US Dollar to Singapore Dollar Exchange Rate Prediction

Haonan Yang (hy2805), Yujia Bai (yb2572), Xi Chen (xc2720), Jiayu Chen (jc6116),

Jingqi Wang (jw4491), Jia Wei (jw4466), Yuqi Liu (yl5478), Lei Bao (lb3492)

Zhiliang Ying, Instructor

GR5261 Statistical Methods in Finance

I Introduction & Data Description

With the increasing growth of trade relations between the US and Singapore, it is clear to note that the variations taking place in the respective exchange rates of these two countries have warranted a great deal of attention from the investors due to the pronounced effect they are having on investment strategies and the making of portfolio adjustments. Create a predictive model of exchange rate variations. Dollar (USD) and the Singapore Dollar (SGD) is critical. These datasets have been compiled from sources such as the World Bank, Yahoo Finance, and FRED (Federal Reserve Economic Data), covering a period given to us from the year 2012 through 2022. It includes essential financial indicators like daily exchange rates (USD to SGD, EUR, JPY, CNY), S&P 500, The Straits Times Index (STI), U.S, Dollar Index (DXI), and the macroeconomic variables of the countries, such as Gross Domestic Product (GDP), inflation, and interest rates.

II Objective

This study develops a prediction model that is capable of forecasting and fitting the future exchange rate of the US Dollar (USD) and Singapore Dollar (SGD) against some key economic indicators, which include but are not limited to GDP, interest rates, inflation rates, gold prices, and trade balances. The study uses Long Short-Term Memory (LSTM) models to contribute insights for investors, financial institutions, and policymakers. The focus of the work is to establish the inter-linkages between economic indicators towards developing robustness for a forecasting framework that is adaptive to the dynamism of financial markets.

III Methodologies & Result

1. Data Preparation

In the data cleaning process, missing values and infinite values were removed in data cleaning, and the feature scale was standardized. The dataset was projected to orthogonal components in the process of maximizing the variance, using Principal Component Analysis (PCA). Lasso Regression was used in the feature selection mechanism in order to downsize less important features. Then the feature importance will be identified from Random Forests to give the most important predictors. SP&500, STI, Gold Price, USD to EUR, CNY Exchange Rate are the key features identified that may require further analysis.

2. Sentiment Analysis

Our sentiment analysis uses NLP for the extraction of insights from 250,000 public news items and market opinions with regard to USD and SGD, classifying sentiments either as positive, negative, or remaining neutral in order to enhance exchange rate forecasting. On our part, the VADER library is used to score the sentiment and weight the effects of each tweet in the sentiment analysis. We used the Loughran-McDonald dictionary and a FinBERT model for the analysis of the specialized financial text. (Xiao and Ihnaini, 2023)

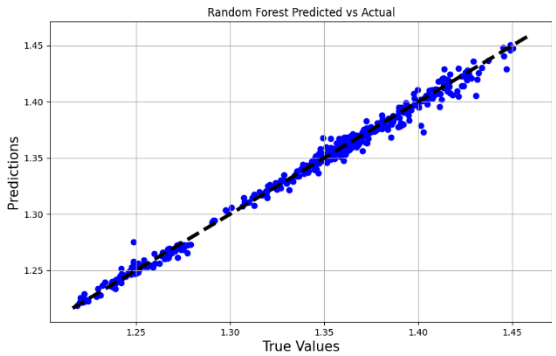
This scatter plot graph depicts a series of part-time currency impact scores sentiment analysis for USD and SGD news carried out from January 2019 through to July 2022. The value on the y-axis, closer to 1, is good for the exchange rate, while those closer to -1 mean bad. A value of 0 is indicative of news without effect. From the chart, it can be seen that USD in blue is constant, hence there is a more bullish sentiment. On the flip side, SGD in orange reflects a broader range of good and bad sentiments; hence, there is more volatility. The dataset includes about 12,000 pieces of news, all in favor of USD, and about 9,000 not in favor, reflecting the pattern of the relative stable attitude in comparison to USD and the more notable fluctuation for SGD.

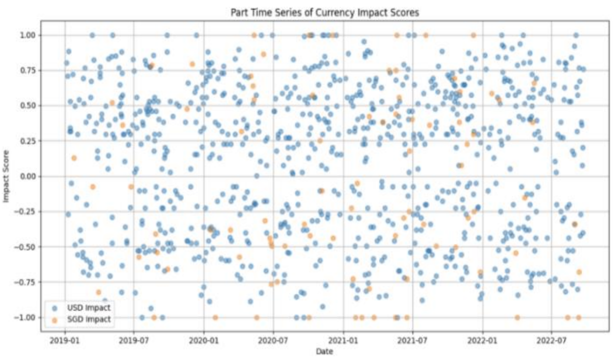
3. Random Forest Regression:

In this section, We use the feature that selected in previous. We can quantify its level of importance through the mean decrease in impurity estimation of a feature's ability to decrease the uncertainty in the data, normally computed using Gini impurity or Entropy in classification and variance reduction in regression.

