Comparative Analysis Based on Machine Learning Model Predicting Exchange Rate Fluctuations

Ziqian Niu

College of Information Science, Massey University, 151 Dairy Flat Highway, Auckland 632
New Zealand, New Zealand

23013675@massey.ac.nz

Abstract. The foreign exchange market has long been a focal point in the financial sector, attracting considerable attention. On a macro level, accurately predicting exchange rate trends is crucial for shaping effective economic and financial policies. For micro-entities, effectively managing and mitigating foreign exchange risks is a key challenge. Thus, forecasting foreign exchange trends holds significant importance across various fields. This article will use the method of model comparison to bring actual data to analyze the three models in the field of foreign exchange forecasting. The three different machine learning models, namely the Linear Regression Model (LSTM), Random Forest Model, and Long Short-term Memory Model, are widely used in the field of artificial intelligence, to predict the foreign exchange rate in a certain period. And choose the best foreign exchange prediction model while comparing their respective advantages. From the perspective of model prediction performance on the same set of foreign exchange data, the LSTM model became the optimal prediction model with the smallest mean square error (2.54e-05) among the three models.

Keywords: Machine learning, exchange rate, LSTM Model, Random Forest Model, Linear Regression Model.

1 Introduction

Foreign Exchange (FX) is the biggest financial market in the world. FX rate forecasting is an important issue as it affects the economic development of a country [1]. The exchange rate is also called the FX market, which refers to the price comparison and exchange rate between the two currencies. The FX market has the characteristics of unified space, high liquidity, and high leverage, which has caused investors to receive high risks while obtaining high profits.

According to Luo Jiemei, a currency does not increase or decrease in absolute, but its valuation may fluctuate relatively to another currency [2]. Exchange rate prediction is crucial for companies, influencing decisions on hedging, short-term financing, investment, and capital budgeting. Hedging decisions hinge on exchange rate forecasts, while international financing involves considering both financing rates and exchange rate changes. These predictions also inform national economic policies and facilitate

B. Siuta-Tokarska et al. (eds.), *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, Advances in Economics, Business and Management Research 300,
https://doi.org/10.2991/978-94-6463-542-3_55

global cooperation in FX research. Accurate forecasting helps avoid potential economic instability, enabling wise investment decisions and fostering enterprise and national development. This essay will use three different machine learning methods to predict the data group after sorting out the latest real-time exchange rate data, checking numerical features, and processing missing values. The paper is designed to provide a more accurate method of foreign exchange rates.

2 Literature Review

FX rates are influenced by many factors, including political, economic, social, and natural factors. It has been shown in the literature that due to intensified geopolitical risks, the conflict hurt FX rates [3]. Therefore, predicting FX rates remains challenging for researchers to construct highly accurate exchange rate forecasting models [4]. After reviewing relevant information, it was found that artificial intelligence (AI) applications are critical for capturing the linear and nonlinear behaviors of finance variables [5]. So, using machine learning to predict FX rates is a widespread application. Following previous research reports, it has been shown that in previous traditional research, time series models have been widely used in the field of financial forecasting [6]. However, due to the complexity and nonlinearity of financial time series, many traditional models cannot fit exchange rates well. From the type point of view, data in the financial field are mainly time series data and financial time series have many high-dimensional features, a lot of noise, and a strong dependence on time variables [7]. Moreover, the characteristics of time series data will change significantly over time. Traditional measurement methods or equations containing parameters are not suitable for analyzing such data, which brings challenges to FX forecast research. In many previous research reports, scholars mostly used combination research methods to predict FX rates, such as the Support Vector Regression (SVR) Model, Support Vector Machine (SVM) Model, and Particle Swarm Algorithm Neural Network (PSO-NN) Model. This article wants to explore the role of a single-machine model in FX rate forecasting. This article aims to apply three different machine learning models, namely the Linear Regression (LR) Model, Random Forest (RF) Model, and Long Short-term Memory (LSTM) Model, to a set of official exchange rate data from August 30, 2004, to February 22, 2024 [8]. By using three machine learning models to track the short-term exchange rate prediction task of a specific currency, it can be seen the performance of the three models in the field of exchange rate prediction and their respective characteristics and applicable places thus comparing the advantages and disadvantages of the three models and select the model that can quickly and accurately grasp the changes in FX rates, providing appropriate assistance to people from all walks of life who need more accurate control of FX rates.

