

A Mini Project report entitled

On

Bitcoin Price Prediction Using Various Regression Models

In partial fulfillment of the requirements for the award of

BACHELOR OF TECHNOLOGY
In
Computer Science and Engineering

submitted by

Komal Soni (18E51A0565)

Mohammed Abdul Hakeem (18E51A0576)

Sheri Bai Abhishek Reddy (18E51A05A7)

Under the Esteemed guidance of

Mrs K Veena, M Tech

Assistant Professor

Department of CSE



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

HYDERABAD INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(Affiliated to JNTU, Hyderabad, TS, NBA Accredited)

2021-2022

HYDERABAD INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(Affiliated to JNTUH, Hyderabad, TS, INDIA, NBA Accredited)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

This is to certify that the mini project work entitled “**Bitcoin Price Prediction Using Various Regression Models**” is a bonafied work carried out by Komal Soni bearing Roll No. 18E51A0565, Mohammed Abdul Hakeem bearing Roll No. 18E51A0576, Sheri Bai Abhishek Reddy bearing Roll No. 18E51A05A7 in partial fulfillment of the requirements for the degree **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE** by the Jawaharlal Nehru Technological University, Hyderabad, during the academic year 2020-2021. The matter contained in this document has not been submitted to any other University or institute for the award of any degree or diploma.

Internal Supervisor

Mrs K Veena, M Tech
Assistant Professor
Department of CSE
HITAM

Head of the Department

Dr Dara Vikram
Professor & HoD
Department of CSE
HITAM

External Examiner



HYDERABAD INSTITUTE OF TECHNOLOGY AND MANAGEMENT

DUNDIGAL – 500 043, HYDERABAD, TELANGANA STATE

Department of Computer Science and Engineering

DECLARATION

We '**Komal Soni(18E51A0565), Mohammed Abdul Hakeem (18E51A0576), Sheri Bai Abhishek Reddy (18E51A05A7)**', are students of '**Bachelor of Technology in Computer Science and Engineering**', session: **2021 - 2022**,**Hyderabad Institute of Technology and Management**, Dundigal, Hyderabad, Telangana State, hereby declare that the work presented in this project work entitled '**Bitcoin Price Prediction Using Various Regression Models**' is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of engineering ethics. It contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

student name: Komal Soni, Abdul Hakeem , S. Abhishek
Roll Nos.: (18E51A0565), (18E51A0576), (18E51A05A7)

Date:

Acknowledgement

An endeavor of a long period can be successful only with the advice of many well wishers.

We would like to thank our chairman **Sri.Arutla Prasanth** for providing all the facilities to carry out project work successfully.

We would like to thank our Honorable Principal **Dr. S Sudhakara Reddy** who has inspired lot through their speeches and providing this opportunity to carry out our mini/major project successfully.

We are very thankful to our Head of the Department, **Dr Dara Vikram**, and B-Tech project coordinator **Mrs P Swathy, M Tech**.

We would like to specially thank my internal supervisor , **Mrs K Veena, M Tech** for our technical guidance, constant encouragement and enormous support provided to us for carrying out our mini project work.

We wish to convey our gratitude and express sincere thanks to all **D.C (Departmental Committee)** and **P.R.C (Project Review Committee)** members, non-teaching staff for their support and Co-operation rendered for successful submission of our major project work.

We also want to express our sincere gratitude to all my family members and my friends for their individual care and everlasting moral support.

Komal Soni (18E51A0565)
Mohammed Abdul Hakeem (18E51A0576)
Sheri Bai Abhishek Reddy (18E51A05A7)

Contents

1	Introduction	1
2	Theoretical Background & Literature Survey	4
2.1	Literature Survey	4
2.2	Theoretical Background	5
2.2.1	Extra tree regressor	6
2.2.2	Gradient boosting machine regressor	7
2.2.3	Random Forest regressor	7
2.2.4	Decision tree regressor	7
2.2.5	Linear regressor	7
2.2.6	Light Gradient Boosting Machine regressor	8
2.2.7	Extreme Gradient Boosting regressor	8
2.2.8	Stochastic Gradient Descent Regressor	8
2.2.9	Kernel Ridge Regressor	8
2.2.10	Bayesian Ridge Regressor	9
2.2.11	Support Vector Machine Regressor	9
3	Requirement Analysis	10
3.1	Hardware Requirements	10
3.2	Software Requirements	10
4	Design	11
4.1	Design Goals	11
4.2	System Architecture	12
5	Methodology	14
5.1	Dataset	14
5.2	Data Preprocessing	15
5.2.1	Data Cleaning	15
5.2.2	Data Normalization	16
5.2.3	Data Splitting	16
5.3	Data Analysis	17
5.4	Machine Learning Models Training	18
5.5	Machine Learning Models Testing	19
6	Results	20
6.1	Visualization of Results	21
7	Conclusion	27
8	Future Work	28

List of Figures

4.1	Flow chart of regression-based models	12
4.2	System Architecture	13
5.1	Overview of Bitcoin Dataset	15
5.2	Dataset Features Analysis	17
5.3	Price Distribution Analysis	17
5.4	Heatmap of Bitcoin dataset	18
6.1	Extra Tree Regressor Actual vs Predicted Bitcoin Price	21
6.2	Gradient Boosting Machine Regressor Actual vs Predicted Bitcoin Price	21
6.3	Random Forest Regressor Actual vs Predicted Bitcoin Price	22
6.4	Decision Tree Regressor Actual vs Predicted Bitcoin Price	22
6.5	Linear Regressor Actual vs Predicted Bitcoin Price	23
6.6	LGBM Regressor Actual vs Predicted Bitcoin Price	23
6.7	XGBoost Regressor Actual vs Predicted Bitcoin Price	24
6.8	Stochastic Gradient Descent Regressor Actual vs Predicted Bitcoin Price	24
6.9	Kernel Ridge Regressor Actual vs Predicted Bitcoin Price	25
6.10	Bayesian Ridge Regressor Actual vs Predicted Bitcoin Price	25
6.11	Support Vector Machine Regressor Actual vs Predicted Bitcoin Price	26

List of Tables

6.1	Regression-based Models results comparision	20
-----	---	----

Abstract

Cryptocurrency is a fascinating area of research developed due to the rapid development of financial technologies. One of the well-received cryptocurrencies is Bitcoin. Bitcoins can be seen from an economic as well as computer science angle. This makes bitcoin an exciting field of study. Studying its behaviour is of great importance to countries as many of them are now legalizing this form of virtual currency. Thus predicting its price and analysing the trend in the open market will be boon not only to an economically growing country but also to every person who wants to invest in Bitcoins.

In 2017, a significant number of individuals profited from the staggering growth of the price of Bitcoin from 800 USD in January to almost 20,000 USD in December. Because the cryptocurrency market being relatively new when compared to traditional markets such as stocks, foreign exchange, and gold, there is a significant lack of studies in regard to predicting its price behavior. This project is interested in evaluating a number of regression-based algorithms in predicting the price of the Bitcoin (BTC) against United States Dollar (USD).

Bitcoin is the fastest-growing cryptocurrency. We have seen drastic changes in bitcoin prices over time. To make predictions for such changes, we use machine learning techniques over real-time data recorded for every 24-hour time interval since the presence of bitcoin beginning in the year 2009. We select eleven different regression models, analyze these models and obtain the best regression-based model for bitcoin price prediction. The results obtained from the study depict that the Bayesian ridge regressor outperforms all the other regression-based models followed closely by the Linear regressor.

