PRESTIGE INSTITUTE OF MANAGEMENT AND RESEARCH, INDORE

(An Autonomous Institution Established in 1994, Accredited Twice Consecutively with Grade A "NAAC'' (UGC) ISO 9001: 2008 Certified Institute, AICTE / UGC Approved Programs affiliated to DAVV, Indore)

(Session 2023 – 2023)

Machine Learning Project Report on

"Rupees to Dollar Exchange Rate Prediction"

Faculty Mentor:
Mr. Atul Astay

Submitted By:
Manvi Saxena
Pradduman Sugandhi
Pranjal Pandya
Prateek Somani
Rhytham Jain
Rishi Soni
Sakaar Mathur

Class:
BBA Business Analytics
VI Semester (A)

Contents

INTRODUCTION

This capstone project deals with the IND – USD exchange rate prediction. We have used machine learning, deep learning, and their applications to predict the exchange rate. Data was downloaded from yahoo finance and it comprised of past 10 years starting from March,2013 to March,2023.

Machine Learning - Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. By statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics.

HOW IT WORKS –

A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabeled, your algorithm will produce an estimate about a pattern in the data.

An Error Function: An error function evaluates the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

A Model Optimization Process: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this "evaluate and optimize" process, updating weights autonomously until a threshold of accuracy has been met.

METHODS –

Supervised Machine Learning - is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately.

Unsupervised Machine Learning - uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

Semi-supervised learning - offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm.

Deep Learning - Deep Learning is a subset of Machine Learning that is based on artificial neural networks (ANNs) with multiple layers, also known as deep neural networks (DNNs). These neural networks are inspired by the structure and function of the human brain, and they are designed to learn from large amounts of data in an unsupervised or semi-supervised manner. Deep Learning models can automatically learn features from the data, which makes them well-suited for tasks such as image recognition, speech recognition, and natural language processing.

Variable: A variable is a quantity or characteristic of an individual in a study.

There are 7 variables in the dataset – Date, Adj Close, Close, High, Low, Open, Volume. There are 2606 entries.

Independent Variables: The variable that are not affected by the other variables are called independent variables. Here ‘Open’, ‘High’, ‘Low’ are the independent variables.

Dependent Variables: The variables which depend on other variables or factors. We expect these variables to change when the independent variables, upon whom they depend, undergo a change. ‘Close’ is the dependent variable.

METHADOLOGY

Tools used:

1. **Jupyter Notebook**

Jupyter Notebook is a web application that provides creation and modification of live codes, visualizations, equations and plaintexts. This opensource notebook supports multiple programming languages. It uses for numerical simulation, information visualization, machine learning, statistical modeling, and many more functionalities. We used this notebook for writing readable codes for developing models and visualizing the results and accuracy errors.

2. **Pandas**

This is although an opensource library which provides data structures and data analysis tools. The important note about pandas is its high performance and easy to use especially for manipulating operations in numerical tables and time series data. Though pandas used to store the currency data in dataframe where it then divided in X and Y dimensions and made it ready for scaling and other preprocessing operations.

3. **Numpy**

This open-source library is doing the computing, it supports multi-dimensional arrays and matrices with so many functions which make working easy with arrays and matrices. Since arrays can increase the speed and efficiency in data analysis tasks this library can help a data scientist a lot. In the forecasting of these data, Numpy is used to store data in arrays. Since the prediction models require Numpy arrays as their parameters to operate fast and decrease the training and prediction time.

4. **Matplotlib**

Matplotlib is a plotting library that is used to create variety of graphs for different purposes. The salience of matplotlib is its easiness in use. A good quality plot can be produced with few lines of code. So matplotlib is used to plot the output historically and also to show the accuracy and difference between the prediction and real output.

5. **Yfinance**

Yfinance is a popular opensource library developed by Ran Aroussi as a means to access the financial data available on Yahoo Finance. Yahoo Finance offers an excellent range of market data on stocks, bonds, currencies and cryptocurrencies. It also offers market news, reports and analysis and additionally options and fundamentals data- setting it apart from some of it's competitors.

6. **Seaborn**

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

7. **StandardScaler**

Python sklearn library offers us with StandardScaler() function to standardize the data values into a standard format. we initially create an object of the StandardScaler() function. Further, we use fit_transform() along with the assigned object to transform the data and standardize it.

