T.C.
FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ
Faculty Of Enginnering Computer Engineering Department

DATA MINING PROJECT: MILESTONE

PREDICTING CRYPTOCURRENCY

İREM ERÇEL ,1921221038

AHMET YANIK, 1821221001

Dr. Gönül ULUDAĞ

STEP 2: MILESTONE

Our Dataset: https://www.kaggle.com/datasets/akhil14shukla/binance-coin-prices-per-day

a) Motivation: What problem are you tackling, and what's the setting you're considering?

The aim is predicting cryptocurrency prices. This is a challenging task due to the inherent volatility and complexity of the cryptocurrency market. The setting involves historical data of cryptocurrency prices along with associated features such as trade count, volume, and open, high, low, and close prices. Our goal is to predict the price accurately, recognizing the difficulty of the task and planning to refine the model by moving to finer time granularity (hourly or minute data) in the future.

b) Method: What data mining techniques have you tried and why?

Exploratory Data Analysis (EDA): Understanding the distribution of your data, checking for outliers, and visualizing relationships between features.

Time Series Analysis: Given the temporal nature of your data, time series analysis techniques such as decomposition, autocorrelation analysis, and trend identification might be helpful.

Feature Engineering: Creating new features such as price change percentages, moving averages, or technical indicators that might capture patterns in the data.

Data Standardization and Normalization: Ensuring that features are on similar scales, especially when using algorithms sensitive to feature scales.

Baseline Algorithm: Applying a baseline algorithm (e.g., linear regression, decision tree) to establish a benchmark performance and compare the effectiveness of more sophisticated models.

Principal Component Analysis (PCA): As you've mentioned, PCA for dimensionality reduction to capture the most important features and reduce noise.

Machine Learning Models: Trying various regression models such as linear regression, decision trees, random forests, or even more advanced models like gradient boosting or neural networks.

c) Preliminary experiments: Describe the experiments (Preprocessing & Feature Engineering) that you've run, the outcomes, and any error analysis that you've done. You should have tried at least one baseline algorithm.

its included preprocessing steps like handling missing values, converting date columns to datetime objects, and creating the 'Price Change' feature as described earlier.And also experimented with different strategies for handling outliers, scaling features.

For feature engineering, we created the 'Price Change' feature to capture the percentage change in closing prices. This could be a valuable feature for predicting future prices.

In terms of baseline algorithms, we will apply a simple linear regression model to establish a baseline for comparison.

Error analysis would involve assessing the performance of model, examining residuals, and identifying patterns or trends that the model might be missing.

The cryptocurrency market is highly unpredictable, and accurate predictions are challenging. We will Continue refinement of our models

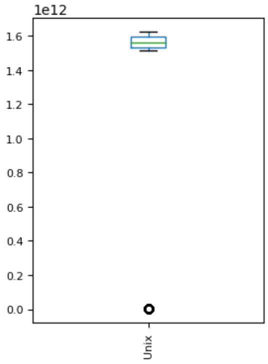
d) Next steps: Given your preliminary results, what are the next steps that you're considering?

PREPROCESSING & FEATURE ENGINEERING STEPS:

1) Preprocessing:

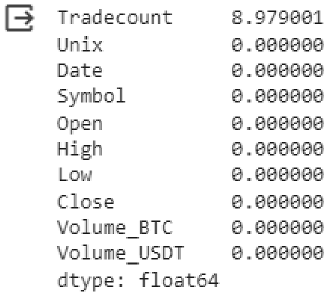
i. Check whether there are any outliers? Remove outliers (if there are any) in your data.

There is 1 outlier in the unix column.



ii. How many columns include missing values?

There are 124 values in just one column called tradecount.



```
Tradecount     8.979001
Unix           0.000000
Date           0.000000
Symbol         0.000000
Open           0.000000
High           0.000000
Low            0.000000
Close          0.000000
Volume_BTC     0.000000
Volume_USDT    0.000000
dtype: float64
```

iii. Explain your method to handle each of those missing values.

Only had null values in 1 row. So I filled them in by averaging the values.

iv. Explain if you need to apply any kind of transformations and/or encoding.

