CRYPTOCURRENCY TRADING PROJECTION

EAS503 Programming and Database Fundamentals for Data Scientists

Final Project- Group 7

University at Buffalo
School of Engineering and Applied Sciences

TEAM MEMBERS

Venkateswararao Para
50442199

Yeswanth Nelakurthy
50441785

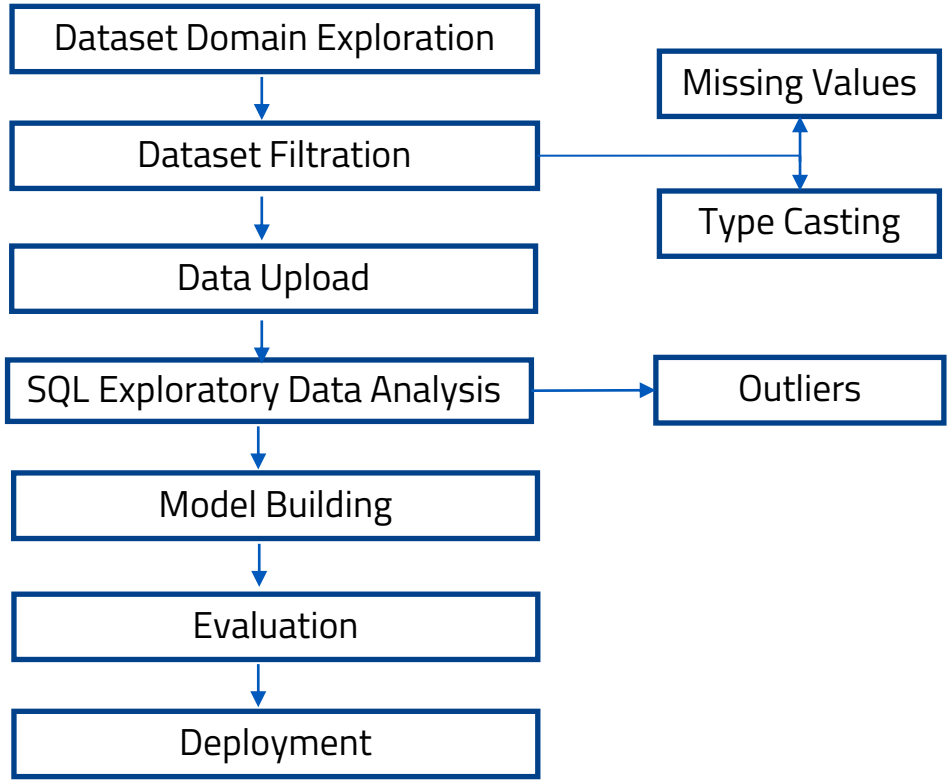
Akhil Katha
50442207

Ganesh Reddy
50442295

MOTIVATION

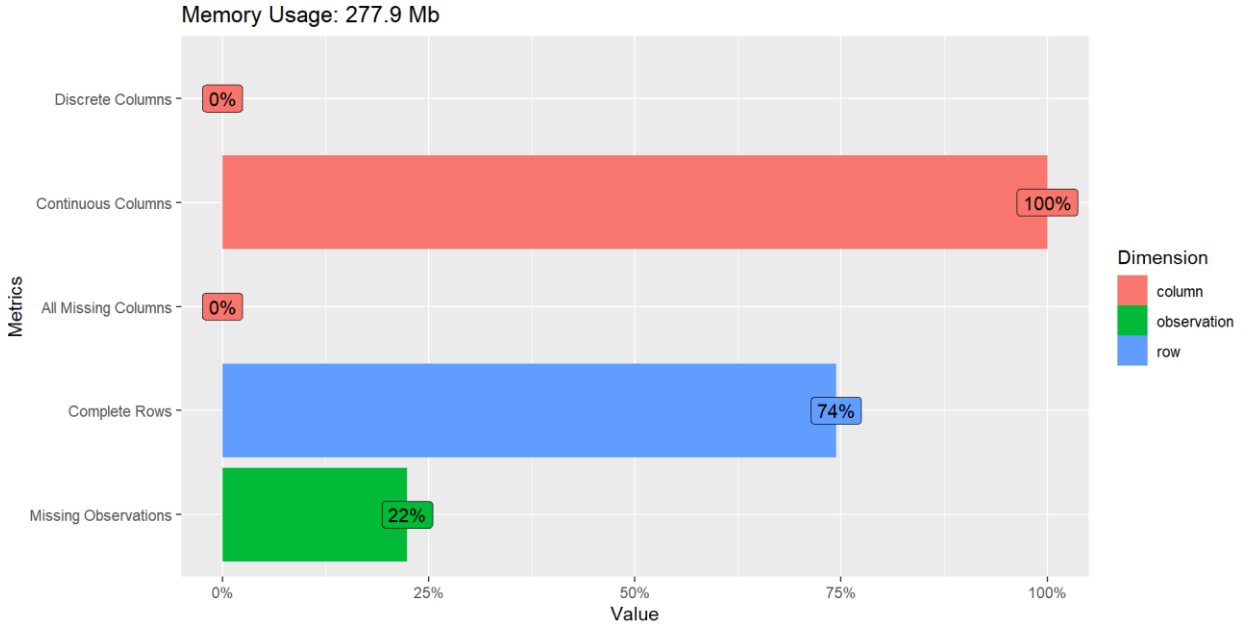
- Cryptocurrencies are here for a very long time and are volatile. In the past few years, prices fluctuated with huge oscillations. Therefore, constructing a trading strategy is extremely difficult.
- The project aim is to build a model that helps to make this trading estimation of buy or sell.
- The current dataset CSV file has bitcoin exchanges for the time period from January 2012 to September 2020. And a moving average concept is utilized to construct a target variable.

FLOW CHART


Dataset Domain Exploration
Dataset Filtration
Data Upload
SQL Exploratory Data Analysis
Model Building
Evaluation
Deployment
Missing Values
Type Casting
Outliers

DATA SUMMARY

Rows: 4857377
Columns: 8
Missing Values: 1243608 * 8
Total Observations: 4857377 * 8
