Currency Market Insights: Machine Learning for JPY-USD Exchange Rate Prediction

Brown University

Yu Gu

Github: https://github.com/KelvinG991/JPY-USD-prediction-project

Introduction

Over the course of history, exchange rates have played a pivotal role in international finance, serving as crucial indicators for countries to adjust monetary policies and foster economic stability. Similarly, for individuals, a familiarity with exchange rates is essential for risk management and potential losses mitigation. Unlike the stable eras of the gold standard or the Bretton Woods agreement, contemporary exchange rates exhibit constant fluctuations, presenting a significant challenge in predicting future rates. This project aims to leverage historical data to forecast exchange rates between two of the most heavily traded currencies, the USD and the JPY. There are some previous works on this topic on Kaggle, but they explain exchange rates only through past exchange rates. This project is enhanced by incorporating formal economic models for a comprehensive understanding.

In all of the machine learning models in my project, the following features are included:

- Ex: exchange rate between JPY and USD at this time period (unit: JPY/USD)
- R1: the interest rate of the US at this time period (unit: %)
- R2: the interest rate of Japan at this time period (unit: %)
- ExLag1-10: Ex 1-10 month(s) ago (unit: JPY/USD)
- R1Lag1-10: R1 1-10 month(s) ago (unit: JPY/USD)
- R2Lag1-10: R2 1-10 month(s) ago (unit: JPY/USD)

All the data is collected from the Federal Reserve Bank of St. Louis database. Original website of the data in can be found in the reference section.

This project draws inspiration from the interest rate parity condition formula below, in which Ex^e is the expected exchange rate and is calculated by ExLag1-10 in this model:

$$R_1 = R_2 + \frac{Ex^e - Ex}{Ex}$$

This economic model posits a robust connection between domestic and foreign interest rates and the exchange rate. Put simply, the idea is that interest rates and exchange rates should exhibit a correlation to eliminate arbitrage opportunities. In this model, changes in interest rates would influence exchange rates, and vice versa, creating a balance. The rationale behind this relationship is to prevent speculators from exploiting discrepancies in interest rates across countries, which could lead to arbitrage opportunities. Should the model fail to hold, it would create openings for speculators to capitalize on market inefficiencies, potentially resulting in disorder within the foreign exchange market.

EDA

Figure 1 illustrates three significant jumps and three stable periods within the dataset. These jumps correspond to key historical events, namely the collapse of the Bretton Woods System (1971-1973), President Carter's intervention to halt the Dollar decline and the 2nd Oil Shock (1977-1979), and the Plaza Accord and Louvre Accord (1985-1987). Since these events are historical occurrences, their impact on future predictions is considered negligible. Consequently, it is advisable to exclude data preceding the third jump in 1988 and concentrate on the dataset during the ongoing stable period that followed. This approach aims to enhance the relevance of the data for more accurate and current predictions.

.

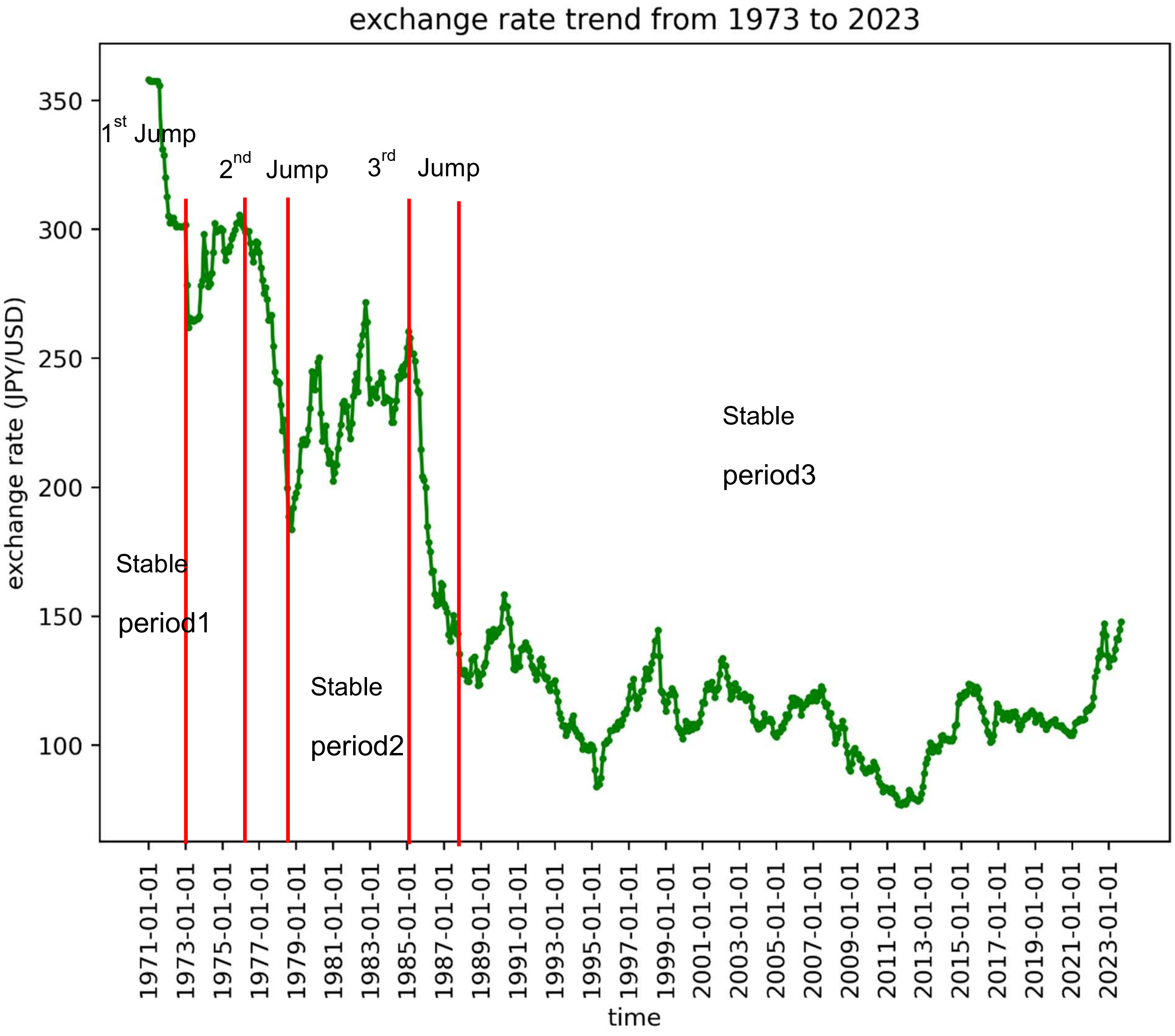

Figure 1 – 3 major jumps and 3 stable periods in the exchange rate history

In Figure 2, a compelling linear relationship is evident between present exchange rates and their historical counterparts. This linear association remains robust, even when comparing present data with data from nine months ago. The consistency in this linear trend suggests that linear regression models may outperform non-linear models in capturing and predicting exchange rate dynamics.