8. **Dataset**

The dataset is downloaded from yahoo finance of last 10 years (21 March,2013 –20 March,2023) with 2606 entries. The data has following attributes:

Adjusted Close

Close

High

Low

Open

Volume

The ‘Open’, ’High’, ’Low’ attributes used as inputs of the training algorithms and the attribute which is ‘close’ is used for output and forecasting each day close price.

Why we used Linear Regression:

Linear Regression - Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

While training the model we are given: **x:** input training data (univariate – one input variable(parameter)) **y:** labels to data (Supervised learning) When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values. **θ1:** intercept **θ2:** coefficient of x Once we find the best θ1 and θ2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

y = Dependent variable

x = Independent variable

a0 = Intercept of line

a1 = Linear regression coefficient

Linear regression is important because it provides a scientific calculation for identifying and predicting future outcomes. The ability to find predictions and evaluate them can help provide benefits to many businesses and individuals, like optimized operations and detailed research materials. When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by

adding too many variables to the model, which reduces model generalizability. Occam's razor describes the problem extremely well – a simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

INTERCEPT IN REGRESSION

The **intercept** (sometimes called the "constant") in a regression model represents the mean value of the response variable when all of the predictor variables in the model are equal to zero.

The slope indicates the steepness of a line and the intercept indicates the location where it intersects an axis. The slope and the intercept define the linear relationship between two variables, and can be used to estimate an average rate of change. The greater the magnitude of the slope, the steeper the line and the greater the rate of change.

MODEL CONSTRUCTION

DATA PREPROCESSING

For training an algorithm in machine learning we need to clean and prepare the data to be fit and can be used for training. This preprocessing has many steps which some of them are different from one approach to another.

Data is split into two parts, training and testing. Training portion is 80% and the testing portion is 20% which is a common practice. It starts from 60% training and 40% testing and goes to a much greater training portion. For this purpose, the 'train_test_split' object of Scikit-Learn is used to divide data based on K-Fold Cross validation technique.

TRAINING, TESTING AND VALIDATION

While working on algorithms to monitor the model accuracy the data should separate into training, testing and validations sets. The training set is used for the learning process which the algorithm learn the signals of the training examples, the correlation between inputs, their weights and the output. As described earlier the dataset should be divided into two parts in cross-validation, though

the testing portion is kept aside for testing the accuracy of the model after it is trained. Though in this study based on most common practices the data is divided into 80% for training and 20% for testing. The first training is conducted in the training part, the models are trained based on the examples in this portion.

NORMALISING THE DATA

Data normalization is the process of reorganizing data within a database so that users can utilize it for further queries and analysis. Simply put, it is the process of developing clean data. This includes eliminating redundant and unstructured data and making the data appear similar across all records and fields. For the purpose of data normalization, several scaling techniques can be used, though the most fit technique or function for this project is 'StandardScaler'.

The StandardScaler function of sklearn is based on the theory that the dataset's variables whose values lie in different ranges do not have an equal contribution to the model's fit parameters and training function and may even lead to bias in the predictions made with that model. Therefore, before including the features in the machine learning model, we must normalize the data through standard scaling. The Standard Scaler works by eliminating the mean from the features and scaling them to unit variance.

LINEAR REGRESSION

Linear regression is a technique used to model the relationships between observed variables. The idea behind simple linear regression is to "fit" the observations of two variables into a linear relationship between them. Graphically, the task is to draw the line that is "best-fitting" or "closest" to the points \((x_i,y_i),\) where \(x_i\) and \(y_i\) are observations of the two variables which are expected to depend linearly on each other.

Linear Regression is generally classified into two types:

1. Simple Linear Regression
2. Multiple Linear Regression

In Simple Linear Regression, we try to find the relationship between a single independent variable (input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line.