Chapter 1

Introduction

Bitcoin is a network that runs on a protocol known as the blockchain. While it does not mention the word blockchain, a 2008 paper by a person or people calling themselves Satoshi Nakamoto first described the use of a chain of blocks to verify transactions and engender trust in a network.² The key intention was to create a transaction system free from intervention by any central or monetary authority, be based on a mathematical algorithm instead of “third-party trust”, payments can be done electronically in a protected, verifiable and incontrovertible way. The application of this idea implies a payment system that all transactions happen directly between the owner and the receiver and is broadcast through a P2P network. Despite the information being public, the identity of the user is anonymous.

The blockchain has since evolved into a separate concept, and thousands of blockchains have been created using similar cryptographic techniques. This history can make the nomenclature confusing. Blockchain sometimes refers to the original Bitcoin blockchain. At other times, it refers to blockchain technology in general, or to any other specific blockchain, such as the one that powers Ethereum.

Any given blockchain consists of a single chain of discrete blocks of information, arranged chronologically. In principle, this information could include emails, contracts, land titles, marriage certificates, or bond trades. In theory, any type of contract between two parties can be established on a blockchain as long as both parties agree on the contract. This takes away any need for a third party to be involved in any contract and opens up a world of possibilities including peer-to-peer financial products, such as loans or decentralized savings and checking accounts, wherein banks or any intermediary are irrelevant.

Blockchain’s versatility has caught the eye of governments and private corporations; indeed, some analysts believe that blockchain technology will ultimately be the most impactful aspect of the cryptocurrency craze.

In Bitcoin's case, the information on the blockchain is mostly transactions. Bitcoin is really just a list. Person A sent X bitcoin to person B, who sent Y bitcoin to person C, etc. By tallying these transactions up, everyone knows where individual users stand. It's important to note that these transactions do not necessarily need to take place between humans.

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by Arthur Samuel in 1959 as- Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed. With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance. A machine has the ability to learn if it can improve its performance by gaining more data.

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

Regression is a supervised machine learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum." The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Regression-based models are the most prominent data-driven models widely used for the prediction task. The output of regression models is continuous rather than categorical. Regression models use the available data to find the relationship between the variables. A wide variety of regression models are available which are obtained after modifications in general regression model. All these models have their own merits and can be used for predictions. Bitcoin came into existence in 2009, but in the year 2013, active trading of bitcoin started. Since then, bitcoin prices have seemed to vary drastically making the prediction of bitcoin prices a challenging because of the price volatility and dynamism. In this project, we predict bitcoin prices based on the real-time data achieved using API key. We use several regression-based models for predictions. The result of the project will help us uncover the best regression-based model for bitcoin price prediction.

Chapter 2

Theoretical Background & Literature Survey

2.1 Literature Survey

Machine learning is a subset of artificial intelligence, which is capable of predicting the future based on past data. The research conducted by Hitam et al in their work "Comparative Performance of Machine Learning Algorithms for Cryptocurrency Forecasting" shows that machine learning models are known to deliver results similar to the actual results and also improve the accuracy of the results. Hence, Machine learning models have various advantages over other forecasting models.

Hamayel et al in their work "Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms" tests Long Short-Term Memory, Gated Recurrent Unit, bi-Long Short-Term Memory models performance on three different cryptocurrencies Bitcoin, Litecoin, and Ethereum. Model evaluation was done using the Mean Absolute Percentage Error. The results obtained from their study proved that the best model in prediction for all types of cryptocurrencies is the Gated Recurrent Unit.

Miura R. et al in their study "Artificial Neural Networks for Realized Volatility Prediction in Cryptocurrency Time Series" uses Realized volatility time series analysis for models like Artificial Neural Networks, Multi-Layer Perceptron, Gated Recurrent Unit, Long Short-Term Memory, Support Vector Machine, and ridge regression to predict bitcoin prices and compares the results with Auto-Regressive Realized Volatility. Their study summarizes that ridge regression performs better than all the other tested models based on model's Root Mean Squared Error values.

The studies of Mallqui et al and their research work "Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine

learning techniques” include prediction of bitcoin price direction and forecasting bitcoin prices including minimum, maximum, and closing prices based on daily data. Using Recurrent Neural Networks and tree classifier combined for prediction and Support Vector Machine for forecasting provided the best results.

The work of Patrick Jaquart et al in their work ”Short-term bitcoin market prediction via machine learning” includes various feature sets like technical, blockchain-based, sentiment, interest-based, and asset-based features to analyse that Recurrent Neural Networks and gradient boosting classifier performs best among various models and also proves that technical features are the best suitable feature set for prediction of bitcoin price.

S. McNally et al in their research ”Predicting the Price of Bitcoin Using Machine Learning”, predicts bitcoin price based on various models like, Bayesian optimised Recurrent Neural Network, Auto-Regressive Integrated Moving Average and Long Short-Term Memory. The outcomes of their study proves that Long Short-Term Memory outperforms all models and Auto-Regressive Integrated Moving Average forecast cannot outperform any non-linear deep learning model.

All the research work on bitcoin price prediction focuses on different models like and Long Short-Term Memory, Recurrent Neural Networks, Artificial Neural Networks, Gated Recurrent Unit, Support Vector Machine, Multi-Layer Perceptron, etc. So far, no research work tests all general categories of regressors for bitcoin price prediction. Hence, in this study we aim at testing eleven different regressor models for prediction of bitcoin prices.

2.2 Theoretical Background

Machine Learning is teaching the computer to perform and learn tasks without being explicitly coded. This means that the system possesses a certain degree of decision-making capabilities. Machine Learning can be divided into three major categories:

Supervised Learning In this ML model, our system learns under the supervision of a teacher. The model has both a known input and output used for training. The teacher knows the output during the training process and trains the model to reduce the error in prediction. The two major types of supervised learning methods are Classification and Regression.

Unsupervised Learning Unsupervised Learning refers to models where there is no supervisor for the learning process. The model uses just input for training. The output is learned from the inputs only. The major type of unsupervised learning is Clustering, in which we cluster similar things together to find patterns in unlabeled datasets.

Reinforcement Learning Reinforcement Learning refers to models that learn to make decisions based on rewards or punishments and tries to maximize the rewards with correct answers. Reinforcement learning is commonly used for gaming algorithms or robotics, where the robot learns by performing tasks and receiving feedback.

Machine learning tasks are of two categories. They are:

Classification In Classification, the output is discrete data. In simpler words, this means that we are going to categorize data based on certain features. For example, differentiating between Apples and Oranges based on their shapes, color, texture, etc. In this example shape, color and texture are known as features, and the output is “Apple” or “Orange”, which are known as Classes. Since the output is known as classes, the method is called Classification.

Regression In Regression, the output is continuous data. In this method, we predict the trends of training data based on the features. The result does not belong to a certain category or class, but it gives a numeric output that is a real number. For example, predicting House Prices is based on certain features like size of the house, location of the house, and no. of floors, etc.

In this study, we aim at comparing the performance of various regression-based models. Selected models for bitcoin price prediction are, Extra Tree Regressor, Gradient Boosting Machine Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regressor, LGBM Regressor, XGBoost Regressor, Stochastic Gradient Descent Regressor, Kernel Ridge Regressor, Bayesian Ridge Regressor, and Support Vector Machine Regressor.

2.2.1 Extra tree regressor

It is a tree-based ensemble method for supervised regression problems. The basic idea behind the extra tree regressor is to compute a meta estimator that fits several randomized decision trees (hence, the name, Extra tree) on various sub-samples of the dataset. In the extreme case, we obtain trees whose structures do not depend

on the output values of the learning sample. This algorithm provides accuracy and computational efficiency. Extra tree regressor takes several parameters as input; we can control the strength of randomization, splitting parameter, maximum depth, minimum sample leaves, etc. by adjusting those parameters based on requirements.

