A PROJECT ON

"Crypto Crystal Ball: Using Machine Learning to Forecast the Price of Trending Digital Currencies"

SUBMITTED IN
PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE COURSE OF DIPLOMA
IN BIG DATA ANALYSIS (PG-DBDA)

SUBMITTED BY:

Anisha Gupta(220943025006)
Raunak Mudgal (220943025032)

UNDER THE GUIDANCE OF:
Ms. Trupti Joshi

CERTIFICATE

This is to certify that the project work under the title 'Crypto Crystal Ball: Using Machine Learning to Forecast the Price of Trending Digital Currencies' is done by Anisha Gupta & Raunak Mudgal in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.(PG-DBDA).

Ms Trupti Joshi
Project Guide

Date: 13/3/2023

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Ms Trupti Joshi mam and Mr. Tushar B. Kute sir.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form. Last but not the least we thank the entire faculty and the staff members of KNOW IT Cdac , Pune for their support.

FROM-
Anisha Gupta
Raunak Mudgal
DBDA Sept 2022 Batch,

TABLE OF CONTENTS

1. Introduction

1.1 Introduction and Objectives:

In recent years, the world of digital currencies has gained immense popularity and has become a significant player in the global financial market. With the emergence of new cryptocurrencies, the market has become more complex, and predicting the price of these currencies has become increasingly difficult. In this context, the use of PySpark(Mlib) algorithms has emerged as a promising approach for predicting the prices of these currencies.

The Crypto Crystal Ball is one such tool that utilizes machine learning techniques to forecast the prices of trending digital currencies. This tool employs various factors such as historical data, market trends, and other relevant information to make accurate predictions about the future price of cryptocurrencies. By using advanced algorithms, the Crypto Crystal Ball provides users with insights into the market's behavior, allowing them to make informed decisions about buying and selling digital currencies.

- A cryptocurrency is a digital currency, which is an alternative form of payment created using encryption algorithms. The use of encryption technologies means that cryptocurrencies function both as a currency and as a virtual accounting system. To use cryptocurrencies, you need a cryptocurrency wallet
- Cryptocurrency such as Bitcoin are more popular these days among investors.
- To predict the bitcoin price accurately, firstly , we identify the daily trends in the bitcoin price while gaining insight into the optimal features surrounding the bitcoin price. Secondly, using the available information, we will predict the sign of the daily price change with highest possible accuracy.
- Data used in the project is structured in nature from year 2020.1 to 2021.5. It was collected from www.kaggle.com. To assess the accuracy of the developed prediction model and identify ways to improve its performance.This research uses Decision Tree Regressor, Random Forest Regressor, GBT Regressor, Linear Regression and K-Means.

Here, we will explore the workings of the Crypto Crystal Ball, its benefits, and the impact it can have on the world of digital currencies. We will delve into the PySpark(Mlib) techniques used by the tool and how it can help traders and investors stay ahead of the curve in the volatile world of cryptocurrencies.

The main feature will be PySpark (Mlib) , in which we will be using algorithms such as Decision Tree Regressor Algorithm, Linear Regression Algorithm, Random Forest Algorithm, Gradient Boosting Regressor and K-Means Clustering Algorithm which will predict accurately the mean square error and mean absolute error and Also, will find which algorithm gives a faster and efficient result by comparatively-comparing each one of them on the basis of relative measure.

1.2 Why does this problem need to be solved?

The problem of accurately predicting the prices of digital currencies is critical because it can greatly impact traders, investors, and the overall market. As cryptocurrencies gain more popularity and mainstream adoption, more people are investing in them, and the market's volatility has increased.

Predicting the price of cryptocurrencies accurately is challenging due to several factors, including their decentralized nature, the lack of regulatory oversight, and their susceptibility to market sentiment and speculation. The ability to accurately forecast the price of these digital currencies can enable traders and investors to make informed decisions and minimize their financial risks.

Moreover, accurate price predictions can help reduce market manipulation, fraudulent activities, and increase investor confidence in the market. Machine learning algorithms have the potential to analyze vast amounts of data and identify patterns that humans may not be able to detect, making them useful in predicting the prices of cryptocurrencies accurately. Therefore, the development of tools such as the Crypto Crystal Ball can significantly benefit the crypto market by providing accurate insights, helping traders and investors make better decisions, and increasing overall market stability.

