Technical Report on Premier Project by Group NLP in Hamoye Internship HDSC Winter '22

Introduction

Foreign exchange rate, or FOREX rate, is the rate at which a particular currency exchanges with another. In simpler terms, it is the value of one currency in terms of another. The foreign exchange market determines FOREX rates which often fluctuates based on transactions between traders. For instance, the foreign exchange rate of Naira to Dollar is 415.76. This implies that it takes 415.76 Naira to buy a Dollar.

As we must have heard, data is the new oil. Data science technology is a discipline that works on datasets using different methods and processes to extract useful insights. It helps to identify and study the patterns of occurrence of particular subjects of study. As of now, data science has transformed the way professionals of different fields make observations and decisions. What better way to take action if not based on trends and measurable insights.

For this project, the NLP team of the HDSC Winter '22 used a raw dataset containing foreign exchange rates information of about 21 countries. The dataset provided the FOREX rates of 20 years spanning from 2000 to 2009. The team used the information from this data to understand the relationships between FOREX rates trends and predict changes across selected countries.

Methods of Analysis

Most industry observers report that data preparation steps for business analysis or machine learning consume 70 to 80% of the time spent by data scientists and analysts. The data preparation pipeline for the FOREX rates datasets consists of the following steps:

1. **Extract and load the data**

   The data for our project group is titled: PP22/T616 on Hamoye's medium page with a scope of Trade, Agriculture and Finance. It falls under the topic "Foreign Exchange Rates". The dataset for this project was accessed from the kaggle directory. Jupyter Notebook was used as the primary tool for carrying out data cleaning and analysis. On the other hand, Python was utilized as the choice of querying language. All relevant libraries from the Python machine learning libraries were successfully imported. Then, we loaded our dataset which was already downloaded to a local directory in Comma Separated Value (.csv) file format.

2. **Date transformation**

   Data cleaning is the process of ensuring data is correct, consistent and usable. You can clean data by identifying errors or corruptions, correcting or deleting them, or manually processing data as needed to prevent the same errors from occurring.
   In cleaning our data, we queried the dataset to know more about it using the info() method. By doing so, it was observed that the dataset contained 5217 rows (data ranging from index 0 to 5216) and 24 columns containing the various countries we are to study. It was also observed that the Time series column located at index 1 was in an object data type. Hence, it was pertinent to convert it into a datetime format to make it usable.

   Likewise, other columns ranging from index 2 to 23 which were originally in object format were converted into float data type. The data unit (index 0) which had “unnamed" as a column name was dropped from the dataset using the drop method. This was done because it was a replica of our index column. This process reduced our dataset from its original 24 columns to 23 columns.

   Using the isnull() method, the total number of missing values was accounted for. These missing values ranged from 197 to 201 in different country columns. Hence, it is knowledgeable to drop such values using the dropna() function. Dropping these missing values leaves the dataset with only valid and significant values. After following these steps judiciously, the total dataset accounted for a new value of 5015 rows and 23 columns. Upon completing the cleaning process, the dataset became ready for exploration and further analyses.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) techniques continue to be a widely used method in the data discovery process today. It helps data scientists to perform initial investigations on a provided data so as to discover patterns, spot anomalies, and check assumptions using summary statistics and graphical representations. Data scientists can use exploratory analysis to ensure the results produced are valid and applicable to any desired business outcomes and goals.

Upon completing an EDA and drawing insights, one can use its features for more sophisticated data analysis or modeling. There are a number of tools used for performing an EDA. However, Python is the preferred choice for this project. Python is an object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and libraries, makes it very attractive for rapid application development and model training and building.

Sequential Process for Performing Data Visualization on the Provided Dataset

• Data conversion from different currencies to US Dollars (USD). This is achieved by taking the inverse of each provided value. To evaluate the trend between currencies, it is important to have all currencies in the same format.

• Check the average value of each currency in USD for the four quarters within the years provided. This gives further insight into the annual changes occurring for each currency on a quarterly basis.

• Check for noticeable trends within the provided dataset. Some resulting trends indicated that the Euro has been appreciating until 2015 when it depreciated from 0.75 to 0.94. However, it appreciated again in 2019 to a value of 0.89. The United Kingdom's Pounds appreciated from 2000 to 2009, although it has been depreciating in respect to USD till 2019. New Zealand has been doing well, as it maintained a high value from 2000 to 2003.