Discrete Column: Target Variable



MOVING AVERAGE

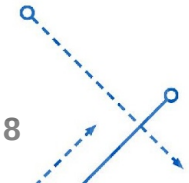
- It is a time series technique for analyzing and determining trends in the data.
- From the close feature and its mean, we are bringing out the short-term moving average (SMA1) and long-term moving average (SMA2).
- To create a target variable, the short-term moving average (SMA1) and long-term moving average (SMA2) are used.
- For the trading strategy if SMA1 > SMA2 and the signal value is 1 then we are going to buy and if SMA1<SMA2 and the signal value is 0 then we are going to sell.

DATA MODEL

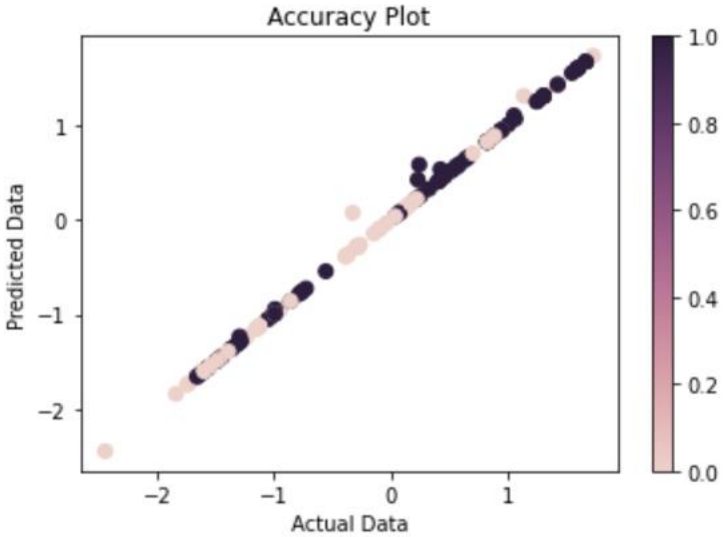
- The data in the CSV file is in the form of a collection of strings which we have split into respective columns, dealt with missing values and converted to float.
- Once a list of tuples has been created, we have created a history data table and inserted the data using execute many which has 8 columns and 3.6 million rows.
- We derived the target table from the history data table, which contains around 5 columns.
- In the Target table we derived SMA1, and SMA2 using their close feature and moving average. The signal is derived from SMA1 and SMA2.

