Intelligent Financial Portfolio Construction: Machine Learning Based Optimization

Hasan Mohammed Sami
Student Id: 012192026

A Thesis

in

The Department

of

Computer Science and Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Computer Science and Engineering

United International University

Dhaka, Bangladesh

October, 2022

Approval Certificate

This thesis titled "**Intelligent Financial Portfolio Construction: Machine Learning Based Optimization"** submitted by **Hasan Mohammed Sami**, Student ID: **012192026**, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on October, 2022.

Board of Examiners

1.

______________________________ Supervisor

Dr. Mohammad Nurul Huda
Professor & Director – MSCSE Program
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

2.

______________________________ Head Examiner

Dr. Md. Saddam Hossain Mukta
Associate Professor & Undergraduate Coordinator
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

3.

______________________________ Examiner-I

Dr. A.K.M. Muzahidul Islam
Professor
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

4.

______________________________ Examiner-II

Dr. Hasan Sarwar
Professor
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

5.

______________________________ Ex-Officio

Dr. Salekul Islam
Professor & Head
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

Declaration

This is to certify that the work titled "**Intelligent Financial Portfolio Construction: Machine Learning Based Optimization**" is the outcome of the research carried out by me under the supervision of Dr. Mohammad Nurul Huda, Professor & Director – MSCSE Program, Department of Computer Science and Engineering, UIU.

Hasan Mohammed Sami
Student ID: 012192026
MSCSE Program
Department of Computer Science & Engineering
United International University
Dhaka, Bangladesh

In my capacity as supervisor of the candidate's thesis, I certify that the above statements are true to the best of my knowledge.

Dr. Mohammad Nurul Huda
Professor & Director – MSCSE Program
Department of Computer Science & Engineering
United International University
Dhaka, Bangladesh

Abstract

Investors have always been highly interested about stock price forecasting. It’s observed that lower middle-income people and middle-income people can contribute only 10-15% of their wages in investment. Machine Learning based stock price forecasting is proven to be most efficient for price prediction according to the conventional research processes. The proposed research is conducted in order to derive a quality stock price forecasting technique for lower middle income or middle-income people so that their financial distress could be relieved.

Here, the research introduces chi-square test for finding the differences between observed and predicted prices. For the price prediction the machine learning (ML) tools such as SVR, LogR, XGBOOST, DTR, RFR, and LSTM are introduced, where input features for the ML are obtained from principal component analysis (PCA) and statistical averaging method. Statistical averaging method calculates a new feature from the stock price features open, low, high, adj close, and close, and finally obtains a new feature vector for a ML algorithm combining the stock price features and the new feature. Moreover, Portfolio is constructed observing the higher trend of predicted prices for the different stocks of an investor to reduce the risk of investment.

From the experiments it is observed that the proposed average feature-based method using the chi-square test (confidence 10%) achieved a feature dependency score of 26%, whereas PCA-based features did not achieve minimum benchmark of 10%. Besides, LSTM is found to be the best forecasting method and provides the highest accuracy of predicted prices which are 90.11% and 88.15% for the proposed feature set using the statistical averaging method and conventional price features (open, low, high, adj close, and close), respectively. Moreover, the portfolio created based on statistical average price feature provides a return on investment of 23.32% and reduces the risk by a Sharpe Ratio of 50.32 (standard value should be greater than 1.0).

Acknowledgement

First and foremost, I would like to be cordially oblige to Allah the Almighty, for giving me immense patience and blessings in order to complete and continue my research successfully.

I would primarily express my deepest and sincere gratitude to my thesis Supervisor, Professor Dr. Mohammed Nurul Huda, Program Director, MSCSE, UIU for his continuous and coordinated support in my research work. His dynamic suggestions and varied areas of knowledge coordination have deeply impressed me to follow along my dream to pursue my education towards thesis work. Under his guidance I have been able to carry out the duties for research under true appropriate guidance. I am extremely grateful to him for his continuous support and guidance that has made me successful in this endeavor.

I would also like share my cordial gratitude to Md Shah Jalal, Lecturer of CSE, Sonargaon University for his immense support and guidance. Throughout this research I have received active and confident support from him that has helped me to achieve the study milestones. His presentation and thesis paper organization methods have suggested me to follow along my dream of presenting the research ideas formally and adequately.

Finally, I would express my genuine appreciation and heartfelt obligation to those who have supported me in completion of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Currently, more than half of the people can contribute 10-15% of their wages into stock for additional income [1]. In order to gain secondary profit, they have faced various problems for instances, correct investment, stock choices & backgrounds, price forecasting processes and optimized portfolio performance. Correct investment refers to a situation when an individual would reach their intended earnings goal [2]. Furthermore, stock choice is making the best selection of stocks based on financial conditions [3]. Stock price forecasting is the process of predicting the prices of stocks for a specific period [4]. Consequently, portfolio management is the method by which we will optimize the performances of stocks established by allocation of resources [5].

Correct investment faces the problem in reference to making the best usage of the available information related to stocks. Among the investment choices stock pricing is the most popular which faces the primary problem regarding their forecasted future prices. The selection of stock choice is dependent on financial ratios and the stocks volatility due to unavailability of correct information. Price forecasting faces the problem of high noise, non-stationery and non-linearity of price dataset and sample size of price data. Eventually, portfolio management faces the problem related to minimizing the risk of investment in relation to investment.

1.1 Motivation

Researcher in current era uses machine learning (ML) tools to predict stock prices and stock selection. But the ML tools need the effective features to get higher accuracy. For the stock market prediction, the effective price features for ML are derived from conventional price features such as *'Open'*, *'Low'*, *'High'*, *'Adj Close'*, and *'Close'*. Therefore, extracted price features based on statistical average method (SAM) and PCA-based method are combined with conventional price features to get hybrid features that will provide more accuracy. The higher accuracy will indicate whether an investment is safe or not. Besides, it will inform the investors whether the investment process is effective with less risk and produces the highest return. Inherently, the researchers are motivated to provide the quality return for the people of lower middle class or middle class who contribute small investment in stock market. Moreover, a return on investment (ROI) using ML-based method that is greater than the general market return and highest savings account return rate is needed to be predicted to get a risk-free investment. Portfolio management recommends the best investment plan for each individual based on their income, budget, age, and risk tolerance. Portfolio management reduces the risks associated with investing while also increasing the likelihood of profit.

1.2 Objectives

The paper addresses various issues related to the problems of stock price forecasting, selection of quality stocks and optimized portfolio performance. Here, the following issues will be focused:

- **i)** Firstly, we investigated the various ML based conventional research works for forecasting.
- **ii)** Secondly, we apply chi-square test for determining the feature dependency.
- **iii)** Thirdly, we apply different ML-based methods based on our proposed statistical average price features and PCA-based features, and compare their accuracies.
- **iv)** Finally, portfolio optimization has been attained using the trend of predicted prices.

1.3 Organization

The thesis report is organized as follows:

The 2nd chapter have discussed about the related works and scholarly evidence gathered for conducting the research.

The 3^{rd} chapter has suggested the adopted proposed methods to conduct the research.

The 4^{th} chapter has documented the obtained results and analysis performed in this research.

Finally, the 5^{th} chapter examined the obtained results in order to prove the research target through conclusion, limitations and future works.

Chapter 2
Literature Review

Investment into stocks is always associated with risk and return. Selection of stocks is established by financial ratios. Nowadays machine learning remains an important price prediction technique for stock price forecasting. Therefore, our research should have necessary evidences to support ML based forecasting processes. As a consequence, we are discussing various experimental evidences regarding our research in literature review.

2.1 Literature Review

Stock price features are important prediction variables used by ML processes for price forecasting [6]. It's exhibited by Cakra & Trisedya that only one price feature is needed to appropriately forecast stock prices [7]. Furthermore, Li et al showed that stocks behave differently in same period when they have different financial ratios [8]. Similarly, Mehtab et al and Fisher et al showed that when multiple price features are used for prediction, forecasting accuracy was higher [9], [10]. This thesis involves supervised learning-based regression algorithms to forecast stock prices [1], [2]. Furthermore, Hammoudeh et al showed that using predicted prices effective portfolio could be built that could be used successfully implemented for secondary earning [11]. Additionally, Zhu et al suggested how the portfolio performance accuracy had been evaluated for different portfolios [12]. Yu et al and Chen MY showed how tree-based algorithms and classification-based regression also successfully predicted stock prices [13], [14]. Despite of so many effective ML based stock price forecasting we didn't find any of these researches tried to forecast stock prices using additional or derived price features. Although Shahi et al suggested that the stock price features are very close in value to each other, hence the requirement for additional price feature is very limited [15]. None of the above-mentioned researches also didn't accompany feature dependency through chi square testing.

As our research suggests future price predictions hence supportive issues mentioned by Sumaiya & Aswani have shown the benefit of chi square test for relevant feature selection in ML based predictions [16]. As Cai et al and Hussein & Ozyurt have shown the importance of feature extraction for stock price prediction, hence there have been various

ML based methods used for feature selection [17], [18]. Firstly, Zhong & Enke showed the computational benefit of forecasting using only one feature rather than many features by the usage of principal component analysis (PCA) [19]. Long et al showed that when all the price features are averaged then a new feature comes into existence which has lower standard deviation and better forecasting dependency [20]. Salih & Adnan has shown that PCA is capable of reducing dimensionalities for major big databases [21]. According to Salih & Adnan PCA helps make better suggestions for predictions. Our target is to make better price prediction in this research. As the number of features are only 5 and the prices are very close to each other, our research would convert 5:1 price feature using PCA which will help in the prediction process. Hence for this research we considered PCA based feature and average price feature as two additional features for making price predictions using ML methods. But as our research motivation is inclined with good return on investment hence selection of assets is also important. Consequently, in this research we also emphasized regarding theoretical financial backgrounds to make proper asset selection.

To summarize we need to make choices of stocks based on earnings per share (EPS) and current ratio which are the major financial ratios used for stock selection according to Fin et al [21]. EPS refers to earnings of each stocks for each company. According to Lin et al if overall EPS is positive, it means that the company has a positive cash flow of money. Similarly, a greater than 1 current ratio stands for assets being more than company debt. As Amenc & Goltz showed that stocks have a volatility indicator named Beta and Oh et al showed that in case of good prediction algorithm a stock price could be forecasted despite of its high degree of volatility [22], [23]. According to Oh et al, beta is the relationship between a stocks price movement and its relation with major stock index movement. Because of that, we need to emphasize the quality of prediction methods which are initiated using ML algorithms.

Stock price forecasting is dependent on ML methods which obtains a set of rules in order to make quality predictions according to Vadlamudi [24]. Moreover, Asad suggests that supervised learning executes quality stock price forecasting [25]. Due to stock price dataset being noisy, Huang & Liu showed that its best to maintain a selective accuracy range for each predicted price dataset [27]. Both Huang & Liu and Batra & Daudpota have suggested that it’s better to consider an error margin of at least 30% for the predicted price dataset to be truly considered as positive [27], [26]. Hasnain et al showed that each value defined by confusion matrix plays an effective role for predicted values [28]. Gray et al showed that in case of noisy dataset how a threshold range can help estimate better accuracies, precision, recall and F1 scores [29]. The quality parameter of predicted values is judged by the quality of classification report score assigned by the confusion matrix according to Deng et al [30]. For forecasted values mean absolute error (MAE) performs as a better predictor rather than root mean squared error (RMSE) as suggested by Parades et al [31]. Similarly, Willmott & Matsuura have shown that in case of climate research MAE plays a more accurate role in comparison to RMSE [32]. Siew & Nordin showed that the regression techniques will have adequate difference in price dataset to make MAE an important error evaluator [33]. After evaluating the prediction accuracy using MAE, the predicted prices will be used for portfolio optimization.

In case of portfolio-based optimization various parameters like Sharpe Ratio, efficient frontier, and Python based optimization library like PyPortfolioOpt has been used in order to make investment process beneficial. Farinelli et al and Bailey et al also suggested that both the Sharpe Ratio and efficient frontier are important for portfolio profit generation [34], [35]. Mansini et al and Sadati et al have shown that linear programming-based optimization provides best portfolio performance [36], [37]. The suggested PyPortfolioOpt library defined by Yudin brings the absolute best result of portfolio performance for minimum level of risk using Linear Programming based optimization [38]. Guerard et al suggested that for Portfolio based earning, return on investment with minimum risk and maximum return affects most positively [39]. The research originality lies with obtaining positive ROI using predicted prices obtained using proposed price dataset. The ML methods quality thus is needed for effective evaluation.

2.2 Research Findings

ML methods are widely well known for stock price forecasting. Popular ML method xgboost can predict stock prices with high degree of accuracy (71.23%) using a single price feature obtained from ‘Adj Close’ price feature [40]. But using the real price feature of ‘Close’, xgboost can predict stock prices with better degree of accuracy (79.12%) [41]. Eventually xgboost performs best when any random features are considered for stock price forecasting (87.11%) [42]. Similarly for logistic regression when multiple price features are considered for forecasting, the prediction accuracy became better instead of just one price feature [43], [44]. In case of decision tree regression only one price feature was considered for stock price forecasting [45], [46]. But in case of random forest-based regression when multiple price features are used for forecasting, accuracy became better [47], [48]. Moreover, the well-known support vector machine-based regression that learns from classification has used only one feature for stock price forecasting [49], [50], [51]. Finally, we observed that in case of LSTM which is the most popular ones among all other stock price forecasting methods, gets higher accuracy with multiple price features [52], [53], [54]. Our portfolio-based investment requires to make returns greater than S&P 500 based returns and best savings account rates [55], [56]. Observing the Table 1, we emphasize on issues considering the opportunity gaps observed through these researches.

a) Statistically suitable price features were not obtained using ML methods.

b) Forecasted prices were not used for portfolio optimization in most researches.

c) Feature importance were not explored in terms of forecasting dependency.

