HO CHI MINH CITY NATIONAL UNIVERSITY
UNIVERSITY OF ECONOMICS AND LAW

FINAL ESSAY

Subject: Machine Learning
Lecturer: Master Phan Huy Tam

TOPIC: FORECASTING VND EXCHANGE RATE

Student: Huỳnh Ngọc Phương Thảo ID: K214142083

HCM City, 17th of June, 2024

INSTRUCTOR'S COMMENTS:

Ho Chi Minh City, 17th of June, 2024

Instructor scores

Table of Contents

1. Introduction

In today's globalized economy, exchange rates play a crucial role, influencing every aspect of life, from international trade and foreign investment to personal financial decisions. As the dominant currency in the international monetary system, the USD (US Dollar) has become a particularly important focal point, attracting the attention of investors, businesses, and policymakers. Therefore, accurate forecasting of USD exchange rates has become a strategic topic of immense value in various fields. Businesses involved in import and export rely on exchange rates to price their products, calculate profits, and manage foreign exchange risk. Forecasting USD exchange rates enables them to make sound business decisions, optimize profits, and mitigate losses due to exchange rate fluctuations. International investors consider USD exchange rates when making investment decisions in different countries. Forecasting exchange rates helps them assess potential returns, minimize risks, and make effective investment decisions. Central banks of various countries hold USD as part of their foreign exchange reserves to ensure financial stability and fulfill international obligations. Forecasting USD exchange rates assists them in managing reserves effectively, preventing losses from exchange rate volatility. Exchange rates are a critical factor influencing monetary policy decisions in different countries. Forecasting USD exchange rates empowers policymakers to adjust interest rates, exchange rates, and other monetary tools effectively, contributing to macroeconomic stability. USD exchange rate fluctuations reflect the relative strength of the US economy compared to other economies. Exchange rate forecasting provides valuable insights into future economic trends, guiding investment, business, and government decisions. Exchange rates are subject to significant fluctuations due to various economic, political, and market sentiment factors. Forecasting exchange rates enables investors, businesses, and individuals to navigate this volatility, make informed decisions, and mitigate risks. Conclusion, forecasting USD exchange rates is crucial for various sectors, ranging from international trade and financial investments to personal finances. Understanding the trends in USD exchange rates empowers investors, businesses, and governments to make sound decisions, optimize profits, and minimize risks. Consequently, research and forecasting of USD exchange rates deserve greater attention and focus to contribute to sustainable economic and social development.

2. Describe the dataset

The exchange rate dataset between the Vietnamese Dong (VND) and the US Dollar (USD) is a valuable resource for professionals in finance, economics, and

market research. Documenting data since 1/2004, this dataset records the daily fluctuations of the VND/USD exchange rate, reflecting economic, political, and market changes that influence the value of the currency.

Date	VND=X
01/01/2004	
01/02/2004	15147
01/05/2004	15148
01/06/2004	15150
01/07/2004	15153
01/08/2004	15129
01/09/2004	15176
01/12/2004	15158
1/13/2004	15163
1/14/2004	15162
1/15/2004	15164
1/16/2004	15187
1/19/2004	15143
1/20/2004	15138

Date: The transaction date, recorded in MM/DD/YYYY format. This column contains information about the date on which the exchange rate was noted.

VND=X: The exchange rate value of VND relative to USD. Each value in this column represents the number of VND equivalent to 1 USD on the respective date.

count	mean	std	min	25%	50%	75%	max
4949	19962.6755	3036.42262	20.89	16525	20912	22675	24871

Count: There are 4,949 entries in the "VND=X" column, indicating the number of observations available for the VND/USD exchange rate.

Mean: The average exchange rate is 19,962.6755 VND per 1 USD. This suggests that on average, one US dollar is exchanged for about 19,963 Vietnamese dong.

Std (Standard Deviation): The standard deviation is 3,036.42262, showing that the exchange rate fluctuates significantly around the mean. This reflects high volatility or instability in the VND/USD exchange rate over the period covered by the data.

