ANALYZING THE IMPACT OF MULTIPLE STOCK INDICES IN PREDICTION OF US DOLLAR INDEX

MSc Research Project
Data Analytics

Mrunali More
Student ID: x18189059

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mrunali More
Student ID:	x18189059
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	17/8/2020
Project Title:	ANALYZING THE IMPACT OF MULTIPLE STOCK INDICES IN PREDICTION OF US DOLLAR INDEX
Word Count:	6203
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	2nd September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

ANALYZING THE IMPACT OF MULTIPLE STOCK INDICES IN PREDICTION OF US DOLLAR INDEX

Mrunali More
x18189059

Abstract

Foreign currency exchange rate and stock prices are the most important factors which play a vital role in the country's economic health. Forex rates are essential for the businesses during the international trading. Increase in the exchange currency rate is liable for a rise in demand for international trading; as a result, the rise in the profit and stock prices of the firm. Recent studies have proved that correlation exists between the stock prices and forex rates. This research study aims to forecast the US dollar index prices, considering values of four stock indices in the USA as the external factors. Time series models ARIMA and Facebook's Prophet along with machine learning model Extreme Gradient Boosting and Long Short-Term Memory (LSTM) neural network is applied for the forecasting. Result of these techniques is evaluated using mean absolute percentage error, while the result of multivariate ARIMA is used as a benchmark. Comparing the result of all techniques, we can see that the Prophet model outperformed by achieving the lowest mean absolute percentage error rate following ARIMA and LSTM models. The predicted results are following the actual trend of prices but need to improve the volatility of the output. Based on the findings of this analysis, stock prices can be successfully used to predict US dollar index rates for the USA market.

1 Introduction

1.1 Background and Motivation

The US dollar index estimates the value of the US dollar price corresponding to the prices of six other currencies namely, Euro, Swiss Franc, Japanese Yen, Canadian dollar, British Pound, and Swedish Krona.[1] The index price will grow as the dollar rises against those currencies and drop if the same weakens against those currencies.[2] Trading with a correct strategy at the right moment can be profitable for the traders (Shrestha et al. 2018). The US dollar index enables traders to monitor the value of the American dollar against the basket of chosen currencies in a single transaction.[2] Therefore, forecasting of the financial markets has become a difficult task. Recent studies show that the application of artificial intelligence, machine learning and the neural network has the ability to control such nonlinear and complex features of financial data, providing good prediction results.

[1] https://www.investopedia.com/terms/u/usdx.asp
[2] https://www.dailyfx.com/us-dollar-index

1.2 Relationship Between currency exchange rates and stock prices

Many researchers have investigated the linkage or co-movement between the currency exchange rates and stock prices. (Lin 2012) studied the co-movement between stocks and forex rate prices of the Asian market by applying the autoregressive distributed lag model. The outcomes of this analysis reveal that currency exchange rates and stock prices show strong correlation during a crisis period. (Lin 2012) also reviewed the co-movement between stocks and forex rate changes across various industries.

(Cumperayot, Keijzer and Kouwenberg 2006) observes that, many developing countries that have undergone a severe recession, a sharp fall in the stock market dramatically raises the risk of significant currency loss on the same day. (Cumperayot, Keijzer and Kouwenberg 2006) also found proof of spillover [3] of currency market crisis in the same region, but limited evidence suggests that currency market crisis extends from one country to another.

1.3 Research question and objective

"How far the predictions of US dollar index prices can be improved with the help of time series analysis and using prices of US stock indices (such as NASDAQ, NYSE, DJIA and S&P 500) as external factors?"

This study aims to answer the above research question by combining the prices of stock market indices like NYSE, NASDAQ, S&P 500 and Dow Jones Industrial Average (DJIA) with US dollar index prices. Since the target data is time-series, the analysis is performed using statistical time series algorithms along with machine learning techniques. The results of this analysis will be beneficial for US businesses and investors during international trading. The objectives of the study to address the research question are:

- Data pre-processing and merging of stock prices data with exchange rates.
- Scaling of data using the normalization method.
- Determining the relationship between stock prices and exchange rates using appropriate techniques.
- Implementation of time series (ARIMA and PROPHET), machine learning algorithms (XGBoost) and neural network on transformed data.
- Evaluation of outputs of implemented models.

The remaining paper is structured as follows, section 2 presents the related work about the currency exchange rate prediction followed by section 3 methodology. Section 4 describes the implemented techniques followed by section 5 which represents results and experimental evaluation. Finally, section 6 provides the conclusion and future work.

2 Related Work

This section covers a detailed overview of previous work in foreign currency exchange rate analysis in the form of literature review. It provides reviews for researches carried

[3] Spillover is a side effect of something else happened in unrelated context

out using different time series algorithms and machine learning techniques to predict the currency exchange rate and factors affecting the result.

2.1 Review of influencing factors for exchange rates

Prediction of the currency exchange rate is essential for businesses and investors. The recognition of factors that are affecting the exchange rate will help them to avoid financial loss and risk. The cost of any country's currency is calculated by market forces focused on commerce, consumption, tourism, and geopolitical threats.[4] Many researches in the past have used various factors to forecast currency exchange rate. Summary of these researches is given in Table 1.

Table 1: Summary of Datasets used for Exchange Rate Prediction

Authors	Data Source
(Galeshchuk 2016)	Forex rates of EUR/USD, GBP/USD, and USD/JPY
(Ahmed and Straetmans 2015)	macroeconomic factors
(Ranjit et al. 2018)	Tweets by traders
(Jin et al. 2013)	news articles + stock prices
(Smith 2012)	Google insights search

In the research by (Galeshchuk 2016) has employed multi-layer perceptron on exchange rates data as a data source for prediction. Exchange currency rate is a time-series data which majorly depends on the exchange rate of the previous day. (Galeshchuk 2016) has gathered data for exchange rates EUR/USD, GBP/USD, and USD/JPY in three different steps: daily data, monthly data, and quarterly data. The results of this analysis are determining low prediction error in daily data than monthly and quarterly data. In addition, (Ahmed and Straetmans 2015) are predicting exchange rate cycles using traditional macroeconomic factors as data source like Excess Return on Carry Trade (EROCT), Relative Purchasing Power Parity (RPPP), TED spread, Relative Stock Market Returns (RSMT) and Term Spread. The bull (rising prices) and bear (decline in prices) are identified using a nonparametric algorithm which classifies ups and downs in series and executed dynamic probit model for the prediction of forex rates (Ahmed and Straetmans 2015). Result of (Ahmed and Straetmans 2015) analysis proves that the breaches of the Uncovered Interest Rate Parity (UIRP), RPPP and RSMR shows predictive capacity over currency fluctuations, but only for small period.