PySpark is a powerful tool that enables users to process and analyze large datasets using distributed computing, making it an ideal solution for this problem.

PySpark is a Python library that runs on top of Apache Spark, a distributed computing framework. PySpark allows users to write code in Python and execute it on a distributed computing environment, enabling faster processing of large datasets. The distributed computing capability of PySpark makes it possible to perform computations on multiple machines in parallel, significantly reducing the time required to process large datasets.

2. Problem Definition and Algorithm:

2.1 Problem Definition

The problem definition of our project is to develop a machine learning-based tool, the Crypto Crystal Ball that can accurately forecast the prices of trending digital currencies. The tool will utilize historical data, market trends, and other relevant information to predict the future price of digital currencies accurately.

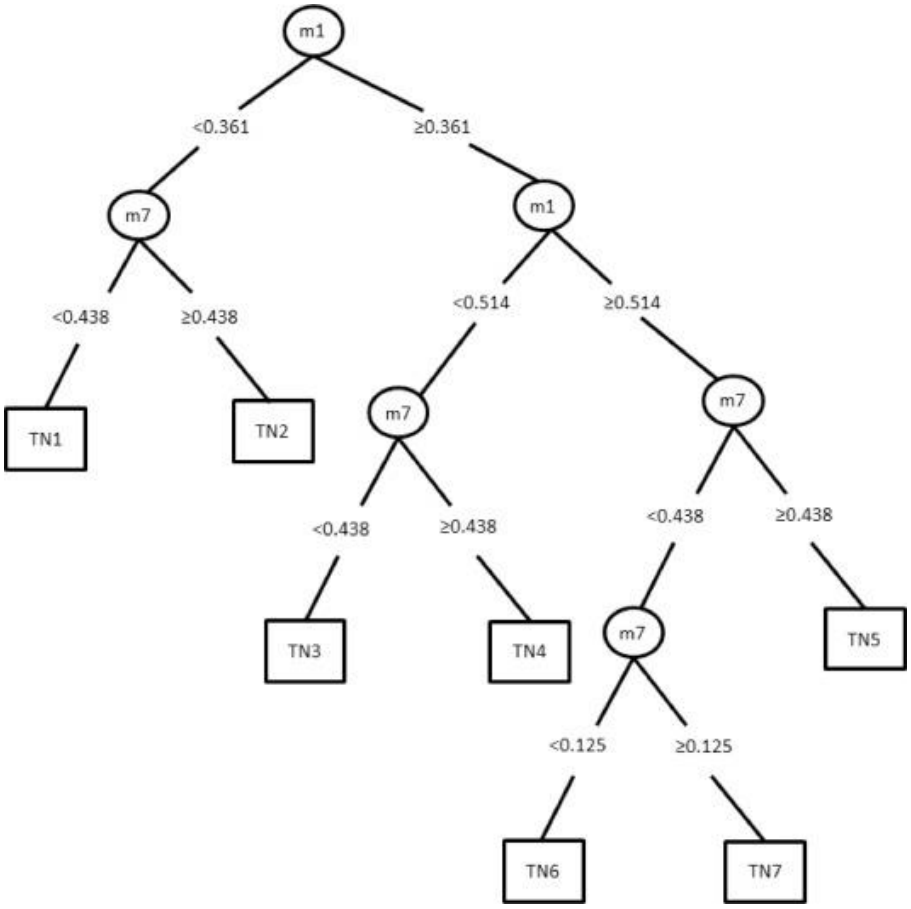
The main goal of the project is to help traders and investors make informed decisions about buying and selling digital currencies, ultimately maximizing their returns while minimizing their risks. The tool will also help reduce market manipulation and fraudulent activities by providing accurate price predictions.

To achieve this goal, the project will involve collecting and preprocessing large datasets of historical data related to digital currencies, building and training machine learning models using PySpark, and deploying the models in a scalable, distributed computing environment. The final product will be a user-friendly tool that provides real-time price predictions and insights into the behavior of digital currencies in the market.