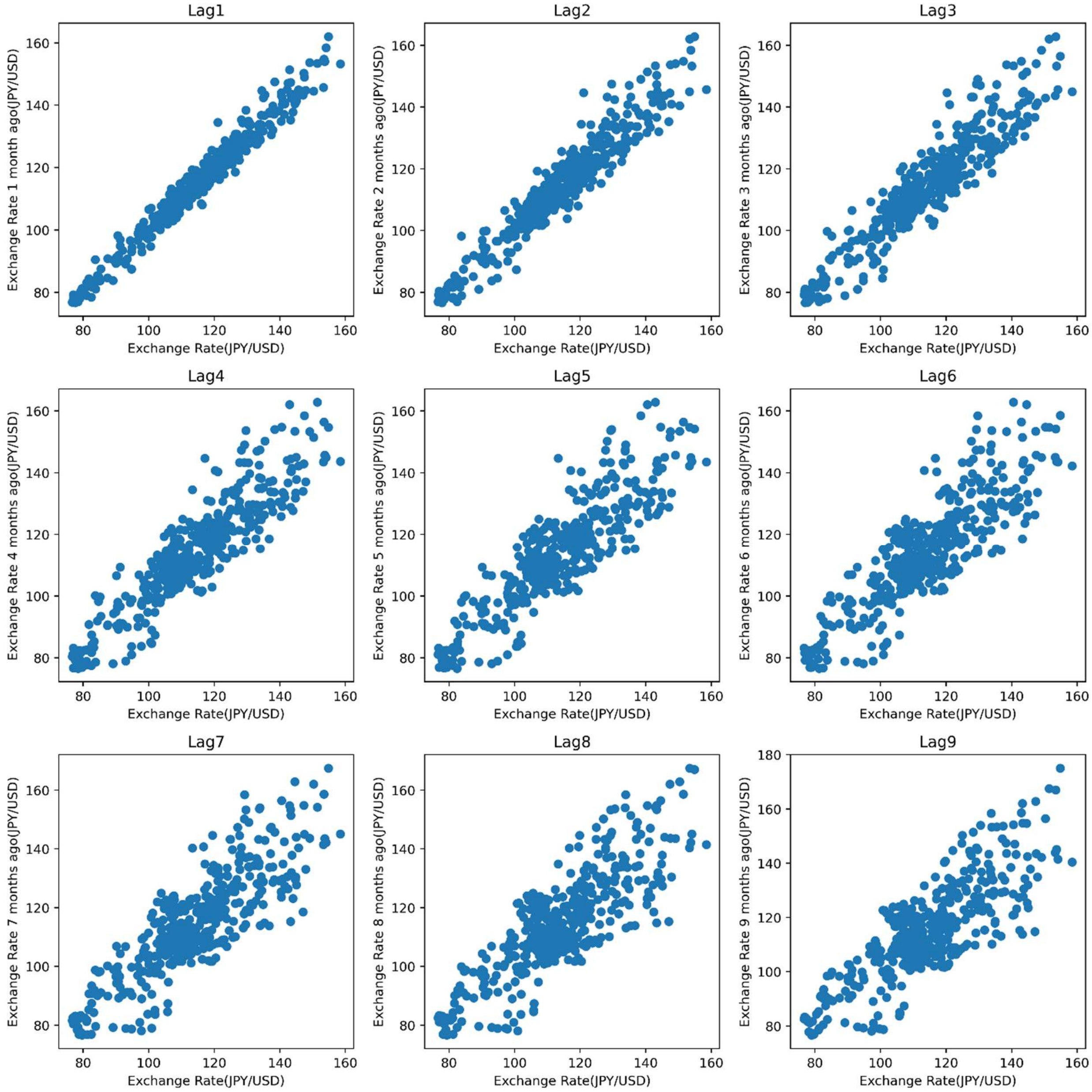

Figure 2 – comparation between past exchange rate and present one

An intriguing observation emerges from figure 3: the US exchange rate appears to be unrelated to the exchange rate when it hovers around 0. However, a notable shift occurs when the exchange rate surpasses approximately 2%. At this point, a positive relationship becomes discernible between these two variables. This peculiar phenomenon suggests that the influence or correlation between the US exchange rate and the overall exchange rate is contingent on the specific range of exchange rate values.

