Prediction of Foreign Exchange Rate Using Data MiningEnsemble Method

Msc
Data Analytics

VIMALRAJ KUMAR
x15009025

School of Computing
National college of Ireland

Supervisor: Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	VIMALRAJ KUMAR
Student ID:	x15009025
Programme:	Msc Data Analytics
Year:	2016
Module:	Research Project
Lecturer:	Catherine Mulwa
Submission Due Date:	22/08/2016
Project Title:	Prediction of Foreign Exchange Rate Using Data Mining Ensemble Method
Word Count:	6000

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 encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	15th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. 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):	

Prediction of Foreign Exchange Rate Using Data Mining Ensemble Method

VIMALRAJ KUMAR
15009025
MSc in Data Analytics

15th September 2016

Area 'Advanced Data Mining'

Can we efficiently predict foreign currency exchange rate by considering factors (interest and inflation rates and public debts) using data mining ensemble method?

Abstract

This project presents the implementation prediction that can accuracy predict the foreign exchange rate. how the prediction accuracy can be improved by developing an ensemble model of the deep learning algorithm, Distributed Random forest and generalised linear model using sparkling water (Spark +H20). According to the researchers of literature review from 2000-2016, there are several models that has been used for predicting the foreign exchange rate. Among which Artificial Neural Network, Linear regression, Support vector machine, Arima are best-suited models for predicting the time series data.By having the above models as a base for this project and also by considering this project with the time series data of exchange rate, an additional feature of an ensemble is done on the output of these three models. This model is also implemented on the big data for getting better accuracy and faster predictions. Each model by using sparkling water produces the accuracy of more the 90% and the ensemble models increase the accuracy of the model by 3% with the accuracy of 93%. The evaluation and the results of this model clearly delivers the ensemble method, on the sparkling water which can efficiently improve the accuracy and performance of the model.

Contents

1 Introduction

Money is the most important and deciding factor in our day to day life and we should know how to use it efficiently. In order to use the money efficiently, countries and their people have started to invest in several businesses. While investing, they have also started to invest in other countries and shares which mainly depend on the currency exchange rate. The well-planned business will always win. Similarly, people who plan and invest by understanding the circumstances of the financial market(i.e. the ups and downs of the financial market) also wins. One of the most important deciding factors of the financial market is country's currency exchange rate which is also known as the Foreign exchange rate.

Foreign Exchange rate market is one of the most difficult markets in the world that is effectively forecast because of its volatile and unpredictable nature. Hence, predicting of this Foreign Currency Exchange Rate (FOREX) has become important in the financial sector. By predicting the exchange rate effectively, the vulnerability of the trade investment can be decreased which provide smooth international trade and most importantly bringing maximum profit to the investors.

The main idea of this project is to predict the FOREX rate efficiently and also in a reliable manner.To achieve this, the best data mining models are selected based on the past research and the ensemble model created out of it for the better accuracy. This entire model is developed on the big data using Sparkling water (Spark +H2o) to predict the FOREX rate quickly and efficiency.

In order to demonstrate this project, the historical dataset of United states of America, United kingdom and Canada with their financial factors including interest rate,public debts, Gross domestic product, Consumer price index (Inflation rate), Shares and gold price are considered. The prediction of Exchange rate of Indian rupees - INR will be done against USD - United States Dollars, CAD - Canadian Dollar, and GBP - United Kingdom Pound are achieved with this ensemble model.

1.1 Research Question

Can we efficiently predict foreign currency exchange rate by considering factors (interest and inflation rates and public debts) using data mining ensemble method?

The above-mentioned research question for this project is to demonstrate whether the data mining ensemble method by considering financial factors (interest and inflation rates and public debts) can efficiently predict foreign currency exchange rate.

1.2 Research Objective

- Literature review of prediction models for foreign currency exchange rate which considers several factors developed using data mining ensemble method (2000-2016).
- Design, analysis and implementation of a prediction model for foreign currency exchange rate which considers factors (interest and inflation rates and public debts) using data mining ensemble method.
- Can using the sparkling water framework improves the performance of the proposed prediction model?
- Comparison of accuracy of the independent model verses the proposed prediction model.
- Implemented prediction model, Analysed data and comparing the accuracy of independent model vs the prediction model.