3 Research Methodology

3.1 Data Collection and Description

When using machine learning to predict FX rates, it needs to go through five steps: 1. the process of collecting data related to FX rates. 2. Clean and process the obtained data, including missing value handling, outlier handling, and data normalization. 3. Based on the collected data, extract relevant features and the engineering process of influencing factors. 4. The process of selecting and evaluating three models after comparing their strengths and weaknesses, such as calculating prediction accuracy and error rate.5. Optimize and adjust the models based on the evaluation results.

Before starting this research, it is necessary to find a set of real-time updated FX rate data on Kaggle. This dataset contains daily forex exchange rate data from 2004 to the present. It includes columns currency, base currency, currency name, exchange rate, and date, which effectively ensure the authority and accuracy of data. The first step in organizing the data is to read the data set and display the first few rows of the data. Table 1 (The structure of the autoencoder) shows the first few sets of FX data.

Table 1. The structure of the autoencoder

Number	Currency	Base Currency	Currency Name	Exchange Rate	Date
0	Zimbabwean Dollar (ZWL)	EUR	Zimbabwean Dollar	348.491878	22/02/2024
1	Gibraltar Pound (GIP)	EUR	Gibraltar Pound	0.857472	22/02/2024
2	Haitian Gourde (HTG)	EUR	Haitian Gourde	143.373210	22/02/2024
3	Croatian Kuna (HRK)	EUR	Croatian Kuna	7.608491	22/02/2024
4	Honduran Lempira (HNL)	EUR	Honduran Lempira	26.729297	22/02/2024

Then conduct a descriptive statistical analysis of the data, calculating descriptive statistical indicators using defined functions to describe stats. And use Unique Values to count the number of unique values for non-numeric features. It can be seen from the running results that the numerical characteristic of this set of data is the exchange rate. The total number of data is 341,390 records, the average is about 4,765.30, and the standard deviation is about 111,038.40, which shows that the data distribution is highly volatile. Furthermore, the minimum value of this set of data is 2.071254e-05, and the maximum value is 4.881841e+06, indicating the existence of extremely high values. There are 5638 dates in the data, covering a considerable period. At the same time, the data contains 171 currency codes and 169 currency names. The number of currency names is close to the number of currency codes. This shows that there may be a few currency names that do not completely correspond to the codes. Fig. 1 (The structure of the autoencoder, Picture credit: Original) clearly and intuitively reflects the general trend of the fluctuations of these five currencies (Australian Dollar, Japanese Yen, Canadian Dollar, Great Britain Pound, Schweizer Franken) in the ten years from 2004 to 2024.