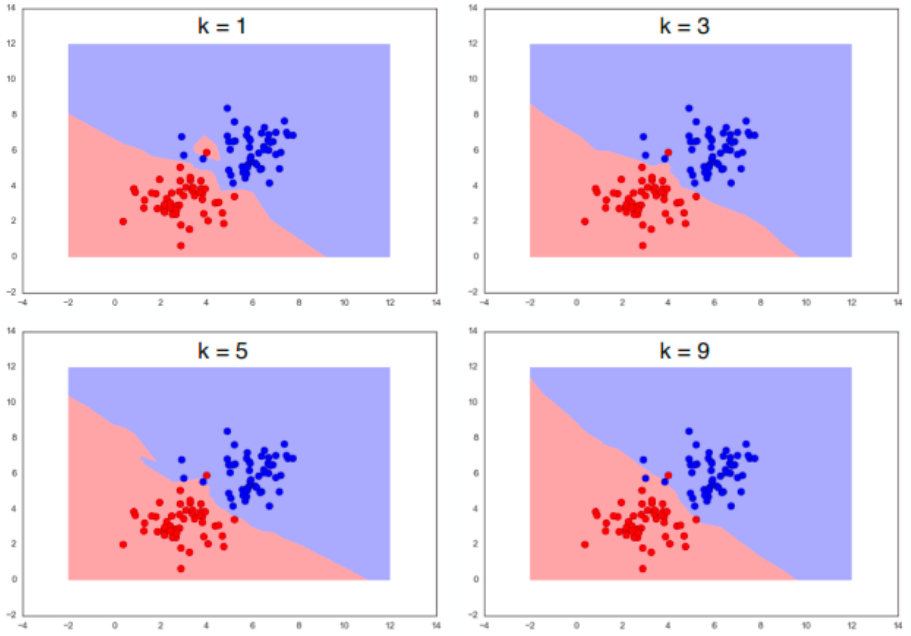
K - NEAREST NEIGHBOR

1

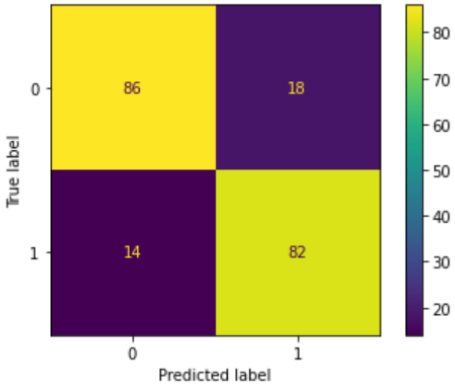


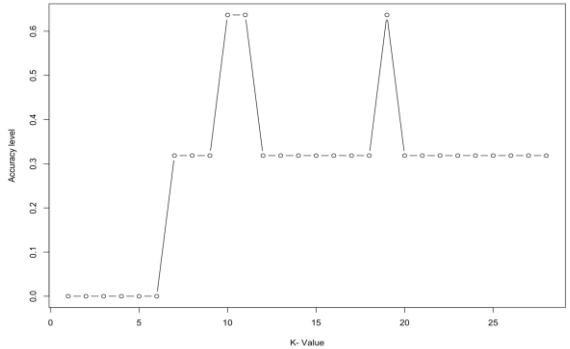
ANALYTICAL RESULTS





ACCURACY METRICS



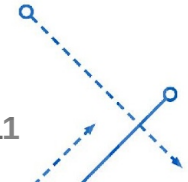


ACCURACY PLOT

- The k-nearest neighbor model achieves an impressive accuracy of 92.36% with a k - value of 9. When updated, this k value begins to improve the accuracy of the model
- The accuracy plot depicts that the model achieves high accuracy for k values 9, 11 and 19.