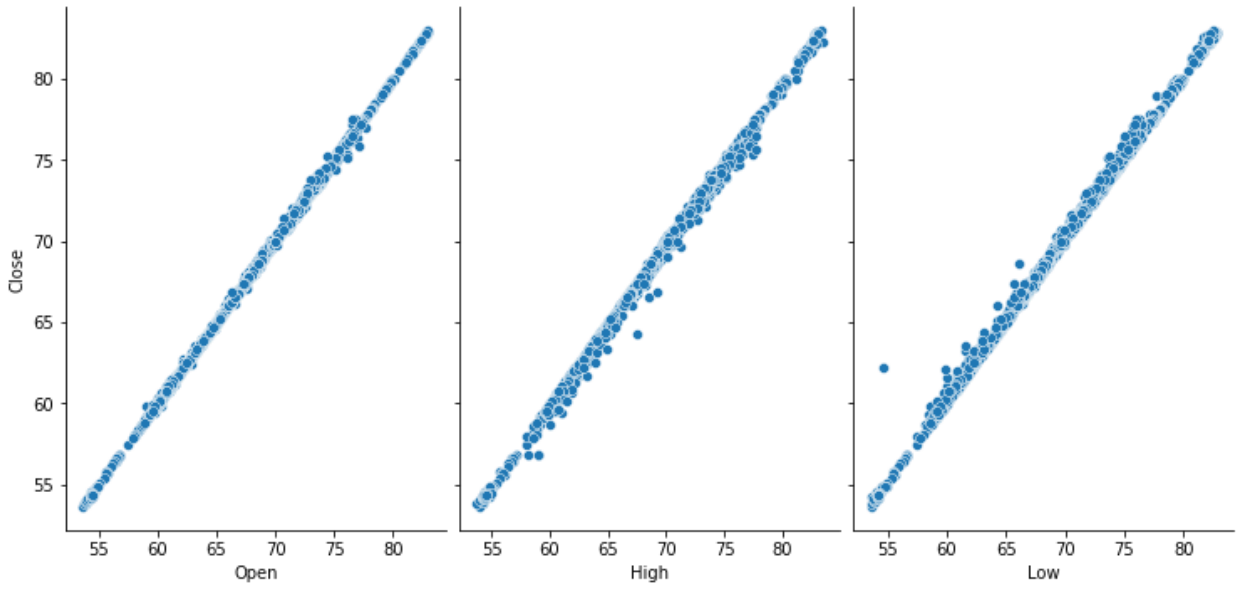
The same equation of a line can be re-written as:

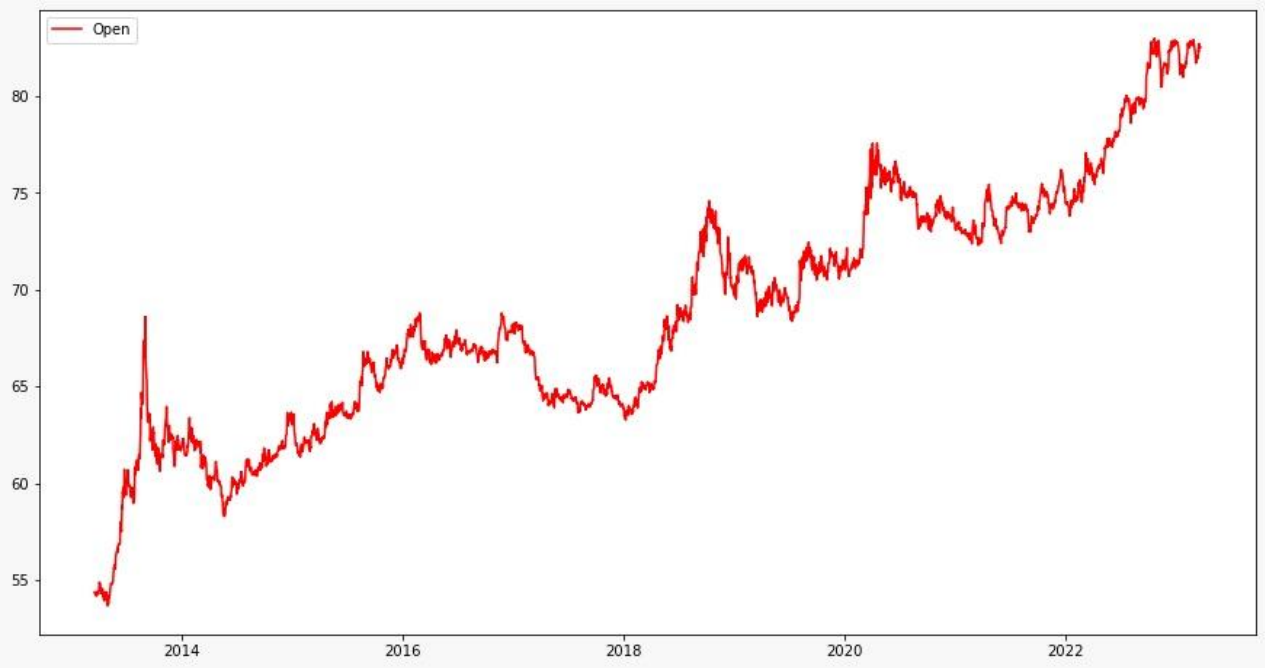
$$Y = \beta_0 + \beta_1 X + \epsilon$$

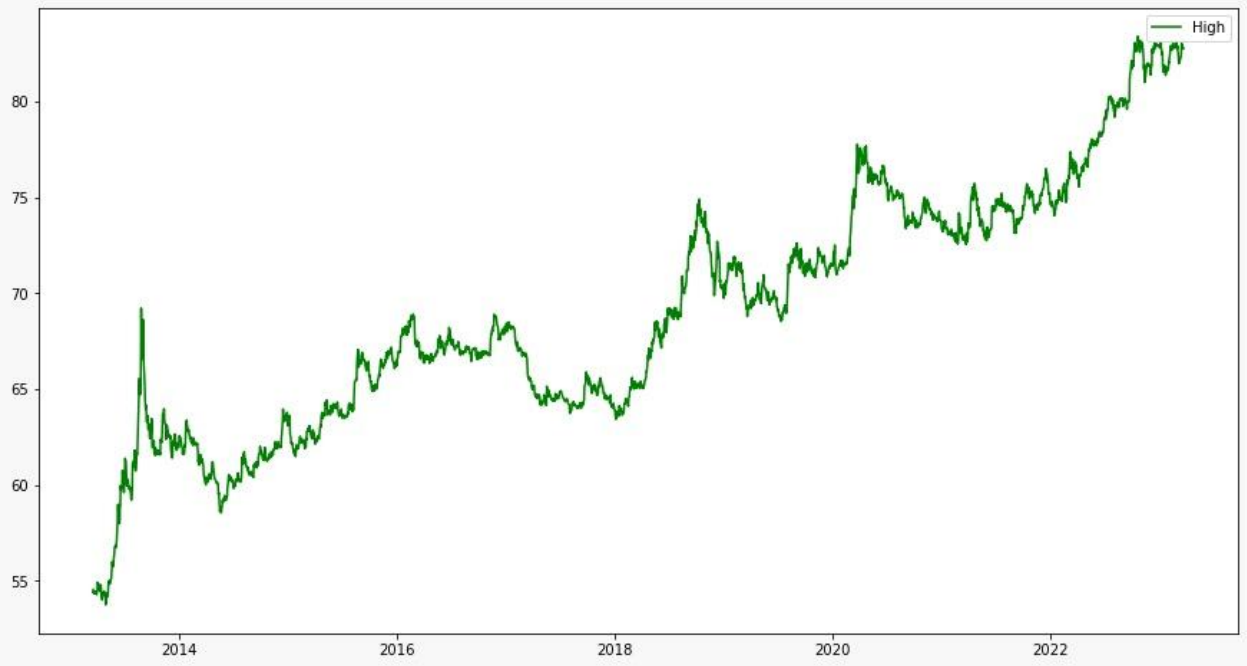
In Multiple Linear Regression, we try to find the relationship between **2 or more independent variables (inputs)** and the corresponding dependent variable (output). The independent variables can be continuous or categorical.

