2018-07-30

Tailored Ensemble Approach for Stock Price Prediction

Dhaliwal, Manmeet

UNIVERSITY OF CALGARY

Tailored Ensemble Approach for Stock Price Prediction

by

Manmeet Dhaliwal

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA

JULY, 2018

Abstract

The stock market is one of the most vital components of a free-market economy, as it provides companies with access to capital in exchange for providing investors partial ownership. By trading in stocks, investors have an opportunity to earn capital gain and investment income. We focus on improving capital gains by tackling the challenge of stock price prediction.

In this research, technical indicators were used to predict stock price. We proposed a Tailored Ensemble Approach (TEA) to improve accuracy. In this approach, various regression models (e.g. SVR, DT, etc.) were considered as base models. Also, multiple feature sets were developed to use to train each regression model. Our ensemble approach finds the combinations of regression models and feature sets, through a validation process, that best predicts future price for each stock. Once the set of combinations are constructed for each stock, a mean ensemble system is used to predict future price.

S&P 100 stocks dataset, spanning from the beginning of 2007 to the end of 2016, was used to evaluate this research. After evaluation, using various configurations, the proposed system consistently outperformed all base regression models used in this study. It was also demonstrated that a customized approach which tailors a set of regression models and feature sets based on stock and prediction period has significant impact on improving the accuracy of the stock price prediction system.

Acknowledgements

I would like to thank my supervisor, Dr. Reda Alhajj for putting his trust in me to be his student for Master program. His great insights in topics such as Data Analysis, Data Mining techniques, among many other, has been very valuable in providing me perspectives on tackling many challenges, including this thesis, through various strategies. Thanks to Dr. Jon George Rokne for his great insight in stock market trends, which helped me broaden my perspective on stock market analysis. We have had some very interesting talks regarding stock market among other topics, and his feedback on my research has been a very valuable asset.

I would like to thank all my friends in the database lab. Thanks to Animesh for being a great friend, helping and motivating me throughout my research. Thanks to Kashfia, we have had great days and some very interesting conversations. Thanks to Sahil and Serhan for all the wholesome fun we have had. Thanks to Alper for his constructive feedback on my research. Thanks to Tamer, for helping me with various things. I had a wonderful time TAing for you. Thanks to Abed and Salim for the good days we shared. I would also like to thank everyone in and around the lab for keeping up with my shenanigans. I wish the best for your endeavors.

Finally, I would like to thank my family, including my parents and my sister, for providing me with lot of great opportunities by moving to Canada and their unwavering support has been a great motivation.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
TEA	Tailored Ensemble Approach
NASDAQ	National Association of Securities Dealers Automated Quotation
NYSE	New York Stock Exchange
TSX	Toronto Stock Exchange
CSE	Canadian Securities Exchange
S&P	Standard & Poor
DJIA	Dow Jones Industrial Average
IPO	Initial Public Offering
FTSE	Financial Time Stock Exchange
ROI	Return on Investment
MSFT	Microsoft Corporation
TSLA	Tesla Inc
AAPL	Apple Inc
IBM	International Business Machines Corporation
EMA	Exponential Moving Average
SMA	Simple Moving Average
SVR	Support Vector Regression
DT	Decision Tree
SGD	Stochastic Gradient Descent
MLP	Multi-layer Perceptron
MAPE	Mean Absolute Percent Error
SVM	Support Vector Machine

Chapter 1 Introduction

1.1 Stock Market

The stock market is, first and foremost, financial institutions established to help businesses and investors come together to buy, sell and trade stocks for purpose of capitalizing enterprises in need of cash infusions. It is one of the most vital components of a free-market economy, as it provides companies with access to capital in exchange for giving investors a slice of ownership. It is a collection of markets and exchanges, such as National Association of Securities Dealers Automated Quotations (NASDAQ, 2018), New York Stock Exchange (NYSE, 2018), Canadian Securities Exchange (CSE, 2018), Toronto Stock Exchange (TSX, 2018), and many more. NYSE is the largest and most prestigious exchange. Companies listed on the NYSE must meet initial listing requirements and comply annually with maintenance requirements. For example, for U.S companies to remain listed on NYSE, they must keep their price above $4 per stock and their market capitalization above $40 million (NYSE, 2018).

The stock market indices are used to represent a nation's stock market, exchange's stock market, etc. Standard & Poor's 500 (Staff, 2018) is an American stock market index based on the market capitalization of 500 large companies having common stock listed on the NYSE or NASDAQ. S&P 500 is widely regarded as the most accurate representation of large-cap American companies (Staff, 2018). There are many other indices, such as Dow Jones Industrial Average (DJIA) (MarketWatch, 2018), Financial Time Stock Exchange 100 (FTSE 100) (Financial Times, 2018), etc.

A private company can raise money through an Initial Public Offering (IPO) by offering stocks, representing ownership certificates of the company. The worth of the company "going public" and the number of stocks being issued determines the opening price of the IPO. After IPO, traders and investors continue to trade a company's stock because the perceived value of company changes over time. Investors can make or lose money based on their perceived value of a stock in future. Stock price changes over time based on the demand and supply for that stock. For example, if a company is improving in their sales, there might be lot of demand for it, which

leads to increase in price for that company's stock. On the other hand, if a company reported a considerable amount of loss, this might lead to low demand for that company's stock but high supply as investors might be selling all or portion of stocks they own for that company. This would lead to stock price falling, as demand decreases and supply increases.

There are two main types of gains investors can get from investing, capital gain and investment income. Investment income comes from dividends in stock market. Dividend is a sum of money paid regularly by a company to its shareholders out of its profits. Capital gain comes from selling the stocks at a higher price than the purchase price. In this thesis, we focus on increasing the capital gains or decreasing the losses by improving the accuracy of stock price prediction and providing various investing opportunities.

1.2 Return vs Risk

Stock market provides one of the best investing opportunity to get a high return on investment (ROI). Usually, with higher return, higher risk is involved as well. Higher return would often mean investing in a stock with high volatility. If a stock is volatile, chances of the price dropping are high as well. Let's take Tesla (TSLA) (Tesla Inc., 2018) and Microsoft Corporation (MSFT) (Microsoft Corp., 2018) stocks for example.

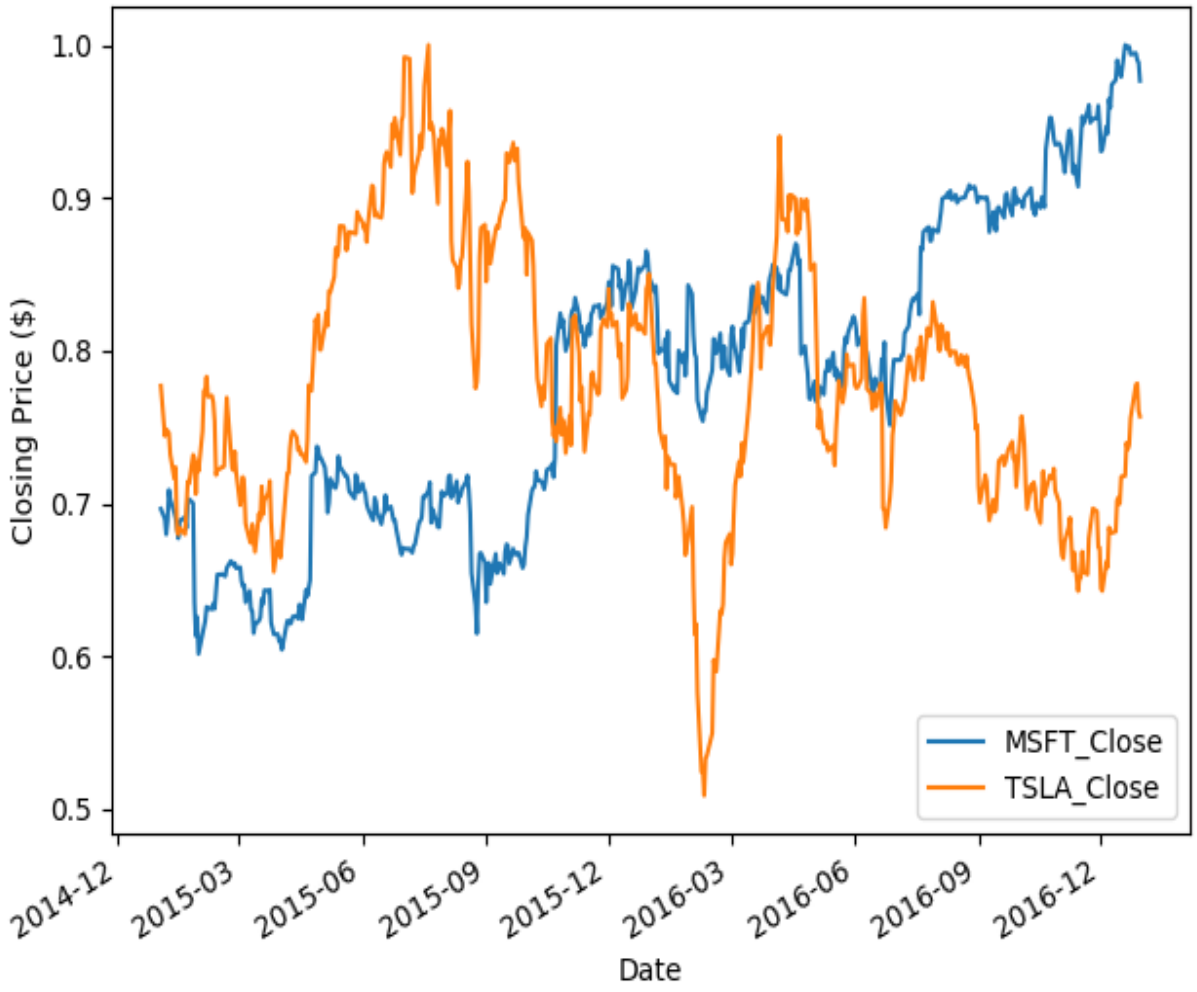

Figure 1-1: Normalized Historical Prices for MSFT and TSLA

TSLA is more volatile than MSFT, and based on the Figure 1-1, there are more opportunities to earn high return on TSLA than MSFT. MSFT is lot more stable than TSLA, this shows that opportunities to get higher ROI through MSFT might be lower than TSLA, but there is less risk involved as well. There are many different investing strategies used by investors, whether they are based on risk level, length of investment, or other factors (Connors & Alvarez, 2009) (Graham, 2006).

1.3 Investors

Warren Edward Buffett (Wikipedia, 2018) is one of the most popular investors in the world. He has a net worth of US $86.5 billion as of February 17, 2018. He is the chairman and chief executive

of Berkshire Hathaway Inc (Berkshire Hathaway Inc., 2018). Berkshire Hathaway is an American multinational conglomerate holding company headquartered in Omaha, Nebraska, United States. The company wholly owns companies like GEICO, Dairy Queen, etc., and owns variable portions of Pilot Flying, American Express, The Coca-Cola Company, among many other. Berkshire Hathaway has averaged an annual growth in book value of 19.0% to its shareholders since 1965 (compared to 9.7% from the S&P 500) (Wikipedia, 2018).

While, there are many who made a considerable fortune through stock market, there are also those who have lost their entire life savings through stock market, either because of market crash, like the one in 2008 (Farmer, 2012) or by making bad investment decisions through the normal trading sessions. We focus on improving investment decisions for normal trading sessions, to provide investors a better chance at getting higher ROI.

1.4 Importance of investing

It is becoming more important than ever to invest these days. With low interest rate on savings account and high inflation rate, the value of money depreciates over time. Let's assume you leave $100 in a bank account with 1% interest. After a year, you will receive 1 dollar from interest, increasing your deposit in bank account to $101. With 2.14% inflation rate for 2017 in US (Statista, 2018), an item that would have costed $100 at start of 2017 would now cost $102.14 at the end of year. Since the purchasing power of your cash went down and the interest was not high enough to keep up with inflation, the same amount of cash you had at the start of 2017 is worth less at the end of 2017. Even ignoring inflation, it is better to invest your money to build, and grow over time, optimizing the efficiency of your assets.

Investing in stocks is a risky opportunity. The amount of risk is associated with the profit goal. Aiming for higher profits might require taking higher risk opportunities and vice-versa. To reduce the risk, we want to improve the accuracy of stock price prediction. If we know the future stock price with certain level of confidence, it could give us an opportunity to increase the capital gains, or in a tough situation reduce our losses.

1.5 Literature Introduction

There has been much research done on stock price prediction (Hadavandi, Shavandi, & Ghanbari, 2010) (Bollen, Mao, & Zeng, 2011) (Tsai, Lin, David C., & Chen, 2011). There are three main types of analysis performed for predicting stock price/movement:

1. fundamental analysis (Dechow, Hutton, Meulbroek, & Sloan, 2001), relies on the statistics of the macroeconomics data such as interest rates, money supply, inflation rates, and foreign exchanges rates as well as the basic financial status of the company
2. technical analysis (Kara, Acar Boyacioglu, & Baykan, 2011), which is based on studying historical stock prices
3. data mining (Deng, Mitsubuchi, Shioda, Shimada, & Sakurai, 2011), which makes use of large amount of data available from numerous possible sources such as twitter, news, etc.

Fundamental analysis focuses on predicting price of a company based how the company is performing, what the interest rates are and so on. Let's consider company performance as one of the factors for predicting stock price. If company is performing well, we expect it to continue doing so, resulting in higher stock price in future. On the other hand, if a company continues reporting worse performance than expected, the stock price might fall as well.

Technical analysis is based on looking at the historical data to predict future prices. The idea here is that history will repeat itself. For example, if high volume movement resulted in rise of stock price in past, we expect stock price to rise during similar stock movement.

Data mining and artificial intelligence methods make use of many sources of data to predict the outcome. Many research papers consider sentiment value of public, through social media sites like twitter, to predict stock price movement. The hypothesis is that, if public has positive sentiment, stock price will rise and vice-versa.

There are two main prediction categories when it comes to stock market:

1. regression, which is predicting stock price
2. classification, which is predicting stock price movement

Regression analysis is focused on predicting the actual stock price. To convert this problem to classification, rather than predicting the price, stock price movement is predicted. There are usually two labels for the classification problem, positive and negative movement. Some research papers create further classes such as label for increase in stock price by a certain percentage.

1.6 Contribution

In this thesis, we propose a Tailored Ensemble Approach (TEA) for stock price prediction, which makes use of historical data of stocks. From the historical data, we extracted variety of features, such as exponential moving average (EMA), simple moving average (SMA), volatility, and so on. The base models are machine learning regression algorithms such as Support Vector Regression (SVR), Bayesian Ridge Regression, Multi-layer perceptron (MLP), etc. Different variations of these algorithms were trained as the base models. These variations were dependent upon the feature set used for training each model. For training, evaluation, and testing, we used 10 years' worth of historical data from S&P 500 spanning from start of 2007 to end of 2016.

The contribution of this thesis can be summarized as follows:

1. various regression algorithms are considered, and many different variations of each regression models are created based on feature sets
2. evaluation procedure is performed on the regression models to find out which regression models are good at predicting each stock
3. threshold value is optimized to find out how many of the best models should be considered to predict the stock price for each stock
4. ensemble system picks the best variations of models for predicting stock price of each stock

The novelty of this approach comes from the design of the system. Tailoring the mean ensemble model to use best configurations of base model and features set based on prediction period and stock gives our system better opportunity to improve the accuracy of stock price prediction and provides wider range of investing opportunities by working with hundreds of stocks. This allows the system to be scaled up easily by including more models, feature sets, stocks, and prediction periods. By including more stocks and prediction periods, we are also increasing the range of

opportunities for the investors. Including more models and feature set will potentially increase the performance of the proposed approach as well.

1.7 Thesis Structure

Rest of the thesis is structured as follows:

1. Chapter Chapter 2 provides background and related work on stock market
2. Chapter Chapter 3 details the data preprocessing stage
3. Chapter Chapter 4 describes our methodology
4. Chapter Chapter 5 describes the evaluation process
5. Chapter Chapter 6 concludes our research and discusses future work

Chapter 2 Related Work

There are three main types of analysis performed when it comes to stock price prediction, fundamental, technical, and based on data mining.

Fundamental analysis is based on looking at the financial and economic indicators to predict company's performance over a long-term period (Tsai, Lin, David C., & Chen, 2011) (Graham, 2006). Warren Buffett is also a big proponent of long term value investing (Martin, 2018). Fundamental indicators include financial indicators and economic indicators. Few examples of financial and economic indicators are; revenues, future growth, return on equity, profit margins, inflation, money supply, unemployment rate, etc. (StockCharts, 2018).

Technical analysis is based on looking at historical data to find patterns for making future predictions (Patel, Shah, Thakkar, & Kotecha, 2015). Technical analysis tends to be used for short-term predictions, as patterns change quite often and technical indicators do not have strong predictive power for long term predictions. Technical analysis includes indicators such as closing price, volume of stock traded, volatility of stock, etc. (Trading Technologies International, Inc., 2018).

There are also approaches based on data mining, where data from various sources, such as Twitter, news, etc. is gathered to predict stock prices and/or momentum (Asur & Huberman, 2010). These approaches have the advantage of considering the factors that technical and fundamental analysis would not have, such as; sentiment of public, popularity of a stock, public's mood towards a new product, and so on.

2.1 Fundamental Analysis

(Tsai, Lin, David C., & Chen, 2011) proposed an ensemble approach for stock movement prediction. The hybrid methods of majority voting and bagging are considered. In majority voting (Kittler, Hatef, Duin, & Matas, 1992), the class which receives the largest number of votes is selected as the final classification decision. In bagging (Breiman, 1996), several classifiers are trained independently by different training sets via the bootstrap method. Then the results of classifiers are aggregated via an appropriate combination method, such as majority voting.

