Machine Learning Model for Predicting Exchange Rate - Sri Lanka

Punsisi Kosgahakumbura

May 2023

Table of Contents

1. Introduction to the Dataset

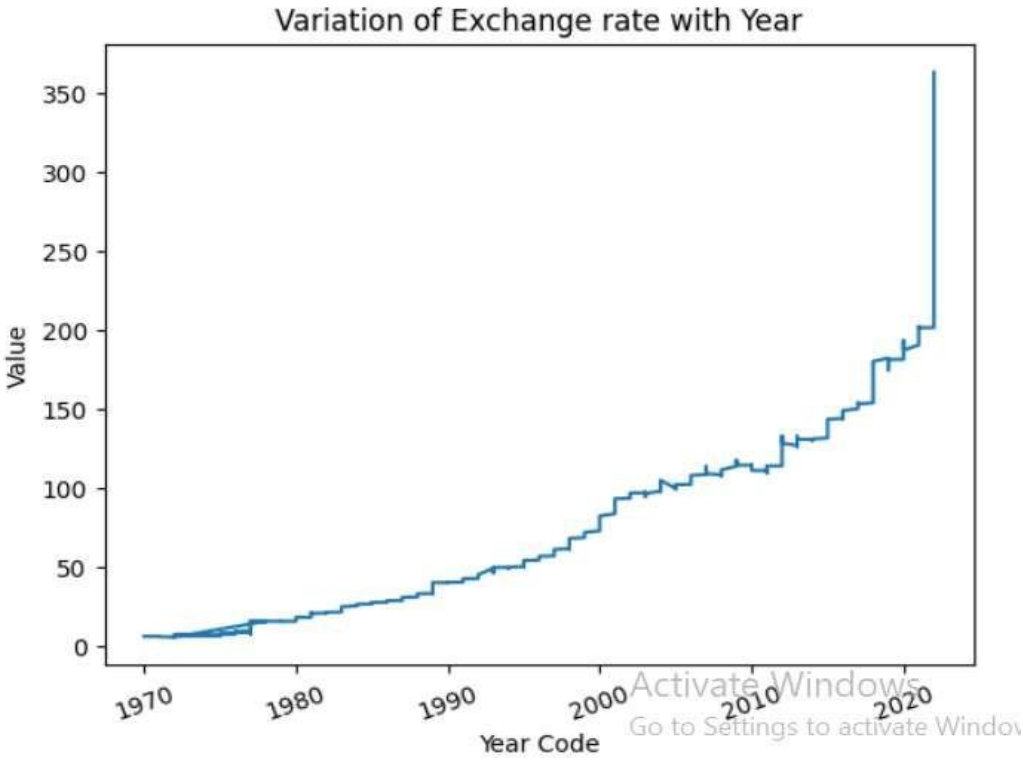
The "exchange-rates_lka.csv" dataset contains data about annual exchange rates, Standard Local Currency Units (SLC) per US dollar. Annual and monthly exchange rates, Local currency Units (LCU) per US dollar. It consists of 813 rows and 17 columns with the predictor variable 'Value' and 16 attributes given below.

#	Column	Non-Null Count	Dtype
0	Iso3	814 non-null	object
1	StartDate	814 non-null	object
2	EndDate	814 non-null	object
3	Area Code	813 non-null	float64
4	Area Code (M49)	813 non-null	object
5	Area	814 non-null	object
6	Element Code	813 non-null	object
7	Element	813 non-null	object
8	ISO Currency Code	813 non-null	object
9	Currency	813 non-null	object
10	Year Code	813 non-null	float64
11	Year	814 non-null	object
12	Months Code	813 non-null	float64
13	Months	813 non-null	object
14	Unit	1 non-null	object
15	Value	814 non-null	object
16	Flag	813 non-null	object

These variables play crucial roles in understanding and analyzing exchange rate data, enabling researchers to study trends, patterns, and fluctuations in exchange rates for different regions and time periods.

