HDSC SPRING' 23 Premiere Project Presentation

BY

Team TENSORFLOW

TIME SERIES ANALYSIS OF FOREIGN EXCHANGE RATES OF CHINA-YUAN/US DOLLARS (2000 – 2019)

Project Description

The climbing rates of international trade and financial developments have been pegged to the exchange rate for some time now. In developing countries, the input structure of production depends on imported capital and intermediate goods, so an increase in exchange rates makes import production inputs more expensive and thus negatively affects economic growth. So, we will be analysing the foreign exchange rates of different currencies around the world compared to the US Dollars. (i.e. If the China-Yuan is given as 6.08/US$ at a given period; This denotes 1US$ is equivalent to 6.08 US$ at the said period).

Aims and Objectives

The aim of this project is to identify the nature of data on the foreign exchange rates of China-Yuan/US Dollars over a given period of time by a sequence of observations and to forecast/predict future rates of the currency.

Objectives of this project

1. To identify underlying patterns such as trend, seasonality, and cyclicity that may exist in the data.
2. To forecast/predict future value rates of China-Yuan/US Dollars.
3. To compare the actual rates of this currency with the forecasted/predicted/expected value rates.

Flow process

The steps taken in the project is illustrated in the flowchart below;

Data Gathering

The dataset was obtained from Kaggle (via the link).

Data Preparation

The following process was used in preparing the data:

1. Data Collection: The data collected was a structured data which consists of 5217 rows and 23 columns. The dataset contains foreign exchange rates from 22 different currencies across the world which was gathered on a daily basis between “2000-01-03” and “2019-12-31”.
2. Data Transformation: The data collected have some missing values named ‘ND’ that needs to be filled before proceeding to data analysis. Although there are different methods of handling missing data however the best method to handle the missing data, in this case, is to fill with the preceding values using the pandas ‘ffill’ code.
3. Data Visualization: The trend of the data was shown to understand the pattern or direction the series exhibits using line plot. Also, boxplot was used to check for outliers in the data

Exploratory Data Analysis (EDA)

EDA is applied to investigate the data and summarize the key insights. It helps to give basic information and an understanding of the data. The cleaned data contains no duplicates and has no missing value. The minimum rate of China-Yuan/US$ was recorded to be 6.0402 while the maximum rate was 8.28. This means that the US Dollar has been a valuable currency compared to the China-Yuan. Furthermore, the mean was recorded to be 7.199286 China-Yuan/US$ which means on average 1 US Dollar is equivalent to 7.199286 China-Yuan.

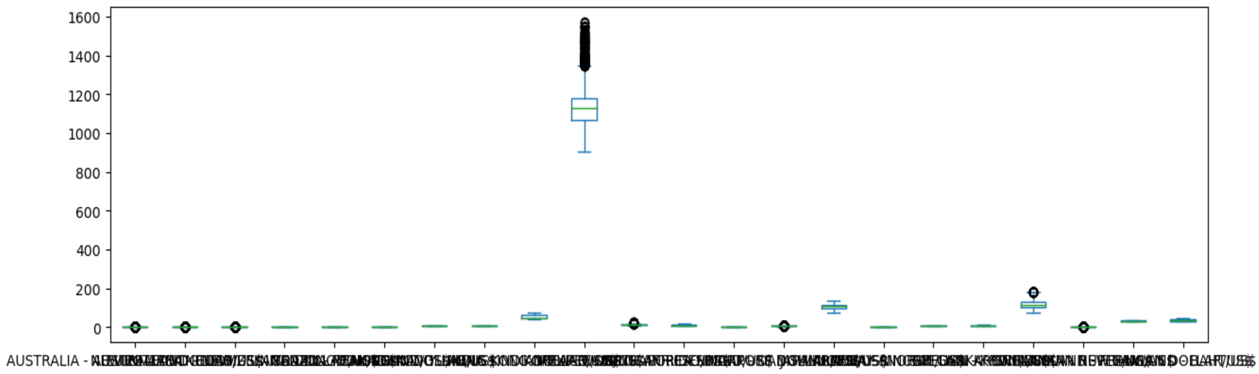

Figure 1: Boxplot to show if outliers are present in the data

From the above plot, we can see the data contains no outlier, so we can proceed with analysis