Furthermore, two type of ensembles, that are homogeneous (e.g. an ensemble of neural network) and heterogeneous (e.g. an ensemble of neural network, decision tree, and logistic regression), are considered as well.

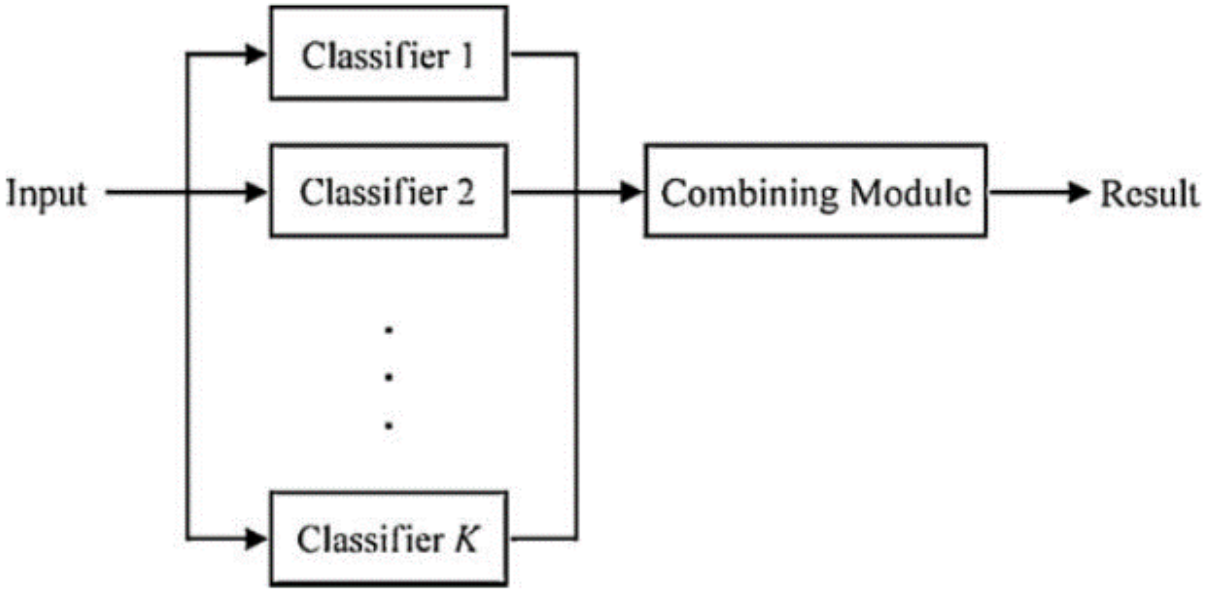

Figure 2-1: An example of classifier ensemble

Figure 2-1 (Kim, Kim, & Lee, 2003) shows a basic version of an ensemble system. Input is received by each individual classifier. The output from these classifiers is then combined using any of the combining modules (e.g. majority voting) and to produce the final result.

Taiwan Economic Journal dataset was used for the experiment. Quarterly data ranging from the second quarter of 2002 to the third quarter of 2006 was collected. A list of financial ratios and economic indicators shown in Table 2-1 (Tsai, Lin, David C., & Chen, 2011) were used for the study.

Table 2-1: The Finance ratios and economic indicators

Financial ratios	
Capital structure	Debt ratio Long-term capital
Amortization capability	Current ratio Quick ratio Interest cover
Business operation capability	Total asset turnover ratio Fixed asset turnover ratio Inventory turnover ratio
Profitability	Accounts receivable turnover ratio Return on assets Margin before interest and tax Net assets per stock Return on stockholder's equity
Cash flows	Cash flow ratio
Others	Constant net assets growth ratio Net assets growth ratio after tax Frequent interest growth ratio after tax Return on total assets growth ratio Return ratio of the last quarter
Economic indicators	Deposit interest rate Currency transferring rate (US dollars to Taiwan dollars) Discount rate Money supply Consumer price index Wholesale price index Unemployment rate Bond trading amount Total assets of listed companies Taiwan stock index Industrial Production Index

Accuracy for classifiers was calculated using the formula below.

$$Prediction\ accuracy = \frac{a+d}{a+b+c+d}$$

Where values of a, b, c, d are based on the confusion matrix in Table 2-2 (Tsai, Lin, David C., & Chen, 2011).

Table 2-2: Confusion matrix

Actual	Predicted	
	Negative	Positive
Negative	*a* (correct)	*b* (incorrect)
Positive	*c* (incorrect)	*d* (correct)

Where a is the number of correct prediction that an instance is negative; b is the number of incorrect predictions than an instance is positive; c is the number of incorrect predictions than an instance is negative; d is the correct predictions than an instance is positive.

Table 2-3 (Tsai, Lin, David C., & Chen, 2011) are the results after testing the various ensemble models vs. the base models. A homogeneous voting ensemble (using variations of MLP) gave the best ROI, while a heterogeneous ensemble with 7 classifiers gave the best accuracy result.

Table 2-3: Prediction accuracy and return on investment of the prediction models

	Prediction accuracy (%)	Investment performance	Rank of prediction accuracy	Rank of return on investment
MLP	63.33	4457.617	7	5
CART	59.44	3570.672	11	10
LR	60.28	3532.55	10	11
Voting (homo.)	66.25	4836.862	4	1
Voting (hetero.)	66.39	4637.826	2	2
Bagging (homo. 3×)	64.44	4448.412	6	6
Bagging (homo. 5×)	62.50	4132.984	9	8
Bagging (homo. 7×)	62.64	4005.882	8	9
Bagging (hetero. 3×)	65.00	4262.104	5	7
Bagging (hetero. 5×)	66.67	4545.209	1	3
Bagging (hetero. 7×)	66.39	4525.891	2	4

2.2 Technical Analysis

(Hadavandi, Shavandi, & Ghanbari, 2010) focused on price prediction through integration of genetic fuzzy systems and artificial neural network. They use stepwise regression analysis (SRA) (Esfahanipour & Aghamiri, 2010) to determine factors which have most influence on stock prices. Raw data is divided into k clusters by means of self-organizing map (SOM) (Mangiameli, Chen, & West, 1996) neural network. Finally, all clusters are fed into independent genetic fuzzy systems (GFS) models with ability of rule-based extraction and database tuning.

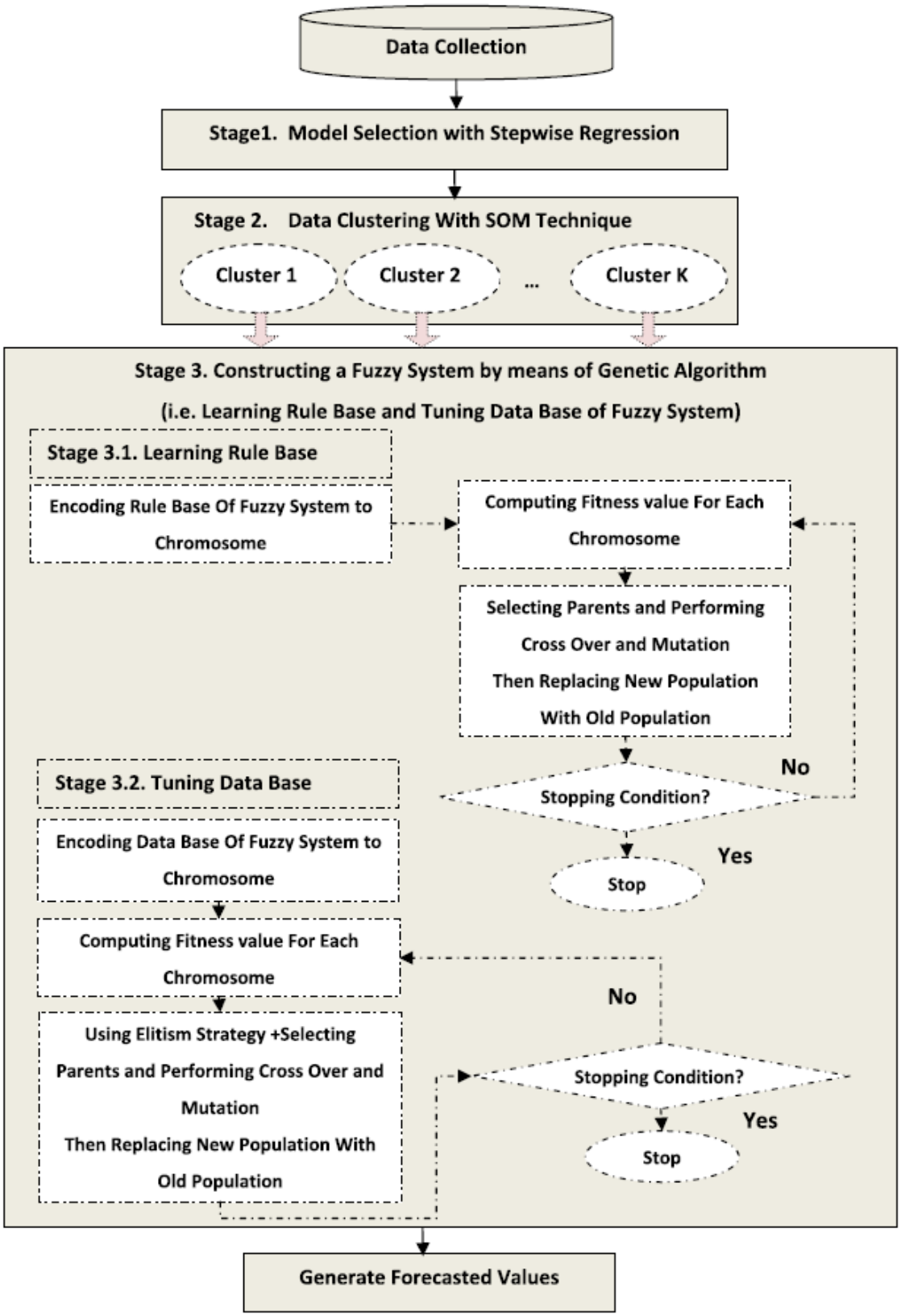

Figure 2-2: The procedure of making CGFS model

Figure 2-2 (Hadavandi, Shavandi, & Ghanbari, 2010) shows three main stages to construct CGFS model.

1. The first stage is variable selection stage; they use stepwise regression analysis to choose the key variables that are to be considered in the model. Stepwise regression method determines the set of independent factors that most closely determine the dependent variable.
2. Stage 2 consists of data clustering by SOM. SOM is a nonhierarchical clustering which outperforms hierarchical algorithms in clustering messy data and has better accuracy and robustness (Mangiameli, Chen, & West, 1996)
3. Stage 3 consists of constructing a Fuzzy System by means of Genetic Algorithm (GA). Mamdani-type fuzzy rule-based system is used in this research (Cordón, 2011). Evolutionary process for evolving knowledge base of fuzzy rule-based system consists of two general stages; stage 3.1 evolves rule base of fuzzy system and stage 3.2 tunes database of fuzzy system.

For evaluation, daily historical data for IBM, DELL, British airlines, and Ryanair airlines was collected. The variables used for the proposed approach were Open price, Close price, High Price, and Low price. Table 2-4 (Hadavandi, Shavandi, & Ghanbari, 2010) shows how data was split for training and testing.

Table 2-4: Training and test datasets

Training and test datasets.

Stock type	Stock name	Training data			Test data		
		From	To	#Observations	From	To	#Observations
IT sector	IBM corporation	10 Feb 2003	10 Sep 2004	400	13 Sep 2004	21 Jan 2005	91
	DELL Inc.	10 Feb 2003	10 Sep 2004	400	13 Sep 2004	21 Jan 2005	91
Airline sector	British airlines	17 Sep 2002	10 Sep 2004	503	11 Sep 2004	20 Jan 2005	91
	Ryanair airlines	6 May 2003	6 Dec 2004	400	7 Dec 2004	17 Mar 2005	71

To evaluate their proposed approach and compare to other approaches, they used a common evaluation statistic called Mean Absolute Percent Error (MAPE).

$$MAPE = 100 * \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i - P_i|}{Y_i} \qquad 2.1$$

Where Y_i is the actual value and P_i is the forecasted value of the ith test data obtained from CGFS and N is the number of test data.

Table 2-5: CGFS evaluation vs. previous methods based on MAPE value

Method	Stock name			
	IT sector		Airline sector	
	IBM corporation	DELL Inc.	British airlines	Ryanair airlines
HMM [19]	1.219	1.012	2.629	1.928
Hybrid of HMM, ANN and GA [20]	0.849	0.699	1.646	1.377
Hybrid of HMM and fuzzy logic [21]	0.779	0.405	1.529	1.356
ARIMA [21]	0.972	0.660	1.573	1.504
ANN [21]	0.972	0.660	2.283	1.504
CGFS (the proposed method)	0.63	0.534	1.413	1.241

Table 2-5 (Hadavandi, Shavandi, & Ghanbari, 2010) shows that the proposed approach (CGFS) has lower MAPE than the other approaches and as such outperforms them all. Therefore, CGFS can be used as a suitable forecasting tool to deal with stock price forecasting.

(Kara, Acar Boyacioglu, & Baykan, 2011) focused on predicting the momentum of stock price rather than the price itself. They developed two efficient models; three-layered feedforward artificial neural network (ANN) and support vector machines (SVM) for classification. Two comprehensive parameter setting experiments for both models were performed to improve their prediction accuracy. Each prediction was classified into one of two labels; 0 or 1. 0 represents prediction that stock prices will be lower than the current price, 1 represents stock prediction price rising than the current price.

The data used in this paper is the direction of daily movement in the Istanbul Stock Exchange (ISE) National 100 Index. The data set covers the period from January 02, 1997 to December 31, 2007. Technical indicators and their formulas are shown in the Table 2-6 (Kara, Acar Boyacioglu, & Baykan, 2011) below.

Table 2-6: Selected technical indicators and their formulas

Name of indicators	Formulas
Simple 10-day moving average	$\frac{C_t+C_{t-1}+\cdots+C_{t-10}}{10}$
Weighted 10-day moving average	$\frac{((n)\times C_t+(n-1)\times C_{t-1}+\cdots+C_{10})}{(n+(n-1)+\cdots+1)}$
Momentum	$C_t - C_{t-n}$
Stochastic K%	$\frac{C_t-LL_{t-n}}{HH_{t-n}-LL_{t-n}} \times 100$
Stochastic D%	$\frac{\sum_{i=0}^{n-1} K_{t-i}\%}{n}$
RSI (Relative Strength Index)	$100 - \frac{100}{1+\left(\sum_{i=0}^{n-1} Up_{t-i}/n\right)/\left(\sum_{i=0}^{n-1} Dw_{t-i}/n\right)}$
MACD (moving average convergence divergence)	$MACD(n)_{t-1} + 2/n + 1 \times (DIFF_t - MACD(n)_{t-1})$
Larry William's R%	$\frac{H_n-C_t}{H_n-L_n} \times 100$
A/D (Accumulation/Distribution) Oscillator	$\frac{H_t-C_{t-1}}{H_t-L_t}$
CCI (Commodity Channel Index)	$\frac{M_t-SM_t}{0.015D_t}$

Results show that their model performed better than the same models implemented in different papers because of the set of technical indicators models are trained on, analysis of models' parameter and selection of efficient parameter values. Neural network model of (Diller, 2003) had an accuracy of 60.81%, while (Altay & Satman, 2005) had an accuracy of 57.80%. While the proposed system in paper had an accuracy rate of 75.74%, much higher than the listed approaches. The neural network model also had a better accuracy than the SVM tested, which had an accuracy of 71.52%.

2.3 Data Mining

(Bollen, Mao, & Zeng, 2011) proposed an approach to include twitter mood in the stock price prediction model. Two tools were used for mood tracking. OpinionFinder, which measures

positive vs. negative mood and Google-Profile of Mood States (GPOMS) which measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). Granger causality analysis is used to test whether public's mood states has predictive information about DJIA closing price. SOFNN model was used to compare the performance based on two sets of inputs: (1) the past 3 days of DJIA values, and (2) the same combined with various permutations of mood time series. Figure 2-3 (Bollen, Mao, & Zeng, 2011) below shows the methodology used in the paper.