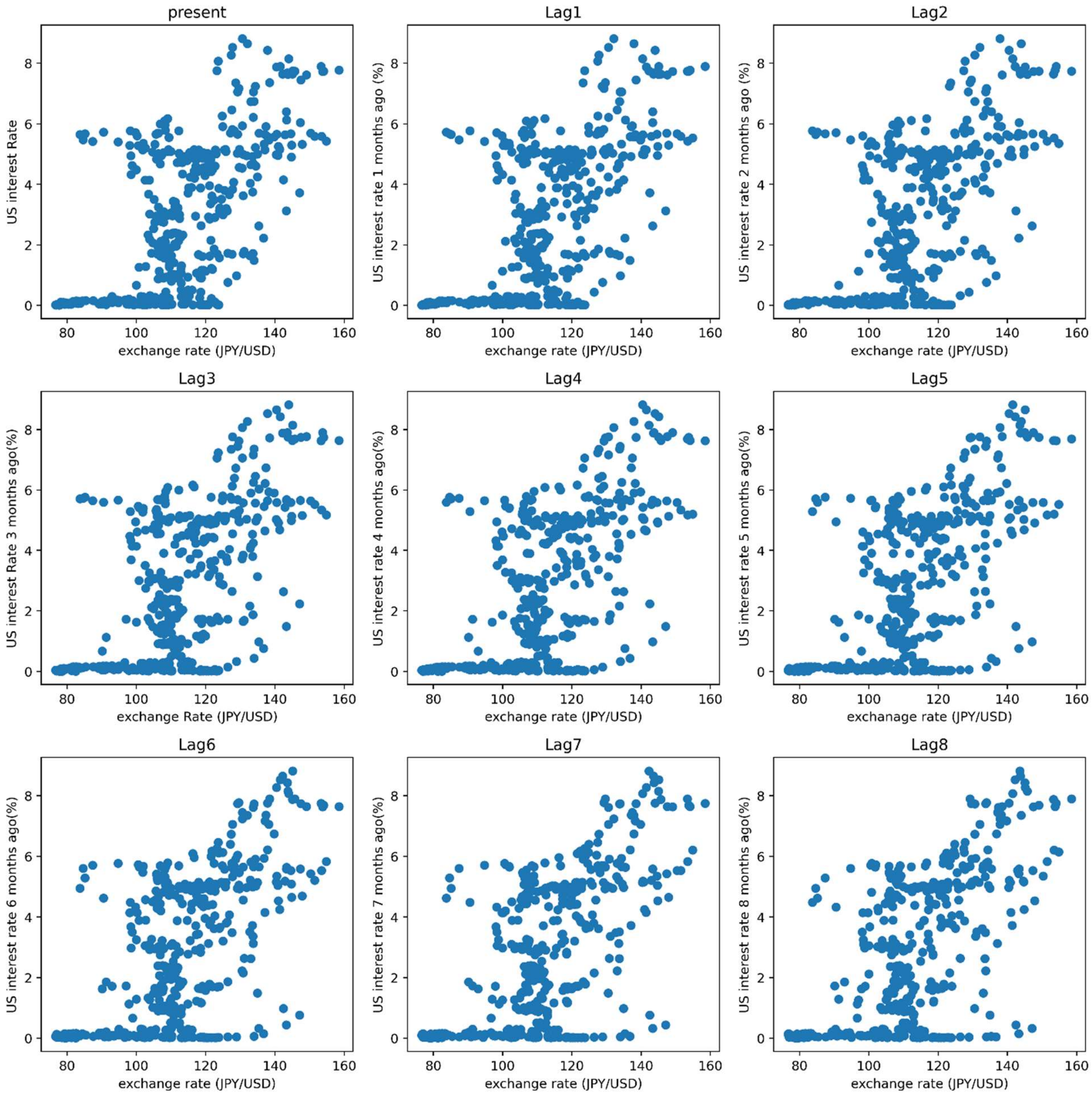

Figure 3 – comparation between present and past US interest rate and exchange rate

Methods

Given that the dataset consists entirely of time series data, utilizing 'TimeSeriesSplit' from scikit-learn seems to be the most appropriate method for splitting. In this model, approximately 60% of the data (283 out of 441) is allocated to the training set, while the validation set (70 out of 441) and the test set (88 out of 441) each receive 20% of the data. The separation of the training, validation, and test sets is depicted in Figure 4, and this temporal division is crucial to mitigate the risk of overfitting. By preventing the algorithm from learning future occurrences in advance, this approach aims to minimize data leakage and enhance the model's generalization performance on unseen data.

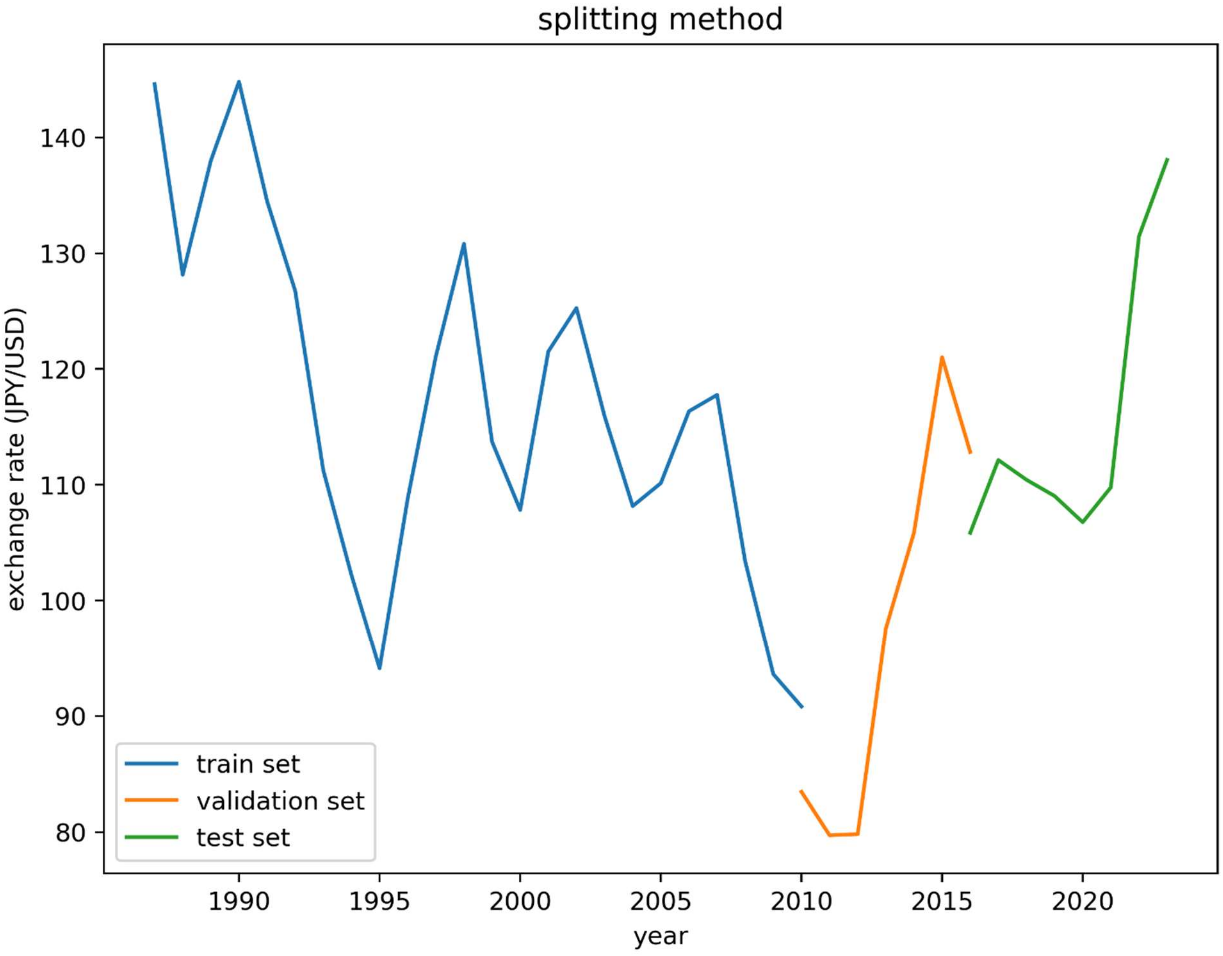

Figure 4 – splitting using sklearn's 'TimeSeriesSplit'

In each dataset, the presence of missing values poses a challenge for subsequent model fitting steps, particularly because most machine learning models employed in this project, with the exception of XGBoost, cannot handle missing values directly. To address this issue, a linear regressor is deployed to estimate the missing values leveraging information from existing features.

