

**COFFEE PRICE PRIDITION USING MACHINE-LEARNING
TECHNIQUES: A CASE OF ETHIOPIAN COMMODITY
EXCHANGE (ECX)**

ABREHAM YEHEYIS GEBISA



**A Thesis Submitted to the Department of Computer Science and
Engineering**

School of Electrical Engineering and Computing

**Presented in Partial Fulfillment of the Requirement for the Degree
of**

Masters of Science in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

July 2019

**COFFEE PRICE PRIDITION USING MACHINE-LEARNING
TECHNIQUES: A CASE OF ETHIOPIAN COMMODITY
EXCHANGE (ECX)**

ABREHAM YEHEYIS

Advisor: DR. TEKLU URGESSA



**A Thesis Submitted to the Department of Department of Computer
Science and Engineering**

School of Electrical Engineering and Computing

**Presented in Partial Fulfillment of the Requirement for the Degree
of Masters of Science in Information Systems Engineering**

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

July 2019

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

APPROVAL SHEET

This Research Thesis entitled “COFFEE PRICE PRIDITION USING MACHINE LEARNING TECHNIQUES: A CASE OF ETHIOPIAN COMMODITY EXCHANGE DATA” has been read and approved as meeting the preliminary research requirements of the Department of Computing in partial fulfillment for the award of the degree of Master Science in Information Systems Engineering,

Adama Science and Technology University.

Adama, Ethiopia.

Submitted By:

_____	_____	_____
Student's name	Signature	Date

Approved by:

_____	_____	_____
Advisor	Signature	Date

_____	_____	_____
Department Chair	Signature	Date

_____	_____	_____
School Dean	Signature	Date

_____	_____	_____
Postgraduate Dean	Signature	Date

DECLARATION

I hereby declare that this MSc Thesis is my original work, has not been presented for a degree in any other university, and all sources of materials used for this thesis have been properly acknowledged.

Name: Abreham Yeheyis

Signature _____

This MSc Thesis has been submitted for examination with my approval as thesis advisor.

Name: Dr. Teklu Urgessa

Signature _____

Date of submission: July, 2019

DEDICATION

I dedicate my thesis work for my mom W/ro Mulumebet Mesfin Seyum

Acknowledgments

First and foremost, My Deepest Thank goes to the Almighty GOD, as He did much and He is always with me. And I would like to thank **Virgin Mary**, mother of Jesus Christ who has taught the deaf to speak and enables the listening ear to hear speech. I am deeply indebted and very glad to express my sincere gratitude and appreciation to my Advisor Dr. Teklu Urgessa for his important and constructive comments, criticisms and professional advice and doing all these on time from the beginning of the proposal writing to the completion of this thesis. Above all for treating me as a colleague. This thesis would have remained a dream had it not be for your support. I am indeed honored to work with you. My thanks also go to the former our program lecturer Ato Tinsae Gizachew for your support and help.

My great thank goes to the W/ro Bizuayehu Girma ECX marketing department manager for your real cooperation and help via Email, Phone and during our meeting, your smiley face and kind response encourages me to hall my work so thanks once again GOD bless you and your family.

I would like to give my thanks to all the respondents and informants for providing me the primary and secondary data. My thanks also goes to Mr. Addis and Mr. Leyewu officers of ECX Coffee Quality Assurance department.

I am grateful thanks My Family for encouraging and supporting me in my entire life journey, how this can be true without your support? GOD bless you all! Finally, to all My Classmates for a time we spend together and ASTU for giving me this scholarship to study my MSc and funding me to work on this thesis.

Table of Contents

Acknowledgments	v
List of Tables	viii
List of Figures.....	ix
Acronyms	x
ABSTRACT	xi
CHAPTER 1: INTRODUCTION.....	12
1.1. Background.....	12
1.2. Background of Ethiopian Commodity Exchange	15
1.3. Motivations	16
1.4. Statement of the Problem.....	16
1.5. Research Questions.....	17
1.6. Significance of the Study	17
1.7. Objectives	18
1.7.1. General objective.....	18
1.7.2. Specific Objectives.....	18
1.8. Scope and Limitation	19
1.9. Structure of the Thesis	20
CHAPTER 2: LITERATURE REVIEW.....	21
2.1. Machine learning	21
2.1.1. Process of Applying Supervised Machine Learning	22
2.1.2. Supervised Learning.....	26
2.1.3. Regression Models	27
2.1.4. Prediction.....	28
2.2. Commodity Exchange.....	29
2.3. Applications of Machine Learning in ECX	29
2.4. Related works	31
CHAPTER 3: METHODOLOGIES.....	34
3.1. Data Source.....	34
3.2. Data Collection Method and Analysis.....	34

3.3.	Data Understanding	35
3.4.	Algorithms	36
3.5.	Tools	38
CHAPTER 4: EXPERIMENTATIONS		43
4.1.	Overview	43
4.2.	Packages	44
4.3.	Data preparation	44
4.4.	Create a Validation Dataset	45
4.5	Build Models	46
CHAPTER 5: EVALUATION AND DISCUSSION		48
5.1.	Select Best Model	48
5.2.	Making Predictions in LSTM	50
CHAPTER 6: CONCLUSION AND FUTURE WORK		59
6.1.	Conclusion	59
6.2.	Future Work	60
References		61
APPENDIX A:		66
APPENDIX B:		68
APPENDIX C:		69

List of Tables

Table 1.1 ECX Commodities Records	19
Table 2.1 Methods to handle missing data	24
Table 2.2 Related work.....	33
Table 3.1 Elements of SciPy.....	40
Table 4.1 Sample of the dataset.....	45
Table 5.1 Algorithms Performance Result	49
Table 5.2. Forecasted confidence table	58
Table 6.1. Coefficient of ARIMA	67
Table 6.2. Training set error measures	67

List of Figures

Figure 1.1 Ethiopian Coffee	13
Figure 1.2. Ethiopian Coffee Map	14
Figure 1.3. Organization of the thesis.....	20
Figure 2.1 Process of Applying Supervised ML to A Real-World Problem.....	26
Figure 5.1 Actual trained and predicted graph	50
Figure 5.2. Predicted graph.....	56
Figure 5.3. Yergachefe special	51
Figure 5.4. Export Jimma	51
Figure 5.5. Gujjii Special.....	52
Figure 5.6. Harar Export.....	52
Figure 5.7. Local by product	53
Figure 5.8. Local byproduct	53
Figure 5.9. Unwashed export.....	54
Figure 5.10. Washed Specialty	54
Figure 5.11. Washed Specialty	55
Figure 5.12. Washed unwashed Export.....	55

Acronyms

AR	Auto-Regressive
ARMA	Auto-Regressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
ANN	Artificial Neural Networks)
BPNN	Back Propagation neural network
CART	Classification and Regression Trees
CHAID	Chi-squared Automatic Interaction Detection
DTR	Decision Tree Regression
ECX	Ethiopian Commodity Exchange
GA	Genetic Algorithms
HIPC	heavily indebted poor countries
IMF	International Monetary Fund
ICO	International Coffee Organization
IS	Information Systems
ID3	Iterative Dichotomize
KNN	K-Nearest Neighbor
LR	Linear Regression
LSTM	Long Short Term Memory
MSE	Mean Square Error
MARS	Multi-Adaptive Regression splines
ML	Machine Learning
OGM	Oxfam Great Britain
RNN	Recurrent neural network
SOM	Self Organizing Maps
SVR	Support Vector Regression
TSF	Time Series Forecasting
VAR	Vector Autoregression

ABSTRACT

Commodity market decision making is a very challenging task because it is difficult to make a good decision in a timely manner even if the decision is made in time may not found to be the right one, it is significant to course correct and moves on. Forecast about commodity market with high accuracy movement produce profits for users of the market.

In this study machine-learning prediction model is developed by comparing different prediction algorithms, the one with high performance to predict the un-for-seen data is selected. An approach to forecast future market track of the Ethiopian Commodity Exchange (ECX) data has been introduced based on chart patterns recognition by using machine learning Regression model. Models are built through different algorithms including Linear Regression, decision tree Regression (CART), SVR, KNN, and LSTM. Where the Results -6614722.039, 0.909470, 0.910699, -0103426 and 0.91523 are obtained respectively, compared and discussed in details in chapters.

Important patterns to support decision making in commodity trading had been found out. In order to visualize the result, ARIMA visualization technique is also introduced Time series prediction is conducted and a fit span of time for the commodity market data is examined. After a series of algorithms comparison and experimental analysis, LSTM is used to predict the trade data. This paper attempted to predict the future market direction based on ECX historical data. to decide whether to buy or hold the commodity using machine-learning techniques.

Key terms: Closing Price, LSTM, ARIMA, Time series data, prediction

CHAPTER 1: INTRODUCTION

This chapter presents introductory part comprising of the background of the study, motivations, statement of the problem, basic research questions, objectives, significance, limitation and scope of the study, and organization of the thesis.

1.1. Background

Ethiopia is located in East Africa. According to (Gregory et al). The total area of the country is around 1,104,300 sq. km. Ethiopia's poverty-stricken economy is based on agriculture; its economic building is unique in Africa. It has no oil or mining zone and virtually no isolated investment. With the republic lacking basic development components, the Ethiopian economy is highly in need of agriculture [1]. In 2012, the country qualified for international aid under Heavily Indebted Poor Countries (HIPC) initiative. Since then, the International Monetary Fund (IMF) is the leading foreign aid provider for the country [2]. Accounting for about forty-fifths of Gross Domestic Product (GDP), and 85% of total employment. the agricultural sector suffers from frequent drought and poor cultivation practices. Coffee is important to the Ethiopian economy with exports of some \$782 million in 2017.

About 2.5 billion cups of coffee is drunk every single day [3]. Most of peoples are very familiar and may not feel comfortable without a cup of coffee in the morning. Nevertheless, some of us have lack of concepts about the origin of coffee. According to the ICO, the scattering of coffee entirely growing and drinking began in the Horn of Africa, where according to a traditional historical tale; the province of Kaffa is the area that coffee trees originated.



Figure 1.1 Coffee

It is recorded that slaves were taken from present-day Sudan into Yemen and Arabia through the great port of its day, Mocha. It is assumed that coffee was nurtured in Yemen by the 15th century or earlier. In a shot to stop its cultivation elsewhere, the Arabs obligatory a ban on the export of fertile coffee beans, a restriction that was eventually avoided in 1616 by the Dutch, who brought live coffee plants back to the Netherlands to be grown in greenhouses [4]. The most well recognized coffee-producing region in southwestern Ethiopia (see Figure 1.2. Ethiopian Coffee Map). Ethiopia is a very complex coffee origin that means in Ethiopia the coffee cultivating area are producing different coffee types.

The best Ethiopia dry-processed coffee (Harrar found in eastern highland) tends to be medium-bodied and brilliantly acidic with rough, fruity or winy tones. The best washed Ethiopian coffee (Yirgacheffe, Sidamo found in the southern part) is light-bodied but explosive with multifaceted

[illegible]

14

1.2. Background of Ethiopian Commodity Exchange

Commodity exchanges are structured markets, which functions as a means for risk controlling and operation costs reduction for both buyers and sellers. Historically, commodity exchanges date back to the middle ages. Until recently, commodity exchanges were rigorously confined to agrarian countries, but with the advent of reasonable technology, many developing countries in Africa have been keen on establishing one. The wave began with Zambia and Zimbabwe in the 1990s and followed by Kenya and Uganda [6]. Then Ethiopia joins the group of exchanges in Africa by the title called the Ethiopian Commodity Exchange (ECX).

The Ethiopia Commodity Exchange (ECX) is a commodities exchange established April 2008 in Ethiopia. In Proclamation 2007-550, which created the ECX, its stated objective was "to ensure the development of an efficient modern trading system" that would "protect the rights and benefits of sellers, buyers, mediators, and the general public [7].

To curb the situation, the government of Ethiopia together with its development partners The World Bank, International Monetary Fund, Oxfam GB, among others launched the Ethiopian Commodity Exchange (ECX) with the aim of creating a well-organized, clear and orderly marketing system to serve the needs of buyers, mediators and, Sellers and promote increased market contribution of Ethiopian small-scale manufacturers. It further planned to create an integrated trading floor for sellers and buyers. The new system expected to develop more secure and consistent schemes for treatment, storing, grading, among other facilities, inspire risk-free payments, and offer a goods transfer system to settle transactions [8], [9]. It also intended to build trust and transparency among all market actors through the dissemination of market information and clearly defined rules for trading and warehousing, as well as the eventual provision of internal dispute settlement services.