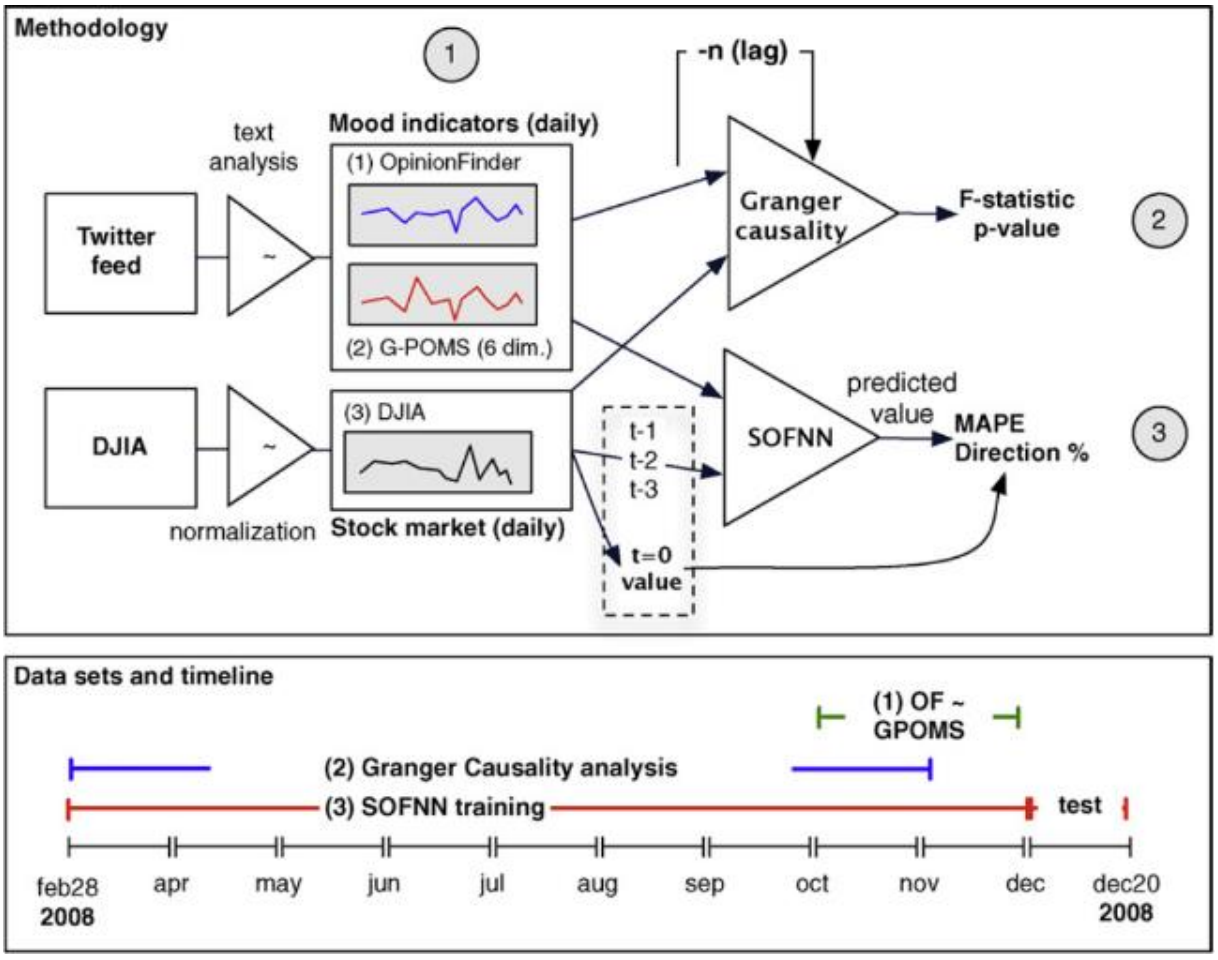

Figure 2-3: 3 phases of methodology and corresponding data sets

First phase deals with creation and validation of OpinionFinder and GPOMS public mood time series from October 2008 to December 2008. Second phase deals with use of Granger causality analysis to determine correlation between DJIA, OpinionFinder and GPOMS public mood from August 2008 to December 2008. Third phase is training of a SOFNN to predict DJIA

values based on various combinations of past DJIA values and OpinionFinder and GPOMS public mood data from March 2008 to December 2008.

For evaluation, twitter data is used for finding public mood with the closing price of DJIA, which is downloaded from Yahoo! Finance. Table 2-7 (Bollen, Mao, & Zeng, 2011) below shows the results of the methodology proposed.

Table 2-7: DJIA daily prediction using SOFNN

Evaluation	I_{0F}	I_0	I_1	$I_{1,2}$	$I_{1,3}$	$I_{1,4}$	$I_{1,5}$	$I_{1,6}$
MAPE (%)	1.95	1.94	1.83	2.03	2.13	2.05	1.85	**1.79***
Direction (%)	73.3	73.3	**86.7**	60.0	46.7	60.0	73.3	80.0

Results show that adding positive vs. negative mood has no impact on the prediction accuracy. While, adding Calm mood, results in highest prediction accuracy. This confirms with the Granger causality analysis, as Calm mood had the highest Granger causality relation with DJIA. The results show that the by including certain mood (e.g. calm) in the feature set, performance of the SOFNN model can be significantly increased.

proposed an approach that uses technical indicators, numerical dynamics (frequency) and overall sentiment analysis of news and comments to predict stock price. Multiple Kernel Learning Regression framework was used for this regression problem.

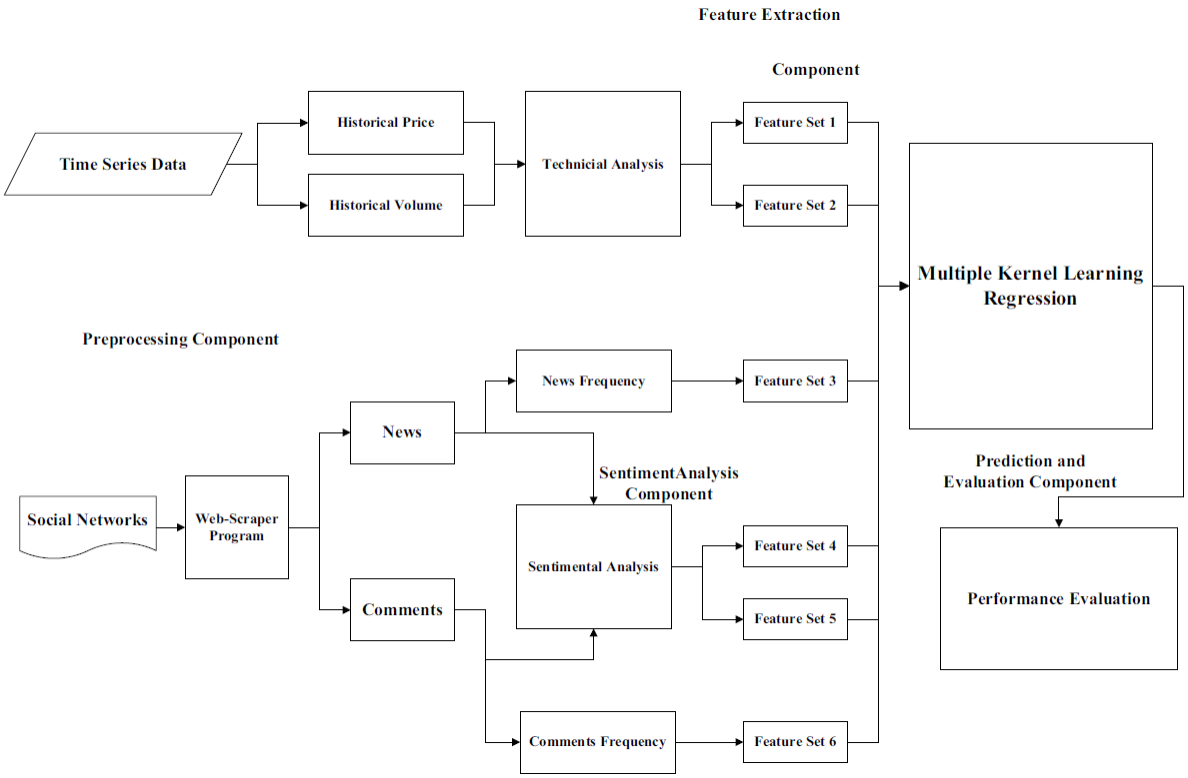

Figure 2-4: Proposed Model

Figure 2-4 (Deng, Mitsubuchi, Shioda, Shimada, & Sakurai, 2011) shows the proposed model in research. After performing technical and sentiment analysis, calculating comments and news frequency, features were fed to MKL Regression model for price prediction. For sentiment analysis, sentiments were divided into three different sentiments, positive, neutral, and negative. As for technical indicators, Price ROC, MACD value, BIAS of stock price was used. Stocks of three companies, Sony, Panasonic, and Sharp were used to evaluate this approach. Table 2-8 shows the list of methods and Table 2-9 shows feature set of those methods used for evaluation.

Table 2-8: A list of methods for experiments

Model	Description
MKL-A	Technical Analysis + Sentiment Analysis + Numerical Dynamics, MKL regression model
SVR-A	Technical Analysis + Sentiment Analysis + Numerical Dynamics, SVR regression model
SVR-ta	Technical Analysis Only, SVR regression model
SVR-sa	Sentiment Analysis Only, SVR regression model
SVR-nd	Numerical Dynamics Only, SVR regression model

Table 2-9: A list of input feature sets (where T is the day)

No.	Feature Set Description
1	Price ROC, MACD value, BIAS of Stock Price, from *(T-6)* to *T*
2	Volume ROC, from *(T-6)* to *T*
3	Frequency of News, from *(T-6)* to *T*
4	Frequency of Comments, from *(T-6)* to *T*
5	Number of positive/negative/neutral news, from *(T-6)* to *T*
6	Number of positive/negative/neutral comments, from *(T-6)* to *T*

MAE, MAPE, and RMSE evaluation measures were calculated to evaluate the models. The proposed approach significantly outperforms other approaches that used one type of feature set rather than including all features sets. It was concluded that using features from multiple various sources is better than only one of them.

Chapter 3 Data Preprocessing

3.1 Data

We collected the daily historical data of S&P 100 stocks spanning from January 01, 2007 to December 31, 2017. S&P 100 index was chosen because it represents approximately 63% of market capitalization of the S&P 500 and almost 51% of the market capitalization of U.S equity markets (Wikipedia, 2018). Also, 87 of the stocks listed in S&P 100 had been listed on stock market long enough to provide us enough historical data for our study. Data was downloaded through Quandl (Quandl, 2018) using the Quandl Python module (Quandl, 2018). Quandl offers many sources of data, few are free but most of them are paid data sets. We collected the historical data from one of the free sources, WIKI Prices (WIKI Prices, 2018).

Table 3-1: Sample historical data for AAPL

Date	Open	High	Low	Close	Volume	Ex-Dividend	Split Ratio	Adj. Open	Adj. High	Adj. Low	Adj. Close	Adj. Volume
1/3/2007	86.29	86.58	81.90	83.80	44225700	0	1	11.09	11.13	10.53	10.77	309579900
1/4/2007	84.05	85.95	83.82	85.66	30259300	0	1	10.80	11.05	10.77	11.01	211815100
1/5/2007	85.77	86.20	84.40	85.05	29812200	0	1	11.02	11.08	10.85	10.93	208685400
1/8/2007	85.96	86.53	85.28	85.47	28468100	0	1	11.05	11.12	10.96	10.98	199276700
1/9/2007	86.45	92.98	85.15	92.57	119617800	0	1	11.11	11.95	10.94	11.90	837324600

Table 3-1 shows the sample data for Apple stock downloaded from Quandl. For each day U.S stock market was open, daily historical data was collected. Date corresponds to the date for which data was collected for a stock. Open is the price stock opened at the start of trading day. High is the highest price stock reached during that trading day, while Low is the lowest price it reached. Close is the price stock closed at the end of trading day. Volume is the number of stocks traded during the trading session of a trade day. Ex-dividend is the time between the announcement and payment of a dividend, but it is irrelevant for our study.

Split ratio is the ratio for stock split (e.g. 2-for-1 would mean 1 stock is being split into 2 stocks). Split ratio helps us determine adjusted open, high, low, etc. Adjusted variables are necessary because un-adjusted stock price after stock split would show huge increase/decrease in stock price without reasonable cause. Let's consider Apple stock at 100 dollars stock price as an example. If Apple issued a stock split of 2-for-1, initially each stock was worth 100 dollars, after

the stock split, each stock will be worth 50 dollars. The stock price has changed from 100 dollars to 50 dollars just because of stock split. Though, there has been some studies performed to find out the market reaction to stock splits (Lamoureux & Poon, 1987) (Easley, O'Hara, & Saar, 2001), but that is a topic for another research. Therefore, we want to adjust the variables if any stock split happened in past. WIKI Prices already has adjusted variables based on stock splits, if any. We only used the adjusted variables (e.g. Adj-Open) and date for feature extraction and methodology.

3.2 Feature Set

Historical stock data can help models make a good future price prediction, but Open, Close, etc. indicators can help us derive much more useful indicators. We can find out the direction stock was moving in, how volatile it has been recently, how much has it changed in past n days, and so on. These indicators help us understand how the stock is been behaving over a period and based on that behavior how did it perform in future. With these indicators, models have a much better understanding of stock behavior and the outcome that behavior leads to. Table 3-2 lists the feature set, including all the derived technical indicators and their formulas.

Table 3-2: List of feature set, including technical indicators and their formulas

No.	Feature	Formula
0	Date	-
1	Open	-
2	High	-
3	Low	-
4	Close	-
5	Volume	-
6	Stock id	Unique, assigned to each stock symbol
7	SMA (t)	$= \frac{1}{n}\sum_{i=0}^{n-1} p_{m-i}$, where m is current date, n is (t), p is the closing price on $m - i$ day in past
8	SMA $(t * 2)$	$= \frac{1}{n}\sum_{i=0}^{n-1} p_{m-i}$, where m is current date, n is $(t * 2)$, p is the closing price on $m - i$ day in past
9	SMA $(t * 3)$	$= \frac{1}{n}\sum_{i=0}^{n-1} p_{m-i}$, where m is current date, n is $(t * 3)$, p is the closing price on $m - i$ day in past

10	EMA (t)	$= \begin{cases} p_m, & t = 1 \\ ((p_m - \text{EMA}(m-1)) * k) + \text{EMA}(m-1), & t > 1 \end{cases}$, where m is date, $k = 2 \div (t+1)$, p_m is the closing price on date m, $(m-1)$ is previous market date
11	EMA $(t * 2)$	$= \begin{cases} p_m, & t = 1 \\ ((p_m - \text{EMA}(m-1)) * k) + \text{EMA}(m-1), & t > 1 \end{cases}$, where m is date, $k = 2 \div ((t*2)+1)$, p_m is the closing price on date m, $(m-1)$ is previous market date
12	EMA $(t * 3)$	$= \begin{cases} p_m, & t = 1 \\ ((p_m - \text{EMA}(m-1)) * k) + \text{EMA}(m-1), & t > 1 \end{cases}$, where m is date, $k = 2 \div ((t*3)+1)$, p_m is the closing price on date m, $(m-1)$ is previous market date
13	Momentum (t)	$= p_m - p_{m-t}$, where p_m is the closing price on date m, and p_{m-t} is the closing price on t number of market days before m
14	Momentum $(t * 2)$	$= p_m - p_{m-t*2}$, where p_m is the closing price on date m, and p_{m-t*2} is the closing price on $t * 2$ number of market days before m
15	Momentum $(t * 3)$	$= p_m - p_{m-t*3}$, where p_m is the closing price on date m, and p_{m-t*3} is the closing price on $t * 3$ number of market days before m
16	Volatility (t)	$= \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(p_{m-i} - \mu)^2}$, where μ is the average stock price for t previous days, n is t, m is current date and p_{m-i} is the closing price on ith day in past
17	Volatility $(t * 2)$	$= \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(p_{m-i} - \mu)^2}$, where μ is the average stock price for t previous days, n is $t * 2$, m is current date and p_{m-i} is the closing price on ith day in past
18	Volatility $(t * 3)$	$= \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(p_{m-i} - \mu)^2}$, where μ is the average stock price for t previous days, n is t*3, m is current date and p_{m-i} is the closing price on ith day in past
19	Volatility historical	$= \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(p_{m-i} - \mu)^2}$, where μ is the average stock price for t previous days, n is total number of records available, m is current date and p_{m-i} is the closing price on ith day in past

In Table 3-2, t stands for number of days. Stock id is an integer we assigned to each stock. Every stock gets a unique id that we will use in our methodology. Simple Moving Average (SMA) is an average of previous n days of closing price. In our study, we calculated three different SMAs based on our prediction period, t. This was done because if we want to predict a year in future, it might be better to look at SMA of past one year, two years, and three years rather than a pre-defined number of days. Exponential Moving Average (EMA) is same as SMA, except EMA gives a higher weighting to recent prices. Since EMA weights recent prices higher, it is better for capturing more recent trends than SMA. Momentum refers to the rate of change on price movements for a stock. Momentum helps identifies trend as well, if a stock has been trending up, then that might mean it will continue doing so in near future. Volatility (standard deviation) refers to the size of changes in a stock's price. If a stock has high volatility, it would mean that the stock price could go rise or fall dramatically. If the volatility is low, it would mean the changes in stock price are lot more stable.

3.3 Normalization

Data needed to be normalized before continuing with methodology. This was because, a change of 100 dollars in a stock's price might be a 10% change, while in other stock's price be a 100% change. Also, the regression models are being trained on data from all stocks collectively, that is why it is necessary to normalize the data. Data for each stock was normalized individual. Following formula was used for normalization.

$$s_{f_i} = \frac{s_{f_i}}{\max(s_f)} \qquad 3.1$$

Where s represents the stock, f is the feature of s, i is the ith value for f, and $\max(s_f)$ returns maximum value for a stock's feature. This formula normalizes all values in the range of 0 and 1. Normalization is performed on all features that require it, exceptions are Date and Stock id.

Volume did not require normalization, but some regression models have slower performance when dealing with large values. That is why Volume was normalized as well. As for Volatility historical, normalization was not needed as each stock only has one value for it. But for

consistency and performance of regression models, Volatility historical was normalized as well based on its values for all stocks.

Chapter 4 Methodology

We proposed a Tailored Ensemble Approach (TEA) to tackle the stock price prediction challenge. An overall architecture of the system is shown below.

