VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering (SCOPE)

MTech-Business Analytics

FALL INTER SEMESTER

CSE3088-Artificial Intelligence and Knowledge based systems

"Time Series Forecasting of Bitcoin Prices using LSTM and RNN with Particle Swarm Optimization and Grey Wolf Optimizer"

by

S Narthana(21MIA1124)

Sammeta Lekhana(21MIA1080)

exploratory data analysis (EDA)

(BITCOIN DATASET)

1. Data Understanding

- Count the number of rows and columns in the dataset.
- Check the data types of each column.
- Calculate the total number of missing values in each column.
- Identify the unique values in categorical columns.
- Calculate the average, minimum, and maximum values for numeric columns.
- Determine the range (difference between minimum and maximum values) for numeric columns.
- Find the most frequent values in categorical columns.
- Check for duplicate rows in the dataset.
- Calculate the correlation between numeric columns.
- Identify the top N rows with the highest Bitcoin prices.
- Calculate the average Bitcoin price per year.
- Determine the number of unique dates in the dataset.
- Group the data by month and calculate the average Bitcoin price for each month.
- Identify the top N rows with the largest daily price changes.
- Determine the earliest and latest dates in the dataset.
- Calculate the average Bitcoin price for weekdays and weekends separately.
- Group the data by year and calculate the total trading volume for each year.
- Calculate the average Bitcoin price for each day of the week (Monday to Sunday).

Analysis Task	Purpose	Output / Result	Inference
Count Rows and Columns	Obtain the dimensions of the dataset.	Number of rows and columns in the dataset.	Understanding dataset size.
Check Data Types	Determine the data types of each column.	Data types of each column in the dataset.	Identifying the nature of variables.
Missing Value Count	Identify the total number of missing values in each column.	Number of missing values in each column.	Assessing the data quality.
Unique Values	Find the unique values in categorical columns.	Unique values in each categorical column.	Understanding categorical variables.
Numeric Stats	Calculate the average, minimum, and maximum values for numeric columns.	Average, minimum, and maximum values for each numeric column.	Understanding the data distribution.
Range	Determine the range (difference between min and max) for numeric columns.	Range (max - min) for each numeric column.	Assessing the data spread.

Analysis Task	Purpose	Output / Result	Inference
Most Frequent Values	Identify the most frequent values in categorical columns.	Most frequent values in each categorical column.	Understanding common categories.
Duplicate Rows Check	Check for the presence of duplicate rows in the dataset.	Number of duplicate rows in the dataset.	Assessing data duplication.
Correlation	Calculate the correlation between numeric columns.	Correlation matrix between numeric columns.	Understanding relationships.
Top N Bitcoin Prices	Find the top N rows with the highest Bitcoin prices.	N rows with the highest Bitcoin prices based on the 'Close' column.	Identifying high price instances.
Average Price per Year	Calculate the average Bitcoin price per year.	Average Bitcoin price for each year based on the 'Close' column.	Understanding price trends by year.
Unique Dates Count	Determine the number of unique dates in the dataset.	Number of unique dates in the 'Date' column.	Assessing the time period covered.

Analysis Task	Purpose	Output / Result	Inference
Average Price per Month	Group the data by month and calculate the average Bitcoin price.	Average Bitcoin price for each month based on the 'Close' column.	Understanding price trends by month.
Top N Daily Price Changes	Identify the top N rows with the largest daily price changes.	N rows with the largest price changes based on the 'Price Change' column.	Identifying extreme price movements.
Earliest and Latest Dates	Determine the earliest and latest dates in the dataset.	Earliest and latest dates in the 'Date' column.	Assessing the time period covered.
Average Price Weekdays/Weekends	Calculate the average Bitcoin price for weekdays and weekends.	Average Bitcoin price for weekdays and weekends based on the 'Close' column.	Understanding price patterns based on weekdays and weekends.
Total Trading Volume per Year	Group the data by year and calculate the total trading volume.	Total trading volume for each year based on the 'Volume' column.	Assessing trading activity by year.
Average Price per Day of the Week	Calculate the average Bitcoin price for each day of the week.	Average Bitcoin price for each day of the week based on the 'Close' column.	Understanding price patterns by day of the week.

2. Data Visualization

- Line Plot: Bitcoin price over time
- Bar Plot: Daily Bitcoin trading volume
- Box Plot: Bitcoin price distribution by year
- Scatter Plot: Bitcoin price vs. trading volume
- Kernel Density Plot: Kernel density estimation of Bitcoin price
- Box Plot: Bitcoin price distribution by month
- Histogram: Distribution of Bitcoin price by year
- Line Plot: Bitcoin price moving average
- Heatmap: Correlation matrix of numerical variables
- Pie Chart: Distribution of Bitcoin price categories

Analysis	Purpose	Expected Output	Result and Inference
Line Plot	Visualize Bitcoin price over time	Line plot of Bitcoin price	Identify trends and patterns in Bitcoin price over time
Bar Plot	Analyze daily Bitcoin trading volume	Bar plot of daily trading volume	Understand the volume of Bitcoin traded on a daily basis

Analysis	Purpose	Expected Output	Result and Inference
Box Plot	Explore Bitcoin price distribution by year	Box plot of Bitcoin price by year	Identify the central tendency and spread of prices per year
Scatter Plot	Examine the relationship between Bitcoin price and trading volume	Scatter plot of price vs. volume	Determine if there is a correlation between price and volume
Kernel Density Plot	Estimate the probability density of Bitcoin price	Kernel density plot of Bitcoin price	Understand the distribution and shape of Bitcoin price
Box Plot	Analyze Bitcoin price distribution by month	Box plot of Bitcoin price by month	Identify any monthly trends or outliers in Bitcoin price
Histogram	Explore the distribution of Bitcoin price by year	Histograms of Bitcoin price by year	Understand the distribution of prices across different years
Line Plot	Visualize Bitcoin price moving average	Line plot of Bitcoin price and moving average	Observe trends and smooth fluctuations in Bitcoin price
Heatmap	Analyze the correlation between numerical variables	Heatmap of correlation matrix	Identify relationships and dependencies between variables