Given that all features are continuous and lack explicit boundaries, a 'StandardScaler' is applied for preprocessing. The 'StandardScaler' transforms the features to have a mean of 0 and a standard deviation of 1, ensuring that all features are on a comparable scale. This standardization helps prevent any particular feature from dominating the learning process due to its scale, promoting a more balanced and accurate estimation of missing values.

In this project, a diverse set of five machine learning algorithms—Lasso, Ridge, Random Forest, KNN, and XGBoost—has been employed. To enhance the robustness of the models and control randomness, five different random states (0 to 4) have been deployed for models that incorporate randomness, specifically Ridge, Lasso, and Random Forest.

It's worth noting that random state cannot be utilized in the splitting step because randomly shuffling time series data during the splitting process may lead to data leakage. The temporal order of time series data is critical, and shuffling the data would disrupt this order, potentially compromising the integrity of the model and its predictive capabilities.

Results

The evaluation metric chosen is the root mean square error (RMSE) due to its ability to provide a detailed understanding of prediction accuracy, which can be more informative compared to R2. To establish a baseline for comparison, the RMSE is calculated by comparing the mean value in the test set with the true values. This baseline serves as a benchmark to gauge the performance improvements achieved by the machine learning models. The hyper parameters tuned and the results are as follows:

Model	Hyper parameter(s) tuned	Best Hyper Parameters	Mean RMSE	Baseline RMSE
Ridge	Alpha: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]	0.0001	3.229	11.9320
Lasso	Alpha: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]	0.01	3.115	
XGBoost	Learning Rate: [0.01, 0.1, 0.2] Max Depth: [3, 4, 5] N_Estimators: [50, 100, 200]	0.2 5 50	5.137	
RF	Max Features: [None,1, 3, 10, 20] Max Depth: [None,1, 3, 10, 20]	None 3	3.101	
KNN	N_Neighbors: linspace(1, 250, 50)	1	9.771	

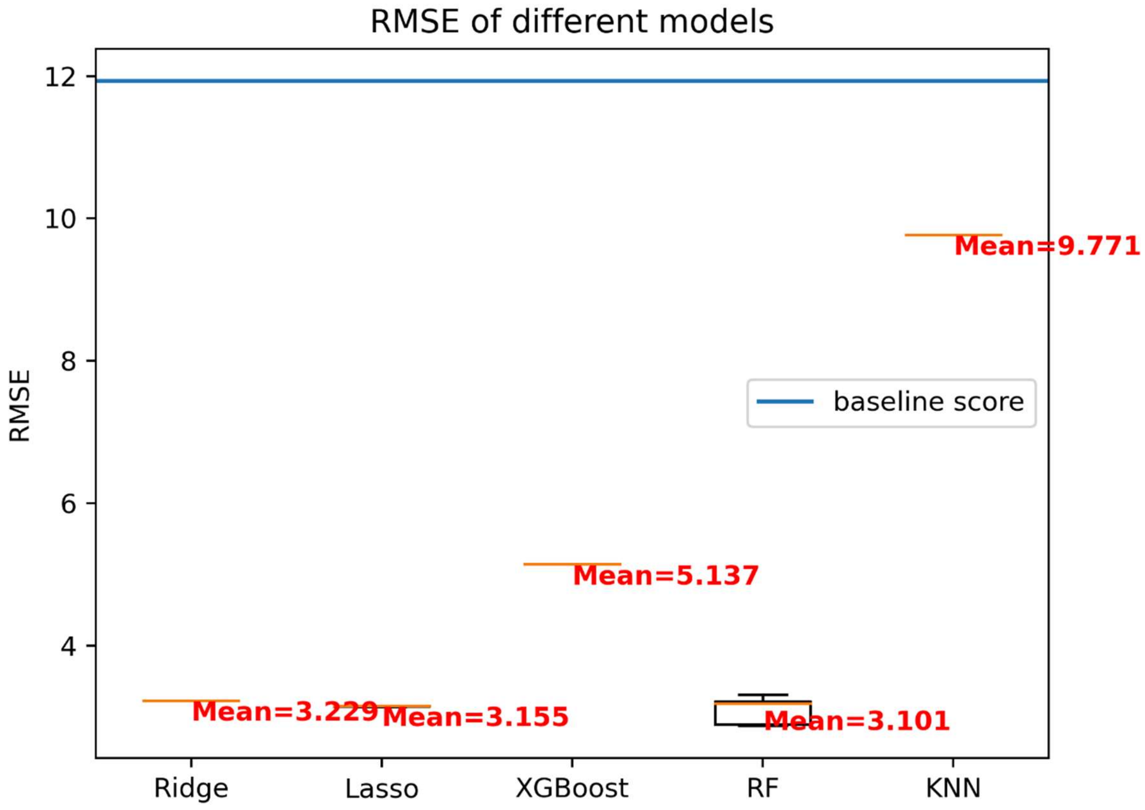

Figure 5 – RMSEs and baseline values of each model

As indicated in the presented results, the RMSEs of all models are significantly below the baseline, underscoring the commendable performance of each. Notably, the Random Forest Regressor (RF) outperforms even the two linear regressors (Ridge and Lasso), which may be unexpected given the strong linear relationships as is discussed in the previous context. Despite a slightly higher standard deviation in the Random Forest model's predictions, this variability remains within acceptable bounds. Next we will discuss mainly the result of the Random Forest model.

Insights from the Random Forest model

In Figure 6, the overall excellent predictive power of the RF is evident. However, a discernible bias appears between 2021 and 2023, and this could be attributed to the exclusion of data related to the Covid-19 pandemic. The model does not incorporate the impact of the pandemic, which may be a significant factor influencing exchange rates during that period.

