A Project Based Seminar Report

on

"Prediction Of Cryptocurrency Prices Using Machine Learning Based On Blockchain Information."

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

Himanshu Agrawal 407001

Sacchi Agrawal 407002

Ankur Sakhala 407007

Amita Bodhe 407023

Under the guidance of

Prof. S. M. Jaybhaye

Sinhgad Institutes

Department of Information Technology

STES's Sinhgad College of Engineering

Vadgaon (Bk.), Off. Sinhgad Road,

Pune 411041.

Semester-VII, Final Year Engineering

2018-2019

Sinhgad Institutes

CERTIFICATE

This is to certify that the project based seminar report entitled "Machine Learning Algorithms for Cryptocurrency Price Prediction" being submitted by Himanshu Agrawal (407001) Ankur Sakhala (407007) Sacchi Agrawal (407002) Amita Bodhe (407023) is a record of bonafide work carried out by them under the supervision and guidance of Prof. S. M. Jaybhaye in partial fulfillment of the requirement for **BE (Information Technology Engineering) - 2015 Course** of Savitribai Phule Pune University, Pune in the academic year 2018-2019.

Date: 12/10/2018

Place: Pune

Prof. S. M. Jaybhaye
Seminar Guide

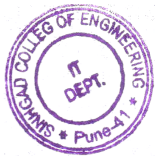
Prof. G. R. Pathak
Head of the Department

Dr. S. D. Lokhande
Principal

This Project Based Seminar report has been examined by us as per the Savitribai Phule Pune University, Pune requirements at STES's Sinhgad College of Engineering, Pune-411041 on . . . . . .

Internal Examiner

External Examiner



ACKNOWLEDGMENT

We are highly indebted to my guide Prof S. M. Jaybhaye for her guidance and constant supervision as well as for providing necessary information regarding the seminar report and also for her support in completing the seminar report. We would like to express my special gratitude and thanks to Staff Members of department of Information Technology for giving us such attention and time.

This acknowledgment would be incomplete without expressing our thanks to Prof. G. R. Pathak, Head of the Department (Information Technology) for his support during the work.

We would like to extend our heartfelt gratitude to our Principal, Dr. S. D. Lokhande who provided a lot of valuable support, mostly being behind the veils of college bureaucracy.

We would also like to express our gratitude towards our parents and friends for their kind co-operation and encouragement which helped us in completion of this report.

Himanshu Agrawal
Ankur Sakhala
Sacchi Agrawal
Amita Bodhe

ABSTRACT

In recent years,Bitcoin ecosystem has gained the attention of consumers, businesses, investors and speculators alike.As a result of blockchain-network-based feature engineering ,macro-economic factors of actual market and machine learning algorithms optimization, we can obtain up-down Bitcoin price movement classification accuracy of roughly 55 percent. This research is concerned with predicting the price of Bitcoin using machine learning. The goal is to ascertain with what accuracy can the direction of Bit- coin price in USD can be predicted. The price data is sourced from the Bitcoin Price Index . The task is achieved with varying degrees of success through the implementation of a Bayesian optimised recurrent neural network (RNN) and Long Short Term Memory (LSTM) network.

CONTENTS

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION TO PROJECT TOPIC

1.1 Introduction to Project

Prediction of mature financial markets such as the stock market has been researched at length. Bitcoin presents an interesting parallel to this as it is a time series prediction problem in a market still in its transient stage. As a result, there is high volatility in the market and this provides an opportunity in terms of prediction. In addition, Bitcoin is the leading cryptocurrency in the world with adoption growing consistently over time. Due to the open nature of Bitcoin it also poses another paradigm as opposed to traditional financial markets. It operates on a decentralised, peer-to-peer and trustless system in which all transactions are posted to an open ledger called the Blockchain. This type of transparency is unheard of in other financial markets. Given the complexity of the task, deep learning makes for an interesting technological solution based on its performance in similar areas. Tasks such as natural language processing which are also sequential in nature and have shown promising results. This type of task uses data of a sequential nature and as a result is similar to a price prediction task. The recurrent neural network (RNN) and the long short term memory (LSTM) flavour of artificial neural networks are favoured over the traditional multilayer perceptron (MLP) due to the temporal nature of the more advanced algorithms.

Various studies on statistical or economical properties and characterizations of Bitcoin prices refer to its capabilities as a financial asset, the relationship between Bitcoin and search information, such as Google Trends and Wikipedia, and wavelet analysis of Bitcoin and BGT-Blockchain Google trends.We here will be bridging the gap between these two by Machine Learning.

1.1.1 Machine Learning

Data mining can be defined as the extraction of implicit, previously unknown and potentially useful information from data. Machine learning provides the technical basis for data mining.

The purpose of this research is to predict the direction of the price of Bitcoin. As this is a task with a known target it is a supervised machine learning task although some pre-processing can take advantage of unsupervised learning methods. The super- vised algorithms explored include Wavelets and the wavelet discrete transform, several type of artificial neural networks including the Multi-Layer perceptron (MLP), Elman Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM). In terms of pre-processing, random forests were used for feature selection while Bayesian optimisation was performed to optimize some the parameters of the LSTM.

1.1.2 Bitcoin

Bitcoin is a digital cryptocurrency and payment system that is entirely decentralized, meaning it is based on peertopeer transactions with no bureaucratic oversight. Transactions and liquidity within the network are instead based on cryptography. The system first emerged formally in 2009 and is currently a thriving open-source community and payment network. Based on the uniqueness of Bitcoin payment protocol and its growing adoption, the Bitcoin ecosystem is gaining lots of attention from businesses, consumers, and investors alike.

1.1.3 Blockchain

Decentralization can be specified by the following goals: (i) Who will maintain and manage the transaction ledger? (ii) Who will have the right to validate transactions? (iii) Who will create new Bitcoins? The blockchain is the only available technology that can simultaneously achieve these three goals. Generation of blocks in the Blockchain, which is directly involved in the creation and trading of Bitcoins, directly influence the supply and demand of Bitcoins. Combination of Blockchain technologies and the Bitcoin market is a real-world example of a combination of high-level cryptography and market economies.

1.2 Motivation behind project topic

If you were to pick the three most ridiculous fads of 2017, they would definitely be fidget spinners , artificial intelligence and, yes, cryptocurrencies. I am actually not a hodler of any cryptos. So, while I may not have a ticket to the moon, I can at least get on board the hype train by successfully predicting the price of cryptos by harnessing deep learning, machine learning and artificial intelligence (yes, all of them!). Machine Learning - A computer is said to learn if its performance P on a class of tasks T improves with experience E.Having n number of applications like Image recognition,Sentiment analysis,Natural Language processing,Captcha,Pred analysis and even driverless cars,this field/domain totally amuses one to try something in it. Availability of built in libraries like WEKA to code in JAVA,SCIKITS LEARN in Python,variety of datasets available on kaggle,user friendly editors like Jupyter Notebook and IDE?s like Spyder/Pycharm it become very easy for one to implement his/her ideas.

Bitcoin as a currency is in a transient stage and as a result is considerably more volatile than other currencies such as the USD. Interestingly, it is the top perform- ing currency four out of the last five years. Thus, its prediction offers great potential and this provides motivation for research in the area.

1.3 Aim and Objective(s) of the work

The aim of this project is find out with what accuracy the direction of the price of Bitcoin can be predicted using machine learning methods.

1. Proper data collection and finding out the perfect number and types of variables that can be the best suit.

2. To chose the most appropriate algorithm and ways to validate it,optimize it,evaluate it on various parameters for accuracy and errors.