Min: The minimum value is 20.89, which may be an outlier as it's highly unlikely for the exchange rate to be this low under normal conditions. It might be necessary to review this data point for accuracy.

25% (First Quartile): The first quartile value is 16,525 VND/USD, meaning that 25% of the observations are equal to or less than 16,525 VND/USD.

50% (Median): The median of the data is 20,912 VND/USD, indicating that half of the data points have an exchange rate lower than 20,912 VND/USD and the other half have a higher rate. This also reflects the central tendency of the data without being influenced by outliers.

75% (Third Quartile): 75% of the observations have values below 22,675 VND/USD. This shows that the majority of the data falls below this exchange rate.

Max: The highest value recorded is 24,871 VND/USD, indicating the peak exchange rate reached during the observation period.

From these statistics, it is evident that the VND/USD exchange rate experiences substantial fluctuations, ranging from very low to quite high, with significant swings during the study period. These fluctuations may be due to major economic shifts, political changes, or global financial market factors.

variable	dtype	count	unique	missing value
VND=X	float64	4949	2422	7

The results you shared from the describe() function in Python provide a general overview of the "VND=X" column in your dataset. Here's what each parameter means:

variable: The name of the column being analyzed, here it is "VND=X".

dtype: The data type of the column, which in this case is float64. This indicates that the values in the column are floating-point numbers.

count: The number of non-null values in the column, representing the number of valid observations. For "VND=X", there are 4949 valid entries.

unique: The number of unique values in the column. For "VND=X", there are 2422 unique values, indicating that there have been various different exchange rates between VND and USD during the period covered by the data.

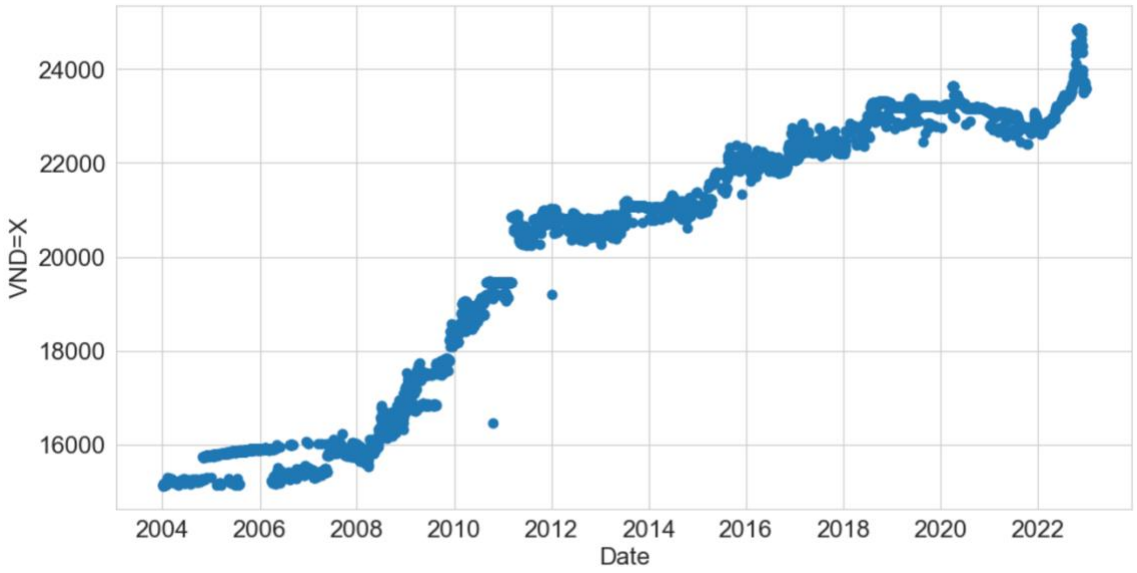
missing value: The number of missing values in the column, here it is 7. This means that there are 7 days in the dataset for which the VND/USD exchange rate information is missing.

df.dropna(inplace=True): This line removes all rows in the DataFrame df that contain any NaN values. The inplace=True parameter means that the changes are made directly to df without needing to assign the result back to df. Essentially, after this operation, df will no longer have any rows with missing data.

df.reset_index(): This line resets the index of the DataFrame df. When you drop rows from a DataFrame, the index can become non-sequential. Resetting the index rearranges the index in a continuous sequence starting from 0. However, without assigning it back to df or using inplace=True, this operation will not change df itself but only return a new DataFrame with the index reset.