2) Feature Engineering

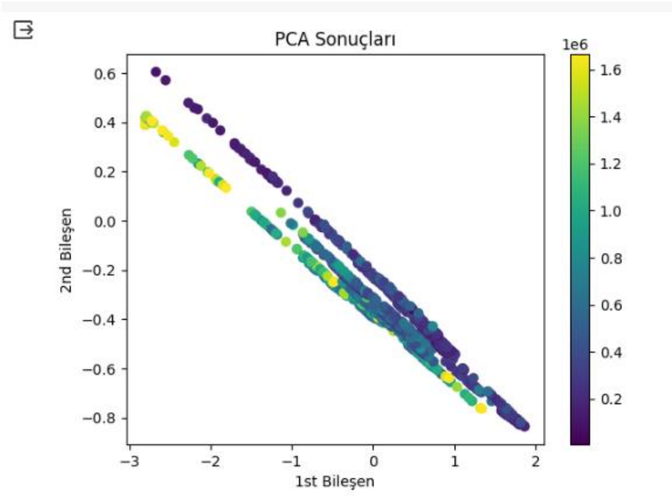
i. Determine which features are the most valuable and whether you need to create a new feature?

Calculated the correlation between each feature and the target variable (cryptocurrency price). Features with high correlation more valuable.

ii. Select (i.e., filter) or create features that make data mining algorithms work.

A new column called 'Price_Change' has been created, representing the percentage change of the closing prices in the 'Close' column. This column contains the percentage change of each day's closing price compared to the previous day. A subset of the DataFrame consisting of the 'Date', 'Close', and 'Price_Change' columns are printed to the screen.

iii. Apply dimensionality reduction i.e. PCA to your data.



Principal Component Analysis (PCA) is a dimensionality reduction technique used to reduce variability in multidimensional data sets and reveal important features (components). PCA is widely used in analysis such as modelling, visualization and classification, especially in high-dimensional data sets.

1. Data Mining Algorithms:

a. Linear Regression:

- **Advantages:**
  - Simple and easy to understand.
  - Computationally efficient.
  - Provides coefficients that can be interpreted.
- **Disadvantages:**
  - Assumes a linear relationship between features and target.
  - Sensitive to outliers.
- **Brief Definition:** Linear Regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation.
- **Parameters to be Tuned:**
  - The main parameter is the learning rate in the context of stochastic gradient descent.

b. Decision Tree:

- **Advantages:**
  - Easily interpretable.
  - Handles non-linearity well.
  - No need for feature scaling.
- **Disadvantages:**
  - Prone to overfitting.
  - Sensitive to small variations in the data.
- **Brief Definition:** Decision Trees recursively split the data based on features to make decisions.
- **Parameters to be Tuned:**
  - Maximum depth of the tree, minimum samples split, and minimum samples leaf are common parameters to tune.

c. Random Forest:

- **Advantages:**
  - Reduces overfitting by combining multiple trees.
  - Handles missing values and outliers well.
- **Disadvantages:**
  - Can be computationally expensive.
- **Brief Definition:** Random Forest is an ensemble learning method that constructs a multitude of decision trees and merges them together to get a more accurate and stable prediction.
- **Parameters to be Tuned:**
  - Number of trees in the forest, maximum depth of the trees, minimum samples split, and minimum samples leaf.

4. Hyperparameter Tuning:

We can use techniques like Grid Search or Random Search to find the best hyperparameters for each algorithm.