Analysis	Purpose	Expected Output	Result and Inference
Pie Chart	Understand the distribution of Bitcoin price categories	Pie chart of Bitcoin price categories	Gain insights into the proportion of Bitcoin price ranges

3. Data Pre-processing

- Check for missing values
- Convert 'Date' column to datetime
- Extract year, month, and day from 'Date' column
- Check for duplicate rows
- Check for outliers in 'Close' column
- Remove outliers
- Normalize 'Close' column using Min-Max scaling

Analysis	Purpose	Expected Output	Result Inference
Check for missing values	Identify any missing values in the dataset	Number of missing values for each column	Determine the extent of missingness in the dataset

Analysis	Purpose	Expected Output	Result Inference
Convert 'Date' column	Transform the 'Date' column into a datetime format	Updated 'Date' column with datetime format	Enable convenient manipulation of date-related features
Extract year, month, and	Extract individual components (year, month, day) from the 'Date' column	New 'Year', 'Month', and 'Day' columns	Enable analysis and modeling based on specific time components
Check for duplicate rows	Identify and quantify the presence of duplicate rows	Number of duplicate rows	Assess the impact of duplicate rows on the dataset
Check for outliers in	Detect potential outliers in the 'Close' column	Outliers flag for each data point	Determine the presence and extent of extreme values in the 'Close' column
Remove outliers	Remove the identified outliers from the dataset	Updated dataset without outliers	Improve the robustness and accuracy of subsequent analysis
Normalize 'Close'	Apply Min-Max scaling to normalize the 'Close' column	Scaled values for 'Close' column	Ensure comparability and consistency of 'Close' values

4. Feature Engineering

- Calculating the daily percentage change in Bitcoin price using the pct_change() function.
- Computing the 7-day rolling mean of the Bitcoin price using the rolling() function with a window size of 7.
- Estimating the exponential moving average (EMA) of the Bitcoin price using the ewm() function with a span of 30.
- Creating lagged variables by shifting the Bitcoin price by 1 day and 7 days using the shift() function.
- Removing rows with missing values after feature engineering using the dropna() function.

Feature Engineering Step	Purpose	Output/Result Inferences
Calculate daily percentage change	Capture the daily price movement as a percentage, which can help identify trends and volatility.	New feature column 'Price_Change' with daily percentage change.
Compute 7-day rolling mean	Smooth out short-term fluctuations, providing a trend indicator over a 7-day period.	New feature column 'Rolling_Mean' with the rolling mean values.
Estimate exponential moving average (EMA)	Identify long-term trends in the Bitcoin price by giving more weight to recent data points.	New feature column 'EMA' with the exponential moving average values.

Feature Engineering Step	Purpose	Output/Result Inferences
Create lagged variables (shift by 1 day and 7 days)	Incorporate past Bitcoin prices as features, capturing historical patterns and dependencies.	New feature columns 'Lagged_Price_1' and 'Lagged_Price_7' with shifted Bitcoin prices.
Remove rows with missing values	Ensure the dataset has complete data for subsequent analysis and modeling.	Updated dataset with missing rows removed.

5. FEATURE SELECTION

feature selection method called "Feature Importance" using the Random Forest algorithm

Analysis	Purpose	Output	Inference
Feature Importance	Identify the most important features for prediction	Feature importance scores for each variable	1. "Adj Close" has the highest importance in predicting Bitcoin price.
			2. "High" and "Open" are also significant features.

Analysis	Purpose	Output	Inference
			3. "Low" has relatively lower importance compared to other features.
			4. "Volume" has very low importance and may not contribute significantly to the prediction.

6. PCA (Principal Component Analysis)

Linear Discriminant Analysis (LDA)

Technique	Purpose	Output	Result Inference
PCA	Dimensionality Reduction	Transformed dataset in reduced dimension	- Reduced dimension representation of the dataset - Explained variance ratio
LDA	Dimensionality Reduction & Classification	Transformed dataset in reduced dimension	- Reduced dimension representation of the dataset - Improved class separability

For PCA:

- Purpose: PCA is applied to reduce the dimensionality of the dataset while capturing the most important information. It helps to identify patterns and relationships among the features.
- Output: The transformed dataset is obtained with a reduced number of dimensions. This output can be used for further analysis or visualization.
- Result Inference:
    - The reduced dimension representation obtained from PCA can be used to visualize the Bitcoin price data in a lower-dimensional space.
    - The explained variance ratio indicates the amount of variance in the original data explained by each principal component. Higher values indicate that the principal component captures more information from the original data.

For LDA:

- Purpose: LDA is applied for dimensionality reduction while considering the class labels. It aims to find a lower-dimensional representation that maximizes class separability.
- Output: The transformed dataset is obtained in a reduced dimension space, considering the class information. This output can be used for further analysis or visualization.
- Result Inference:
    - The reduced dimension representation obtained from LDA can help visualize the Bitcoin price data while considering the different classes or categories.
    - The reduced dimension space obtained from LDA is expected to have improved class separability, which can be beneficial for classification tasks related to Bitcoin price prediction.

7. k-means clustering

K-means clustering	Discover clusters or patterns in Bitcoin prices	Clusters or patterns identified in the data	K-means clustering can help identify distinct groups or patterns within the Bitcoin price data, aiding in market analysis.