Although, foreign exchange rate majorly depends on economic factors, in research carried out by (Ranjit et al. 2018) and (Jin et al. 2013) are predicting exchange rate by performing sentimental analysis. (Ranjit et al. 2018) are using exchange rates along with tweets by traders as a data source. Lexicon based sentimental analysis was conducted on the tweets to extract the sentiments as the market sentiments directly affect the exchange rate price movements. The artificial neural network is applied to exchange rates in which results are overfitting. Naïve Bayes model is utilized in the sentimental analysis, which predicts the prices with 90.63% accuracy (Ranjit et al. 2018). Whereas (Jin et al. 2013) are using news articles along with stock prices and exchange rates applying linear

[4] https://www.investopedia.com/ask/answers/08/what-is-foreign-exchange.asp

regression model for the prediction. The Latent Dirichlet Allocation method is applied to detect appropriate phrases from news stories by searching through a target keyword, then used custom sentiment dictionaries to measure movement values for interest rates and inflation and AFINN dictionary is used to extract sentiments (Jin et al. 2013). This analysis is employed on currencies of four countries out of which the model predicted good results for two countries with recall equals to 1.

Alternatively, (Smith 2012) collected Google insights search volume data for keywords searches like economic crisis, financial crisis, inflation, and recession for the foreign currency volatility prediction. He calculated volatility for short term (natural long of one-week data) and long term (moving average of four weeks) using the GARCH model. This analysis by (Smith 2012) rejected the hypothesis that a GARCH 's conditional variance is an impartial predicting variable of volatility in the exchange rate. It also rejected the hypothesis that data of Google insight search volume have predictive capacity beyond the GARCH.

2.2 Review of time series techniques

The statistical ARIMA (AutoRegressive Integrated Moving Average) model has been successfully used in the past for the analysis and prediction of time series data like stock prices prediction, weather forecasting, tourism demand analysis, etc. (Tseng et al. 2001) have tried to improve the results of conventional ARIMA techniques by considering fuzzy regression. (Tseng et al. 2001) compared four models, ARIMA, fuzzy ARIMA, Chen fuzzy time series and Watada fuzzy time series. The fundamental principle of the fuzzy theory is that the parameter variance in the model generates residuals between estimators and findings, and the probability function is used to deal with actual measurements. Fuzzy ARIMA gives better results than ARIMA when data provided to fuzzy ARIMA is less than the data required for the ARIMA model. This model is suited for decision-makers to get the best and worst possible situations. It also handles outliers in the data. Besides ARIMA, ARCH family-based techniques are also the most common models for volatility prediction in financial data due to their ability to detect clustering and consistency of variability over time series (Lahmiri 2017). In the research, (Lahmiri 2017) has implemented traditional GARCH and EGARCH models and hybrid GARCH/EGARCH-ANN models to predict the volatility in US/Canada and US/Euro exchange rates. The findings revealed that hybrid models outperform the conventional GARCH and EGARCH models. Also, GARCH performance better among GARCH and EGARCH models.

2.3 Review of artificial intelligence techniques

Statistical ARIMA technique has been commonly utilized for over two centuries for forecasting of time series. In recent years, machine learning techniques are also catching the attention of researchers for the prediction of time series and financial data. The most popular machine learning methods are support vector machine, random forest, decision tree and neural networks (Ramakrishnan et al. 2017). Machine learning has been successfully implemented for financial predictions like the stock market, interest rates, bankruptcy, and forex rates (Majhi, Panda and Sahoo 2009). (Kamruzzaman, Sarker and Ahmad 2003) have applied support vector machine (SVM) model comparing the outputs of different SVM kernels for the prediction of the Australian dollar against six different currencies. The results showed that polynomial and radial kernel had given the best results while predicting a trend in data. The polynomial kernel is also reducing overall prediction error. Whereas (Ramakrishnan et al. 2017) are checking the impact of commodities prices such as palm oil, crude oil, gold price and rubber price on the Malaysian exchange rate, with the help of SVM along with neural network and random forest model. This analysis reveals that random forest technique performs better than SVM and neural network because of its capacity of managing random data.

Further (Sun, Liu and Sima 2020) introduced Light GBM a novel gradient boosting model to forecast the cryptocurrency. In comparison to conventional GBDT (gradient boosting decision tree) based approaches, Light GBM would move tree vertically, while other algorithms would move tree horizontally, making it an efficient tool for processing large-scale data and features (Sun, Liu and Sima 2020). Light GBM model comparatively performs better than SVM and RF in robustness, making it an efficient forecasting technique when handling a large number of data instances and features at the same time. The approaches implemented by (Sun, Liu and Sima 2020) is also more effective for 2-week prediction.

In the analysis by (Shrestha et al. 2018), the results of four neural networks namely, Simple Recurrent Neural Network, Long Short Term Memory (LSTM), Muti-Layer Perceptron (MLP) and Gated Recurrent Unit (GRU) are compared to predict Nepalese currency against US dollar, Euro and Pound. Close prices of exchange rates are considered as a dependent variable and previous day's prices of open, close, high, and low are considered as independent vasriables. LSTM model predicted good results than others with the lowest mean absolute percentage error. (Özdemir and Bogosyan 2018) have tried to generate a trading signal for cryptocurrency using neural networks. Technical indicators commonly used for the prediction of financial asset movements. (Özdemir and Bogosyan 2018) implemented three neural networks, a recurrent neural network with one output, one classification neural network with three outputs as buy, wait and sell and second classification neural network with five outputs as a strong buy, buy, wait, sell, and strong sell. But the results of this study are not satisfactory and needs improvement as the wrong prediction of trading signals can lead to a huge loss.