2.2.2 Gradient boosting machine regressor

Gradient Boosting Machine Regressor is a combination of two functionalities. First, steepest descent minimization, and second, stagewise additive expansion. In order to minimize the steepest descent, we can use the least-squares, least absolute deviation, and Huber-M loss function. The stagewise additive expansion is obtained by including additive components like the regression tree. Gradient Boosting Machine Regressor produces robust, interpretable, and competitive procedures. It is widely used for mining less than clean data.

2.2.3 Random Forest regressor

A combination of tree predictors is chosen as a random forest if they satisfy two conditions. Firstly, each tree predictor depends on the values of a random vector that is sampled independently, and secondly, the combination of tree predictors has the same distribution for all trees in the forest. As the number of trees in the forest increases, the generalization error converges to a limit as it depends on each of the trees in the forest and the correlation between them. This improves predictive accuracy and also helps control the overfitting problem. Random forest regressor has various parameters like max samples and bootstrap. In order to modify the random tree regressor, changes are to be made to its parameters. Internal estimates monitor error, strength, and correlation can be used to evaluate model performance.

2.2.4 Decision tree regressor

Decision Tree Regressor takes a dataset and keeps dividing the dataset. At each division of the dataset, an associated decision tree is built simultaneously. The outcome of such a process is a tree with nodes and leaf nodes. The root of the decision tree represents the best predictor. Each edge in a decision tree represents a possible attribute value. Leaf nodes on the other hand provide the final results. The decision tree regressor deals with continuous numeric values.

2.2.5 Linear regressor

The term linear in Linear Regression, according to algebra, provides a linear relationship between variables. Linear regression is a statistical procedure for calculat-

ing the value of a dependent variable based on the value of independent variable/s. The goal of the Linear Regression is to minimize the residual sum of squares errors, which is the linear approximation of the observed target and predicted target.

2.2.6 Light Gradient Boosting Machine regressor

Light Gradient Boosting Machine LGBM Regressor is a leafwise expansion algorithm. The selection of leaf nodes is made so as to maximize delta loss. This helps in reducing losses and improves accuracy. The 'Light' in LGBM regressor accounts for its high-speed computation with lower memory requirements to run. LGBM regressor is used to work with large datasets, in smaller datasets, the LGBM regressor suffers from overfitting.

2.2.7 Extreme Gradient Boosting regressor

Extreme Gradient Boosting XGBoost regressor like gradient boosting regressor is also a combination of two functionalities. First, Loss function minimization, and second, stagewise additive expansion of an objective function. The loss function is used to control the complexity of the algorithm. While the objective function determines nodes to be selected. In order to reduce overfitting and to obtain better training speed, XGBoost Regressor implements randomization techniques. The goal here is to reduce the computational complexity required to determine the best split. XGBoost uses pre-sorted data stored in a compressed column-based structure, and selective subset attributes selection based on the gain index which aids in the parallel computation of best split for all considered attributes.

2.2.8 Stochastic Gradient Descent Regressor

Stochastic Gradient Descent Regressor uses online feature selection and computes the gradient of the loss of the sample and updates the model hence reducing the learning rate eventually. It adds a penalty to the loss function which shrinks model parameters to zero vector. This process consumes lesser time compared to other batch processing algorithms. It is also efficient, and easy to implement. It is sensitive to feature scaling and requires a number of hyperparameters such as regularization parameters and the number of iterations.

2.2.9 Kernel Ridge Regressor

Kernel Ridge Regression chooses independent variables as regressors and performs ridge regression with a potentially infinite number of non-linear transformations.

Kernel Ridge Regression can be described as a blend of ridge regression and classification with kernel trick. Kernel trick is a process in which the model operates in high dimensions by computing the inner products between all pairs of data in the feature space of the dataset. Kernel Ridge Regression is widely used for pattern analysis.

2.2.10 Bayesian Ridge Regressor

Bayesian Ridge Regression is a probabilistic approach to the regression model. It allows regularization of parameter tuning based on the chosen dataset which is obtained by including uninformative priors over the hyperparameters of the model. Regularization of the model is obtained by maximizing the log marginal likelihood. Instead of setting precision manually, Bayesian Ridge Regression allows the use of gamma distribution to set precision based on the dataset under consideration.

2.2.11 Support Vector Machine Regressor

Support vector machine regression is a regression model which is non-parametric. The regression hyperplane in support vector machine regression is obtained by optimizing support vectors which consist of distance from the nearby data points. ϵ -insensitive loss instead of quadratic loss is used to measure the empirical error in support vector machine regression. Support vector machine regression uses the sequential minimal optimization method to minimize the cost function.

Chapter 3

Requirement Analysis

3.1 Hardware Requirements

Processor : Any Processor above 500 MHz

RAM : 4 GB

Hard Disk : 500 GB

System : Pentium IV 2.4 GHz

Any system with above or higher configuration is compatible for this project.

3.2 Software Requirements

- Operating system : Windows 7/8/9/10
- Programming lang : Python
- IDE : Jupyter Notebook
- Tools : Anaconda

Chapter 4

Design

4.1 Design Goals

The goal of this project is to predict the highest and closing price of Bitcoin on a given day based on the Bitcoin data of several preceding quarters. It is technically challenging to predict the accurate price, mainly due to lack of seasonality and highly volatile nature of the cryptocurrency market. This is primarily a statistic prediction drawback. To predict bitcoin price using regression-based models, we gather information about regressor models and then we implement data gathering, pre-processing, and model building using the Python programming language and various libraries such as pandas, scikit-learn, lightgbm, XGBoost, seaborn, and statsmodel. The constant increase in bitcoin usage has become an extremely serious problem, with the development of technology and hi-tech tools having a significantly greater impact on the bitcoin price. The large amounts of information also pose's a challenge to analyze such data and identify similarities or relations between the data. Also there is a challenge of inconsistency that can occur in the data due to incompleteness in the dataset. Therefore, there is an urging need of proper techniques to analyze large volumes of data to get some useful results out of it. So the main aim of this project is to propose a general and effective approach to predict the bitcoin price using regression-based techniques.

