**Introduction**

Global financial system depends badly on foreign exchange market and sometime it is referred to as foreign exchange. There is an area where money is traded and prices of goods is called exchange rates. The GBP/USD pair representing exchange rate between British Pound and US Dollar. The dollar line is the most traded currency pairs among many currency pairs traded in this market.

Its ability to predict futures value of GBP/USD exchange rate for wide range of stakeholders are invaluable. For example traders and investors depends on accurate forecasts to enable them to make informed decisions and to safeguard their assets from currency fluctuations and to benefit from them also government and financial authorities oversight and take these forecasts into consideration to better tell their investment and monetary policy with they can develop connections or foreign markets in international trade for risk management and budgeting and strategic planning and creating them income and financial stability protection against wide range of currency up and down.

As there is great risk in any accurate predicting of future exchange rates. researchers and practitioners have been looking for experienced analytical techniques. These have proven data mining and machine learning methods to be very useful tools because patterns and insights can be obtained from big data.

Two such validated methods is main strengths of this study is clustering and regression using K-means random forest. An unsupervised learning method called K implies that data are grouped according to similarity. It can be used to look for patterns or trends in past events that may control currency prices when forecasting future exchange rates. For example it can classify days according to common patterns in the amount of money moving and this information can then be used to train predictive models.

On other hand RFM is a supervised learning method that builds a large number of decision trees and increases accuracy by combining their predictions This method more confidently predicts GBP/USD rate by training model that is provided on historical data. Its strength lies in handling big data with lots of instability which is why it works so well with the complex changing aspects of forex market. This study aims to provide detailed knowledge of variables that drive GBP/USD exchange rate movements and to provide predictive models that can help inform investment decisions using this advanced framework and provide insightful information for on international monetary policy.

**Methodology**

The project methodology conforms to six-step Cross-Industry Standard Procedure for Data Mining (CRISP-DM).

**Business logic:** The project objectives and requirements from a business perspective and then translate this knowledge into data mining problem definition and planning importance.

**Data Understanding:** Initial data collection that is followed by activities to better identify data and identify data quality issues identify initial awareness from the data or identify subgroups of interest to generate hidden information theory.

**Data preparation:** This phase includes all activities to create final data from original raw data such as data preparation and case selection and data transformation and normalization for modeling.

**Modeling:** Selecting and applying various modeling techniques and adjusting model parameters to their appropriate ranges. In this application, K-means clustering is used to identify models and Random Forest Regression is used for its predictive capabilities.

**Evaluation:** Wide evaluation of model is required past to implementation. To evaluate performance of model.This function uses several metrics such as inertia and silhouette score and mean absolute error MAE and root mean squared error RMSE.

**Deployment:** This phase can include anything from creating reports to developing an iterative data mining process. In this case specification and conclusion of model are part of process.

**Data Collection and Understanding**

The first stage of data analysis process in this case GBP/USD exchange rate project and it is careful data collection. This special phase is collecting multiple datasets from reputable and authentic and historical and financial sources that a focus on Yahoo’s finances. The dataset selected for this study includes important financial information such as trading volume and daily fluctuations in exchange rates all of which are carefully recorded with open, high, low, close, adjusted closed price for each trading day.

This dataset requires a multi-split approach that combines graphical and statistical analysis techniques to fully capture its tones and complexities Graphical analysis provides a clearer understanding of market behavior by budding and modeling data over time. To identify trends and volatility this requires a chart of daily closing prices over a specific period of time. In difference statistical analysis examines data sets in depth using statistical methods to analyze and understand relationships and patterns in data.

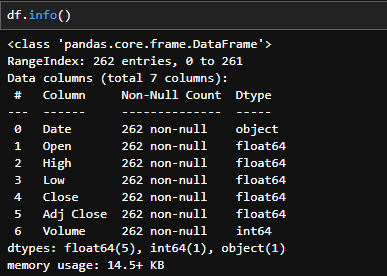
The first research phase is important because it sets overall framework for project. This includes identifying appropriate stages of analysis and selecting appropriate software and analytical tools understanding effects or consequences that may arise. This phase does not involve data collection not only but also provides accuracy and consistency. To this end summary statistics are performed to obtain an overview of data including average exchange rates and variances and trading volumes in addition to classifying data according to criteria that have appropriate feel and such as for time differences and market conditions and or anything abnormal.