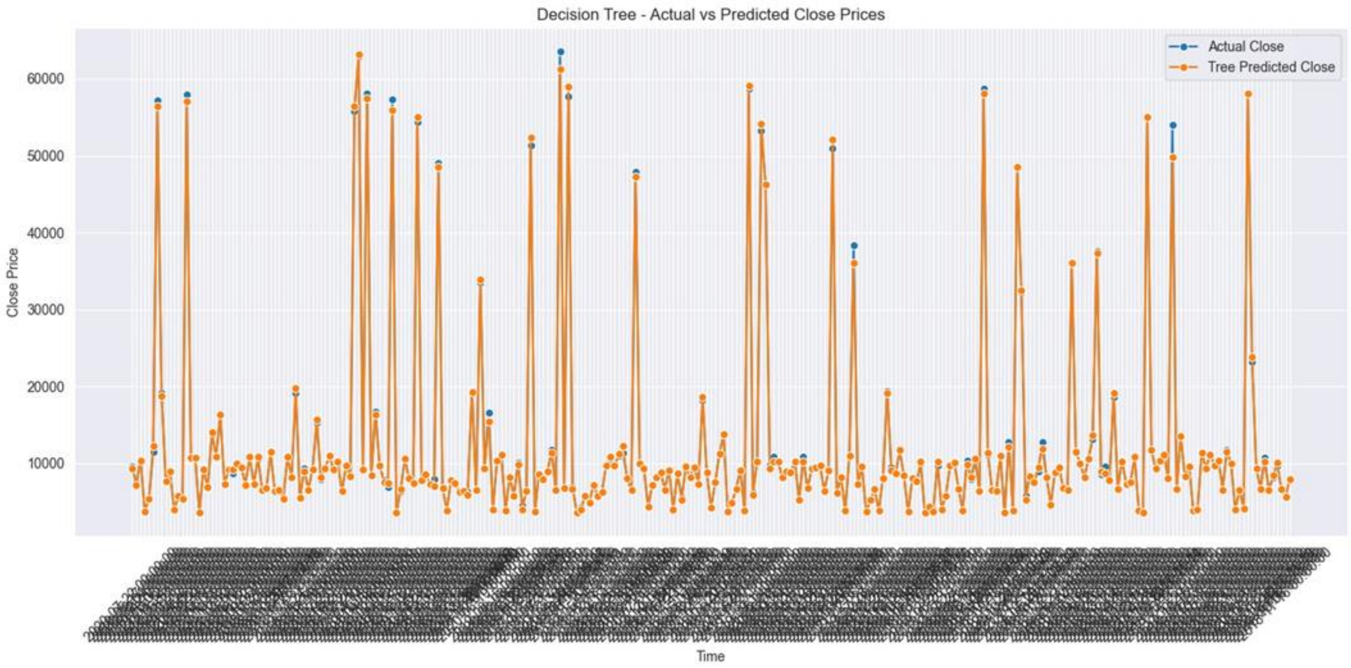
5. Comparison:

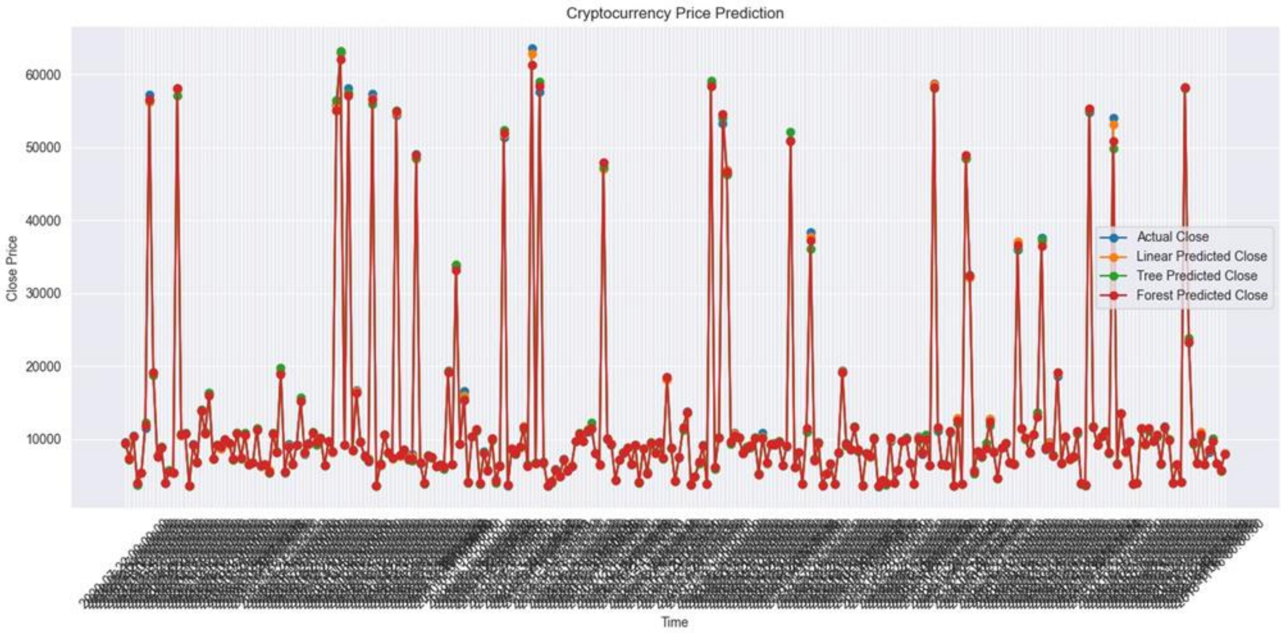
Numerical Comparison:

Evaluated performance metrics such as Mean Squared Error (MSE) or R-squared for each algorithm.

Visual Comparison:

Plot actual vs. predicted values for each algorithm to visually assess their performance.





FINAL WRITEUP

İrem Erçel,1921221038

Ahmet Yanık, 1821221001