3. Technology used also played an important role in deciding the success of a project.To select and have hands on these required stuff.

4. Understanding the algorithm used at its full depth so that we can optimize it,generalize it in future.

The purpose of this study is to achieve a considerable amount of accuracy and use all the proposed research to find out the best algorithm for cryptocurrency price prediction.

Chapter 2

LITERATURE SURVEY

HUISU JANG,JAEWOOK LEE,"An Empirical Study on Modeling and Prediction of Bitcoin Prices With Bayesian Neural Networks Based on Blockchain Information"Bitcoin has recently attracted considerable attention in the fields of economics, cryptography, and computer science due to its inherent nature of combining encryption technology and monetary units. This paper reveals the effect of Bayesian neural networks (BNNs) by analyzing the time series of Bitcoin process. We also select the most relevant features from Blockchain information that is deeply involved in Bitcoin?s supply and demand and use them to train models to improve the predictive performance of the latest Bitcoin pricing process. We conduct the empirical study that compares the Bayesian neural network with other linear and non-linear benchmark models on modeling and predicting the Bitcoin process. Our empirical studies show that BNN performs well in predicting Bitcoin price time series and explaining the high volatility of the recent Bitcoin price.

Satoshi Nakamoto,"Bitcoin: A Peer-to-Peer Electronic Cash System"A purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. Digital signatures provide part of the solution, but the main benefits are lost if a trusted third party is still required to prevent double-spending. We propose a solution to the double-spending problem using a peer-to-peer network.The network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work. The longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the largest pool of CPU power. As long as a

majority of CPU power is controlled by nodes that are not cooperating to attack the network, they'll generate the longest chain and outpace attackers. The network itself requires minimal structure. Messages are broadcast on a best effort basis, and nodes can leave and rejoin the network at will, accepting the longest proof-of-work chain as proof of what happened while they were gone.

Alex Greaves, Benjamin Au,"Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin"Bitcoin is the world?s leading cryptocurrency, allowing users to make transactions securely and anonymously over the Internet. In recent years, The Bitcoin the ecosystem has gained the attention of consumers, businesses, investors and speculators alike. While there has been significant research done to analyze the network topology of the Bitcoin network, limited research has been performed to analyze the network?s influence on overall Bitcoin price. In this paper, we investigate the predictive power of blockchain network-based features on the future price of Bitcoin. As a result of blockchain-network based feature engineering and machine learning optimization, we obtain up-down Bitcoin price movement classification accuracy of roughly 55%.

EVITA STENQVIST,JACOB LOONO ,"Predicting Bitcoin price fluctuation with Twitter sentiment analysis"The decentralized cryptocurrency, arguably, by design, a pure free market commodity ? and as such, public perception bears the weight in Bitcoins monetary valuation.This thesis looks toward these public perceptions, by analyzing 2.27 million Bitcoin-related tweets for sentiment fluctuations that could indicate a price change in the near future. This is done by a naive method of solely attributing rise or fall based on the severity of aggregated Twitter sentiment change over periods ranging between 5 minutes and 4 hours, and then shifting these predictions forward in time 1,2, 3 or 4 time periods to indicate the corresponding BTC interval time.The prediction model evaluation showed that aggregating tweet sentiments over a 30 min period with 4 shifts forward, and a sentiment change threshold of 2.2%, yielded a 79% accuracy.

Chapter 3

METHODOLOGIES

3.1 Data Collection

We use the Bitcoin transaction data available on the CS224W website, which contains every Bitcoin transaction made prior to April 7, 2013. All transactions are available on a public ledger, and this data set is a large text file containing a line for every transaction. Each line includes the transaction id, sender, recipient, value (in BTC),and a timestamp. Transactions involving multiple senders and multiple receivers are represented by multiple lines with the same transaction id. This file comes with a matching file containing one line for each group of addresses that appear together in some transaction, hence belonging to the same user or entity. We used this file to transform our data set into a simplified one indexed by user entity rather than address using the Union Find algorithm, based on the intuition that for any given transaction, only one entity could be the sender of that transaction, even if multiple sending accounts were used. The original data set contains nearly 37 million transactions between roughly 6 million addresses. Once reduced, we obtained the same number of transactions between just over 3 million unique users. We represent the data in a directed graph, where each node is a user and each edge is a transaction (from sender to receiver). Bitcoin mining is represented in the data as a transaction from one user to itself, and is hence manifested in the graph as a self-loop. This representation of the data allows us to explore various properties of the graphs and also use them as features in price prediction. In addition, to perform price prediction we needed a complete historical listing of prices for Bitcoin. We acquired from api.bitcoincharts.com the entire histories of several Bitcoin exchanges, where each line in a given history contains the timestamp and the exchange rate in USD. From these transactions we computed the average price of Bit-

coin across all transactions as the official Bitcoin price at 15 second intervals dating back to just after the first Bitcoin was mined.

3.2 Feature Extraction

We compiled several network-based features to develop our supervised machine learning algorithms. The feature families we selected using the following approach: Features were developed for time-frames of one-hour, one-day, one-week and one-month prior to prediction time. The goal is to predict the price of Bitcoin in USD one hour in advance. All features are computed using only information available an hour prior to the target prediction time, and for node features we used both in and out degree. The following features were extracted:

1. current Bitcoin price
2. net flow per hour
3. number of transactions per hour
4. mean transaction value
5. median transaction value
6. median and average node in- and out-degree
7. alpha constant of power law
8. total number of Bitcoin mined
9. number of new addresses
10. mean initial deposit amount among new addresses
11. number of transactions performed by new addresses

Many of these features are somewhat correlated with the price of Bitcoin, though of those most are only loosely related. For instance, we see the total number of mined Bitcoin plotted against the current price. In addition, we identified three addresses that had by the far the greatest influence on the network, with over 10% of all Bitcoin ever sent passing through at least one of these addresses. Creatively, we labeled these as A,B, and C, and of these the largest transactor (by over a factor of 2) was A. We believe this address to be associated with Mt.Gox, which was the worldś largest Bitcoin exchange prior to their 2014 collapse. From each of these nodes, for the hour prior to prediction, we collected the following features:

1. total Bitcoin passing through
2. net Bitcoin flow (received minus sent)
3. number of transactions
4. closeness centrality

In doing so, we hoped not only to extract important information for our classifiers and regressors, but also to gain a general idea of how the largest

players in the Bitcoin market affect its valuation. We can gain some idea of the activity of the differing addresses by seeing their net balance, the total Bitcoin coming in minus that going out summed over a year long period. We collected data every hour from February 1, 2012 to February 1, 2013 to form our training set, and from February 1, 2013 to April 1, 2013 to form our test set. It was necessary to use data temporally after the training set data for testing in order to correctly simulate the real world scenario of predicting price based on past data.

3.3 Feature Selection

Choosing which features to use is an important aspect of any regression or classification optimization. To prune features, we calculated the mutual information between each feature and the output, pruning all insignificant features from the model. The figure below gives the equation for I, the mutual information, where A is our feature data and B our output data.

$$I(A;B) = \sum_{b \in B} \sum_{a \in A} p(a,b) * log\left(\frac{p(a,b)}{p(a)p(b)}\right)$$

s

3.3.1 Barro's model and BTC economics

Barro's model provides a simple Bitcoin pricing model under perfect market conditions. In this model, Bitcoin is assumed to possess currency value and is exchangeable with traditional currencies, which are under central bank control and can be used for purchasing goods and services. The total Bitcoin supply, SB, is represented by

$$S_B = P_B B$$

where PB denotes the exchange rate between Bitcoin and dollar (i.e. dollar per unit of Bitcoin), and B is the total capacity of Bitcoins in circulation..