- The Value variable holds the recorded value of the exchange rate. It represents the numerical value of the exchange rate between the specified currency pairs.
- The ISO3 variable serves as a country code, uniquely identifying each country or territory. It enables the differentiation of exchange rate data based on specific regions.
- The StartDate and EndDate variables indicate the time range for the recorded exchange rate data. They allow for the identification of the duration over which exchange rates were measured.
- The Area variable provides the name of the area or country for which the exchange rate data is recorded. It helps in distinguishing exchange rate information based on specific regions.
- The Element variable describes the category or element of the data, which, in this case, pertains to exchange rates. It provides valuable context to understand the nature of the recorded data.
- The ISO Currency Code and Currency variables represent the currency code and name, respectively, used for the exchange rates. These variables allow for the identification of the currency pairs involved in the exchange rate calculations.
- The Year and Months variables specify the year and month for which the exchange rate data is recorded. They enable the temporal analysis of exchange rate fluctuations over different time periods.

The variation of Exchange Rate in Sri Lanka is given by the following graph.



2. Methodology

A. Data Pre-processing

- Handling missing values

To understand the presence of null values in the dataset, the command "data.isnull().sum()" was executed. The resulting output indicated the number of null values for each variable in the dataset.

```
Iso3                  0
StartDate             0
EndDate               0
Area Code             1
Area Code (M49)       1
Area                  0
Element Code          1
Element               1
ISO Currency Code     1
Currency              1
Year Code             1
Year                  0
Months Code           1
Months                1
Unit                813
Value                 0
Flag                  1
dtype: int64
```

Upon inspecting the dataset, it was observed that the 0th index row indicates the presence of missing values in each column, including Area Code, Element Code, Element, ISO Currency Code, Year, Year Code, Months Code, Months, and Flag. Hence, it is recommended to drop the 0th row from the dataset as it does not contain any useful information. Considering the Unit variable having 813 missing values out of 814 total entries, it would be appropriate to drop the entire column from the dataset. Indeed, based on the information provided, it appears that there are no missing values to handle in the "Exchange Rates" dataset for Sri Lanka. This indicates that the dataset is complete and does not require any further preprocessing steps related to missing data.

- Data Transformation

- The "Value" variable represents the exchange rate, it should ideally have a numeric data type instead of being classified as an object. It would be appropriate to ensure that the "Value" variable is properly converted to a numeric data type for accurate analysis and modeling of exchange rates.
- To convert the "Element Code" variable to numeric representation, the values 'LCU' and 'SLC' can be replaced with 0 and 1, respectively. This transformation ensures that the "Element Code" variable is accurately represented as a numeric data type.
- To convert the "ISO Currency Code" variable to numeric representation, the values 'LNR' and 'LKR' can be replaced with 0 and 1, respectively. This conversion ensures that the "ISO Currency Code" variable is accurately represented as a numeric data type.

- Data Reduction

In order to consider only monthly exchange rates, the dataset underwent a filtering process where all the rows with annual exchange rates were removed. This step ensured that only the rows representing monthly exchange rates remained in the dataset for further analysis.

B. Feature Engineering

	Iso3	StartDate	EndDate	Area Code (M49)	Area	Element Code	Element	ISO Currency Code	Currency	Year	Months	Unit	Value	Flag
count	813	813	813	813	813	813	813	813	813	813	813	0	813	813
unique	1	634	635	1	1	2	2	2	2	53	13	0	682	1
top	LKA	1973-01-01	1976-12-31	'144	Sri Lanka	LCU	Local currency units per USD	LKR	Sri Lanka Rupee	1972	Annual value	NaN	5.952370005000000	X
freq	813	4	4	813	813	760	760	714	714	26	106	NaN	25	813

- After conducting feature analysis on the dataset, it was observed that the features "ISO3," "Area Code (M49)," "Area," and "Flag" contained only unique values and did not exhibit any relevant patterns or frequencies. Hence, these features were deemed irrelevant for predicting the exchange rate and were subsequently removed from the dataset.

- It was determined that the information about the "Element," "Currency," "Year," and "Months" was already represented by the corresponding features such as "Element Code," "ISO Currency Code," "Year Code," and "Months Code." Therefore, the features "Element," "Currency," "Year," and "Months" were dropped from the dataset. This decision was made to avoid redundancy and reduce the dimensionality of the data while still retaining the necessary information for predicting exchange rate.

- The descriptive statistics of the remaining features and “Value”, as given below.