VISUALIZATION

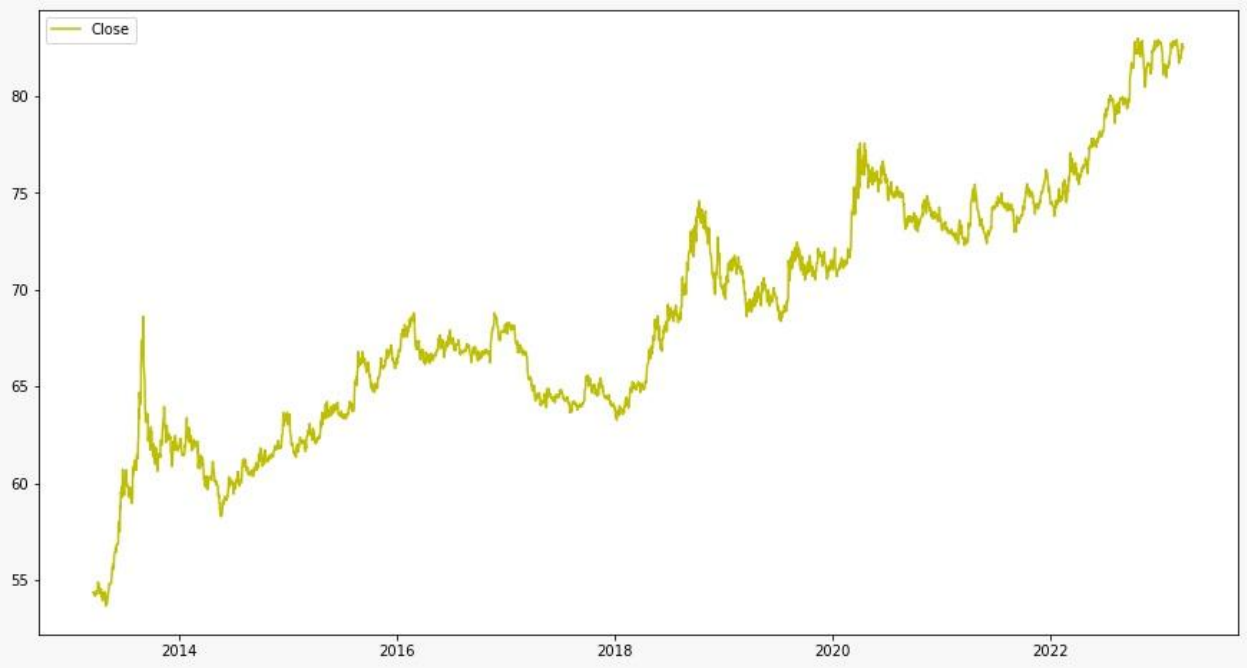
With a simple definition, visualization is the most convenient way of communicating a message. This has been heard too much that a picture worth thousands of words. Since 40 visualization is also creating images, graphs, and visionary shapes, it makes it easy to make complex topics easy for understanding. Therefore, in this study, visualization is used for clarification of the learning process and the accuracy improvement in training phase. In addition, both the actual values which is the real output is comparatively visualized with predicted output to show the difference in very precise and clear sight. Matplotlib from python libraries is used for doing this helpful task.