The total Bitcoin demand depends on the general price level of goods or services, P; the economy size of Bitcoin, E; and the velocity of Bitcoin, V, which is the frequency at which a unit of Bitcoin is used for purchasing goods or services. The total demand of Bitcoin, DB, is described as followed by:

$$D_B = \frac{PE}{V}$$

The market equilibrium with the perfect market assumption is acquired when the supply and the demand of Bitcoin is the same amount. The equilibrium is therefore achieved at

$$P_B = \frac{PE}{VB}$$

This equilibrium equation implies that in the perfect market, the Bitcoin price in dollars is affected proportionally by the general price level of goods or services multiplied by the economy size of Bitcoin,which can be determined indirectly from the global macroeconomic indexes in actual markets and inversely by the velocity of Bitcoin multiplied by the capacity of the Bitcoin market,extracted from the Blockchain platform.

3.3.2 Types of Variables

There are mostly three types of variables namely,Response variables which includes log price of BTC,Blockchain information like block size,transactions per block,confirmation time,hash/difficulty rate,macro-ecnomic factors consisting of SandP500, DOW30,NASDAQ,CrudeOil,Gold,Nikkei225,VIX and Global currnecy ratio i.e IPY,CHE,CNY,EUR.

3.3.3 Dependent and Independent Variables

The independent variable for this study is the closing price of Bitcoin in US Dollars taken from the Coindesk Bitcoin Price Index. Rather than focusing on one specific exchange this price index takes the average prices from five major Bitcoin exchanges; Bitstamp, Bitfinex, Coinbase, OkCoin and itBit. The closing price is chosen over a three-class dummy classification variable representing price going up, down or staying the same for the following reason; the use of a regression model over a classification model offers further model comparison potential through the capture of the root mean squared error (RMSE) of the models. Classifications are then made based on the prediction of the regression model e.g. price up, price down or no change. Additional performance metrics include accuracy, specificity, sensitivity and precision.

The dependent variables are taken from the Coindesk website, Blockchain.info and from the process of feature engineering.

As a general rule, the greater the mutual information, the more informative the feature. Many of the features we computed proved uninformative, and had mutual information scores barely above random chance. The features that proved the most informative and whose inclusion improved our results were:

1. Current Bitcoin price
2. Net Bitcoin flow for A, B, and C
3. Closeness centrality for A, B, and C
4. Mined bitcoin in the last hour
5. Number of transactions among new addresses in the last hour
6. Mean node degree in the last hour
7. Net flow in the last hour

Interestingly, none of the features that looked farther back than the last hour were at all informative. Unsurprisingly, by far the most informative feature for price prediction was current price.

3.4 Network Algorithms

Union Find The Union Find algorithm is used to determine contiguous subsets within a network. This algorithm is applied as a preprocessing technique to relate multiple accounts related to a single owner entity. The algorithm is broken down recursively calling the Union and Find functions.

1. Union function: merges two disjoint subsets into the union of those subsets.

2. Find function: checks to see which other entities are part of the same subset as any of those already merged.

3.5 Learning Algorithms

Linear Regression : Linear Regression (LR) is a predictive model that formulates a line of best fit between a scalar dependent variables and multiple explanatory variables. Linear fit occurs by minimizing the mean squared error between the predicted and actual output. Below we detail the hypothesis, parametrization (which can be extended in vectorized form), cost function, and goal of linear regression.

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\, J(\theta_0, \theta_1)$

Logistic Regression Logistic Regression is a predictive regression model in which the dependent variable is categorical. In the simple case (as in our problem formulation)

where there are only two catogories, Logis- tic Regression uses Maximum Likelihood Es- timation to formulate the probabilities in which Logistic Regression will take on a particular class, with an iterative algorithm such as Newton?s method used to obtain the fitted model. Logistic Regression is typically seen as a robust baseline method for classification. Below is the objective and MLE formulation for Logistic Regression, as well as the formulation of the logistic function.

$$\max_{\theta} \sum_{i=1}^{n} \log p(y_i|x_i, \theta).$$

$$p(y = 1|x) = \sigma(\theta^\top x) = \frac{1}{1 + \exp(-\theta^\top x)}.$$

Support Vector Machine Support Vec- tor Machine (SVM) is a discriminative clas- sifier that generates a separating hyper- plane. Error tolerance budget is included to make separating hyperplane robust in case of inseparable class data. Linear decision boundaries are augmented to more complex boundary shape through kernel implementation (e.g. polynomial, Gaussian and radial kernel). SVM obtains decision boundary by creating a margin which maximizes the functional and geometric margins between classes. SVMs have received attention for the impressive performance in classification. SVM can also be augmented for regression problems as Support Vector Regression (SVR) optimizing response variable distance from the decision boundary. Below we provide formulation of the optimization and constraint model for SVM.

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Resampling by Bootstrap and Cross Validation : Resampling refers to a method used for model validation.We have two-cross-validation, and bootstrap.We identify advantages and disadvantages of each method and select the appropriate method for the empirical analysis of this study.

A bootstrap method is one of the sampling techniques that new data set is sampled from the original data set with the replacement. A typical bootstrap works as follows-

1 . We have the original data setD with the number of N.

2 . Below following step is repeated B times for particular large number to produce B different bootstrap data set, Z1; Z2; . . . ; ZB here, Data set Zi with the size N is generated by sampling from the original data set D with the replacement.

3 . The machine is trained from each bootstrap data set.

4 . Accuracy of the machine is calculated by averaging each bootstrap data set.

$$Accuracy = \frac{1}{B}\sum_{j=1}^{B}\frac{1}{N}\sum_{i=1}^{N}(1 - Loss(\hat{y}_i^j, y_i))$$

where yi is an i-th true training output data, yij i is an i-th estimated output from the bootstrap data Zj, and Loss() is a loss function.

A cross-validation randomly divides the original data set into K equal-sized parts without the replacement. We fit the machine learning model to the K - 1 parts leaving out particular set k and acquire a prediction error for the left-out k part. Total prediction accuracy is combined after the procedure is repeated for each part to leave. A general procedure is as follows:

1 . We divide the original data set into K partial equal-sized data.

2 . We can compute the total accuracy:

$$accuracy_K = \sum_{k=1}^{K}\frac{n_k}{N}\frac{1}{n_k}\sum_{i=1}^{n_k}(1 - Loss(\hat{y}_i^k, y_i))$$

where N is the total number of the original data set, others have same definition with in the bootstrap description.

3 . The estimated standard deviation of the cross-validation:

$$\hat{SE}(CV_K) = \sqrt{\frac{\sum_{k=1}^{K}(Err_k - \bar{Err_k})^2}{N-1}}$$

Err_k is the k-th loss, $\sum_{i=1}^{n_k} Loss(\hat{y}_i^k, y_i)$.

Bootstrap is adequate to validate a predictive model performance, to use an ensemble method, and to estimate of bias and variance of the trained model. Bootstrap creating the cloned multiple samples with the replacement is not originally developed for model validation. It can give more biased results. Therefore, we employ the cross-validation technique to our model validation. Cross-validation can create high-variance problems when data size is small. Our data size is sufficient to overcome the problem. We employ the 10-fold cross-validation methods generally used for model validations.

Statistical Analysis of Algorithms :

Table 3.1: Statistical Analysis of Algorithms

Response variable	**Log Price-RMSE**	**Log Volatility-RMSE**
Linear regression	0.0935	0.4823
Bayesian NN	0.0039	0.2325
Support Vector regression	0.3201	0.5297

The Root Mean Square Error of three different algoithms can be analysed from the above table.Bayesian NN having the least RMSE for log price and log volatility,it can be chosen as the best amongst.

