Name: Sainyukta Gujar (22060641026)

Sanika Hande (22060641035)

Topic: Application of Regression in Banking and Finance

**Multiple Linear Regression**

**Aim:** The aim of this study is to investigate the relationship between the USD exchange rates with various other currencies (USD/JPY, USD/GBP, USD/AUD, USD/CNY, USD/ZAR, USD/CAD, USD/MXN, USD/CHF, USD/IDR, USD/IQD, USD/IRR) and the exchange rate of USD to Indian rupee (USD/INR).

**Introduction:** The study will use a multiple linear regression model to examine how changes in the USD exchange rates with other currencies affect the USD/INR exchange rate. The model will be used to identify which exchange rates have a significant impact on the USD/INR exchange rate and to what extent. First, EDA is performed on the dataset to get the insights of the data. Further a Multiple Linear Regression model is fitted on the dataset taking USD/INR as the independent variable and USD/JPY, USD/GBP, USD/AUD, USD/CNY, USD/ZAR, USD/CAD, USD/MXN, USD/CHF, USD/IDR, USD/IQD, USD/IRR as the dependent variables. Then it is verified if the assumptions of Linear Regression model are satisfied.

**Data Characteristics:** The dataset contains 5140 observations. All the variables in the dataset are quantitative since the variables stand for exchange rates of USD with various currencies. In total there are 12 variables out of which 11 are the independent variables and one is the dependent variable.

**Exploratory Data Analysis:** Before starting with EDA, the necessary libraries are imported and data is loaded in the IDE. The *dataset.head ()* command is used to retrieve the first five rows from the dataset. Then *dataset.describe()* is used to obtain the summary of the dataset which includes the mean, median, standard deviation, etc. for all the variables in the dataset. Further, the presence of missing values is checked using *dataset.isnull ().sum ()* and concluded that no missing value is present in the dataset.

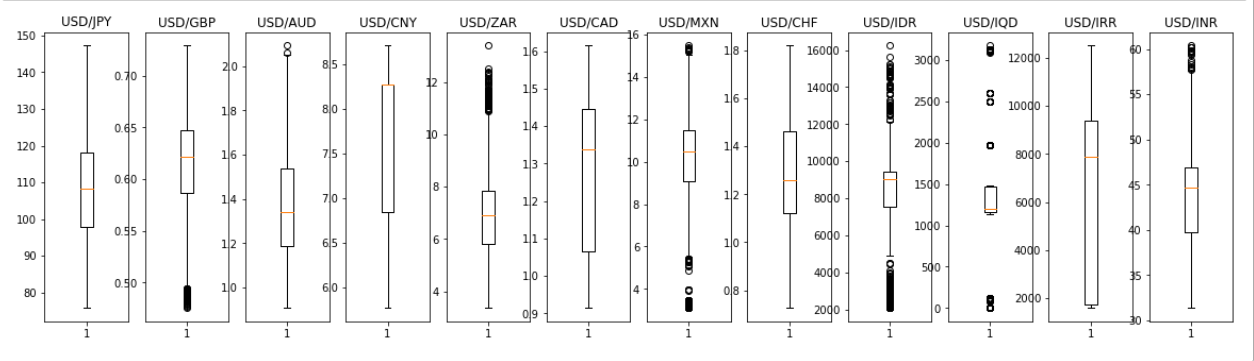


Fig. (1)

As shown in Fig. (1) above, the presence of outliers is detected using the box and whiskers plot and it can be concluded that the dataset is not much affected by the presence of outliers and it can be explained as the fluctuations in the exchange values due to volatilities in the market.