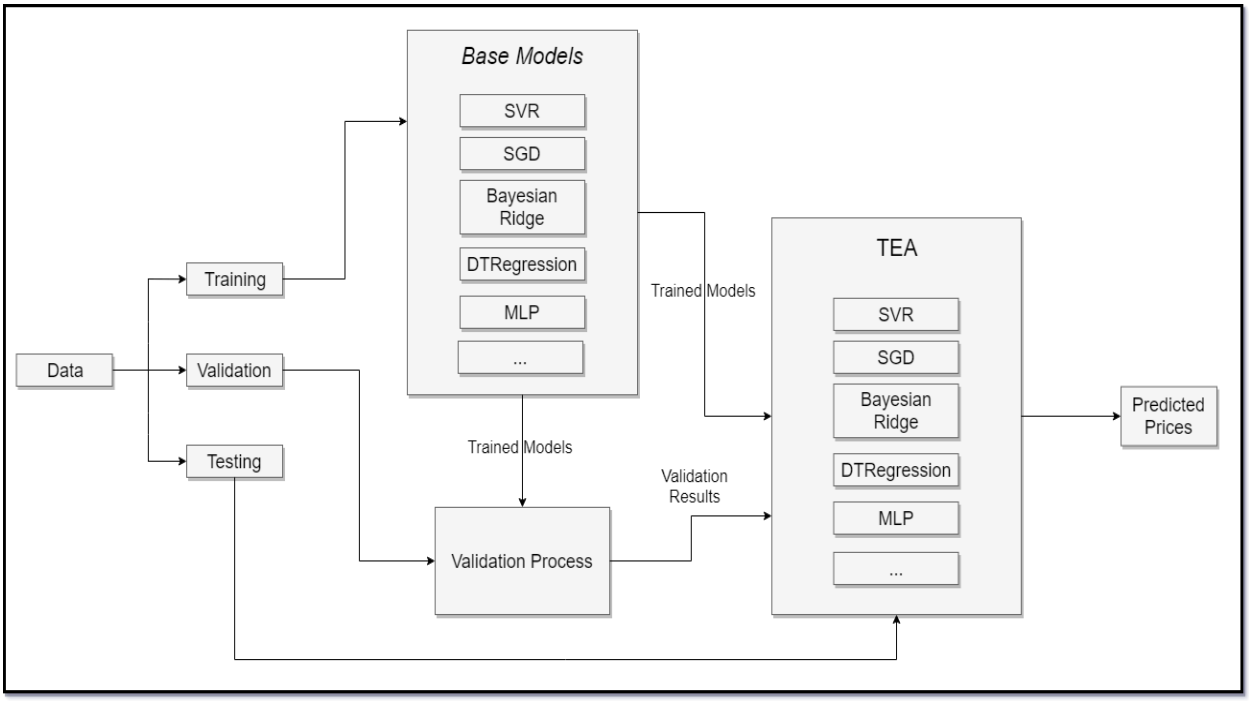

Figure 4-1: Overall system

Figure 4-1 shows the overall structure of the system. Data is split into Training, Validation, and Testing set. Distribution statistics of the data will be discussed in Chapter 5. Training data is used to train Regression models such as SVR, MLP, etc. Variations of each model are based on parameter setting or feature set used for model training. Each one of trained regression model and its variations are validated using the Validation dataset. Trained models, their validation results, and testing data is passed onto the ensemble system, which aggregates the results from selective subset of base models to calculate the final predicted price. Regression models are discussed in 4.1. Feature sets used for training the models are discussed in 4.2. Training Process is discussed in 5.4. Validation Process is discussed in 4.3. Lastly, the TEA is discussed in 4.4.

4.1 Regression Models

We chose many regression models, such as SVR, DT, etc. for the ensemble system. This was done to get variety of results based on several types of algorithms, e.g. linear, non-linear, probability based, etc.

4.1.1 Support Vector Regression (SVR).

SVR attempts to minimize the generalization error bound as to achieve generalized performance. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. SVR has been applied in various fields - time series and financial predictions, approximation of complex engineering analysis, convex quadratic programming and choices of loss functions, etc. (Chi-Jie, Lee, & Chiu, 2009) (Wu, Ho, & Lee, 2004). SVR is the most common application form of SVMs. An overview of the basic ideas underlying support vector machines for regression and function estimation has been given in (Basak, Pal, & Chandra, 2007).

To train SVR model, you must give parameter settings, such as kernel, epsilon, gamma, etc. We left most of the settings to default. For kernels, we trained three variations of SVR model, one for each of the popular kernel, ‘poly’, ‘rbf’, and ‘linear’ (Basak, Pal, & Chandra, 2007).

Figure 4-2 (Scikit, 2018) shows the results of SVR using various kernels. Linear kernel gives us a straight line based on the data because of the linear function. Polynomial kernel allows learning of non-linear models by developing polynomial function for the data. RBF kernel is lot more flexible than linear and polynomial kernel, which allows it to model the functions more closely to the data itself.

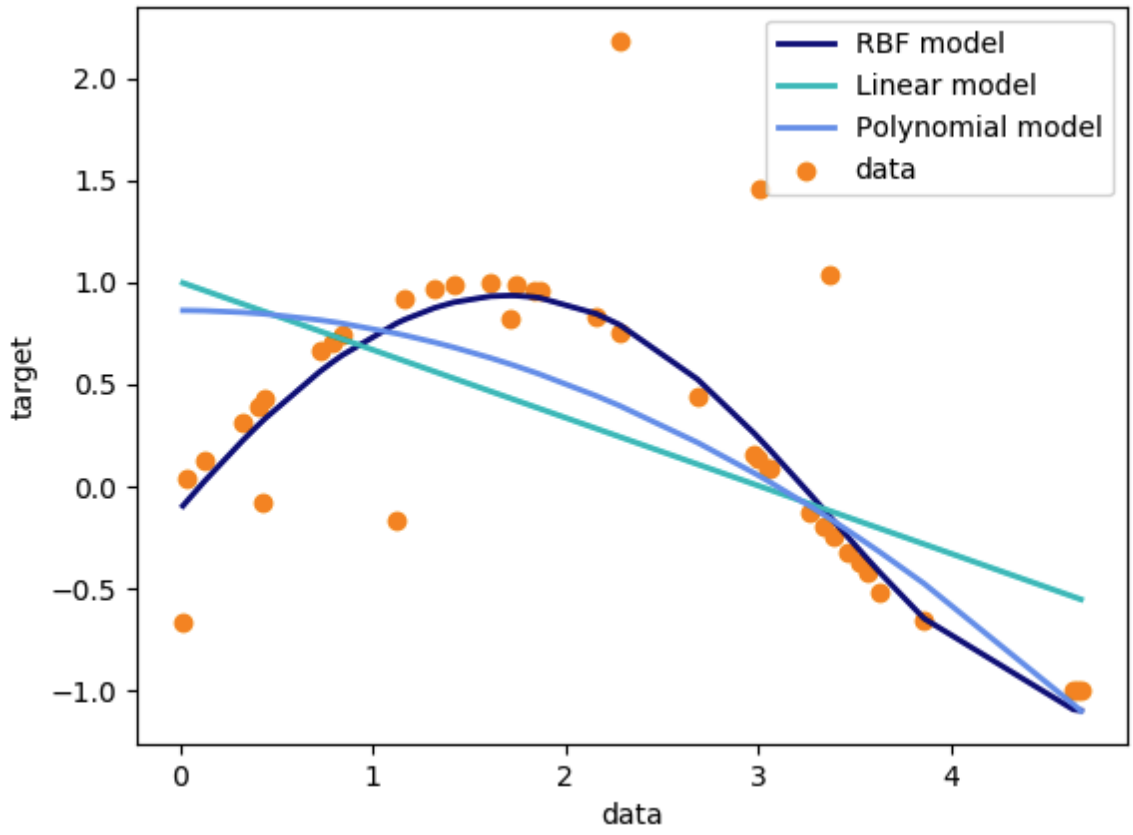

Figure 4-2: Support Vector Regression Kernels

4.1.2 Stochastic Gradient Descent Regression (SGD)

SGD (Scikit, 2018) (Bottou, 2010) is a stochastic approximation of the gradient descent optimization and iterative method for optimizing a function. SGD updates a set of parameters in an iterative manner to minimize an error function. It is well suited for regression techniques on data which has thousands or more number of rows. SGDregressor used in this study implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties to fit linear regression models. The linear model is fitted by minimizing a regularized empirical loss with SGD. Its parameters were set to default as well.

4.1.3 Decision Tree Regression

Decision tree (Tso & Yau, 2007) learning uses a decision tree to go from observation about an item to conclusions about the item's target value. Decision tree is a tree like structure that leads to a conclusion based on possible consequence. Below is an example of decision tree.

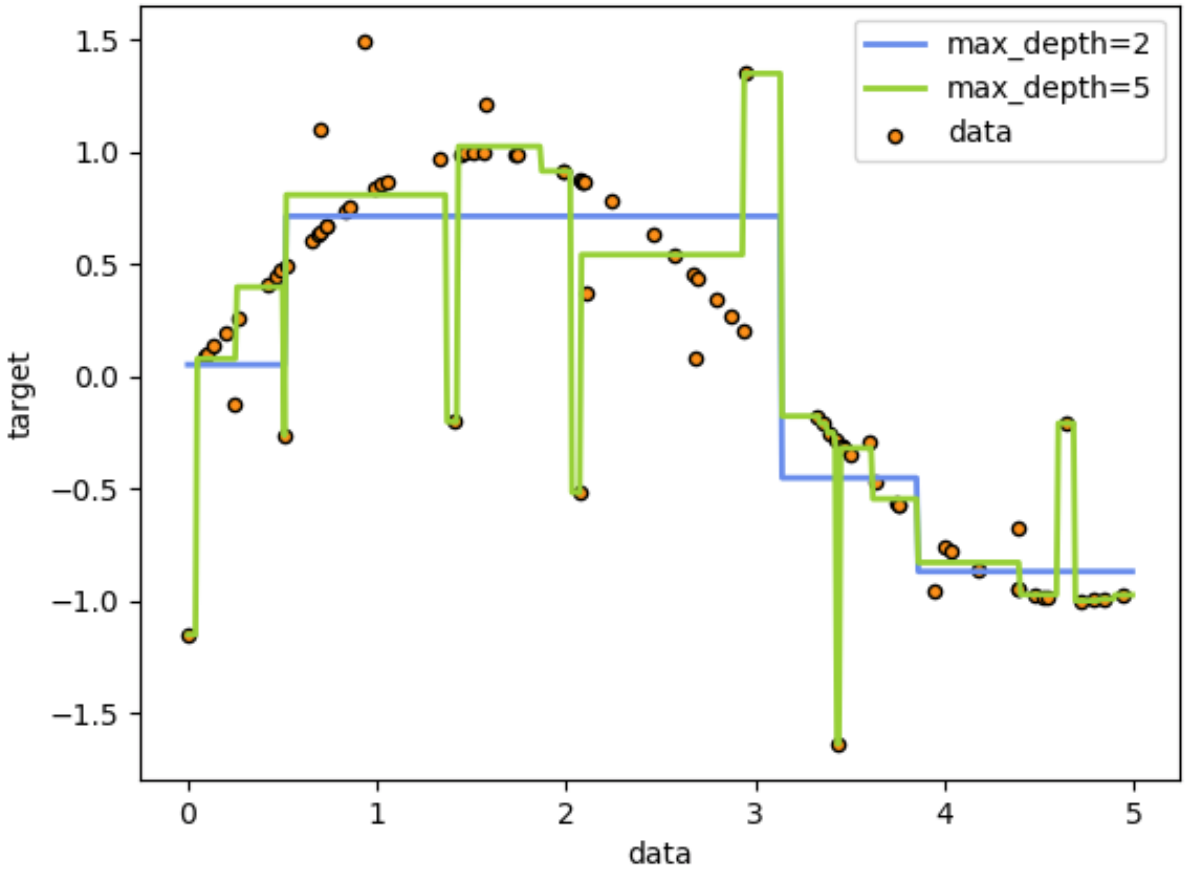

Figure 4-3: Decision Trees

Figure 4-3 (Scikit, 2018) shows decision trees trained for regression models with different depth setting. As we increase the depth of decision tree, it learns more details of the training data. If the maximum depth of the tree is set high, the decision tree learns minute details of the training data and learn from noise, i.e. they overfit.

Decision tree learning is one of the predictive modelling approaches used in statistic, data mining and machine learning. Decision tree where the target variable can take continue value are called regression trees. Like other base models, the parameter values for decision tree regressor are set to default as well.

4.1.4 Bayesian Ridge Regression

Bayesian linear regression is essentially a linear regression model with priors on the coefficients and on the noise so that in absence of data, the priors can take over (Raftery , Madigan, & Hoeting, 1997). Priors (or prior probability distribution) is the probability of an event to take place

before any evidence is taken into consideration. On the other hand, ridge regression (Hoerl & Kennard, 1970) is a technique for analyzing multiple regression data that suffer from multicollinearity. Multicollinearity is a phenomenon in which one predictor variable can be linearly predicted from others with substantial degree of accuracy. But this also brings up an issue, and that is when you are fitting a regression model, predictors could be correlated with other predictors in the model (Farrar & Glauber, 1967). This could result in estimates being very sensitive to minor adjustment in model. Ridge regression is used for dealing with this problem (Maddala & Lahiri, 1992).

4.1.5 Multi-layer Perceptron (MLP)

MLP (Agirre-Basurko, Ibarra-Berastegi, & Madariaga, 2006) is a supervised learning algorithm that learns a function to convert the feature set into the output. There can be several hidden layers between feature set and the output, which has some function to help perform the conversion. Let's assume the input vector is x, the hidden layer activations h, and the output activation y. You have some function f, that maps from x to h and another function g that maps from h to y. So, the hidden layer's activation is f(x) and the output of network is g(f(x)). The idea is that two functions (g(f(x)) can compute things that f and g cannot do individually. MLP has the capability to learn non-linear models, which gives a chance to have a good variation in our list of base models.

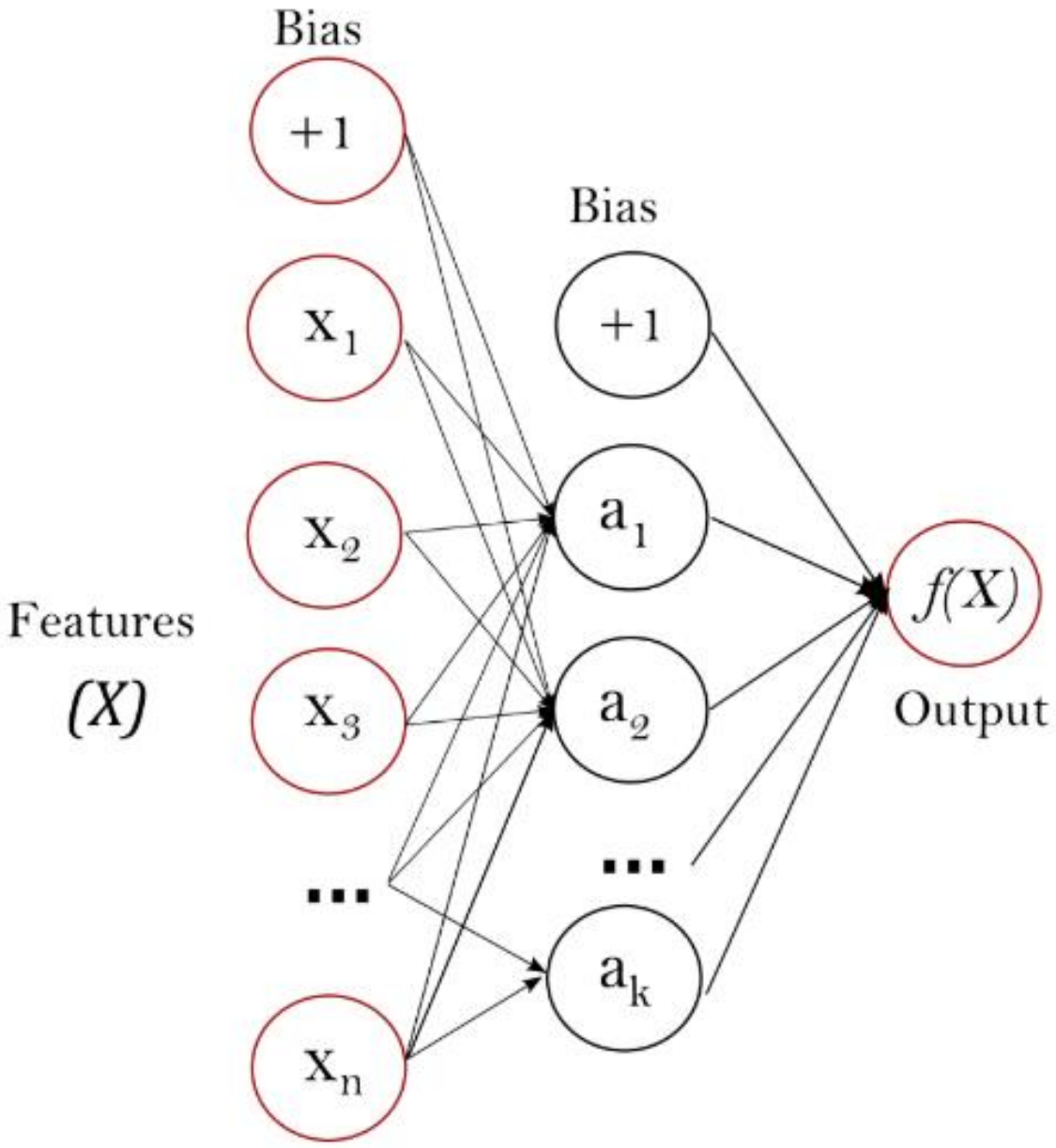

Figure 4-4: Multi-layer Perceptron

Figure 4-4 (Scikit, 2018) shows MLP that consists of one hidden layer. The leftmost layer, known as the input layer, consists neurons representing the input features. Each neuron in hidden layer transfers the values from the previous layer with a weighted linear summation, followed by a non-linear activation function. The output layer receives the values from the last hidden layer and transforms them into output value.

4.2 Feature Sets

Subsets from complete list of Feature Set (Table 3-2) were formed to train the models. A list of feature set is shown below in Table 4-1.