2.2 Algorithm Definition

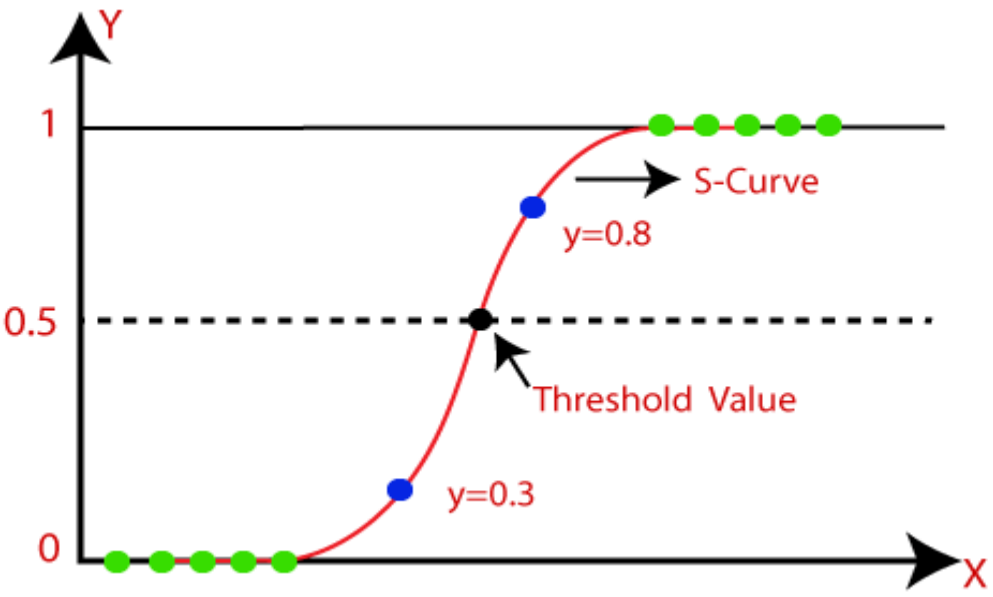
Decision Tree Regressor: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce

meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.



Logistic Regression: Logistic Regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the output of categorical dependent variables. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or false, etc, but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is used for solving classification problems. In logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

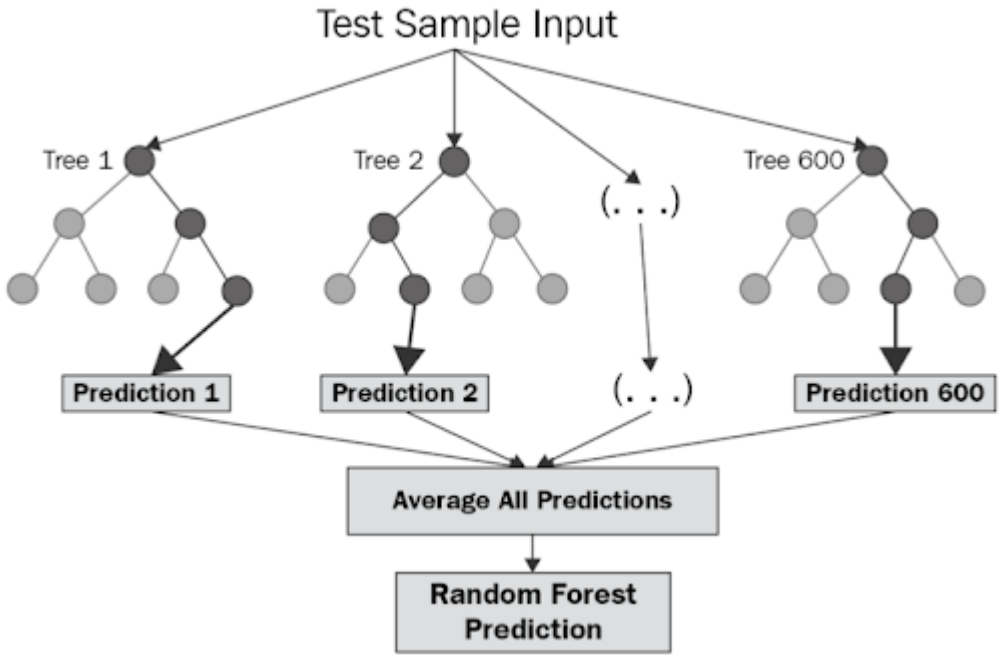


Random forest Regressor: Random Forest Regressor is a UnSupervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by