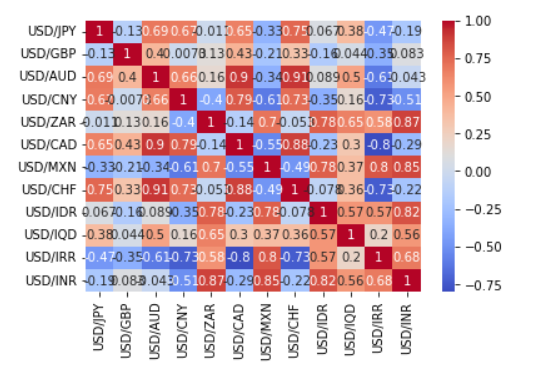


Fig. (2)

Lastly, from the Fig. (2) given above a correlation plot which indicates correlation between all the variables in the dataset is obtained where the colour red indicates high positive correlation between the variables and the colour blue indicates high negative correlation between the variables.

**Fitting the Model:** After performing EDA, the data is then split into training set and testing set in the ratio of 80:20. A Multiple Linear Regression model is then fitted on the training set. The efficiency of the model is then checked by comparing the calculated predicted values for the test set and the actual test set values of the dependent variables.

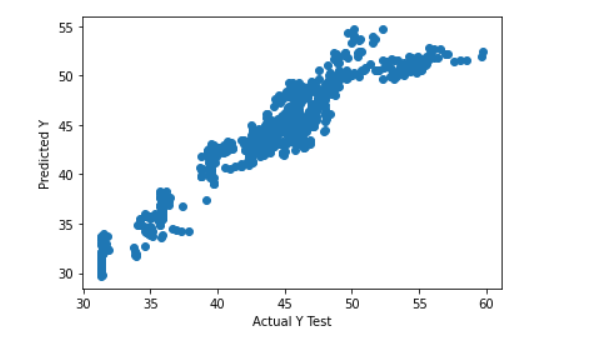


Fig. (3)

Fig. (3) shown above is the scatterplot of the Actual Y values and predicted Y values of the test set.

Finally, the values of R square, Mean Absolute Error, Mean Square Error and Root Mean Square Error are calculated to check if the model is the best fit for the dataset.

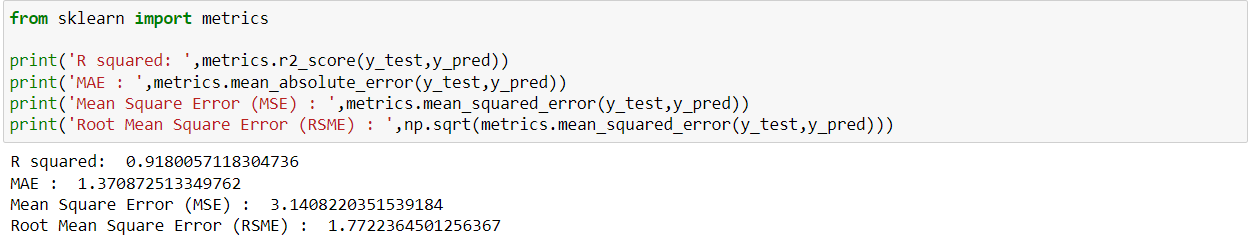


Fig. (4)

From Fig. (4) it can be concluded that the R-squared value is 0.918 which means that model explains 91.8% variation in the target variable. A higher R Square value indicates a better fit of the model on the data. The MAE value is 1.37 which explains that the difference between the predicted values and the actual values of the data is off by 1.37 units. The MSE value of 3.14 indicates that the model’s predictions deviate from the actual values by 3.14 units. From the RMSE value of 1.77 it can be concluded that on an average the model’s predictions deviate by 1.77 units.