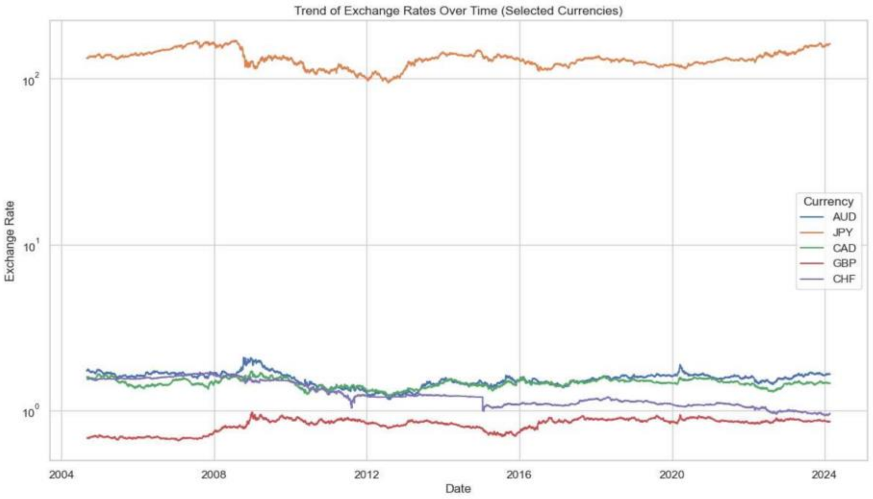

Fig. 1. Trend of Exchange Rates Over Time (Selected Currencies) (Photo/Picture credit: Original).

3.2 Data Pre-processing

Before forecasting future exchange rates, it is necessary to ensure that the data is clean and suitable for modeling. Data cleaning includes four steps. 1. To handle missing values. The research is supposed to check whether the data has missing values. If there are missing values, to decide whether to delete or fill them is needed. 2. To handle outliers requires identifying and handling possible outliers, because the presence of outliers may hurt the prediction performance of the model. 3. Data type conversion requires ensuring that the type of each column of data is suitable for subsequent analysis. For example, date columns should be converted to date classes. 4. The final data filter is used to determine the relevant records that need to be filtered from the entire existing data set when making a forecast for a specific currency. After checking according to these four steps, it was found that there are 79 missing values in the currency name column. However, since the focus of this study is to predict FX rates, choosing to ignore these missing values is a good choice. In terms of exchange rate, it is observed through descriptive statistics and distribution charts that there are extreme values in the data. This is common in financial data, so no processing is done. Making short-term forecasts is a common and practical need in finance. For a specific currency selection, the availability and completeness of data, the attention and liquidity of the market, and volatility are important in deciding what to forecast. Taking the above considerations into account, running the code shows that the Australian dollar has the most data points and more obvious fluctuations, which can provide more opportunities for short-term predictions. In the data preparation process, focus on the exchange rate data of AUD and prepare time series data to adapt to the model. At the same time, based on the time series data, some derived features, such as lag features, need to be generated to help the model capture time dependence. In this set of data, the Australian dollar (AUD) has the most data points so choose it for short-term predictive analysis. This essay will use the data from the last 30 days as the test set to evaluate the predictive accuracy of the model so that even if it uses historical data to predict future exchange rate changes, the model can still simulate real situations. This training set includes 5085 records, and the test includes 31 records. Next, feature engineering will be performed based on the selection, creating some derived features, such as lag features (including the exchange rate of the previous day and the exchange rate of the previous seven days), Table 2 is the specific expression of the hysteresis characteristics (The structure of the autoencoder). And finally preparing to select an appropriate machine learning model for training.

Table 2. Lagging characteristics of Australian dollar exchange rate data

Number	Currency	Base Currency	Exchange Rate	Date	Lag (1 day)	Lag (1week)
341364	AUD	EUR	1.7487	08/09/2004	1.7430	1.7237
341356	AUD	EUR	1.7719	09/09/2004	1.7487	1.7296
341351	AUD	EUR	1.7616	10/09/2004	1.7719	1.7315
341346	AUD	EUR	1.7618	13/09/2004	1.7616	1.7474
341340	AUD	EUR	1.7455	14/09/2004	1.7618	1.7459

3.3 LR Model

Based on data collection and sorting, the LR Model, RF Model, and LSTM Model of machine learning were applied to the sorted data set. The LR model is relatively simple to implement and can provide a baseline performance for future comparison of the performance of more complex models, so choose to use this model first. During testing, this article chose not to separate features and target variables from the test set since the size of the test set is small and used for final validation. Instead, predictions are made directly throughout the test period.