• Generate a plot to determine the country with the highest currency/USD. From the generated plot, it can be deduced that the United Kingdom's Pounds are the highest currency value relative to US Dollars.

• Also, generate a plot to determine the country with the lowest currency/USD. From the generated plot, it can be observed that Korean Won is the lowest currency value relative to US Dollars.

• Check the relationship between the lowest and highest trending currencies (i.e. United Kingdom's Pounds and Korean Won).

• Observe each currency's improvement in US Dollars. From the diagram, positive growth in currencies is depicted in "Blue", negative growth in currencies is depicted in "Red".

• Identify trending currencies according to the last date of 2019 provided in the dataset. Doing this shows the strongest to the weakest currencies as provided in the dataset.

• Generate a correlation heat map plot for the currencies. A heat map is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions. From the heat map generated, an observation which was accounted for was that of the Euro and Danish Krone currencies having a perfect positive correlation. This is clearly explained in the attached link https://en.wikipedia.org/wiki/Denmark_and_the_euro.

• Use scatter plot to check the correlation between Euro and Denmark currencies. A pure simple regression line is observed from the above plot.

• Generate a plot to show a set of currencies having a high correlation. This is done to show the relationship between the provided currencies which have been presented in dollar formats.

• Obtain the maximum, minimum, and last day price of each currency. Also, indicate the years at which these events occur. This gives insight into significant fiscal highlights which have occurred within the provided data timeframe.

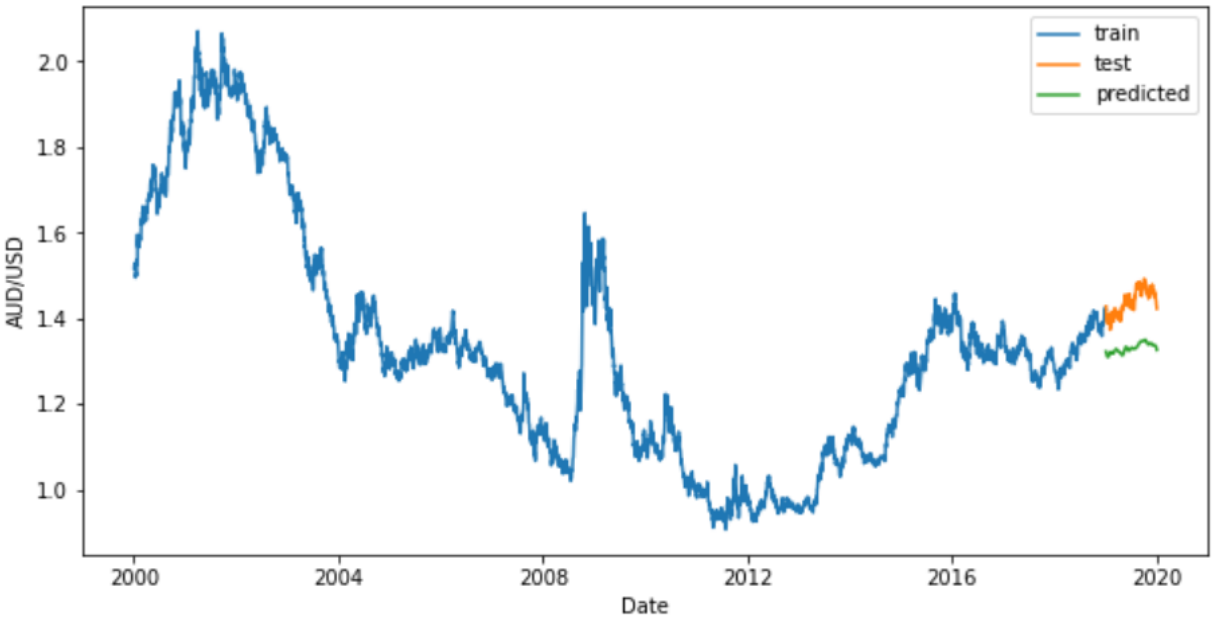
Prediction Models

We chose to work with 4 selected currencies based on the most traded currencies data from investopedia. Chosen countries are United Kingdom (Pounds), Euro Area (Euro), Australia (AUD), and New zealand (NZD)

Techniques used for the forecasting and model building are the ADFULLER test, autocorrelation plot, and the Prophet module. We chose these libraries because we observed that our datasets were not stationary, and they would be the best option for predictive analysis.