**Checking Assumptions of Linear Regression:**

* Linearity: By plotting scatterplots for all independent variables against the dependent variable, it is concluded that non linearity is present in the dataset. We can get rid of non-linearity in the data by using transformations like log-transformation, square root- transformation on the independent variables.
* Independence: The Durbin-Watson test is used to check the assumption of independence in the dataset. Here the value is 0.016 indicating that autocorrelation is present in the dataset and the absence of independence. However other transformation methods like Cochrone-Orcutt or Hildreth-Lu can also be used to eliminate autocorrelation.
* Homoscedasticity: The assumption of Homoscedasticity is then addressed and the Breusch Pagan test is used to detect Homoscedasticity. Since the p-value is not greater than 0.05 we reject the null hypothesis and conclude that homoscedasticity is not present in the dataset. The WLS method is then used to get rid of heteroscedasticity after which the R square value changes to 0.961 from 0.91 implying the model to be a better fit.
* Normality: The Shapiro-Wilk test is used to detect normality of Residuals in the dataset. Since the p-value is less than 0.05 the assumption of normality of residuals is not satisfied. This can be possible because of the outliers present in the dataset.
* Multicollinearity: From the correlation matrix it can be observed that multicollinearity is present in the dataset. To get rid of multicollinearity a ridge regression model is fitted on the dataset and the value of MAE is obtained which reduces from 1.37 to 1.34 indicating a reduction in difference between predicted and actual values.

**Conclusion:** Although the values of R square, MAE, MSE and RMSE indicate that the model is the best fit for the data. It can be observed that the assumptions of Multiple Linear Regression are not satisfied and hence it is not always necessary that a model can be said to be the best fit for the data based on only the values of R square, MAE, MSE and RMSE. There is a need to check for the assumptions of Linear Regression and then fit the model to get the best results. As in this case the R square value increases to 0.961 and the MSE decreases to 1.34 implying the best fit model.

**Logistic Regression**

**Aim:** The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. In this project, a logistic regression model is built for deciding whether a campaign will be successful in getting a client to sign up for the term deposits.

**Introduction:** Banking and finance are two closely related fields that form the backbone of modern economies. Banks are financial institutions that offer a wide range of financial services, including accepting deposits, providing loans, facilitating payments, and managing investments. Finance, on the other hand, involves the management and investment of money, including financial analysis, risk management, and investment strategy development. The banking and finance sectors are essential for providing individuals and businesses with access to capital and financial services that enable them to invest, grow, and manage their financial resources effectively. Moreover, these sectors play a critical role in promoting economic development and stability by facilitating the flow of capital and investment across different sectors and industries**.**

Term deposits are a type of savings account offered by banks and financial institutions, which allow customers to earn a fixed interest rate on their deposits over a specified period. They are also known as fixed deposits or time deposits. Term deposits typically have a maturity date ranging from several months to several years, during which time the funds cannot be withdrawn without incurring a penalty. The interest rate on term deposits is typically higher than that of regular savings accounts, making them an attractive option for individuals and businesses looking to earn a guaranteed return on their savings. Term deposits offer a low-risk investment option, as they are insured by government deposit insurance programs in many countries, providing a level of security for depositors.

**Data Characteristics:** The data is about a Portuguese banking institution’s direct marketing campaign (phone calls). There are 17 variables and 45211 values.

* The input variables or the independent variables are as follows:
* age
* job: type of job ('admin, 'blue-collar', 'entrepreneur',' housemaid',' management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
* Marital: marital status ('divorced', 'married', 'single', 'unknown';)
* education: (“unknown”, secondary", "primary”, “tertiary")
* Default: has credit in default? (“yes”, “no")
* balance: average yearly balance
* Housing: has a housing loan? ("yes”, “no")
* Loan: has a personal loan? (“yes”, “no")
* contact: contact communication type ("unknown", "telephone", "cellular")
* day: last contact day of the month
* month: last contact month of the year ("Jan", "Feb", "Mar", ..., "Nov", "Dec")
* duration: last contact duration, in seconds
* campaign: number of campaigns for this client
* p-days: number of days that passed by after the client was last contacted from a previous campaign ( -1 means the client was not previously contacted)
* previous: number of contacts performed before this campaign and for this client
* p-outcome: outcome of the previous marketing campaign (‘failure',' non-existent', 'success')
* The output variables are as follows:
* y: has the client subscribed to a term deposit? ("yes", "no")

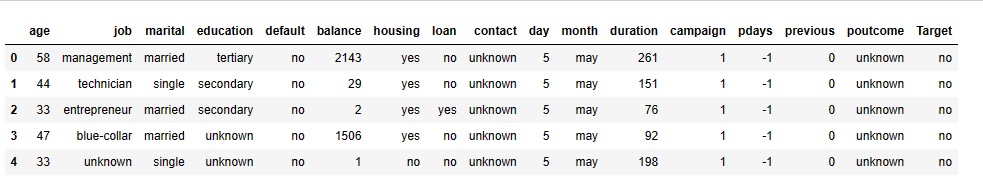


Fig. (4)

**Exploratory Data Analysis:**

1) Checking for Null- values



Fig. (5)

From Fig. (5), we can see that there are no missing values in the dataset.