LR is a regression analysis that uses a least squares function called an LR equation to model the relationship between one or more independent variables and a dependent variable. Such a function is a linear combination of one or more model parameters called regression coefficients. It predicts the value of unknown data by using another related known data value. It mathematically models the unknown or dependent variables and the known or independent variables as linear equations. The results predicted using the LR Model are shown in Fig. 2 (The structure of the autoencoder) and the Mean square error is 2.5647167705179283e-05.

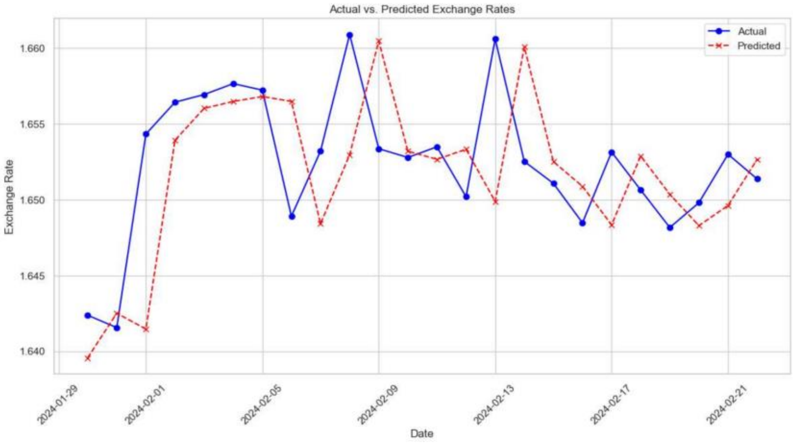

Fig. 2. Use the LR Model to predict outcomes (Picture credit: Original).

3.4 RF Model

The RF Model is a (parallel) ensemble algorithm composed of decision trees. By combining multiple weak classifiers, the final result is voted or averaged, so that the result of the overall model has high accuracy and generalization performance, and also has good stability. The excellent performance of RF is mainly due to Random and Forest. One makes it resistant to over-fitting, and the other makes it more accurate.

Next, apply the RF Model to predict the same set of data. The results are shown in Figure 3 (The structure of the autoencoder) and the mean square error is 3.229817795843547e-05. It is not difficult to see that the mean square error (MSE) of the RF model prediction is about 3.27e-05. Compared with the previous LR model, this value is slightly larger, indicating that on this specific task, the prediction of the RF model is accurate. The performance is slightly lower than that of the LR model.

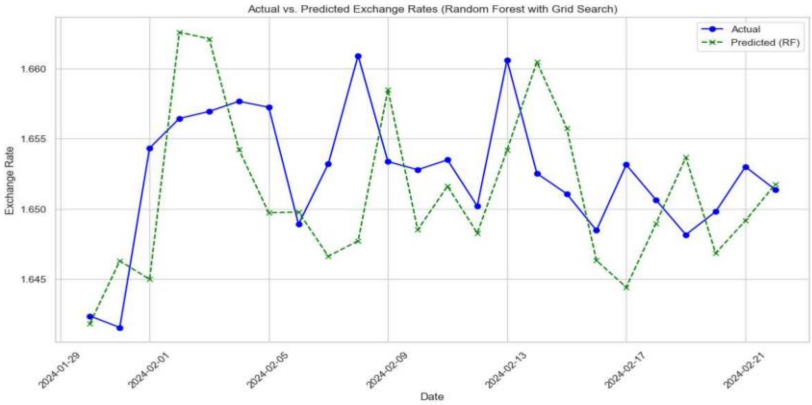

Fig. 3. Use RF Model to predict outcomes (Picture credit: Original)

3.5 LSTM Model

LSTM is a neural recurrent network that solves the vanishing gradient problem in ordinary RNNs through additional units, and input and output gates. Intuitively, the LSTM is a large nonlinear time-series model [9]. Time series forecasting using LSTM networks (LSTM) is a common method in the field of deep learning. LSTM can learn long-term dependencies in time series data and is very suitable for complex sequence forecasting tasks, such as exchange rate forecasting in financial markets. Next, the LSTM model will be used to predict the short-term exchange rate of the Australian dollar (AUD). Before training the LSTM model, some preprocessing steps need to be performed on the data. Since the LSTM model is sensitive to the scale of the input data, it is usually necessary to normalize the data to a smaller range, such as [0, 1], to normalize the data. The LSTM model requires a three-dimensional array as input in deep learning libraries such as Keras, usually in the form of (samples, time steps, and features), so the data also needs to be reshaped. The results are shown in Fig. 4 (The structure of the autoencoder) and the mean square error is 2.540252392699937e-05.