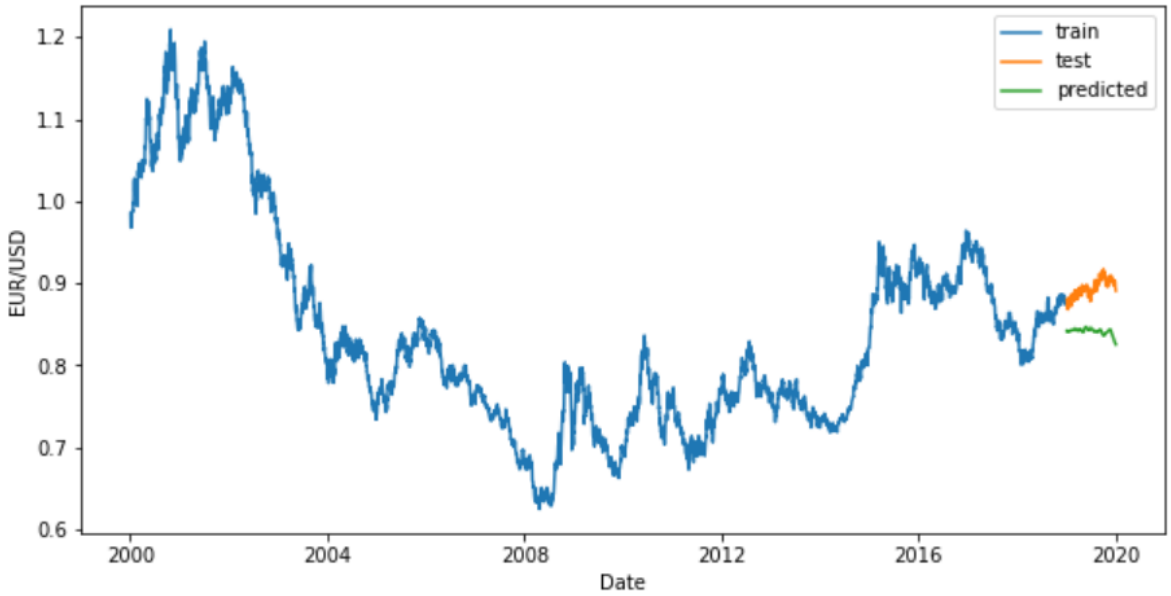
Results and observations

1. Australian Dollar FOREX prediction



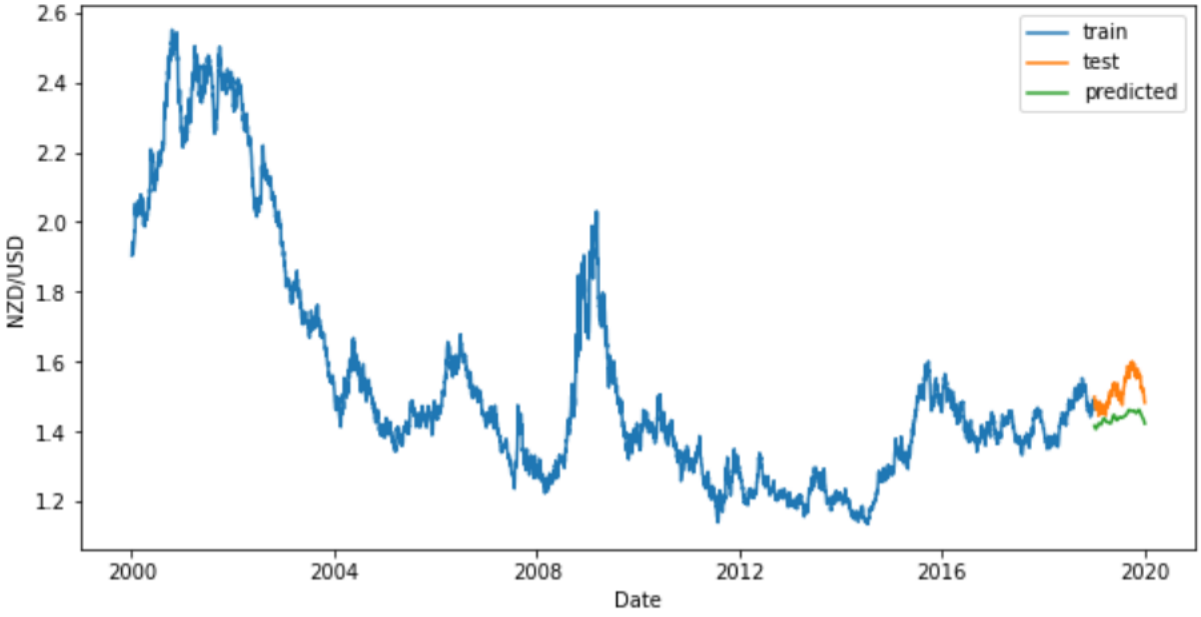

```
The mean absolute error for AUD is : 0.10825145741490602
The mean squared error for AUD is :  0.012149893273436841
```