2) Finding a Summary of the dataset:

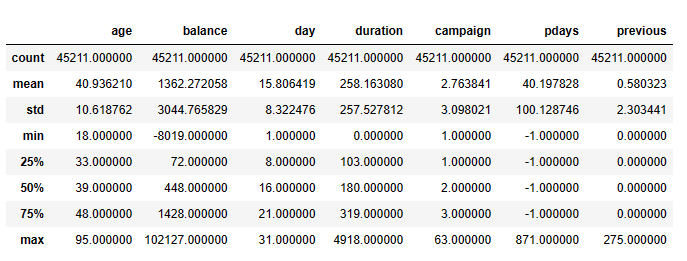


Fig. (6)

3) Checking for outliers

Here, we will check for the outliers of numeric values in the dataset

Age:

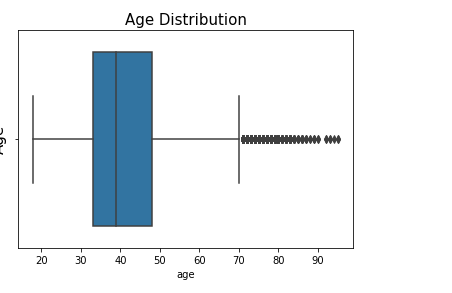


Fig. (7)

From Fig. (7), it can be seen that there are outliers present. And further analysis shows that there are 487 outliers and the age above 70.5 are outliers. Here, we will not remove the outliers as here we predict whether the customer will subscribe to the term deposit or not. So the outliers can be used to check whether people above specific age are interested or not so that we can decide on whether or not to include people of a specific age to include in the further campaign.

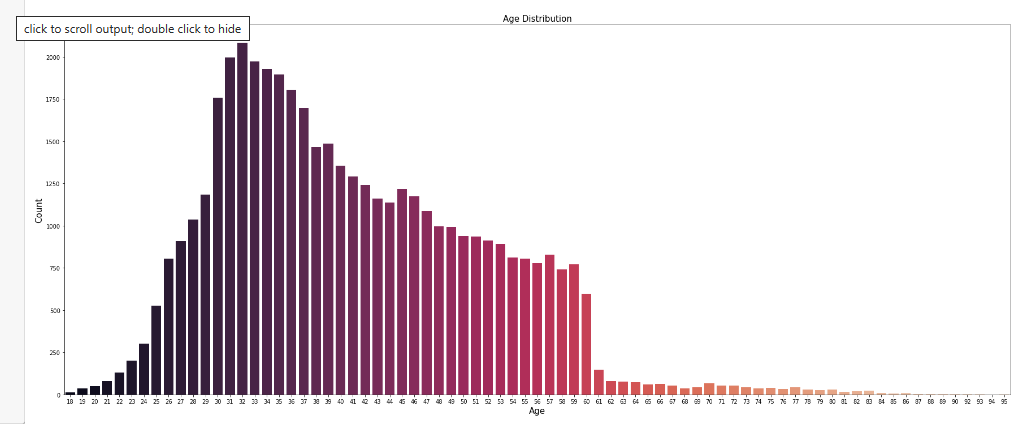


Fig. (8)

Fig. (8) shows the distribution of different age groups. We can see that most of the people lie between the age group 25-35.

