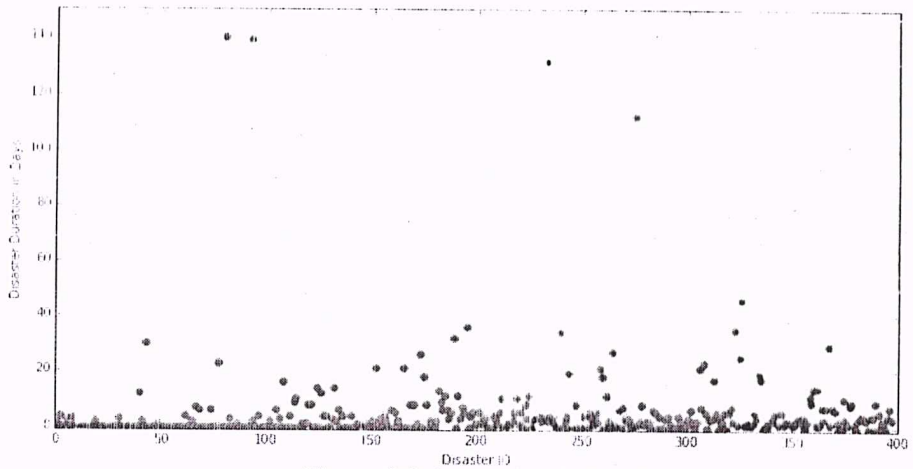



Figure 6.8: Anomolies in thedisaster data

Similarly to the Related Assets section, this data does not relate specificallyto any one company. Because of this, we will look atprice changes of the Dow Jones index, rather than any of its component companies.

# CHAPTER 7

## REPORT CONCLUSION

A large body of work was presented in this report. Two of the most widely used methods, Fundamental Analysis and Technical Analysis showed little promise in the experiments carried out. Technical Analysis specifically shows little to no potential of ever producing any statistically significant result when the correct methodology is applied.

Machine learning methods were then tested on a wide range of data sources. The result of some models looked hopeful, but ultimately failed when they were put through realistic trading simulations. This highlights that the stock market is prone to differences between theory and practice.

If there is anything that this report shows, it is that profitable stock market prediction is an extremely tough problem. Whether it is possible at all ultimately remains an open question. After completing the project, it is the firm belief of the author that the only viable trading strategy for a casual investor is a passive buy and hold strategy in index funds and ETFs.

# CHAPTER 8

## RESULT

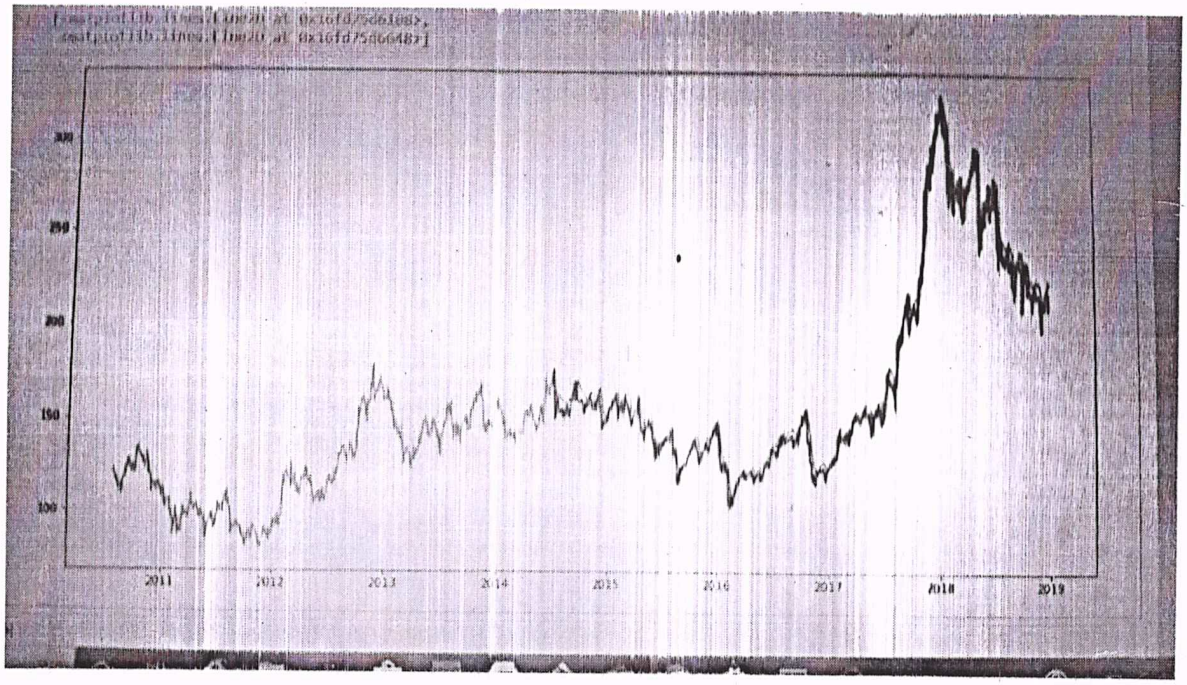



Fig. 8.1 The Result Occurred from the applied datasheet of a company.

# REFERENCES


[1] EM-DAT the international disaster database. http://www.emdat.be/database. Accessed 2015-02-26.

[2] Quantopian. https://www.quantopian.com/ Accessed: 2015-03-15.

[3] Investopedia.com. April 2015. http://www.investopedia.com/terms/s/slippage.asp.

[4] Quandl, January 2015. URLhttps://www.quandl.com/.

[5] http://www.kaggle.com

[6] http://www.uci.com For NSE dataset.

[7] Online Stock Trading Guide. Head and shoulders pattern, March 2015.

[8] S. A. R. Nai-Fu Chen and Richard Roll, Economic Forces and the Stock Market, The Journal ofBusiness, vol. 59, no. 3, pp. 383-403, (1986).

[9] E. F. Fama, Random Walks in Stock Market Prices, Financial Analysts Journal, vol. 51, no. 1,pp. 75-80, (1995).