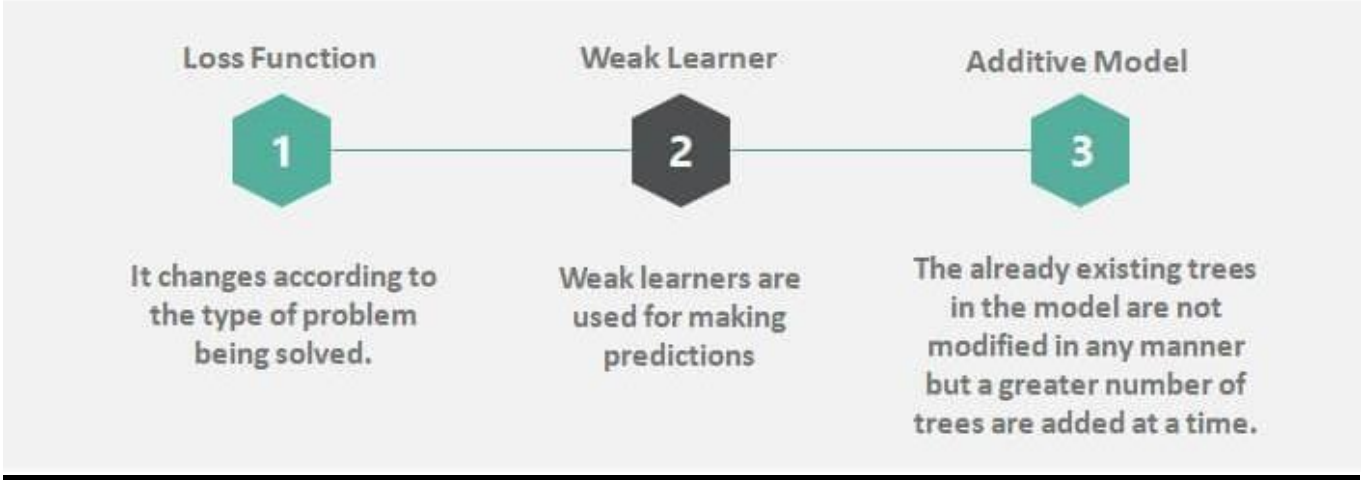
using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.



Gradient Boosting Regressor:

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.
There is a technique called the **Gradient Boosted Trees** whose base learner is CART (Classification and Regression Trees).



3.Cryptoucrrency Analysis Classification:

There are several types of analysis that can be done on our cryptocurrency project. We can classify these analyses into the following categories:

Price analysis: You could analyze historical Bitcoin prices and trends to identify patterns and make predictions about future price movements. This could involve looking at factors such as supply and demand, market sentiment, and global events that could impact the cryptocurrency market.

Volume analysis: You could also look at trading volume data to identify trends and patterns in buying and selling behavior. This could help you identify market trends and make predictions about future price movements.

Sentiment analysis: Another approach would be to analyze social media data and news articles to gauge public sentiment around Bitcoin. This could involve using natural language processing techniques to identify sentiment and make predictions about how people will react to news and events.

Market cap analysis: You could also analyze Bitcoin's market capitalization, which is the total value of all Bitcoin in circulation. This could help you identify trends and patterns in investor sentiment and make predictions about future price movements.

Network analysis: Finally, you could analyze Bitcoin's network data to gain insights into the behavior of users and miners. This could involve looking at factors such as transaction volume, mining difficulty, and network hash rate to make predictions about Bitcoin's future performance.

Technical analysis: This type of analysis involves examining historical price and volume data to identify trends and patterns that can be used to predict future price movements. Technical analysis techniques include chart analysis, trend analysis, and statistical analysis.

Machine learning-based analysis: This type of analysis involves using machine learning algorithms to analyze large datasets and identify patterns that can be used to predict future price movements. Machine learning-based analysis techniques include regression analysis, classification analysis, and clustering analysis.

4. Platform used :

- **PYTHON**

We have used the Python programming language to build our project. Using various libraries of python we have created interactive graph and interface. The algorithms of machine learning are build using PySpark library.

- **TABLEAU**

For data visualization we have used Tableau Public. Using this we created various graph of the features.
analyzing your data in a distributed environment. PySpark supports most of Spark’sfeatures such as Spark SQL, Data Frame, Streaming, MLlib (Machine Learning) and Spark Core.

- **PySpark:**

PySpark is an interface for Apache Spark in Python. It not only allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively.