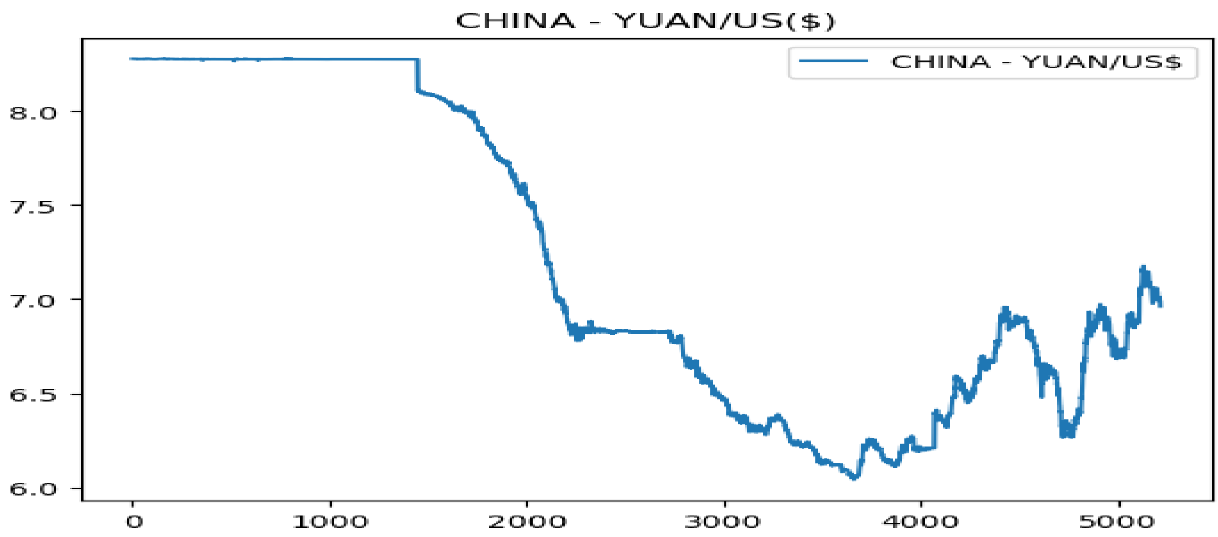

Figure 2: Trend of China-Yuan/US$

Figure 2 above shows the pattern of the time series data over time.

Model Training, Validation, and Evaluation

The dataset was prepared for time series analysis and was divided into training and testing datasets. Baseline Mean Absolute Error (MAE) was calculated to give an insight into what the model's accuracy should be at max. Linear Regression, Auto-Regressive model, and ARIMA model were used to train the dataset and their performance was assessed using evaluation metrics such as mean absolute error (MAE), root mean square error (RMSE), residual sum of square (RSS), and r2_score.

Linear Regression Model

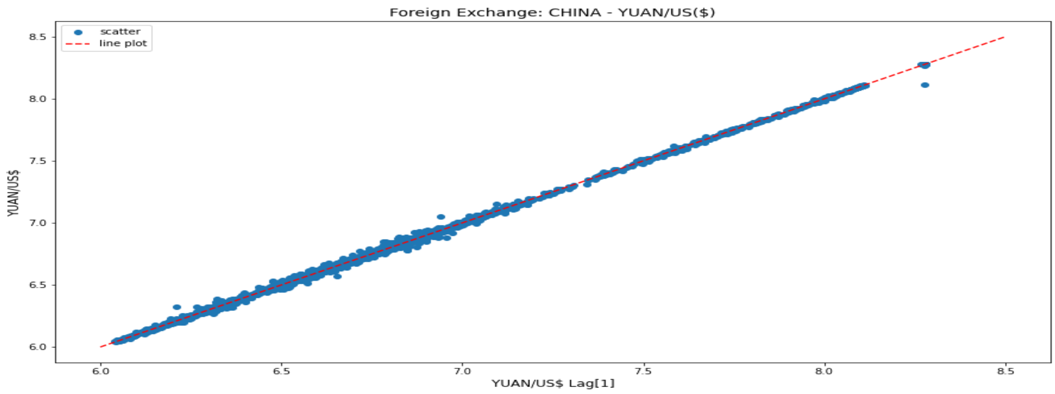

Figure 3: Line plot of China-Yuan/US$

In figure 3 above, the line plot was used to check if the relationship between the variables is linear such that a straight line is the line of best fit. This shows that linear regression can be used for the time series analysis. Linear regression was used to train the dataset and the performance was evaluated.

Auto-Regressive model

For AR model to be used, the lag value (p) must be gotten using autocorrelation and/or partial autocorrelation plot. From the acf and pacf plot below (figure 4 and 5) a lag value (p) is estimated to be 35.