The commodity market is essentially a non-linear system that is very hard to model with any sensible accuracy. Farmers have been trying to find a way to predict commodity prices and productivity. To achieve those objectives, some researchers used the techniques of elementary analysis, wherever commerce rules are advanced based on the information related to economic

science, industry, and company. Some authors said that fundamental analysis assumes that the price of a commodity depends on its intrinsic value and expected return on investment.

1.3. Motivations

The motivated is that as we know, all the information about today commodity trading of all specific trader the price is predictable via orally or theoretically which leads to wrong conclusion Thus, if we can obtain just a piece of partial information, we can expect to improve the current prediction lot. With the growth of the Internet, social networks, and online social interactions, getting daily user predictions is a feasible job. Thus, my motivation is to design a prediction model incorporating historical data and user's predictions to make a stronger model that will benefit everyone.

1.4. Statement of the Problem

Currently, predictions are based on absolute theories and behaviors this leads to unavailability of Prediction tracking nothing of the request provide consistent predictions for the coffee scheduling market / Prices. If you are scheduling to capitalize on coffee X today, then there is no application can forecast its behavior at that instance No Word on the performance of the Forecasts. There is no application accessible today that can offer us assurance on their forecasts or at least confirm the percent exactness that their system can deliver. Today, the outstanding challenge of consuming a databank is to produce useful rules from raw data in a database for manipulators to make conclusions, and these rules may be secreted deeply in the raw data of the database. Usually, the method of rotating data into understanding relies on manual exploration; this is becoming unreasonable in many fields as data dimensions raise exponentially.

The problem with predicting coffee prices is that the volume of data is too bulky and vast. but This study uses the machine learning techniques or methods which is the Regression and prediction approach on the historical data available on ECX database and tries to help the farmers and The Organization by predicting the future market directions to build their decision on to sell that commodity in order to achieve profit.

For the organization, knowing the future market direction may help them to make the farmers more productive by providing the right information at the right time this can make society strong economically.

1.5. Research Questions

1. What are the machine learning algorithms approaches being appropriate to predict the futures markets?
2. For the next three years' Ethiopian coffee market price will go up, go down or stay as it is?

1.6. Significance of the Study

The main significance of this work is to predict the ECX data that could outperform the buy-and-hold strategy and lower risk. This strategy ought to be effective for any market security, given a reasonable amount of historical training data. The second goal of this work is to learn about computational finance techniques. As the world becomes more a connected because of globalization and the web, the power to trace and live these connections will increase. In today's society, more data is being recorded than ever, yet the information is gleaned from these records is in its infancy. Learning the techniques that can be used to parse and classify big data sets was the ultimate objective, due to the importance it will play in the future of the world.

1.7.Objectives

1.7.1. General objective

By analyzing and preprocessing the Ethiopian Commodity Exchange Data (Closing Price) prediction using machine learning techniques

1.7.2. Specific Objectives

- Review literature on related to machine learning and prediction is conducted
- Data preparation
- Machine –learning for prediction algorithms are compared
- The performance of the algorithms is performed
- The selected algorithms perform the prediction.
- The results are interpreted
- Further recommendation for future work is forwarded

1.8. Scope and Limitation

As mentioned in the above the aim of this study is predict the continuous values for ECX historical data which include from May 31-2011 – April 02-2018. A data after and before the stated data is excluded in our study and the ECX trade market database have seven differ crops or commodities as shown in Table 1.1

No .	<i>Commodities Name</i>	<i>No of total records in DB</i>	<i>no of commodities record</i>	<i>In percent</i>
1	Coffee	139,556	100,323	71.8%
2	Green mung bean		168	0.12%
3	Haricot beans		8006	5.7%
4	Maize		88	0.06%
5	Pea beans		659	0.5%
6	Sasame		30299	21.7%
7	Wheat		13	0.009%

Table 1.1 ECX Commodities Records

Our study focuses only on the coffee sector because Ethiopia is known to be the origin place for coffee. It is the major export commodity cultivated in the country. Coffee grown in Ethiopia is recognized all over the world for its first-rate quality and flavor. Today, Ethiopia stands as the biggest coffee producer and exporter in Africa and amongst the leading in the world [6] and the second reasons we keep our scope in coffee is from a total of 139,556 records 100,323 records taken, 71.8% are coffee records where the other six commodities takes 28.2% of total recodes. During experimentation, we analyze the data by modeling different algorithms and show their results, we take one best prediction algorithms, and we recommend it with its result.

the limitation in this area is, its difficult to get articles to in reference in countryside The performance of machines or computer to process the data takes more than weeks to get result this makes challenged to get what we need and the power interruption during job, availability of internet access and budget are the major limitations during our thesis work.

1.9. Structure of the Thesis

In the below Figure, we explain the frame of the thesis in a simple way to give the reader a general appearance. Chapter 1 will give the reader information about the problem, background, and motives for study. This chapter is also included the significance of the study. Chapter 2 reviews literature that has studied about commodity price predictions that have done ever in the past. Chapter 3 clarify methodologies for continuous data regression and predictions. Chapter 4 models the algorithms for experimentations and compares the results obtained. The results of the algorithms discussed and high-performed algorithms selected in Chapter-5. In the last chapter recommendations and conclusions are takes place and by stressing the thesis statement, and to leave some final ideas to the reader.

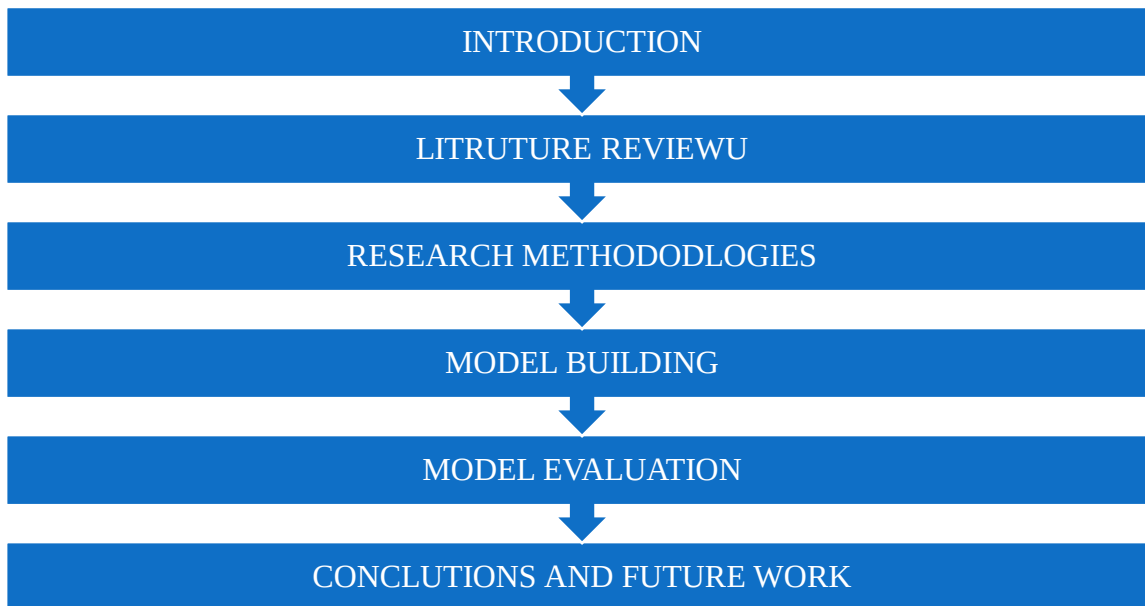


Figure 1.3. Organization of the thesis

CHAPTER 2: LITERATURE REVIEW

2.1. Machine learning

Machine learning studies how to automatically learn to make an accurate prediction based on past observation. There are several applications for machine-learning(ML), the most significant of which is data analyzing. People are usually susceptible to create mistakes throughout analyses or, possibly, once attempting to determine relationships between multiple options. This makes it troublesome for them to search for solutions to certain issues. Machine learning will usually be with success applied to those issues, to increase the efficiency of systems and the designs of machines. Every instance in any dataset utilized by machine learning algorithms is depicted as exploitation a similar set of options. The features may be continuous, categorical or binary. If instances are given with known legendary (the corresponding correct outputs) then the training is named supervised, in distinction to unsupervised learning, where instances are unlabeled. By applying these unsupervised (clustering) algorithms, researchers hope to discover unknown, but useful, classes of items [10]. Another kind of machine learning is reinforcement learning [11]. The coaching info provided to the learning system by the atmosphere (external trainer) is within the style of a scalar reinforcement signal that constitutes a measure of how well the system operates. The learner is not told that actions to require, however rather should discover that actions yield the most effective reward, by trying each action in turn.

Inductive machine-learning is the method of learning a group of rules from instances example in a training set or more, generally speaking, creating a classifier that can be used to generalize from new instances.

2.1.1. Process of Applying Supervised Machine Learning

The process that we are applying in supervised ML to a real-world problem is described in Figure 2.1 to discuss each of them the first step is collecting the dataset. If an essential expert is available, then we could suggest which fields (attributes, features) are the most informative. If not, then the most effective methodology is that of “brute-force,” which means measuring everything available in the hope that the right (informative, relevant) features can be isolated. However, a dataset collected by the “brute-force” methodology is not directly appropriate for induction. It contains in most cases noise and missing feature values and therefore requires significant pre-processing [12].

The second step is Data preparation and data pre-processing. What can be wrong with data? There is a hierarchy of problems that are often encountered in data preparation and preprocessing like Impossible and unlikely values have been inputted, missing values and irrelevant input features are present in the data at hand. Impossible values ought to be checked for by the information handling software, ideally at the point of input so that they can be reentered. These errors are generally straightforward, such as coming across negative prices when positive ones are expected. If correct values cannot be entered, the problem is converted into a missing value category, by simply removing the data.

The data preparation and data preprocessing Depending on the circumstances, researchers have a variety of ways to decide on from to handle missing data [13] & [14] Have introduced a survey of up to date techniques for outlier detection. These researchers have identified the techniques’ advantages and disadvantages. Instance choice is not, only used to handle noise however to address with the impracticability of learning from terribly massive datasets. Instance selection in these datasets is an optimization problem that attempts to maintain the mining quality while minimizing the sample size [15]. It reduces knowledge and allows a data-mining algorithm program to operate and work effectively with terribly massive datasets.

There is a variety of procedures for sampling instances from a large dataset [16]. Random sampling, which selects a subset of instances randomly and stratified sampling, which is applicable when the class values are not uniformly distributed in the training sets.

Instances of the minority class are selected with greater frequency in order to even out the distribution. Other techniques for handling imbalanced datasets can be found in [17]. Incomplete knowledge is associate degree inescapable drawback in handling most universe knowledge source.

Generally, there are some vital factors to be taken under consideration once processing\ unknown feature values. One of the foremost vital ones is that the supply of “unknown-ness”:

- (i) A value is missing because it was forgotten or lost;
- (ii) A certain feature is not applicable for a given instance (e.g., it does not exist for a given instance);
- (iii) For a given observation, the designer of a training set does not care about the value of a certain feature (so-called “don’t-care values”).

Depending on the circumstances, investigators have a number of methods to choose from to handle missing data according to [18].

Methods	Descriptions
The technique of ignoring cases with unknown feature values	This technique is simplest: ignore any instances that have a minimum of one unknown feature worth.
Most common feature value	The value of the feature that happens most frequently is chosen to be worth all the unknown values of the feature.
Most common feature value in the class	This time the value of the feature, which occurs most commonly within the same class is selected to be the value for all the unknown values of the feature.
Mean substitution:	Researchers substitute a feature’s mean value (computed from available cases) to fill in missing data values on the remaining cases. A more sophisticated solution than using the “general” feature means is to use the feature mean for all samples belonging to the same class to fill in the missing value

Regression or classification methods:	A regression or classification model based on the complete case data for a given feature is developed. This model treats the feature because of the outcome and uses the opposite options as predictors.
Hot deck inputting:	The most similar case to the case with a missing value is identified, and then a similar case's Y value for the missing case's Y value is substituted.
Method of treating missing feature values as special values:	“Unknown” itself is treated as a replacement value for the choices that contain missing values. Feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible [19]. This reduces the spatial property of the information and permits data processing algorithms to control quicker and additional effectively.