Chapter 4

SOFTWARE REQUIREMENT SPECIFICATION

4.1 FUNCTIONAL REQUIREMENTS

This system performs real-time extraction and analysis of Blockchain and sentiments data for accurate prediction of Cryptocurrency prices.

4.1.1 System Feature I:

1. Fetch and Display Holdings.

UseCaseName	Fetch and Display Holdings.
Priority	Essential.
Precondition	Manually add transactions or Binance Log in
Basic Path	1.Fetch data using Manual process .
Alternative Path	Binance Connect
Post condition	Holdings are displayed in form of graphs.
Other	The extracted data has blockchahin info,tweets name as well as related effect.

Table 4.1: Fetch and Display Holdings.

2. Getting Twitter data/CSV file.

UseCaseName	Getting twitter data/CSV file
Priority	Essential.
Precondition	CSV file and Twitter API connection.
Basic Path	1.Applying Cleaning, Text Similarity and Stemming algorithm.
Alternative Path	None
Post condition	Dataset containing sentiments words and related date and time based on blockchain info and bitcoin trends.
Other	None.

Table 4.2: Getting Twitter data/CSV file.

3. Predict and view results.

UseCaseName	Predict and view results.
Priority	Essential.
Precondition	Pre-processed dataset.
Basic Path	Sentiments related classification using machine learning or classification algorithm and Time series analysis of blockchain info.
Alternative Path	None
Post condition	Classify data into positive, negative and neutral sentiment successfully.
Other	None.

Table 4.3: Predict and View Results

4. Finding percentage of Matched Patterns.

UseCaseName	Finding percentage of Matched Patterns.
Priority	Essential.
Precondition	The extracted tweets are classified into type of sentiments.
Basic Path	1.Show type of representation graph from user
Alternative Path	None.
Post condition	User get graphical and statistical results for selected time intervals and graphs.
Other	None.

Table 4.4: Finding percentage of Matched Patterns.

5. Train-Test-Validation

UseCaseName	Train-Test-Validation
Priority	Essential.
Precondition	Proper CSV file which is cleaned and with all KPI's extracted.
Basic Path	1.70-30 percent split of data set is done. 2. Validation and verification of data is done for greater accuracy.
Alternative Path	None
Post condition	retrieval of prediction results.
Other	None.

Table 4.5: Test-Train-Validation

6. Login

UseCaseName	User Login
Priority	Essential.
Precondition	Valid user records in database
Basic Path	1.Update user details by admin. 2.Delete user.
Alternative Path	None
Post condition	Update to the database.
Other	None.

Table 4.6: User Login

7. Time Series Prediction

UseCaseName	Time Series Prediction
Priority	Essential.
Precondition	amp; CSV file and twitter data with all extracted KPI's and keywords fields.
Basic Path	amp; 1.Data set validation.
Alternative Path	amp; None.
Post condition	amp; Prediction results will be generated.
Other	amp; None.

Table 4.7: Time Series Analysis.

4.1.2 System Feature II:

Generates real time results by continuously fetching real time data. System can analyse large amount of data effectively and efficiently. Performs a topic related task on user‘s behalf.

4.2 EXTERNAL INTERFACE REQUIREMENTS

4.2.1 User Interfaces :

The interface between user and the system include many provisions from where they can access the whole system.

4.2.2 Software Interfaces :

The system can use Windows as the operating system platform. To run this application we need Python V 3.0 and below HTML V 5 and Apache handler as server api on hosted server. To store data we need Mysql database. At client side this system can be accessed through web application.

4.2.3 Hardware Interfaces:

The entire software requires a completely equipped computer system including monitor, keyboard, and other input output devices. The data is manually entered to the system and the outputs are produced using various algorithms. It also requires a server where this application can be hosted.

4.2.4 Communication Interfaces

The system makes use of internet services and hence any web browser and a proper networking should be enabled within the entire premise.

4.3 NON FUNCTIONAL REQUIREMENTS:

4.3.1 Performance Requirements :

System must be capable of storing and processing the streams received or generated without dropping any data. It Should be able to delete, modify and update the models created at any time.

4.3.2 Operating System :

Software must be run on all operating systems like Windows 7 or later, MacOS 10.13 or later.

4.3.3 Response Time :

MySQL gives nearly real time data (if internet connection is consistent) and when integrated with Twitter API we get real time output.

4.3.4 Safety/Security Requirements :

Heterogeneous data sources.

Data extracted from different proprietary sources. Integrating them so they can be used together in an integrated way during the analysis process.

4.3.5 Software Quality Attributes

Adaptability:

The proposed system is adaptable to different sources of social media data.

Reliability:

As the system uses Hadoop Distributed File System, it makes a default three replication of input data. Hence the system is reliable in case of any data loss issues.

Scalability:

New machines can be easily added or removed as and when required.

Robustness:

System can be used for both structured as well as unstructured data.

4.4 SYSTEM REQUIREMENTS:

4.4.1 Database Requirements:

- MySQL

4.4.2 Software Requirements

- Client Side System requirements (Software) :

  For Desktop Platforms :

  1. Windows 7 or later, MacOS 10.13 or later
  2. Browser Requirements : Chrome / Firefox (access through website)

- Server Side System Requirements (Software) :

  1. Python 3.0
  2. Anaconda Navigator
  3. MySQL Database
  4. Apache web server or equivalent
  5. Server Side Scripting Support : HTML 5 or above

4.4.3 Hardware Requirements :

Sr. No.	Parameter	Minimum Requirement	Justification
1	Processor	64 bit (Intel/ AMD)	Recommended for best performance
2	RAM	4 GB	Min. Requirement of physical memory
3	Hard Disk	2GB	Min. Memory required to be free

Chapter 5

DESIGN AND MODELLING

5.1 Architecture Diagram

Prediction model for time series analysis usually consists of two phases namely **1) Model Creation** and **2) Using the Model with Streaming data**.

There is a slight line between overfitting and generalization which becomes a matter of concern.The train set is fed to ML algo and finally a model is built which can then give accurate results for the test set.