Table 4-1: Feature sets used for training

Feature Set	Features
1	Open, High, Low, Close, Volume, SMA (t), SMA (t ∗ 2), SMA (t ∗ 3), EMA (t), EMA (t ∗ 2), EMA (t ∗ 3), Momentum (t), Momentum (t ∗ 2), Momentum (t ∗ 3), Volatility (t), Volatility (t ∗ 2), Volatility (t ∗ 3), Volatility (historical)
2	Open, Momentum (t), Momentum (t ∗ 2), Momentum (t ∗ 3)
3	Open, Volatility (t), Volatility (t ∗ 2), Volatility (t ∗ 3), Volatility (historical)
4	Open, High, Low, Close, Volume
5	Open, SMA (t), EMA (t), Momentum (t), Volatility (t)

Feature Set 1 includes all the features used in this study. Feature Set 2 focused on training models based on stock momentum features, feature set 3 focused on volatility, feature set 4 focused on looking at stock's moving average (simple and exponential), momentum, and volatility same period in past as we are predicting in future. Lastly, feature set 5 contains only the features that we received from raw data and were not derived from it. Through the research in stock market, we have concluded that not every stock should be predicted using the same feature set. Let's take TSLA (Tesla Inc., 2018) and MSFT (Microsoft Corp., 2018) for example. Tesla has never reported profitable quarter, in fact it has reported loss for every single quarter so far, but still its stock price has increased tremendously over past few years. On the other hand, Microsoft reports profitable quarters and its stock price has been increasing at a stable rate. It would make sense to use Profitability as feature when predicting stock price for Microsoft, but not for Tesla. Though, we are not using fundamental analysis indicators, we hope to see that models trained with certain technical indicators are better suited for certain stocks. Each of the regression model was trained on every feature set listed in Table 4-1.

4.3 Validation Process

In validation process, we want to see the performance of each trained model on individual stock in the dataset. Validation is a crucial step for TEA as it allows the system to determine which

models are better suited for certain stocks. Not only we determine the importance of each model, we also determine which combination (s) of a model and feature set is best for predicting the price of a certain stock through their performance for that stock using the validation dataset.

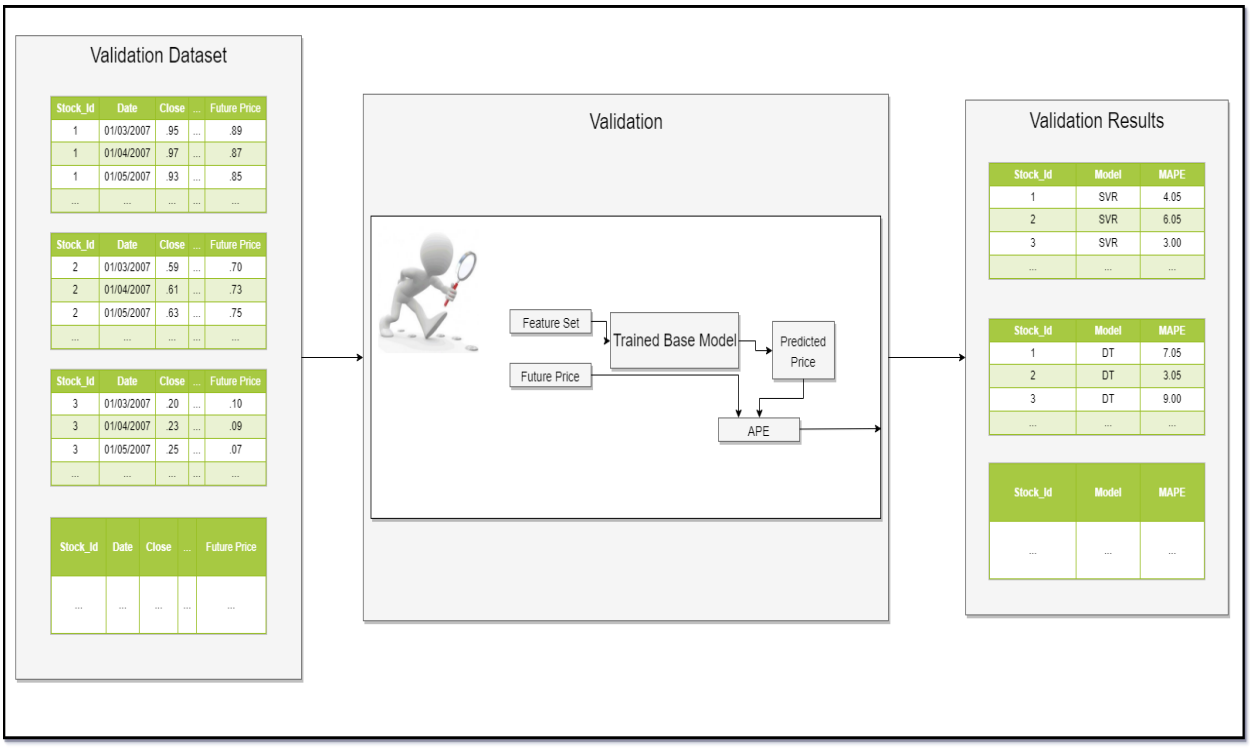

Figure 4-5: Validation Process

Figure 4-5 shows the validation process. Validation dataset is used for validating the performance of each trained model. We want to find out how each model performs for each individual stock in our dataset. For every record in validation set for a stock, Absolute Percent Error (APE) was calculated.

$$APE_i = 100 * \frac{|Y_i - P_i|}{Y_i} \qquad 4.1$$

Y_i is the actual stock price, and P_i is the predicted price for the i^{th} daily historical record for a stock in the dataset. APE calculates the absolute difference between predicted price and

actual price. To measure the overall performance of a trained model for a stock, average of all APE values for that model and stock were taken, which is also known as MAPE.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} APE_i \qquad 4.2$$

MAPE calculates the average performance of a trained model for a stock in our dataset. This helps us determine which combination (s) of model and feature set performs better for certain stock.

4.4 Tailored Ensemble Approach (TEA)

We implemented a Tailored Ensemble Approach (TEA). Rather than considering all models for the ensemble system, it is important to look at which models are good at predicting a stock. This is important to consider as not every model might provide us with best results for every stock in the dataset. Stocks have different pattern, and certain models might be better at predicting certain stocks. Also, same model trained on different feature set might give us different results. That is why it is important to consider which feature set is important for predicting the price for each unique stock in our dataset. This variability is not only based on stock symbols, but also for prediction periods. When predicting for different prediction periods, combination (s) of certain model and feature set that give us the best results during short-term might not provide us with similar performance in long-term predictions. Therefore, it is also important to consider which combination of models and feature sets are best for each prediction period, ranging from short-term to long-term.

This is where one of our feature, Stock id, comes into play. Through the validation results, we found the error rates of models for each Stock id. Based on those validation results, we can find out which models are the ones our ensemble system should consider for predicting the price of given Stock id.

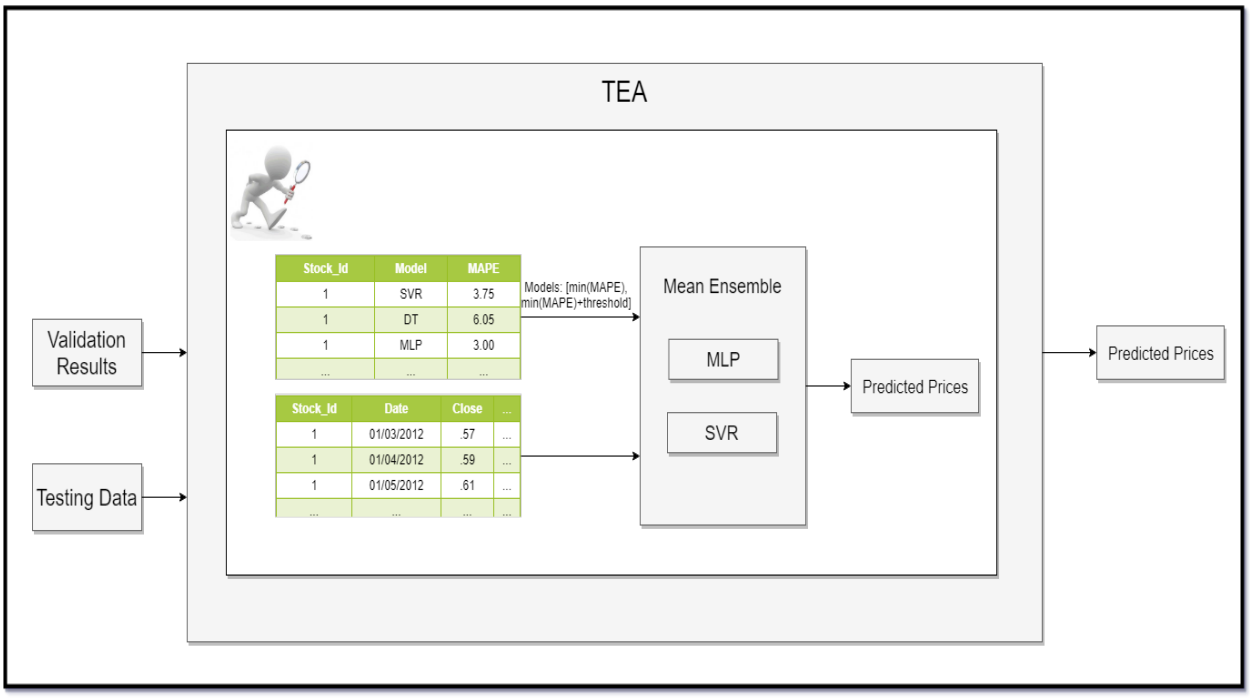

Figure 4-6: Tailored Ensemble Approach

Figure 4-6 shows the Tailored Ensemble Approach (TEA). As the figure shows, in one of the iteration, system will be predicting future prices for stock where id equals 1. The validation results for that stock shows that MLP is the best one, followed by SVR, and so on. Rather than just considering the best model, we want to consider few of the best ones. This is because few models could have the error rate very close to the best one. If that is the case, then it would be better to consider the results of all those models as well. For this, we set a threshold value, so if any of the other models are at most threshold percent worse than the best one, they will be considered in the ensemble system as well. We ran an experiment to optimize the threshold value, so we get the best possible predicted prices.

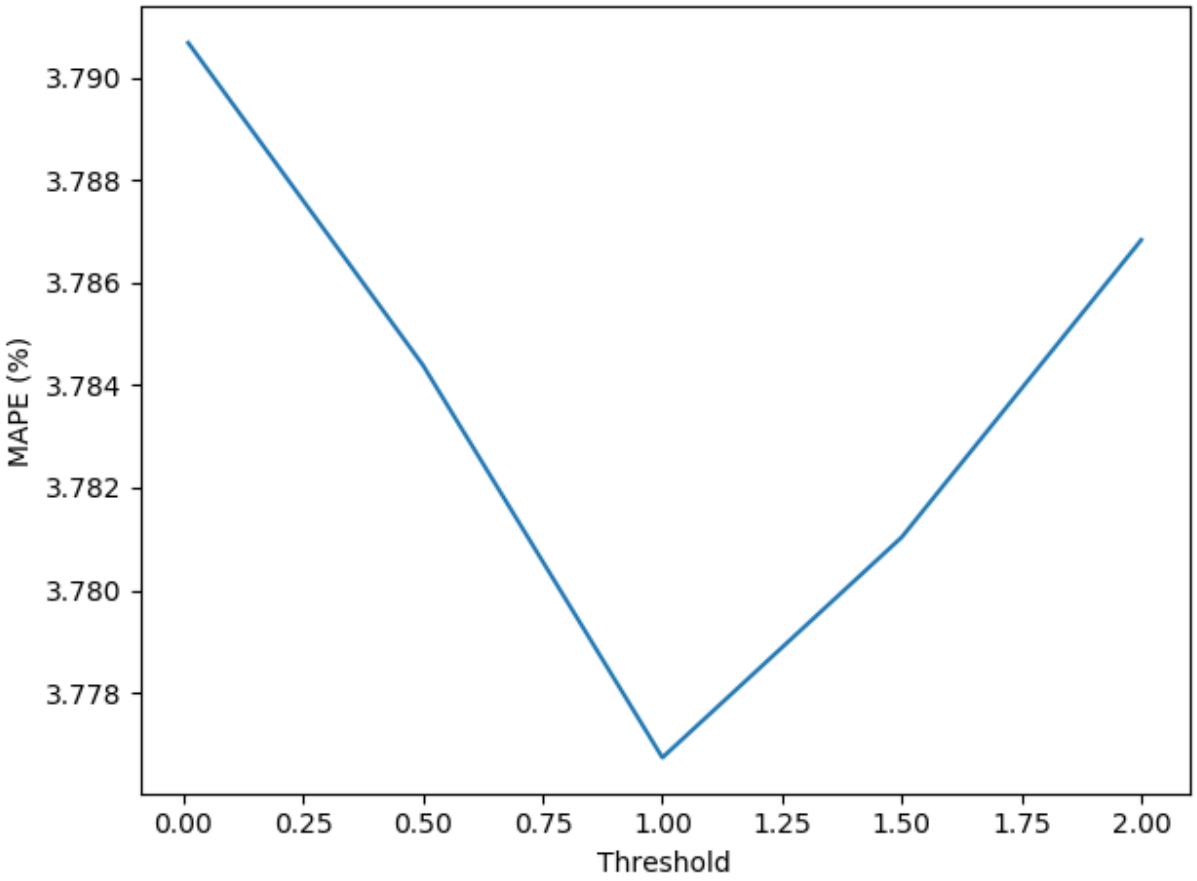

Figure 4-7: TEA MAPE vs. Threshold

Figure 4-7 shows the result of the experiment. In the experiment, prediction period was set to 31. We ran this optimization process only for a single long-term prediction period. Optimizing threshold for each prediction period would have been too exhaustive given our evaluation process. So, once the threshold was optimized, we used the same threshold value for all prediction periods. The threshold values of (0, 0.25, 0.5, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00) were tried. Increasing the threshold value meant more models were considered for predicting the stock price by the system. As the threshold value increases to 1.00, MAPE of TEA decreases. Increasing the threshold value past 1.00 increases the MAPE of TEA. Therefore, we decided that 1.00 was the best threshold value to use for our system. The threshold values were only tested for a long-term prediction period. When using the system for investing, it would be better to optimize the threshold value based on the term of investing.

Once the threshold value was optimized, the selected models were considered in a mean ensemble system for that stock, which produced the predicted prices. This system allows separate set of models to be considered in mean ensemble system based on stock for which we are predicting future price. This tailored approach will pair different stocks with separate set of models that are best suited for them.

Chapter 5 Evaluation

For evaluation, we wanted to compare the performance of proposed methodology with base methodologies over many variations. This was done to check the performance consistency of the proposed methodology and confirm that the results are not coincidental.

5.1 Cross-validation: Holdout

Cross-validation is any of various validation techniques to evaluate the performance of predictive models (Kohavi, 1995). Holdout method is one of cross-validation techniques, where the dataset is split randomly into training and testing set. This split is done without replacement. This means if a data point was assigned to training set, it cannot be assigned to testing set as well. Training set is used to train the models, while testing set is used to check the performance of trained models. The ratio of dataset used in training and testing set depends upon the data we are working with and amount of data available. 50/50 split is where 50% of the data is used for training, while remaining 50% is used for testing.

5.2 Data Split in Holdout method

In this study, we used the holdout method but with a slight modification. Rather than splitting the dataset into training and testing set, we split the dataset into training, validation, and testing. Training set served the purpose of training base models in our system, such as SVR, DT, etc. Validation set is used to validate the performance of these base models for each individual stock in our dataset, this process was explained in 4.3. Lastly, the testing dataset was used to test the proposed model and all base models to compare the performance of our proposed model versus the baseline approaches.

Also, in the standard holdout method, data is split randomly into different sets. But this would not be the best option for stock market dataset as it is a time series data. We want to train with the past data to predict the future data. Therefore, rather than randomly splitting the dataset, first few years were assigned for training and remaining years for validation and testing. Also, we wanted to have each stock trained, validated, and tested. That is why, the data was split into training, validation, and testing set at the stock level and eventually aggregated together

into one training, one validation, and one testing set for the entire dataset. This was done because we wanted each stock's pattern to be considered when training the base models. For validation, it was a must that we have validation data for each stock because our system tries to find model (s) which perform best for each stock. Lastly, we wanted to have each stock in our testing set as we wanted to evaluate the performance of proposed model and baseline approaches evenly on all stocks in our dataset.

The first split of dataset was into two subsets, training, validation and testing. The validation and testing subset was further split into validation set and testing set. Alternative months of each year, from the validation and testing subset, were assigned to validation set and testing set. Odd months of each year for used for validation, while even months were used for testing set. We wanted to have variety of data points for the validation and testing process. Since each model for each stock was validated, the validation dataset for validating each model at stock level was relatively small. But this was necessary as we wanted to validate models for each individual stock to find out which models are better suited for each stock. Therefore, variety of data in validation set for each stock gives us the performance of a model for a stock that is based on the bigger picture rather than certain prediction period. Let's consider the prediction for AAPL stock during the stock market crash prediction period. If the models were validated during market crash period, then the models which perform best during that stock market crash period for AAPL stock, will have the best performance. But if we are testing during the period when there was no market crash, models which performed best during crash time, might not be the best models for AAPL stock price prediction anymore. Therefore, variety of data in validation and testing set for each stock was necessary and dataset was split accordingly.

5.3 Repeated Holdout Method

Rather than evaluating the model once, we performed this multiple time on different training, validation, and testing set. This was done to ensure consistency of evaluation results and make sure they were not coincidental. Usually this is done using the K-Fold Cross Validation technique, but in this case since we do not want to randomly split the dataset, or train using future dataset and predict for past dataset, we chose to repeat the holdout method using different splits to

achieve comparable results. Table 5-1 shows the split of dataset into training, validation, and testing.