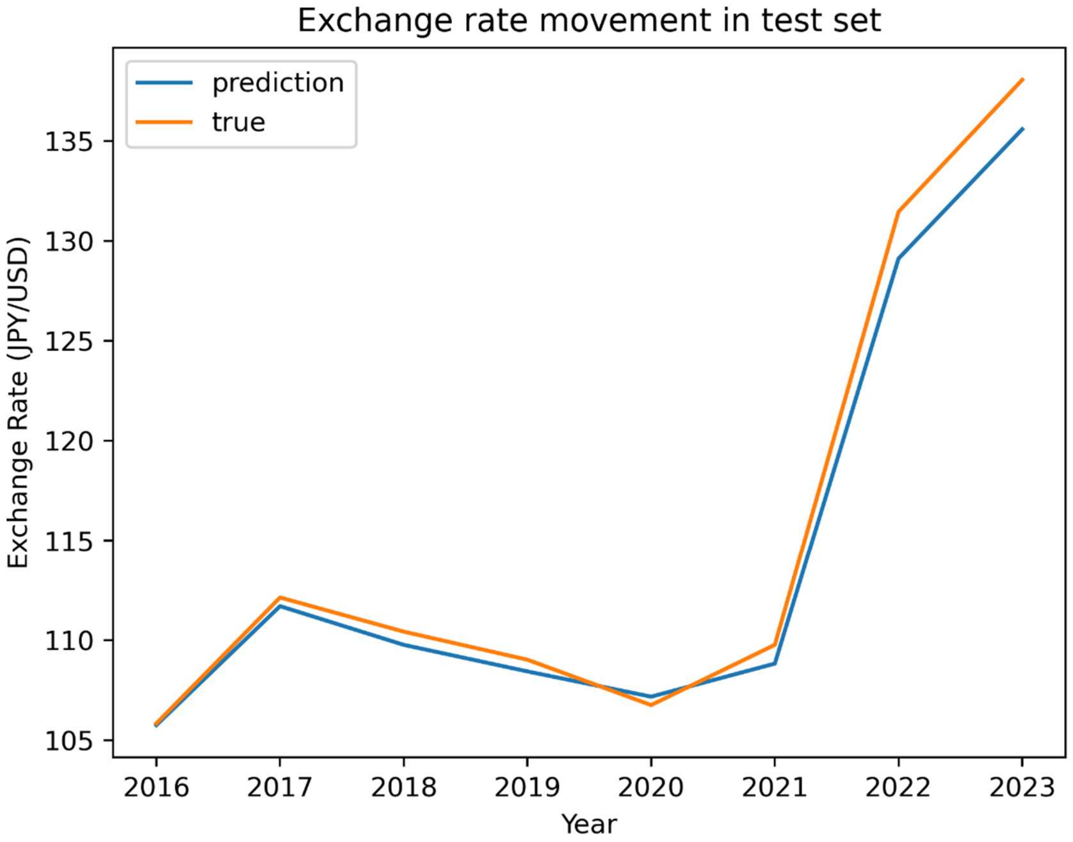

Figure 6 – Overall performance of RF

In this project, the global feature importance is assessed through three different methods: SHAP values, permutation importance, and mean decrease impurity. Specifically, the results for random state = 1 are presented, although it's important to acknowledge that multiple random states were employed (five in total) for the Random Forest model.