Table 1: Research Findings for Stock Price Forecasting

Authors	**ML Methods**	**Name of Applied Features**	**No. of Features**	**Dataset Size**	**Accuracy**	**Time Period for Research**
Y.Yang (2021) [34]	XGBoost	Special Feature from 'Adj Close	1	7500	71.23%	Start: 01/01/2014 End: 31/06/2018
A.B Gumelar et al (2020) [35]	XGBoost	Close	1	16665	79.12%	Start: 01/01/2012 End: 31/12/2015
Dey et al (2016) [36]	XGBoost	Random Feature among 5 Features	1	10220	87.11%	Start: 01/09/2008 End: 11/08/2013
Gong & Sun (2009) [37]	Logistic Regression	Open, High, Low, Close, Adj Close	5	21375	83.01%	Start: 01/01/2005 End: 20/10/2007
Ali et al (2018) [38]	Logistic Regression	Close	1	100280	89.77%	Start: 01/01/2011 End: 01/01/2015
Hindrayani et al (2020) [39]	Decision Tree Regression	Adj Close	1	4600	74.32%	Start: 01/01/2016 End: 31/12/2020
Zhou et al (2019) [40]	Decision Tree Regression	Adj Close	1	11250	71.23%	Start: 01/04/2010 End: 06/29/2016
Vijh et al (2020) [41]	Random Forest Regressor	Open, High, Low, Close, Adj Close	5	10080	70.39%	Start: 04/05/2009 End: 04/05/2019
Polamuri et al (2019) [42]	Random Forest regressor	Adj Close	1	1260	72.32%	Start: 01/01/2013 End: 31/12/2018
Henrique et al (2018) [43]	Support Vector Regression	Close	1	33661	76.64%	Start: 01/03/2017 End: 26/05/2017
Kao et al (2013) [44]	Support Vector Regression	Adj Close	1	12300	71.12%	Start: 06/11/2007 End: 30/11/2011
Guo et al (2018) [45]	Support Vector Regression	Adj Close	1	9500	74.23%	Start: 01/01/2017 End: 03/31/2017
Moghar & Hamiche (2020) [46]	Long Short-Term Memory	Adj Close	1	2843	81.22%	Start: 08/19/2004 End: 12/19/2019
Wang et al (2018) [47]	Long Short-Term Memory	Open, High, Low, Close, Adj Close	5	7250	97.66%	Start: 01/01/2012 End: 06/30/2018
Selvin et al (2017) [48]	Long Short-Term Memory	Open, High, Low, Close, Adj Close	5	11200	94.69%	Start: 10/06/2014 End: 11/28/2014

After observing prediction accuracies obtained from various researches, we can identify that LSTM is supposedly the best ML method for price forecasting. Also, we observe that prediction accuracies become better when the stocks have low volatility and multiple price features are used. At last, this research has focused the explained issues which could be explored in order to obtain better price forecasting and consequently better portfolio optimization.

2.3 Contributions

Our research is focused to ensure qualified price prediction in order to provide convincing stock price forecasting for investors with smaller contribution margins. In this research we have made the following contributions in order to make this analysis scientifically reassuring.

- Extract two additional price features one for statistical average price method and another for PCA-based method.
- Compare the score of chi-square test to detect the feature dependency for the statistical average price method and PCA-based method
- Evaluate the prediction accuracy of ML-based methods for both volatile and nonvolatile stocks.
- Select the best performer in terms of prediction accuracy where LSTM provides the best using the statistical average price method and PCA-based method.
- Evaluate the portfolio performance using actual price model and proposed feature-based model.

Chapter 3

Proposed Methods

The proposed research process follows several step wise processes including stock selection, data preparation, forecasting and finally portfolio creation. The research evaluation is carried out after the proposed research is conducted according to the following norms. The research flowchart in Figure 3.0 explains the complete proposed research methodology adopted in this thesis.

The research follows the following steps in order to propose its final outcome:

a) Step 1: Selection of Stocks:
   i) The dataset is collected from investing.com and macrotrends.net regarding stocks eps, current ratio and beta values.
   ii) The Dataset contains intended values of 42 stocks.
   iii) The proposed method will select assets based on financial ratio and Beta Constraints are:
      1) EPS > 0
      2) Current Ratio > 1 && < 2
      3) Beta > 1.5 OR < 0.5 OR between (1.05 and 0.95)

b) Step 2: Dataset Preparation:
   i) There will be three different price datasets for this research process.
      1) Conventional Feature based Price Dataset – Contains 5 Price Features (‘Open’, ‘High’, ‘Low’, ‘Adj Close’, ‘Close’).
      2) Additional Feature based Price Dataset – Contains 6 Price Features (‘Open’, ‘High’, ‘Low’, ‘Adj Close’, ‘Close’, ‘*PCA Complied 1 Dimensional Price Feature*’). (#PCA generally reduces the dimensionality of multiple features and here the all-price features are reduced to 1 dimensional price feature)
      3) Average Feature based Price Dataset – Contains 6 Price Features (‘Open’, ‘High’, ‘Low’, ‘Adj Close’, ‘Close, ‘*Average Feature*’). (#Average price is determined by averaging the prices of all features for one specific date and then adding it into price dataset as an additional price feature)
   ii) Chi Square Testing: For Feature Dependency each feature is tested in term of its significance. If any price feature exhibits a feature dependency of less than

10% that feature or price dataset should be removed from stock price forecasting.

c) Step 3: Stock Price Forecasting:

  i) This step requires the usage and application of Machine Learning models used by supervised learning techniques in order to predict stocks prices for a specific time period.

    1) Price Dataset Training: Selected price dataset which is used for financial forecasting in order to obtain ML rules.
    2) Price Dataset Testing: The predicted price values will be evaluated against the original price dataset in order to derive accuracy.

d) Step 4: Portfolio Optimization:

  i) This step requires the predicted prices as the background for portfolio formation. The following steps will be defined as constraints for portfolio formation.

    1) The Predicted price (i = n) has to be less than real price (i = 1). (i.e., i refers to the date of the stock).
    2) The best possible ML method in terms of accuracy will be providing highest price accuracy. The best predicted price will be used for portfolio creation.

  ii) The secondary step in Portfolio Optimization will be defining the following parameters applied using PyPortfolioOpt.

    1) Sharpe Ratio > 1.
    2) Stock Return is greater than market return.
    3) Stock has to be positive and integer.

  iii) Portfolio Using the three different predicted prices based on three different price datasets will be formed.

    1) Real Price Portfolio using the same stocks will be formed.
    2) Predicted Price Portfolio will be evaluated against the Real Price Portfolio in terms of accuracy measurement

       Real Price Portfolio Vs Predicted Price Portfolio will be evaluated.

e) Step 5: Research Target Evaluation

  This step will evaluate the predicted portfolio performance in terms of Sharpe Ratio value, return on investment and accuracy measurement between portfolio accuracies.

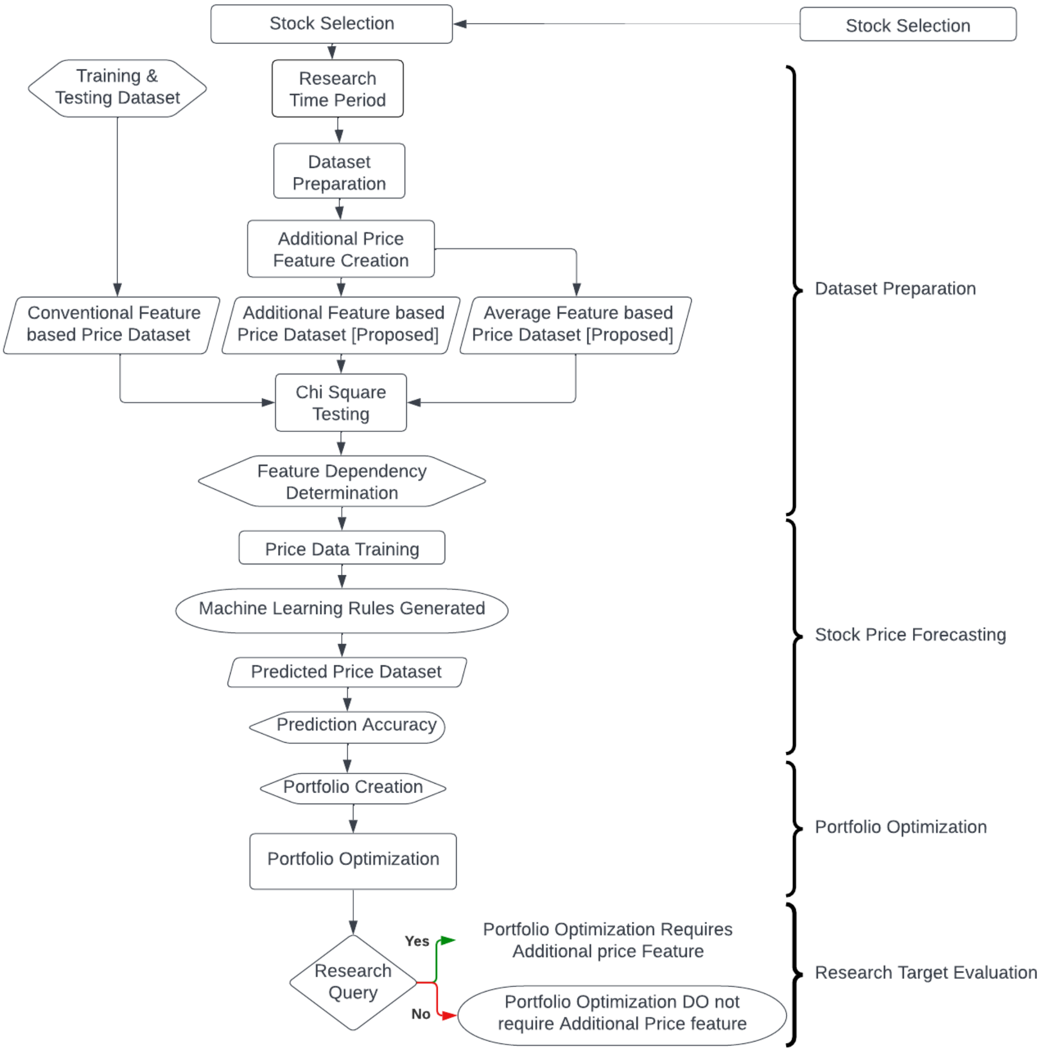

Fig 3.0: Proposed Research Methodology

3.1 Stock Selection

In this research as we will ultimately create a positively returning portfolio, which will fulfill the motivation of the research. This research process will help make better relationship between statistical studies with social benefit. The time period selected for this research will be from March mid till June last 2020 considering a highly volatile period due to COVID 19 pandemic [57]. Table 2 includes the financial ratio and beta values of the stocks which have good financial standing according to investing.com for our research. EPS is earnings per share, which suggests how much money a share earns for a company. Current ratio means the current assets : current liabilities. If current ratio is greater than 1 it means the company has more current asset than current liabilities. Beta is the relationship between a company’s reward and risk. Its shows the relationship between the covariance of asset and market with the variance of market. The explanation of Beta is provided below in Equation (1).

$$\beta = \frac{Covariance\ (Return\ of\ Stock\ i, Return\ of\ Market\ m)}{Variance\ (Return\ of\ Market\ m)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Beta is the relationship between covariance and variance. Covariance between return of stocks i and m. Variance of return of market, m.

Table 2: Stocks with Financial Ratio & Beta

Ticker	Beta	EPS	Current Ratio	Ticker	Beta	EPS	Current Ratio
ALRM	1.15	1.47	4.67	HD	1.03	11.56	1.36
JD	0.92	0.15	1.20	SNPS	1.09	4.27	1.19
LRCX	1.34	17.6	3.31	CELH	2.17	0.08	3.16
NTAP	1.33	3.07	1.65	NLS	1.74	1.1	1.86
NVDA	1.45	6.12	3.92	WYY	1.4	0.24	1.15
CRM	1.17	3.86	1.22	SHOP	1.6	1.62	17.87
SWKS	1.21	4.8	5.17	SPSC	0.89	1.15	4.79
SYNA	1.17	3.2	2.41	CTLT	1.5	1.58	2.56
TSM	0.91	2.23	1.75	CORT	1.03	0.89	10.77
MSFT	0.83	6.2	2.53	ALXN	1.34	4.31	3.81
FB	1.18	8.78	5.51	HZNP	1.14	3.61	3.49
TSLA	2.19	0.52	1.63	ACMR	0.89	0.67	2.62
ETSY	1.65	1.86	4.98	VNET	0.44	0.77	2.24
PYPL	1.12	2.65	1.38	KNSL	0.82	3.00	1.56
IPGP	1.48	1.97	10.09	BR	0.84	4.03	1.51
DE	0.97	8.69	1.89	OSTK	4.41	0.35	1.68
ROK	1.37	8.77	1.48	OIIM	0.8	0.13	4.99
ALB	1.6	3.59	1.38	LPTH	1.45	0.079	2.95
NKE	0.84	1.77	2.66	BRKS	1.97	0.88	3.08
CAT	0.96	6.03	1.54	PTC	1.33	1.12	1.22
ERIE	0.42	5.55	1.31	IRBT	1.47	5.39	3.05

According to the financial ratios of the provided stocks and on the basis of their given volatilities as beta, the stocks in Table 2 will be chosen on the basis of the criterion in Figure 3.1.