Table 2.1 Methods to handle missing data

Generally, options are categorized as relevant: These are features have an influence on the output. The rest cannot undertake their role. Unrelated features are outlined as those features not having any influence on the output. Their values could be generated at random and not influence the output. In addition, Redundant: A redundancy exists at any time a feature can take the role of another (perhaps the simplest way to incur model redundancy).

Feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible [19]. This reduces the spatiality of the info and allows data processing algorithms to control quicker and a lot of effectively. The fact that several options rely upon each other usually unduly influences the accuracy of supervised Machine learning classification models. This problem can be addressed by constructing new features from the basic feature set [20]. This technique is called feature construction/transformation. These new generated options could cause the creation of more concise and correct classifiers. In addition, the invention of significant features contributes to the higher quality of the created classifier and an improved understanding of the learned thought. [10]

In addition, the other the main and critical step in supervised machine learning techniques are Algorithm selection. The choice of that specific learning algorithm we must always use may be an essential step. The classifier's evaluation is most often based on prediction accuracy. There are at least three techniques that are used to calculate a classifier's accuracy. One technique is to separate the data set into training and estimating, the training set may use two-thirds for training and the other third for estimating performance. In another technique, known as cross-validation, the training set is divided into mutually exclusive and equal-sized subsets and for each subset, the classifier is trained on the union of all the other subsets. The average of the error rate of every set is thus associate degree estimate of the error rate of the classifier. Leave-one-out validation may be a special case of cross-validation. All test subsets consist of a single instance. This type of validation is, of course, more expensive computationally, however helpful once the foremost accurate or correct estimate of a classifier's error rate is needed.

If the error rate evaluation is unsatisfactory, a variety of factors must be examined: perhaps relevant features for the problem are not being used, a larger training set is needed, the dimensionality of the problem matter is simply too high, the selected algorithm is inappropriate or parameter tuning is needed. In our study, A standard technique examination of supervised ML algorithms is to perform statistical comparisons of the accuracies of trained classifiers on specific datasets [21].

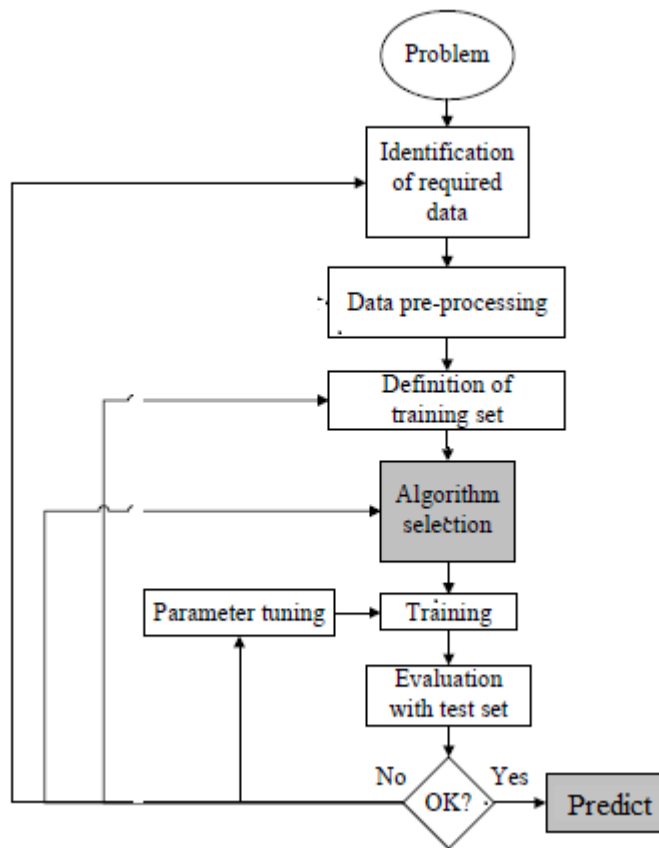


Figure 2.1 Process of Applying Supervised ML to A Real-World Problem

2.1.2. Supervised Learning

Supervised learning explanations for a numeral research activity in machine learning and many supervised learning methods have found application within the process of multimedia system content. The shaping characteristic of supervised learning is that the accessibility of annotated training data. The name appeals to the concept of a supervisor that instructs the learning system on the labels to go along with training examples. Normally, these labels are class labels in classification problems. Supervised learning algorithms induce models from these training data and these models can be used to classify other unlabeled data.

According to (Kotsiantis et al) supervised machine learning is to build a concise model of the distribution of class labels in terms of predictor features [21]. The resulting classifier is then used to assign class labels to the testing instances wherever the values of the predictor options are

framed, but the value of the class label is unknown. This paper describes numerous supervised machine learning classification methods. Of course, a distinct article cannot be an entire review of all supervised ML classification algorithms also known induction classification algorithms, yet we hope that the references cited will cover the major theoretical issues, guiding the reader of this study in interesting research directions and suggesting possible bias combinations that have yet to be explored.

2.1.3. Regression Models

The prediction of continuous values can be modeled by statistical techniques of regression. Many problems can be solved by linear regression, and even more, can be tracked by applying transformations to the variables so that a nonlinear problem can be converted to a linear one [22]. [23] Stated that regression models are the leading predictive models. The regression models are linear or nonlinear regression models for modeling continuous response, and logistic regression models for modeling discrete choice response [22]. These regression models take into account the effect of time. The most common types of time-dependent models are survival models, time-series models, non-stationary models in which the course of the event spell, time-related censored and truncated data, and others [22].

As the sometime Classification is one the major machine learning approach and its process of identifying new objects or events as belonging to one of the known predefined classes, classifying future or unknown patterns in systematic approach it has a mathematical technique like decision tree, linear programming, and statics [24]. The aim of CART is to construct a model by discovering the influence of independent variables on the dependent variable so the unclassified record can be classified. In [25] time series forecasting models are constructed to forecast the future market index by using a back propagation neural network. Models are built in many workflows to look at an appropriate span of your time commodity market data. After that, a unique approach for predicting future market direction is proposed based on chart patterns recognition by victimization data processing classification. Classification methods include SVR, decision tree, LR and k-nearest neighbors; the model aims to predict whether the market index will go up, go down or stay in the next day.

2.1.4. Prediction

The prediction uses various methods of CART approaches such as linear regression, Decision tree regression, and k-nearest neighbors (k-NN) [26]. There are two ways to predict the Commodity [27]

- i. To predict the commodity price and
- ii. To generate the buy or sell signal for the commodity.

Buy or sell signal generation has been used for commodity prediction. There are two types of analysis for buy/sell signal generation Fundamental Analysis and Technical Analysis. As we discussed in chapter one Fundamental analysis is based on company's profile and assets in the market, while technical analysis totally depends on company's stock price in the market, and volume trade on that particular price. Their model has been developed based on technical analysis. We designed or build a different model based on machine learning application techniques to predict the future commodity price.

In other studies, financial time series prediction is an important subject for many financial analysts and researchers as accurate forecasting of different financial applications play a key role in investment decision making. Financial time series prediction is a very important subject for several monetary analysts and researchers as the correct prediction of various monetary applications play a key role in investment decision making. [28] Technical analysis undertakes that previous values of the commodity have an influence on the future development of the market. In technical analysis, technical indicators created by special formulas are used to forecast commodity trends. In previous times, complex machine-learning methods had been presented for time series prediction. Among them, artificial neural networks (ANNs) [29], Support vector machines (SVMs) [30], Genetic Algorithms [31] and Self Organizing Maps (SOM) [32] are the most communal used machine-learning methods in financial time series prediction. Since the first 1990s, ANNs become the most popular machine learning techniques used as different from standard statistical models in financial time series analysis and prediction. In addition, the prediction performance is evaluated using the mean square with error (MSE)

2.2. Commodity Exchange

A commodity exchange is a centralized market place where sellers and buyers meet to transact commodities in an organized fashion with certain clearly specified and transparent “rules of the game”, allowing for maximum effective competition among buyers and sellers [33].

The practicality of a commodity exchange lies in its official capacity to eliminate or reduce the high transaction costs often faced by entities along commodity supply chains in developing countries [34]. A commodity exchange reduces dealing prices by giving services at a lower price than that participant would incur if they were acting outside an institutional framework. These can include – but are not limited to – the costs associated with finding a suitable buyer or seller, negotiating the terms and conditions of a contract, securing finance to fund the transaction, managing credit, cash and product transfers, and arbitrating disputes between contractual counterparties [33]. By doing so, commodity exchanges become relevant to the situation faced by commodity sectors in developing countries in a way that by filling the gap arises due to the structural changes after liberalization that has seen a substantial withdrawal of government support for the agriculture sector [35].

2.3. Applications of Machine Learning in ECX

In this study, we focus on examining Ethiopian coffee prices have become more integrated to world prices after prediction. The Organization is currently trading seven commodities: Coffee, Sesame, haricot Beans, maize, Wheat, green mung bean and pia beans. The strategic importance of coffee to the economy, nevertheless, is evidently reflected in the Exchange’s trading sessions and the volume of the transaction.

In literature, we learn Ethiopia as the origin of coffee Arabica. The country produces a wide variety, highly differentiated, most of which are shade-grown by small farmers [36]. Coffee has accounted for an average of 60% of the total export earnings for the past five decades [37]. We base our analysis on the coffee sector because coffee is a major export commodity and an important source of employment in Ethiopia. Coffee is also the most traded agricultural commodity in the

ECX both in terms of volume and value, and significant reforms have occurred in the coffee sector in the past years.

Therefore, these sectors have to support different methods in our study we try to predict the price of the coffee by using machine-learning applications in different aspects or attributes. The output will help the exchanges do more and the farmers to produce more and to be beneficiary in terms of cost and human resource usage. Certainly, a higher price interdependence between domestic and international prices may signal that part of the objectives of the ECX are being achieved in terms of stronger linkages with external markets and improved market information but does not necessarily mean that it is a welfare-improvement system, particularly among the most vulnerable groups of farmers. Higher volatility interactions from international to local prices may also have serious implications for farmers and low-income consumers by making them more vulnerable to international shocks. [38]

2.4. Related works

Title and Year	Authors	Methodologies	Finding	Gaps
Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm, in March 2013.	Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi Mohammed K. Ali Shatnawi	The <i>k</i> NN algorithmic rule methodology is implemented on the stock data. Also, mathematical calculations and visualization models are provided	According to the results, the kNN algorithm is robust Algorithms than Linear regression algorithm. the prediction results were related to actual stock prices.	error ration in Total squared RMS error is high and Linear regression result is not stated and doesn't predict the future market price
Classification and Prediction in data mining with neural networks, 2003	Serhat OZEKE and nur OSMAN	In this study, they realized the classification and prediction model with back propagation, which is the most popular neural network learning algorithm	They mention In their paper Neural Networks (NN) are drawn in data mining for classification and prediction. Backpropagation is used as a learning algorithm. In their simulation, financial data set is evaluated. And they expectations of bank results and their proposed	Only one algorithm is implemented in the paper

			Neural Network results are compared and some differences are obtained.	
Stock Market Prediction using Artificial Neural Networks. Case Study of TAL1T, Nasdaq OMX Baltic Stock,2015	Hakob GRIGORYAN	Technical analysis with statistical and ML techniques have been implemented and develop some strategies and methods to be helpful for financial time series forecasting. The statistical methods are autoregressive conditional heteroskedasticity (ARCH), autoregressive integrated moving average (ARIMA) & Smooth Transition Autoregressive (STAR) model	Combined prediction model, based on artificial neural networks (ANNs) with main principal component analysis (PCA) for financial time series forecasting. The prediction acts are evaluated using the standard evaluation measure called mean squared error (MSE).	is not include more stock market influencing factors, based specifically on fundamental and technical analyses compared with other prediction models

Application of Data Mining Technology in Predicting The seroprevalence of HBV, HCV, HIV: The case of national blood bank of Ethiopia. 2011	Haftom Gebregziabher	Used J48 Decision tree & naïve algorithms applied in weka	The model has generated class predictions for all the predefined classes of the safe and unsafe groups	In order to enhance the performance of the predictive model, proper research needs to undertake.
--	----------------------	---	--	--

Table 2.2 Related work

CHAPTER 3: METHODOLOGIES

3.1. Data Source

The research data used in this study is historical data taken from the Ethiopian Commodity Exchange full backup of the database for the only coffee sector. The whole data set covers the period from May 31, 2011, to April 02, 2018 100,323 daily observations. Data collection was done through primary and secondary source. Primary data sources included key informants for each attribute. Secondary data sources mainly covered ECX publications, technical document, and annual reports of the Organization. Primary data covered different sources and provided an essential preparation for the interviews. Secondary data helped to crosscheck official information, learn about major events, technical details, historical decisions, and main organizational players and roles. They conjointly supported the exploring of explicit responses throughout interviews.

