ISSN 2085-4579

Finding Features of Multiple Linear Regression On Currency Exchange Pairs

Raymond Sunardi Oetama[1], Ford Lumban Gaol [2], Benfano Soewito[3], Harco Leslie Hendric Spits Warnars [4]
[1,2,3,4] Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia

raymond.oetama@binus.ac.id
fgaol@binus.edu
bsoewito@binus.edu
spits.hendric@binus.ac.id

Accepted 30 June 2022
Approved 13 July 2022

***Abstract*— Due to the prospects for financial gain, forex is always attractive to many people. However, because forex market analysis is not simple, a computer is needed to assist in creating predictions using features that are understandable to people. This study employs the Multilinear Regression technique to identify these kinds of features. The features and prediction target have a very strong correlation with the lowest RMSE is 0.00408 and the highest R2 is 0.99477, the prediction quality is quite outstanding. The outcome will help academics in the forex field use machine learning algorithms to make better predictions.**

Index Terms—Features; Forex; Machine Learning; Multilinear Regression; Predictions;

I. INTRODUCTION

People trade for money by exchanging products for goods, goods for money, and, more recently, money for money. Forex trading is a type of trading in which one country's currency is swapped for another country's currency [1]. Not all currencies, however, are exchanged. EUR (European Union currency), GBP (British pound sterling), AUD (Australian Dollar), NZD (New Zealand Dollar), USD (United States Dollar), CAD (Canadian Dollar), CHF (Swiss franc), and JPY (Japanese yen) are the most widely traded currencies. Currency pairs are used in forex trading; for example, GBP and USD are coded as GBPUSD, EUR and USD are coded as EURUSD, and so on.

A Currency trader must have the knowledge and analytical capabilities of the transactions that take place in the forex market to generate a profit. If a trader takes a BUY action when the price rises, the trader will profit. Meanwhile, traders who take SELL actions will benefit as prices fall [2].

The currency market, on the other hand, is difficult to analyze [3]. This is since manual analysis is limited, but forex prices move in milliseconds. As a result, computers must assist humans in analyzing forex price fluctuations. Because the parameters utilized in the analysis may be no longer relevant in today's world.

The algorithmic trading approach is one way to undertake forex price analysis with the use of an algorithm [4]. Multiple Linear Regression is one of the algorithms employed in this strategy. Multiple Linear Regression is, in fact, an old algorithm. Even so, its capacity to explain which features influence prediction results, on the other hand, is undeniable [5]. This advantage will be used in this study to identify traits that are extremely useful in decreasing forex price forecast errors.

The opening price, the highest price, the lowest price, and the closing price are the four original prices that are used to express forex price fluctuations in Japanese candlesticks [6]. As a result, this research contributes to the discovery of traits that contribute significantly to the reduction of forex price forecast errors. The closing price is used for prediction in this study since it occurs at the end of each day's transaction.

The purpose of this study is to find features that are highly connected with prediction targets that are relevant to the current situation of the forex market and to reduce mistakes in forex price predictions. Whilst limitation of this study is other pairs that aren't limited to GBPUSD and EURUSD require more investigation. Only applies to feature groups, not to feature merging.

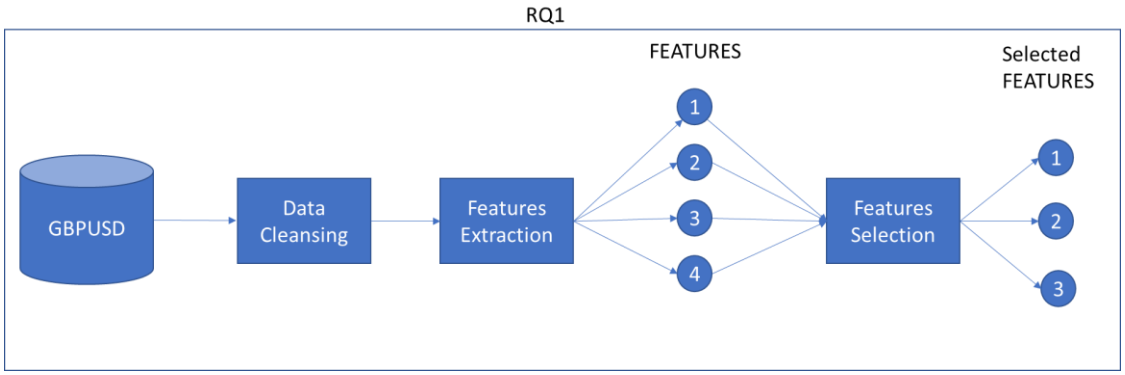


Fig. 1. Data Preparation Process

The purpose of this study is to find features that are highly connected with prediction targets that are relevant to the current situation of the forex market and to reduce mistakes in forex price predictions. Whilst limitation of this study is other pairs that aren't limited to GBPUSD and EURUSD require more investigation. Only applies to feature groups, not to feature merging