1.3 Ensemble Methodology

The ensemble model is a specialised algorithm in the classifier is used to predict the new points by considering the weighted value of the points in the dataset. There are several ensemble methods including Bagging , boosting and randomised tree. The researchers suggest that ensemble models are more accurate than the independent models, that is why the ensemble model is used for predicting the forex. The ensemble of three models output is achieved to increase the accuracy of the model, in which the models considered in ensemble are Artificial Neural Network, Random forest, and Generalised Linear model, similarly Deep learning, Distributed Random forest and Generalised Linear model in Sparkling water.

1.4 Deliverables

- The developed and evaluated prediction model for the model of foreign currency exchange rate which considers factors (interest and inflation rates and public debts) using data mining ensemble method.
- The results of the analysed data and evaluation are done on these models produced the better accuracy of 95% using our ensemble model and also finding out the factor that best contribute the forex.
- The results of the comparing the independent and the prediction model, the ensemble model produced best accuracy than others.

2 Literature review of prediction models for foreign currency exchange rate which considers several factors developed using data mining ensemble method (2000-2016)

2.1 Introduction

In order to have a clear understanding of how to approach the project and to know the technologies, techniques used in the past for predicting the FOREX, the first and best part is starting with literature review. In our literature review, the research papers from 2000 to 2016 are only considered because our ideal goal of bringing accuracy can be achieved using recent technologies. The more valued and cited papers are only considered for our literature review to get strong foundation for implementing the project.

2.2 A review of Foreign exchange currency rate prediction(papers from 2000 - 2016)

There are several traditional forex rates theory models such as Interest Rate Parity, Purchase Power Parity and Balance of International payment for predicting the forex. But these models find it extremely difficult to predict after the implementation of Floating exchange rate. After that, the prediction of forex has always been a challenging task and this led to the discovery of several other models for predicting the forexLin et al. (2013).

According to Lin et al. (2013), Auto regressive moving average (ARMA) is best suited for stationary exchange rate and the models like Auto regressive Conditional Heteroskedastic (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity(GARCH) are suited for the dynamic exchange rate. After these models, there are several Non-linear models that were developed which produced better results than those ARMA, ARCH and GARCH. The several Non-Linear models are Artificial Neural Network(ANN), Genetic Algorithm (GA) and many others are providing better accuracy compared to other models.

The author Lin et al. (2013) also had the same idea like ours in handling the large amount of real-time dynamic data using big data Map reduce concepts of cloud computing. The author used Linear regression prediction model on cloud computing to predict the foreign exchange rate. The model proposed by the author produced an average accuracy of 97% which indicates that the implementation of projects can also produce better accuracy.

There is a great difficulty in predicting the forex because of its high volatility and noise. The author Yu et al. (2005) recognised that Artificial Neural Network (ANN) is powerful in forex prediction than the traditional models. The author's literature review also shows that several types of ANN are used and compared for predicting the Forex. According to author Yu et al. (2005) , the several types of ANN were used in the past such as Multilayer feed-forward network(MLFN), Recurrent network, Clustering Neural Network Model (CNN), General Regression Neural Network (GRNN) for predicting the forex. Similar to our model the author also used ensemble-based method for predicting the forex, as the author felt that single model is not sufficient for predicting the forex accurately. The authors in their implementation, they have used United states dollars(USD) against Deutsche Mark(DEM),Great Britain Pounds(GBP) and Japanese Yen(JPY) for demonstrating their model. The author Kimata et al. (2015) also used Artifical Neural Network for predicting the forex of Solomon Islands dollar (SBD)against United States Dollar (USD), Great Britain pound (GBP), Australian dollar (AUD), New Zealand dollar (NZD), and Japanese yen (JPY). The author used the specialised weighted method, in which USD and AUD have weighted value of 80% and the rest of the currencies shares the remaining portions.

The author Yu et al. (2005) used ensemble model of GLAR and ANN for predicting the forex and their results were outstanding. This idea of an ensemble in our project also helped in providing better accuracy. The author also provided a valuable information of splitting the time series data, as he collected data from January 1971 - December 2000 as the training sets and January 2001 - December 2003 as the testing sets. This process of splitting the data is followed in our project. As the model proposed by the authors has the lowest Normalized Mean Square Error(NMSE) and the Dstat and return rate (R) is high shows that the model implemented by the author is highly efficient. This author model was the main foundation for our model as our model also follows the similar type of ensemble based implementation.