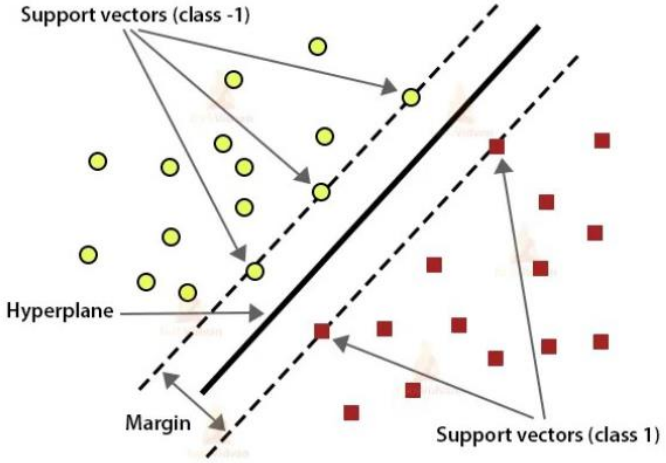
SUPPORT VECTOR MACHINE

2



OVERVIEW

- Support vectors : The points closest to the decision boundary : ie. those that define the decision boundary.
- The goal of the maximal margin classifier is to identify the linear boundary that maximizes the total distance between the line and the closest point in each class.



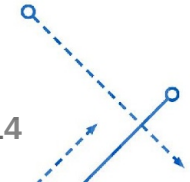
OPTIMIZATION- BEHAVIOUR

Approach:

- The SVM tuning is carried out with Linear Kernel, Polynomial Kernel, Gaussian Radial Basis Function (RBF), and Sigmoid Kernel. Selected is the linear with cost value 10.
- Randomized train test split is performed. A loop of 10 distinct variations along with cross-validation is performed. No difference is noticed in terms of behaviour.
- A perfect hyperplane with many support vectors across the margin.

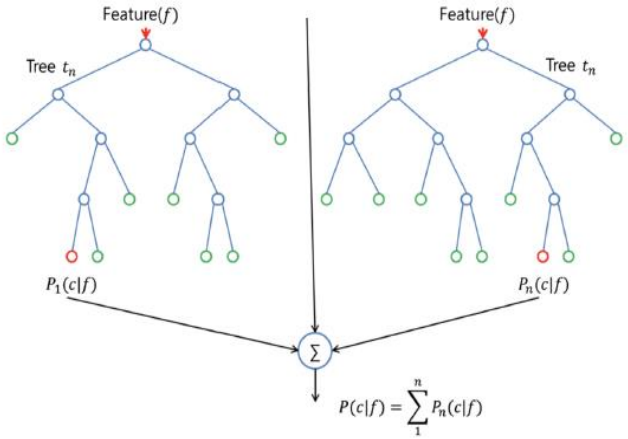
	precision	recall	f1-score	support
0	1.00	0.93	0.97	104
1	0.93	1.00	0.96	96
accuracy			0.96	200
macro avg	0.97	0.97	0.96	200
weighted avg	0.97	0.96	0.97	200

RANDOM FOREST 3



OVERVIEW

- Random forest builds on bagging by creating ensembles of independent decision trees.
- This is done first by training each tree on separate samples in the data, as is done in bagging.
- The added step is that for each tree and at each split, we randomly select a set of predictors. From the predictors, we select the optimal predictor and the optimal threshold for the split.
- Parameters to tune : the number of predictors to randomly select at each split, the number of trees, and the minimum leaf node size.



QUALITATIVE ANALYSIS - RESULTS

Method	Precision	Recall	F1 Score
Random Forest	89%	0.89	0.89
Support Vector Machine	100%	0.93	0.97
K - Nearest Neighbour	86%	0.83	0.84

CONCLUSIONS

The possible reasons of obtaining the highest accuracies across all the models could be the following:

- There might not be enough diversity in the collected data
- Perhaps we need to further enhance the model by exploratory data analysis through hypothesis testing

We systematically followed the right process using principles/methods we were taught in class

THANK YOU!

Any questions?