4.2 System Architecture

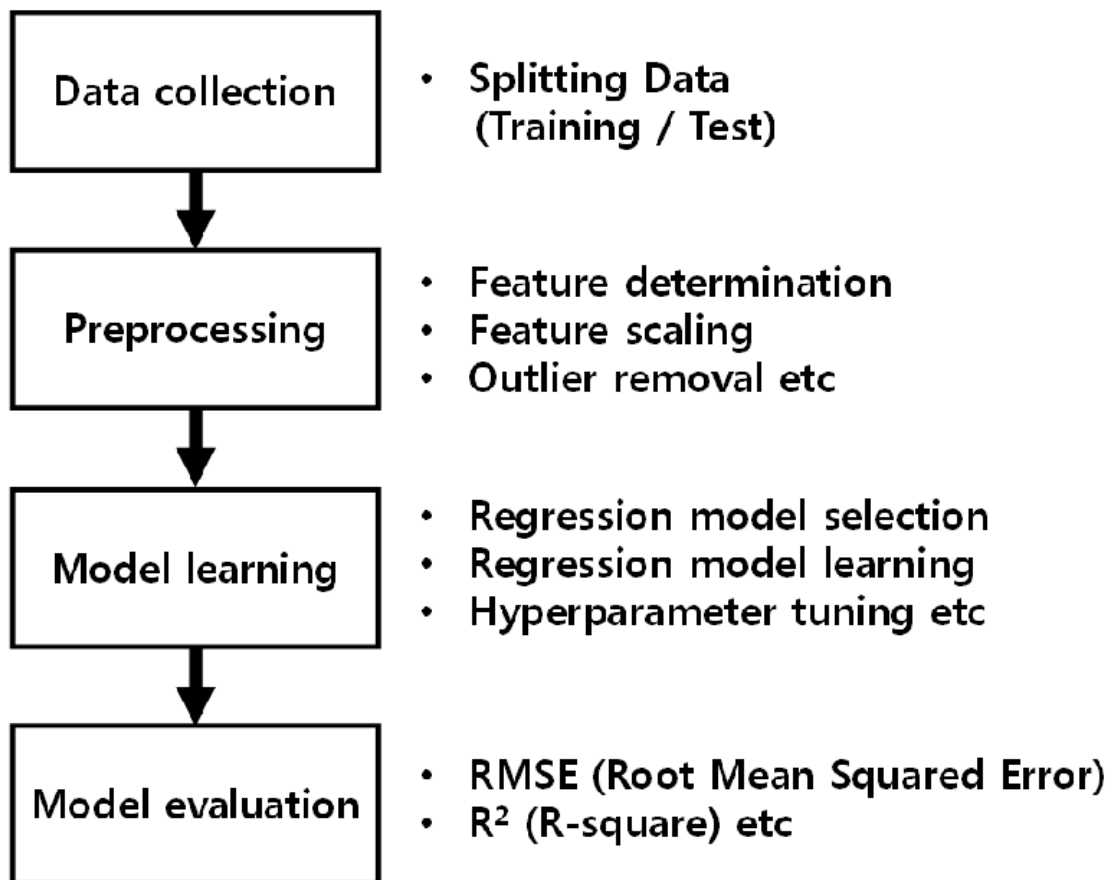


Figure 4.1: Flow chart of regression-based models

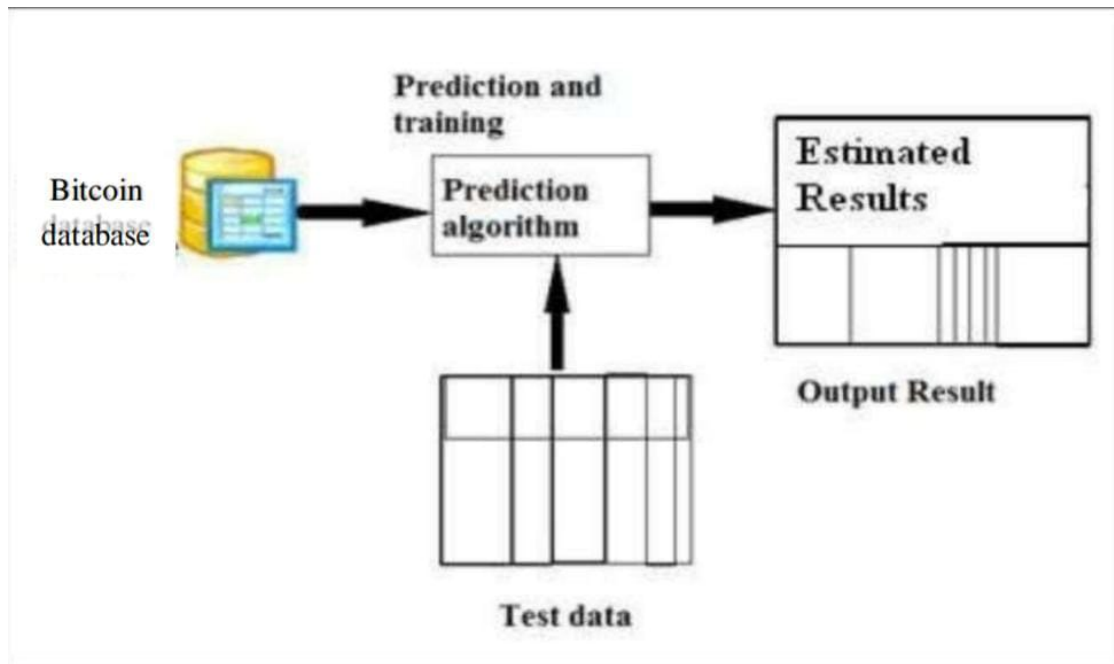


Figure 4.2: System Architecture

The architecture of the proposed system has the following components:

Training data to train the models.

Testing data to apply the models.

Prediction algorithms.

Models evaluation

Comparison of the results.


Chapter 5

Methodology

5.1 Dataset

Several Bitcoin data sets are available online to download for free. Most of them provide the data related to price of Bitcoin on a minute to minute basis. However, the top goal of the project is to create one-day ahead prediction of highest and shutting worth of Bitcoin. So, we will need data such as highest and closing price of Bitcoin for each day over period of several years. The "coinmarketcap.com" API provides the Bitcoin worth knowledge set. This API gives access to Bitcoin exchanges and daily Bitcoin values. It permits users to customise the question whereas victimisation the interface to transfer the historical Bitcoin costs. The data is available in three different formats i.e JSON, XML and CSV. Data is accessed in as a JSON object. This API key provides data from 3, January 2009 to the present date. The dataset contains values like the date (UNIX timestamp), open (USD), high (USD), low (USD), close (USD), and volume (USD).

```
# Fetching data from the website
url = "https://web-api.coinmarketcap.com"
parameter = {"convert": "USD",
             "slug": "bitcoin",
             "time_end": 9999999999,
             "time_start": "1483228800"}
content = requests.get(url=url, params=param).json()
dataframe = pd.json_normalize(content['data']['quotes'])
```



Date	Open*	High	Low	Close**	Volume	Market Cap
Dec 13, 2021	\$50,114.74	\$50,205.00	\$45,894.85	\$46,737.48	\$32,166,727,776	\$883,345,705,454
Dec 12, 2021	\$49,354.85	\$50,724.87	\$48,725.85	\$50,098.34	\$21,939,223,599	\$946,823,114,555
Dec 11, 2021	\$47,264.63	\$49,458.21	\$46,942.35	\$49,362.51	\$25,775,869,261	\$932,873,019,896
Dec 10, 2021	\$47,642.14	\$50,015.25	\$47,023.70	\$47,243.31	\$30,966,005,122	\$892,777,676,312
Dec 09, 2021	\$50,450.08	\$50,797.17	\$47,358.35	\$47,672.12	\$29,603,577,251	\$900,835,068,960
Dec 08, 2021	\$50,667.65	\$51,171.38	\$48,765.99	\$50,504.80	\$28,479,699,446	\$954,312,404,511
Dec 07, 2021	\$50,581.83	\$51,934.78	\$50,175.81	\$50,700.08	\$33,676,814,852	\$957,948,940,989

Figure 5.1: Overview of Bitcoin Dataset

5.2 Data Preprocessing

The primary knowledge collected from the web sources remains within the raw kind of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. The following steps square measure concerned within the process of primary knowledge. A huge volume of information collected through field survey must be sorted for similar details of individual responses..

Data Preprocessing could be a technique that's accustomed convert the raw knowledge—data—information into a clean data set. In alternative words, whenever the info is gathered from totally different sources it's collected in raw format that isn't possible for the analysis. Therefore, bound steps square measure dead to convert the knowledge— the info— the information into a little clean data set. This technique is performed before the execution of reiterative Analysis. The set of steps is understood as knowledge preprocessing.. The process comprises:

- * Data Cleaning
- * Data Normalization
- * Data Splitting

5.2.1 Data Cleaning

The Data obtained from the API key will be in JSON format. We take the JSON data using get () method of requests library and convert it into a pandas data frame. We also change UNIX timestamp into localized timestamp. We then create a new feature in dataset called 'Mean' which holds the value of average of 'Low' and 'High'. Now, drop all the NaN and NULL values from the data frame. Set the index of data frame as 'Date'. We also create a new feature 'Actual' consisting of values of next day's 'Mean' and through this study we will try to predict this value.

5.2.2 Data Normalization

Data Normalization is the process of reducing the values of all the features to a small range of values such that, no information is lost during the process and working with normalized dataset provides better understanding of dataset. For this study, we use MinMaxScaler for normalization.