Flow Diagram

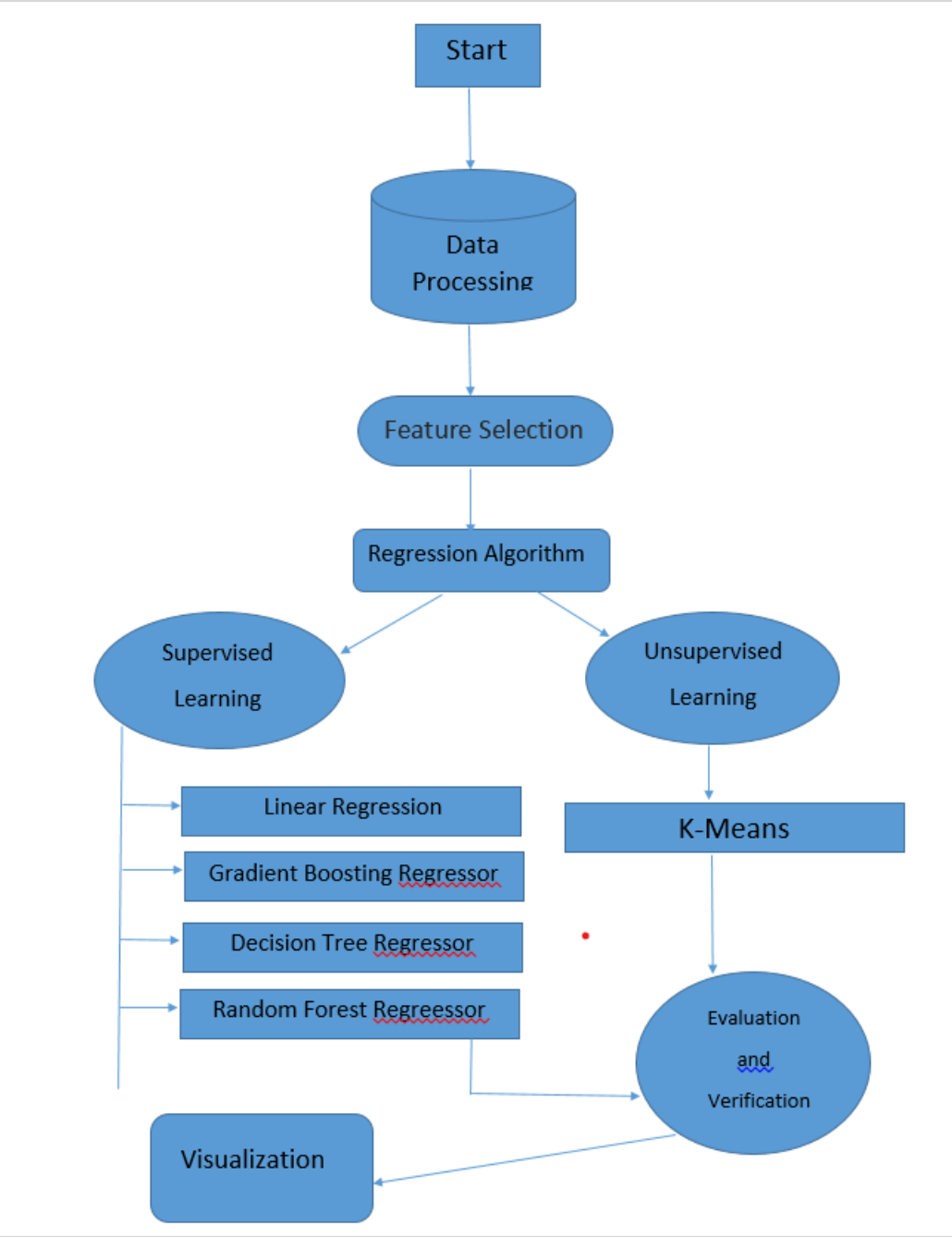


Figure: Process of Bitcoin Price Prediction

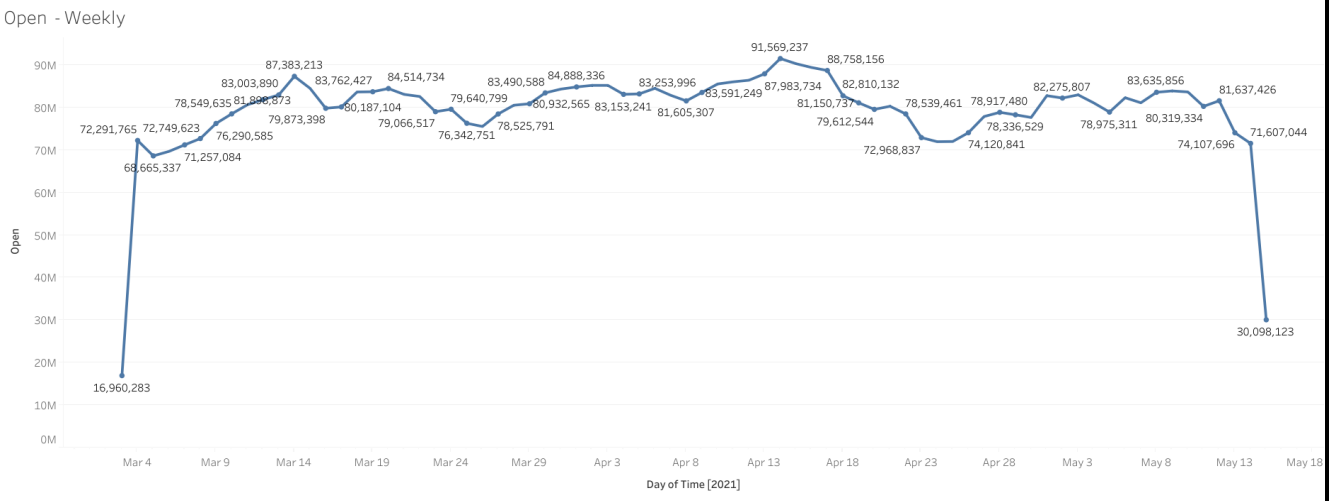
5. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an important step in any data analysis project, including our Bitcoin cryptocurrency prediction project. EDA helps to identify patterns and relationships in the data, as well as to detect any anomalies or outliers that may require further investigation.

For our project, we used Tableau for exploratory data analysis. Tableau is a powerful data visualization tool that allows us to quickly and easily explore large datasets. We used Tableau to create various visualizations that helped us to understand the data better and identify any trends or patterns that may be useful in our analysis.