Job:

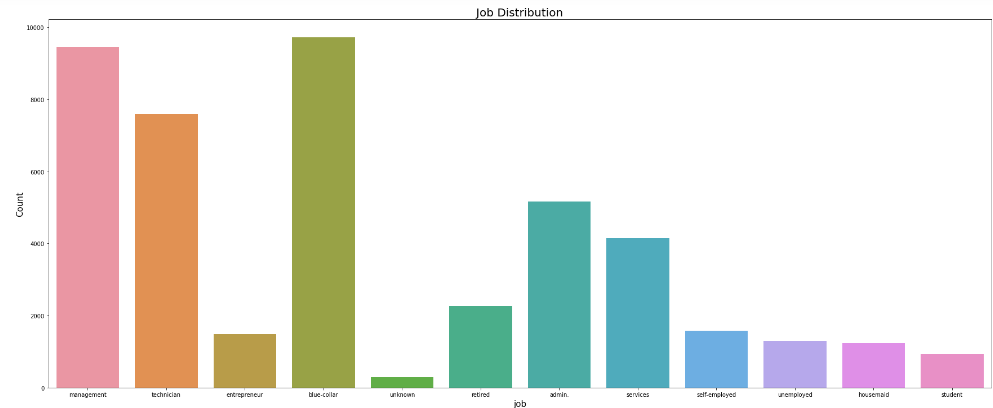


Fig. (9)

Fig. (9) shows the job distribution of different people we can see that most of the people are from management or blue-collar domain.

Marital:

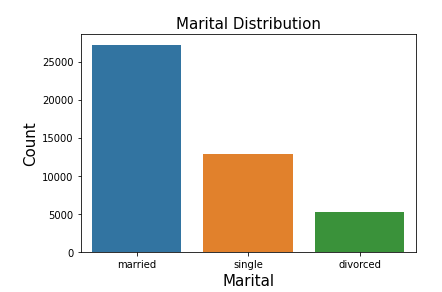


Fig. (10)

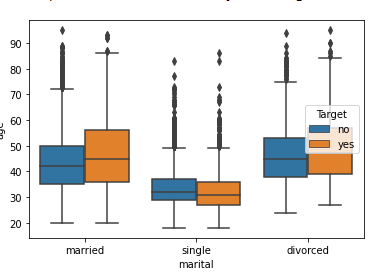


Fig. (11)

Most of the people in the given dataset are married. And the number of married, people subscribing to the term deposits is more than the number of single and divorced people.

Education:

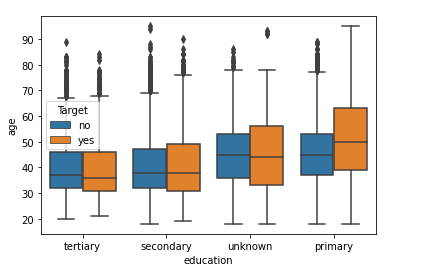


Fig. (12)

Fig. (12) show that number of people subscribing to the term deposit is from primary education.

Similarly, we can say that for default people who are not in default are more likely to subscribe to the term deposit. For the housing loan people who do not have housing loans are more likely to subscribe to the term deposit. Similar is the case for people having personal loans. For the contact type Cellular contact type is the most preferred contact type for contacting people. Most of the people are contacted in the month of May. For the contact duration phone call above 643 seconds are considered an outlier but we will not remove them as the long call duration may indicate that the customer is interested in subscribing to the term deposit. Also, most of the contact is done between the 8 to 21st days of the particular month. The campaign distribution shows the number of contacts performed during each campaign. Also, the p outcome shows that the outcome of the previous campaign does not have any specific effect on the people subscribing to this term deposit. The dependent variable is more distributed for the people who have not subscribed to the term deposit.

4) Correlation Matrix:



Fig. (13)

Since no variables have a correlation greater than 0.5 we can say that multicollinearity is absent. Hence, we can now fit the model.

5) Converting the categorical variables into numeric:

Here, the categorical variables are job, marital, education, default, housing loan, personal loan, contact type, and previous outcome we have encoded them into numeric types using the concept of dummy variables. Also, the output variable is whether the customer will subscribe to the term deposit or not (yes or no). It is also encoded into a numeric type where 1 represents that the customer will subscribe to the term deposit and 0 represents that the customer will not subscribe to the term deposit. This transformation is done using a label encoder.