	Area Code	Element Code	ISO Currency Code	Year Code	Months Code	Value
count	707.0	707.0	707.000000	707.000000	707.000000	707.000000
mean	38.0	0.0	0.862801	1993.711457	7006.478076	66.927268
std	0.0	0.0	0.344301	15.859919	3.452718	62.480269
min	38.0	0.0	0.000000	1970.000000	7001.000000	5.482468
25%	38.0	0.0	1.000000	1978.000000	7003.000000	15.570500
50%	38.0	0.0	1.000000	1993.000000	7006.000000	48.071800
75%	38.0	0.0	1.000000	2008.000000	7009.000000	109.209400
max	38.0	0.0	1.000000	2022.000000	7012.000000	363.148421

- To understand the relationship with the predictor variable and other features the correlation analysis was conducted as follows.

```
Area Code                NaN
Element Code             NaN
ISO Currency Code    0.382819
Year Code            0.929674
Months Code          0.012037
Value                1.000000
Name: Value, dtype: float64
```

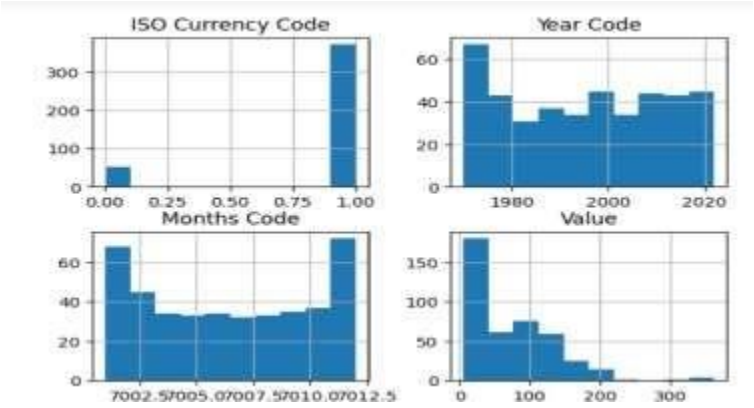
  - The features Area Code and Element Code have NaN correlation coefficients. This suggests that these features do not exhibit any variations or do not show any numeric relationship with the target variable.
  - ISO Currency Code has 0.382819 correlation coefficient, indicating a weak positive relation with the predictor variable ‘Value’.
  - Year Code has a correlation coefficient of 0.929674, indicating a strong positive correlation with the "Value" variable. This suggests that the "Year Code" is likely to be a highly influential predictor of the "Value" variable.
  - Months Code has a correlation coefficient of 0.012037, suggesting a very weak positive correlation with the "Value" variable.

Since Area Code and Element Code do not provide any significant relation to the prediction, they were dropped from the data set.

After conducting the above steps in data preprocessing and feature engineering, it was determined that the most relevant features for predicting exchange rates are "ISO Currency Code," "Year Code," and "Months Code." These features have been identified as significant predictors based on the analysis performed during the preprocessing and feature engineering stage.

C. Model Building

- Split the data.
  - 60% (423) of data is assigned to the training set, 20% (123) to validation set and 20% (123) to test set. The histogram for each numeric variable in the "train_data" dataset as given in below.



  - The correlation of the train data indicates the relation of all the features in the training dataset.

	ISO Currency Code	Year Code	Months Code	Value
ISO Currency Code	1.000000	0.507147	-0.008908	0.379459
Year Code	0.507147	1.000000	-0.001026	0.924721
Months Code	-0.008908	-0.001026	1.000000	0.016225
Value	0.379459	0.924721	0.016225	1.000000

- Model selection and Training the model

As it is a regression problem, the models chosen for training the data were LinearRegression(), RandomForestRegressor(), and DecisionTreeRegressor(). These models were trained using 423 data instances from the dataset.

- Validate the model

All three models were validated using means squared error and the following are the results.

Linear Regression Validation MSE: 173.16581731658724
Random Forest Regressor Validation MSE: 8.62042181995342
Decision Tree Regressor Validation MSE: 1.0109715634937573

The Linear Regression model has the highest validation MSE value of 173.1658, indicating larger prediction errors and a larger average squared difference from the actual values. This suggests that the Linear Regression model may not be the best fit for this particular regression problem. The Random Forest Regressor model performs better with a lower validation MSE value of 8.6204. It demonstrates smaller prediction errors and a smaller average squared difference from the actual values, suggesting better predictive capability compared to the Linear Regression model. However, the Decision Tree Regressor model outperforms both the Linear Regression and Random Forest Regressor models with the lowest validation MSE of 1.0109. It exhibits the smallest prediction errors and the smallest average squared difference from the actual values, indicating the best performance among the three models for this regression task.