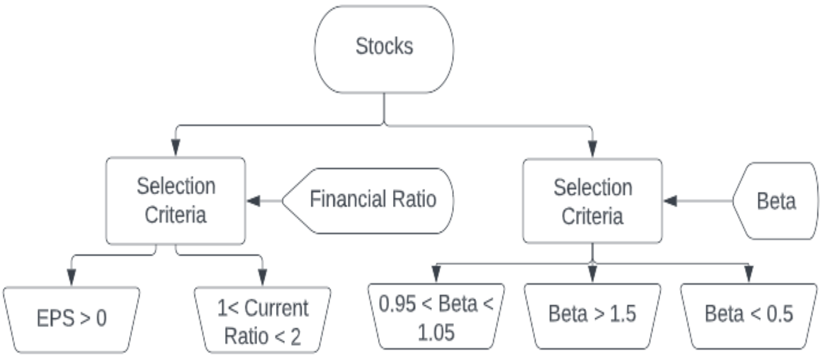

Fig 3.1: Stock Selection Flowchart

3.2 Dataset Collection & Preparation

Among the sample stocks nearly 10 different stocks are selected. Here 4 stocks with high beta, 4 stocks with beta equal or close to 1 and two stocks with low beta are selected. This makes 60% of the stocks considered for our stock to be volatile. In order to effectively train, each stock will have an overall training data size of 2783 days where from 1st January 2010 the training dataset is selected. The testing or forecasting period will be from 15th March 2020 till 30th June 2020. Among the ML based forecasting processes LSTM is found to have predicted price dataset with greater than 80% accuracy. The following Table 3 exhibits a price data sample for the effective research process. The stock financial ratios are collected from investopedia.com/articles/stocks/06/ratios.asp. The suggestion of best performing stocks is collected from investors.com/research/best-stocks-to-buy-now/

Table 3: Sample Price Dataset with Conventional Features

NASDAQ					
Date	**Open**	**High**	**Low**	**Close**	**Adj Close**
22-Jul-10	12,025.37	12,093.02	11,767.19	11,834.11	11,834.11
21-Jul-10	11,914.15	12,060.59	11,812.72	12,059.61	12,059.61
20-Jul-10	11,726.09	11,939.96	11,703.36	11,897.65	11,897.65
19-Jul-10	11,515.00	11,721.22	11,448.97	11,713.15	11,713.15
18-Jul-10	11,561.64	11,629.03	11,322.84	11,360.05	11,360.05
15-Jul-10	11,379.36	11,454.69	11,295.33	11,452.42	11,452.42
14-Jul-10	11,151.21	11,279.97	11,005.93	11,251.19	11,251.19

3.2.1 Feature Extraction

According to the research target and motivation, this thesis aims to evaluate the prediction accuracy by using quality price features for this research.

a) Extracting price dataset of stocks from yahoo.finance.com/stock/stock_ticker_symbol.
b) Extracting the conventional features from those stocks.
c) Creation of additional price feature using ML method and statistical techniques that we have proposed in this research
d) Adding the significant new feature for more accurate price prediction.

3.2.2.1 Conventional Approach

In order to make effective prediction we would primarily conduct the research based on available price features.

a) Applying the conventional price dataset using the 'Open', 'High', 'Low', 'Adj Close', 'Close' from January 1, 2010 or other first dates till March 15, 2020 as Training Period.
b) Apply the 6 supervised Learning method to learn regarding the data in order to make forecast from March 15, 2020 till June 30, 2020.

By following up to this step, we are complying to follow the research process carried out by the conventional researchers from the available dataset presented from yahoo/finance as raw dataset shown in Figure 3.2.

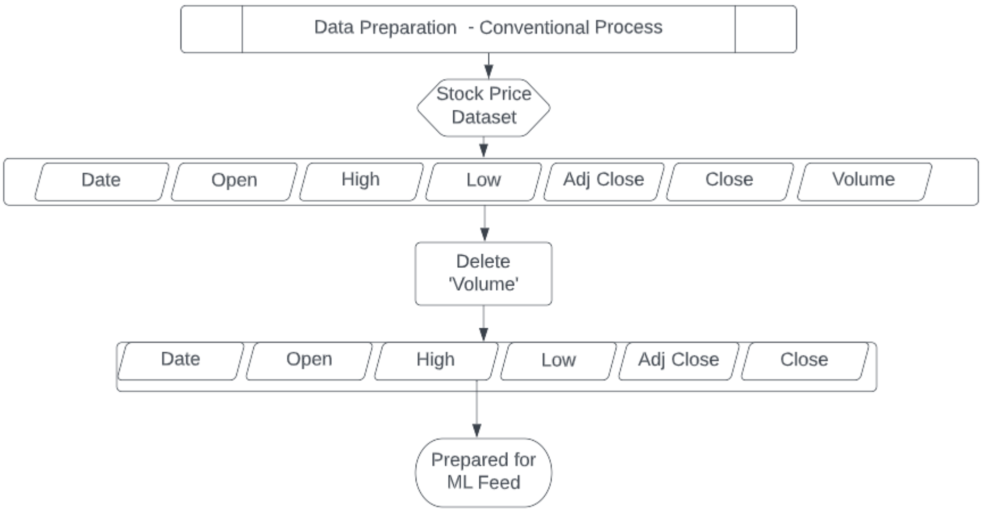

Fig 3.2: Data Preparation Conventional Method

The research originality lies with creation and addition of new features. Our originality exists within the parameter of forecasting accuracy generated using additional price features.

3.2.2.2 Proposed Approach for Features Extraction

In this research we have introduced both ML process and statistical methods for creating additional price features. We have introduced a new price sample considering the price dataset provided. The following method will help us introduce new features into the price dataset.

a) Reducing the dimensions of 5 different price features into 1 price feature and introducing it as an additional price.

b) Averaging prices of all 5 price features and considering the newly created average price as the additional price feature.

Considering the price dataset in the sample from Table 6, the following processes will be followed in order to create new additional price features.

3.2.2.2.1 PCA based Price Feature

The first additional price that will be created using the currently available features will be the PCA processed price feature. The overall thesis qualification depends on the research capability of prediction when an additional price feature is used for price prediction. Moreover, PCA generally is used to reduce the complicacy aligned with multiple features. Our thesis would thus observe if PCA creates a valuable price feature that will help in price prediction. The following Figure 3.3 will determine the process of data preparation using additional price feature defined by PCA.

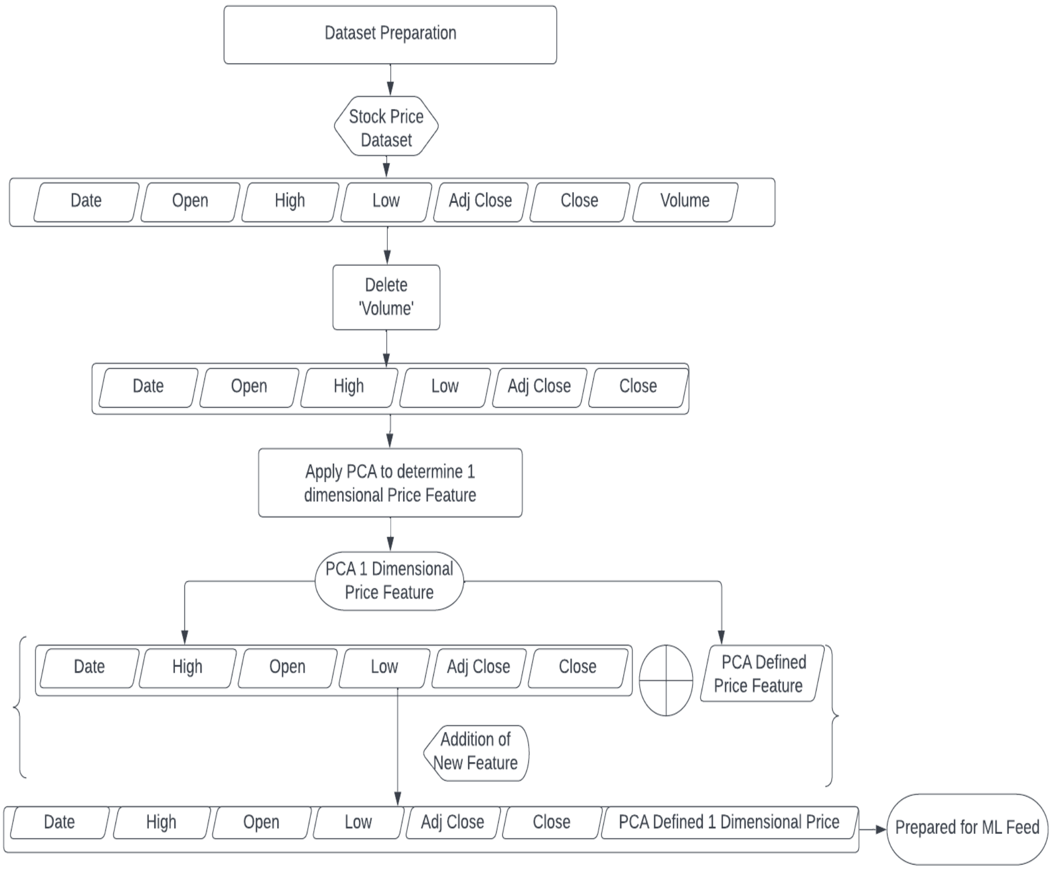

Fig 3.3: Additional Feature based Price Dataset

In this research we apply PCA in order to obtain a new price feature from the currently available price features.

Current Feature Vector:

Table 4: Conventional Features

Feature	**Feature Name**
F1	‘Open’
F2	‘High’
F3	‘Low’
F4	‘Close’
F5	‘Adj Close’

In our first experiments I have used <F1, F2, F3, F4, F5> feature vector for machine learning.

Then, I have derived a new feature, Fpca from <F1, F2, F3, F4, F5>. This new feature Fpca is incorporated in the the previous feature vector <F1, F2, F3, F4, F5> to get a new feature vector <F1, F2, F3, F4, F5, Fpca>.

This new feature vector <F1, F2, F3, F4, F5, Fpca> is used in machine learning for our second experiment.

Just to check the effect of the new feature this experiment is designed. But the addition of new feature did not improve the performance of the machine learning based system.

3.2.2.2.2 Statistical Average Price Feature

Similarly, the average price feature should be added as well in the dataset in order to create one additional comparison. In Figure 3.4, we exhibited the process of adding average feature by averaging all the prices of the conventional features of the given date.

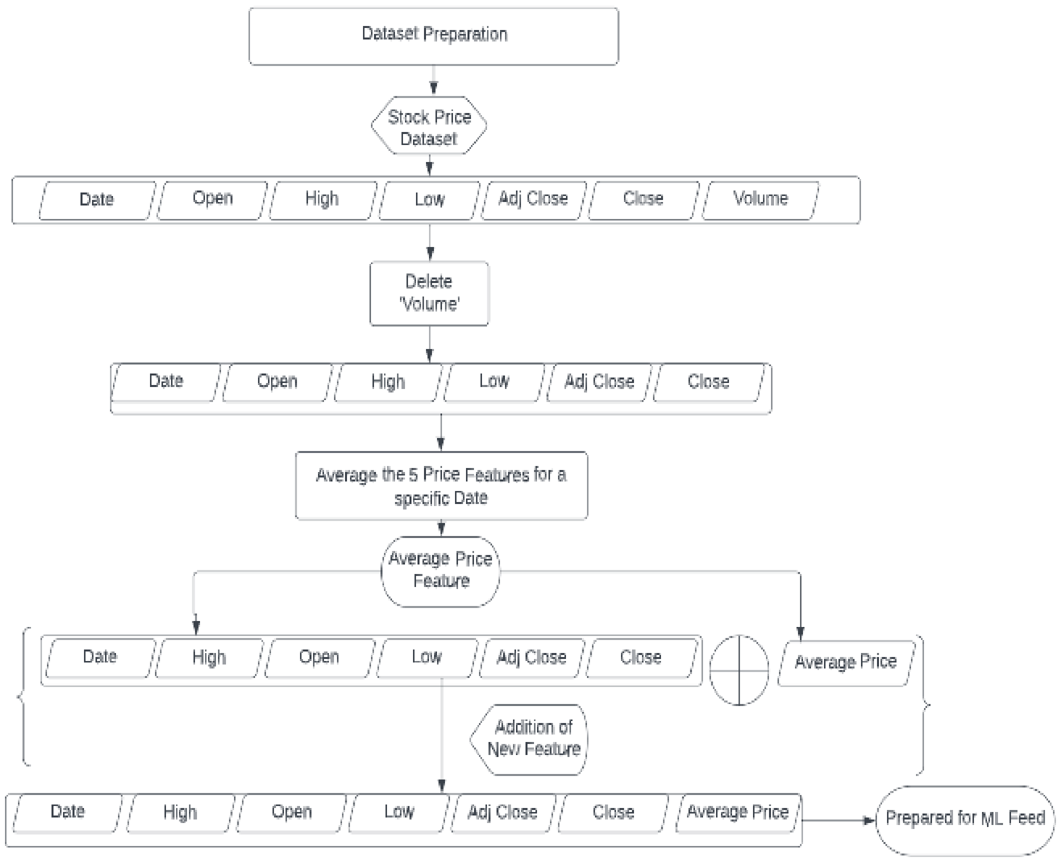

Fig 3.4: Average Price Feature based Price Dataset

This research has utilized both machine learning and statistical processes in order to create additional price features. The research will evaluate the portfolio performance using the predicted prices drawn from the three different price datasets.

a) **Conventional Price Dataset**: This price dataset will have the features of 'Open', 'High', Low', 'Adj Close', 'Close' in this dataset.

b) **Additional Feature based Price Dataset**: The price dataset will have the features of 'Open, 'High', 'Low', 'Adj Close', 'Close' and *'PCA defined Price Feature'* in this dataset.

c) **Average Feature based Price Dataset**: This price dataset will have the features of 'Open', 'High', 'Low', 'Adj Close', 'Close' and '*Average Price'* in this dataset.