Table 5-1: Training, Validation, and Testing dataset for each Holdout id

id	Training (YYYY-MM-DD)	Validation and Testing (YYYY-MM-DD)	Validation	Testing
1	2007-01-01 to 2010-12-31	2011-01-01 to 2016-12-31	Odd months (1, 3, 5, …)	Even months (2, 4, 6, …)
2	2007-01-01 to 2011-12-31	2012-01-01 to 2016-12-31	Odd months (1, 3, 5, …)	Even months (2, 4, 6, …)
3	2007-01-01 to 2012-12-31	2013-01-01 to 2016-12-31	Odd months (1, 3, 5, …)	Even months (2, 4, 6, …)
4	2007-01-01 to 2013-12-31	2014-01-01 to 2016-12-31	Odd months (1, 3, 5, …)	Even months (2, 4, 6, …)
5	2007-01-01 to 2014-12-31	2015-01-01 to 2016-12-31	Odd months (1, 3, 5, …)	Even months (2, 4, 6, …)

In id = 1, the models were trained on first 4 years of data, while validation and testing were performed on remaining 6 years of data. With the increment of id by 1, a year was moved from the validation and testing set to training set. The purpose of doing this was to confirm that the proposed approach will consistently outperforms base models in multiple scenarios. That's why we started with 6 years in the validation and testing dataset and decreased that set until only 2 years were remaining in it. We did not continue at this point, because if 2 years were in validation and testing set, that would give only one year in validation, and one year in testing. Reducing the validation and testing dataset any further would result in lack of data to provide accurate results.

Also, rather than predicting for certain duration in future, we wanted to predict for multiple different time durations. This was done to test whether same models with same configurations have similar performance when predicting for different time durations. The list of prediction periods for stock price is follow: 2, 3, 4, 5, 10 ,15, 21, 31, 42, 52, 63, 126, 189, and 252.

Stock price predictions are split into two main time durations, short-term and long-term. Anything less than a month is considered short-term prediction, and anything over a month is

considered as a long-term prediction. Our short-term predictions consider the time duration ranging from 2 days to 21 days stock market is open. In a month, on average there are 21 days when stock market is open. So, if we are predicting the price change after 21 stock market days, we are predicting one month in advance. Rest of the predictions are based on long-term durations, ranging from 31 days to 252 stock market days. 252 stock market days is equivalent to one year of real time.

5.4 Training

We trained 7 base models for the ensemble system. But each of the base models were trained on all 5 of feature sets shown in Table 4-1. This would result in 35 base models with different configurations. To perform repeated holdout step, we trained each of the base model on 5 different training set shown in Table 5-1. Performing evaluation process for one prediction period (e.g. 2 days in future) would result in training 175 (35 base models * 5 training sets) base models. We ran our proposed methodology for 14 prediction periods, which resulted in 2450 (175 for each time duration * 14-time durations) base models. Training time of each models was relatively close to each other, except SVR models, which took exponentially longer time to converge.

In this study, we chose not to optimize the base models for few reasons.

Optimizing the base models will improve their accuracy, but that also means it will improve the accuracy of the proposed approach itself. Since our approach picks base models best suited for each stock, if the performance of base models increases, then the performance of our ensemble system increases as well. This is because once the base models are optimized and their accuracy increases, the ensemble system will, as usual, pick the best models for each stock. Let's consider AAPL stock, and DT and MLP base models are best at predicting its price, when base models were not optimized. If we were to optimize the base models, the list including best model(s) for AAPL stock would be a new one. If this list does not contain DT or MLP, that would mean that after optimization process there are other models which perform better than the optimized version of DT and MLP and the ensemble system should pick those models instead. DT and/or MLP could still be part of the best models after optimization process, and in that case ensemble system would still consider their results for the prediction. If base models are

optimized, they will produce better results, but the same goes for our proposed approach as well. Also process to optimize models is a much longer process than just training a model. To evaluate our approach, we trained 2450 base models. Optimizing each one of them would be unrealistic. Optimization is a step that should be added when we consider using this approach in an application, but as we discussed, to show improvement of our proposed ensemble approach over base models, it's irrelevant for this study.

5.5 Validation

Through validation process, our system determines which set of model and feature set combination is better suited for each stock in our database. Since we have 2450 trained models and 87 stocks, it is not convenient to show all the validation results. Following table shows which base models were used by our system after validation process for AAPL and AMZN stock.

Table 5-2: Models chose by TEA for AAPL and AMZN during each prediction period

Prediction Period	AAPL	AMZN
2	'bayesian_fs1', 'sgd_fs1', 'svr_linear_fs1'	'bayesian_fs4', 'svr_linear_fs4', 'svr_rbf_fs4'
3	'bayesian_fs1', 'bayesian_fs4', 'sgd_fs1', 'svr_linear_fs1', 'svr_linear_fs4'	'bayesian_fs4', 'svr_linear_fs4', 'svr_rbf_fs4'
4	'bayesian_fs1', 'bayesian_fs4', 'sgd_fs4', 'svr_linear_fs1', 'svr_linear_fs4'	'bayesian_fs4', 'svr_linear_fs4', 'svr_rbf_fs4'
5	'bayesian_fs4', 'mlp_fs4', 'sgd_fs4', 'svr_linear_fs1', 'svr_linear_fs4', 'svr_rbf_fs4'	'bayesian_fs4', 'svr_linear_fs4', 'svr_rbf_fs4'
10	'svr_linear_fs4', 'svr_rbf_fs4'	'bayesian_fs4', 'svr_linear_fs4', 'svr_rbf_fs4'
15	'svr_linear_fs4', 'svr_rbf_fs4'	'svr_linear_fs4', 'svr_rbf_fs4'
21	'bayesian_fs4', 'svr_linear_fs4', 'svr_rbf_fs4'	'svr_linear_fs4'
31	'svr_linear_fs2'	'svr_linear_fs4'
42	mlp_fs2', 'svr_linear_fs2', 'svr_rbf_fs2'	'sgd_fs5', 'svr_linear_fs2', 'svr_linear_fs4', 'svr_linear_fs5'
52	'svr_linear_fs5'	'sgd_fs5', 'svr_linear_fs2', 'svr_linear_fs5', 'svr_rbf_fs2'

63	'svr_linear_fs5'	'sgd_fs5', 'svr_linear_fs5'
126	'svr_linear_fs2', 'svr_linear_fs4'	'svr_linear_fs2', 'svr_linear_fs5'
189	'mlp_fs2'	'sgd_fs5'
252	'dt_fs2', 'mlp_fs2'	'dt_fs5'

Table 5-2 shows the models used by our system when predicting prices for AAPL and AMZN stocks at each prediction period. We can see variety of model and feature set combination being used depending on the stock and prediction period. Feature set 5 seems more common for long-term predictions, while feature set 4 seems more common for short-term predictions, when it comes to AMZN stock. As for AAPL stock, feature set 2 and 5 seems more common for long-term predictions, and feature set 1 and 4 seems common for short-term predictions. If we look at the models being used, Bayesian seems to be one of the popular models for short-term predictions in both stocks. For long-term predictions, we do not see any Bayesian model (or variations of it). This means that Bayesian models is useful in short-term predictions, but not for long-term predictions. SVR model with linear kernel seems to be an excellent choice for long-term and short-term predictions for both stocks. We also see DT, SGD, and MLP regressors being used but only for certain time periods.

These validations results show that based on different time periods and stocks, the models and feature set combination which gives the best performance are different. Even when it comes to same stock, to predict for different time periods, it is important to consider separate set of features and different models. Therefore, validation served a very important role in our proposed approach.

5.6 Testing

To compare the performance of models, MAPE was calculated. The formula for MAPE is as follows.

$$MAPE = (\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i - P_i|}{Y_i}) * 100 \qquad 5.1$$

All the base models (35 in total) and the proposed approach were tested on variation of testing data. Following table shows the overall results.

Table 5-3: MAPE values for top 10 models for each holdout id and prediction period combination

Row	Holdout id	Prediction Period	TEA	svr_rbf_fs4	svr_linear_fs4	svr_rbf_fs5	svr_rbf_fs3	svr_linear_fs5	svr_linear_fs2	svr_rbf_fs2	mlp_fs4	bayesian_fs4
0	1	2	1.21	1.41	1.25	1.42	1.66	1.33	1.41	1.59	1.34	1.24
1	2	2	1.13	1.29	1.18	1.31	1.53	1.26	1.33	1.46	1.23	1.17
2	3	2	1.13	1.27	1.18	1.32	1.49	1.26	1.33	1.45	1.24	1.17
3	4	2	1.16	1.29	1.20	1.31	1.52	1.28	1.36	1.45	1.34	1.19
4	5	2	1.21	1.32	1.24	1.34	1.53	1.32	1.41	1.49	1.37	1.23
5	1	3	1.59	1.76	1.62	1.80	1.99	1.72	1.79	1.93	1.81	1.62
6	2	3	1.49	1.63	1.52	1.66	1.83	1.62	1.68	1.79	1.62	1.53
7	3	3	1.50	1.62	1.52	1.66	1.80	1.62	1.68	1.77	1.55	1.53
8	4	3	1.54	1.64	1.56	1.67	1.82	1.65	1.71	1.78	1.62	1.56
9	5	3	1.58	1.66	1.59	1.71	1.83	1.69	1.74	1.82	1.65	1.60
10	1	4	1.91	2.05	1.93	2.10	2.28	2.03	2.10	2.23	2.18	1.95
11	2	4	1.79	1.89	1.81	1.96	2.09	1.91	1.96	2.08	1.86	1.82
12	3	4	1.79	1.88	1.81	1.95	2.06	1.91	1.96	2.06	1.87	1.82
13	4	4	1.84	1.91	1.85	1.97	2.09	1.95	1.99	2.08	1.89	1.86
14	5	4	1.86	1.93	1.87	2.00	2.10	1.98	2.02	2.09	1.99	1.88
15	1	5	2.16	2.32	2.18	2.32	2.51	2.29	2.33	2.44	2.21	2.20
16	2	5	2.03	2.16	2.05	2.17	2.31	2.15	2.20	2.29	2.13	2.07
17	3	5	2.05	2.13	2.06	2.17	2.29	2.15	2.20	2.28	2.11	2.07
18	4	5	2.09	2.17	2.10	2.21	2.33	2.20	2.24	2.32	2.16	2.11
19	5	5	2.10	2.16	2.11	2.22	2.33	2.22	2.25	2.32	2.17	2.12
20	1	10	3.09	3.17	3.12	3.21	3.45	3.23	3.27	3.34	3.23	3.19
21	2	10	2.99	3.04	3.01	3.08	3.26	3.11	3.13	3.19	3.34	3.08
22	3	10	3.04	3.08	3.06	3.12	3.23	3.15	3.17	3.22	3.08	3.11
23	4	10	3.08	3.08	3.10	3.15	3.23	3.18	3.19	3.23	3.12	3.12
24	5	10	3.19	3.18	3.21	3.26	3.34	3.31	3.30	3.32	3.32	3.22
25	1	15	3.91	3.93	3.95	3.99	4.34	4.07	4.11	4.12	3.95	4.08
26	2	15	3.78	3.80	3.82	3.83	4.08	3.91	3.94	3.95	3.90	3.95
27	3	15	3.77	3.77	3.81	3.81	3.94	3.88	3.90	3.92	3.95	3.91
28	4	15	3.81	3.77	3.83	3.82	3.90	3.89	3.91	3.93	3.89	3.89
29	5	15	4.09	4.03	4.14	4.11	4.21	4.19	4.22	4.24	4.13	4.16
30	1	21	4.64	4.69	4.73	4.76	5.28	4.87	4.93	4.92	5.62	4.95
31	2	21	4.46	4.59	4.61	4.59	4.96	4.71	4.76	4.76	4.53	4.82
32	3	21	4.43	4.48	4.52	4.50	4.65	4.60	4.64	4.62	4.51	4.70
33	4	21	4.43	4.44	4.55	4.45	4.54	4.61	4.64	4.60	4.49	4.63

34	5	21	4.75	4.71	4.87	4.74	4.86	4.93	4.96	4.93	4.73	4.88
35	1	31	5.75	5.95	5.96	6.20	7.05	6.18	6.14	6.31	6.51	6.36
36	2	31	5.57	5.86	5.77	5.99	6.47	5.92	5.93	6.10	6.05	6.16
37	3	31	5.39	5.61	5.59	5.76	5.85	5.68	5.68	5.77	5.63	5.89
38	4	31	5.34	5.48	5.66	5.56	5.53	5.71	5.71	5.71	6.11	5.80
39	5	31	5.64	5.70	5.99	5.75	5.87	6.03	6.02	6.03	5.77	6.02
40	1	42	6.60	7.23	6.94	7.49	8.90	7.21	7.21	7.42	7.03	7.52
41	2	42	6.21	6.91	6.51	7.24	7.84	6.82	6.71	7.01	9.17	7.09
42	3	42	6.09	6.56	6.36	6.80	6.74	6.55	6.48	6.62	7.43	6.78
43	4	42	5.97	6.20	6.40	6.27	6.17	6.50	6.47	6.35	6.45	6.60
44	5	42	6.23	6.35	6.77	6.39	6.54	6.89	6.80	6.72	6.41	6.83
45	1	52	7.33	8.54	7.91	8.71	10.74	8.19	8.21	8.85	8.94	8.78
46	2	52	6.80	8.09	7.27	8.30	9.26	7.55	7.45	8.17	9.77	8.12
47	3	52	6.56	7.43	6.93	7.52	7.49	7.11	7.06	7.35	8.25	7.58
48	4	52	6.29	6.67	6.89	6.77	6.50	6.99	6.94	6.85	6.92	7.15
49	5	52	6.57	6.74	7.31	6.78	6.92	7.39	7.30	7.03	6.77	7.37
50	1	63	8.30	10.03	8.91	10.16	12.60	9.26	9.32	10.46	11.33	10.11
51	2	63	7.63	9.44	8.12	9.53	10.53	8.35	8.44	9.47	10.59	9.24
52	3	63	7.12	8.40	7.61	8.17	7.82	7.70	7.78	8.20	10.36	8.40
53	4	63	6.65	7.11	7.48	7.25	6.93	7.54	7.56	7.59	7.25	7.76
54	5	63	6.97	7.23	7.96	7.34	7.48	7.98	7.91	7.78	7.18	7.99
55	1	126	10.43	16.10	11.63	18.12	16.26	12.26	12.04	13.47	16.54	13.66
56	2	126	9.67	14.55	10.86	15.53	11.36	11.01	11.10	13.35	17.28	12.72
57	3	126	8.40	10.01	10.12	9.96	9.07	10.23	10.35	10.53	10.56	11.10
58	4	126	8.16	9.07	10.64	9.23	9.32	10.68	10.77	9.91	9.39	10.82
59	5	126	8.28	9.91	11.69	9.62	10.01	11.69	11.78	10.54	9.89	11.79
60	1	189	11.31	14.77	13.38	18.14	14.41	13.41	13.64	20.36	14.47	13.78
61	2	189	10.81	14.49	13.35	13.94	11.79	13.16	13.41	14.23	14.42	14.10
62	3	189	9.49	11.05	12.83	11.29	11.03	12.95	13.00	11.88	12.05	13.32
63	4	189	9.50	11.01	14.06	11.54	11.47	14.05	14.18	11.78	11.25	14.19
64	5	189	8.54	11.37	14.76	10.97	11.17	14.59	14.87	11.55	11.26	15.02
65	1	252	11.70	15.29	14.43	14.61	15.15	14.30	14.69	17.09	15.32	14.40
66	2	252	10.86	16.55	14.62	14.97	12.97	14.50	14.69	15.72	16.30	15.22
67	3	252	8.73	11.33	13.94	11.65	11.30	14.16	13.92	11.57	11.93	14.03
68	4	252	8.91	11.43	15.74	10.93	11.40	15.28	15.60	13.39	11.81	15.58
69	5	252	7.64	11.60	15.89	11.62	10.74	14.83	15.72	11.88	11.83	16.01

Table 5-3 shows the MAPE values for top 10 model. Models were ranked based on their average MAPE value throughout all of test cases. So, the best model is the one with lowest average MAPE value. Smallest MAPE value for each row is highlighted in green. The table shows results of each evaluation we performed considering the variability in Holdout Id, prediction period. The green color filled cells shows that the proposed approach is the one with lowest MAPE in almost all cases. In row 24, 28, 29, and 34 we see that a variation of SVR (SVR with RBF

kernel and trained on feature set #4) outperforming our model, though only just slightly as seen through the MAPE values. In those rows, we are working on validation and testing dataset in Holdout Id 4 and 5. In Holdout Id 4, we are validating on 18 months and test on 18 months of data. In Holdout Id 5, we are validating on 12 months and testing on 12 months of data. Compare to other Holdout Id's where we are working with bigger validation and testing data, we might see some inconsistency in results, such as in the rows mentioned above.

As the prediction period increases, error rate is increasing as well. This is true for all models. This shows, further out in future we are trying to predict, models are becoming less and less efficient, and that makes sense. First, it is harder to predict for future, and the further out in future we are trying to predict, more errors we will have. Second, we are working with the technical analysis indicators. These indicators are more commonly known for doing short-term predictions, not long term. That is why, as the prediction period increase, the error rates of models increases as well.