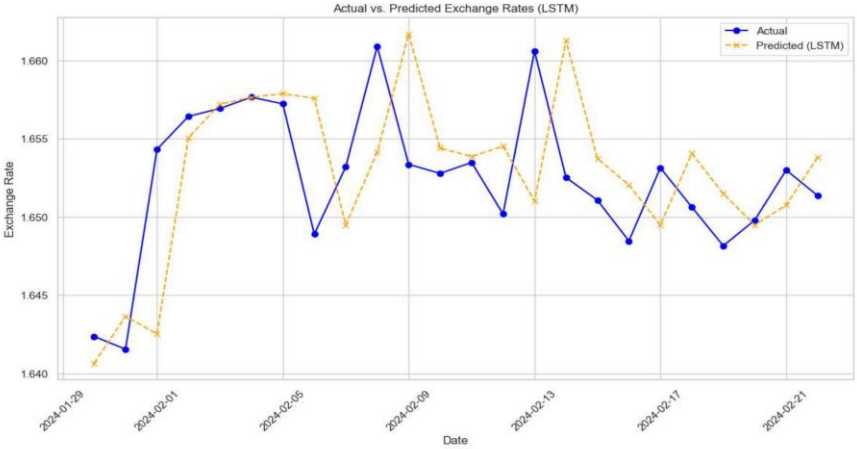

Fig. 3. Use Long Short-term Model to predict outcomes (Picture credit: Original)

4 Results and Discussion

In the task of short-term exchange rate prediction of the Australian dollar (AUD), three different models - LR, RF, and LSTM Model (LSTM) - were compared by running the code. When processing the result data, when the sample size is not large, the evaluation criteria based on small samples are more inclined. In this case, the mean square error will be used for biased estimation. Generally, when the sample size is constant, the indicator used to evaluate the quality of a point estimate θ ˆ is always a function of the distance between the point estimate and the parameter true value θ. The most commonly used function is the square of the distance. Since the estimator θ ˆ has random property, it can be found that the expectation of this function, which is the mean square error given by the following formula:

$$MSE\ (\theta\,\hat{}\,)\ =\ E\ (\theta\,\hat{} - \theta)\ 2 \tag{1}$$

The mean square error is the most general criterion for evaluating point estimates. Naturally, researchers hope that the estimated mean square error is as small as possible. Their mean square errors (MSE) are LR Model: about 2.56e-05, RF Model: about 3.23e-05, LSTM Model: about 2.54e-05.

Comparison based on model performance, these results show that the LR Model and LSTM Model perform equally well and are slightly better than the RF Model. The characteristics of each model are as follows. LR Model showed the best performance very close to the LSTM Model, which indicates that for this specific prediction task, the main trend in the data may be linear, or in other words, the model can capture most of the influence on the exchange rate changing factors. From another perspective, in the problem of predicting FX rates, the LSTM Model still has shortcomings. Compared with traditional RNNs, the LSTM Model has higher computational complexity. Due to the introduction of the gating mechanism and long-term memory mechanism, the LSTM Model requires more parameters and calculation amounts. The complexity of LSTM makes its internal operating mechanism less intuitive and difficult to explain the network's decision-making process, which requires a large amount of data for training. That is, LSTM has more parameters to train, so more data is needed to avoid overfitting. The RF Model is more like a black box model, with relatively poor interpretability and difficulty in explaining the decision-making process of a single tree. At the same time, the computational overhead is relatively large, and building multiple decision trees and integrating their results may require more computing resources. In some cases, the model may also be sensitive to data containing a lot of noise and may even overgrow while the LR Model suffers from several shortcomings when applied to modern data, such as sensitivity to extreme values and cross-correlation (in both the variable domain and the observation domain), and can lead to overfitting. A better solution is to perform a piecewise LR Model, especially for time series data. According to Li Mingchen, the precise forecasting of exchange rates is a crucial aspect of risk management, particularly for businesses involved in global commerce and investment [10]. Therefore, it is important to select a model with high prediction accuracy.