(Majhi, Panda and Sahoo 2009) established two new less complex artificial neural networks, FLANN (functional link artificial neural network) and CFLANN (cascaded functional link artificial neural network) to forecast currency exchange rate. FLANN is a single-layer neural network with a nonlinear input and a single neuron at output and in CFLANN two single-layer FLANNs are cascaded in sequence. FLANN is a replacement for multilayered artificial neural networks (Majhi, Panda and Sahoo 2009). Basic least mean square algorithm (LMS) is also implemented to compare the results. CFLANN model outperforms and proves to be suitable for all types of currency exchange prediction for any number of months in the future.

2.4 Review of hybrid models

In recent years, researchers have also started adopting hybrid models (a combination of two algorithms) to improve the outcomes of machine learning or neural networks. In this way, data and model are given equal priority and allows to check data quality at early stages.[5] The approach of using modelling, simulation, and optimization techniques results in the continuous improvement of model parameters by continuously comparing the model forecasts with original data. (Sespajayadi, Nurtanio et al. 2015)) predicted Euro/USD

[5]https://www.itwm.fraunhofer.de/en/departments/opt/machine-learning-hybrid-models.html

exchange rates using Genetic Algorithm-neural network hybrid (GANN) model. Genetic Algorithms are random search algorithms which aim to extend knowledge of the natural selection process and evolutionary biology to generate high-quality solutions to problems. The predicted results of GANN achieved high accuracy rate, and low root means square error around 0.00044 as compared with actual values. In contrary, a nonlinear ensemble model proposed by (Yu, Wang and Lai 2005) is an integration of generalized linear autoregression (GLAR) with ANN. GLAR model is similar to ARIMA, which also considers external factors while predicting time series results. In this analysis, predictions of GLAR and ANN are combined to get one result and ensembled using principal component analysis (PCA) method to get the second result. The predictions of a nonlinear ensembled hybrid model achieved good accuracy with the low normalized mean square error.

To address the limitations of standard statistical models, (Rout et al. 2014) developed an ARMA (AutoRegressive Moving Average) with differential evolution based training to predict forex rates. The outcomes of this model are compared to four other hybrid evolutionary computing techniques including ARMA-cat swarm optimization (CSO), ARMA-bacterial foraging optimization (BFO), ARMA-particle swarm optimization (PSO), and ARMA- forward backwards least mean square (FBLMS) (Rout et al. 2014). This study has not achieved very good results as compared to other analysis, but among all five hybrid ARMA models, ARMA-DE model performed better than others. The financial time series are originally chaotic, containing non-linearity, determinism, and sensitivity. (Pradeepkumar and Ravi 2016) applied hybrid quantile regression random forest (QRRM) to predict the exchange rates using chaos in data. Reconstructed data containing chaos, lag and embedding dimension are given as input to quantile regression, random forest, and hybrid QRRF models. In this research, QRRF model-generated best predictions than others because of the robustness of random forest model, QR model's heteroskedasticity and low variance because of the bagging process (Pradeepkumar and Ravi 2016).

2.5 State of art in exchange rate prediction

Even though there have been many kinds of research carried out on currency exchange rates, still several studies are exploring novel factors to predict forex prices. (Calice and Zeng 2019) conducted an analysis to determine the impact of sovereign credit default swap (CDS) on the prediction of exchange rates. A credit default swap permits the investor to trade his credit risks with another investor.[6] The details gained from this study offers entirely new insight for a deeper understanding of the currency market dynamics. Portfolio sorts and Fama–MacBeth regression techniques are Utilizing on datasets of 29 countries, (Calice and Zeng 2019) discover that out-of-sample exchange rates can be forecasted significantly by the sovereign term CDS.

Whereas, a study by (Zhou et al. 2020)) proves that economic policy uncertainty between the US and China (Sino-US EPU ratio) has a strong and positive effect on the volatility of Chinese exchange rate. (Zhou et al. 2020) carried out this analysis using GARCH-MIDAS and traditional GARCH models were used to check the robustness of the GARCH-MIDAS. An interesting analysis was conducted by (Njindan Iyke 2020) in which he explored the influence of novel COVID-19 on exchange rate returns and volatility. The disease outbreak is very uncertain, unpredictable, and affects the world economy, but it has provided an experimental setup for the researchers to test the effects

[6] https://www.investopedia.com/terms/c/creditdefaultswap.asp

of natural disaster on financial markets. (Njindan Iyke 2020) used the total number of cases and deaths of 25 most affected countries from December 2019 to May 2020. The analysis results confirmed that COVID-19 has great forecasting capacity over volatility than over returns for a one-day forecast period; however, the opposite is seen for a five-day forecast duration where COVID-19 appears to influence returns rather than volatility (Njindan Iyke 2020).

3 Methodology

Knowledge discovery from databases (KDD) is an exploratory analysis to extract high-level knowledge from low-level data. The KDD includes the process of data storage and access, applying data mining algorithms on large datasets, interpreting the results and visualization; overall, it shows how the relationship between man-machine is developed and supported (Fayyad, Piatetsky-Shapiro and Smyth 1996). KDD has emerged and continues to grow because of the research areas such as expert systems, machine learning, statistics, pattern recognition, artificial intelligence, data visualization, and high-performance computing (Fayyad, Piatetsky-Shapiro and Smyth 1996).

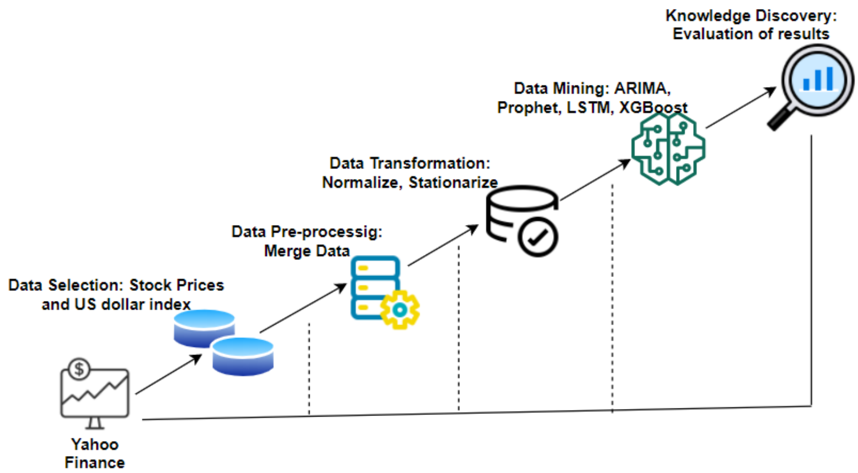

Figure 1: KDD for US Dollar Index Prediction

KDD phases for US dollar exchange rate prediction are as follows:

3.1 Data Selection

The data selected for this study is a historical dataset of US Dollar Index, which contains daily prices of the US dollar relative to the basket of foreign currencies. The daily stock prices of NASDAQ index, NYSE index, Dow Jones Industry Average (DJIA) index and S&P 500 index will act as external factors in this analysis. The US Dollar index prices and stock index prices are collected for a period of 30 years from 01 Jan 1990 to 01 Jan 2020 downloaded directly from Yahoo Finance. The dataset includes six variables, namely, High, Low, Open, Close, Volume, Adjacent Close and Date. We will be working only on close prices as it gives rates when the financial market closes at the end of the day.