3.2. Data Collection Method and Analysis

The main data techniques used in this research study were semi-structured interviews with trade market managers and some coffee experts in coffee quality assurance department at Ethiopian Commodity Exchange head office located at Addis Ababa Ethiopia. Personal interviews brought about one in all the foremost necessary and valuable sources of information the social nature of information systems has led many Information Systems researchers to adopt research approaches that focus primarily on human interpretations and meaning [42]. Interpretive studies advocate a relativistic understanding of the phenomena being studied [43]. Interpretive researchers see the pursuit of meaning and understanding as subjective and knowledge as a social construction [44]. They examine the social reality and subjective meanings held by the individual by eliciting and observing what is important and necessary to them. They are not reporting facts, but their interpretations of other people's interpretations [42]. There is no rigid separation between data assortment and analysis. With the intention that the results of the analysis can facilitate guide the subsequent collection of data. The cycle is repeated and theory is elaborated and checked as the process continues. When conducting informative analysis, it is generally accepted that researchers should interact directly and intensively with the subjects of

their research over a period. The interviewees were chosen for his or her connection to the abstract queries instead of their representativeness. The total range of respondents to interview was reached heuristically, i.e., the decision to stop adding respondents was taken when nothing new was being learned from the interviews and a state of theoretical saturation was achieved. The interviewees were selected based on their closeness to the topics of the study research and their levels of experience in management and organizational issues. It was deemed essential not to limit interviewees to IS/IT staff. All the interviews were conducted in Amharic and for the purpose of this thesis were translated into English by the researcher himself. The information gathered from these interviews was subjective, although an attempt was made to present an account from various perspectives and levels within the organization.

3.3. Data Understanding

This step includes collecting sample data and deciding which data, including format and size, are needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, the plausibility of attribute values, etc. this step includes verification of the usefulness of the data. Once the business problem and the objective are clearly defined, the next step is to take time to understand what kind of data is available and which data is the most appropriate and has the most predictive power that can address the objective. Usually, the data-understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. In order to understand the nature of the data and the attributes at the ECX database brief explanation was made by the market experts. The information about each commodity is stored in a database. Initially, information about the commodity is recorded when the individual arrives at the warehouse which includes the origin of commodity and the information is recorded in the information collection sheet. In fact, all information clerk personnel at the reception area are provided with a centralized form or record format that should be filled when a seller is to offer a commodity and this helps to maintain consistent information. The dataset consists of a total of 139,556 records and 14 attributes. These attributes are Year, Month, Trade Date, Commodity, Commodity Type, Commodity Name, Description, Symbol, Origin, Closing Price, Ton, Total Value, Warehouse, Production Year. We are working with selected 100323 records and 9 attributes only. In

Ethiopian Commodity Exchange office, they trade seven different commodities those are: coffee, green mung beans, haricot beans, maize, pea beans, sesame, and wheat. However, this research paper will focus on one commodity type its Coffee. Coffee has a total 100,323 records after preprocessing the data by the techniques mentioned in the above out of 100,323 records 75 records removed from the dataset because of outliers missing values and redundancies.

3.4. Algorithms

3.4.1. Linear Regression

Machine learning, more specifically the field of predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. In applied machine learning, we will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends. As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical procedure and a machine-learning algorithmic rule. Linear regression is an attractive model as a result of the representation is so simple. The illustration is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, each the input price (x) and also the output value are numeric.

3.4.2. K-nearest neighbor

K-nearest neighbor technique is a machine learning algorithm that is considered as simple to implement [39] According to [31] the K-NN classifiers opening values, high values, low values, closing values The classifier model considers gap worth, high value, low value, closing value and returns of the Market. Index as independent variables and the next day's class as the dependent variable. Returns for a day is calculated as

$$Returns = \frac{v_t - v_{t-1}}{v_{t-1}}$$

Where v_t is the closing value of the index on the current day and v_{t-1} is the closing value of the index of the previous day. If the coming days' come back is positive, the next day's class is category is assessed as "bull" otherwise "bear". The k-NN algorithmic rule identifies 'k' nearest neighbor's within the training data set in terms of the euclidian distance with respect to the day for which price index movement is to be classified. Once nearest neighbors are identified, the classification for that day is simply the majority of the next day's classes of those identified neighbors. The k-NN algorithms try out various k values in the training dataset and find the optimal value of k that produces a good result. Then this classifier model with the optimum value of 'k' is applied to the test data set for categorizing the next day's class. In other words, it simply tries to predict the index association for the next day as either 'bull' or "bear". The output of the classifier is compared with the particular categories of the test data set for justification.

3.4.3. Support Vector Regression

Support vector machine (SVM) analysis is a popular machine-learning tool for classification and regression, first identified by [48]. SVM regression is taken into account as a statistic technique because as a result of it depends on kernel functions. Statistics and Machine Learning chest implement linear epsilon-insensitive SVM (ϵ -SVM) regression, which is also known as L1 loss. In ϵ -SVM regression, the set of training data includes predictor variables and observed response values. The goal is to find a function $f(x)$ that deviates from in by value no greater than ϵ for each training point x , and at the same time is as flat as possible [49]. Support Vector Machine can even be used as a regression method, maintaining all the most options that characterize the algorithmic rule (maximal margin). The Support Vector Regression (SVR) uses similar principles because of the SVM for classification, with only a few minor differences. First, because the output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM, which would have already requested from the problem. Besides this reality, there is also a more complicated reason; the algorithm is more complicated therefore to be taken into consideration. However, the most plan is often the same: to reduce error, individualizing the hyperplane, which maximizes the margin, keeping in mind that part of the error is tolerated.

3.4.4. Decision Tree Regression (CART)

Decision Tree algorithm goes to the family of supervised machine-learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too [50].

The general reason for using DT is to create a training model that may use to forecast grouping or value of target variables by learning decision rules inferred from past data (training data).

3.4.5. Time Series Forecasting (TSF)

Time series forecasting [54] is a process of analyzing time series data and predicting the future outcome. It uses statistical methods to model and explains a time-dependent series of data points. TSF may be a widely used technique in monetary prediction.

3.4.6. Long Short-Term Memory (LSTM)

Time series prediction problems are a difficult type of predictive modeling problem. Unlike regression prognostic modeling, statistic or time-series also adds the quality of a sequence dependence among the input variables. A powerful type of neural network designed to handle sequence dependence is called recurrent neural network. The Long Short-Term Memory network or LSTM network is a type of recurrent neural network used in deep learning because of terribly massive architectures are often with success trained.

In a study, we validate the forecasting reliability of our prediction by developing LSTM networks in Python using the Keras deep learning library to address a demonstration time-series prediction problem.

3.5. Tools

In machine learning, there are various techniques followed in order to achieve a given machine-learning goal. Since this is a study, which attempts to build a model that predicts the coffee prices, there is a need for discussing the techniques used in classification and building a

machine-learning model. For this reason, the detailed discussion of the techniques is followed in each subsection.

3.5.1. Overview of Anaconda

Anaconda is a tool we are used during our experimentation its open source data science package with a community of over 6 million users [45] it is easy to download and install, and it is supported on Linux, MacOS, and Windows. Anaconda is free Python distribution, including over 195 of the most popular Python packages for science, math, and data analysis. Anaconda includes Python 2.7/Python 3.7 and cross-platform Python packages, as well as tools for integration with Excel. Python and Anaconda support a spread of processes inside the scientific data advancement, from getting information or data, manipulating and method data, and visualizing and communicating research results. Because Python can be used in a wide variety of applications, even beyond scientific computing, users can avoid having to learn new software or programming languages when new data analysis needs arise. Python's open source availability enhances research reproducibility and enables users to connect with a large community of fellow users.

3.5.2. Python

For many individuals, the Python programming language has strong attractiveness. Since its first appearance in 1991, Python has become one of the most popular understood programming languages, at the side of Perl, Ruby, and others. Python and Ruby become mainly dominated since 2005 more or less for building websites using their various web frameworks, like Rails (Ruby) and Django (Python). Such languages are often called scripting languages [46]. Why we need to research in Python is has gone from a bleeding-edge or “at your own risk” scientific computing language to one of the most important languages for data science, machine learning, and general software development in academia and industry.

For examining data and collaborative computing and data visualization, Python will inevitably draw comparisons with other open source and commercial programming languages and tools in wide use, such as R, MATLAB, SAS, Stata, and others [45]. In recent years, Python’s enhanced

support for libraries such as pandas and scikit-learn had made it a popular choice for data analysis tasks. Combined with Python's overall strength for all-purpose software system engineering, it is an excellent option as a primary language for building data applications.

I. Libraries

SciPy stands for Scientific Python. SciPy is built on NumPy. It is a collection of packages addressing a number of different standard problem domains in scientific computing. It is one of the greatest valuable libraries for kind of high-level science and engineering modules like discrete Fourier transform Linear Algebra, Optimization, and Sparse Matrices.

Elements of SciPy	Descriptions
scipy.integrate	used for Mathematical integration procedures and differential equation solvers;
Scipy.linalg	Linear algebra routines and medium decompositions extending beyond those provided in numpy.linalg
Scipy.optimize	Function optimizers (minimizers) and root finding processes
Scipy.sparse	Sparse matrices and sparse linear system solvers
Scipy.special	Wrapper round SPECFUN, a FORTRAN library applying many common mathematical functions, such as the gamma function and Scipy.stats Regular continuous and discrete

Table 3.1 Elements of SciPy

Together NumPy and SciPy form a reasonably complete and mature computational foundation for many traditional scientific computing applications.

II. NumPy

NumPy, short for Numerical Python, has long been a cornerstone of numerical computing in Python. It provides the data structures, algorithms and library glue required for many scientific applications relating to numerical data in Python.

Beyond the fast array-processing capabilities that NumPy adds to Python, one of its primary uses in data analysis is as a container for data to be passed among algorithms and libraries. For numerical data, NumPy arrays are more well-organized for storing and operating data than the other built-in Python data structures. In addition, libraries written in a lower-level language, such as C or FORTRAN, can operate on the data stored in a NumPy array without copying data into some other memory representation. Thus, many numerical computing tools for Python either assume NumPy arrays as a primary data structure or else target seamless interoperability with NumPy.

III. Matplotlib

Matplotlib is the furthestmost popular Python library for creating plots and different two-dimensional data visualizations. It was first created by [61] and is now maintained by a large team of developers. It is designed for making plots appropriate for publication. While there are other visualization libraries available to Python programmers, Matplotlib is the most widely used and as such has generally good integration with the rest of the ecosystem. In our study, we use it for visualization purpose.

IV. Pandas

The panda's name itself comes from panel data, an econometrics period for multidimensional structured datasets, and a play on the phrase Python data analysis itself. Pandas provide high-level data buildings and functions intended to create operating with structured or tabular data fast, easy, and expressive. Since its emergence in 2010, it has helped enable Python to be a powerful and productive data analysis environment. Pandas for data manipulation, preparation, and cleaning is such an important skill in data analysis. Data structures with tagged axes supporting automatic or obvious data alignment—this prevents common errors follow-on from misaligned data and working with differently indexed data coming from diverse sources.

Combined time series functionality is done in `panda`. The same data structures handle both time series data and non-time series data, Arithmetic operations and reductions that preserve metadata Flexible handling of missing data, Merge and other relational operations found in popular databases.

V. Statsmodels

`Statsmodels` is a statistical analysis package that was seeded by [47]. Compared with `scikit-learn`, `Statsmodels` contains algorithms for classical (primarily frequentist) statistics and econometrics. This includes and in our study, used submodules are: Regression models: Linear regression, generalized linear models, robust linear models, linear mixed effects models, etc. and Time series analysis: AR, ARMA, ARIMA, VAR, and other models and to Visualization of statistical model results `Statsmodels` is more focused on statistical inference, providing uncertainty estimates and p -values for parameters. `Scikit-learn`, by contrast, is more prediction-focused.

VI. Scikit-learn

For machine learning. Built on `NumPy`, `SciPy`, and `Matplotlib`, this library contains many efficient tools for machine learning and statistical modeling including linear regression, SVM, nearest neighbors, feature selection, matrix factorization, Grid search, cross-validation, metrics, Feature extraction, and normalization.