II. Literature Study

A. *Multiple Linear Regression (MLR)*

MLR is a method for predicting the dependent variable from a set of independent factors [7]. In this study MLR is applied to calculate today's close price based on some features ft+1, ft+2, ... fi while w0 is the intercept and coefficients wt+1, wt+2, … wi as shown on Equation 1:

$$C_t = w_0 + w_1 f_{t+1} + w_2 f_{t+2} \dots w_i f_{t+i} \quad (1)$$

B. *Correlation*

The term "correlation" refers to the relationship that exists between two variables [8]. The correlation value might be anywhere from -1 to 1. The correlation is plus when the first variable is bigger than the second variable is also bigger. Whilst the negative sign means the opposite way. The correlation is stronger when the coefficient value is higher. Correlation can be calculated using the Equation 2:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (2)$$

MLR target variable close price is referred to as y in this study, and features f_1, f_2, … f_i. are referred to as x.

III. Research Methods

The Forex market is investigated. With a daily transaction of over US Dollars 6.59 trillion [9] the currency market is the largest in the financial world. The FX market is open five days a week, 24 hours a day. Because GBPUSD is a popular currency pair among forex traders, it is the currency pair of choice.

The CRISP-DM method is a data mining approach that is well suited for application in the industrial environment [10] This is evident in the Business Understanding module, where the first utility produced is a business problem of price prediction in the Forex market. It will be a near price forecast in the daily period if it is translated into a data analysis problem. The Data Understanding process begins with data collected from yahoo.finance.com, a popular financial data database supplier. The data was collected four years daily transactions from 2018 to 2021, to ensure that the research findings are still relevant to the status of the currency market.

The stages of the Data Preparation process are shown in Fig. 1 to answer RQ1. The data purification process is carried out at this point to see if any nulls remain. Because null data can cause mistakes in the program's results, it will be removed. Python was utilized in this study, and it is a popular data mining program nowadays. The close price is the variable utilized to calculate the prediction target. Where this price will be projected using a variety of alternative qualities that may be derived from the basic price of forex trading, namely the open, high, low, and close prices that have already occurred, starting with D-1, D-2, and so on. The feature must correlate with the close price of at least 95% to be selected. The goal of choosing a high correlation is to reduce prediction errors caused by mistakes in choosing predictive variables that are less connected with the close price.