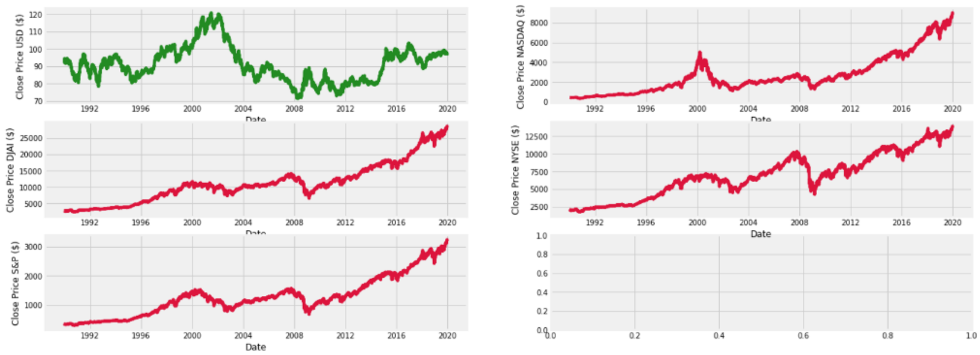

Figure 2: Plot of US Dollar Index and Stock Indices

3.2 Data Pre-processing

3.2.1 Merging of Dataset

For easy access, close prices of all five datasets are stored in one new pandas data frame. Before storing values in a new data frame, we will check whether all five close prices contain data for the same dates and data for missing days is removed. The final dataset contains a total of 7526 values. Figure 3 shows the basic descriptive statistics of final data.

	USD	NASDAQ	DJAI	NYSE	S&P
count	7526.000000	7526.000000	7526.000000	7526.000000	7526.000000
mean	90.988468	2672.059501	11094.943094	6912.690210	1264.494260
std	10.100382	1949.116644	6093.828594	3146.605163	671.534101
min	71.330002	325.399994	2365.100098	1715.060059	295.459991
25%	82.949997	1314.385010	7275.245117	4605.680054	805.032471
50%	89.879997	2166.800049	10500.114746	6773.955078	1191.155029
75%	96.870003	3363.409973	13342.594727	9278.705078	1484.985016
max	120.900002	9022.389648	28645.259766	13944.139648	3240.020020

Figure 3: Descriptive statistics of final data

3.3 Data Transformation

3.3.1 Decomposition of data

Time series data is a combination of level, trend, seasonality, and noise. With the help of decomposition process, we can easily split up all the components of the time series to get a better understanding. Figure 4 below shows the plot of decomposed data, and hence we can see that US dollar data shows yearly seasonality.

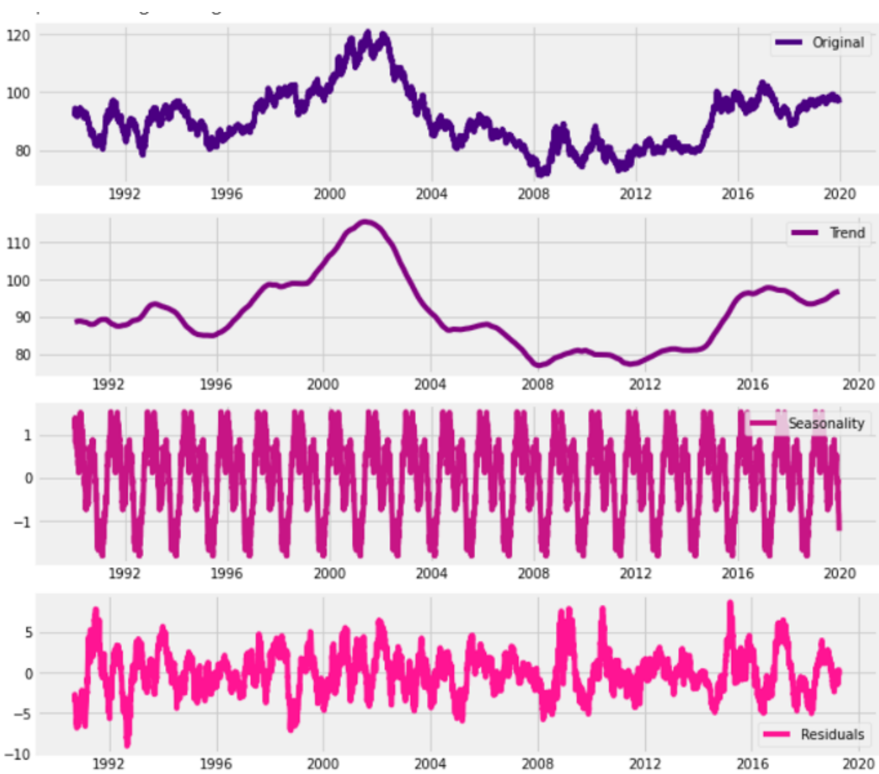

Figure 4: Seasonal Decomposition of data

3.3.2 Normalization of Data

Final data is normalized between 0 to 1 so that all the variables are in the same range. Also, normalized data uses less space and time while processing. The description of normalized data is shown in Figure 5.

	USD	NASDAQ	DJAI	NYSE	S&P
count	7526.000000	7526.000000	7526.000000	7526.000000	7526.000000
mean	0.396580	0.269824	0.332184	0.425022	0.329093
std	0.203760	0.224114	0.231879	0.257305	0.228059
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.234416	0.113716	0.186838	0.236373	0.173056
50%	0.374218	0.211728	0.309550	0.413677	0.304186
75%	0.515231	0.349317	0.417710	0.618497	0.403974
max	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 5: Summary of normalized data

3.3.3 Unit Root Test

Before applying any data mining algorithm, we need to check if time series data is stationary. Statistical properties like variance and mean of stationary data do not vary with time. Stationary data do not contain any trend or seasonality.