3.3 Chi Square Testing for Feature Independency

In this thesis we have considered evaluating the dependency on predicted price dataset with original prices through accuracy measurement by chi square testing. This process will eliminate any doubt if it suggests that newly created price features are required for prediction or not. The output result is more dependent on all of the features rather than just one. In order to proceed with Chi Square testing, we need to evaluate it by hypothesis testing, which can infer using the following conclusions.

The evaluation will occur by defining the difference between predicted training prices and real training prices. Our research threshold should suggest if each feature is at least more than 10% important then they should be good enough for prediction. The following Figure 3.5 will describe the chi square evaluation method to determine the feature independency for forecasted prices drawn using training dataset.

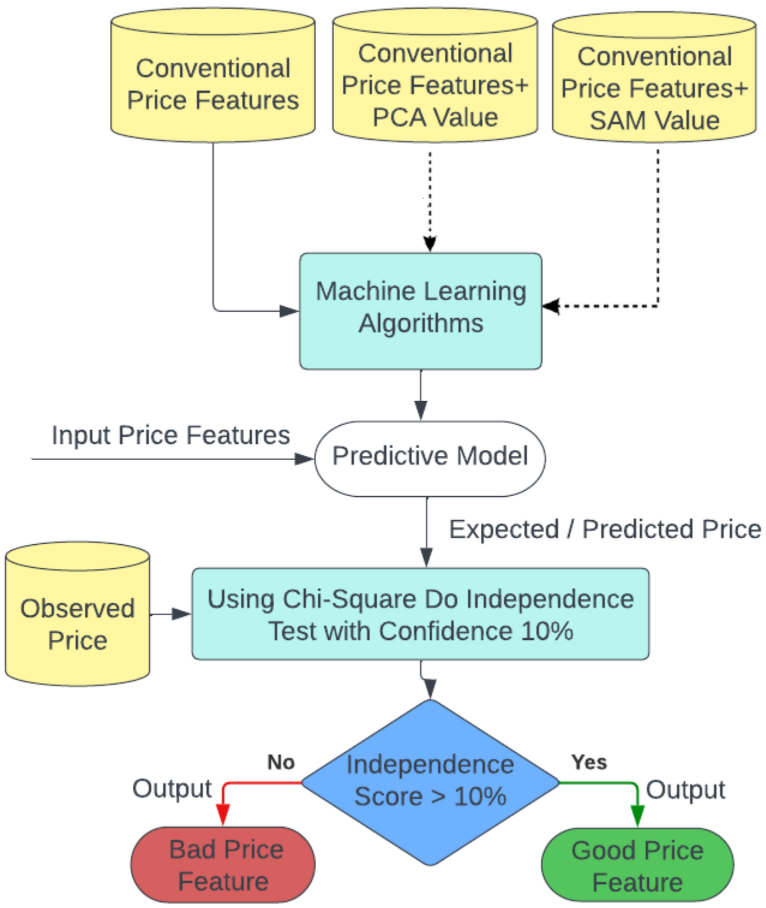

Fig 3.5: Feature Selection Process

Chi Square Testing is used to check the independency of features. If correlation of features is higher, the accuracy of machine learning will be lower.

In the Chi Square testing the difference between observed (Original price) and expected values (predicted price from machine learning) are calculated. Then, for the given degrees of freedom (5-1) for conventional feature data set and confidence interval 0.1 we can find out the critical value from the Chi Square Distribution Table. Now we have to check whether the critical value is greater than our obtained value of Chi Square. If it is, then our value is significant that means features are independent.

3.3.1 Application of Chi Square Test

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to establish whether a difference between the observed and expected data is because of chance or to a correlation between the variables under consideration.

By creating additional price features, this thesis has questioned the importance of presently available conventional price features. Hence, we need to evaluate the prediction quality of each predicted prices generated using forecasting algorithms in order to see if each predicted price features have acceptable feature dependency for prediction. As 6 different supervised learning methods are used each forecasting processes will predict one specific price. The predicted price will then be compared for chi square score evaluation. The average chi square scores of each price will then be averaged and finally represented as percentage in order to exhibit a chi square score exhibited by each prediction algorithm. The step wise process is discussed below:

Step 1: Training Dataset will be used for price prediction using all supervised learning algorithms.

Step 2: Predicted Price Dataset generated in all ML processes.

Step 3: Chi Square Test applied using Chi Square Test in Equation 2

Step 4: Averaging all Chi Square Test Results (i = 1 till n being the very last chi square score) shown in Equation 2

$$\text{Average } \chi 2 = \frac{\sum(\text{Chi Square Scores from 1 till n})}{\text{Total Number of Observations}} \ldots\ldots\ldots\ldots\ldots \text{....Eq (2)}$$

Step 5: Distributing the Chi Square Dependency into Percentage and consider the ones with greater than 10% result.

The significance value 0.1 is used in the experiments. The following Figure 3.6 explains the complete process

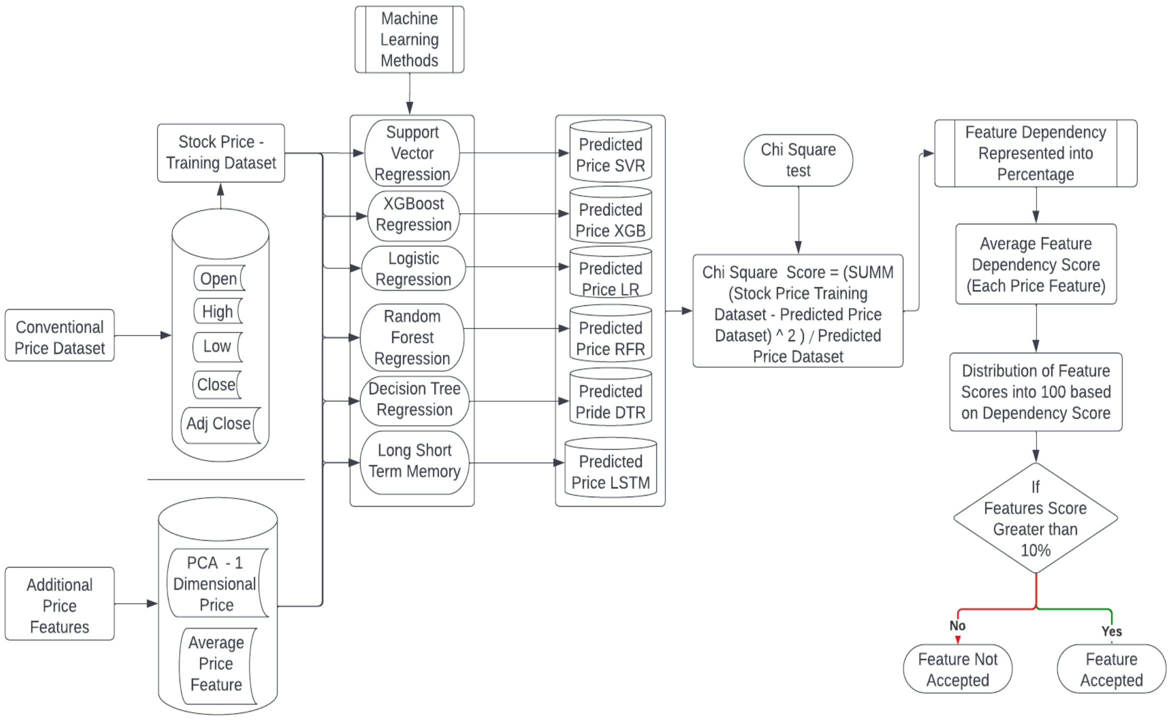

Fig 3.6: Implementation of Chi Square Test for Feature Selection

In the results and analysis chapter we have shown in better details that the additional price feature dataset which consists of PCA based 1 dimensional price that has low feature dependency than 10% making it an inconsistent feature. Whereas we have found that average feature-based price dataset consisting of average price feature have the highest feature dependency by 26%. This illustrates that only conventional feature base price dataset and average feature-based price dataset should be used for stock price forecasting.

3.5 Stock Price Forecasting – Conventional Method

In this research we have performed stock price forecasting using conventional price features. As Selvin et al and Wang et al have shown that stock price forecasting using all 5 price features provides best forecasting accuracy, hence for this research we have used these 5 price features for stock price forecasting [53], [54].

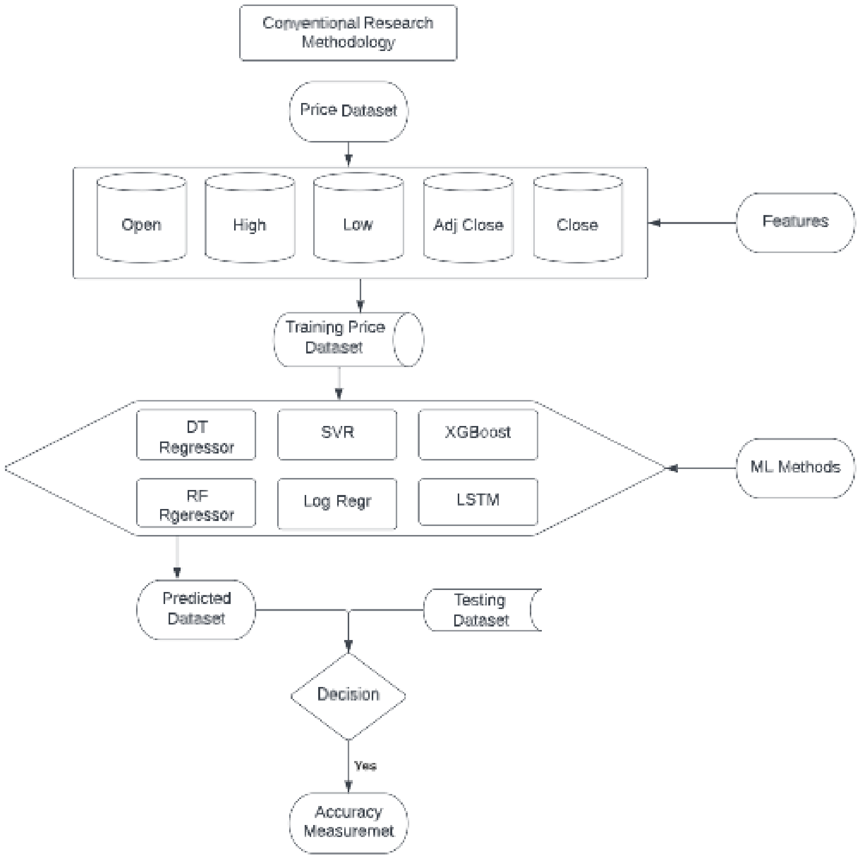

Fig 3.7: Stock Price Forecasting Using Conventional Feature based Price Dataset

3.5.1 Stock Price Forecasting – Proposed Method

The Chi square test is carried out in order to figure out the prediction accuracy defined on the predicted dataset being created on training prices. As price forecasting is completely dataset dependent various models are used for forecasting purposes. The following flowchart describes the stock price forecasting method identifying the best method by accuracy. The research process will be carried out using two different forecasted prices generated from two different price dataset. The following Figure 3.7 exhibits the process of machine learning application in the conventional price dataset and new price dataset in order to forecast stock prices.

Classification model using ML Tools have been constructed in terms of rise or fall of stock prices. Here the classification has been performed in order to correctly identify whether a given stock would rise or fall.

On the other hand, Regression model using ML Tools have been constructed for obtaining expected value in feature evaluation using chi square test.