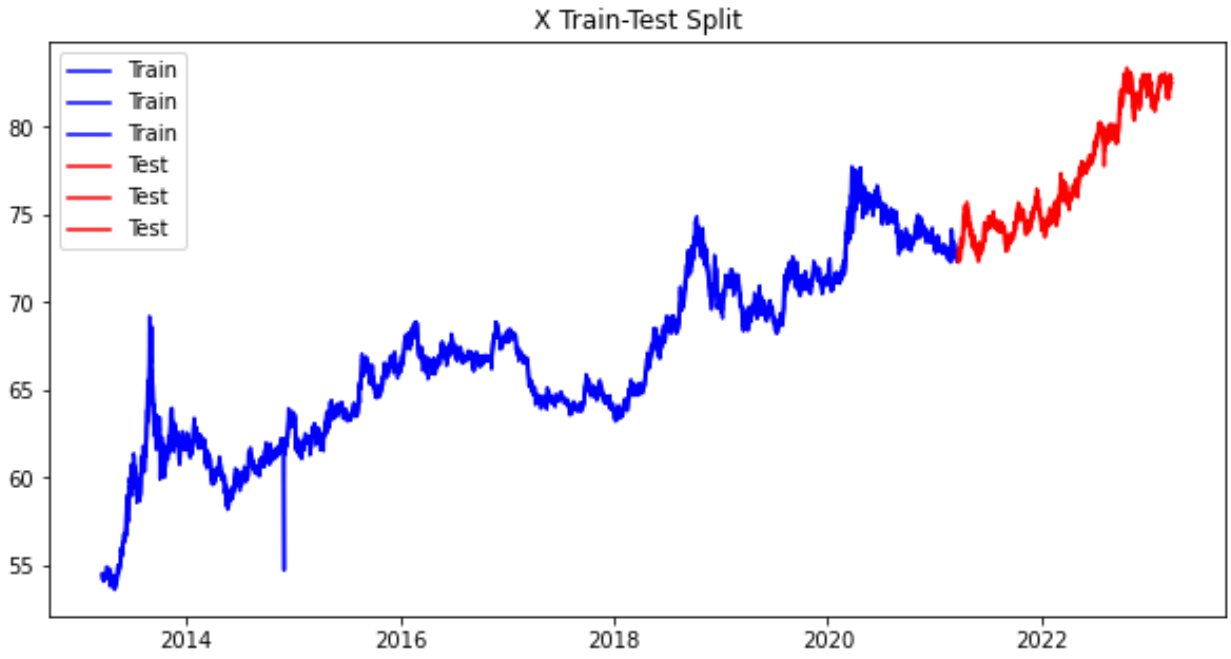


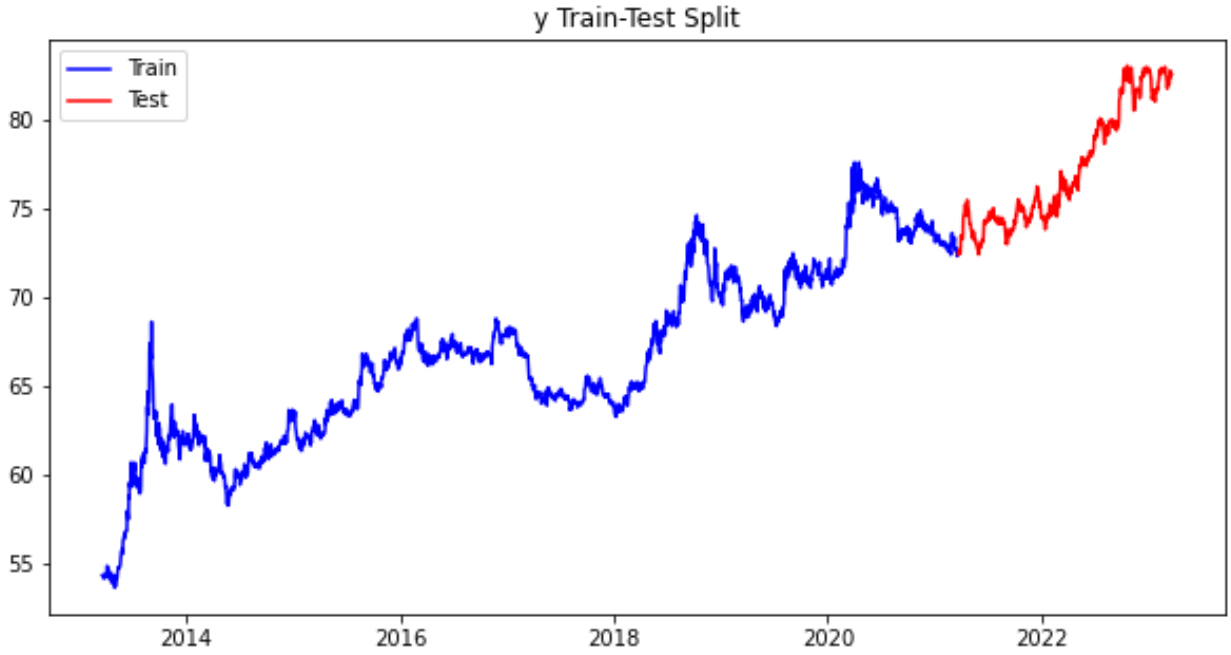

Open
80
75
70
65
60
55
2014
2016
2018
2020
2022

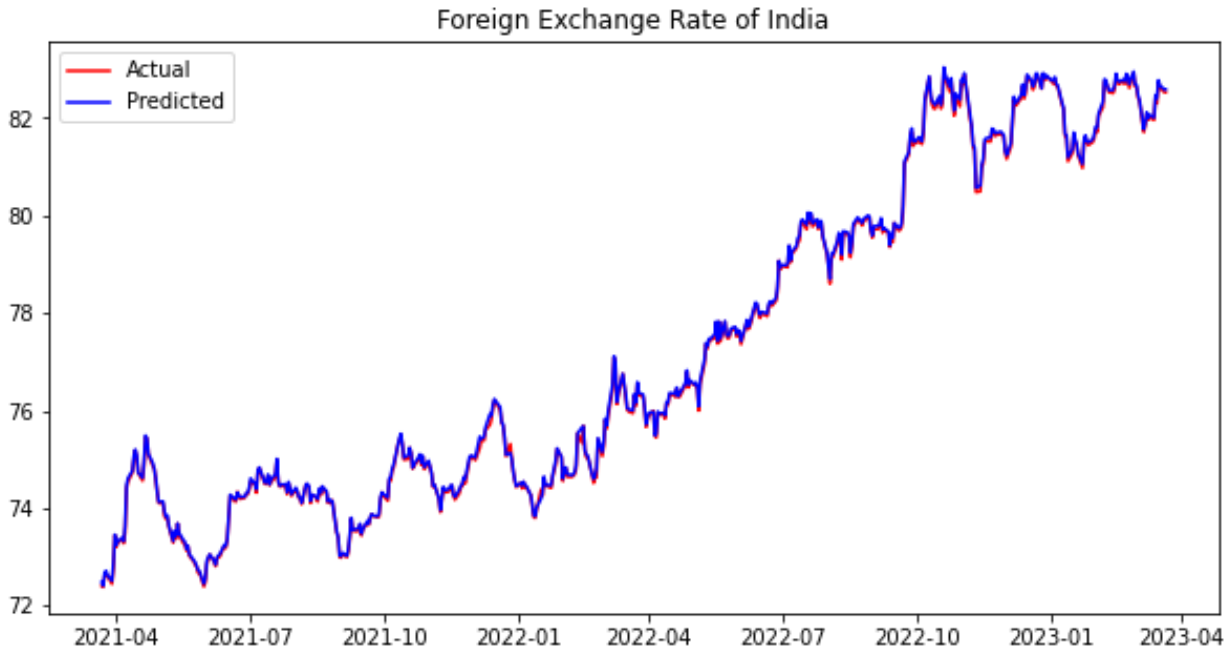

High
80
75
70
65
60
55
2014
2016
2018
2020
2022


Low
80
75
70
65
60
55
2014
2016
2018
2020
2022


Close
80
75
70
65
60
55
2014
2016
2018
2020
2022


X Train-Test Split
Train
Train
Train
Test
Test
Test
80
75
70
65
60
55
2014
2016
2018
2020
2022


y Train-Test Split
Train
Test
80
75
70
65
60
55
2014
2016
2018
2020
2022



CONCLUSION AND FINDINGS

Currency exchange rate forecasting became a challenge for human beings. Almost everybody's economy directly or indirectly connected to foreign currency prices. Therefore, currency exchange is vital for humans in this era. Machine learning and generally artificial intelligence technologies going to help humans predict currency rates. The model's performance were great due to rich data on the training phase. But the fact that different factors still remains which the currency rate has many aspects thus forecasting only time series and historical data is good but not enough.

In conclusion, Python's linear regression program has proven to be a helpful and successful tool for analysing and forecasting exchange rates when used to predict the INR exchange rate. We can create models that correctly predict exchange rate trends and movements by using historical data and statistical methods. It's crucial to keep in mind, though, that exchange rates are frequently

unpredictable due to a variety of external variables, including political developments, general economic conditions, and investor sentiment.

Although the INR exchange rates' possible future direction has been predicted by the linear regression model used in this analysis, its accuracy must be constantly monitored and improved as new data become available. Moreover, it's important to take into account other factors such as inflation rates, interest rates, and government policies when making financial decisions based on exchange rate forecasts.

Overall, the use of linear regression on Python for INR exchange rate prediction is a valuable tool for businesses, investors, and individuals who seek to understand and make informed decisions about the foreign exchange market.

The accuracy of Linear Regression model has come out to be 99.970%, which is almost perfect to the actual values. It means the actual and predicted values are 99% accurate. We used the forex data to predict the value of ‘close’ and it was found that the actual value was 82.7156982421875 and the predicted value was 82.90031626.