To check if the stationarity of data, we perform the Augmented Dickey-Fuller (ADF) test. ADF test checks the null hypothesis to see that the unit root is present in data. The unit root is a characteristic of time series data which makes it non-stationary.[7] Result of the ADF test is shown in Figure 6

```
Results of Dickey-Fuller Test: USD
Test Statistic                   -1.974825
p-value                           0.297719
#Lags Used                        0.000000
Number of Observations Used    7525.000000
Critical Value (1%)              -3.431219
Critical Value (5%)              -2.861924
Critical Value (10%)             -2.566974
dtype: float64
```

Figure 6: Unit Root Test: ADF Test

As we can see, p-value > 0.05, hence we do not reject the null hypothesis, and it proves that data is non-stationary.

The data is then stationarized by taking the difference between the original data and the lagged value of data.

3.3.4 Johansen Co-integration Test

If there is co-integration among the factors of data, then the presence of a long-term relationship between the factors is confirmed. The group of variables is considered to be co-integrated if they are collectively non-stationary, but their linear combination is stationary (Ibrahim 2000).

Johansen co-integration test is performed on data containing US dollar index prices and stock indices prices. Johansen test has two statistics, Trace, and maximum Eigenvalue. Trace checks the null hypothesis that there are at most k co-integrating vectors against the alternative of k or more vectors whereas Eigenvalues check the null hypothesis for k co-integrating vectors against k+1 co-integrated vectors (Ibrahim 2000). We do not reject the null hypothesis as the values of trace are less than critical values which shows that the US dollar index and stock index prices are co-integrated and have long term relationship.

```
Eigen Values:
 [0.00450676 0.00316367 0.00150947 0.0007697  0.00013474]


Critical Values:
 [[65.8202 69.8189 77.8202]
 [44.4929 47.8545 54.6815]
 [27.0669 29.7961 35.4628]
 [13.4294 15.4943 19.9349]
 [ 2.7055  3.8415  6.6349]]


Trace Values:
 [75.99982534 42.01434214 18.17313713  6.80731666  1.01385777]
```

Figure 7: Result of Johansen Co -integration Test

[7] https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/

3.3.5 Granger Causality Test

Granger causality test checks the short-term relationship between the dependent and independent variables in data. It tests the null hypothesis that the past values of time series (X) do not cause the time series (Y).[8] The null hypothesis is rejected as p-values are equal to 0 shown in Figure 8, which confirms that past values of stock prices do affect the US dollar index prices.

	NASDAQ_x	DJAI_x	NYSE_x	S&P_x	USD_x
USD_y	0.0001	0.0	0.0	0.0	1.0

Figure 8: Granger Causality test

3.4 Data Mining

Data mining techniques such as machine learning algorithm extreme gradient boosting, long short-term memory neural network, time series methods multivariate SARIMAX and PROPHET are applied on the stationary data to check the impact of stock prices on foreign exchange rates. The detailed information of implemented models is summarized in section 4.

3.5 Knowledge Discovery

The outcomes of all the techniques are evaluated using mean absolute percentage error, root mean square error and mean absolute error. The details of the experimental evolution are described in section 5.

4 Implementation

4.1 Implementation of multivariate SARIMAX model

ARIMA (Autoregressive Integrated Moving Average) is a statistical model commonly utilized for analysis and prediction of time-series data. ARIMA is a transformed version of the autoregressive moving average (ARMA) model (Mondal, Shit and Goswami 2014).

In this study, we are working on SARIMAX, which is a seasonal multivariate ARIMA model. SARIMAX considers the seasonality of data while fitting the model. The ARIMA model is majorly dependent on order (p, d, q) and seasonal order (P, D, Q). These parameters are described as:

- p indicates a number of autoregressive
- d is an integrated which denotes the number of times difference taken to make data stationary
- q signifies a number of moving averages

[8]https://towardsdatascience.com/granger-causality-and-vector-auto-regressive-model-for-time-series-forecasting-3226a64889a6

- P denotes a number of seasonal autoregressive
- D is a number of seasonal differences
- Q shows a number of seasonal moving averages

Exog parameter is used to provide external factors which are stock indices prices. The Akaike Information Criteria (AIC) is calculated to determine the goodness of fit of the model. The partial autocorrelation function (PACF) and autocorrelation function (ACF) is used to verify values of p and q, respectively.

The result of a model with the lowest value of AIC is considered closer to actual values.

```
                                     Statespace Model Results
==========================================================================================
Dep. Variable:                                 USD   No. Observations:                6020
Model:             SARIMAX(2, 1, 3)x(0, 1, 1, 12)   Log Likelihood               19570.468
Date:                            Sun, 16 Aug 2020   AIC                         -39118.936
Time:                                    16:35:58   BIC                         -39045.228
Sample:                                         0   HQIC                        -39093.341
                                           - 6020
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
NASDAQ        -0.4488      0.059     -7.628      0.000      -0.564      -0.333
DJAI           0.3454      0.095      3.622      0.000       0.158       0.532
NYSE          -2.3305      0.075    -31.034      0.000      -2.478      -2.183
S&P            2.9966      0.172     17.431      0.000       2.660       3.334
ar.L1         -0.7627      0.045    -16.803      0.000      -0.852      -0.674
ar.L2         -0.9066      0.060    -15.204      0.000      -1.023      -0.790
ma.L1          0.7695      0.046     16.577      0.000       0.679       0.861
ma.L2          0.9118      0.061     15.035      0.000       0.793       1.031
ma.L3          0.0045      0.011      0.403      0.687      -0.017       0.026
ma.S.L12      -0.9851      0.003   -311.793      0.000      -0.991      -0.979
sigma2       8.55e-05   1.15e-06     74.479      0.000    8.33e-05    8.78e-05
===================================================================================
Ljung-Box (Q):                       45.95   Jarque-Bera (JB):               924.05
Prob(Q):                              0.24   Prob(JB):                         0.00
Heteroskedasticity (H):               0.57   Skew:                            -0.08
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.91
===================================================================================
```

Figure 9: Summary of SARIMAX

4.2 Implementation of time series Prophet model

Prophet model is open-source software, created by a team of Facebook's data scientists in 2014 available for Python and R. Prophet is designed to manage business time-series data which contains time, daily, weekly observations, large outliers, trend changes, missing data (Yenidoğan et al. 2018). Prophet framework easily handles components like a trend, seasonality (weekly and yearly), weekends and holidays.

$$d(t) = a(t) + s(t) + h(t) + \epsilon \tag{1}$$

Where, d(t) is Prophet time series, a(t) is trend, s(t) is seasonality, h(t) is holiday and e is an error in data.