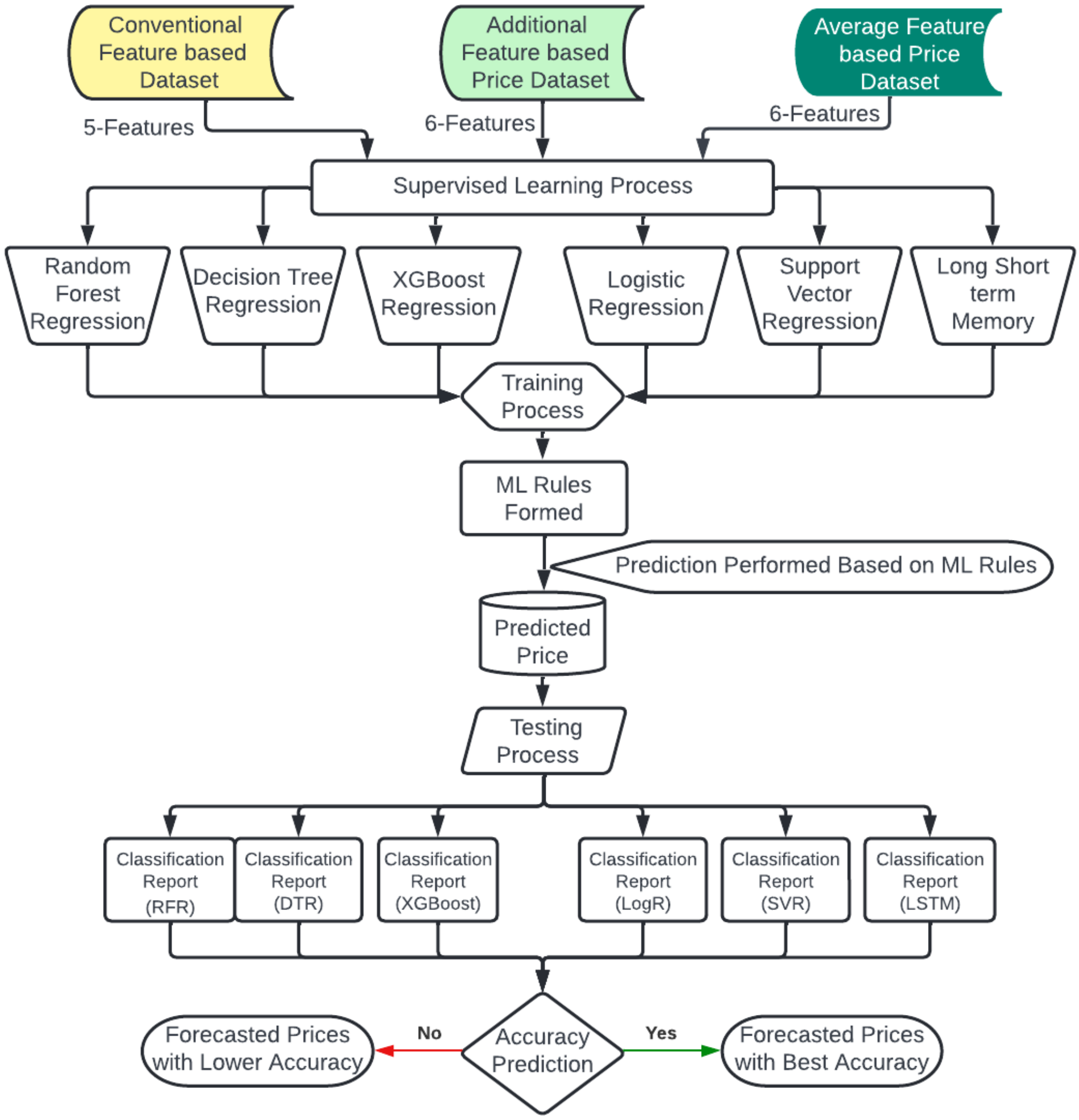

Fig 3.8: Stock Price Forecasting Flowchart

If we refer to our Chapter 4 in section 4.2.4 regarding significant feature selection, we found that the feature obtained by PCA process had a dependency of 9% which is less than the minimum threshold requirement of 10%. Hence for this research purpose we are not including additional feature-based price dataset for stock price forecasting. For this research our price forecasting process using ML is described in Figure 3.7.

Step 1: Include both conventional feature-based price dataset and average feature-based price dataset with 5 and 6 features respectively for stock price forecasting.

Step 2: Feed price dataset of both into each supervised learning algorithms.

Step 3: Using the training price dataset determine the ML rules for price forecasting.

Step 4: Predict prices for testing period.

Step 5: Compare the predicted prices with actual prices for the testing period.

Step 6: Determine the classification report and forecasting accuracy for predicted price dataset.

Step 7: Compare the forecasted price dataset obtained using different supervised learning methods and finally determine the best ML process for price forecasting despite varying stock volatility

Whether a stock is volatile or not depends on the standard deviation of prices of different stocks on different dates. If prices of a particular stock vary randomly, that stock is volatile. If a stock is volatile, the investor will be confused to invest.

Volatility depends on beta value. Lower or higher value of beta indicates the more volatility.

The following results will be attained

a) Best ML method for price prediction using additional price features and conventional price features will be determined
b) We will be able to understand and differentiate if the additional price features create better accuracy in forecasting processes or not, considering which ML processes perform it better.
c) Best predicted prices will be used in the Portfolio development.

For Regression Model, R2 is used. On the other hand, for classification model, Correct rate, Accuracy, F1-Score and Recall are used.

3.6 Portfolio Optimization

Using ML method for price prediction, we have decided to apply these predicted prices to create a financial portfolio that will have least risk and highest return. The portfolio created using the predicted price portfolio will be evaluated against the portfolio developed by real prices for the same time period. The final results will determine if the research queries are answered or not by the following attributes.

a) Portfolio efficiency has to be greater than 0.5 as it will determine if the predicted prices will be able to generate profit.
b) Portfolio should make a return on investment greater than research target.

The two different predicted price datasets obtained using conventional feature-based price dataset and average feature-based price dataset will be used for portfolio creation. The predicted prices should have an indication of growth and would require quality stock allocation and performance accuracy in terms of portfolio optimization. Mostly the portfolio performance needs to be highly accurate in terms of portfolio created using actual prices. Figure 3.8 describes the portfolio optimization flowchart.

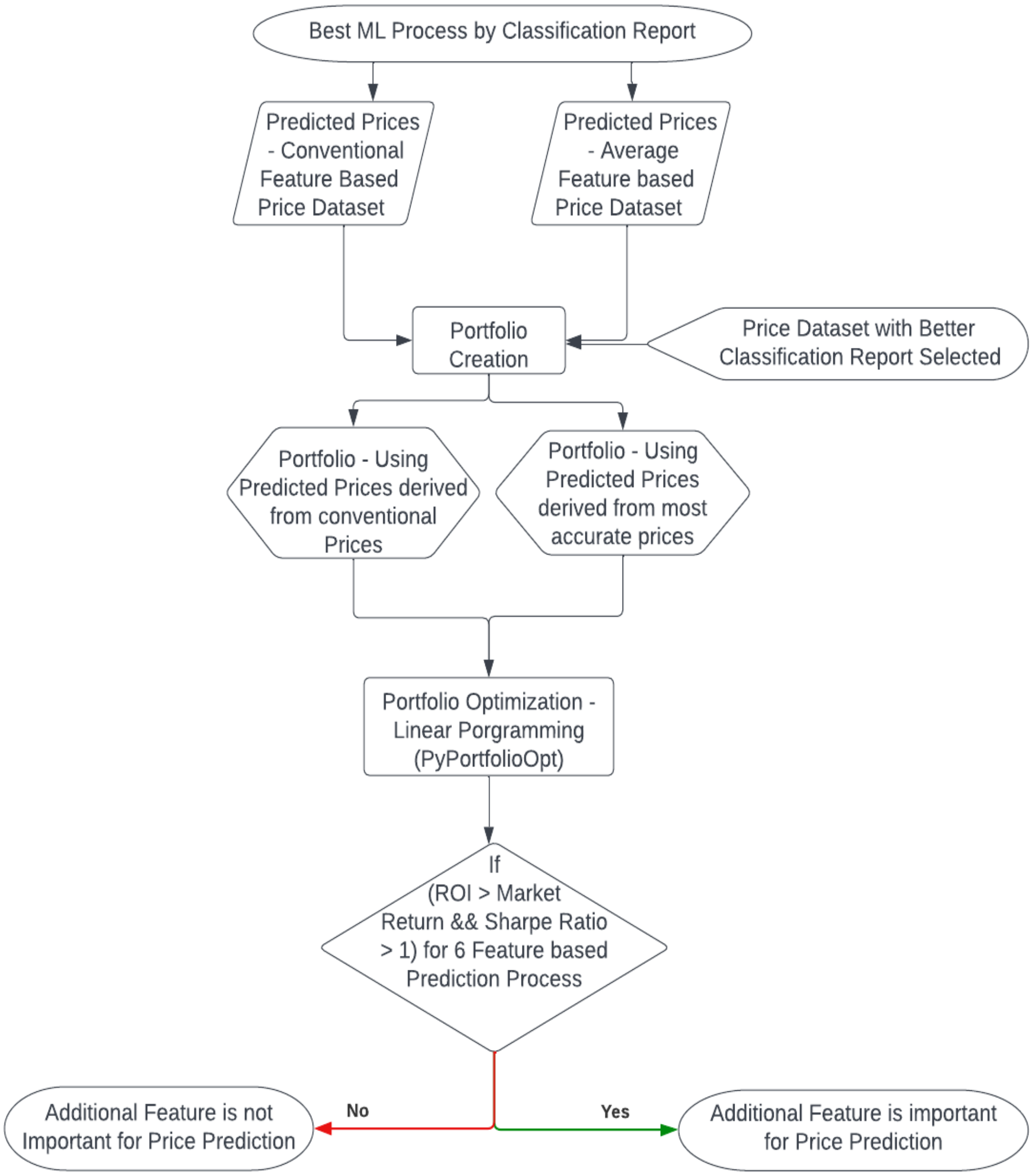

Fig 3.9: Portfolio Optimization Flowchart

The following steps are followed in Portfolio creation and optimization using predicted prices

Step 1: The stocks which are required for selection in portfolio needs to have lower buy price and higher predicted sell price.

Step 2: The stocks with criteria from Step 1 will be selected for Portfolio Creation

Step 3: Using Linear Programming applied by PyPortfolioOpt.py we would make stock allocation and determine

a) Portfolio ROI
b) Portfolio Sharpe Ratio

Step 4: We will create portfolio of same stocks using actual prices for Testing period and then compare the previously built portfolio using predicted prices with the current portfolio in terms of Sharpe Ratio achievement, Performance accuracy and Stock allocation accuracy.

Finally, the portfolio created by the predicted prices will be tested in terms of research objectives. The following observations would be made after creation of Portfolio

a) The Portfolio created using average feature-based price dataset needs to have higher ROI than average market return rate and best savings account rate. This will make an investor coming from the lower middle class or middle class more interested in making investment using additional price feature defined by average process.

Ultimately, the research objective is clarified within the framework of methodology. The complete research process will be evaluated in chapter 4 where the quantitative findings will resolve the research queries.

Using input feature vector to Machine Learning tools we can predict whether a stock is fallen or risen. Those stocks which rise sharply are used to construct a portfolio. The different weights are allocated to different stocks in portfolio so that the risk factor can be minimized and consequently the profit can be maximized. Here, an optimum set of weights for different stocks in portfolio will be determined to minimize the risk factor. Weights are allocated based on the return value and also on the volatility of that stock.

Chapter 4
Results & Analysis

A research is evaluated within the parameters of the results analysis. In order to observe the proposed methods, the derived results are evaluated in this chapter. Firstly, we talked about selection of stocks. Secondly, we discussed regarding conventional and proposed price features. Thirdly we speculated regarding the results obtained about feature dependency tests. Fourthly we talked about the accuracies achieved by ML methods and the best predicted prices. Finally, the results acquired by the portfolio derived by the predicted prices are evaluated.

4.1 Selection of Stocks Based on Volatility and Financial Ratio

As Lin et al have suggested that stocks are best selected within the constraints of market performance parameters and their relation to markets; hence we decided to select stocks based on financial ratios and stock Beta. The following stocks have been generated from the selected files of stocks provided in table 5.

Table 5: Selected Stocks for Research

Selected Stocks	Ticker	Beta	EPS	Current Ratio
High Beta Stocks	CELH	2.2	9.78	1.78
	OSTK	4.41	6.34	1.91
	NLS	1.78	4.32	1.62
	TSLA	2.71	2.31	1.73
Medium Beta Stocks	CAT	0.95	11.64	1.24
	DE	0.98	8.74	1.17
	HD	1.02	2.34	1.31
	TSM	1.01	1.19	1.09
Low Beta Stocks	VNET	0.36	2.31	1.27
	ERIE	0.43	4.65	1.16

4.2 Prepared Price Dataset

A following price dataset is generated on the basis of research suggestion process that shows the following price dataset. This price dataset includes all the conventional price features. Based on ML and statistical process the derived prices also got included below:

Table 6: Price Dataset including Existing & Proposed Features

Existing Features						Proposed Features	
Date	**Open**	**High**	**Low**	**Close**	**Adj Close**	**Average Price**	**1 Dimensional Price**
7/22/2022	12025.4	12093.02	11767.19	11834.11	11834.11	11910.76	11865.97
7/21/2022	11914.2	12060.59	11812.72	12059.61	12059.61	11981.336	12010.15
7/20/2022	11726.1	11939.96	11703.36	11897.65	11897.65	11832.942	11835.36
7/19/2022	11515	11721.22	11448.97	11713.15	11713.15	11622.298	11610.91

In Table 6, primarily we demonstrate the existing price features for stock price prediction. In our research we have proposed two additional price features named as 'Average Price' obtained by statistical average method and '1 Dimensional Price' generated from principal component analysis. Our prediction technique will propose a six variable based price prediction. This process will involve only one feature among the two proposed features for stock forecasting process. In this research the forecasting process will only be preceded when newly formed features show efficient feature dependency.

4.3 Chi Square test – Price Independency

A chi square test is performed in order to show if the real price features or additional price features are important in the prediction process. In this research we have tested the chi square score achievement using predicted prices attained by each price forecasting algorithms for training price dataset.

4.3.1 Conventional Feature based Price Dataset

It's observed that when we use all the 5 price features for making predictions each of them has greater than 10% dependency in making accurate price prediction. It's observed from the Figure 4.0, all the price features contain more than 10% importance in terms of price prediction.
Using the following Figure 4.0, price forecasting is dependent on conventional price features. This research would establish chi square dependency test for additional features based on additional feature-based price dataset and average feature-based price dataset.