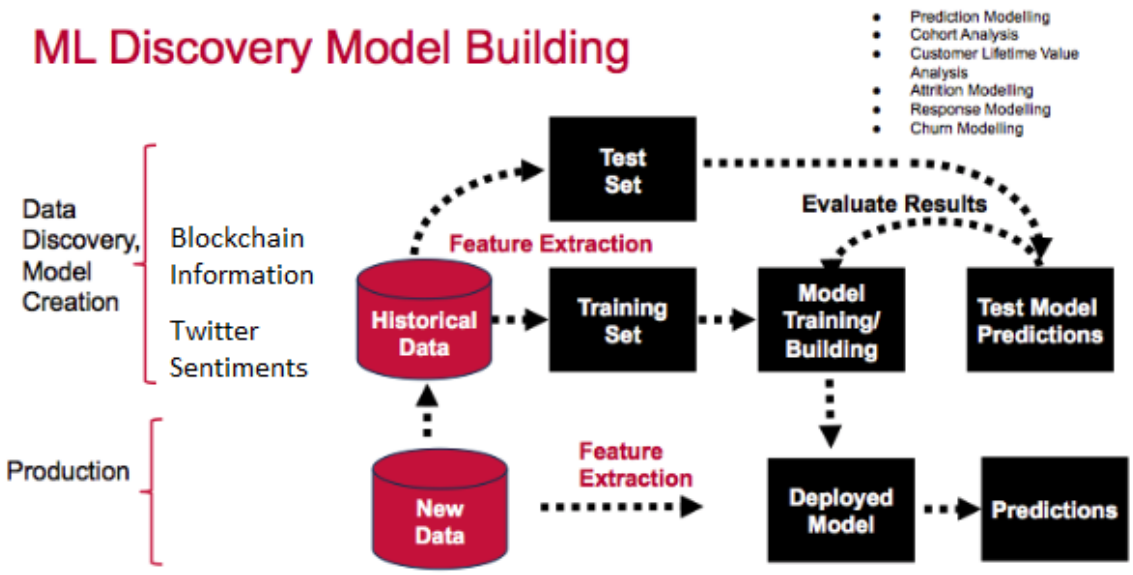


Figure 5.1: Architecture Diagram

5.2 Data Flow Diagram

5.2.1 Data Flow Level 0

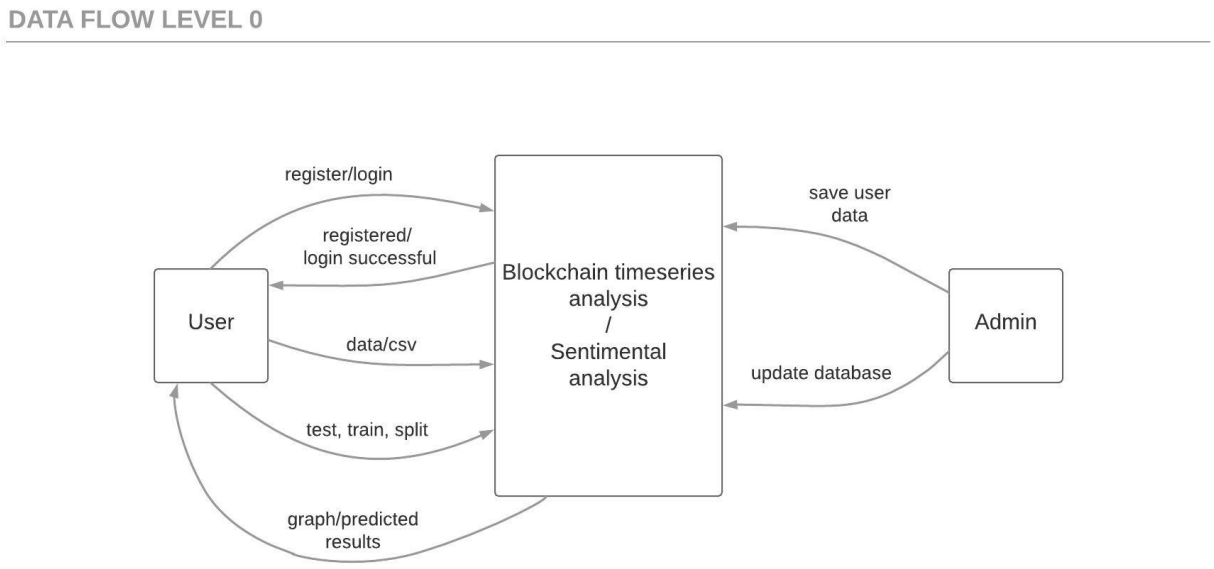


Figure 5.2: Data Flow Diagram Level 0

A data flow diagram (DFD) illustrates how data is processed by a system in terms of inputs and outputs. As its name indicates its focus is on the flow of information, where data comes from, where it goes and how it gets stored.

5.2.2 Data Flow Level 1

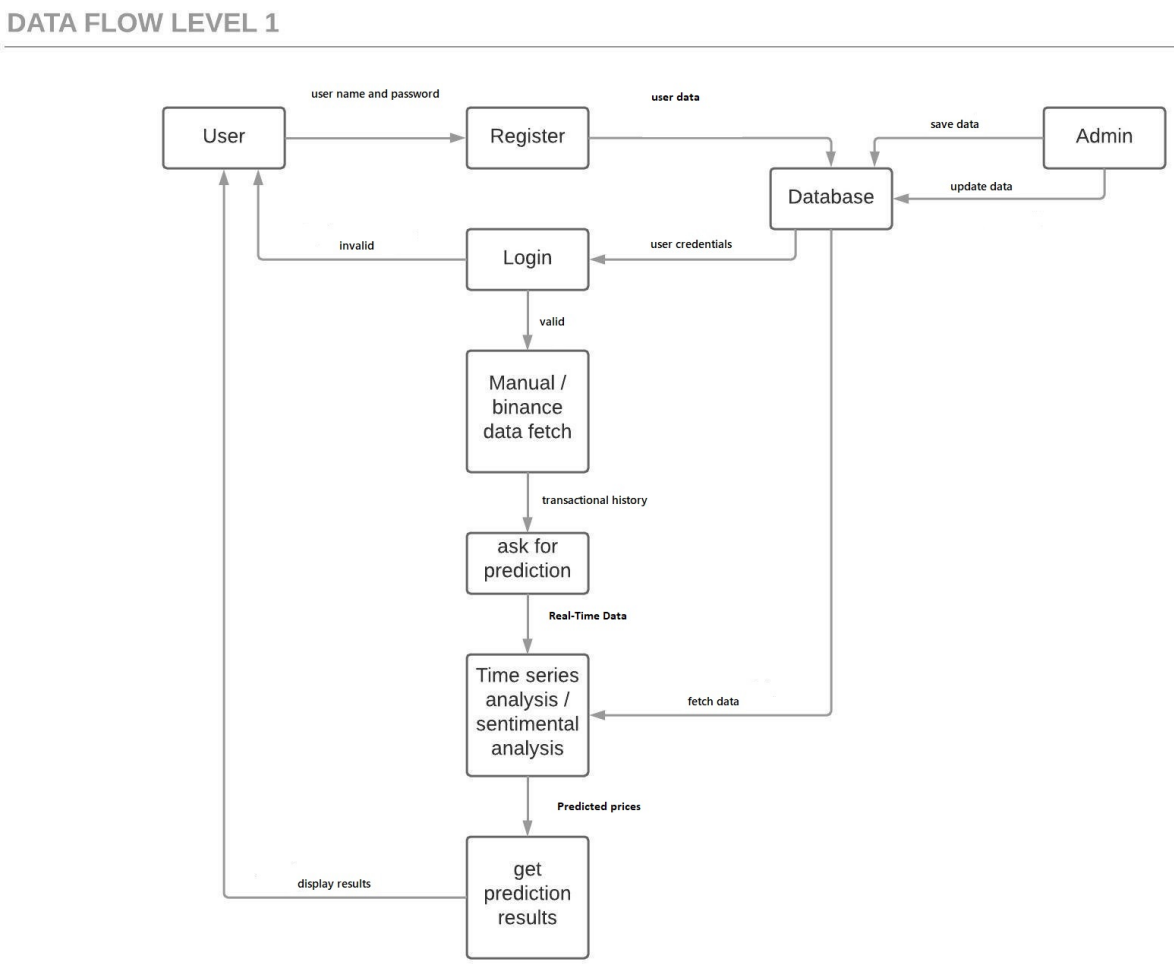


Figure 5.3: Data Flow Diagram Level 1

A level 1 data flow diagram (DFD) is more detailed than a level 0 DFD but not as detailed as a level 2 DFD. It breaks down the main processes into subprocesses that can then be analyzed and improved on a more intimate level.