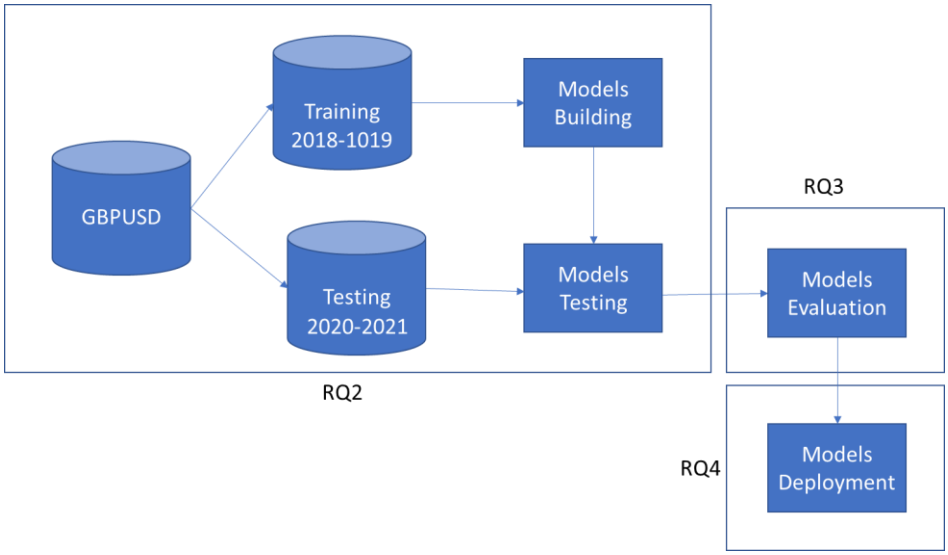


Fig 2. Research Processes From Models Buildings to Models Deployment

The MLR model was utilized. The high level of interpretation is an advantage of this method that is also a concern in this investigation. The MLR approach was chosen because it is more important to get features as well as to understand their contributions to daily close price predictions. After all, this research is still at the beginning of the process of discovering and reviewing features in forex prediction research. The data is split into two portions, training data, and testing data, using a 50:50 split to test the model. GBPUSD data from 2018 to 2019 is part of the training data. GBPUSD data from 2020 to 2021 was used in the testing. The Modeling to Deployment Process is depicted in Fig. 2 to answer RQ2 to RQ4.

The RMSE and R^2 units will be used to compare the prediction quality of each prediction model at the evaluation stage. The RMSE method is used to calculate prediction error in the hopes of reducing the error as much as possible by combining all highly linked information. R^2 is a metric for determining whether the forecast results and factual data are still linear in terms of the MLR algorithm. The higher the amount of linearity, the closer R^2 is to the value of 1. In this study, the optimal model is the one with the minimum RMSE and highest R^2. The best model will be evaluated at the deployment stage by applying it to data for a different currency pair, such as EURUSD. Because it is a widely traded currency pair in the forex market, this currency pair was selected.

Fig 3. Data Preparation Process GBPUSD Chart from 2028 to 2021

IV. Results and Discussion

A. Business Understanding

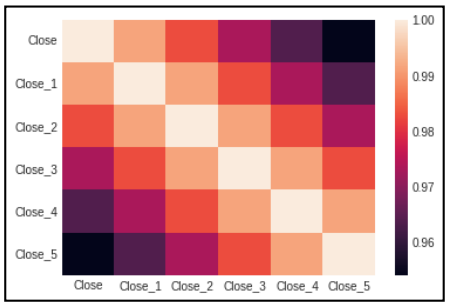
Because forex trading only has four basic prices, the close price will be forecasted utilizing several features that can be derived from these four basic values. From Equation 1, when all features are zero, interception equals the value of the close price. If the features are not all zero, the contribution of each feature to the close price forecast will be visible in their corresponding weights.

B. Data Understanding

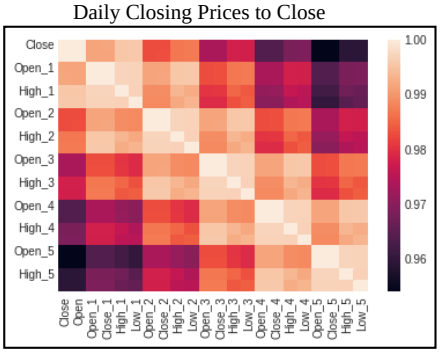
From 2018 through 2021, a popular financial website yahoo.finance.com was used to acquire GBPUSD statistics. As illustrated in Fig. 3, the price of GBPUSD began to rise in early 2018 and then began to fall until the first half of 2020. There was a drop in the first semester of 2020, followed by a strong increase. Of course, this is extremely helpful in determining the accuracy of the MLR prediction model. The data from 2020 to 2021 will be utilized as testing data, while the remaining data will be used as training data. Following that, the GBPUSD trend increased until the middle of 2021, before reversing till the end of 2021.