Table 5-4: MAPE values for top 10 models averaged for each prediction period

Row	Prediction Period	TEA	svr_rbf_fs4	svr_linear_fs4	svr_rbf_fs5	svr_rbf_fs3	svr_linear_fs5	svr_linear_fs2	svr_rbf_fs2	mlp_fs4	bayesian_fs4
0	2	1.17	1.32	1.21	1.34	1.55	1.29	1.37	1.48	1.30	1.20
1	3	1.54	1.66	1.56	1.70	1.85	1.66	1.72	1.82	1.65	1.57
2	4	1.84	1.93	1.85	1.99	2.12	1.96	2.01	2.11	1.96	1.86
3	5	2.09	2.19	2.10	2.22	2.35	2.20	2.25	2.33	2.16	2.11
4	10	3.08	3.11	3.10	3.16	3.30	3.19	3.21	3.26	3.22	3.14
5	15	3.87	3.86	3.91	3.91	4.10	3.99	4.02	4.03	3.97	4.00
6	21	4.54	4.58	4.66	4.61	4.86	4.74	4.79	4.76	4.78	4.80
7	31	5.54	5.72	5.79	5.85	6.16	5.90	5.90	5.98	6.02	6.05
8	42	6.22	6.65	6.60	6.84	7.24	6.79	6.74	6.83	7.30	6.96
9	52	6.71	7.49	7.26	7.62	8.18	7.45	7.39	7.65	8.13	7.80
10	63	7.33	8.44	8.01	8.49	9.07	8.17	8.20	8.70	9.34	8.70
11	126	8.99	11.93	10.99	12.49	11.20	11.17	11.21	11.56	12.73	12.02
12	189	9.93	12.54	13.67	13.18	11.97	13.63	13.82	13.96	12.69	14.08
13	252	9.57	13.24	14.92	12.76	12.31	14.62	14.92	13.93	13.44	15.05
0	2	1.17	1.32	1.21	1.34	1.55	1.29	1.37	1.48	1.30	1.20
1	3	1.54	1.66	1.56	1.70	1.85	1.66	1.72	1.82	1.65	1.57
2	4	1.84	1.93	1.85	1.99	2.12	1.96	2.01	2.11	1.96	1.86
3	5	2.09	2.19	2.10	2.22	2.35	2.20	2.25	2.33	2.16	2.11

4	10	3.08	3.11	3.10	3.16	3.30	3.19	3.21	3.26	3.22	3.14
5	15	3.87	3.86	3.91	3.91	4.10	3.99	4.02	4.03	3.97	4.00
6	21	4.54	4.58	4.66	4.61	4.86	4.74	4.79	4.76	4.78	4.80
7	31	5.54	5.72	5.79	5.85	6.16	5.90	5.90	5.98	6.02	6.05
8	42	6.22	6.65	6.60	6.84	7.24	6.79	6.74	6.83	7.30	6.96
9	52	6.71	7.49	7.26	7.62	8.18	7.45	7.39	7.65	8.13	7.80
10	63	7.33	8.44	8.01	8.49	9.07	8.17	8.20	8.70	9.34	8.70
11	126	8.99	11.93	10.99	12.49	11.20	11.17	11.21	11.56	12.73	12.02
12	189	9.93	12.54	13.67	13.18	11.97	13.63	13.82	13.96	12.69	14.08
13	252	9.57	13.24	14.92	12.76	12.31	14.62	14.92	13.93	13.44	15.05

Table 5-4 shows the averaged MAPE values based on prediction period. Smallest MAPE value for each row is highlighted in green. Now for each prediction period, we are performing repeated holdout method, and the results we will achieve by doing that would not be inconsistent and truly show which model is the best. Our approach outperforms base models in almost every single prediction period.

Table 5-5: Overall Results for each model. Sorted in ascending order

Model (1)	MAPE	Model (2)	MAPE
TEA	5.17	svr_linear_fs3	6.60
svr_rbf_fs4	6.05	svr_linear_fs1	6.70
svr_linear_fs4	6.12	mlp_fs1	6.72
svr_rbf_fs5	6.15	sgd_fs5	6.77
svr_rbf_fs3	6.16	sgd_fs2	6.88
svr_linear_fs5	6.20	sgd_fs3	6.94
svr_linear_fs2	6.25	bayesian_fs3	7.00
svr_rbf_fs2	6.31	bayesian_fs1	7.09
mlp_fs4	6.33	dt_fs5	9.18
bayesian_fs4	6.38	dt_fs4	9.36
mlp_fs5	6.40	dt_fs2	9.49
sgd_fs4	6.44	dt_fs1	9.81

mlp_fs3	6.45	dt_fs3	9.85
mlp_fs2	6.47	svr_poly_fs1	25.25
bayesian_fs5	6.53	svr_poly_fs4	25.87
svr_rbf_fs1	6.54	svr_poly_fs3	25.96
sgd_fs1	6.59	svr_poly_fs2	29.78
bayesian_fs2	6.60	svr_poly_fs5	30.89

Table 5-5 shows overall results of each model throughout the entire evaluation process, sorted in ascending order based on MAPE value. They have been averaged on holdout ids and prediction periods. TEA had the lowest MAPE and as such is the best model in the overall results as well. TEA has outperformed other base models consistently in various scenarios. Variations of SVR using polynomial kernel did the worse by far. This is because, we did not consider optimization process for each model because it would have been impossible to do for the evaluation process. Because of that, SVR with polynomial kernel by default had the worst default configuration resulting in bad results. SVR models seems to be most popular ones when it comes to performance. Even considering that, our system uses variety of models, not just variations of SVR as shown in 5.5.

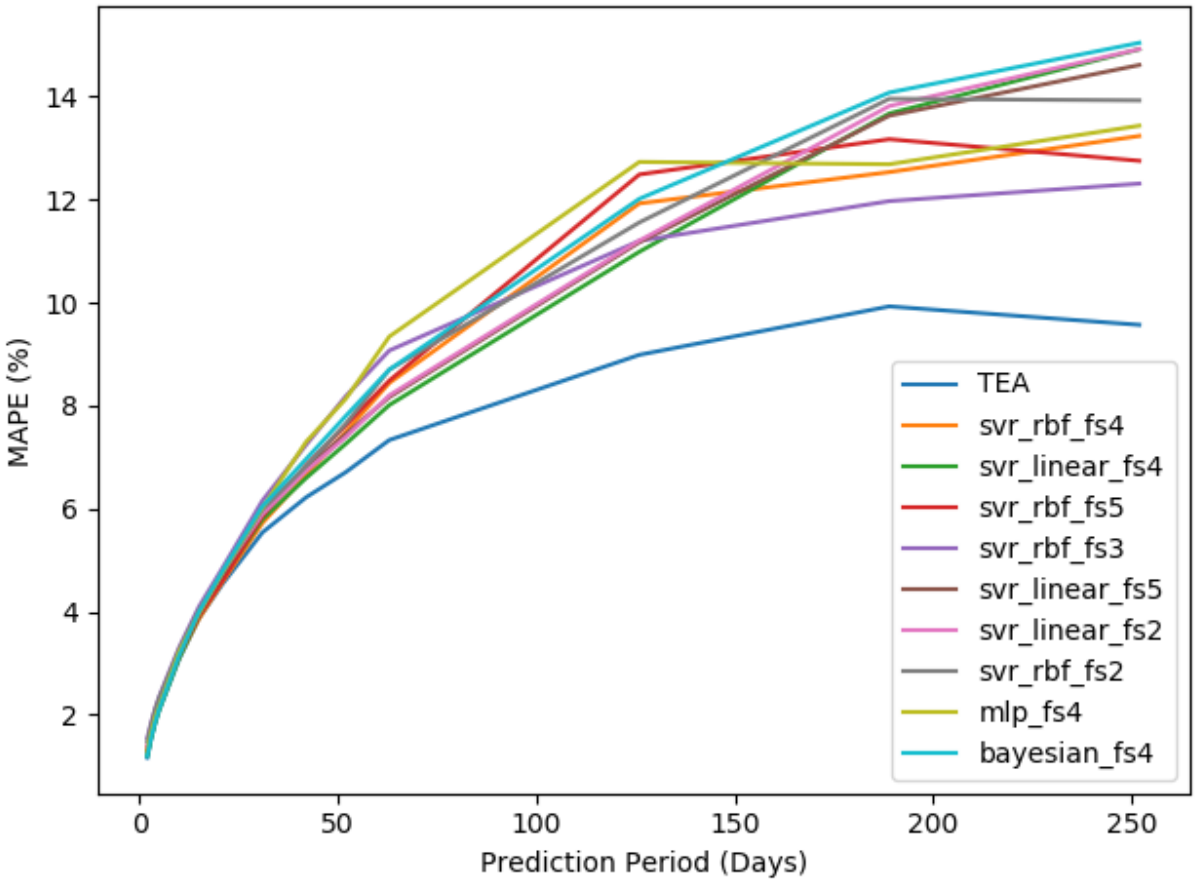

Figure 5-1: Model Errors vs Prediction Period

Figure 5-1 shows that not only the proposed approach outperforms other models, as the errors rates are increases because of predicting further out in future, the performance difference between the proposed approach and other base model increases as well, in support for proposed approach. Because of the tailored approach, our ensemble system picks models based on how they perform for different prediction periods as well. If a feature set is better suited for long-term predictions, then our model will tend to pick models trained on that feature set when it comes to predicting for long-term. We can clearly see that the performance of proposed approach decreases significantly slower than the base models when we increase the prediction period, but there is lot of overlap in results during short-term predictions.

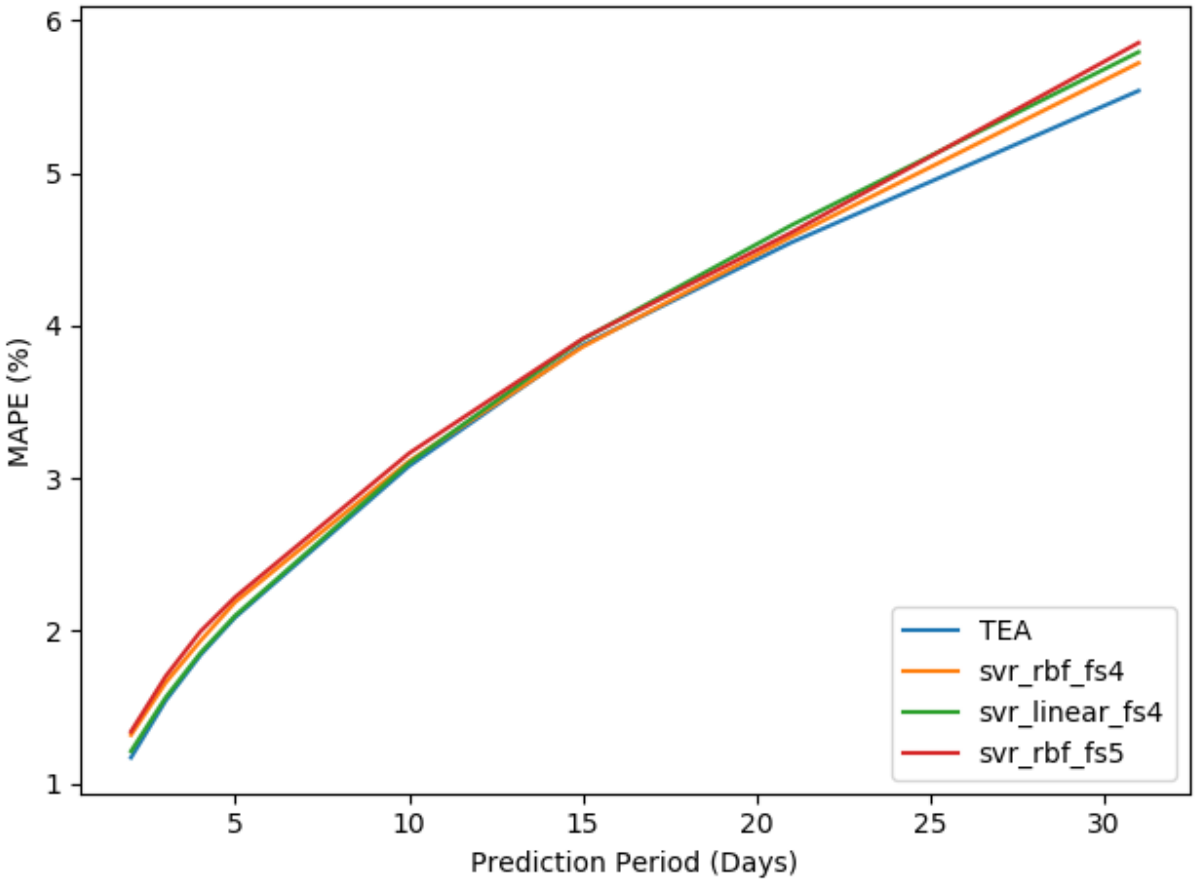

Figure 5-2: Model Errors vs. Prediction Periods for short-term

Figure 5-2 shows the results for short-term predictions. The error values are very close to each other for most of the models. Our model does outperform base models in almost every case except one, but the difference is very small. There could be a few reasons for this. One is that the error values are already quite small for short-term predictions, and improvements for short-term predictions will be relatively small as well. It could also be the case that to get considerable improvement, we might have to collect more data from various sources rather than historical data only (further discussed in 6.2).

One of the other reason could also be that during short-term predictions, most models have similar MAPE values, but during long-term, the error values varies based on different models.

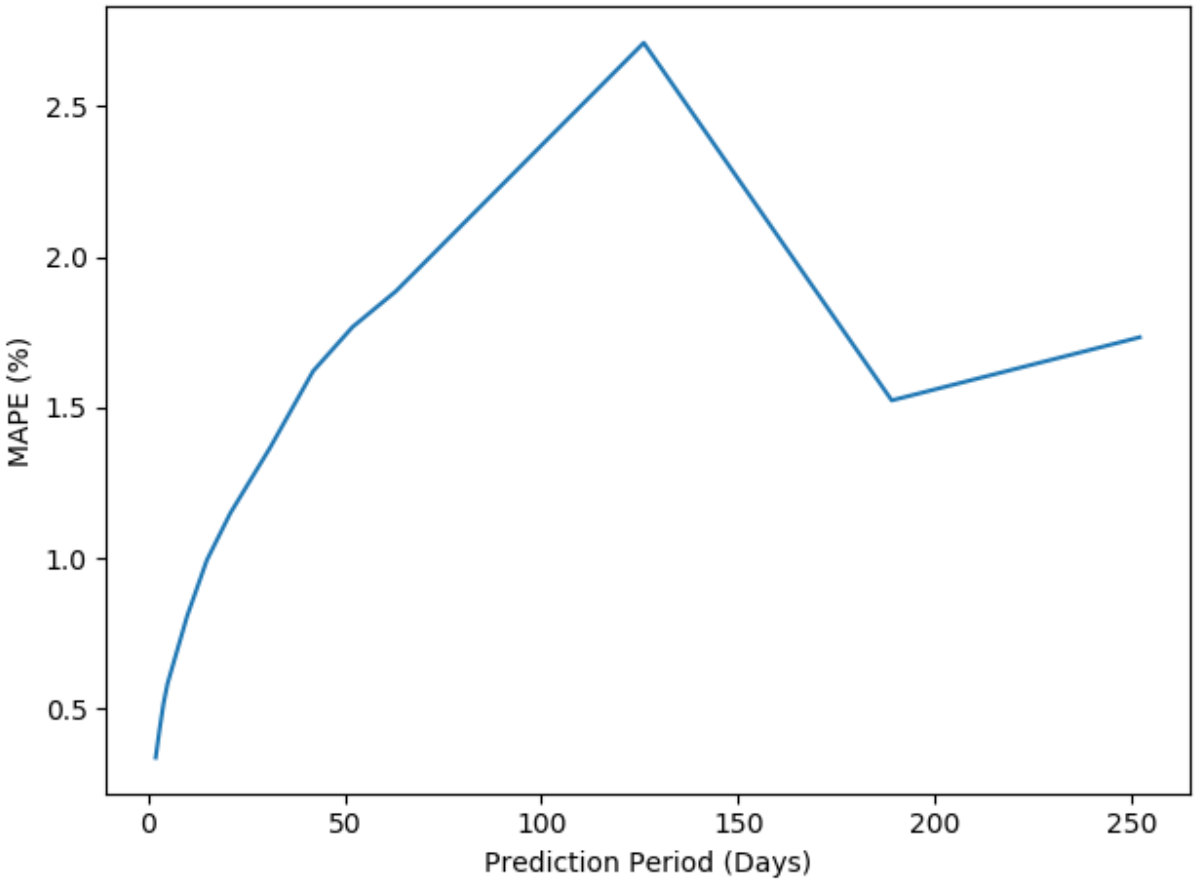

Figure 5-3: Standard Deviation vs. MAPE of base models

Figure 5-3 shows the standard deviation of error values at each prediction period. As the number of days increase, the standard deviation between error values is also increasing as well. This shows that as the number of days in future we are predicting for is increasing, the error values for different models varies. In short-term, since the standard deviation is so low, even if our ensemble system improves the performance, it would be relatively low as well. But for long-term, models are not showing similar performance, this provides our ensemble system a better opportunity to combine the results of multiple models to produce better results. We see a dip in standard deviation at 189 days (9 months real time) mark. This shows that models have more similarity in their performance at prediction period 189 than at 126.