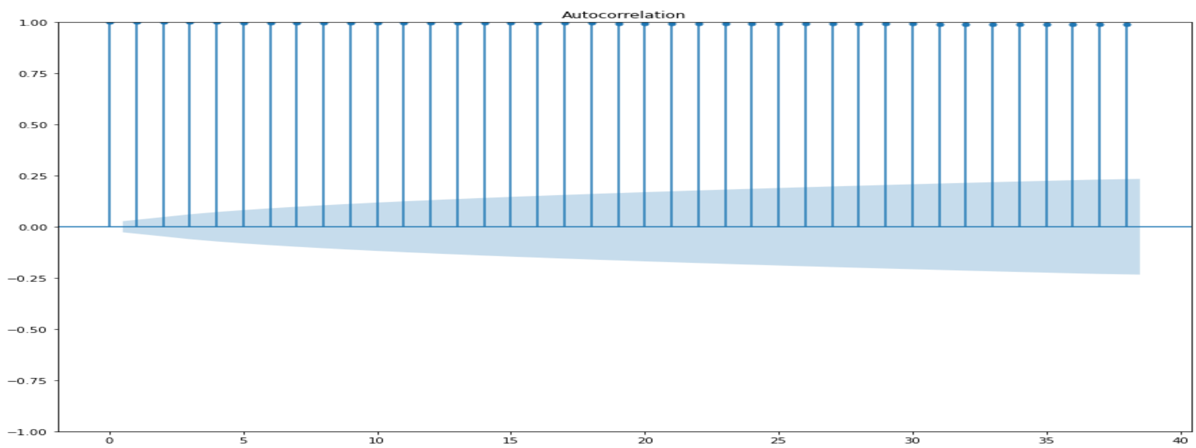

Figure 4: Autocorrelation function (acf) plot of China-Yuan/US$

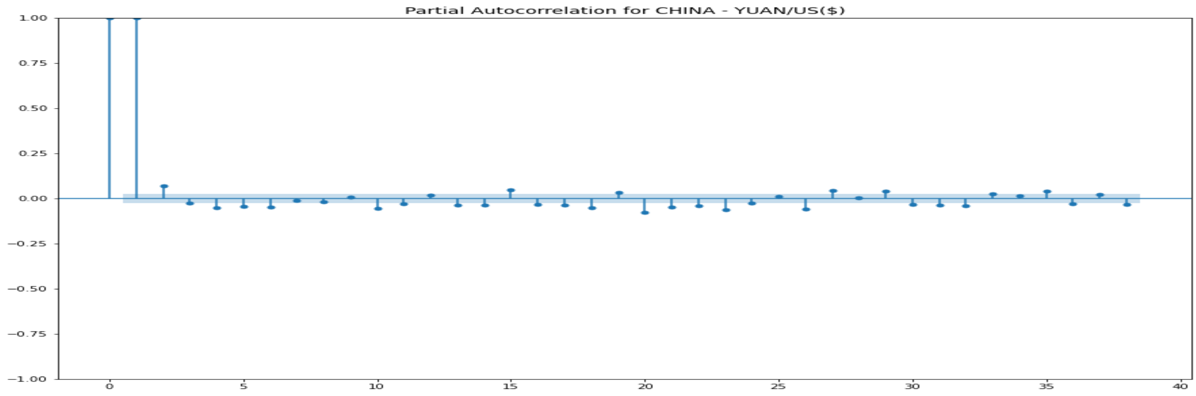

Figure 5: Partial autocorrelation function (pacf) plot of China-Yuan/US$

After training the model, the model was evaluated however the MAE was gotten to be 0.3750 i.e. there is about 38% error in prediction. This shows that the model did not perform well. Therefore, hyper-parameter tuning was used to select a lag 'p' over a range of 1 to 38. The lag value with the least training MAE was selected for walk-forward validation and the model was evaluated. A MAE value of approximately 0.01 was gotten which means our prediction is about 99% accurate.

ARIMA model

The ARIMA model was used to train the data and the trend, seasonality, residuals, and plot diagnostics was plotted as shown below.