There are many methods for identifying outliers, but one popular method is using IQR (Interquartile Range).

3. Data Visualization



General Trend: The chart exhibits a general upward trend in the USD/VND exchange rate from 2004 to 2022. This indicates a continual depreciation of the VND relative to the USD over the long term.

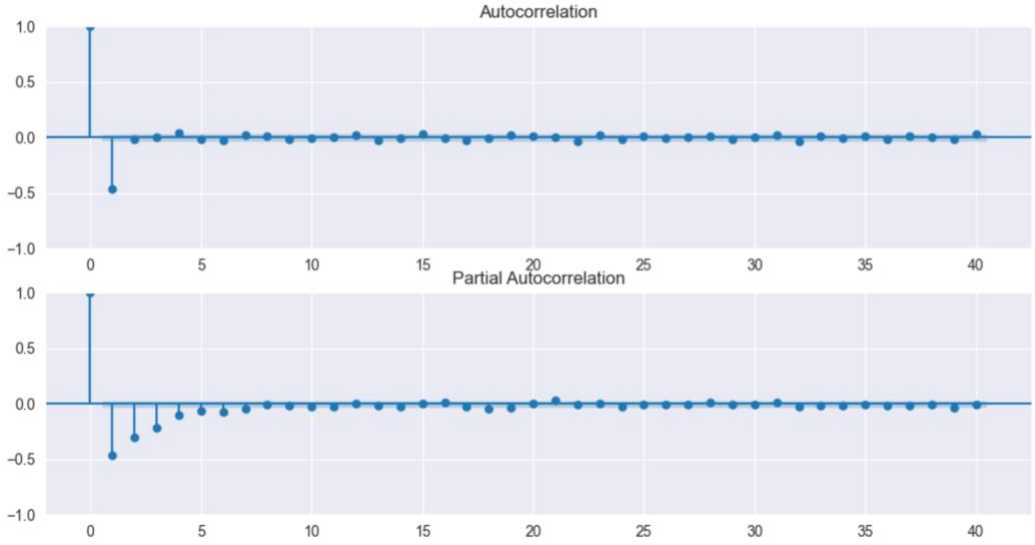
Stability and Volatility: The period from 2004 to about 2008 shows the exchange rate being relatively stable, with minor fluctuations. From 2009 to 2014, there is a noticeable sharp increase in the USD compared to the VND, indicating strong volatility during this period. After 2014 to near 2020, the exchange rate becomes more stable again with smaller fluctuations, though the upward trend of USD appreciation continues. From 2020 to 2022, there is another sharp rise in USD value, possibly due to macroeconomic factors or monetary policies.

Significant Drops: There are points on the chart showing sudden drops in exchange rate (vertical blue lines). These could be due to data errors or unusual economic events. If these are accurate data points, they require detailed analysis to understand the causes.

Recent Trends: The end of the chart shows a very strong increase in the USD value relative to the VND, particularly in 2022. This could be the result of global economic instability or changes in the monetary policies of the countries involved.

This chart provides important insights into how exchange rate fluctuations can reflect broader economic trends and underscores the need for careful economic monitoring and analysis.