Identifying any inconsistencies missing quantities or unimportant materials that may compromise quality of data is possible through this hard classification and initial analysis. It is important to address these problems’ role from beginning in order to preserve objectives of analysis. It provides guidance for subsequent preparation steps and which may include handling missing values reducing data and changing variables. By providing a solid foundation for data analysis this preparation enhances validity and reliability of projects findings.

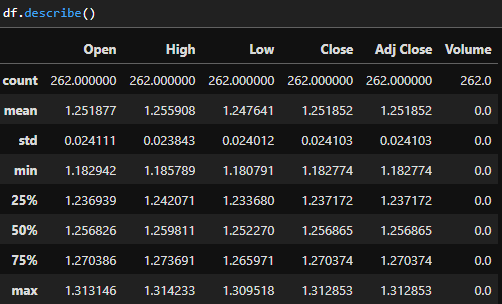
**Data Preprocessing**

First important step to generate clean data from raw data is data preparation. For GBP/USD exchange rates previous process included following steps.

**Cleaning:** Improving data quality by finding and retrieving and removing errors or any inconsistencies.



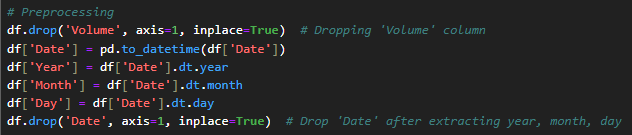
With the help of **df.info()** we can check if there any null values present in dataset and datatype of rows.



And with help of describe method we can find mean standard deviation and all other things to make data prepare for the model and to remove outliers.

**Feature Selection:** The selection of most important attributes for inclusion in the models is called feature selection. Volume is not included in this work for sake of uniformity.

**Feature engineering:** This is process of forcing useful additional information from available data and such as breaking down a date into its month, day, and year.



Here we did both feature selection and feature engineering to make data in good form.

**Normalization:** Using standardscaler to ensure that each feature contributes equally to analysis. This is especially important for K-means models in clustering where performance of method can be affected by scalability.

A black background with red and blue text

Description automatically generated

**Model Development**

The model development phase is when actual data mining takes place. Due to predictive nature of GBP/USD exchange rate, two models were developed:

• The supervised random forest regression learning method was chosen because of its robustness against overfitting and its ability to handle nonlinear data by constructing multiple decision trees and working by combining results. This following steps were important in random forest modeling.

• Feature selections, with inputs between the target variable Close. Modifying parameters, deciding thickness and depth of trees in the forest. Pattern training involves detection of trends in prior data.

• K-means Clustering: K-means was used as an unsupervised learning method to find genetic patterns or clusters in data that could provide information for predictive models. The action taken was as follows:

• Decide how many decision clusters ‘k' to search in data. The normalized inputs of random forest model are used for feature selection, but the target variable is not included. Clustering is process of repeatedly assigning each data point and according to feature similarity and into one of ‘k’ groups. Definition of clusters to illustrate structure of data and provide direction for regression model.

**Model Evaluation and Results**

This time several metrics were used to evaluate execution of both models. An improvised retreat into woods.

**MAE, or Absolute Error**

Calculates amount of error in a series of predictions regardless of direction of error because scores are linear and each individual difference has equal weight in the mean.

**Mean square root of error (RMSE)**

This metric is used to evaluate square root of difference in the mean square root of difference between actual watching and forecasts. Larger errors are given slightly higher weights and which can be helpful when larger errors is highly undesirable. The random forest model produced the smallest average prediction error and showed the greatest accuracy MAE 0.0103 and RMSE 0.0140.

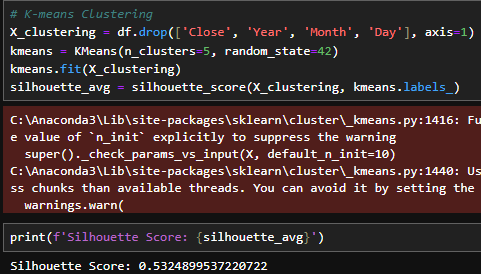
A screen shot of a computer program

Description automatically generated

Here we fitted model and find out the mae and rmse to check model result and it gave values in 0s which shows very good result of our model but here our model can overfit with this so we have to check this model on other dataset too.

**K-means Clustering: Score of the silhouette**

This metric compares shape of object and its own cluster with other clusters. An object with a high silhouette score corresponds well to its own cluster and poorly to neighboring clusters. A Silhouette score of 0.3093 indicated low group involvement in K-means model.