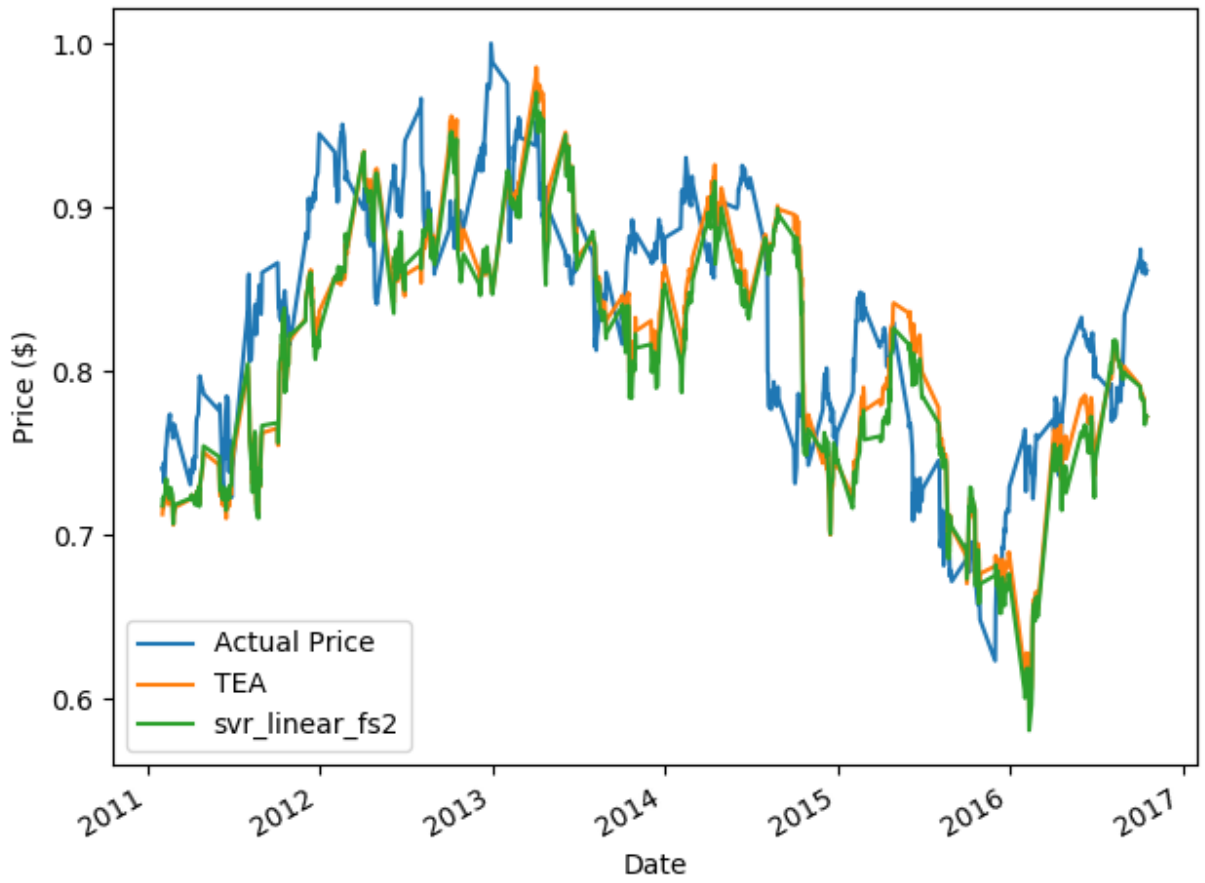

Figure 5-4: Predicted Prices for IBM (Prediction period = 52)

Figure 5-4 shows the actual prices for IBM stock, predicted prices by TEA and best base model, SVR trained with linear kernel on feature set 2. In this case, testing dataset ranged from start of 2011 to end of 2016. As we can see that, TEA follows the actual price more closely than the variation of SVR model. MAPE value for TEA for IBM stock is 5.81% calculated for this testing dataset. MAPE for variation of SVR model is 6.00%. This shows that TEA performs better than the best base model of the system.

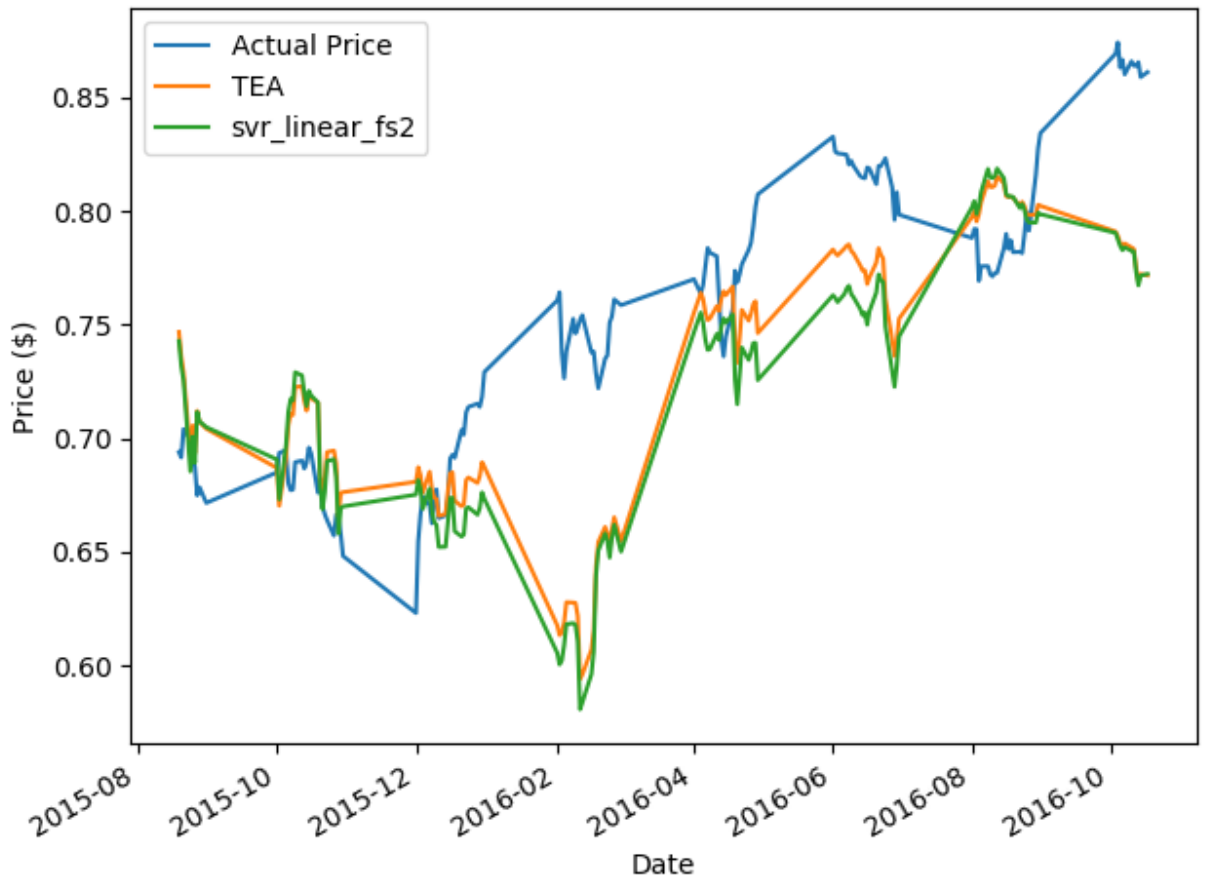

Figure 5-5: Predicted Prices for IBM stock in detail (Prediction Period = 52)

Figure 5-5 shows the prediction prices for IBM for a subset of testing dataset to better show the comparison between the proposed approach and best base model in the system. In this graph, we can see that our system follows the actual price more closely than the variation for SVR. Since ours is a tailored approach, that provides a good opportunity to our system to truly pick best set of model and feature set combination, and this can be seen throughout our results.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This thesis focuses on the difficult challenge of stock price prediction. Although there has been much research done on this topic, opinions are conflicted. Some research suggests that the price patterns can be predicted, while others suggest that it cannot. In this research, we assume that stock price can be predicted, and performance can be improved by combining multiple models in an ensemble system.

There are many models and features that can be used to tackle this challenge. Models such as, SVR, DT, and MLP have their own advantages and disadvantages. Some of them focus on linear problems, while others on non-linear. The same applies when it comes to developing feature sets. Some feature sets are better suited for certain stocks or prediction period.

In this research, we proposed a Tailored Ensemble Approach (TEA). This is an ensemble approach based on stock and prediction period and considers only the best models to predict their price. We had many combinations of models and feature sets that we trained for our system. Based on the performance of these combinations, given stock and prediction period, TEA chooses only the most accurate combinations for predicting prices of each stock and the given prediction period. To decide how many models TEA should choose for predicting prices, we experimented to obtain the optimal threshold value. Our experiment showed that a threshold value of 1 was the optimal option for our system. This meant that for a given stock, if the most accurate base model had the error rate of 4%, with a threshold value set at 1, TEA will consider all base models that have error rate at most 1% higher than the most accurate one (i.e. errors ranging from 4% to 5%).

We performed extensive testing to evaluate our model. Our dataset contained historical data for the stocks in S&P 100. We used the holdout method for evaluation, where we trained and tested our model on training, validation, and testing set. We repeated the holdout method by using various combinations of training, validation, and testing sets. We also tested on multiple prediction periods, ranging from short-term to long-term predictions. Mean Absolute Percent

Error (MAPE) was calculated to compare the performance of base models and the proposed TEA. TEA outperformed almost every model in the system when considering each unique prediction period and holdout iteration. Looking at the overall results, models have MAPE value ranging from 5.17 to 30.89. Our model is most accurate at a 5.17 MAPE value, while the second most accurate model is at 6.05 MAPE. This is a very significant, especially when it comes to stock price prediction.

Our research showed that the same model and feature set does not work for every stock and prediction period. Models and feature sets that have high accuracy for one stock in the short-term period might not provide similar performance during long-term predictions for the same stock. Also, certain models and feature sets might be most accurate for some stocks but might not even be one of the top performing for other stocks. The proposed Tailored Ensemble Approach (TEA) solves this problem by tailoring combinations of models and feature sets that are most accurate for each stock and prediction period.

6.2 Future Work

There are several ways to extend this research to further improve the system.

First, we can collect data from various sources. We have worked with technical indicators (i.e. historical data) in this study. Technical indicators are generally used for short-term predictions, that is why we see high prediction accuracy during short-term periods. We can also consider collecting data for fundamental analysis. This would include fundamental indicators, e.g. debt ratio, profit, unemployment rate, cash flow, etc. Including fundamental indicators could further improve the system during long-term predictions and even might prove useful for short-term predictions. We can also consider other sources of information, such as sentiment analysis from tweets. Since our ensemble approach picks set of models based on their performance for each stock, some stock might be better predicted using multiple sources of information.

Second, we can consider clustering similar stocks together and designate combinations of models and feature sets that have the higher accuracy for each cluster. This clustering process could be based on various factors, such as correlation, stock sector, etc. This would also let us optimize each combination of model and feature set based on the cluster. We were not able to

perform optimization in current study because of the sheer number of models and feature sets, but with clustering approach, number of combinations we need to train will drastically decrease and optimization process will be feasible. Given the decrease in number of combinations, we can apply this approach to much bigger datasets, such as NASDAQ market.

Lastly, to apply this approach for investing in stocks, we recommend making a few adjustments. In this research, we optimized the threshold value only for one prediction period, because of time constraint. When using this approach for investing in stocks, it would be better to optimize the threshold value based on the prediction period. Also, it would be worth considering incremental learning to update the models with the latest data. Rather than using the models trained on data many years ago, it would be useful to keep updating them with the most current data to predict future prices.

Chapter 7 Bibliography

Agirre-Basurko, E., Ibarra-Berastegi, G., & Madariaga, I. (2006). Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. *Environmental Modelling and Software, 21*(4), 430-446.

Altay, E., & Satman, M. (2005). Stock market forecasting: artificial neural network and linear regression comparison in an emerging market. *Journal of Financial Management & Analysis, 18*(2), 18.

Anon. (2018). *Magnifying glass with white stick figure*. Retrieved 06 01, 2018, from https://threequartersandcounting.files.wordpress.com/2016/09/magnifying-glass-with-white-stick-figure-bending-over1.jpg

Asur, S., & Huberman, B. (2010). Predicting the future with social media. *In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 01*, 492-499.

Basak, D., Pal, S., & Chandra, D. (2007). Support Vector Regression. *Neural Information Processing-Letters and Reviews, 11*(10), 203-224.

Berkshire Hathaway Inc. (2018). *BERKSHIRE HATHAWAY INC.* Retrieved 05 01, 2018, from http://www.berkshirehathaway.com/

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter Mood as a Stock Market Predictor. *Journal of Computational Science, 2*(1), 1-8.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT*, 177-186.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Chi-Jie, L., Lee, T.-S., & Chiu, C.-C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems, 47*(2), 115-125.

Connors, L., & Alvarez, C. (2009). *Short term trading strategies that work.* [Los Angeles, Calif.]: Tradingmarkets Publishing Group.

Cordón, O. (2011). A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *International Journal of Approximate Reasoning, 52*(6), 894-913.

CSE. (2018). *CSE*. Retrieved from https://thecse.com/

Dechow, P., Hutton, A., Meulbroek, L., & Sloan, R. (2001). Short-sellers, fundamental analysis, and stock returns. *Journal of Financial Economics, 61*(1), 77-106.

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011). Combining technical analysis with sentiment analysis for stock price prediction. *In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference*, 800-807.

Diller, A. (2003). Predicting direction of ISE national-100 index with back propagation trained neural network. *ournal of Istanbul Stock Exchange, 7*(25-26), 65-81.

Easley, D., O'Hara, M., & Saar, G. (2001). How Stock Splits Affect Trading: A Microstructure Approach. *The Journal of Financial and Quantitative Analysis, 36*(1), 25-51.

Esfahanipour, A., & Aghamiri, W. (2010). Adapted Neuro-Fuzzy Inference System on indirect approach TSK fuzzy rule base for stock market analysis. *Expert Systems with Applications, 37*(7), 4742-4748.

Farmer, R. (2012). The stock market crash of 2008 caused the Great Recession: Theory and evidence. *Journal of Economic Dynamics and Control, 36*(5), 693-707.

Farrar, D., & Glauber, R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92-107.

Financial Times. (2018). *Markets.ft.com*. Retrieved from https://markets.ft.com/data/indices/tearsheet/summary?s=FTSE:FSI

Graham, B. (2006). *The intelligent investor.* New York: Harper.

Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems, 23*(8), 800-808.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Kara, Y., Acar Boyacioglu, M., & Baykan, Ö. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications, 38*(5), 5311-5319.

Kim, E., Kim, W., & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems, 34*(2), 167-175.

Kittler, J., Hatef, M., Duin, R., & Matas, J. (1992). On combining classifiers. *IEEE trnsactions on pattern analysis, 20*(3), 226-239.

Kohavi, R. (1995). A study of cross-validation and bootsrap for accuracy estimation and model selection. *Ljcai, 14*(2), 1137-1145.

Lamoureux, C., & Poon, P. (1987). The Market Reaction to Stock Splits. *The Journal of Finance, 42*(5), 1347-1370.

Maddala, G., & Lahiri, K. (1992). *Introduction to econometrics* (2 ed.). New York: Macmillan.

Mangiameli, P., Chen, S., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research, 93*(2), 402-417.

MarketWatch. (2018). *Dow Jones Industrial Average*. Retrieved from https://www.marketwatch.com/investing/index/djia

Martin, E. (2018). *How Warren Buffett's winning investing strategy can be applied to any purchase you make*. Retrieved 05 01, 2018, from https://www.cnbc.com/2018/05/04/warren-buffett-invests-for-the-long-term.html

Microsoft Corp. (2018). *Microsoft Corp.* Retrieved 05 01, 2018, from https://www.marketwatch.com/investing/stock/msft

NASDAQ. (2018). *Nasdaq*. Retrieved from https://www.nasdaq.com/

NYSE. (2018). *Nyse.com*. Retrieved from https://www.nyse.com/index

NYSE. (2018). *NYSE: Choosing the Right Listing*. Retrieved 05 01, 2018, from https://www.nyse.com/get-started/international/choosing-the-right-listing

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications, 42*(1), 259-268.

Quandl. (2018). *Get Financial Data Directly into Python*. Retrieved 05 01, 2018, from https://www.quandl.com/tools/python

Quandl. (2018). *Quandl*. Retrieved 05 01, 2018, from https://www.quandl.com/

Raftery , A., Madigan, D., & Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Stastical Association, 92*(437), 179-191.

Scikit. (2018). *Decision Tree Regression*. Retrieved 06 01, 2018, from http://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html#sphx-glr-auto-examples-tree-plot-tree-regression-py

Scikit. (2018). *Neural network models (supervised)*. Retrieved 06 01, 2018, from http://scikit-learn.org/stable/modules/neural_networks_supervised.html

Scikit. (2018). *Stochastic Gradient Descent*. Retrieved 06 01, 2018, from http://scikit-learn.org/stable/modules/sgd.html

Scikit. (2018). *Support Vector Regression using linear and non-linear kernels*. Retrieved 06 01, 2018, from http://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html#sphx-glr-auto-examples-svm-plot-svm-regression-py

Staff, I. (2018). *Standard & Poor's 500 Index - S&P 500*. Retrieved 05 01, 2018, from https://www.investopedia.com/terms/s/sp500.asp

Statista. (2018). *Projected U.S. inflation rate 2010-2023 | Statistic*. Retrieved 05 01, 2018, from https://www.statista.com/statistics/244983/projected-inflation-rate-in-the-united-states/

StockCharts. (2018). *Fundamental Analysis*. Retrieved 05 01, 2018, from http://stockcharts.com/school/doku.php?id=chart_school:overview:fundamental_analysis

Tesla Inc. (2018). *Tesla Inc.* Retrieved 05 01, 2018, from https://www.marketwatch.com/investing/stock/TSLA/historical?siteid=mktw&date=&x=0&y=0

Trading Technologies International, Inc. (2018). *Technical Indicator Definitions*. Retrieved 05 01, 2018, from https://www.tradingtechnologies.com/help/x-study/technical-indicator-definitions/list-of-technical-indicators/

Tsai, C.-F., Lin, Y.-C., David C., Y., & Chen, Y.-M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing, 11*(2), 2452-2459.

Tso, G., & Yau, K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy, 32*(9), 1761-1768.

TSX. (2018). *TSX*. Retrieved from https://www.tsx.com/

WIKI Prices. (2018). *WIKI Prices*. Retrieved 01 01, 2018, from https://www.quandl.com/databases/WIKIP/usage/quickstart/api

Wikipedia. (2018). *Berkshire Hathaway*. Retrieved 05 01, 2018, from https://en.wikipedia.org/wiki/Berkshire_Hathaway

Wikipedia. (2018). *S&P100*. Retrieved 06 01, 2018, from https://en.wikipedia.org/wiki/S%26P_100

Wikipedia. (2018). *Warren Buffett*. Retrieved 05 01, 2018, from https://en.wikipedia.org/wiki/Warren_Buffett

Wu, C.-H., Ho, J.-M., & Lee, D.-T. (2004). Travel-time prediction with support vector regression. *IEE transactions on intelligent transaction systems, 5*(4), 276-281.