CHAPTER 4: EXPERIMENTATIONS

4.1. Overview

ECX 10 years' data given to the machine learning system to make an accurate prediction this contains 100,323 columns and 9 rows the attributes have 5 categorical values and 4 numerical attributes preprocessing and cleaning data is done. The main objective of the experimental work is to model the different algorithms for 10 years ECX Data and depending on the performances of the algorithms recommending the best algorithms and visualize the data

Briefly described in chapter three the software we used for implementation is an anaconda. The working environment for loading dataset, visualizing the data with different visualization techniques, modeling algorithms and performance evaluations are done in Jupyter notebook Jupiter Notebook in python is a web application that allows us to create and share codes and documents. It provides the location, where you can document your code, run it, look at the outcome, visualize data and see the results without leaving the environment. This makes it a handy tool for performing end-to-end data science workflows – data cleaning, statistical modeling, building and training machine learning models, visualizing data, and many other uses. Notebook documents contain the inputs and outputs of a collaborative session similarly as further text that accompanies the code however isn't meant for execution. In this means, notebook files can serve as a complete computational record of a session, interleaving executable code with explanatory text, and rich representations of resulting objects. These documents are inside JSON files and are saved with them. ipynb extension. Since JSON is may be a plain text format, they will be version-controlled and shared with colleagues. The underlying reason for selecting and discussing the selected techniques goes in addressing the Predefined research objectives stated. Which is; predicting the coffee prices. Being this the rationale, it necessitates justifying models and experiments to be carried out in the course of the knowledge discovery process.

4.2. Packages

In the experimentation environment, different python packages and libraries are needed to load data, preprocess data and to visualize the results. All packages and libraries used for demonstration is the last updated version of anaconda it is 3.7.1. We worked in two different anaconda application environment for Logistic Regression (LR), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART) and Support Vector Regression (SVR) algorithms are modeled and executed in boot root IDE and LSTM is in tensor flow environment. Because boot root application does not support Keras function which works needed in LSTM algorithm.

4.3. Data preparation

The procedures that we follow during experimentation is after we install the software and packages we Load the dataset directly from the file repository. We used pandas to load the data to the working environment. And We summarize the dataset to look in a few different ways, dimension of a dataset, how many instances and attributes are the data contains, with the shape property and eyeball of the data statistical summary of all attributes are done as summarization of data We extract a quick idea of how many instances (rows) and how many attributes (columns) the data contains with the shape property. It is also always a good idea to actually review our data. Our dataset contains categorical and numerical values. The next thing was Visualizing the dataset. We visualize the data in two methods, univariate and multivariate by using pandas and matplotlib. To better, understand each attribute of a data set a univariate plot suggests, we're only examining one variable. We begin with some univariate plots, that is, plots of each individual variable. Given that the input variables are numeric, we create box and whisker plots of each. A box plot summarizes however, every attribute is distributed. It also draws a line for the median and a box around the 25th and 75th percentiles. Whiskers tell us how the data is spread, and the dots outside the whiskers give candidate outlier values. To get a clearer idea of the distribution of the input attributes: We create a histogram of each input variable to get an idea of the distribution. Since histograms group data into bins and give us an idea of how many observations each bin holds. The shapes of the bins tell us whether an attribute is Gaussian, skewed, or has an exponential distribution. It also hints us about outliers. Where Multivariate plots to better understand the relationships between continuous values attributes.

Now we can look at the interactions between the variables. First, we look at scatter plots of all pairs of attributes. This is useful to identify structured relationships between input variables.

1	Trade_Date	Commodity	Commodity_Type	Commodity_Name	Symbol	Origin	ClosingPrice
2	3/8/2010	Coffee	UnWashed	Export	2UYCB3	Yirgacheffe B	558
3	3/16/2010	Coffee	UnWashed	Export	2UYCB3	Yirgacheffe B	560
4	3/26/2010	Coffee	UnWashed	Export	2UYCB3	Yirgacheffe B	532
5	3/29/2010	Coffee	UnWashed	Export	2UYCB3	Yirgacheffe B	520
6	2/18/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	548
7	3/1/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	540
8	3/12/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	540
9	3/16/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	546
10	3/18/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	540
11	3/19/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	540
12	3/22/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	541
13	3/24/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	535
14	4/5/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	488
15	4/6/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	488
16	4/16/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	480
17	4/23/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	470
18	4/26/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	470
19	5/4/2010	Coffee	UnWashed	Export	2UYCB4	Yirgacheffe B	475
20	5/19/2010	Coffee	UnWashed	Export	2UYCB4D	Yirgacheffe B	530
21	5/26/2010	Coffee	UnWashed	Export	2UYCB4D	Yirgacheffe B	530
22	6/3/2010	Coffee	UnWashed	Export	2UYCB4D	Yirgacheffe B	540
23	6/9/2010	Coffee	UnWashed	Export	2UYCB4D	Yirgacheffe B	565

Table 4.1 Sample of the dataset

After understanding our data, we converted the categorical attributes like "commodity", "commodity_type" to "continuous the concept is "one-hot-encoding".APPENDIX B:We use "date" as it is in a machine-learning model. Thus, we created two features month_of_the_year and day_of-the_month. Predicting closing price is not a classification problem but a Regression

The target attribute is the closing price, which is continuous. The evaluation metric for Regression is R2.

4.4. Create a Validation Dataset

We need to know that the model we created is any good quality. We used statistical methods to estimate the performance of the models that we create on unseen data. We also want a more concrete estimate of the performance of the best model on unseen data by evaluating it on actual

unseen data. That is, hold back some data that the algorithms will not get to see and we used this data to get a second and independent idea of how accurate the best model might actually be. We split the loaded dataset into 10 fold with the random seed of 30.

4.5 Build Models

In model building, we used two applications of the integrated development environment in anaconda those are base root and tensor flow. Base root application supports Logistic Regression (LR), K-Nearest Neighbors (KNN), classification and Regression Trees (CART) and Support Vector Regression (SVR) functions and the created tensor flow application support Keras function to LSTM algorithm.

This is a decent mixture of simple linear is LR and nonlinear (KNN, CART, SVR, LSTM) algorithms. We model the algorithms within their function and evaluate each model in turn with Kfold Cross Validation we state for $K = 10$ we split our data into 10 fold cross-validation. This cross-validation is primarily used in applied machine learning to estimate the skill of a machine-learning model on unseen data. In order to estimate how the model is expected to perform in general when used to make predictions on data. It is a standard method because it is simple to understand and it generally results in a less biased estimate of the model skill than other methods. In addition, we tend to reset the random number seed before every run to make sure that the evaluation of each algorithm is performed using exactly the same data splits. It ensures the results are directly comparable. Each observation within the data-set sample is assigned to a personal cluster and stays in this cluster for the period of the procedure. This means that each sample is given the opportunity to be used in the hold out set one time and used to train the model 9 times. This approach involves casually dividing the set of observations into k terms, or folds, of approximately equal size. The first fold is treated as a validation set, and therefore the method is fit on the remaining $k - 1$ folds. We build a model for four different linear and nonlinear algorithms, set up test harness to use Kfold validation and evaluate the algorithms in turn. in LSTM Environment the Keras library is installed and the data is normalized with the MinMaxScaler pre-process class from scikit-learn and we can split the data into train and test and convert an array of values into a data set matrix we transfer the compiled train/test data into the expected structure by using `numpy.shape()`. The LSTM network has three layers with one

input, a hidden layer with 32 LSTM blocks or neurons, and an output layer that makes a single value prediction. The default sigmoid activation perform is employed is used for the LSTM blocks. The network is trained for 500 epochs and a batch size of 1 is used. Once the model is fit, we estimate the performance of the model on the train/test datasets.

CHAPTER 5: EVALUATION AND DISCUSSION

5.1. Select Best Model

In the above chapter, we build models in different algorithms and here we come up with their test results need a way to choose between machine learning models. Moreover, our goal is to estimate the likely performance of a model on out-of-sample data. We trained and tested on the same data. We split the dataset into 10 folds so that the model can be trained and tested on different data one by one. Then, testing accuracy could be a higher estimate than training accuracy of out-of-sample performance. Nevertheless, we have a big problem with the train/test split. It provides a high variance estimate since dynamic that observations happen to be within the testing set can significantly change testing accuracy. Testing accuracy, can change a lot depending on which observation happens to be in the testing set? Hence, we have to use k-fold cross-validation to solve this problem.

K-fold cross-validation Dataset is split into 10 "folds" of equal size. Each fold acts as the testing set 1 time and act as the training set 10-1 times. Average testing performance is employed because of the estimate of out-of-sample performance. Also known as cross-validated performance. Benefits a more reliable estimate of out-of-sample performance than train/test split. Reduce the variance of a single trial of a train/test split. Hence, with the benefits of k-fold cross-validation, we are able to use the average testing accuracy as a benchmark to decide which is the most optimal set of parameters for the learning algorithm. It does not work well when data is not uniformly distributed. It is hard to mount the quality of a given model without quantifying its performance over training and testing. This is usually done using some kind of performance metric, whether it is through calculating some type of error, the goodness of fit, or some other useful measurement. For this study, the researcher used to calculate the Coefficient of determination [48], R^2 , to quantify our model's performance. The constant or the coefficient of determination for a model could be a helpful statistic in regression analysis, as it often describes how "good" that model is at making predictions.

The values for result range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable. A model with Result of zero perpetually fails to predict the target variable, whereas a model with an R^2 of 1 perfectly predicts the target variable. Any value between Zero and One indicates what percentage of the target variable, using this model, the features can explain. A model may be given a negative result in addition, that indicates that the model is no better than one that naively predicts the mean of the target variable.

In our modeled algorithms, we get the performance as follows

No.	Algorithms	Performance
1	Linear Regression	-6614722.039
2	K-nearest neighbor	0.909470
3	Decision tree regression	0.910699
4	Support vector regression	-0.103426
5	Long Short Term Memory	0.91523

Table 5.1 Algorithms Performance Result

This means Linear regression and support vector regression is very bad to predict the target variable because the result is under zero whereas the KNN, LSTM, and CART algorithms are a well-performed algorithm result in can predict the unseen data especially LSTM algorithm can predict perfectly. To display the result, the system takes more than 12 hours.

During our experimentation, the result looks Long Short Term Memory has better estimated high score. We create a plot of the model evaluation results and compare the spread and the mean Result of each model. There is a population of accuracy measure for each algorithm because each algorithm was evaluated 10 times (10-fold cross-validation). The LSTM algorithm is a model with a Result of close to 1 perfectly predicts the target variable and was a high-performance model based on our tests. We fit LSTM model directly on the validation set and summarize the results as a final estimated Result score, the performance result is 92% this means to predict the unseen data of ECX can be highly performed in long short-term memory

5.2. Making Predictions in LSTM

By using LSTM, we predict our target variable

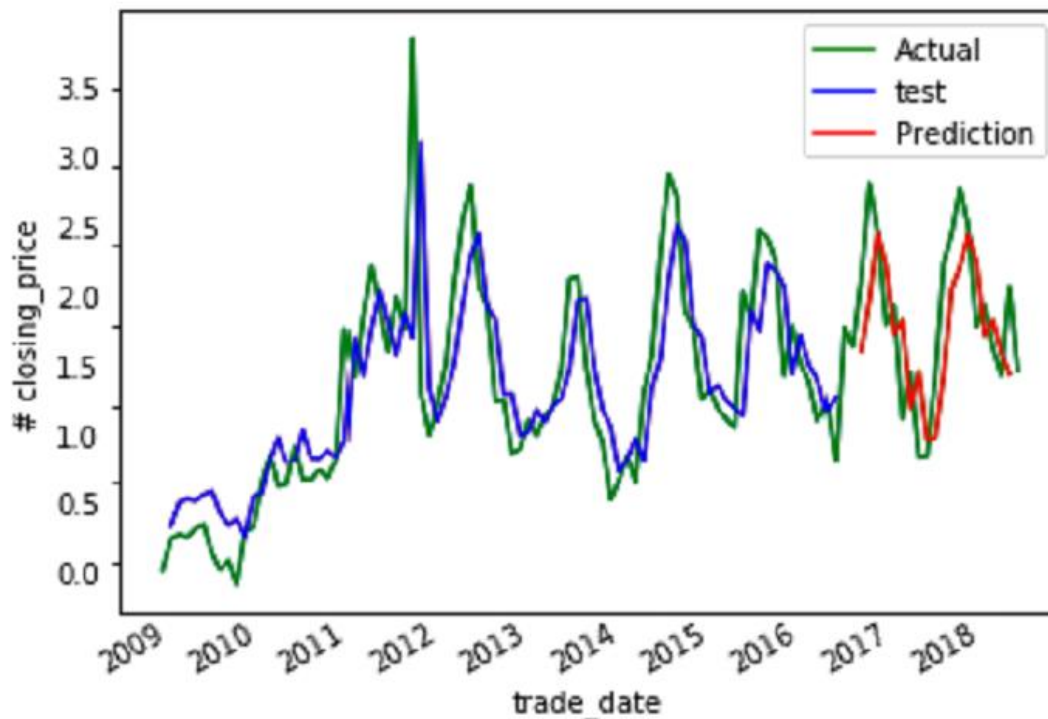


Figure 5.1 Actual trained and predicted graph

The line plot is showing the observed values compared to the rolling forecast predictions overall, our forecasts align with the true values as well. This graph showing in Figure 5.1 for all trade data's, where for some specific origins and commodity types, the forecasted results are stated bellow.