The silhouette score in code of 0.5324899537220722 indicates that on average the data points are relatively far from other clusters and correspond reasonably well to their clusters. This positive result is slightly above 0.5. There is an important structure within groups and although there may be some overlap or groups are not very far from each other. It’s not a perfect score, which would be close to 1 and but its also not a negative score and which would be close to 0 or negative.

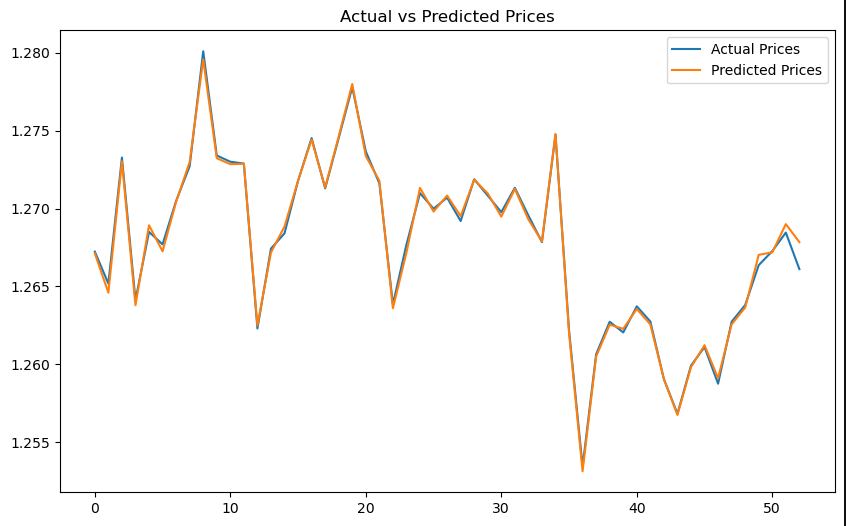
**Illustrations and explanations**

Actual vs. Predicted Values and Visualization of Clusters graphs and among other visualizations provided clear insights into data structure and model performance.

The results of analysis and drawings give a clear picture of structure and performance of model.

**Actual vs. predicted difference**

This line graph shows a direct comparison between predicted value of random forest regression model and actual closing value of time series. The goal is to check how well the model predicts actual values. The accuracy of model in predicting exchange rates is indicated by goodness of fit of two lines.



**Cluster Visualization**

Using normalized features open and adjusted close values the scatter plot shows clusters found by K-means algorithm One-day trading data values are indicated by each location in plot and color coding identify groups. Points in different groups should be well apart and while points in same group should be close to each other. The story makes it easier to understand behavior of market and to see relationship between different trading days.

**Residual Plot**

The figure shows the residual plot. In regression analysis the graph like this is used to show residuals and or errors and between actual and projected values. Residuals are shown on the y-axis and while expected values are shown on x-axis. For residual values of 0 the residuals should be randomly distributed along horizontal axis. If residuals are randomly distributed and show no discernible pattern model is fine. If patterns appear and it indicates that the model lacks some detailed structural information in data. Your residual plot shows a small curve and indicating that the model may need to be adjusted to nature of data.