Prophet model requires a unique data frame containing two special columns, namely 'ds' and 'y' to handle seasonality and time series. The 'ds' column stores date-time series and 'y' column contain time series values (dependent variable values) (Yenidoğan et al. 2018). External factors (independent variables) are added using 'add_regressor' function.

4.3 Implementation of Long Short-Term Memory neural network

Long-term memory is a type of deep neural network that is commonly used in time series analysis. Unlike the conventional neural network, this model adds a hidden layer that is developed by the sequential information of a time series, with output depending on the hidden layer (Bao, Yue and Rao 2017). Another benefit of the LSTM is that it can quickly understand long-term dependencies because of its memory cells. A memory cell is comprised of four units: an input gate, an output gate, a forget gate and a self-recurrent neuron (Bao, Yue and Rao 2017).

In this analysis, we added the date and its features to data as external factors because LSTM is not a time series model. A sequential five-layer LSTM neural network is implemented using Keras library in python. The number of neurons and epochs are specified, and a suitable optimizer is used to reduce the error and to obtain accurate forecasted results.

```
Model: "sequential_16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_64 (LSTM)               (None, 11, 64)            16896
_________________________________________________________________
lstm_65 (LSTM)               (None, 11, 128)           98816
_________________________________________________________________
lstm_66 (LSTM)               (None, 11, 64)            49408
_________________________________________________________________
lstm_67 (LSTM)               (None, 32)                12416
_________________________________________________________________
dense_23 (Dense)             (None, 1)                 33
_________________________________________________________________
dropout_19 (Dropout)         (None, 1)                 0
=================================================================
Total params: 177,569
Trainable params: 177,569
Non-trainable params: 0
```

Figure 10: Summary of LSTM

4.4 Implementation of machine learning algorithm XGBoost regression

Extreme gradient boosting (XGBoost) is one of the most popular machine learning algorithms used to solve classification and regression problems. Boosting is an ensemble technique to aggregate all the poor models to make them better and robust.[9] Gradient boosting is an improved version of the decision tree algorithm. In XGBoost, gradient boosted trees grow parallelly and reduce the required time (Basak et al. 2019).

In this analysis, we added the date and its features to data as external factors same as LSTM because even XGBoost is not a time series model. We implemented XGBoost with default parameters and using hyperparameters with a list of possible values along with 3-fold cross validation.

[9]https://www.datasciencelearner.com/gradient-boosting-hyperparameters-tuning/

5 Evaluation

5.1 Experiments with multivariate SARIMAX and Results

Multiple experiments are performed on SARIMAX by changing the order (p, d, q) and seasonal order (P, D, Q, m) of a model. We calculated AIC for a different combination of orders and implemented model with orders which have least AIC values. The outputs of the experiment are evaluated using root mean square error, mean absolute percentage error, and mean absolute error.

As shown in the Table 2, even though a model with order (2,1,3) (0,1,1,12) do not have lowest AIC, it performed better with a low error rate than the model with the lowest AIC value.

Table 2: Results of SARIMAX

p,d,q	P,D,Q	AIC	RMSE	MAPE	MAE
(0,1,0)	(0,0,0,12)	-39406.321	7.774	0.070	6.732
(2,1,3)	**(0,1,1,12)**	**-39118.936**	**7.252**	**0.059**	**5.663**
(3,0,2)	(1,1,1,12)	-38928.439	7.508	0.061.	5.883
(0,1,1)	(0,1,0,12)	-35088.266	57.605	0.543	51.937

In Figure 10, we can see the plot of predicted values and actual values. From the plot, we can say that predicted output follows the same trend as the original values, but could not capture the volatility from 2015 till 2018.

Figure 11: Actual and Predicted values of ARIMAX

Diagnosis of residual is carried out to check the goodness of fit of SARIMAX model shown in Figure 11. It contains Standardized Residual, Histogram, Normal Q-Q plot and Correlogram plot. From the standardized residual plot, we can see the absence of seasonality in the residuals. The correlogram plot reveals that there is low autocorrelation between the lagged versions of itself. The histogram displays that the residuals are normally distributed. The q-q plot illustrates that the residuals are following a linear trend. These findings help us determine that the model provides a satisfactory fit to our data.[10]