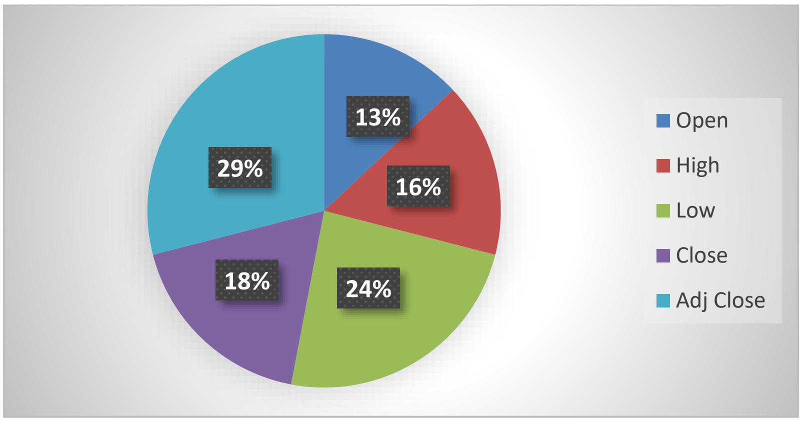

Fig 4.0: Chi Square Test – Conventional Price Dataset

Figure 4.1 shows the feature dependency of conventional features based on Chi square test for supervised learning algorithms. It's clearly observed that none of the features are less than 10% hence all the features are important for price prediction.

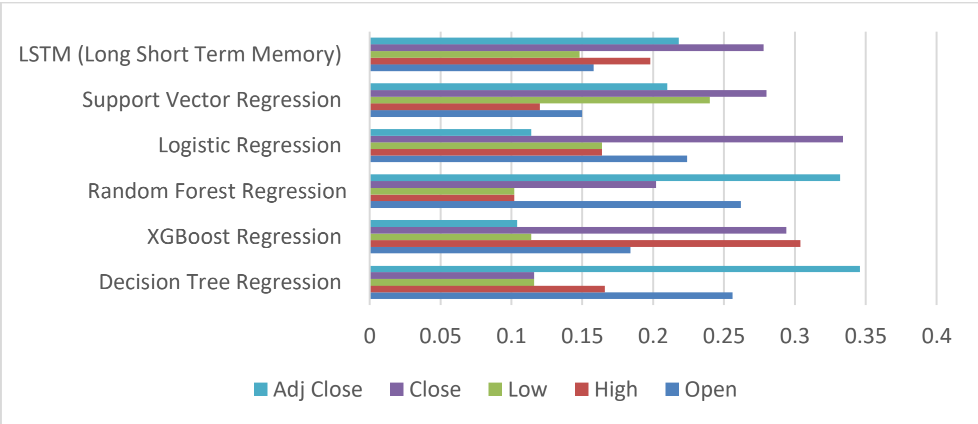

Fig 4.1: Chi Square Test for Conventional Feature based Price Dataset

4.3.2 Additional Feature based Price Dataset

In this thesis we have derived six price features including a PCA derived price that represents a one-dimensional price which is best representing all the other prices. PCA generally reduces the computational complicacy in reference to prediction but we have seen that PCA is usually known to sacrifice a certain degree of accuracy in relation to prediction. The chi square test result for 6 Price features is demonstrated below in order to show the importance of all the price features. Figure 4.2 illustrates that in several ML methods PCA based price feature have less than 10% value, showing incompetent feature dependency for price prediction.

Fig 4.2: Chi Square Test – Additional Feature based Price Dataset

4.3.3 Average Feature based Price Dataset

So far after evaluating the research findings we didn't find any price prediction performed using average prices of all the 5 price dataset features. The importance of average price should be exhibited by the chi square test if the prediction dependency becomes higher than 10% for each process. Figure 4.3 exhibits that average price feature remains as the most dominant feature considering none of the prediction methods has shown that feature to be less than 10% in terms of prediction accuracy.

Fig 4.3: Chi Square Test – Average Feature based Price Dataset

4.3.4 Significant Feature Selection

Using Chi Square test both the Average price added dataset with PCA incorporated Price added dataset are put forward in order to understand the consistency of these tests. It's observed that as PCA based price feature does not pass the 10% feature dependency hence it does not remain as an important price feature for prediction exhibited by Figure 4.4.

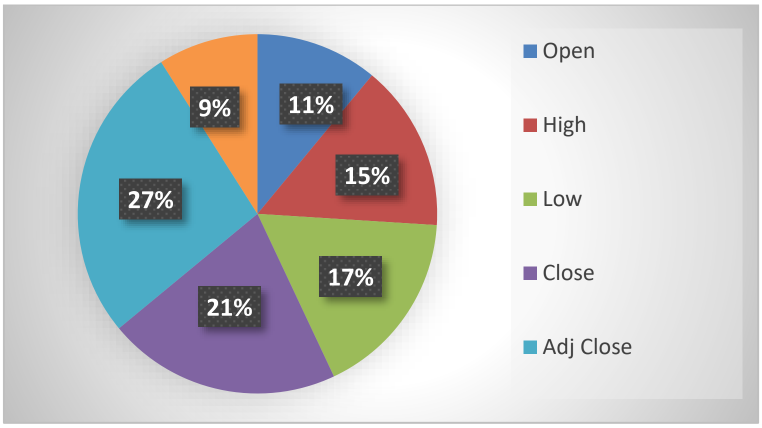

Fig 4.4: Chi Square Dependency – Additional Feature based Price Dataset

Similarly, we have observed that the average price feature remains an important price feature. Figure 4.5 shows that average price feature exhibits the highest feature dependency in comparison to all other price features. Hence average price feature could be an important contributor for price forecasting processes.

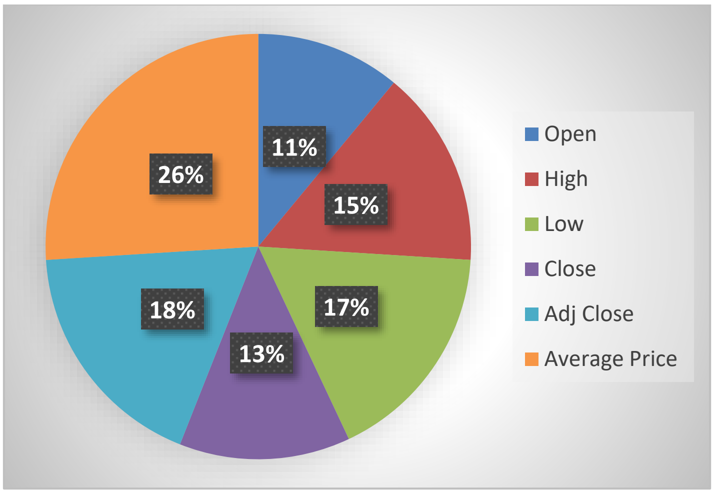

Fig 4.5: Chi Square Dependency – Average Feature based Price Dataset

Using PCA to predict prices produced different predicted prices for the training dataset. In chi square test, the additional price feature created by PCA is shown to be less than 10%. Also, we have shown that average price feature remains the most effective one among all others in financial prediction process with a feature dependency of 26%. Hence, PCA based price feature included in additional feature-based price dataset should not be added in this research.

4.4 Stock Price Forecasting

In this research we have applied different types of ML methods based on new proposed features for stock price forecasting. In terms of forecasting/predicted accuracy we have derived the best regression based predicted price for portfolio creation.

Table 7: Accuracy of Proposed Vs Conventional Price Dataset

ML Methods	XGBoost		SVR		RFRegr		LogRegr		DTR		LSTM	
Dataset	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**
CELH	67.7	63.1	78.1	79.8	67.3	63.5	63.3	63.9	64.1	61.1	**90.5**	90.1
OSTK	65.6	61.8	79.6	76.5	63.1	63.7	62.7	63.5	64.1	63.1	**87.3**	86.1
NLS	66.2	63.1	81.7	77.7	67.4	67.2	62.8	64.3	65.5	61.8	87.1	**88.5**
TSLA	67.8	62.5	78.8	76.1	67.5	68.3	62.7	63.4	66.4	61.6	**88.9**	87.2
CAT	64.7	62.5	79.2	75.6	62.5	69.8	65.4	62.7	63.6	59.8	**90.7**	86.3
DE	64.7	61.4	79.8	78.1	67.4	67.3	62.4	63.7	62.5	59.6	91.2	**91.3**
HD	63.1	61.5	79.8	79.3	67.2	68.9	67.4	61.2	63.2	63.9	**89.4**	88.9
TSM	63.1	60.6	77.7	76.6	67.2	63.6	66.1	65.1	66.2	61.3	**91.6**	86.9
VNET	64.6	62.7	80.9	76.1	66.5	67.1	62.4	60.2	66.6	59.5	**92.1**	90.4
ERIE	67.8	63.3	78.9	78.7	63.6	64.2	62.7	61.6	63.6	63.3	**91.7**	85.4
Average	65.52	62.24	79.45	77.4	65.97	66.34	63.78	62.93	64.6	61.5	**90.1**	88.1

We have seen that LSTM makes the best possible forecasting using both conventional feature-based price dataset (CFD) and average feature-based price dataset (AFD). Similarly, the classification report also indicates that among all the aforementioned supervised learning methods LSTM predicts the stock prices with best accuracies. In Table 7 we will demonstrate the price predictions based on obtained accuracies of various ML methods. If we observe carefully among the 10 stocks of variant volatility 8 of them had a higher accuracy generated using the proposed price dataset whereas 2 of them had an accuracy greater using conventional feature-based price dataset. This proves that our proposed price dataset is better for making price forecasting. Our research originality lies with provision of average price feature as an additional price feature that helps make the forecasting of prices much better.

Table 8: Average Results by Classification Parameter

Classification Report	Precision		Recall		F1 Score		Accuracy	
Dataset	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**
XGBoost Regression	63.4	64.45	59.2	61.51	61.2	64.3	62.2	65.5
Support Vector Regression	75.6	77.61	72.3	73.42	73.4	76.6	77.4	79.5
Random Forest Regression	67.4	64.4	63.4	61.32	65.5	64.5	66.3	66.01
Logistic Regression	62.4	62.3	58.7	59.8	61.5	61.2	62.9	63.8
Decision Tree Regression	61.3	64.5	58.3	61.3	60.4	63.1	61.5	64.6
LSTM	**87.1**	**89.12**	**83.2**	**85.4**	**85.6**	**88.7**	**88.1**	**90.2**

In Table 8, we demonstrate the average results obtained by each supervised machine learning methods which are used for making effective evaluations regarding these processes prediction quality. The results are explained in terms of the measurement parameters demonstrated by classification report. Using LSTM, that all the scores achieved in the classification report are over than 80%. It seems that LSTM remains the best prediction method when the prediction was performed using both conventional feature-based price dataset and average feature-based price dataset.

Among the six different supervised learning-based ML processes which are predominantly used for price prediction LSTM remains the best one with 88.1% and 90.1% accuracies for conventional price dataset and proposed price dataset. It's seen in case of conventional feature-based price dataset each classification features are exclusively lower than the average feature-based price dataset. This proves that average feature-based price dataset remains a good indicator for price prediction. Adjusted R^2 remains as the best fit for ML tools. The overall values for R^2 for each ML tools for predicted prices using both AFD and CFD are presented below:

Table 9: Adjusted R^2 Score fore ML tools

Adjusted	SVR		LogR		RFR		DTR		Xgboost		LSTM	
R Square	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**	**AFD**	**CFD**
CELH	0.85	0.74	0.71	0.69	0.77	0.76	0.69	0.61	0.73	0.72	0.88	0.88
OSTK	0.8	0.8	0.74	0.64	0.77	0.74	0.74	0.69	0.71	0.71	0.92	0.87
NLS	0.78	0.76	0.69	0.64	0.74	0.71	0.65	0.65	0.69	0.7	0.91	0.91
TSLA	0.82	0.77	0.73	0.67	0.75	0.75	0.72	0.69	0.69	0.65	0.88	0.84
CAT	0.82	0.74	0.71	0.66	0.79	0.71	0.7	0.61	0.72	0.65	0.94	0.9
DE	0.79	0.76	0.74	0.7	0.73	0.72	0.73	0.63	0.7	0.71	0.88	0.84
HD	0.85	0.76	0.68	0.67	0.76	0.72	0.69	0.64	0.74	0.67	0.92	0.88
TSM	0.79	0.81	0.72	0.67	0.78	0.74	0.65	0.66	0.73	0.65	0.87	0.9
VNET	0.84	0.79	0.68	0.69	0.75	0.74	0.65	0.62	0.71	0.66	0.93	0.9
ERIE	0.85	0.81	0.7	0.69	0.79	0.7	0.64	0.65	0.68	0.65	0.93	0.91

4.4.1 Prediction Results – Conventional Feature based Price Dataset

Six popularly known ML methods are used for stock price forecasting. Primarily using conventional price features forecasting is concluded. Figure 4.6 exhibits the classification report exhibited by predicted prices generated from conventional feature-based price dataset. It shows that LSTM exhibits the highest classification report value in comparison to other ML methods.