A graph of a graph showing the value of a product

Description automatically generated with medium confidence

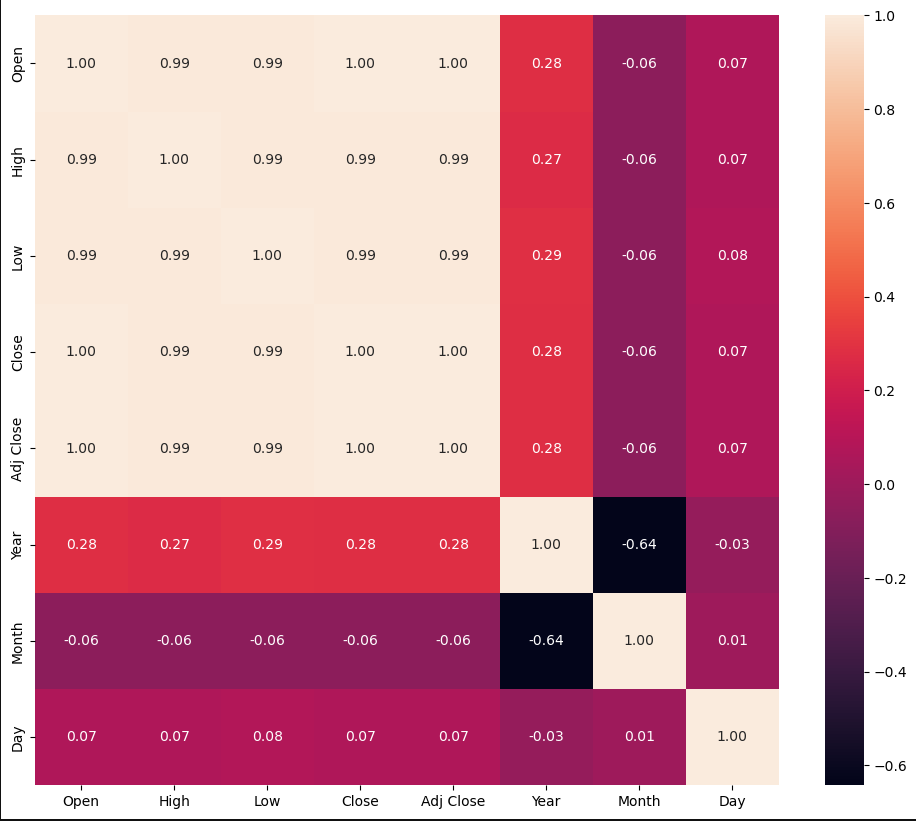
**Heatmap**

• This figure is a heat map of relationship. This graph of correlation matrix of a data set is an effective way to show direction and strength of correlations between variables. Correlations between two variables are represented by each cell in the grid.

• The aim and strength of the interaction are indicated by the color and intensity of cells. Strong positive or negative correlations are indicated by values around 1 or -1 while values closer to 0 indicate any linear relationship.

• The heat map you shared shows rows and the first five labeled characters so that Open and High and Low and Close Adj Close are strongly correlated with each other and indicating that these variables move in same direction.

• This is a common practice in financial accounts and where these are in time on trading day. Although there is little or no correlation between month and day and year shows a positive moderate correlation with these values.



**Discussion on Implications of Results**

**Statistical summary:** The statistical analysis shows that exchange rates vary within a small period of time during the period analyzed and with an average value of around 1.25 for open, high, low and close prices.

**Random forest regression model:** A random forest regressor was trained to predict near future values using historical data. With lowest mean absolute error (MAE) of about 0.0002476 and root mean square error (RMSE) of about 0.0003576, model showed exceptional accuracy in its prediction of experimental data set.

**Comparing numerical prices to numerical values:** The good agreement between actual and forecasted prices indicates that model is able to adequately capture the movement of exchange rates.

**K-means Clustering:** K-means clustering was used to remove date features and objective variables close value characters also examined pattern recognition. Silhouette score of about 0.5325 indicates that K-means showed a good structure and which shows patterns different or clustering in data based on target variables possible.

**Cluster visualization:** Groups of data and showing similar trading patterns or price movements over different periods and are scatter plotted showing clusters based on adjusted close and open prices.

**Conclusion and Future Work**

The studies highlight superb potential of machine learning techniques to enhance predictive capabilities in finance and with particular emphasis on identifying key market patterns that determine GBP/USD exchange rate ups and downs. Specific attention was paid to compelling forecasting performance of random forest model and raising possibility that machine learning can be an effective tool for economic forecasting.

Despite these successes and study also acknowledges challenges encountered and especially in application of clustering techniques. The observed decrease in performance of clustering methods points towards inherent complexity of financial markets and which are influenced by a multitude of factors that can change dynamically over time. This complexity makes it difficult to capture full extent of market behaviors and suggests that current models might only be scratching the surface of what can be achieved with more sophisticated analytical tools.

Even with these achievements paper also notes challenges that existed and especially when implementing clustering methods. The observed decline in aggregate technical performance reflects the true complexity of financial markets and with many strongly variable effects. These complexities make it difficult to fully capture market behaviors.

Study makes several recommendations for future research directions to improve our understanding and ability to predict market developments. It encourages search for more undeveloped machine learning algorithms that are better equipped to handle complex financial data. Furthermore this including additional variables in analysis such as economic data and for example it can provide a more complete understanding of factors affecting market movements. The study also highlights potential benefits of better identifying and understanding financial market structures by testing different clustering methods.

**References**

1. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2006.

2. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2009.

3. K. C. A and A. James, "A Survey on Stock Market Prediction Techniques," 2023 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, India, 2023.

4. Ni, Y., Day, MY. & Huang, P. Trading stocks following sharp movements in the USDX, GBP/USD, and USD/CNY. Financ Innov 6, 35 (2020).

5. Mohan, Sumathy & .Sridhar, L.s & Ambrose, A. & Sudha, Nellore. (2021). Forex trading strategy : an empirical study on the currency pair GBP/USD,. 6. 20-23.