5.3 Entity Relationship Diagram

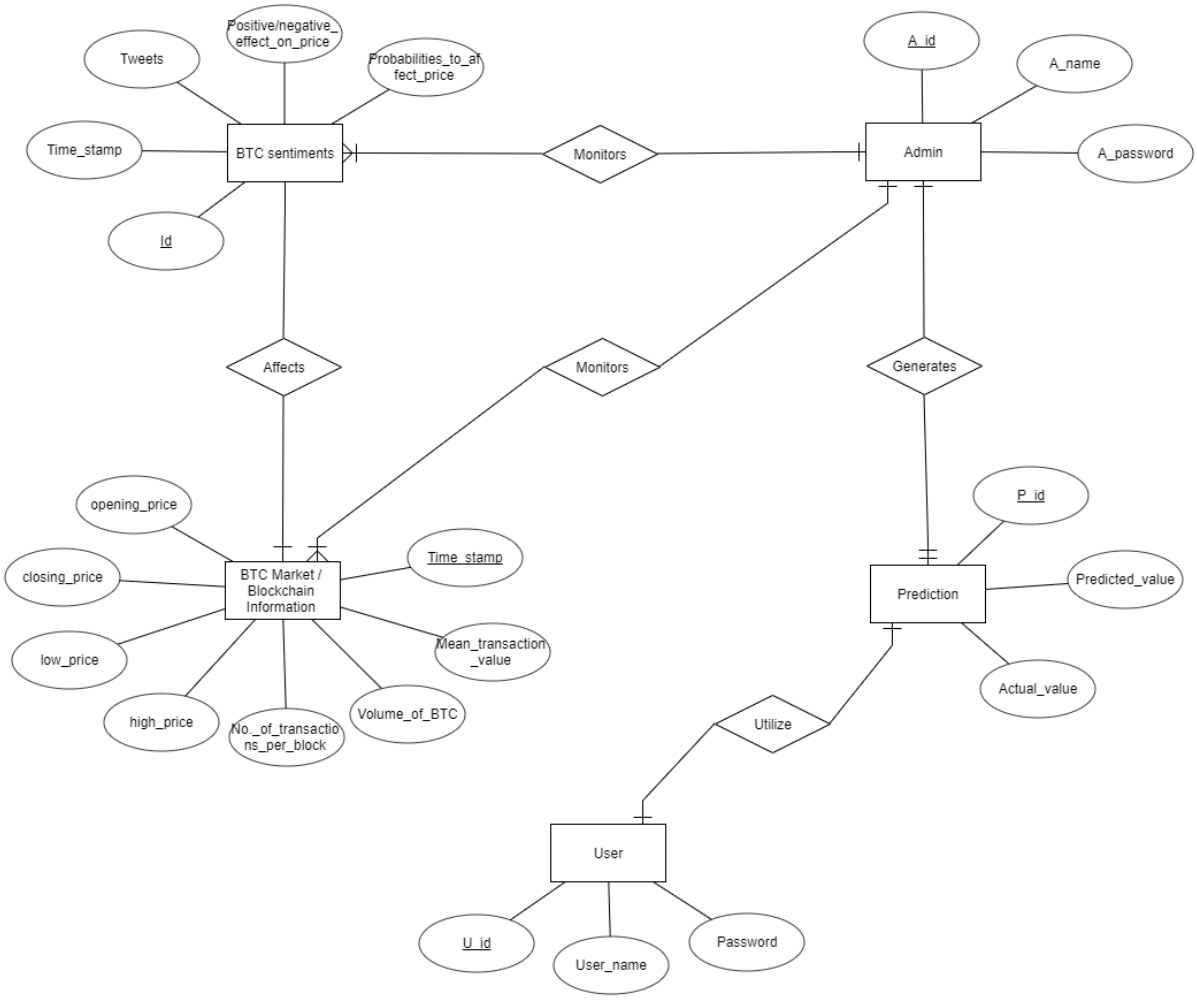


Figure 5.4: Entity Relationship Diagram

An entity relationship model describes interrelated things of interest in a specific domain of knowledge.Database Schema Of the System can be analysed in a better way as of which attribute is related and how.Here we have 5 major Entities with various attributes and relationships between them.

5.4 Use Case Diagram

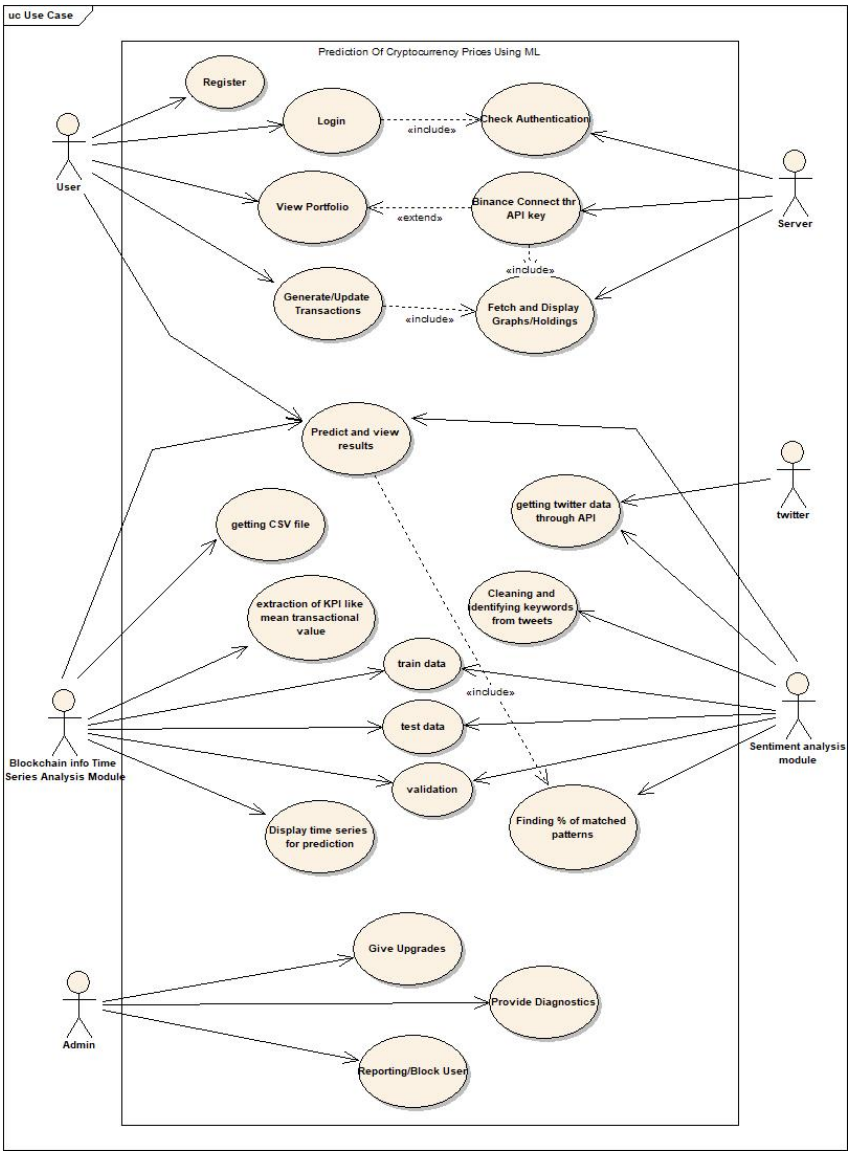


Figure 5.5: Use case diagram for proposed system

The Diagram shows the actors and respective Use Cases of the system.Relationships between them becomes feature of the system.