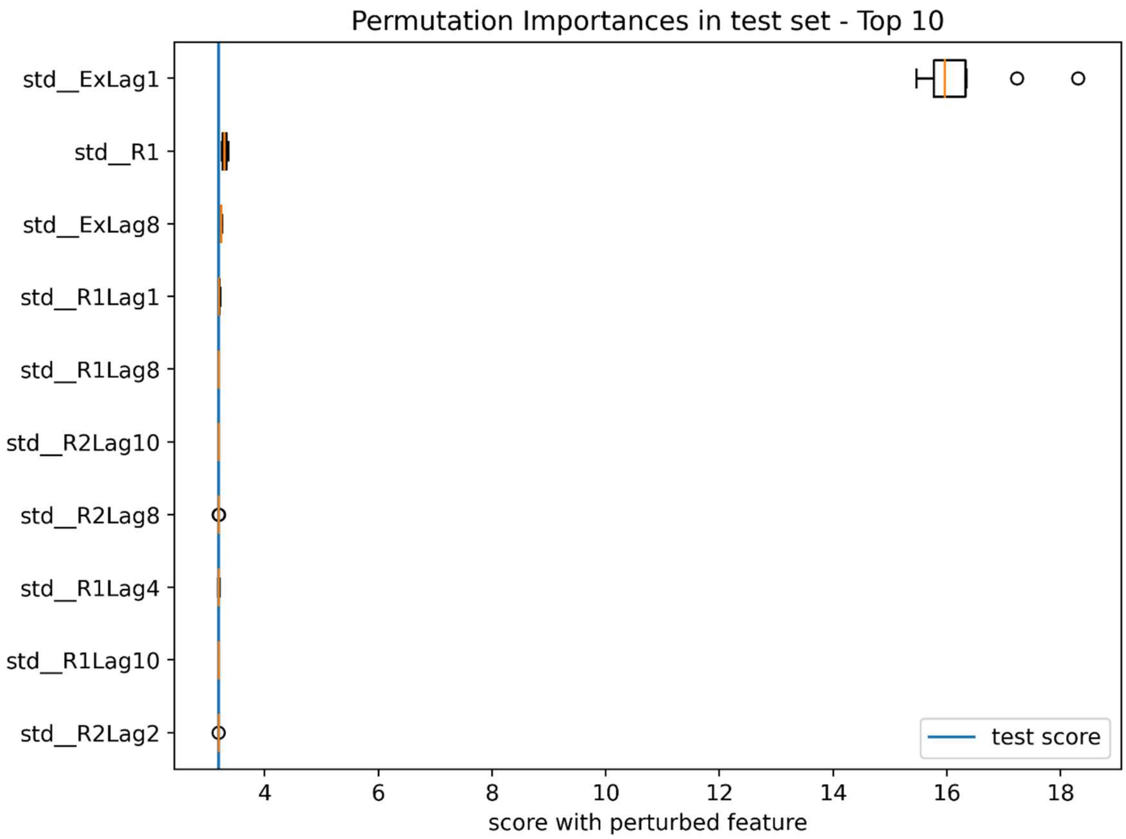

Figure 7 – the top 10 most important features in permutation test

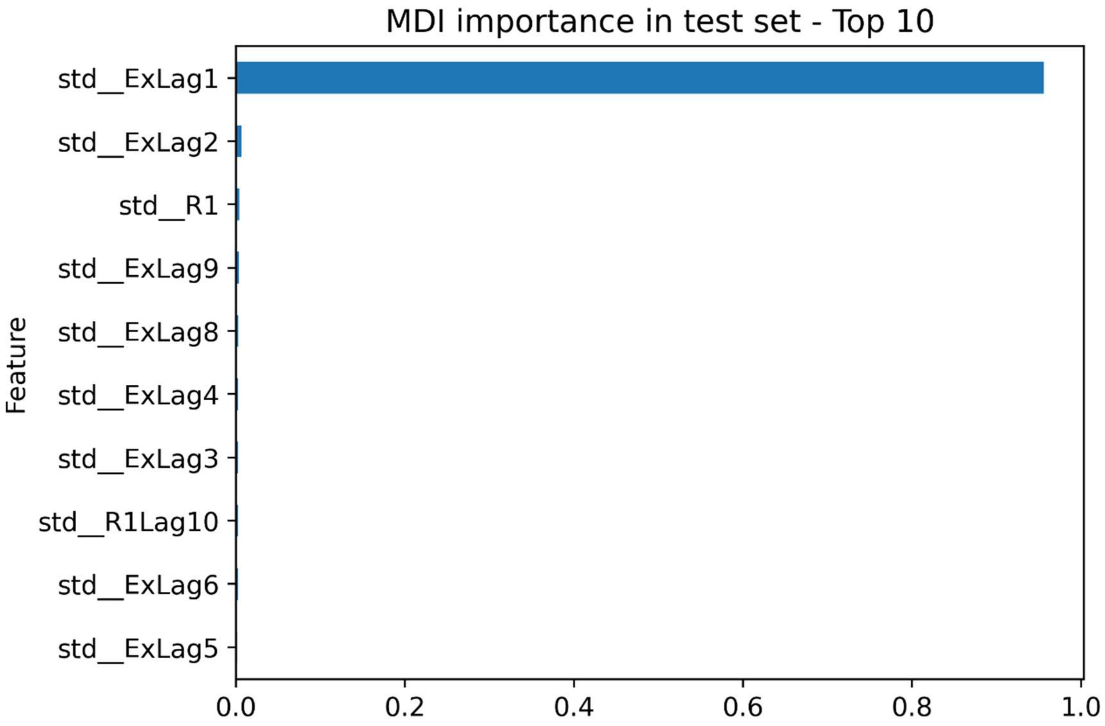

Figure 8 – the top 10 most important features in mean decrease impurity test

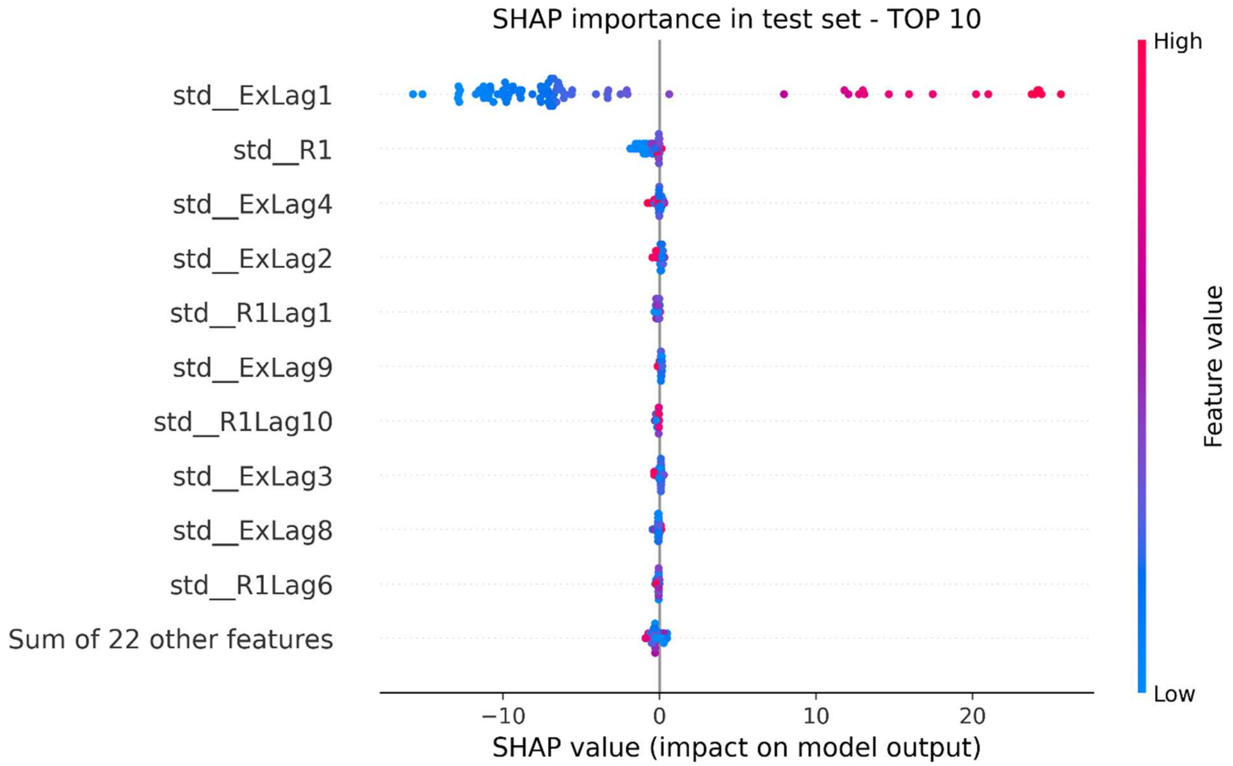

Figure 9 - the top 10 most important features in SHAP test

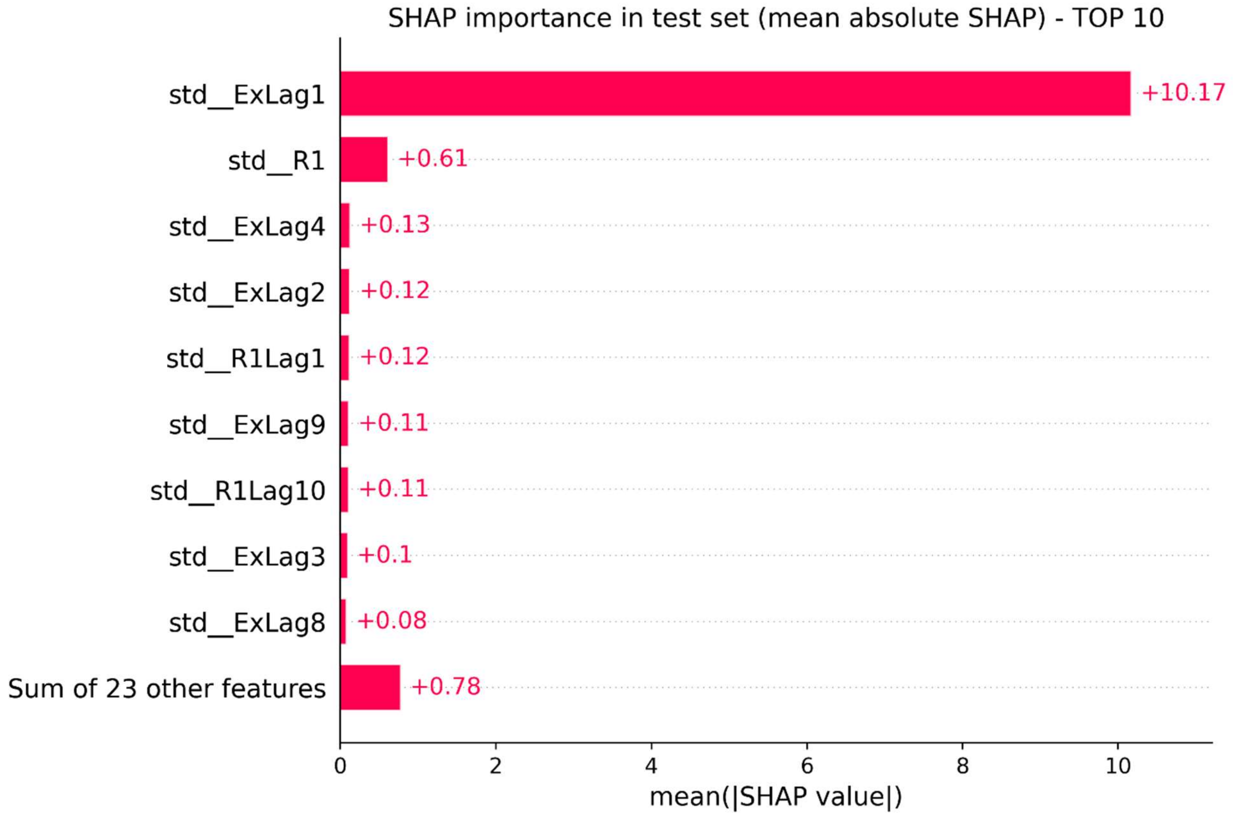

Figure 10 - the top 10 most important features in SHAP test (mean absolute)

The figures presented above reveal the overwhelmingly high importance of the exchange rate one month ago (ExLag1). This observation underscores the strong influence of the immediate past exchange rate on future predictions. The high importance of ExLag1 suggests a clear trend in exchange rate behavior: when the exchange rate in the previous month is decreasing, there is a high likelihood that the exchange rate in the future will also decrease. This correlation points to a persistent trend in exchange rate movements, where recent trends strongly influence future patterns. As for other features, they have little importance on the model except the present interest in the US.

The observation that the model heavily relies on the exchange rate in the last month for predicting the future exchange rate, while giving less weight to other factors such as the interest rate in Japan, raises interesting economic considerations. The prevalence of a strong reliance on the most recent exchange rate alone may be reflective of behavioral biases or limitations in the available data. The limited impact of the interest rate in Japan could indeed be influenced by the extremely low interest rates in recent times, as illustrated in Figure 11. When interest rates are close to zero, their effectiveness in influencing exchange rates may diminish, and other factors, such as recent exchange rate trends, may gain more prominence in shaping expectations.