- Model Evaluation

Since Linear Regression model was not a good, fitted model to the used data set, only the Random Forest Regressor and the Decision Tree Regressor were used to evaluate the model using 142 test data set. The mean squared error and the model accuracy were considered as the parameters for model evaluation.

- Mean squared error of the models.
  Random Forest Regressor Test MSE: 1.6892269180950443
  Decision Tree Regressor Test MSE: 0.6165764329212695

- Model Accuracy
  RandomForestRegressor()>>>0.9996363109403558
  DecisionTreeRegressor()>>>0.9998391068519548

The above results also indicate that the Decision Tree regressor has the best performance and the lowest prediction errors compared to Random Forest Regressor model.

3. Results and Conclusion

A. Discussion of the significance of selected features

ISO Currency Code is significant as it provides the specific currency associated with the exchange rate of Sri Lanka which provides the insights about the currency specific factors that impact the exchange rate. The Year Code represents the specific year in which the prediction occurs. It indicated the annual trends and patterns of Sri Lanka's exchange rate. Month code also plays a crucial role as it indicates the monthly seasoning and economic variations.

By considering the ISO Currency Code, Year Code, and Month Code, the prediction model can take into account currency-specific dynamics, annual trends, and monthly variations, thus providing a more accurate and tailored prediction of Sri Lanka's exchange rate.

B. Selecting the best model for the prediction.

After conducting the validation and evaluation it was found that the Decision Tree Regressor performed the best at both the validation and evaluation stages by having lowest mean squared error and the highest accuracy level. Based on these results, the Decision Tree Regressor model was selected for further prediction of the Exchange Rate. It is expected to provide more accurate and reliable predictions, given its superior performance during the validation and evaluation stages.

C. Predicting values for user inputs

By using the Decision Tree Regressor model, the exchange rate for the year 2023 is predicted by using the following inputs.

Enter ISO Currency Code: 0
Enter Year Code: 2023
Enter Months Code: 7001
Predicted Value: 201.7362

The exchange rate for the month of January in the year 2023 with the currency code of LNR is equals to 201.7362.

Enter ISO Currency Code: 1
Enter Year Code: 2023
Enter Months Code: 7006
Predicted Value: 358.939031578947

The exchange rate for the month of June in the year 2023 with the currency code of LKR is equals to 358.939031578947.

D. Overall performance and effectiveness of the machine learning model

The overall performance and the effectiveness of the machine learning model for predicting Exchange Rate of Sri Lanka cab be evaluated using several factors and, in our case we selected model's accuracy and mean squared error (MSE) to evaluate.

Model's performance was validated and evaluated using MSE, which measures the average squared difference between the predicted exchange rate values and the actual values. A lower MSE value indicates better performance with low prediction error. The Decision Tree Regressor model had the lowest MSE value at both validation stage and evaluating stage. Furthermore, the model's accuracy was also assessed, which provides a measure of how well the model's predictions align with the actual exchange rate values. The Decision Tree Regressor model demonstrated a high accuracy level, further supporting its effectiveness in capturing the patterns and trends in the exchange rate data. Overall, based on the model's performance, including its low MSE and high accuracy, the Decision Tree Regressor model proves to be effective in predicting the exchange rate.

4. References

[1] A. Kharwal, "Currency Exchange Rate Prediction with Machine Learning | Aman Kharwal," *thecleverprogrammer*, May 22, 2021. https://thecleverprogrammer.com/2021/05/22/currency-exchange-rate-prediction-with-machine-learning/ (accessed Jun. 24, 2023).

[2] "An Easy Guide to Stock Price Prediction Using Machine Learning," *Simplilearn.com*. https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning

[3] L. Nielsen, "MACHINE LEARNING FOR FOREIGN EXCHANGE RATE FORECASTING." Available: https://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Nielsen_Laurids-Gert_01424460.pdf

[4] A. Bajaj, "Performance Metrics in Machine Learning [Complete Guide]," *neptune.ai*, May 02, 2021. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide#:~:text=Performance%20metrics%20are%20a%20part

[5] scikit learn, "sklearn.tree.DecisionTreeClassifier — scikit-learn 0.22.1 documentation," *Scikit-learn.org*, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[6] "sklearn.metrics.mean_squared_error — scikit-learn 0.24.2 documentation," *scikit-learn.org*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html