C. Data Preparation

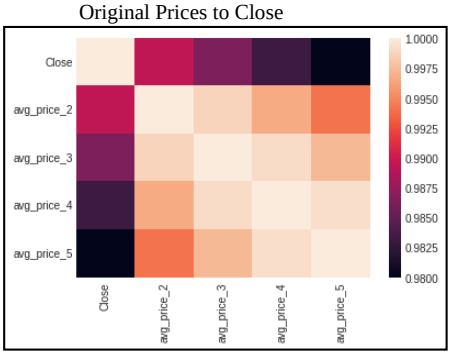
To begin, search for data that is missing. Only on May 22, 2019, did there be no data. To avoid analysis problems, the data was then removed. After that, data is extracted from four different feature categories, including 5 prior daily closing prices, five previous daily original prices, Average prices, and Standard deviation of prices. Firstly, 5 prior daily closing prices. Close 1 (price closed yesterday), Close 2 (price closed two days ago), Close 3 (price closed three days ago), and so on until the fifth day ago are used to create this feature group. The limit has been set to 5 days ago, allowing users to use data for one trade week (five working days). These feature groups were chosen because they were all highly connected with a price that was within 96 percent of one another as shown in Fig. 4. The second feature group is Five previous daily original prices. This feature group is a more thorough version of the previous five daily closing prices feature groups. By including the four essential components of forex trading's basic pricing. Specifically, open 1, high 1, low 1, close 1 on one day ago, open 2, high 2, low 2, close 2 on two days ago, open 3, high 3, low 3, close 3 on three days ago, and so on till five days ago. All these feature groups were chosen because they had a very strong correlation at near prices, with a minimum correlation of 96 percent, as seen in the heatmap data. The next feature group is average prices. This feature group is generated using the average formula, where average 2 is the average close price from one day to two days ago until average 5 is the average close price from one day to five days ago. This feature group was also chosen because it has a very strong association with.



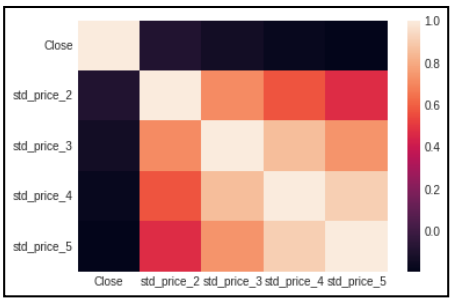
(a) Correlation Heatmap of 5 Previous Daily Closing Prices to Close



(b) Correlation Heatmap of 5 Previous Daily Original Prices to Close



(c) Correlation Heatmap of Average Prices to Close



(d) Correlation Heatmap of Standard Deviation of Prices to Close

Fig 4. Feature Groups Heatmaps

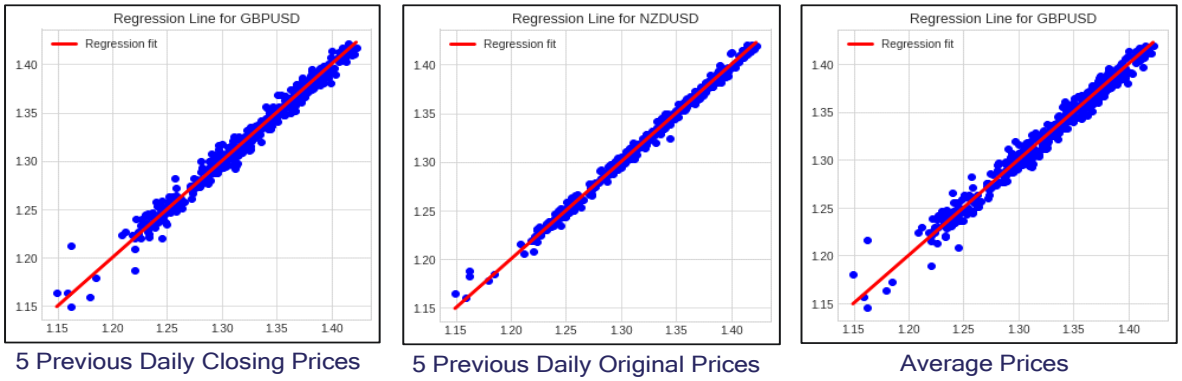


Fig 5. Data Preparation Process

close prices, with a minimum correlation of 98 percent. Because it takes at least 2 days to calculate the average, the Average Prices feature group starts from 2 days ago. The last feature group is the standard deviation of prices. Starting with the standard deviation 2 days ago and ending with the standard deviation 5 days ago.

The standard deviation feature group begins 2 days ago, as the average pricing group, because calculating the standard deviation takes at least 2 days. The standard deviation pricing feature group was not chosen based on the heatmap in Fig. 4, because it correlates with less than 95%.

D. Modeling

The modeling approach begins by dividing the data into training data from 2018 to 2019 and testing data from 2020 to 2021. The attributes that will be used for daily close price forecasts are calculated using training data. 5 prior daily closing prices, 5 previous daily original prices, and average prices are the three features chosen. The MLR Modeling program was then constructed using the Python program, which produces three models with the same name as the feature group. Afterward, the model is evaluated using data and quantified using RMSE and R^2.