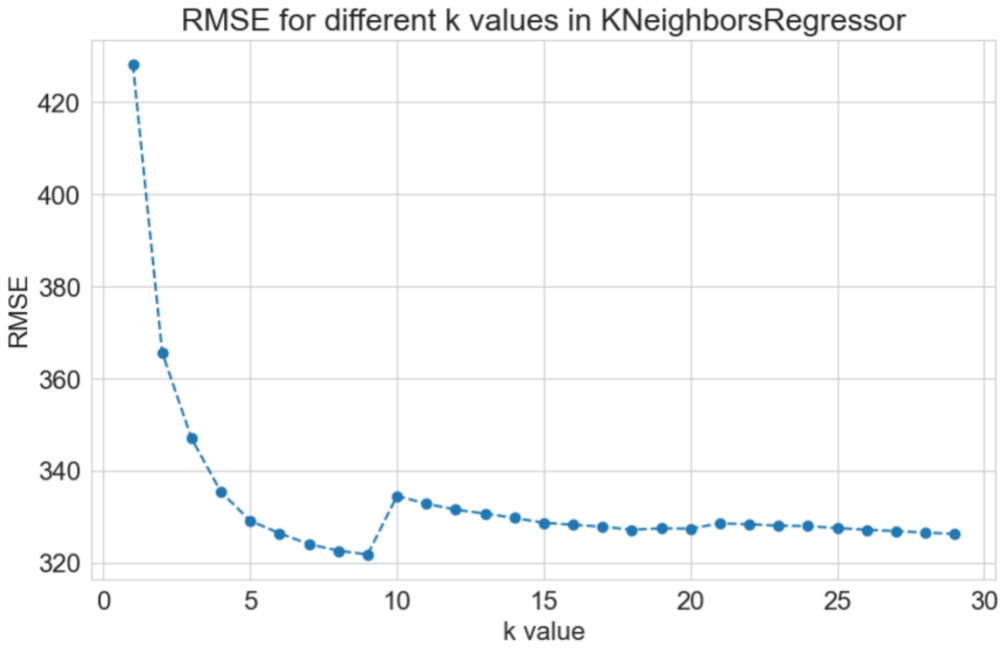
ADF Statistic: -0.4847593777363119
p-value: 0.8949564446343037



Autocorrelation (Top Chart): This chart shows the correlation of the series with itself at different lags. The autocorrelation at lag 0 is always 1 because the series is perfectly correlated with itself. The significant spike at lag 1 suggests a strong correlation from one period to the next. As the lag increases, the autocorrelation values decline but remain above zero, indicating some level of persistence in the series.

Partial Autocorrelation (Bottom Chart): This chart displays the correlation of the series with itself at different lags but after removing the effects of any correlations due to the terms at shorter lags. There is a significant spike at lag 1, which indicates a significant correlation at lag 1 even after accounting for the influence of intervening lags. The near-zero values at higher lags suggest that the direct correlation at these lags is minimal when indirect correlations through intermediate terms are removed.

4. Using KNN model in forecasting



The image depicts a plot showing the RMSE (Root Mean Squared Error) values for different k values in a KNeighborsRegressor. Here are some observations:

At k=1, the RMSE is the highest, indicating poor performance. This is expected because with k=1, the model is highly sensitive to noise in the data.

As k increases from 1 to around 9, there is a steep decline in RMSE. This suggests that the model's performance improves significantly as more neighbors are considered, which helps to reduce the noise sensitivity.

After k=9, the RMSE values start to stabilize and fluctuate slightly around a lower value. This indicates that the model's performance doesn't change much with further increases in k.

Based on this plot, a k value of around 9 seems to be the most effective for minimizing RMSE while maintaining stability in predictions.

The optimal range for k appears to be between 9 and 12, where the RMSE is lowest and relatively stable. Choosing a k value in this range is likely to give you a good balance between bias and variance, providing a robust model.

The model is configured to account for seasonality with a cycle of 52, which is typical for weekly data and assumes an annual repetition pattern.

Seasonal Analysis: Setting up seasonality in the model is crucial, especially if the data exhibits clear seasonal variations, which can significantly impact forecasting effectiveness.

Model Improvement: You could experiment with adjusting the p and q parameters to see if this improves the model's performance, especially if the current values do not provide adequate forecasting accuracy.

Testing and Adjustment: Testing and adjusting the model parameters based on further analysis (such as ACF and PACF plots after differencing) can be helpful to ensure the model best fits the data.