The first step of our EDA process was to import the Bitcoin dataset into Tableau. Once the data was imported, we started by exploring the different variables in the dataset. We created histograms and box plots to visualize the distribution of the data and identify any outliers or anomalies. We also used scatter plots to explore the relationship between different variables and identify any correlations.

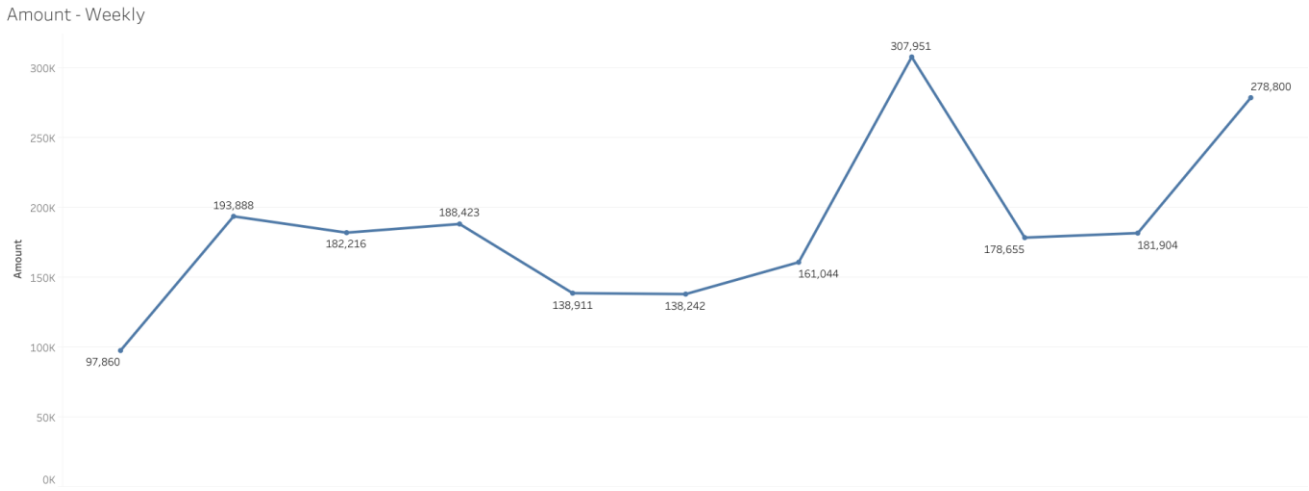
Next, we created time series plots to explore the Bitcoin price over time. We used line charts to visualize the Bitcoin price trend over time and identify any patterns or trends..