5.5 Class Diagram

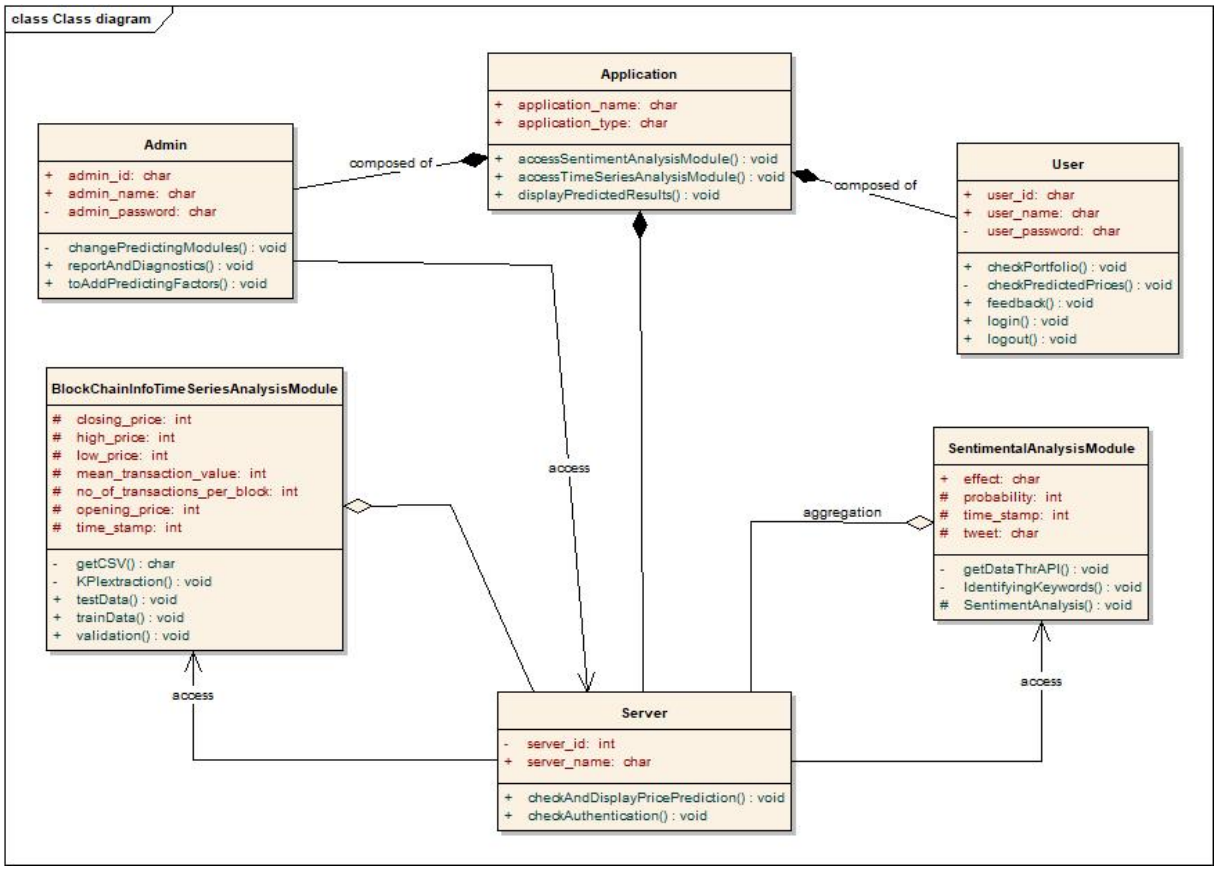


Figure 5.6: Class diagram for proposed system

The system consists of 6 classes: User, Admin, Application,Server and Modules.Relationships between them attributes and Operations are shown.Application is composed of Admin,Server and User.Server is directly associated with/is aggregation of two prediction modules.

5.6 Activity Diagram

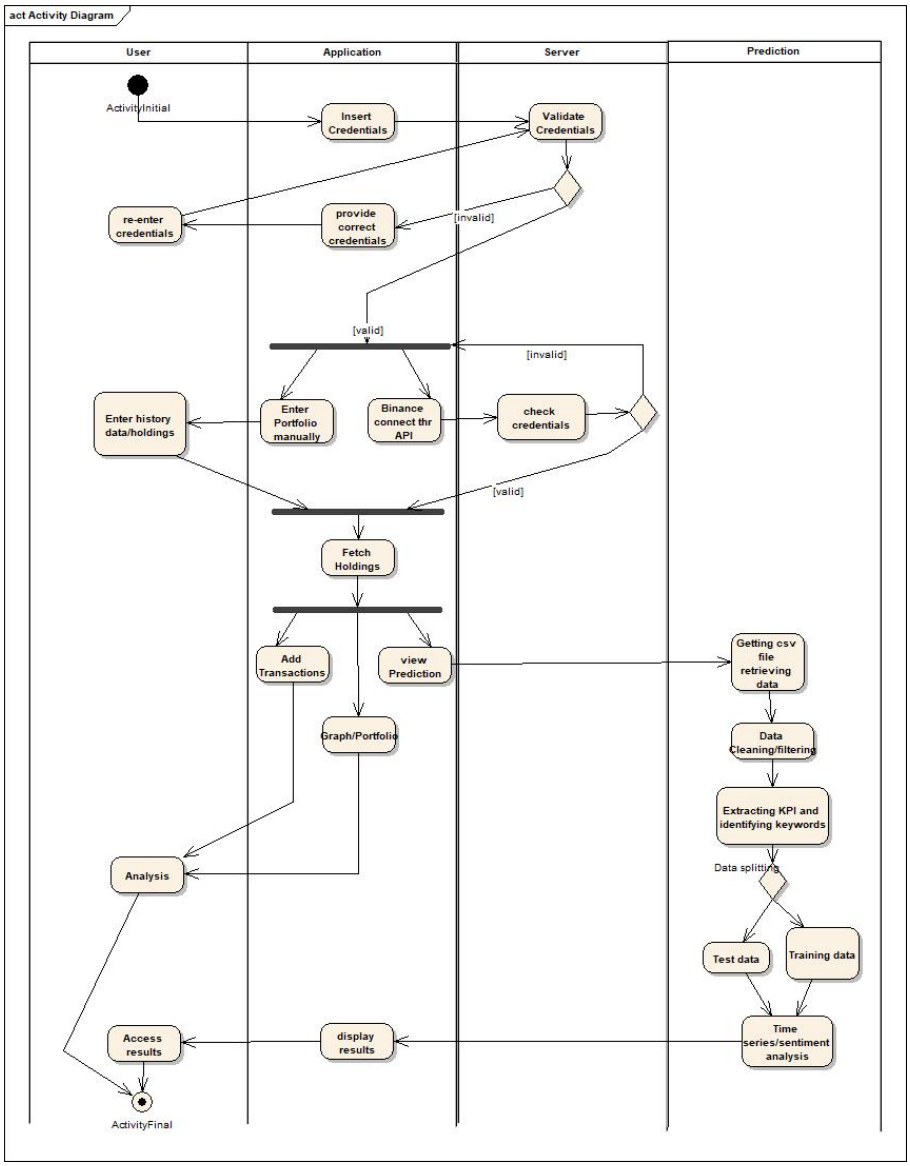


Figure 5.7: Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.Swim lanes shows involvement of various actors in different actions.