2. Euro Area FOREX prediction



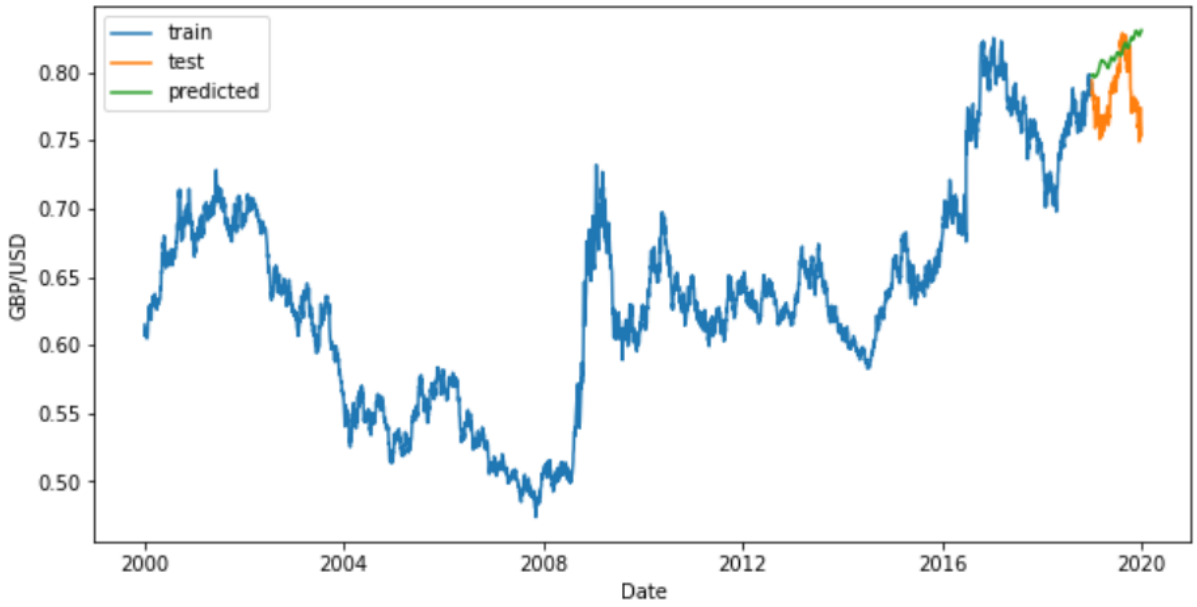

```
The mean absolute error for EUR is : 0.052294241246066145
The mean squared error for EUR is :  0.0028860002268576985
```

3. New zealand Dollar prediction




```
The mean absolute error for NZD is : 0.07874937886778015
The mean squared error for NZD is :  0.0071250916076285985
```

4. United Kindom Pounds FOREX prediction




```
The mean absolute error for GBP is : 0.030984974020624636
The mean squared error for GBP is :  0.001317804282518754
```

Conclusion

The foreign exchange market has a direct impact on wages, cross-border investments, and the economy as a whole. Financial institutions, companies, governments, and other entities use this market to adjust their currency holdings. This indicates the need to obtain accurate evaluations and predictions of market trends. This study proposes a means of utilizing data science technology to evaluate and predict such trends within a country's economy relative to its US dollar equivalent.

NLP group members

Nnona Ebuka John
Aman Vishwakarma
Fatima Alhassan
Sanskriti Jain
Manoj Kumar
Ahmed Anifowoshe
Odubanjo Sulaimon Abiodun
Abiodun Shittu
Gbolahan Alao
Bright Odike
Jeremiah Zhao

Emeke Ihedilionye
Timothy Oladapo
Shereen Onyango
Saheed Olurebi
Oluwaseyi Ogunlana
Olugbuyi Peter
Ola Ade
Nicholas Aniefiok Udoh (APL)
Idara Effiong
Halimat Atanda (QA)