MinMaxScaler scales all the data features in the range $[0, 1]$ or else in the range $[-1, 1]$ if there are negative values in the dataset. This scaling compresses all the inliers in the narrow range $[0, 0.005]$. In the presence of outliers, StandardScaler does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation. This leads to the shrinkage in the range of the feature values.

We determine the dependent variable Y as ‘Actual’ and independent variable X as ‘Open’, ‘Close’, ‘High’, ‘Low’, ‘Volume’, and ‘Mean’ and then normalize X and Y using MinMaxScaler.

5.2.3 Data Splitting

We split the dataset based on its length into two sets, with training set consisting of 90% of data and test set consisting of 10% of data.

```
train_size=int(len(df) *0.9)
test_size = int(len(df)) - train_size
train_X, train_y = X[:train_size].dropna(),
                  y[:train_size].dropna()
test_X, test_y = X[train_size:].dropna(),
                y[train_size:].dropna()
```


5.3 Data Analysis

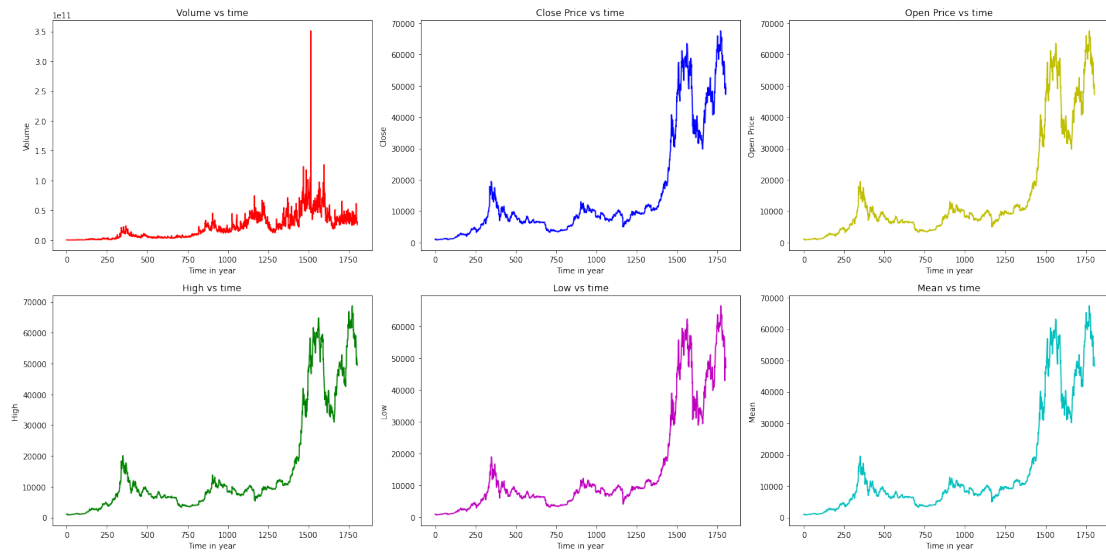


Figure 5.2: Dataset Features Analysis

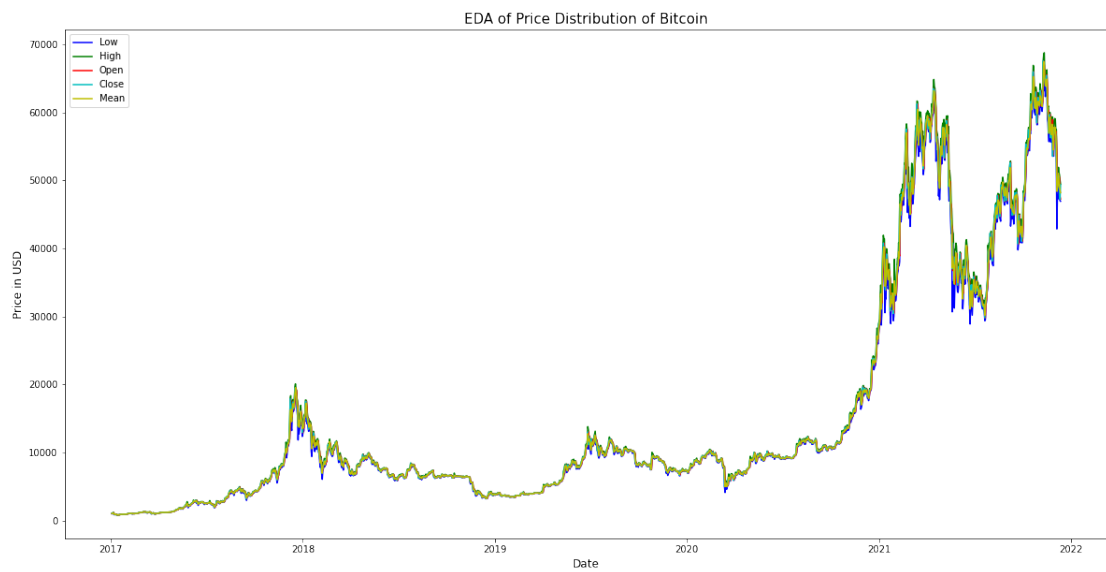


Figure 5.3: Price Distribution Analysis

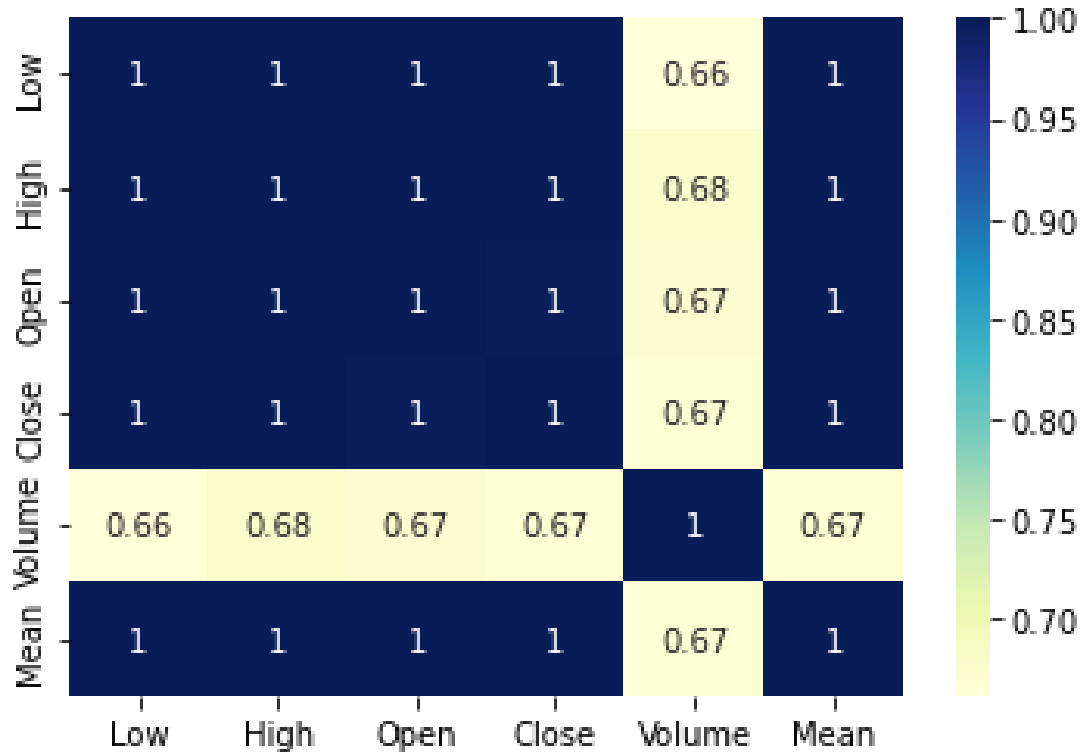


Figure 5.4: Heatmap of Bitcoin dataset

Figure 5.2 is the data distribution plot of the dataset. It displays the changes of all the features with respect to the time (UNIX timestamp). We can adhere from the Figure 5.2 that the price of bitcoin is volatile in nature. We split the dataset based on its length into two sets, with training set consisting of 90% of data and test set consisting of 10% of data. From Figure 5.2, we can also observe that plots for Close, Open, High, Low and Mean appear to be very similar. Hence, we build another plot for verifying the same. Figure 5.3 plots Close, Open, High, Low and Mean versus Time and we observe that there exist little to no difference in all these values. Figure 5.4 shows the heat map of the dataset which explains the correlation between various features of the dataset. We observe that, correlation of Close, Open, High, Low and Mean is equal to 1, which means all these features are highly correlated and the correlation index of Volume is 0.66 to 0.68 which is relatively less than 1. Therefore, from Figure 5.2, Figure 5.3 and Figure 5.4 we can ignore the features Close, Open, High, and Low and use Mean as an equivalent measure.