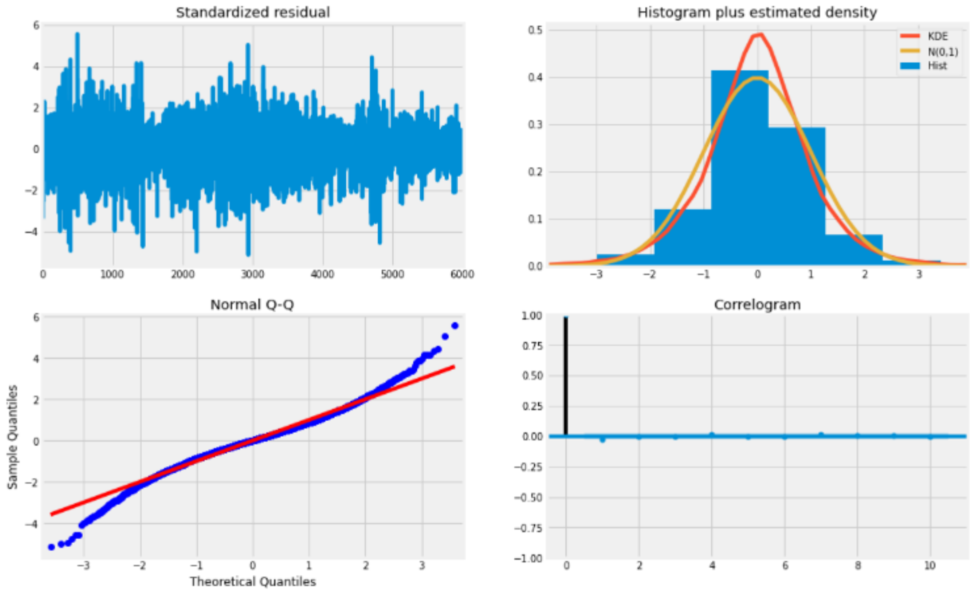

Figure 12: Diagnostic plot of residuals of SARIMAX

5.2 Experiments with PROPHET and Results

PROPHET is a fully automatic model. With less processing steps, it handles stationarity, trend, and seasonality of data. The model was experimented with default hyperparameters and adding new parameter yearly seasonality. PROPHET is the fastest model, takes very less time for processing as compared to other models. Results of experiments are displayed in the Table 3. Model after adding yearly seasonality parameter has improved, and an error rate (MAPE) is decreased by 0.2%.

Table 3: Results of Prophet

Parameters	**RMSE**	**MAPE**	**MAE**
Default	5.837	0.052	4.949
Yearly Seasonality = 365	**5.707**	**0.050**	**4.833**

The plot of predicted and original values is displayed in Figure 13. Prophet model could achieve trend as per actual values, and the volatility is improved as compared to SARIMAX. We can also observe that the volatility of predicted output is little shifted than the original data.

[10]https://www.kaggle.com/magiclantern/co2-emission-forecast-with-python-arima-v2

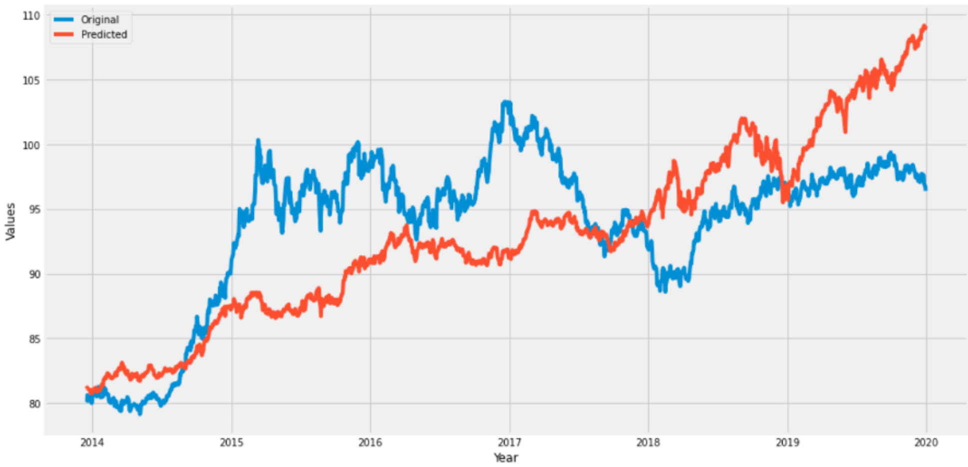

Figure 13: Actual and Predicted values of Prophet

5.3 Experiments with LSTM and Results

Experiments with LSTM is carried out by changing the number of neurons in input and hidden layers. Model architecture has an input layer, three hidden layers and an output layer. The model is trained for 80 batch size and 50 epochs using 'adam' optimizer and 'relu' activation function. The different number of neurons and its output is described in the Table 4.

Table 4: Results of LSTM

Neurons	Epochs	RMSE	MAPE	MAE
32-64-128-32	50	20.168	0.192	18.489
50-50-70-35	50	10.843	0.093	8.790
32-128-64-32	50	8.102	0.072	6.904
64-128-64-32	**50**	**7.814**	**0.070**	**6.713**

Even though the error rate of LSTM reduced from 19% to 7%, the predicted output is showing a one-year seasonality as we can see in Figure 14. The volatility of the predicted values is not correlated with the actual values.

Figure 14: Actual and Predicted values of LSTM

5.4 Experiments with XGBoost and Results

Extreme gradient boosting was implemented on the dataset with default parameters and selected hyperparameters(n_estimators, max_depth, min_child_weight and learning rate), and results are evaluated using RMSE, MAPE and MAE. Table 6 shows the results of experiments. A model with default parameters has error rate (MAPE) 23%, which was reduced by 15% after using tuned hyperparameters and 3-fold cross-validation.

Table 5: Results of XGBoost

Parameters	**RMSE**	**MAPE**	**MAE**
Default	30.713	0.236	22.491
Tuned hyperparameters	**10.686**	**0.086**	**8.158**
Hyperparameters + 3-fold cross validation	**10.686**	**0.086**	**8.158**

The predicted values of XGBoost is almost aligned with actual values from the year 2014 till 2017 can be seen in Figure 15. After 2017, the trend of the predicted output is increasing with very low volatility, whereas actual values are showing a rising trend with variations in volatility.

Figure 15: Actual and Predicted values of XGBoost

5.5 Comparative Results

From Figure 16, we can see that predictions of multivariate SARIMA, and Prophet are almost the same. It could capture the small movement in the data with little lag. A large gap is seen between original values and predicted values by multivariate SARIMA from the year 2015 to 2018. Predicted values of Prophet model reduce this difference. Whereas the outcome of Extreme Gradient Boosting fails to capture drop in trend from the year 2017 – 2018. Furthermore, LSTM model has failed to capture movements in the data and shows yearly seasonality.

Table 6: Best results of all models

Model	RMSE	MAPE	MAE
ARIMA	7.252	0.059	5.663
Prophet	**5.707**	**0.050**	**4.833**
LSTM	7.814	0.070	6.713
XGBoost	10.686	0.086	8.158

Figure 16: Comparative plots of results of all models

6 Conclusion and Future Work

This study explores the relationship between stock index prices (such as NASDAQ, NYSE, DJIA and S&P 500) and US dollar index. The result of this analysis confirms that the stock index prices have a positive impact on US dollar index prices and stock prices can improve the prediction of the US dollar index. We performed the granger causality test, the hypothesis is rejected, and the outcome revealed that past values of stock prices are useful for prediction of the US dollar index. Granger causality test also proves the short-term relationship between the variables. To check the long-term relationship, we performed the Johansen co-integration test. The null hypothesis of Johansen co-integration is not rejected which confirms the long-term relationship between data.