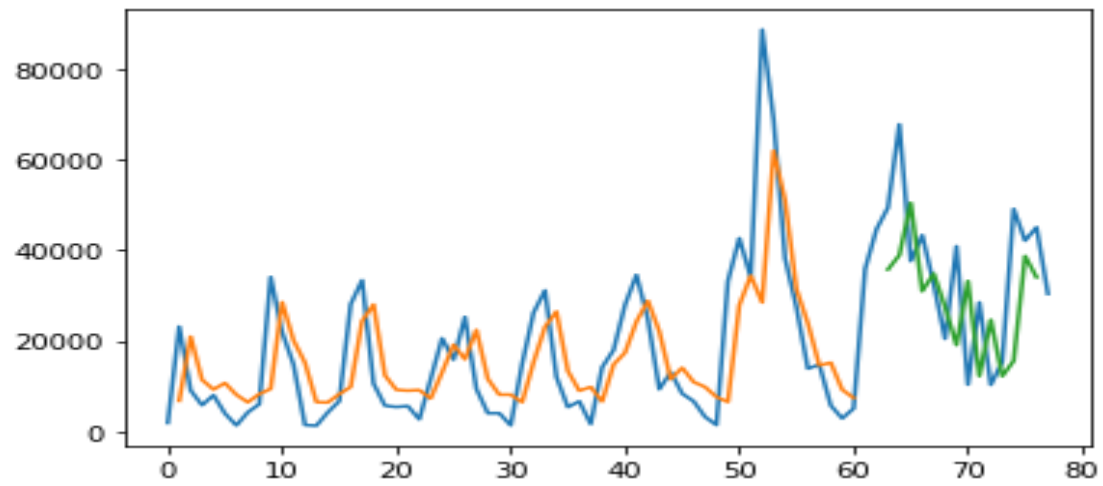


Figure 5.2. Yergachefe special

38745.08247389879
19459.37708172995

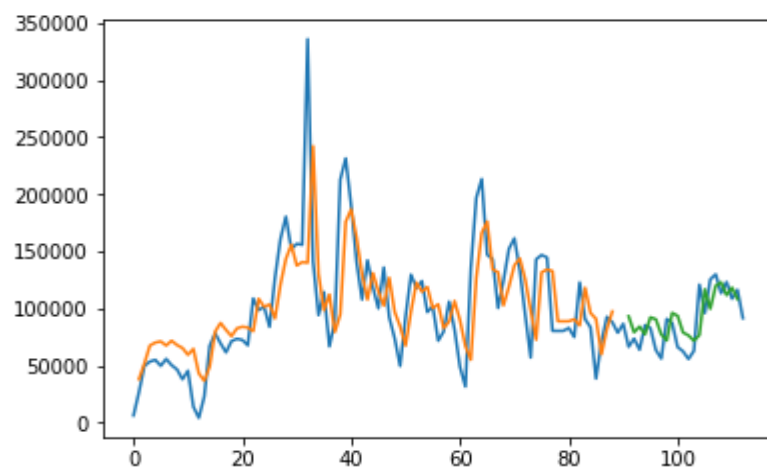


Figure 5.3. Export Jimma

12.87448764621674
15.675301025477498

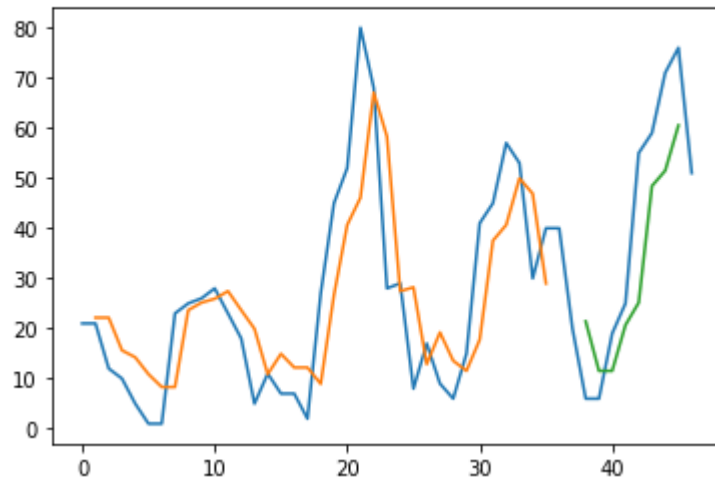


Figure 5.4. Gujjii Special

36821.57281056133
47140.35986323389

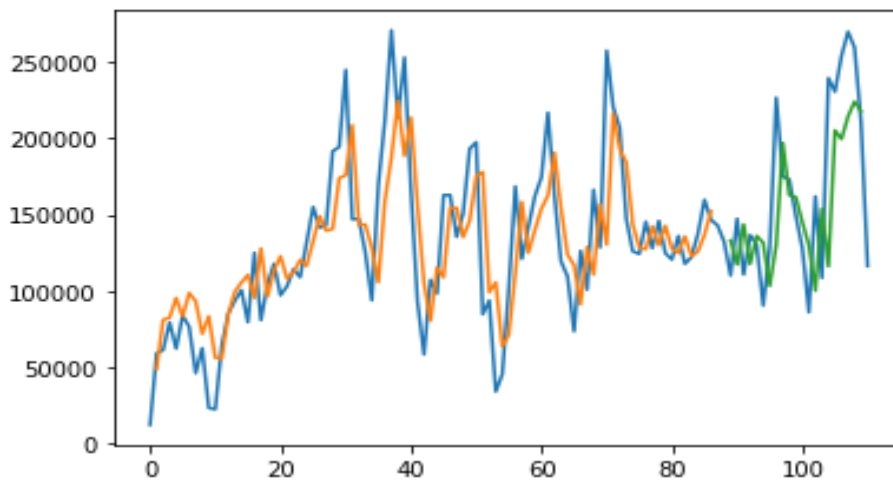


Figure 5.5. Harar Export

14648.660547832224
26793.048025903303

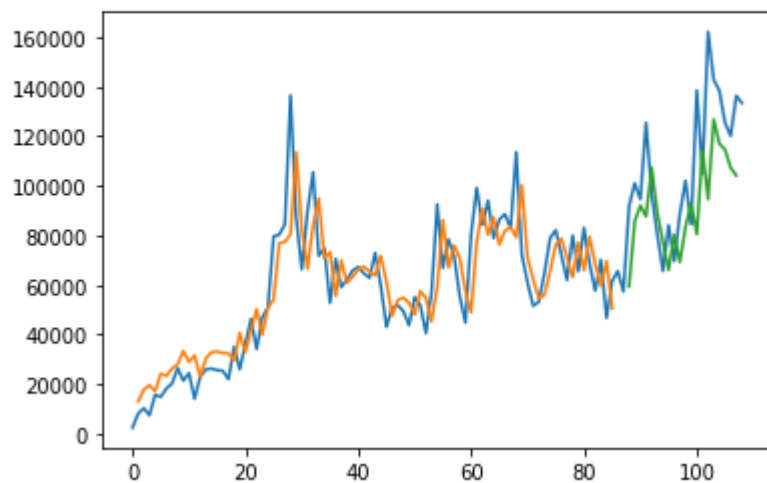


Figure 5.6. Local by product

3725.6002995581616
3786.3552409787694

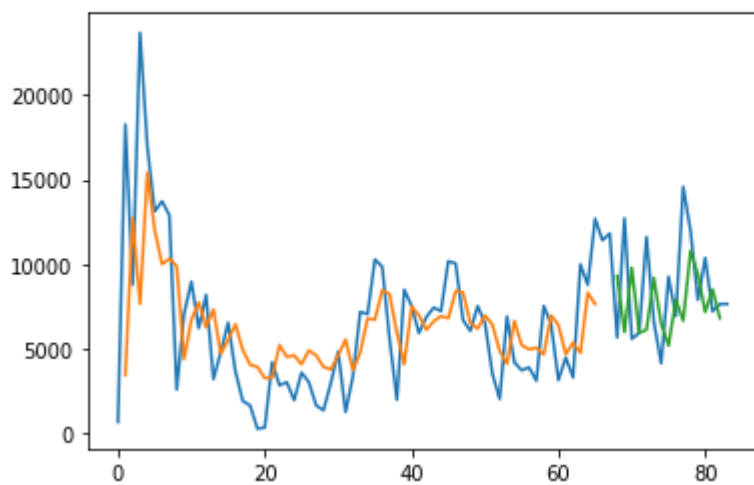


Figure 5.7. Local byproduct

118.97171980337157
85.82955938231704

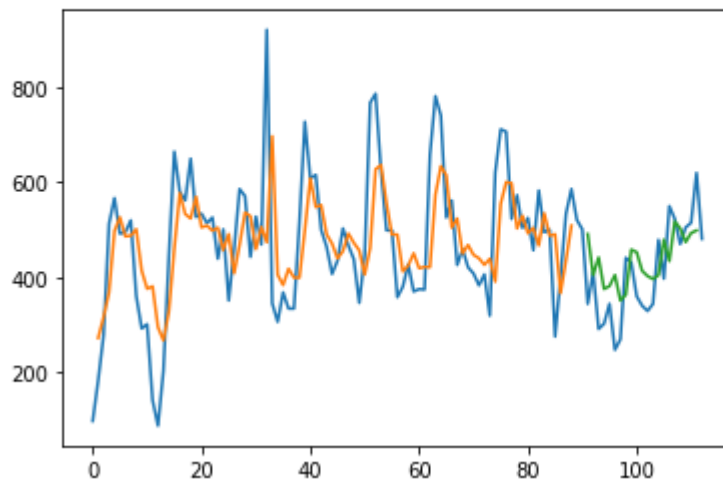


Figure 5.8 Unwashed export

60682.518691923906
141159.6253225663

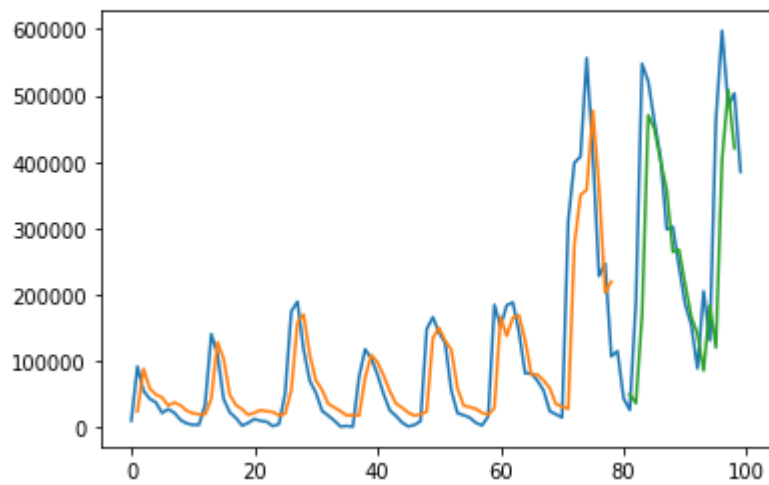


Figure 5.9. Washed Specialty

43.344180079401816
99.1298891531467

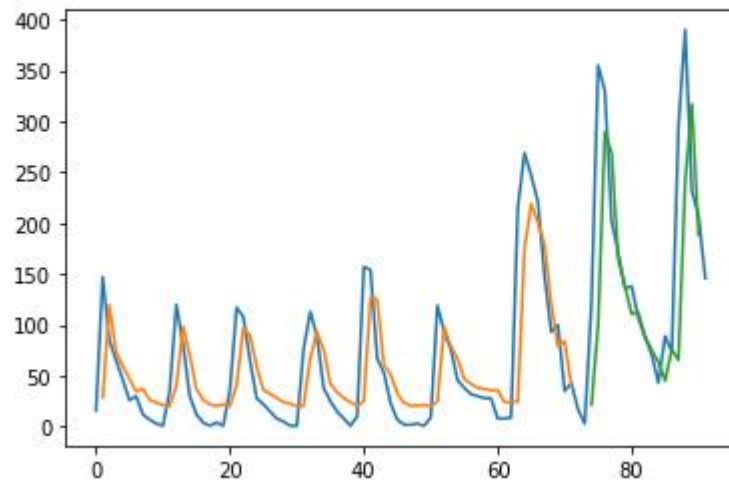


Figure 5.10. Washed Specialty

165607.63147256037
129518.78240223406

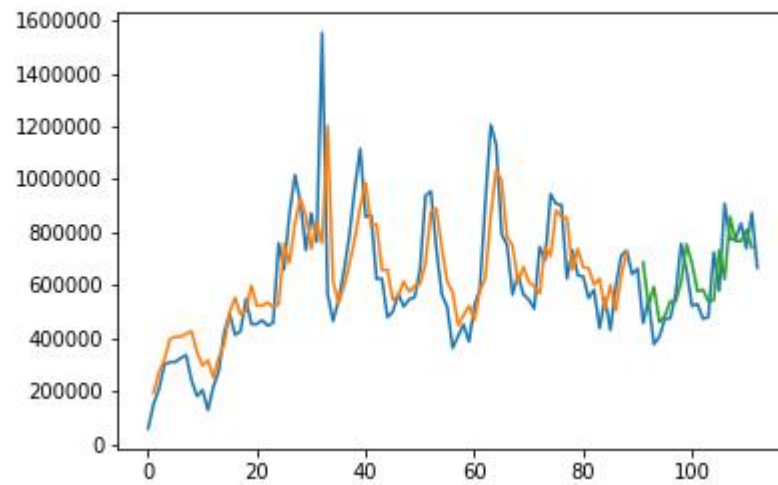


Figure 5.11. Washed unwashed Export

We applied one of the most commonly used method time series forecasting known as ARIMA, which stands for Autogressive Integrated moving average in R studio. By loading the dataset from excel file using readxl and to place our data in time series format we used ts () function and to get the optimal auto ARIMA model we used auto. ARIMA () function then plotted. Forecasted data from the auto ARIMA model.

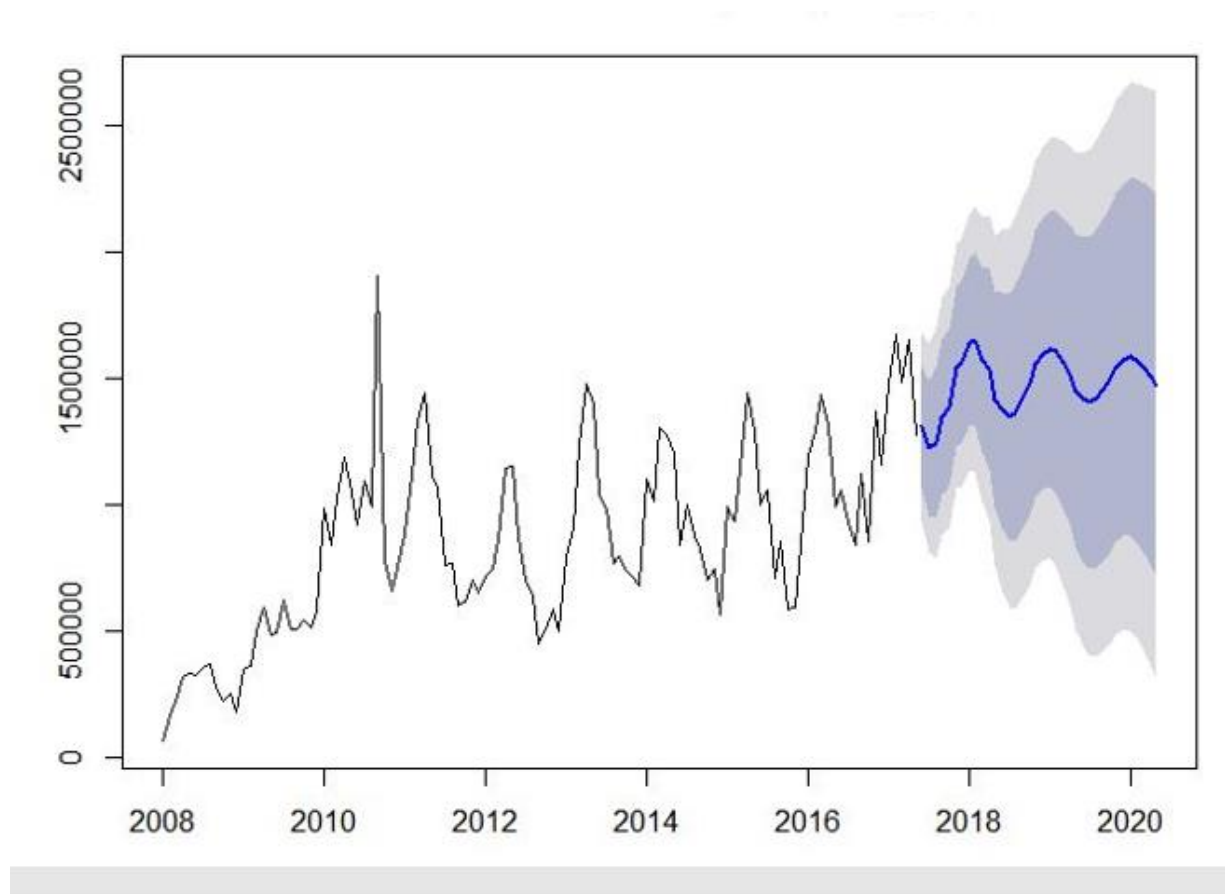


Figure 5.2. Predicted graph

Where the black line actual values and the blue one is a future predicted values and gray part is the confidence level. Whatever happens, the values do not exceed below or above this confidence for the coming three years, the results of forecasted values are stated below in a table with high and low confidence results.

PREDICTED DATE	Forecasted price	Low	High
Jun-17	1308664	1064972	1552357
Jul-17	1225098	954008.1	1496187
Aug-17	1246614	947739.5	1545489
Sep-17	1346900	1038395	1655404
Oct-17	1385333	1071299	1699368
Nov-17	1547054	1229614	1864493
Dec-17	1567320	1245341	1889299
Jan-18	1641237	1311027	1971447
Feb-18	1655402	1310238	2000566
Mar-18	1574697	1206855	1942539
Apr-18	1537095	1141113	1933078
May-18	1412774	987849.9	1837699
Jun-18	1373737	908645.7	1838828
Jul-18	1344419	855082.5	1833755
Aug-18	1363362	856705.8	1870017
Sep-18	1420064	903322.5	1936806
Oct-18	1479015	955631.6	2002398
Nov-18	1556317	1027558	2085077
Dec-18	1594971	1060021	2129922
Jan-19	1619154	1075363	2162945
Feb-19	1608526	1051655	2165398
Mar-19	1562362	987723.6	2136999
Apr-19	1513490	917679.9	2109300
May-19	1452202	834443.4	2069961
Jun-19	1417551	778130.3	2056972
Jul-19	1403518	747188.3	2059847
Aug-19	1417597	748494.2	2086701

Sep-19	1453809	775531.2	2132087
Oct-19	1498157	812807.4	2183507
Nov-19	1543734	852052.6	2235416
Dec-19	1572633	874079.9	2271185
Jan-20	1583680	876616.9	2290744

Table 5.2. Forecasted confidence table

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1. Conclusion

To increase the productivity of the farmers of coffees, different strategies are introduced. Among those strategies, forecasting the future market directions is the major point. Today Predictions are based on absolute theories and behaviors. The problems with forecasting commodity prices are that the volume of data is too bulky and vast so machine learning techniques are used. This working model the Machine Learning algorithms for prediction of Ethiopian commodity exchange data that could outperform the buy-and-hold strategy and lower risk. The approach to predict the future market direction of the ECX data has been introduced based on chart patterns recognition by using machine learning Regression model. Models are built through different methods including linear regression, Decision tree regression (CART), SVR, KNN, and LSTM.