Finally, we used Tableau to create interactive dashboards that allowed us to explore the data in more detail. These dashboards included multiple visualizations that could be filtered and sorted based on different criteria, such as date or price.

Overall, Tableau was a powerful tool that allowed us to explore the Bitcoin cryptocurrency dataset in depth and identify any trends or patterns that may be useful in our analysis. By visualizing the data in different ways, we were able to gain a better understanding of the data and identify any areas that may require further investigation.

5.1 Tableau Summary:



Our exploratory data analysis using Tableau involved creating weekly and daily line charts to visualize the trends and changes in Bitcoin cryptocurrency prices over time. The weekly line chart showed that the price of Bitcoin had a steady increase in the first half of the year, followed by a decline in the second half. The daily line chart provided a more detailed view of the Bitcoin price fluctuations, showing significant volatility in the early part of the year and a stabilization in the second half. Overall, the visualizations helped us to identify key trends and patterns in the Bitcoin cryptocurrency market.

6. Results and discussion:

- The actions performed in this work are done by the Laptop with an intel corei5 8thGen processor and developed the code using python (Spark MLib). The algorithms used in this work are Linear Regression, Random Forest Regressor, Decision Tree Regressor ,Gradient Boosting Regressor and K-Means Clustering.

6.1 Algorithms used with Accuracy

PySpark is an interface for Apache Spark in Python. It not only allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively

Algorithms Used	Root Mean Squared Error
Decision Tree Regressor	628.423
Random Forest Regressor	773.024
GBT Regressor	444.304
Linear Regression	15.2454
K-Means Clustering	27720.4

7. Future work And Conclusion

7.1 Future Work:

There are several potential areas for future work in your Bitcoin cryptocurrency prediction project. Here are some ideas:

Feature engineering: You can explore and experiment with different sets of features and combinations of features to see if they improve the model's accuracy. For example, you could add technical indicators such as moving averages or trading volumes.

Hyperparameter tuning: You can tune the hyperparameters of the regression models to try to find the best possible combination of parameters that improves the model's accuracy. You can use grid search or random search techniques to search for the optimal hyperparameters.

Ensembling methods: You can explore ensemble methods such as bagging, boosting, and stacking to improve the accuracy of the model. You can experiment with different combinations of models to see which ones perform best.

Deep learning: You can explore deep learning models such as LSTM, GRU, or Convolutional Neural Networks (CNN) to see if they can outperform the traditional regression models. These models are particularly useful for time-series data like cryptocurrency prices.

Real-time prediction: You can develop a real-time prediction system that can predict Bitcoin prices based on new data. You can use streaming data sources such as Kafka and Spark Streaming to continuously update the model and make real-time predictions.

Interpretability: You can explore techniques for interpreting the model's predictions and gaining insights into the factors that are driving the predictions. This can be particularly important for making informed decisions based on the model's predictions.

Other cryptocurrencies: You can extend your analysis to other cryptocurrencies such as Ethereum, Litecoin, or Ripple. This will require collecting data on these cryptocurrencies and applying similar analysis techniques.

7.2 Conclusion:

.

In this project, we explored the Bitcoin cryptocurrency dataset and applied machine learning algorithms to predict its price. We started with exploratory data analysis using tools such as Tableau to visualize the data and gain insights into its patterns and trends. We then preprocessed the data using techniques such as feature engineering, normalization, and vectorization.

We trained various regression models such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor and K-Means Clustering on the preprocessed data to predict the future price of Bitcoin. We evaluated the models using various performance metrics such as RMSE, MAE. Based on our experiments, we found that the Random Forest model outperformed the other models in terms of prediction accuracy.

Overall, this project provided us with a hands-on experience in working with cryptocurrency datasets, exploring and analyzing the data, and applying machine learning algorithms to predict their future prices.