The time series algorithms multivariate seasonal ARIMA (SARIMAX) and Prophet, machine learning algorithm XGBoost and long short-term memory (LSTM) neural network are implemented to predict the US dollar index using stock indices prices as predictors. The results of all the techniques are evaluated using root mean square error, mean absolute percentage error and mean absolute error. The prediction error rate (MAPE) of all the models ranged from 5% to 8%, where Prophet performed slightly better than others. The findings of this analysis will be beneficial for the investment market, currency trading, investors and international businesses. The investors and traders would be at a better place while trading/investing when they already have the speculation of future currency exchange rate, which would, in return, help them to avoid losses.

The future work would include the improvement of the volatility of results using GARCH algorithms. This analysis can also be done using stock prices and exchange rate data of different countries, also check whether the stock indices of other countries have any impact on the US dollar index. The impact of different industries like technology, finance, pharmaceutical etc. on the US dollar index can also be evaluated.

7 Acknowledgement

I would like to thank my thesis supervisor Dr. Muhammad Iqbal for his support and assistance with my queries and guiding me in the right direction. I would also like to thank my friends, family and National College of Ireland for their support.

References

Ahmed, Jameel and Stefan Straetmans (2015). "Predicting exchange rate cycles utilizing risk factors". In: *Journal of Empirical Finance* 34, pp. 112–130.

Bao, Wei, Jun Yue and Yulei Rao (2017). "A deep learning framework for financial time series using stacked autoencoders and long-short term memory". In: *PloS one* 12.7, e0180944.

Basak, Suryoday et al. (2019). "Predicting the direction of stock market prices using tree-based classifiers". In: *The North American Journal of Economics and Finance* 47, pp. 552–567.

Calice, Giovanni and Ming Zeng (2019). "The term structure of sovereign credit default swap and the cross-section of exchange rate predictability". In: *International Journal of Finance & Economics*.

Cumperayot, Phornchanok, Tjeert Keijzer and Roy Kouwenberg (2006). "Linkages between extreme stock market and currency returns". In: *Journal of International Money and Finance* 25.3, pp. 528–550.

Fayyad, Usama, Gregory Piatetsky-Shapiro and Padhraic Smyth (1996). "From data mining to knowledge discovery in databases". In: *AI magazine* 17.3, pp. 37–37.

Galeshchuk, Svitlana (2016). "Neural networks performance in exchange rate prediction". In: *Neurocomputing* 172, pp. 446–452.

Ibrahim, Mansor H (2000). "Cointegration and Granger causality tests of stock price and exchange rate interactions in Malaysia". In: *ASEAN Economic Bulletin*, pp. 36–47.

Jin, Fang et al. (2013). "Forex-foreteller: Currency trend modeling using news articles". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1470–1473.

Kamruzzaman, Joarder, Ruhul A Sarker and Iftekhar Ahmad (2003). "SVM based models for predicting foreign currency exchange rates". In: *Third IEEE International Conference on Data Mining*. IEEE, pp. 557–560.

Lahmiri, Salim (2017). "Modeling and predicting historical volatility in exchange rate markets". In: *Physica A: Statistical Mechanics and its Applications* 471, pp. 387–395.

Lin, Chien-Hsiu (2012). "The comovement between exchange rates and stock prices in the Asian emerging markets". In: *International Review of Economics & Finance* 22.1, pp. 161–172.

Majhi, Ritanjali, Ganapati Panda and Gadadhar Sahoo (2009). "Efficient prediction of exchange rates with low complexity artificial neural network models". In: *Expert systems with applications* 36.1, pp. 181–189.

Mondal, Prapanna, Labani Shit and Saptarsi Goswami (2014). "Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices". In: *International Journal of Computer Science, Engineering and Applications* 4.2, p. 13.

Njindan Iyke, Bernard (2020). "The disease outbreak channel of exchange rate return predictability: Evidence from COVID-19". In: *Emerging Markets Finance and Trade* 56.10, pp. 2277–2297.

Özdemir, Aykut and Seta Bogosyan (2018). “Neural Network based Trading Signal Generation in Cypto-Currency Markets”. In: *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. IEEE, pp. 1–4.

Pradeepkumar, Dadabada and Vadlamani Ravi (2016). “FOREX rate prediction using chaos and quantile regression random forest”. In: *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*. IEEE, pp. 517–522.

Ramakrishnan, Suresh et al. (2017). “Forecasting Malaysian exchange rate using machine learning techniques based on commodities prices”. In: *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*. IEEE, pp. 1–5.

Ranjit, Swagat et al. (2018). “Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis”. In: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, pp. 1173–1177.

Rout, Minakhi et al. (2014). “Forecasting of currency exchange rates using an adaptive ARMA model with differential evolution based training”. In: *Journal of King Saud University-Computer and Information Sciences* 26.1, pp. 7–18.

Sespajayadi, Ary, Ingrid Nurtanio et al. (2015). “Technical data analysis for movement prediction of Euro to USD using Genetic Algorithm-Neural Network”. In: *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, pp. 23–26.

Shrestha, Shruti et al. (2018). “Comparison of algorithms in Foreign Exchange Rate Prediction”. In: *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*. IEEE, pp. 9–13.

Smith, Geoffrey Peter (2012). “Google Internet search activity and volatility prediction in the market for foreign currency”. In: *Finance Research Letters* 9.2, pp. 103–110.

Sun, Xiaolei, Mingxi Liu and Zeqian Sima (2020). “A novel cryptocurrency price trend forecasting model based on LightGBM”. In: *Finance Research Letters* 32, p. 101084.

Tseng, Fang-Mei et al. (2001). “Fuzzy ARIMA model for forecasting the foreign exchange market”. In: *Fuzzy sets and systems* 118.1, pp. 9–19.

Yenidoğan, Işil et al. (2018). “Bitcoin forecasting using ARIMA and prophet”. In: *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. IEEE, pp. 621–624.

Yu, Lean, Shouyang Wang and Kin Keung Lai (2005). “A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates”. In: *Computers & Operations Research* 32.10, pp. 2523–2541.

Zhou, Zhongbao et al. (2020). “Can economic policy uncertainty predict exchange rate volatility? New evidence from the GARCH-MIDAS model”. In: *Finance Research Letters* 34, p. 101258.