In this study, we model prediction algorithms in machine learning application and comparison are done and the result is also discussed. During compilation where four algorithms compared in same model and root boot environment and LSTM is compered in its sensor flow environment and their result is where Linear regression and support vector regression is below zero which means very bad result and KNN and decision tree is relatively better compared to LR and SVR where LSTM is highly performed in a model. We predict the data with high performed LSTM algorithm and forecast the future three years' market directions we used ARIMA model in Better Performance.

6.2. Future Work

In this study, we suggest or recommend as future work, since one commodity price may be concurrently related with other commodity prices so we recommend using fuzzy correspondence to analyze the commodity price data to commodity trader can more precisely determine the cause of commodity price change.

The other is to develop a varying go-between or agent to allow the commodity trader to define different parameter such as commodity buying/selling times and insert them into the system, therefore, the system will Change the commodity trader to the best times to buy or sell Commodity. Another is to develop more accurate agent to predict the commodity price at any given time.

Reconsidering the factor affecting the behaviors of the commodity market such as trading volume, news and financial report, which may affect commodity price, can be another rich field for future study. For further studies, there is a need to formulate a model that is generic enough to suit the full range of commodities at Ethiopian commodity exchange market. This study carefully selected one commodity as the basis of the research thought the total listed commodities are seven so can be predicted for all commodities.

References

- [1] Gregory Dicum and Nina Luttinger, *The Coffee Book*, New York New Press., 1999.
- [2] The central intelligence agency, in *the world factbook*, February 04, 2017.
- [4] Food info site, in *www.food-info.net*, 2010. Vishal Parikh, Parth Shah., *Stock Prediction, and Automated Trading System*, *www.csjournalss.com*, Sep-Mar 2015.
- [5] "Economy Watch for money, Ethiopia Economic Structure, Ethiopia," March 31, 2010.
- [7] The proclamation, a proclamation to provide for the establishment of the Ethiopian commodity exchange authority, no. 551, 2007.
- [8] Gabre-Madhin, E., *The Ethiopian Commodity Exchange (ECEX): Making the market work for all*, In *Presentation to a UNCTAD Expert Meeting on the Trade and Development Implications of Financial Services and Commodity Exchanges.*, 2007.
- [9] Alemu, D., and G. Meijerink, *The Ethiopian Commodity Exchange: An Overview*. Ethiopian Pulses, Oilseeds and Spices Processors Exporters Association and Wageningen University, 2010.
- [10] Jain AK, Murty mn, Flynn p, "data clustering: a review. *ACM Comput surveys*, vol. 31, no. 3, p. 264–323, 1999.
- [11] R. S. Sutton and A. G. Barto: "Reinforcement Learning: " 1997
- [12] shichaozhang, cheng qi Zhang, and Xiangyang, *data preparation for DATA MINING, Applied Artificial Intelligence*, vol. 17, p. 375–381, 2003.
- [13] Batista G, Monard MC, *An analysis of four missing data treatment methods for supervised learning.*, *Appl Artif Intell*, vol. 17, p. 519–533, 2003.
- [14] Hodge, V, & Austin, J, *A survey of outlier detection methodologies.*, *Artif Intell Rev*, vol. 22, no. 2, p. 85–126, 2004.
- [15] Liu, H and Motoda, H, " *Instance selection and constructive data mining*," *Kluwer, Boston*, 2001.