E. Model Evaluation

The comparison of the three models as MLR models is shown in Fig. 5, by comparing their regression lines. Because the bulk of the forecast outcomes are quite close to or stick to their respective regression lines, the three models are very good at portraying the regression line. The three models reveal several residuals at the bottom when compared to nutritional volatility in the first half of 2020. The model with the 5 prior daily original prices features group has the best performance with the least amount of residual of these three models. A comparison of RMSE and R^2 will be used to further support these findings. As shown in Fig. 6, the RMSE model 5 previous daily original prices have the best performance, with an RMSE of 0.00408, as compared to the model 5 previous daily closing prices, which has an RMSE of 0.00736 and the Average Prices model, which has an RMSE of 0.00809. This is since the 5 prior daily original pricing model has a lower rate of forecast errors. Next, as shown in Fig. 7, Model 5 previous daily original pricing had the best performance, with R^2 = 0.99477 being the closest to 1. As a result, the best model in this investigation is the model using 5 previous daily original prices.

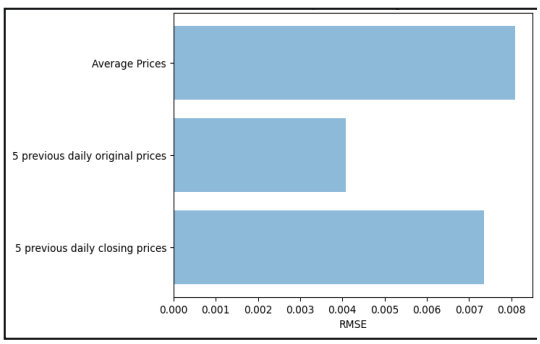


Fig 6. RMSE comparison among Three Feature Groups

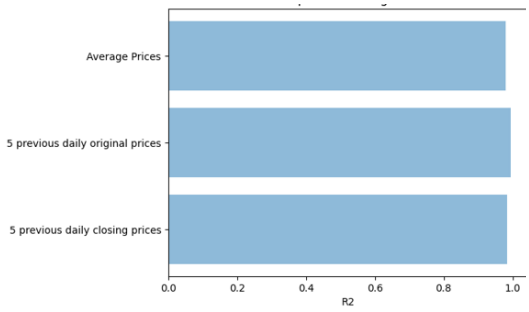
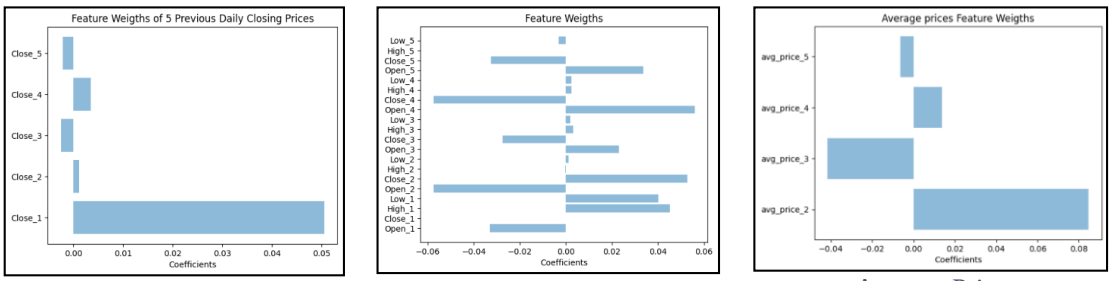


Fig 7. R^2 comparison among Three Feature Groups

able to maintain its shape when price swings occur, as they did in the first half of 2020.