We use the formula Return = $\frac{\text{Previous Value}}{\text{Current Value}-\text{Previous Value}}$ to calculate the percentage change in various columns that are likely related to financial metrics, such as GDP figures, interest rates, export and import values, and prices of different commodities. Other additive models include the Random Forest Regression Model, which forms predictions as a mean of the predictions from the collection of base models and is used in our modeling exercise and could be formally described as: $g(x) = f_0(x) + f_1(x) + f_2(x) + f_3(x) + \ldots$ w here the final model g is the sum of simple base models f_i and a base classifier is a simple decision tree.

The scatter plot of Random Forest Predicted vs. Actual shows a very close alignment of the USD/SGD exchange rates predictions with the actual rates. It shows high model accuracy, as can be seen by the best-fit line that is constant and accurate. The RandomForestRegressor model is a low mean-squared error (MSE) at about 0.000614, explaining the variance of approximately 99.20% to affirm that the model is good. The feature importance analysis shows the dominance of DXI over influencing exchange rate prediction against other variables like Json, STI, and Gold Price. We further established that the inordinacy of multicollinearity was not present in feature selection since the linear regression we used also supports the validation of a model in a robust manner through cross-validation.





4. LSTM theory

As a form of recurrent neural network, LSTMs have the potential of capturing long-term dependencies taking place in sequence data. Thus, LSTMs are ideal models for prediction from financial time series of exchange rates. Unlike their traditional RNN counterparts, which were troubled by the vanishing gradient problem, LSTMs come with gates controlling the information flow, selectively keeping or throwing away the information, so as to allow better prediction in complex, non-linear settings of the financial system (Jie Du et al., 2020).

1. Forget Gate (f_t): Decides what information is discarded from the cell state. It looks at the previous hidden state (h_{t-1}) and the current input (x_t), and outputs a number between 0 and 1 for each number in the cell state (C_{t-1}). 1 means "completely keep this" while 0 means "completely get rid of this."

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

2. Input Gate (i_t): Decides what new information is added to the cell state. It operates in two parts: a sigmoid layer which decides which values to update, and a tanh layer which creates a vector of new candidate values, $\widetilde{C_t}$, that could be added to the state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

3. Update to Cell State (C_t): Combines the old state (C_{t-1}), what we decided to forget earlier, with the new candidate values, scaled by how much we decided to update each state value.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

4. Output Gate (o_t): Decides what the next hidden state (h_t) should be. The hidden state contains information about previous inputs. The hidden state is also used for predictions. The output gate looks at the current input (x_t) and the current cell state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

We settled on the following six: the GDP in billions of USD, the S&P 500 index, the Dollar Index (DXI), the Straits Times Index (STI), the price of gold, short moving average, long moving average, USD to JPY exchange rate, and USD to CNY exchange rate—all supported by our specially created 'sentiment.'. This is a broader view and brings to considerations different features of economic conditions that influence the exchange rate.

5. LSTM Model Construction and Training

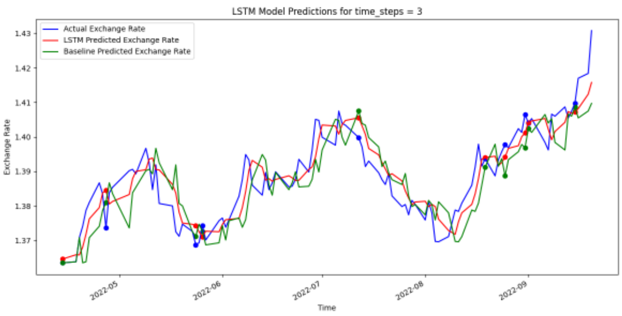
At the model preparation phase, the input features are normalized in such a way to bring consistency in their invariance across different scales, which will make the training unhelpful. We will use the `MinMaxScaler` from `sklearn.preprocessing`to scale each feature into range [-1, 1]. To capture market dynamics in their appropriate representation, our training process uniquely incorporates the approach of a T-3 rolling window. This is whereby the model is retrained every time with the data of three days previous to each target day. Such techniques make the model adaptive to learning from current trends, hence ensuring relevant temporal context for every prediction made to inform those predictions. Such a setup wouldn't just allow better prediction accuracy; the model could also generalize across different market conditions as it would be trained on the latest available data.

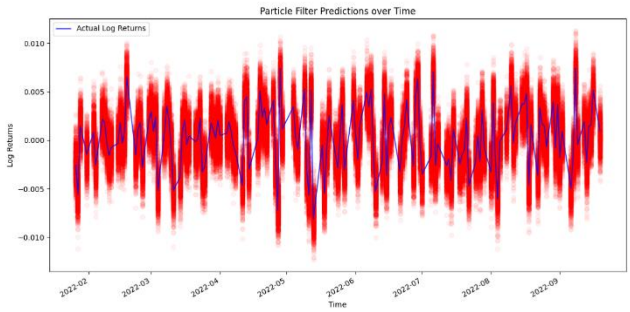
The architecture is advanced and has been developed specifically for the use in analysis of financial time series, represented by the configuration of several 200-unit LSTM layers, with the use of L2 regularization to evade overfitting and, at the same time, to maintain long-term dependencies within data, which are important for proper prediction. So, the dropout layers with a dropout rate of 0.3 are placed between LSTM layers to allow strong generalization not depending on any single data point. The idea of attention is integrated to allow the model to focus on important parts of the data—very useful for financial statements where the recent event might become significant moving forward. Furthermore, bidirectional LSTM layers process data in both a forward and backward direction in an input sequence; thus, with this, they enhance the model's input by adding knowledge from the past and future context of each bidirectional LSTM layer. It increased the ability to find more complex patterns that could have been unnoticed when treated in a one-way scanning loop.