5.4 Machine Learning Models Training

The training dataset is used for training all the selected models which are:

Extra Tree Regressor

Gradient Boosting

Machine Regressor
Random Forest Regressor
Decision Tree Regressor
Linear Regressor
LGBM Regressor
XGBoost Regressor
Stochastic Gradient Descent Regressor
Kernel Ridge Regressor
Bayesian Ridge Regressor
Support Vector Machine Regressor

The training data is then fitted into each of these models, and parameters are set based on each model's requirements.

5.5 Machine Learning Models Testing

After the models are trained on the training dataset, we use testing dataset for predictions using the predict method. The results from the predictions are stored for future reference. To evaluate the performance of all the selected models and to provide a comparison between the models, we use two different metrics. The first comparison metric appointed is score and the second comparison metric chosen is Root Mean Square Error. The score method computes the r-squared value. R-squared value can be defined as the coefficient of determination and hence, provides the accuracy of a model. Root Mean Square Error or RMSE is a risk function that determines the square root of the average squared difference between the predicted and the actual value of a feature or variable.

Chapter 6

Results

Index	Model Name	Accuracy Score	RMSE
1	Extra Tree Regressor	98.40619233122447	0.12934686088
2	Gradient Boosting Machine Regressor	97.7946186680286	0.15215274232
3	Random Forest Regressor	98.32092472096599	0.13276176607
4	Decision Tree Regressor	97.47328695888719	0.16286055039
5	Linear Regressor	99.18191692466188	0.0926694619
6	LGBM Regressor	95.99145243335282	0.20513107593
7	XGBoost Regressor	97.70733885359043	0.15513436264
8	Stochastic Gradient Decent Regressor	92.81290405993839	0.27467216985
9	Kernel Ridge Regressor	98.44817636148377	0.127631866
10	Bayesian Ridge Regressor	99.18239168659156	0.09264256835
11	Support Vector Machine Regressor	82.40662266460114	0.4297466955

Table 6.1: Regression-based Models results comparison

Based on the score and RMSE appointed as the model evaluation metrics we compare the results from all the selected models. The results show that Bayesian Ridge Regression outperforms all the other models with score of 99.182% and RMSE value 0.09464 giving a close combat is Linear Regression with score of 99.181% and RMSE value 0.09466.

6.1 Visualization of Results



Figure 6.1: Extra Tree Regressor Actual vs Predicted Bitcoin Price

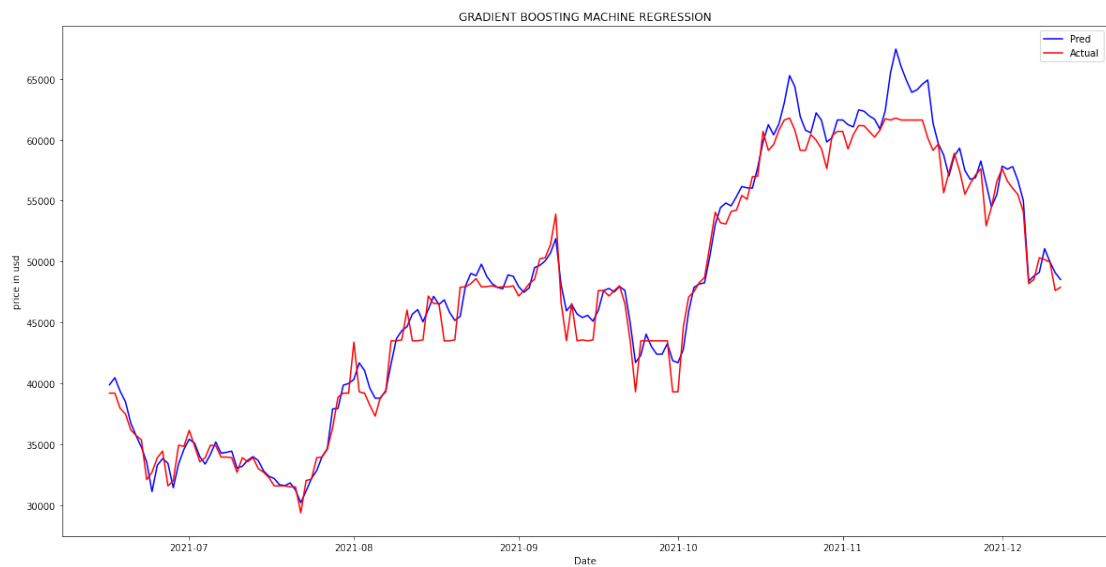


Figure 6.2: Gradient Boosting Machine Regressor Actual vs Predicted Bitcoin Price