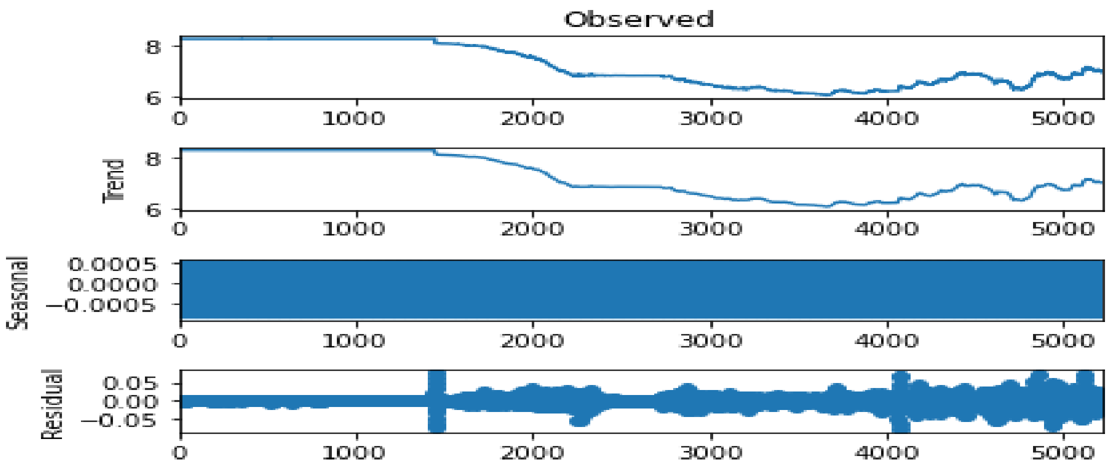

Figure 6: Trend, Seasonal, and Residual Plot of the data

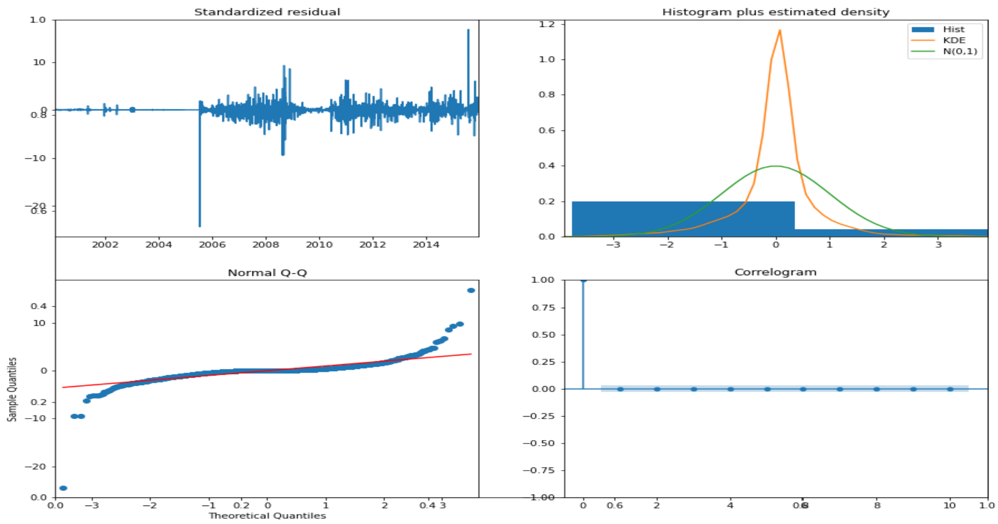

Figure 7: Standardized residual, Histogram and Correlogram plots

The standardized residual plot and the histogram plot above confirms the stationarity and normal distribution of the model. Also the correlogram shows that the lag values (p, 0, q) at (5, 0, 0) have no predictive strength. Walk forward validation was done on the model and the performance was assessed.

Model Inference and Deployment

Models	Evaluation metrics			
	MAE	RSS	RMSE	r-squared
Linear model	0.010836	0.27367	0.016191	0.994079
AR model (with walk-forward Validation)	0.01098	0.274814	0.016224	0.994054
ARIMA (with walk-forward Validation)	0.010807	0.273345	0.016181	0.994086

The table above shows the performance of the three models. In terms of performance (MAE), ARIMA model has the best performance although the differences between it MAE value and linear regression is minimal (which can be said to be negligible). However, because ARIMA model has a high run time, linear regression is selected for deployment (for faster prediction). For reusability and extensive use for several currencies in the foreign exchange dataset, a make_prediction function was created to forecast the future values of the currency.

In addition, since linear regression equation is

$$y = \theta_0 x + \theta_1$$

Where θ_0 = coefficient or slope of the line of best fit

θ_1 = intercept

x = independent variable

Therefore, a linear regression equation for our model's prediction is

$$Yuan/US(\$) = 0.999931(Yuan/US(\$) - L_1) + 7.6e^{-5}$$

Where, $Yuan/US(\$)$ = Yuan/US$ to be predicted,

$Yuan/US(\$) - L_1$ = Yuan/US$ rate of the immediate previous date.

The codes used in Explanatory Data Analysis, model training, validation and evaluation can be assessed on the GitHub repository link below.

GitHub: https://github.com/Dotzyman/Tensorflow-Capstone

Conclusion and Recommendation

In conclusion, linear regression model has a short run time with good performance metrics and was selected to predict the exchange rate value of any currency in our data using the make_prediction function. However, certain factors—such as political events, economic issues, inflation, market sentiment, and even war—influences foreign exchange rates which might cause a rate to suddenly rise or fall thereby affecting the accuracy of the model's projection. Therefore, we recommend that these factors should also be taking into consideration even after the model's prediction. We also recommend that data should be gotten from reliable source because discrepancies and inaccuracies may impair model's prediction accuracy.