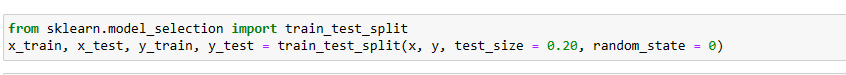
**Fitting the model:**  

Fig. (14)

Here we divide the data into a training set and a testing set in the ratio of 80:20. Where we will train our model on 80% dataset and using the trained model, we will test our dataset on the remaining 20% to check whether the actual values and predicted values are accurate or not.

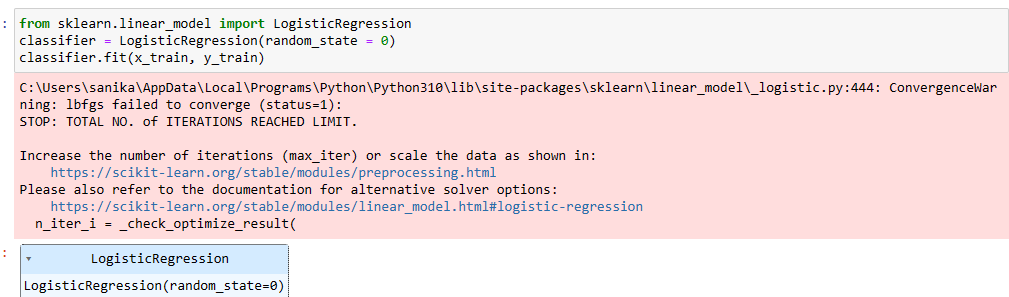


Fig. (15)

The logistic regression model is built here.

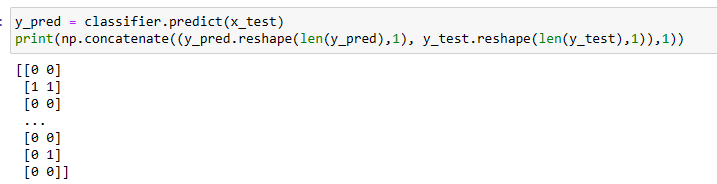


Fig. (16)

Using the above we have predicted whether the customer will subscribe to the term deposit or not based on the test set and we have merged the actual and predicted outcome to compare the performance of our model. Here, [0.0] states that we predicted that the customer will not subscribe to the term deposit and the actual value also says the same. Similarly [1,1] states that our customer will subscribe to the term deposit and the actual value also says the same. Whereas, [0,1] says that we predicted that the person will not subscribe to the term deposit but the actual values say that the customer has subscribed to the term deposit. Here, our model has made an error in prediction.

**Accuracy of the model:**

Using the confusion matrix and accuracy score we have checked the accuracy of our model.

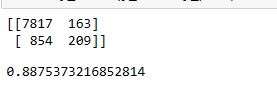


Fig. (17)

Here, the accuracy of our model is 88% which implies that the logistic regression model is a good fit for our data while the confusion matrix implies that our model predicted that 7817 people will subscribe to the term deposit, and in reality, also 7817 people subscribed. Whereas our model predicted that 163 people will subscribe to the term deposit but in reality they do not subscribe to the term deposit. Similarly, our model predicted that 854 people will not subscribe to the term deposit but in reality, they subscribed to the term deposit. The model also predicted that 209 people will not subscribe to the term deposit and in reality also they did not subscribe to the term deposit.

**Conclusion:**

The accuracy score of our logistic regression model is 88.75% indicating that the model makes correct predictions 88 % of the time. Hence our model is a good fit for our data. From our confusion matrix, we can see that 7817 values are True Positive,163 values are True Negative, 854 values are False positive, and 209 values are False Negative.

Other regression models like Decision Trees and Random Forests can also be fitted in order to compare which type of regression model gives the best accuracy Classification techniques like KNN and Naïve Bayes can also be used.