5.7 Sequence Diagram

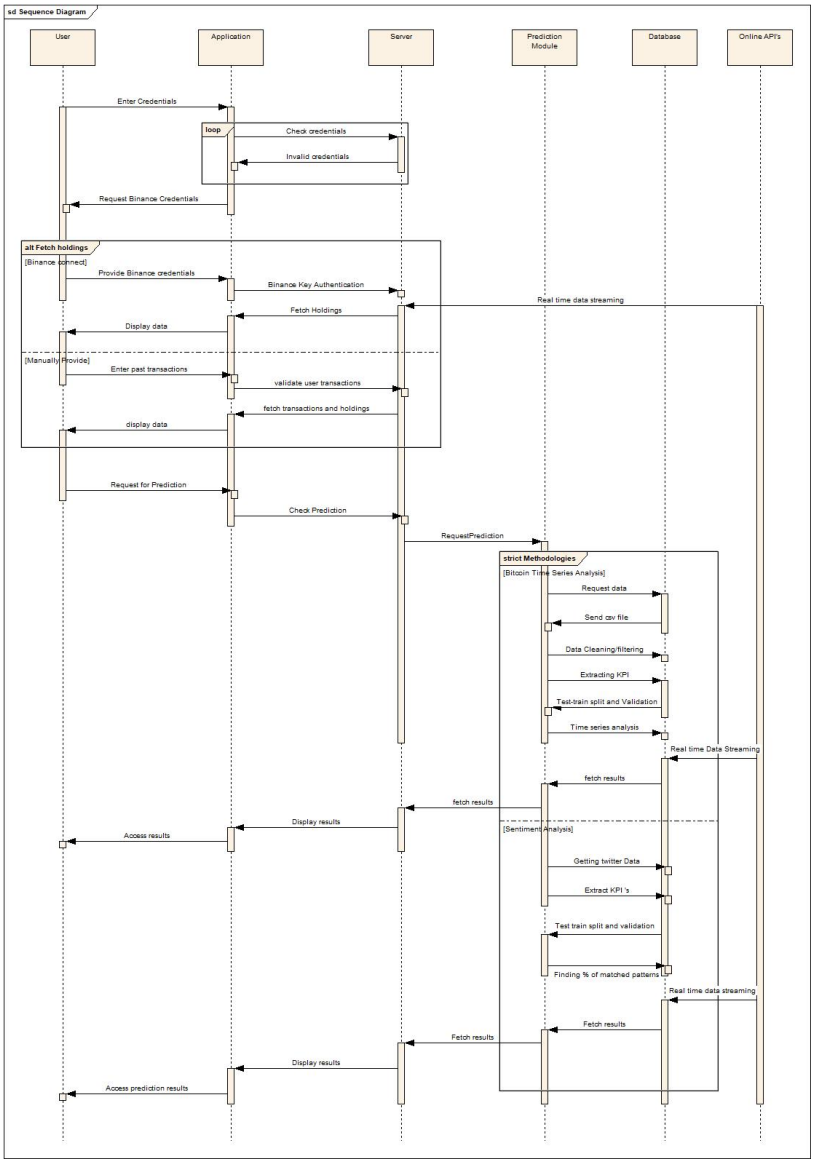


Figure 5.8: Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

5.8 Project Plan

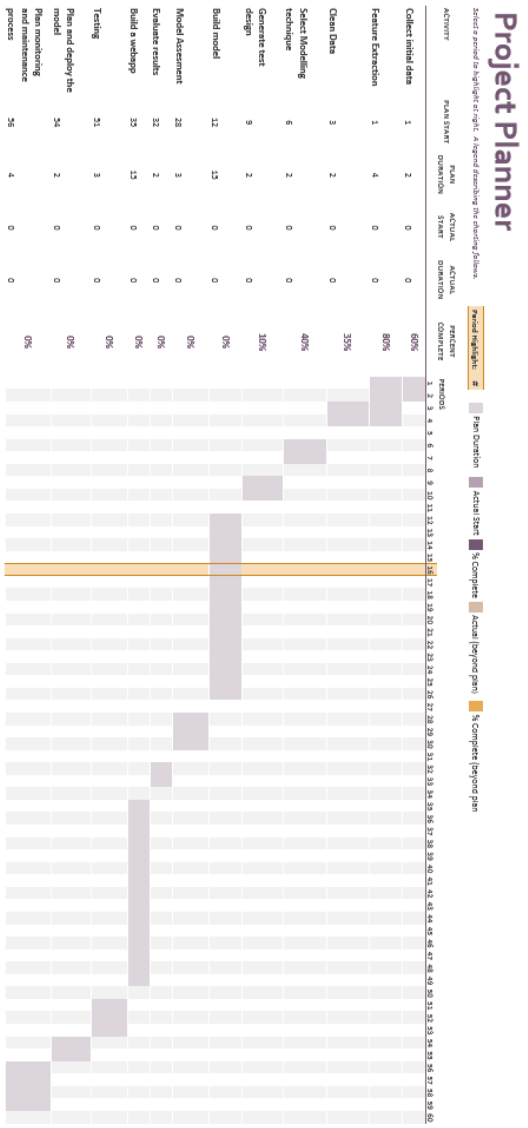


Figure 5.9: Project Plan for the final Implementation of the project.

Microsoft project planner helps to assist a project manager in developing a plan, assigning resources to tasks, tracking progress, managing the budget, and analyzing workloads

Chapter 6

CONCLUSION , FUTURE WORK AND REFERENCES

6.1 Conclusion

Looking back we can conclude that Deep learning models such as the RNN and LSTM are evidently effective learners on training data with the LSTM more capable for recognising longer-term dependencies. However, a high variance task of this nature make it difficult to transpire this into impressive validation results. As a result it remains a difficult task. There is a fine line to balance between overfitting a model and preventing it from learning sufficiently. The goal can be achieved by adopting other extended machine learning methods or considering new input capabilities related to the variability of Bitcoin. Such study will contribute to rich Bitcoin time series analysis in addition to existing Bitcoin studies. This seminar is undertaken to explain what all factors affect the price of BTC and how we can use them in proper Machine Learning algorithms and to evaluate the accuracy of Prediction.This study has found that ML bridges the gap between Statistics and BTC price.

6.2 Reference

1. S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.
2. H. Jang , J. Lee - "An Empirical Study on Modeling and Prediction of Bitcoin Prices With Bayesian Neural Networks Based on Blockchain Information," December 4, 2017.
3. R. J. Barro, "Money and the price level under the gold standard," Econ. J.,vol. 89, no. 353, pp. 13-33, 1979.
4. P. Ciaian, M. Rajcaniova, and D. Kancs, "The economics of Bitcoin price formation," Appl. Econ., vol. 48, no. 19, pp. 1799-1815, 2016.
5. Y. B. Kim1, Jun Gi Kim2, W. Kim3, Jae Ho Im3, T. H. Kim1, Shin Jin Kang2, Chang Hun Kim3,"Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies," August 17, 2016.
6. A. Greaves, B. Au,"Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin," December 8, 2015.
7. E. Stenqvist,J. Lonno - "Predicting Bitcoin price fluctuation with Twitter sentiment analysis",Kth Royal Institute Of Technology,Stockholm, Sweden 2017.
8. J. C. Soldevilla Estrada,"Analyzing Bitcoin Price Volatility," University of California, Berkeley ,May 5, 2017.
9. S. McNally, "Predicting the price of Bitcoin using machine learning," Ph.D. dissertation, School Comput., Nat. College Ireland, Dublin, Ireland, 2016.
10. I. Madan, S. Saluja, and A. Zhao, "Automated Bitcoin trading via machine learning algorithms," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep., 2015.
11. Blockchain Info. https://blockchain.info/
12. Coinbase API. https://www.coinbase.com/docs/api/overview
13. OKCoin API. https://www.okcoin.com/about/publicApi.do