Model: SARIMAX(0, 1, 1)x(0, 1, 1, 52)

This represents a SARIMAX model with both non-seasonal and seasonal components: (0, 1, 1) are the non-seasonal parameters indicating no autoregressive (AR=0) components, one differencing (I=1), and one moving average (MA=1) component.

(0, 1, 1, 52) are the seasonal parameters, with the cycle repeating every 52 time units (potentially weeks), also with no seasonal AR, one seasonal differencing, and one seasonal MA component.

ma.L1 and ma.S.L52 are coefficients for non-seasonal and seasonal MA components:

ma.L1 = -0.6821 with a P>|z| of 0.000, indicating that this parameter is statistically significant.

ma.S.L52 = -0.9994 also with a P>|z| of 0.000, showing that this seasonal MA component is also highly significant.

The std err values show the standard deviation of the estimates for the coefficients, with lower values indicating more precise estimates.

Log Likelihood: -31253.115 - Indicates the likelihood of the model fitting the data; higher is better.

AIC (Akaike Information Criterion): 62512.231 - A measure evaluating the model based on the model fit and the number of parameters. Lower is better.

BIC (Bayesian Information Criterion): 62531.715 - Similar to AIC but penalizes the number of parameters more heavily, helping to avoid overfitting.

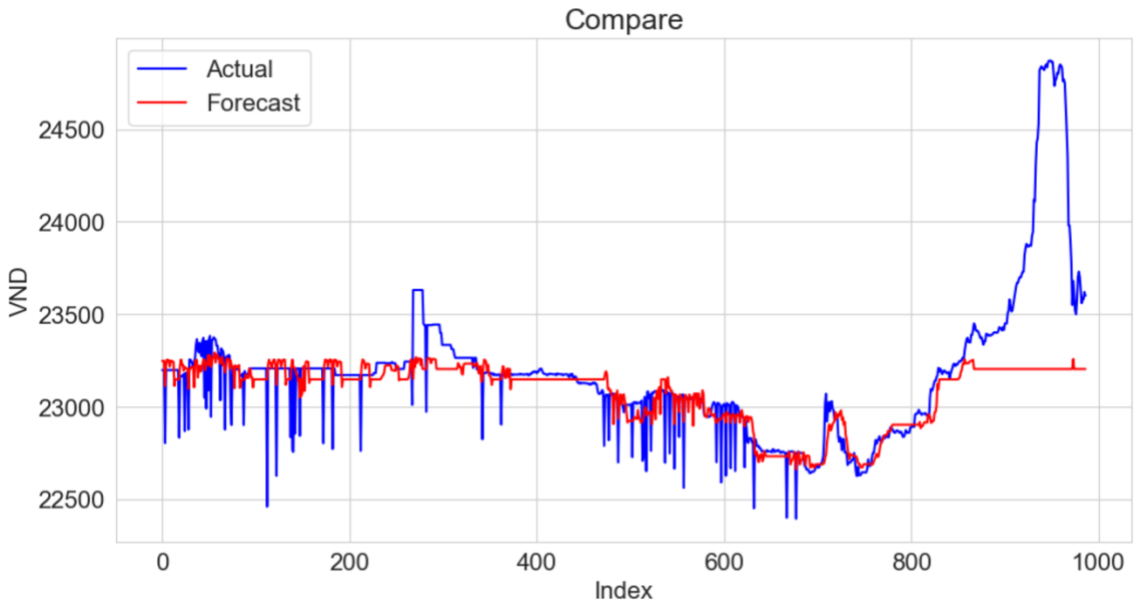
Ljung-Box (Q): Tests whether residuals are randomly distributed, lacking autocorrelation patterns.

Jarque-Bera (JB): Tests whether residuals are normally distributed.

Both tests suggest Prob(Q) and Prob(JB) both are 0.0, indicating that the residuals are not completely random nor normally distributed.

The model parameters indicate that both MA components are highly statistically significant, but additional statistics on residuals suggest the model may not perfectly capture or explain the data's entire dependency structure. Improvements could include examining additional parameters, refining the model to better account for any residual autocorrelation or non-normality, and possibly incorporating more complex seasonal patterns or additional explanatory variables if applicable.