Figure 6.3: Random Forest Regressor Actual vs Predicted Bitcoin Price

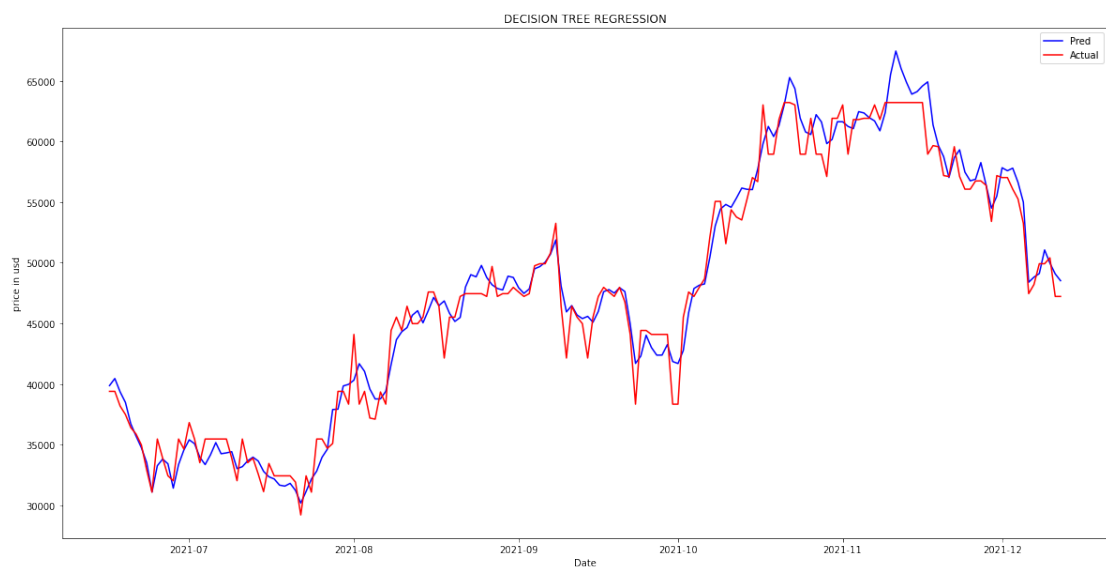


Figure 6.4: Decision Tree Regressor Actual vs Predicted Bitcoin Price

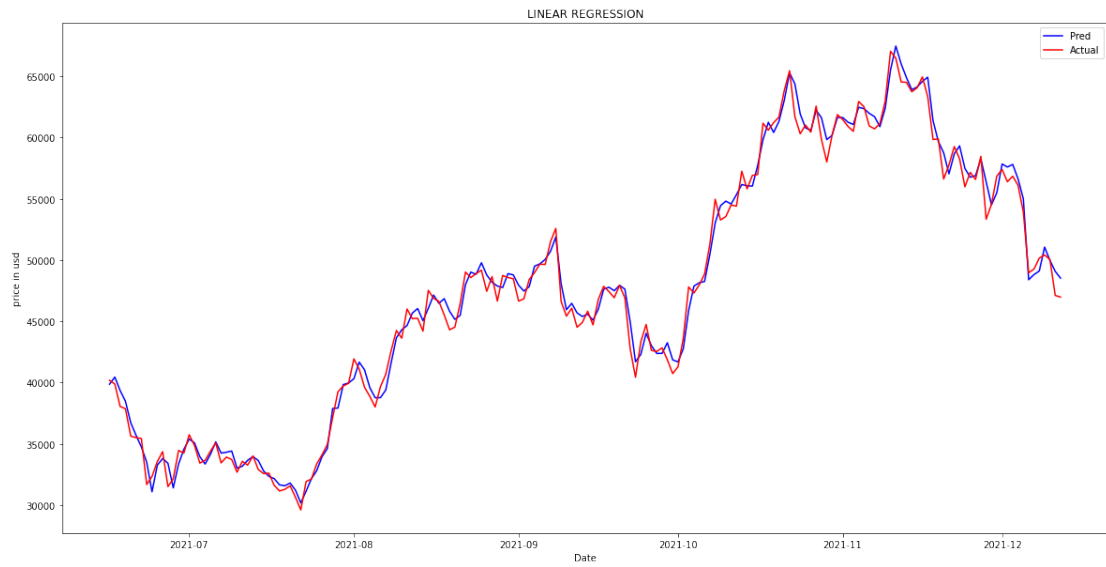


Figure 6.5: Linear Regressor Actual vs Predicted Bitcoin Price

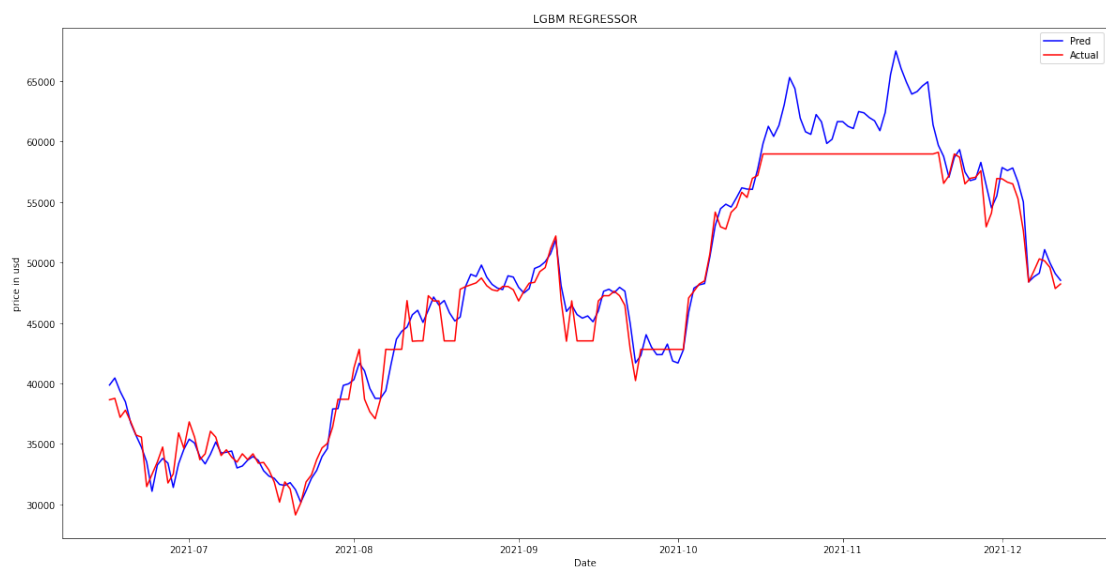


Figure 6.6: LGBM Regressor Actual vs Predicted Bitcoin Price



Figure 6.7: XGBoost Regressor Actual vs Predicted Bitcoin Price

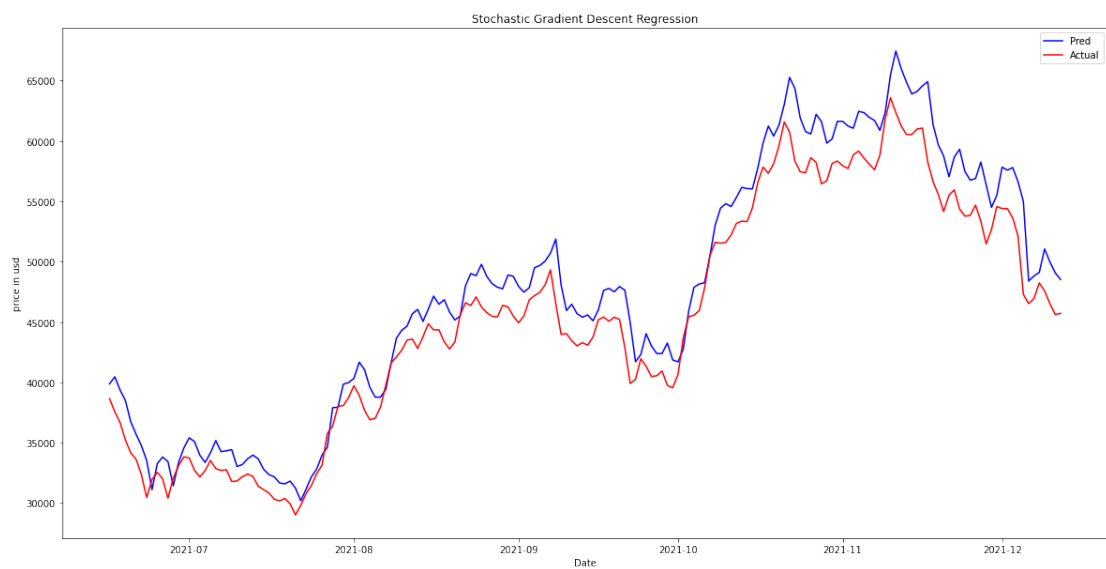


Figure 6.8: Stochastic Gradient Descent Regressor Actual vs Predicted Bitcoin Price