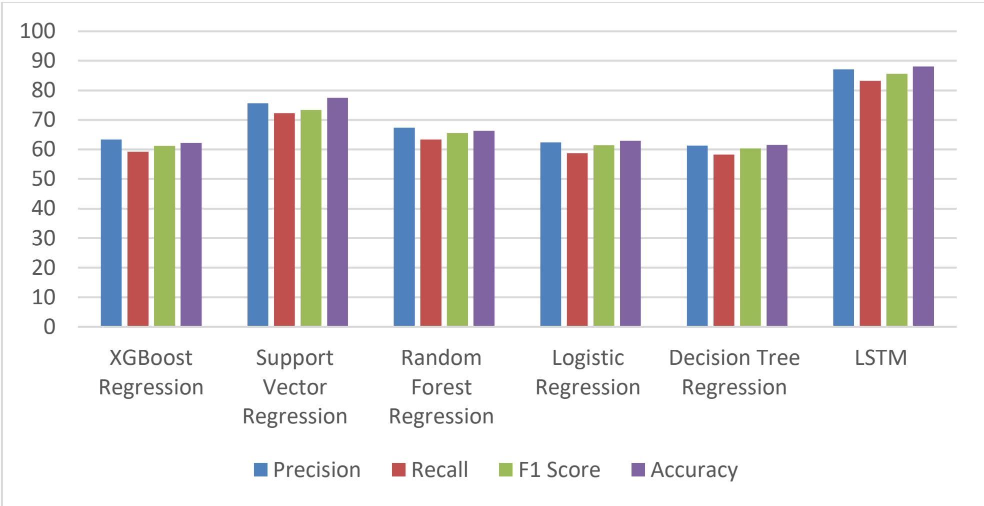

Fig 4.6: Classification Results for Conventional Feature based Price Dataset

LSTM is a machine learning algorithm that could be used for predicting the prices of stocks and commodities. It's observed that in average of 10 different stocks, LSTM has been able to attain the highest accuracy for each stock with values greater than 80% as explained in Figure 4.7.

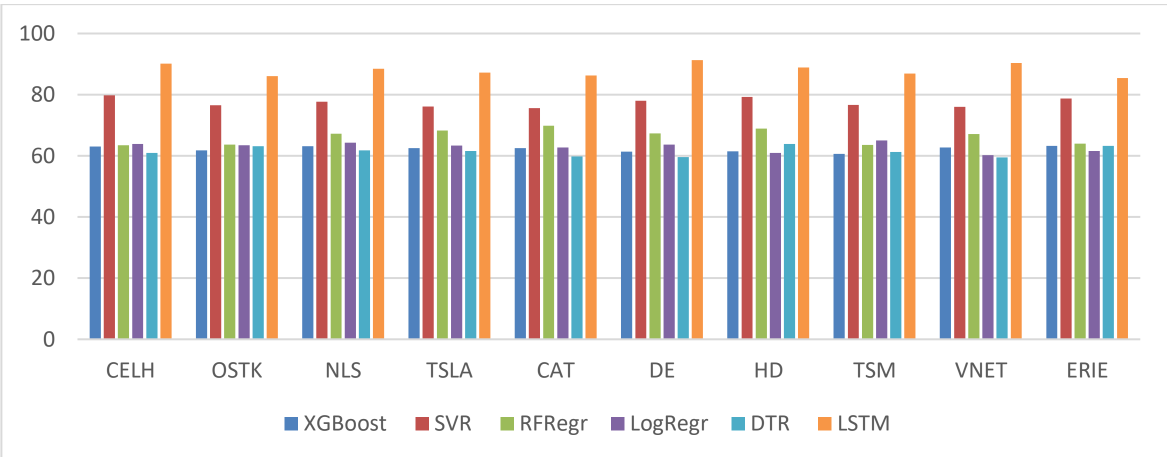

Fig 4.7: Accuracy Graph for Conventional Feature based Various Stocks

4.4.2 Prediction Results - Average Feature based Price Dataset

It's observed in Figure 4.8 that in case of supervised learning methods LSTM makes better predictions rather than other forecasting methods in case of stock price predictions when average price feature is also included for price prediction along with conventional price features.

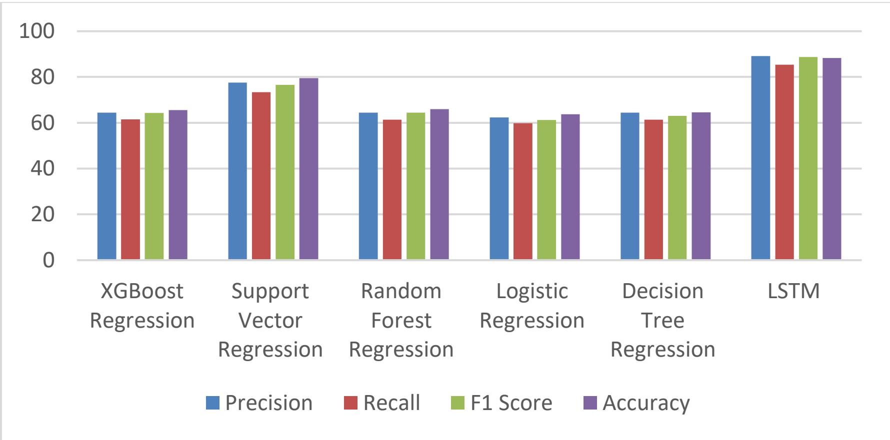

Fig 4.8: Classification Results for Average Feature based Price Dataset

Average price is an additional price feature that is used by supervised learning methods used for ML based regressions. In this study, the average price did have the lowest standard deviation in comparison to their real price counterparts. The new average price would be considered a good prediction feature only if the precision, accuracy score with the 6 price features inclusive of average becomes better. The Prediction accuracy generated by average feature-based price dataset is exhibited below in Figure 4.9 for all the 10 Stocks.

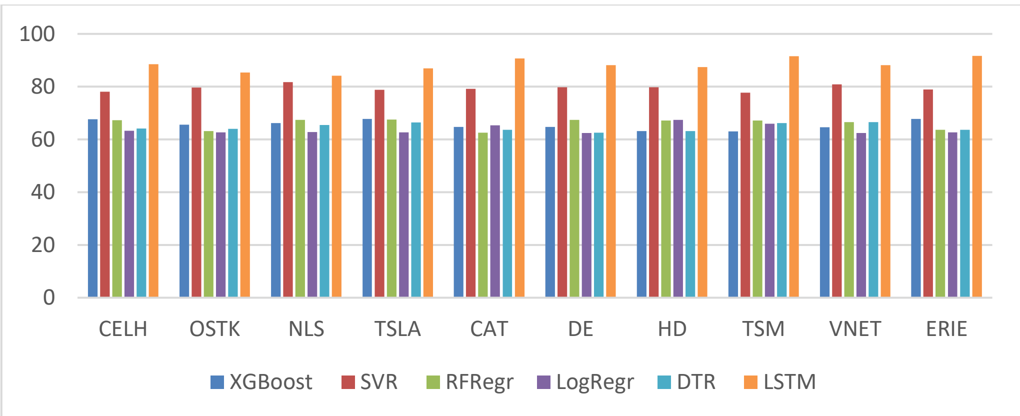

Fig 4.9: Accuracy graph for Average Feature based Various Stocks

4.4.3 Conventional & Proposed LSTM based Forecasting Accuracy

It’s seen that among the various supervised learning methods for identifying the best possible prediction processes these prediction methods should be evaluated by accuracy scores and MAE values. As our proposed method predicts price with better accuracy and the classification report also suggests better results; hence for the purpose of portfolio development using predicted prices we need to include the best possible prices predicted by the supervised learning methods in terms of accuracy and MAE scores as shown in Figure 4.10 and 4.11.

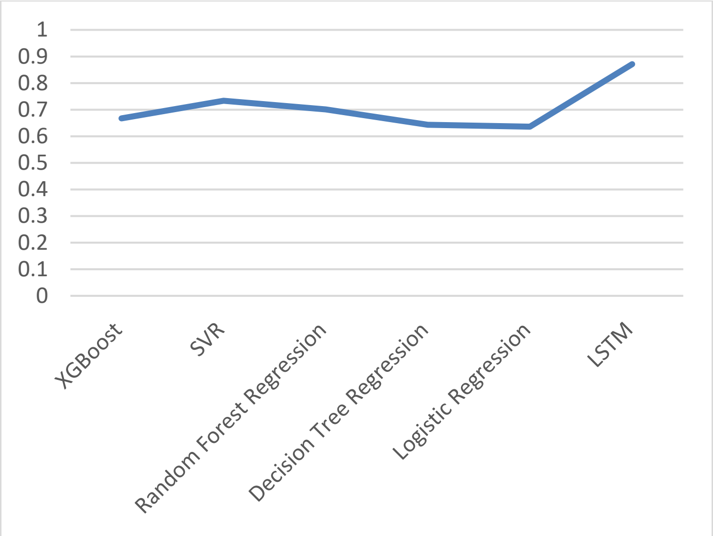

Fig 4.10: Accuracy & MAE of ML Algorithms

Fig 4.11: MAE of Forecasting Method

Comparing both accuracy and MAE we have understood that LSTM is by far the best prediction methods with highest accuracy and lowest error. The predicted prices thus attained by LSTM will have the best quality in terms making portfolio. Moreover, if the predicted prices are used for portfolio development with better accuracies, then financial forecasting and ROI will be practical. According to Figure 4.10 and 4.11 we observe that average accuracy attained are around 90% and average MAE is less than 10% in case of LSTM. Figure 4.11 shows that the best forecasting process is attained using LSTM.

4.5 Portfolio Optimization

In this article, we look at how to optimize the risk and return parity of a portfolio. The process of selecting the best possible risk/return parity using PyportfolioOpt.py is described in more detail in the section on portfolio managers' portfolio optimization. The stepwise portfolio optimization process is exhibited below:

4.5.1 Portfolio Creation

In our research we have concluded that LSTM makes the best price predictions. The predicted prices will be used against the real prices from 15th March 2020 till 30 June 2020. In our research process we will create two different portfolios and both portfolios need to have the following constraints.

a) The first real price of the stock has to be less than the predicted last price of the stock.
b) The last price of the stock needs to be at least 10% greater than first price of the stock.
c) The Predicted price dataset should have a positive slope and a slope greater than the market slope.

All the above attributes are covered using the Beta and Sharpe Ratio of the stock in terms of its return potential. The following Table 10 explains the positive ROI achieved by predicted prices and the potential Sharpe ratio achievement due to increased final predicted price. This will indicate which stocks we need to consider for building a qualified portfolio. In Table 8 we exhibited the real prices and predicted prices obtained using both price dataset. The predicted prices obtained using Conventional Feature based Price Dataset shows lower similarity rather than the forecasted price obtained using Average Feature based Price Dataset.

Table 10: Forecasted Stock Price Using LSTM

Stocks	Real Price (03/15/2020)	Predicted Price (06/30/2020) Conventional Price Dataset	Predicted Price (06/30/2020) Average Price Dataset	Average Growth Percentage	Condition
CELH	4.44	5.22	5.34	0.189189189	Accept
OSTK	1.69	2.83	2.79	0.662721893	Accept
NLS	1.22	2.78	2.81	1.290983607	Accept
TSLA	554.65	671.43	653.48	0.194365816	Reject
CAT	231.45	233.65	230.39	0.002462735	Reject
WYY	2.4	3.5	3.43	0.44375	Accept
HD	109.11	113.41	112.16	0.033681606	Reject
TSM	102.35	104.56	103.57	0.016756229	Reject
VNET	4.96	5.76	5.43	0.128024194	Accept
ERIE	34.55	31.46	33.17	-0.064688857	Reject

According to Table 9 we observe that various stocks have been considered for portfolio development. The primary condition for selection in order to create a portfolio with these stocks will be their growth beyond 10%. TSLA made more than 10% overall predicted return but in reality, the contribution margin being large will not make it a good enough investment alternative. Hence, we found that only the stocks with accepted conditions will be used for portfolio development.

4.5.2 Portfolio Optimization

The convex optimization process used by PyPortfolioOpt will analyze all the redundancies using the minimum variance under the benchmark of optimum gain. The created portfolio thus will fulfill the following criterions along with stock distribution. Two portfolios created by predicted prices are measured against accuracy in terms of Portfolio Efficiency. An efficient portfolio can be built using average price as an additional feature. The new feature introduced by averaging the prices of a portfolio makes the stock prediction more efficient. Similarly, the allocation of stocks also became better with the introduction of average prices of all real prices as a feature in the dataset. The following figure in 4.16 explains the efficient portfolio.

Table 12: Portfolio Results

Summary of Proposed Research (Investment of $1000)		
Portfolio Parameters	**Accuracy**	**Benchmark**
Predicted Portfolio	90.11%	Real Portfolio will have 100% Accuracy
Two Months Gain	23.32%	Yearly S&P return is 10.5%
Expected Annual Return	23.32%	Yearly Highest Savings Account Rate is 7.8%
Annual Volatility	48.30%	S&P Annual Volatility is 22.35%
Portfolio Accuracy	73.31%	Has to be greater than 50%
Approximate Gain	212.6	Gain has to be more than $105
Actual Gain	233.2	Gain has to be more than $105
Sharpe Ratio	50.32	Sharpe Ratio needs to be more than 1

According to Table 12, ROI is explained in terms of performance. It's observed that the portfolio formed using our proposed price dataset makes more effective return in terms of return of investment.

Hence, we can deduce that introducing average prices as an additional price feature helps make both price predictions and portfolio accuracy better.

a) Produce better price predictions when price predictions are conducted using multiple features. (Average feature-based price dataset using LSTM produced an accuracy of 90.1% whereas conventional price features produced an accuracy of 88.1%).

b) The portfolio derived using the predicted prices of proposed price dataset have a ROI of 23.31% in comparison to 21.61%. This proves that our proposed price dataset is a better price dataset that would make better portfolio and better price forecasting.

Chapter 5
Drawbacks & Conclusion

5.1 Limitations of Research

We have not considered the following issues those could be further evaluated for preceding researches

a) In this research we didn't apply deep neural network-based stock price prediction.
b) Cloud based price derivation could be applied in order to make prediction processes more efficient.
c) Modified supervised learning-based regression processes could be used for betterment of price prediction.
d) Further researches could be performed using stocks with lower standard deviation and better financial conditions.