5 Conclusion

This study selected a set of data sets, and after sorting and optimizing the data, intuitive conclusions were drawn. LR Model about 2.56e-05. RF Model is about 3.23e-05 and the LSTM Model: is about 2.54e-05. These results show that the performance of the LR Model and the LSTM Model are equivalent and slightly better than the RF Model. In the comparison between the LSTM Model and the LR Model, the LSTM Model is better. The advantages of the LR Model include its simplicity, high interpretability, and fast training time, making it preferred when resources are limited or fast results are required. As a powerful nonlinear model, the RF Model can usually handle complex data relationships and feature interactions well. In this case, although its MSE is slightly higher, it is still a competitive choice, especially when dealing with possible nonlinear modes. The advantages of the RF model also include resistance to overfitting and the ability to provide feature importance estimates that aid in further analysis. LSTM exhibits similar prediction accuracy to LR, which may indicate that there are certain complex dependencies in the time series that are successfully captured by the LSTM model. Although LSTM is generally more demanding in terms of training time and resource consumption than linear models and RFs, its ability to handle time series data with long-term dependencies may provide additional value for certain application scenarios. In this prediction activity for the Australian dollar, the LSTM Model has the best prediction effect. Given the excellent performance of the LSTM Model and LR Model in this prediction, future research may consider combining the two models with relevant knowledge in the financial field to apply them to different stages of predicting exchange rates, in order to more accurately predict the future trend of exchange rates.

References

1. Das, Pragyan P., Ranjeeta Bisoi, and P. K. Dash.: Data decomposition based fast reduced kernel extreme learning machine for currency exchange rate forecasting and trend analysis. Expert Systems with Applications, 96: 427-449 (2018).
2. Jiemei, L.: FX rate analysis and prediction based on machine learning algorithms. Tsinghua University, (2020).
3. Hossain, Ashrafee T., Abdullah-Al M., and Samir S.: The impact of geopolitical risks on FX markets: evidence from the Russia–Ukraine war. Finance Research Letters, 59: 104750 (2024).
4. Jujie, W, Dong, Y, and Jing, L.: A novel multifactor clustering integration paradigm based on two-stage feature engineering and improved bidirectional deep neural networks for exchange rate forecasting. Digital Signal Processing, 143: 104258 (2023).
5. Ahmed, S.: Artificial intelligence and machine learning in finance: A bibliometric review. Research in International Business and Finance, 61: 101646 (2022).
6. Boyu, Z.: Combination research of machine learning in exchange rate forecast - A comparative analysis based on FA, GARCH and SVM. Zhejiang University, (2021).
7. Yanni, R.: Research on improved machine learning models in FX forecasting. Southwestern University of Finance and Economics, (2022).
8. Kaggle. https://www.kaggle.com/datasets/asaniczka/forex-exchange-rate-since-2004-updated-daily . Forex Exchange Rates Since 2004 (Updated Daily), last accessed 2024/2/21.
9. Ito, K., Hitoshi I., and Yoshihiro K.: LSTM forecasting FX rates using limit order book. Finance research letters, 47: 102517 (2022).
10. Mingchen, L.: Adding double insurance to your investments: Evidence from the exchange rate market. Advanced Engineering Informatics, 60: 102416 (2024).