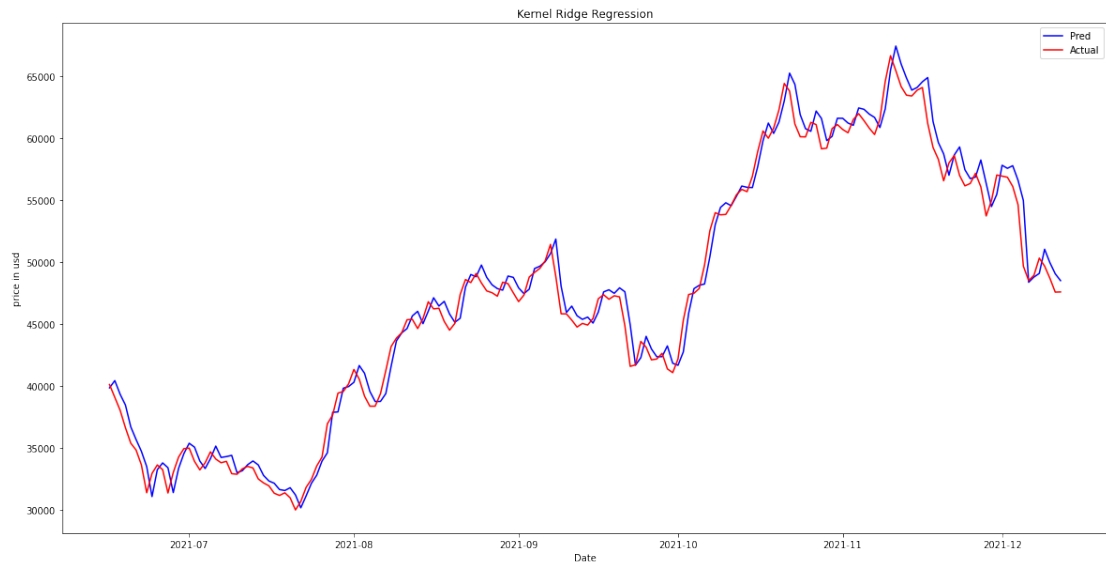


Figure 6.9: Kernel Ridge Regressor Actual vs Predicted Bitcoin Price

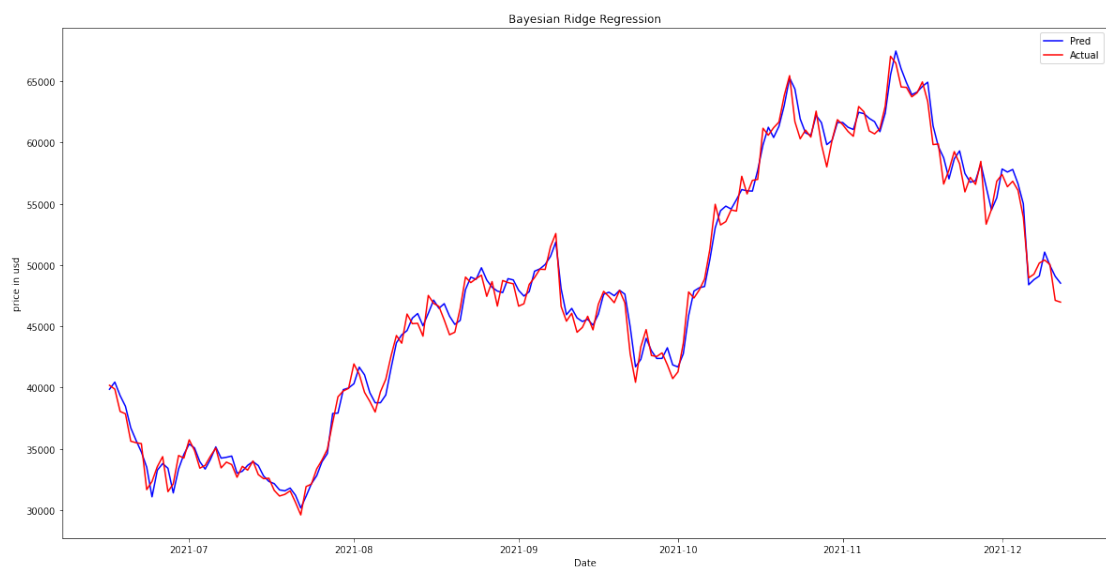


Figure 6.10: Bayesian Ridge Regressor Actual vs Predicted Bitcoin Price



Figure 6.11: Support Vector Machine Regressor Actual vs Predicted Bitcoin Price

The graphs below depicts actual bitcoin prices versus predicted bitcoin prices for all the selected models. We can interpret from the graphs that Bayesian Ridge Regressor outperforms all the other regression based predictive models in predicting the price of bitcoin.

Chapter 7

Conclusion

In this project, we predict the price of bitcoin using eleven different regression-based techniques named as Extra Tree Regressor, Gradient Boosting Machine Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regressor, LGBM Regressor, XGBoost Regressor, Stochastic Gradient Descent Regressor, Kernel Ridge Regressor, Bayesian Ridge Regressor, and Support Vector Machine Regressor. Using RMSE and accuracy score as metrics, we evaluate all the models and obtain promising results with accuracy scores as high as 99.182 and RMSE scores as good as 0.09264 obtained by Bayesian ridge regressor. Linear regressor also provided metrics very close to Bayesian ridge regressor. The accuracy score of 82.406 and RMSE score of 0.429 for Support Vector Machine Regressor are the least obtained values among all the models. With this project, we can conclude that all the regression-based models perform extraordinarily well for predicting the price of bitcoin.

Chapter 8

Future Work

The future scope for this project includes the following possibilities:

Forecasting Bitcoin Price using Machine Learning.

Creating a user interface to access Bitcoin historical data along with Bitcoin forecasting data.

Embedding Bitcoin price prediction and forecasting model into a mobile application for ease of access.

Facilitating backup creation for the mobile application.

Bibliography

- Xu, M., Chen, X. Kou, G. A systematic review of blockchain. *Financ Innov* 5, 27 (2019).
- Kayal, P., Rohilla, P. (2021). Bitcoin in the economics and finance literature: a survey. *SN business economics*, 1(7), 88.
- Araghinejad S. (2014) Regression-Based Models. In: *Data-Driven Modeling: Using MATLAB in Water Resources and Environmental Engineering*. Water Science and Technology Library, vol 67. Springer, Dordrecht.
- Hitam, N.A.; Ismail, A.R. Comparative Performance of Machine Learning Algorithms for Cryptocurrency Forecasting. *Indones. J. Electr. Eng. Comput. Sci.* 2018, 11, 1121–1128.
- Hamayel, M.J.; Owda, A.Y. A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. *AI* 2021, 2, 477–496.
- Miura R., Pichl L., Kaizoji T. (2019) Artificial Neural Networks for Realized Volatility Prediction in Cryptocurrency Time Series. In: Lu H., Tang H., Wang Z. (eds) *Advances in Neural Networks – ISNN 2019*. ISNN 2019. *Lecture Notes in Computer Science*, vol 11554. Springer, Cham.
- Mallqui, D.C.A., Fernandes, R.A.S.: Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. *Appl. Soft Comput.* 75, 596–606 (2019) Website.
- S. McNally, J. Roche and S. Caton, "Predicting the Price of Bitcoin Using Machine Learning," 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), 2018, pp. 339-343.

- Geurts, P., Ernst, D. Wehenkel, L. Extremely randomized trees. *Mach Learn* 63, 3–42 (2006).
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Breiman, L. Random Forests. *Machine Learning* 45, 5–32(2001).
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, WeiChen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017.
- LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149–3157.
- Bentéjac, Candice Csörgő, Anna Martínez-Muñoz, Gonzalo. (2019). A Comparative Analysis of Website.
- Tsuruoka, Yoshimasa Tsujii, Jun'ichi Ananiadou, Sophia. (2009). Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty.
- Peter Exterkate, Model selection in kernel ridge regression, *Computational Statistics Data Analysis*, Volume 68, 2013, Pages 1-16, ISSN 0167-9473.
- Section 3.3 in Christopher M. Bishop: *Pattern Recognition and Machine Learning*, 2006.
- Kanka Goswami, G L Samuel, Support vector machine regression for predicting dimensional features of diesinking electrical discharge machined components, *Procedia CIRP*, Volume 99, 2021, Pages 508-513, ISS 2212-8271.
- P. Rivas-Perea, J. Cota-Ruiz, D. Chaparro, J. Venzor, A. Carreón and J. Rosiles, "Support Vector Machines for Regression: A Succinct Review of Large-Scale and Linear Programming Formulations," *International Journal of Intelligence Science*, Vol. 3 No. 1, 2013, pp. 5-14.
- Li G, Zrimec J, Ji B, et al. Performance of Regression Models as a Function of Experiment Noise. *Bioinformatics and Biology Insights*. January 2021.