The model effectively tested training-validation splits through a 90/10 ratio and uses EarlyStopping to ensure that training is stopped at the right time without overfitting. The low MSE, RMSE, and high R-squared statistics are measures of performance, as they display the model may have good prediction skills for the USD/SGD exchange rates.

Where our LSTM model was rigorously tested against actual exchange rate predictive performance and a T-3 baseline, which in real terms can be thought of as three days prior using data as a simplistic predictive model. On a test instance, let's say on 25th August 2022, the exchange rate on that particular instance was 1.3978, while the LSTM model made a prediction of 1.3941449, and the baseline model predicted 1. From this comparison, it can be deduced that the LSTM model is closely aligned to the actual rates, a characteristic not shared with the baseline model.

The average MAE between the LSTM predictions and actual values was then computed at 0.004454995089310877, below baseline(T-3) with a great margin of 0.0062259615384615465. This, means that the MAE of the LSTM model will be less than the MAE of the baseline model; hence, acting to underline further the superior accuracy and hence the ability to model the complexities of exchange rate movements effectively. T-3 data, being one of the robust frameworks used, supports and assists LSTM to draw historical trends into the framework for prediction with increased accuracy.





Particle Filter Predictions and Market Dynamics Visualization. This graph demonstrates the predictions made by the particle filter using a T-3 rolling window approach. In this approach, features are transformed into log returns, which serve as inputs for predicting exchange rate movements. The predictions are visualized as red dots (representing the predicted spread) contrasted against the actual log returns, which are depicted as a blue line.Trend Capture: Particle filter should be able to follow actual log returns well, so the model should follow underlying market trends accurately in view of the log-return transformation of features.

Among other, log returns are also involved in financial time series analysis operations. Log returns are especially relevant for models like the particle filter that predict future market movements based on historical data. In the presented scenario, log returns of each of the selected features are calculated using mathematical formulas that measure relative changes from periods to periods. The calculation is performed directly in the pre-processing step of the data pipeline. For every point in time t, the code loops through all records, calculating log returns using the formula, and then creates a new column in the dataset to store the results. Later, those transformed data frames are being used as input to the rolling window of the particle filter as follows:

For Incorporation into Particle filter, Rolling window is T-3. The particle filter is designed in a way that log returns of 3 previous time periods are used to predict the next time period. Thus, for every t, it considers. We slides a window of 3 series over the calculated Log Return frame and inserts them into the function algorithm as its input. Prediction done, the window slides one more to later 3 records.

This figure effectively demonstrates the effectiveness of particle filters in time series forecasting, especially in capturing and reacting to market fluctuations. By looking at the density and spread of the forecast points, we can assess the confidence level of the model and the reliability of the forecasts.

IV Conclusion

In conclusion, this study represents a remarkable achievement in the forecasting of USD/SGD exchange rates through advanced machine learning techniques. The most important goal of ours was to create a forecasting model that could encompass multiple economic indicators and integrate the sentiment extracted from news data around the globe to produce valuable information for investors, financial institutions, and policymakers. The steps that have been taken, such as PCA, feature importance, sentiment analysis through the VADER and FinBERT model, have all led to a sound model. The utilization of LSTM models has played a key role in it since LSTMs are capable of capturing long-term dependencies in time-series data, which is crucial for predicting financial markets as they are inherently volatile. The information we can extract from our LSTM model combined with sentiment analysis illustrates how the interactions between different economic stimuli and the sentiment from news channels are dynamic. The rolling windows approach and log return were effective methods in refining our predictions, and furthermore, our model's predictive power measured by MSE, MAE and R squared was significantly high and much more than what a traditional model or a straight baseline can produce. Future studies can add to this work and utilize that as a foundation for real-time data streams and identify more predictors that can have an impact on the exchange rates.

In conclusion, this study is a very important step in our comprehension of the factors that drive the exchange rates between the US and Singapore, as well as predictive methodologies for financial forecasting. These methods can be continuously perfected to allow financial arsenal to effectively navigate in this highly complex systems.

Reference

Xiao, Q., & Ihnaini, B. (2023, March 20). *Stock trend prediction using sentiment analysis*. PeerJ. Computer science. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10403218/

Random Forest regression. Random Forest Regression · GitBook. (n.d.). https://apple.github.io/turicreate/docs/userguide/supervised-learning/random_forest_regression.html

Jie Du, Yingying Cheng, Quan Zhou, Jiaming Zhang, Xiaoyong Zhang, & Gang Li. (2020). Power Load Forecasting Using BiLSTM-Attention. IOP Conf. Ser.: Earth Environ. Sci. 440 03211.