- [16] Reinartz T, "A unifying view on instance selection, data mining and knowledge discovery," *Kluwer Academic Publishers*, vol. 6, p. 191–210, 2002.
- [17] Jankovic N, Stephen S, "The class imbalance problem: a systematic study.," *Intell Data Anal*, vol. 6, no. 5, p. 429– 450, 2002.
- [18] Batista G, Monard MC, "An analysis of four missing data treatment methods for supervised learning.," *Appl Artif Intell*, vol. 17, p. 519–533, 2003.
- [19] Yu L, Liu H, "efficient feature selection via analysis of relevance and redundancy.," *JMLR*, vol. 5, p. 1205–1224, 2004.
- [20] Markovitch S, Rosenstein D, "Feature generation using general construction functions.," *Mach Learn*, vol. 49, p. 59–98, 2002.
- [21] Kotsiantis, Sotiris, *Supervised Machine Learning: A Review of Classification Techniques, Informatica (Ljubljana)*, vol. 31, 2007.
- [22] Jiawei Han and Micheline Kamber “Data Mining: Concepts and Techniques” Academic Press, Morgan Kaufmann, 2001
- [23] Nissan Levine and Jacob Zahavi “The Data Mining and Knowledge Discovery Handbook “(pp.1189-1220) January 2010
- [24] Malkiel, Burton G., The efficient market hypothesis and its critics., *Journal of Economic Perspectives*, pp. 107-111, 2003.
- [25] "Pujana Paliyawan Stock Market Direction Prediction using Data Mining Classification Conference Paper, December 2014.
- [26] Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi, Mohammed K. Ali Shatnawi March, Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm, vol. 3, no. 3, 2013.
- [27] Quinlan, J. R., Induction of Decision Trees, *Mach. Learn.* , vol. 1, no. 1, pp. 81-106, 1986.
- [28] Hakob GRIGORYAN, Stock Market Prediction using Artificial Neural Networks Case Study of TAL1T, Nasdaq OMX Baltic Stock, *Database Systems Journal*, vol. VI, no. 2, 2015.
- [29] Dehnad, Kosrow., Behavioral Finance, and Technical Analysis., *The Capco Institute Journal of Financial Transformation*, vol. 32, Aug (2011).

- [30] Kotsiantis, S. B., Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, vol. 31, pp. 249-268, 2007.
- [31] SUBHA, M.V., and THIRUPPARKADAL NAMBI, S., Classification of Stock Index movement using k-Nearest Neighbours (k-NN) algorithm, vol. 9, no. 9, September 2012.
- [34] Rogstadius, J., Visualizing the Ethiopian Commodity Market., *Linköping*, 2009..
- [35] Dempsey, J, A Case Thesis of Institution Building & Value Chain Strengthening to Link Ethiopian Cooperative Coffee Producers to International Markets. Addis Ababa., 2006.
- [36] Shumate, Z., K. Urgessa, and Z. Kebebew., Analysis of Market Chains of Forest Coffee in Southwest Ethiopia, 2012.
- [38] Vapnik, V., "The Nature of Statistical Learning Theory," *Springer*, 1995.
- [39] Lukasz, G., Tire, M. and Negru, V, Neural Predictions of Stock Price Fluctuations Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), *14th International*, pp. 505- 512, 2012.
- [40] Abhishek, K., Khairwa, A., Pratap, T. and Prakash, S., A stock market prediction model using Artificial Neural Network., *Computing Communication & Networking Technologies (ICCCNT), 3rd International Conference*, pp. 1-5, 2012.
- [41] Walsham, G, The emergence of interpretivism in IS research, *Information systems research*, vol. 6, no. 4, p. 376–394., 1995.
- [42] Wanda J. Orlikowski & Jack J. Baroudi, Studying Information Technology in Organizations: Research Approaches and Assumptions: *The Institute of Management Sciences*, 1991.
- [43] WALSHAM G; Interpreting Information Systems in Organizations., *Wiley, Chichester*, 1993.
- [45] Reilly Media, O., "Python for Data Analysis 2nd Edition," *Inc*, 2017.

- [46] Gordon V.Kass, Chi-square Automatic Interaction Detector (CHAID) was a technique in, 1980.
- [51] Larvae, Nada., "Data Mining in Medicine," *Selected Techniques and Applications.*, 1998.
- [52] Last, Mark and Kandel, Abraham. , "Automated Perceptions in Data Mining.," 2002.
- [53] Two Crows Corporation., Introduction to Data Mining and knowledge discovery," 1999.
- [54] Han, J., Kamber, M., and Pei, J., Data Mining Concepts and Techniques., *Morgan Kaufmann Publishing* 2012.
- [55] Madhu Sanjeevi, Decision Trees Algorithms, <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-treesalgorithms-b93975f7a1f1>, 2017.
- [56] Breuer, T., Jandačka, M., Pühr, R., Rheinberger, K., and Seeger.M., The Interaction Dynamics between Market and Credit Risk, 2011.
- [57] "Chinmay Pradhan, Heading a data analytics company".
- [58] Rogers, Greg, and Joyner, Elen., "Mining your Data for Health care Quality improvement," 2001.
- [59] Witten IHaF, Eibe., Practical Machine Learning Tools and Techniques with Java Implementations, *USA: Academic Press*, 2000.
- [60] Pujana Paliyawan, stock market direction prediction using data mining classification, *February arpn journal of engineering and applied sciences*, vol. 10, no. 3.

Cite reference

- [3] International coffee organization ico.org
- [32] "<https://www.youtube.com/watch?v=gK2IAzN1lcM>".
- [33] UNCTAD," 2009
- [37] <https://machinelearningmastery.com/linear-regression-for-machine-learning/>".
- [44] "<https://opensource.com/article/18/4/getting-started-anaconda-python>".
- [47] https://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination.
- [48]<https://www.mathworks.com/help/stats/understanding-support-vector-machineregression.html>".
- [49] "Jonathan Taylor Professor of Statistics number of regression analysis models popular in".
- [50] "<http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/>".

APPENDIX A:

Codes Used In Rstudio and Its Results

```
#Auto ARIMA Forecasting in R
Library (read XL)
Library (forecast)
#load the dataset from excel using read XL
Data <- read.xlsx ('E: /Thesis/Ecx.xlsx')
#use ts () function to place in time series format
PRICE <- ts (data$` Closing Price`, frequency = 12, start = c(2008,1))
#Plot the time-series formatted data
Plot (PRICE)
#use the auto. ARIMA () function to get the optimal auto ARIMA model
autoarima1 <- auto. ARIMA (PRICE)
#get the forecasted data for a period of 17 weeks =4 months @ 4.2? Weeks/month
forecast1 <- forecast (autoarima1, h=36)
#show forecasted values
forecast1
#plot the forecasted data from the auto ARIMA model
Plot (forecast1)
#plot the residuals over time to see congruence or variance
Plot (forecast1$residuals)
#plot the residuals (sample vs. theoretical)
Qqnorm (forecast1$residuals)
#
Acf (forecast1$residuals)
Pacf (forecast1$residuals)
#get accuracy by MAPE and other leading indicators - each dataset is different method1
Summary (autoarima1)
```

#method2

Accuracy (autoarima1)

#replot? he forecast

Plot (forecast1)

Series: PRICE

ARIMA (3,1,2)(1,0,0)[12]

Coefficients:

ar1	ar2	ar3	ma1	Ma2	sar1
1.2687	-0.2618	-0.3737	-1.7814	0.9414	0.1660
0.0923	0.1500	0.0920	0.0510	0.0454	0.1006

Table 6.1. Coefficient of ARIMA

Sigma^2 estimated as 3.615e+10: log likelihood=-1519.2

AIC=3052.4 AICc=3053.48 BIC=3071.43

Training set error measures:

Training	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
set	23506.3	184160.3	127813	1.17636	15.57149	0.517219	-0.02874

Table 6.2. Training set error measures

APPENDIX B:

Anaconda Jupiter Notebook Code

```
ecx.commodity = 'commodity_' + ecx.commodity.map(str)
ecx.commodity_type = 'ctype_' + ecx.commodity_type.map(str)
ecx.commodity_name = 'cname_' + ecx.commodity_name.map(str)
ecx.symbol = 'symbol_' + ecx.symbol.map(str)
ecx.origin = 'origin_' + ecx.origin.map(str)
# Create dummies
commodity_dummies = pd.get_dummies(ecx.commodity)
ctype_dummies = pd.get_dummies(ecx.commodity_type)
cname_dummies = pd.get_dummies(ecx.commodity_name)
symbol_dummies = pd.get_dummies(ecx.symbol)
origin_dummies = pd.get_dummies(ecx.origin)

ecx['month'] = ecx.index.month
ecx['month'] = 'month_' + ecx.month.map(str)

ecx['day'] = ecx.index.day
ecx['day'] = 'day_' + ecx.day.map(str)

month_dummies = pd.get_dummies(ecx.month )
day_dummies = pd.get_dummies(ecx.day)

ecx = ecx.drop(["day", "month", "commodity", "commodity_type", "commodity_name",
"symbol", "origin", "closing_price"], axis=1)
ecx = pd.concat([ecx, day_dummies, month_dummies, commodity_dummies, ctype_dummies,
cname_dummies,symbol_dummies, origin_dummies], axis=1
```

APPENDIX C:

```
# convert an array of values into a dataset matrix
def create_dataset(dataset, look_back=1):
    dataX, dataY = [], []
    for i in range(len(dataset)-look_back-1):
        a = dataset[i:(i+look_back), 0]
        dataX.append(a)
        dataY.append(dataset[i + look_back, 0])
    return numpy.array(dataX), numpy.array(dataY)

# fix random seed for reproducibility
numpy.random.seed(2)

# normalize the dataset
scaler = MinMaxScaler(feature_range=(0, 1))
dataset = scaler.fit_transform(dataset)

#split into train and test sets
train_size = int(len(dataset) * 0.80)
test_size = len(dataset) - train_size
train, test = dataset[0:train_size,:], dataset[train_size:len(dataset),:]

# reshape into X=t and Y=t+1
look_back = 1
trainX, trainY = create_dataset(train, look_back)
testX, testY = create_dataset(test, look_back)

# reshape input to be [samples, time steps, features]
trainX = numpy.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
testX = numpy.reshape(testX, (testX.shape[0], 1, testX.shape[1]))

model = Sequential()
model.add(LSTM(4, input_shape=(1, look_back)))
model.add(Dense(1))

model.compile(loss='mean_squared_error', optimizer='adam')
model.fit(trainX, trainY, epochs=100, batch_size=1, verbose=2)
```

```

# make predictions
trainPredict = model.predict(trainX)
testPredict = model.predict(testX)

# invert predictions
trainPredict = scaler.inverse_transform(trainPredict)
trainY = scaler.inverse_transform([trainY])
testPredict = scaler.inverse_transform(testPredict)
testY = scaler.inverse_transform([testY])

# calculate root mean squared error
trainScore = math.sqrt(mean_squared_error(trainY[0], trainPredict[:,0]))
print("Train Score: %.2f RMSE' % (trainScore))
testScore = math.sqrt(mean_squared_error(testY[0], testPredict[:,0]))
print("Test Score: %.2f RMSE' % (testScore))

# shift train predictions for plotting
trainPredictPlot = numpy.empty_like(dataset)
trainPredictPlot[:, :] = numpy.nan
trainPredictPlot[look_back:len(trainPredict)+look_back, :] = trainPredict

# shift test predictions for plotting
testPredictPlot = numpy.empty_like(dataset)
testPredictPlot[:, :] = numpy.nan
testPredictPlot[len(trainPredict)+(look_back*2)+1:len(dataset)-1, :] = testPredict

# plot baseline and predictions
print(trainScore)
print(testScore)
plt.plot(scaler.inverse_transform(dataset))
plt.plot(trainPredictPlot)
plt.plot(testPredictPlot)
plt.show()

from statsmodels.tsa.stattools import adfuller
def test_stationarity(timeseries)

#Determine rolling stastics

```

```

movingAveragee=timeseries.rolling(Window=12).mean()
movingSTD=timeseries.rolling(Window=12).std()

#plot rolling stastics
rig = plt.plot(timeseries, colors='blue', label='Original')
orig = plt.plot(movingAveragee, colors='red', label='Rolling Mean')
orig = plt.plot(movingSTD, colors='black', label='Rolling Std')
plt.legend(loc = 'best')
plt.title('Rolling Mean & Standard Deviation')
plt.show(block=False)

#perform Dickey-fuller test:
from statsmodels.tsa.stattools import adfuller
print ('Result oF Dickey-Fuller Test')
dfctest = adfuller(timeseries['closing_price'], autolag= 'AIC')
dfoutput= pd.series(dfctest[0:4], index['Test Statistic', 'P-Value', '#lags Used', 'Number of
Observesions used'])
for key, value in dfctest[4].items():
    dfoutput['Critical value (5%)' %key] = value
print(dfoutput)

```