This analysis highlights the strengths and limitations of your current model and suggests areas for further refinement to enhance forecasting accuracy.



The image shows a comparison between the actual VND (Vietnamese Dong) values and the forecasted values over time. Here are some observations:

Trend Agreement: The forecast (red line) generally follows the overall trend of the actual data (blue line), capturing the major movements in the exchange rate.

Initial Fit: In the initial part of the series, the forecast closely aligns with the actual values. This suggests that the model is performing well in periods with relatively stable or less volatile data.

Volatility: During periods of high volatility, particularly around the 800th index mark, the forecast deviates significantly from the actual values. The forecast line remains relatively flat, failing to capture the sharp movements in the actual data. This indicates that the model may struggle with highly volatile periods or sudden spikes.

Lag in Forecast: There are instances where the forecast seems to lag behind the actual values, especially during sudden shifts. This lag suggests that the model may not be able to quickly adapt to rapid changes in the market.

Period of Stability: There are also periods where the forecast remains flat for extended durations, even though the actual values are changing. This flat forecast could be a result of the model over-smoothing the data or not reacting adequately to new information.

Possible Overfitting or Underfitting: The disparity in some regions, especially during volatile periods, suggests the need to refine the model. It might be overfitting to past stable trends and underfitting to new, volatile data points.

Model Improvement: To improve the model's performance, consider exploring more complex models or additional features that capture volatility better. Techniques like adding external regressors or using models designed for high volatility data (e.g., GARCH models) might be beneficial.

Re-evaluate Parameters: Reassess the model parameters and consider incorporating more recent data or performing a rolling forecast to continually update the model with the latest data.

Hybrid Approach: Using a combination of models (ensemble methods) may help in capturing both the stable trends and the volatile spikes more effectively.

Feature Engineering: Introduce new features that may help the model understand and react to changes in the trend, such as lagged variables, rolling statistics, or macroeconomic indicators.

5. Conclusion

The K-Nearest Neighbors algorithm is a simple yet effective supervised learning algorithm used for classification and regression tasks. Its concept is intuitive: it assumes that similar data points are close to each other in the feature space.

During the training phase, the algorithm simply stores the feature vectors and their corresponding labels in memory. This makes the training phase computationally inexpensive. When given a new, unseen data point, the algorithm calculates the distances between this point and all other points in the training set. The most common distance metric used is the Euclidean distance, although other metrics can also be used depending on the problem domain. After calculating the distances, the algorithm selects the K closest data points (neighbors) to the new data point. For classification tasks, the algorithm assigns the class label that appears most frequently among the K neighbors to the new data point. In regression tasks, the algorithm computes the average (or weighted average) of the labels of the K neighbors and assigns this value to the new data point. Finally, the performance of the model is evaluated using metrics such as accuracy (for classification) or mean squared error (for regression). The value of K is a hyperparameter that can be tuned to optimize model performance.

One of the key advantages of the KNN algorithm is its simplicity and ease of implementation. It also doesn't make any assumptions about the underlying data distribution, which makes it particularly useful for non-linear data. However, it can be computationally expensive, especially when dealing with large datasets, as it requires calculating distances between the new data point and all points in the training set.

In the context of forecasting USD exchange rates, the KNN algorithm can be applied by using historical exchange rate data as features and predicting future exchange rates based on the K nearest historical data points. This approach allows for flexible modeling of exchange rate fluctuations and can provide valuable insights for investors, businesses, and policymakers. However, it's important to note that the performance of the KNN algorithm may vary depending on the specific characteristics of the exchange rate data and the choice of hyperparameters such as the value of K

REFERENCES

1. https://luanvan.co/luan-van/tieu-luan-ty-gia-hoi-doai-tac-dong-len-xuat-nhap-khau-50594/
2. https://dangcongsan.vn/cung-ban-luan/giai-toa-ap-luc-dieu-hanh-ty-gia-663918.html
3. https://vietstock.vn/2022/01/nhin-lai-dien-bien-ty-gia-nam-2021-757-921901.htm