5 Previous Daily Closing Prices 5 Previous Daily Original Prices Average Price

Fig 8. Comparison of the Weights of the Features

The features analysis is then performed to illustrate the benefits and drawbacks of the five prior daily original pricing models as the winning model. Fig. 8 shows the comparison of the weights of the features used in each model's predictions. In comparison to the other two models, the winning model has more features. Although there isn't much of a difference between RMSE and R^2, adding characteristics will improve the prediction ability. Additionally, the winning model's weight distribution is more uniformly distributed, whereas the other models are dominated by attributes that are closer to the close price. It is necessary to have this steadiness to forecast the rise and fall of the GBPUSD price. The winning model's drawback is that, due to the numerous aspects involved, it is less

From Fig. 9, the truth of the GBPUSD closure price is compared to the three models' predictions of the GBPUSD close price. Overall, the truth map from 2020 to 2021 is excellent. Although there may be some variances, the three models can all map extremely well. The model with the 5 previous daily closing prices features group performs best in detecting price variations in the first half of 2020. This is since this model is more flexible because it just uses a single rule, namely the previous five days' close price. This model does not have as many features or prices as the other five daily original pricing models.

F. Model Deployment

Fig 9. Comparison of Truth and Multiple Linear Regression on GBPUSD using 5 previous daily original prices.

Fig 10. Data Preparation Process Comparison of Truth and Multiple Linear Regression on EURUSD using 5 previous daily original prices.

The 5 previous daily original pricing models will now be tested on new data, specifically EURUSD, one of the most prominent currency pairings in the forex market. The overall findings at the time of the EURUSD truth mapping were still extremely good. This model still has prediction errors during price changes, as seen in Fig. 10 for EURUSD in the first half of 2020. The linearity of this model on EURUSD appears to be extremely good, as almost all forecast results are close to a linear line, with RMSE=0.00289, which is lower than the RMSE of this model on GBPUSD, and R^2=0.99538, which is higher than the R^2 of this model on GBPUSD. As a result, the EURUSD model's deployment process was effective, with even better results than the GBPUSD model.

V. CONCLUSION AND FUTURE STUDY

A. *Conclusion*

The 5 previous daily closing prices, 5 previous daily initial prices, and average prices can all be used to anticipate the daily closure of GBPUSD. These three feature groups were chosen because they have a strong relationship with the daily close price. Where the prediction has an extremely low error rate and nearly perfect linearity qualities. To build machine learning, the CRISP-DM approach is utilized, as well as extraction and feature selection, to create machine learning to predict daily close prices. The model was created using the Python language and the MLR algorithm on the GBPUSD currency pair. The best model is MLR with 5 prior daily original pricing feature groups, where RMSE is the lowest and R2 is the highest. The deployment process was pronounced effective because the 5 prior daily original pricing models performed better in another currency pair, namely EURUSD than it did in the GBPUSD currency pair.

B. *Future Study*

This study will be extended to find more features using more algorithms to give more options before applying these features to the expert system. Another possible study is to extract not just features but also rules using classification algorithms.

REFERENCES

[1] M. S. Islam, E. Hossain, A. Rahman, M. S. Hossain, & K. Andersson, “A review of recent advancements in forex currency prediction,” Algorithms, vol. 13, no. 8, 2020.

[2] C. Ma, & X. Wang, “Strategic interactions and negative oil prices. Annals of Financial Economics,” vol. 16, no. 03, 2021.

[3] A. Schrimpf, & V. Sushko, “FX trade execution: complex and highly fragmented,” BIS Quarterly Review, 2019.

[4] A. Vo, & C. Yost-Bremm, “A high-frequency algorithmic trading strategy for cryptocurrency,” Journal of Computer Information Systems, vol. 60, no. 6, pp. 555-568, 2020.

[5] A. Alshanqiti, & A. Namoun, “Predicting student performance and its influential factors using hybrid regression and multi-label classification,” IEEE Access, vol. 8, 2020.

[6] A. Heinz, M. Jamaloodeen, A. Saxena, & L. Pollacia, “Bullish and Bearish Engulfing Japanese Candlestick patterns: A statistical analysis on the S&P 500 index,” The Quarterly Review of Economics and Finance, vol. 79, pp. 221-244, 2021.

[7] B. Shyti, & D. Valera, "The regression model for the statistical analysis of Albanian economy," Int. J. Math. Trends Technol, vol. 62, no. 2, pp. 90-96, 2018.

[8] P. Schober, C. Boer, & L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," Anesthesia & Analgesia, vol. 126, no. 5, pp. 1763-1768, 2018.

[9] J. Guo, D. Ranasinghe, & Z. Zhang, "Developments in Foreign Exchange and Over-the-counter Derivatives Markets," RBA Bulletin, 2021.

[10] C. Schröer, F. Kruse, & J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," Procedia Computer Science, vol. 181, pp. 526-534.