In this research, we have used only average of five feature/PCA based feature as a derived price feature with conventional feature for stock market prediction. In near future normalized feature will be used for experiments. A thorough investigations will be done to check whether the inclusion of normalization has provided better performance or not. It is expected that those experiment values will be used in our upcoming journal submission.

5.2 Conclusion

The selection of stocks for investment, considering the price pattern remains the principal concern for financial forecasting. Our research method has investigated successfully the best possible ML method that have derived price prediction with high degree of accuracy. Despite the above-mentioned drawbacks, we have achieved the following milestones in this research process that have created the suggested results.

a) PCA based price feature exhibited a feature dependency of 9% making it inconsiderable for stock price forecasting.
b) Average price feature exhibited a feature dependency of 26% making it the most considerable price attribute for stock price forecasting
c) Multiple price feature including average price feature made highest possible prediction accuracy for stock price forecasting exhibiting a value of 90.11%.

d) Portfolio derived using predicted prices of average feature-based price dataset exhibited a ROI of 23.32% in a period of two months.

The above evidences proved that the price dataset derived using additional price feature of average price could make better price forecasting. The research target is certainly achieved by making more than effective market return rate of 10% (S&P 500 Yearly return rate) and highest savings account return rate of 8% (Certificate of deposit return rate) by both conventional and proposed price features for predicting stock prices. Hence, by this research a person making small investment can dream of making a secondary earning through ML based forecasting in US stock market.

5.3 Discussion

This research has been conducted in order to address various issues. We can thus propose the following issues for this research:

a) Two additional price features have been proposed using PCA and statistical averaging technique. The chi square test has proven the feature dependency for each price features. It's observed that proposed PCA based price feature obtains the lowest feature dependency of 9% and proposed average price feature exhibited highest price dependency of 26%. It means that for price forecasting using average feature-based price dataset would provide a high degree of forecasting accuracy.

b) In terms of forecasting accuracy LSTM provided the highest accuracy as with lower standard deviation-based price feature helped the neural network model to learn better regarding the price data and make better predictions. Hence the price forecasting accuracies are around 2% higher using proposed price dataset and portfolio-based prediction parameters are better by around 3% for ROI.

In this research we have observed that the additional price feature derived using averaging process exhibits lower standard deviation. In case of LSTM, such price dataset with lower standard deviation helps neural networks to learn more regarding future price forecasts. Hence through complete evaluation we can suggest that the proposed price feature-based price prediction and portfolio optimization have made the prediction process effective and efficient.

References

[1]. Andersson, F.W., Johansson, D., Karlsson, J. et al. The characteristics of family firms: exploiting information on ownership, kinship, and governance using total population data. Small Bus Econ 51, 539–556 (2018). https://doi.org/10.1007/s11187-017-9947-6

[2]. Gottschlich, J., & Hinz, O. (2014). A decision support system for stock investment recommendations using collective wisdom. Decision Support Systems, 59, 52-62. https://doi.org/10.1016/j.dss.2013.10.005

[3]. C. Hargreaves and Y. Hao, "Does the use of technical & fundamental analysis improve stock choice?: A data mining approach applied to the Australian stock market," 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), 2012, pp. 1-6, doi: 10.1109/ICSSBE.2012.6396537.

[4]. Cvitanić, J., & Karatzas, I. (1992). Convex Duality in Constrained Portfolio Optimization. The Annals of Applied Probability, 2(4), 767–818. http://www.jstor.org/stable/2959666.

[5]. Haq, A. U., Zeb, A., Lei, Z., & Zhang, D. (2021). Forecasting daily stock trend using multi-filter feature selection and deep learning. *Expert Systems withApplications*, *168*, 114444. https://doi.org/10.1016/j.eswa.2020.114444

[6]. Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2015, pp. 147-154, doi: 10.1109/ICACSIS.2015.7415179.

[7]. Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 57(5), 102212. doi:10.1016/j.ipm.2020.102212

[8]. Mehtab, S., Sen, J., Dutta, A. (2021). Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models. In: Thampi, S.M., Piramuthu, S., Li, KC., Berretti, S., Wozniak, M., Singh, D. (eds) Machine Learning and Metaheuristics Algorithms, and Applications. SoMMA 2020. Communications in Computer and Information Science, vol 1366. Springer, Singapore. https://doi.org/10.1007/978-981-16-0419-5_8

[9]. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669. https://doi.org/10.1016/j.ejor.2017.11.054

[10]. Hammoudeh, S., Nguyen, D. K., Reboredo, J. C., & Wen, X. (2014). Dependence of stock and commodity futures markets in China: Implications for portfolio investment. Emerging Markets Review, 21, 183-200. https://doi.org/10.1016/j.ememar.2014.09.002

[11]. Zhu, L., & Fan, Y. (2010). Optimization of China's generating portfolio and policy implications based on portfolio theory. Energy, 35(3), 1391-1402. https://doi.org/10.1016/j.energy.2009.11.024

[12]. Yu, H., Chen, R., & Zhang, G. (2014). A SVM Stock Selection Model within PCA. Procedia Computer Science, 31, 406–412. doi:10.1016/j.procs.2014.05.284

[13]. Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. Expert Systems with Applications, 38(9), 11261–11272. doi:10.1016/j.eswa.2011.02.173

[14]. Shahi TB, Shrestha A, Neupane A, Guo W. Stock Price Forecasting with Deep Learning: A Comparative Study. Mathematics. 2020; 8(9):1441. https://doi.org/10.3390/math8091441

[15]. Sumaiya Thaseen, I., & Aswani Kumar, C. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. Journal of King Saud University - Computer and Information Sciences, 29(4), 462–472. doi:10.1016/j.jksuci.2015.12.004

[16]. X. Cai, S. Hu and X. Lin, "Feature extraction using Restricted Boltzmann Machine for stock price prediction," 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE), 2012, pp. 80-83, doi: 10.1109/CSAE.2012.6272913.

[17]. Hussein, M. & Özyurt, F. (2021). A New Technique for Sentiment Analysis System Based on Deep Learning Using Chi-Square Feature Selection Methods . Balkan Journal of Electrical and Computer Engineering , 9 (4) , 320-326 . Retrieved from https://dergipark.org.tr/en/pub/bajece/issue/65264/887339

[18]. Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications, 67, 126–139. doi:10.1016/j.eswa.2016.09.027

[19]. Salih Hasan, B. M. ., & Abdulazeez, A. M. . (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining, 2(1), 20–30. Retrieved from https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032

[20]. Long, W., Lu, Z., & Cui, L. (2018). Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems. doi:10.1016/j.knosys.2018.10.034

[21]. Lin, F., Liang, D., & Chen, E. (2011). Financial ratio selection for business crisis prediction. Expert Systems with Applications, 38(12), 15094–15102. doi:10.1016/j.eswa.2011.05.035

[22]. Amenc, N., & Goltz, F. (2013). Smart Beta 2.0. The Journal of Index Investing, 4(3), 15–23. doi:10.3905/jii.2013.4.3.01

[23]. Oh, K. J., Kim, T. Y., Min, S.-H., & Lee, H. Y. (2006). Portfolio algorithm based on portfolio beta using genetic algorithm. Expert Systems with Applications, 30(3), 527–534. doi:10.1016/j.eswa.2005.10.010

[24]. Vadlamudi, S. (2017). Stock Market Prediction using Machine Learning: A Systematic Literature Review. American Journal of Trade and Policy, 4(3), 123-128. https://doi.org/10.18034/ajtp.v4i3.521]

[25]. M. Asad, "Optimized Stock market prediction using ensemble learning," 2015 9th International Conference on Application of Information and Communication Technologies (AICT), 2015, pp. 263-268, doi: 10.1109/ICAICT.2015.7338559.

[26]. R. Batra and S. M. Daudpota, "Integrating StockTwits with sentiment analysis for better prediction of stock price movement," *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018, pp. 1-5, doi: 10.1109/ICOMET.2018.8346382.

[27]. Huang, H., Liu, X., Zhang, Y., & Feng, C. (2022). News-driven stock prediction via noisy equity state representation. *Neurocomputing*, *470*, 66-75. https://doi.org/10.1016/j.neucom.2021.10.092

[28]. M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," in *IEEE Access*, vol. 8, pp. 90847-90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[29]. D. Gray, D. Bowes, N. Davey, Y. Sun and B. Christianson, "The misuse of the NASA metrics data program data sets for automated software defect

prediction," *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)*, 2011, pp. 96-103, doi: 10.1049/ic.2011.0012.

[30]. Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340-341*, 250-261. https://doi.org/10.1016/j.ins.2016.01.033

[31]. Frías-Paredes, L., Mallor, F., Gastón-Romeo, M., & León, T. (2018). Dynamic mean absolute error as new measure for assessing forecasting errors. *Energy conversion and management*, *162*, 176-188.

[32]. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

[33]. H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend," *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, 2012, pp. 1-5, doi: 10.1109/ICSSBE.2012.6396535.

[34]. Farinelli, S., Ferreira, M., Rossello, D., Thoeny, M., & Tibiletti, L. (2008). Beyond Sharpe ratio: Optimal asset allocation using different performance ratios. Journal of Banking & Finance, 32(10), 2057–2063. doi:10.1016/j.jbankfin.2007.12.02

[35]. Bailey, David H. and López de Prado, Marcos and López de Prado, Marcos, The Sharpe Ratio Efficient Frontier (April 1, 2012). Journal of Risk, Vol. 15, No. 2, Winter 2012/13, Available at SSRN: https://ssrn.com/abstract=1821643 or http://dx.doi.org/10.2139/ssrn.1821643

[36]. Mansini, R., Ogryczak, W., & Speranza, M. G. (2014). Twenty years of linear programming-based portfolio optimization. European Journal of Operational Research, 234(2), 518–535. doi:10.1016/j.ejor.2013.08.035

[37]. Sadati, M. E. H., & Nematian, J. (2013). Two-level Linear Programming for Fuzzy Random Portfolio Optimization through Possibility and Necessity-based Model. Procedia Economics and Finance, 5, 657–666. doi:10.1016/s2212-5671(13)00077-4

[38]. Yudin, A. (2021). Essential Financial Tasks Done with Python. In Basic Python for Data Management, Finance, and Marketing (pp. 231-275). Apress, Berkeley, CA.

[39]. Guerard, J. B., Markowitz, H., & Xu, G. (2015). Earnings forecasting in a global stock selection model and efficient portfolio construction and management.

International Journal of Forecasting, 31(2), 550-560. https://doi.org/10.1016/j.ijforecast.2014.10.003

[40]. Y. Yang, "Market Forecast using XGboost and Hyperparameters Optimized by TPE," 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), 2021, pp. 7-10, doi: 10.1109/AIID51893.2021.9456538.

[41]. Gumelar, A. B., Setyorini, H., Adi, D. P., Nilowardono, S., Widodo, A., Wibowo, A. T., ... & Christine, E. (2020, September). Boosting the Accuracy of Stock Market Prediction using XGBoost and Long Short-Term Memory. In 2020 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 609-613). IEEE.

[42]. Dey, S., Kumar, Y., Saha, S., & Basak, S. (2016). Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting. PESIT South Campus.

[43]. Gong, J., & Sun, S. (2009, June). A new approach of stock price prediction based on logistic regression model. In 2009 International Conference on New Trends in Information and Service Science (pp. 1366-1371). IEEE.

[44]. Ali, S. S., Mubeen, M., Lal, I., & Hussain, A. (2018). Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX). Asian Journal of Empirical Research, 8(7), 247-258.

[45]. Hindrayani, K. M., Fahrudin, T. M., Prismahardi Aji, R., & Safitri, E. M. (2020). Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression. 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). doi:10.1109/isriti51436.2020.9315

[46]. Zhou, F., Zhang, Q., Sornette, D., & Jiang, L. (2019). Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. Applied Soft Computing, 105747. doi:10.1016/j.asoc.2019.105747.

[47]. Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167, 599-606. https://doi.org/10.1016/j.procs.2020.03.326.

[48]. Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. Int. J. Recent Technol. Eng, 8(1), 1224-1228

[49]. Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock Price Prediction Using Support Vector Regression on Daily and Up to the Minute Prices. The Journal of Finance and Data Science. doi:10.1016/j.jfds.2018.04.003.

[50]. Kao, L., Chiu, C., Lu, C., & Yang, J. (2013). Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing, 99, 534-542. https://doi.org/10.1016/j.neucom.2012.06.037

[51]. Y. Guo, S. Han, C. Shen, Y. Li, X. Yin and Y. Bai, "An Adaptive SVR for High-Frequency Stock Price Forecasting," in IEEE Access, vol. 6, pp. 11397-11404, 2018, doi: 10.1109/ACCESS.2018.2806180.

[52]. Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science, 170, 1168-1173. https://doi.org/10.1016/j.procs.2020.03.049

[53]. Y. Wang, Y. Liu, M. Wang and R. Liu, "LSTM Model Optimization on Stock Price Forecasting," 2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2018, pp. 173-177, doi: 10.1109/DCABES.2018.00052.

[54]. S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1643-1647, doi: 10.1109/ICACCI.2017.8126078.

[55]. Darrough, M. N., & Russell, T. (2002). A Positive Model of Earnings Forecasts: Top Down versus Bottom Up. The Journal of Business, 75(1), 127–152. https://doi.org/10.1086/323507.

[56]. Fölster, S. (1999). Social insurance based on personal savings. *Economic Record*, *75*(1), 5-18.

[57]. Yan, B., Stuart, L., Tu, A., & Zhang, T. (2020). Analysis of the Effect of COVID-19 on the Stock Market and Investing Strategies. Available at SSRN 3563380.