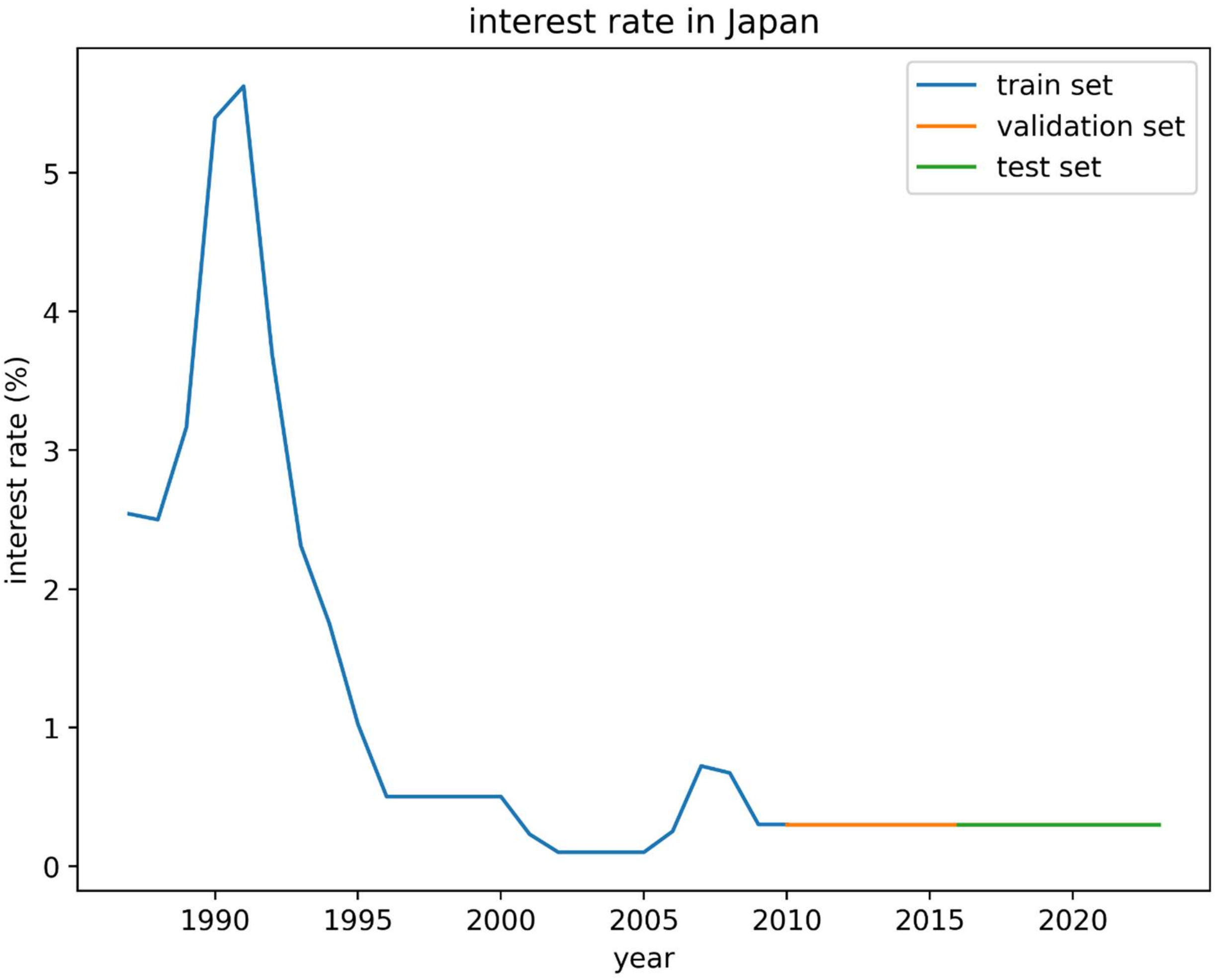

Figure 11 – extremely low and stable interest rates in Japan in recent years

The consistency in local feature importance results, as indicated by SHAP values, further reinforces the dominance of the exchange rate one month ago (ExLag1) in shaping predictions.

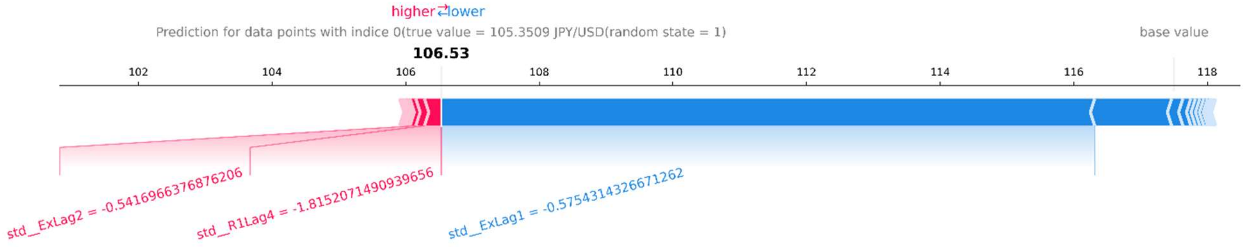

Figure 12 – local feature importance 1

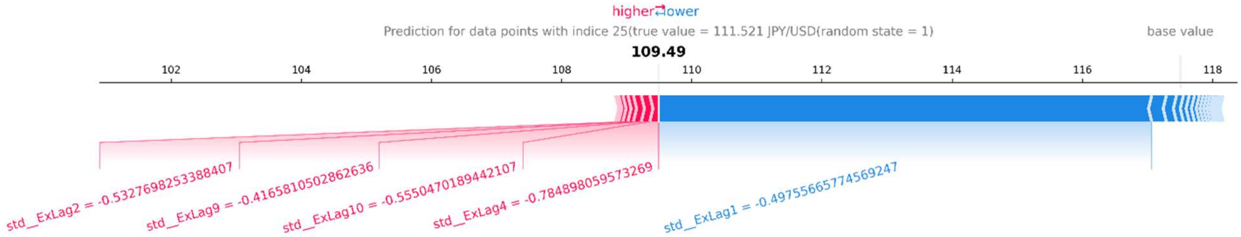

Figure 13 - local feature importance 2

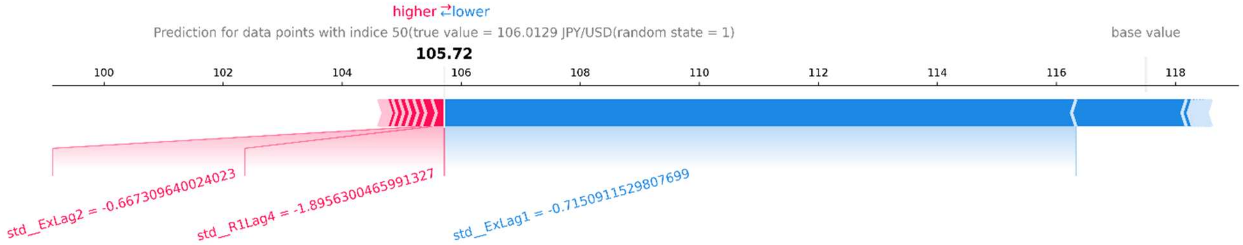

Figure 14 - local feature importance 3

Outlook

The identification of a potential weak spot in the project, specifically the abundance of missing data in the test set, is a valid concern. The reliance on linear regression imputation for the interest rate data in Japan, which extends beyond the available data in the Federal Reserve Bank's database (ends by 1/4/2017), introduces a potential bias in the test score evaluation. To address this issue and obtain a more accurate test score, acquiring additional and more recent data on the interest rate in Japan from another data source is required.

Increasing the number of estimators in the Random Forest regressor is another valid strategy to potentially enhance predictive power. By adding more estimators, the model can capture more complex patterns in the data, leading to a potentially more accurate regressor. While increasing the number of estimators in a Random Forest introduces a potential risk of overfitting, a sufficient number of them can effectively smooth out noisy artifacts produced by individual estimators, resulting in a more accurate regressor. It's possible that some added estimators may introduce noise and worsen the ensemble, but overall, the quality of the model is expected to improve.

Incorporating additional economic models alongside the existing framework is also a good approach. In the current project, the focus is primarily on the equilibrium of the foreign exchange market. However, the goods market and the domestic and foreign money markets also exert a non-negligible influence on the exchange rate. For further insights, refer to chapters 5, 6, and 7 in *International Finance* as listed in the reference page, for details about the AA-DD model.

References

3-month Treasury Bill Secondary Market Rate, discount basis. FRED. (2023a, December 1). https://fred.stlouisfed.org/series/TB3MS

Interest rates, discount rate for Japan. FRED. (2019, April 29). https://fred.stlouisfed.org/series/INTDSRJPM193N

Japanese yen to U.S. Dollar Spot Exchange Rate. FRED. (2023b, December 4). https://fred.stlouisfed.org/series/EXJPUS

Krugman, P. R., Obstfeld, M., Melitz, M. J., & Krugman, P. R. (2018). *International finance: Theory and policy*. Pearson Education.