The author Yao and Tan (2000),Zhang (2003) provided a great indication that Artificial neural network is best suited for forecasting the foreign exchange rate. The authors demonstrated their model by predicting the foreign currency rate of United states Dollars (USD) against Deutsch Mark(DEM), British Pound(GBD), Japanese Yen(JPY), Australian Dollar(AUD) and Swiss Franc(CHF). The author also compared their implemented model with the traditional Autoregressive Integrated Moving Average(ARIMA) which is used traditionally for predicting the Time series data. In this comparison, ANN produced a better accuracy than ARIMA with ARIMA of 50% and ANN of 73%.

Similarly, Kamruzzaman and Sarker (2003) also compared the Artificial Neural Network (ANN) with Auto-regressive Integrated Moving Average (ARIMA) for predicting the Forex. The author also proposed that ARIMA is a traditional model for time series prediction which almost used for two decades and the ANN is the most capable model for handling the time series data. Like Yao and Tan (2000) also demonstrated the prediction for forex using the United States America Dollars (USD),Singapore Dollar (SGD), New Zealand Dollar(NZD),Great Britain Pound (GBP), Japanese Yen(JPY), and Swiss Franc(CHF). The author used three models of ANN, Bayesian regression, scaled conjugate gradient and back propagation for prediction and they are compared. In their model implementation, the scaled conjugate gradient of ANN performed better than others.

The author Leung et al. (2000), observed and compared the General Regression Neural Network (GRNN) with several other models including Artificial Neural Networks type of multilayered feedforward network (MLFN) with several layers of hidden layers, several random walk models, and Multivariate transfer function for forecasting of foreign exchange rate. The author's general regression neural network (GRNN) is used for predicting the forex of three different countries currencies i.e Canadian Dollars (CAD),Great Britain Pounds(GBP)and Japanese Yen(JPY). The author's models of GRNN produced better accuracy than the existing ANN.

The model General Regression Neural Network (GRNN) implemented by the Chen and Leung (2004) is same as that of Leung et al. (2000), for the forecasting of foreign exchange rate. As the author felt that GRNN is strong structure and consumes less time for the training and prediction he also used GRNN for rectification of the error. The author developed two stage of error correction for bringing better accuracy in the artificial neural network.

According to author Liaw and Wiener (2002), the random forest can be used for both classification and regression and also stated that it is one of the best classifier models which is very robust against over-fitting. The author also demonstrated the regression based example using Boston House data and their output produced better accuracy results.Segal (2004) used random forest regression in their implementation of machine learning bench-marking. Among all the regression techniques, random forest was performing well compared to others for their implementation.

The authors Urrutia et al. (2015) implementation of mathematical model for predicting the exchange rate of Philippines is very supportive and encouraging to our project as the author also had the same idea of ours, in considering the main factors that influence the exchange rate like interest rate, inflation rate, import and export of trade using Auto-regressive integrated Moving Average (ARIMA). The author implemented several mathematical models to justify the results and have done a detailed study of each factor using the Multiple Linear regression (MLR). The author concluded that Interest rate and Labour rate has the main contribution.

As the author Urrutia et al. (2015) suggested that considering the factor influencing, the exchange rate in its prediction plays an important role. I wanted to do further research with Patel et al. (2014) to understand the main factors really contributing to the exchange rate prediction. The authors theoretically provided a promising evidence that the main factors influencing the exchange rate are Inflation rate, Interest rate, Capital Account balance, Role of speculators, Cost of Manufacture, Debt of the Country, Gross Domestic Product, and political stability Cărbureanu (2011).

These experimental results give a clear idea that the model proposed by the authors accomplished better prediction accuracy and performance.

2.3 Identified the gaps

The authors Leung et al. (2000),Kamruzzaman and Sarker (2003),Yu et al. (2005),Kimata et al. (2015),Lin et al. (2013) have done a great amount of work by developing several prediction models for predicting the forex in the field of finance. As most of the authors except Urrutia et al. (2015) and Yao and Tan (2000) are really focused on predicting the forex with several models and finding out which one better suits for forex prediction. They should also consider other attributes which also mainly contributes to the prediction of forex.

2.4 Conclusion

As these literature review is backbone of our project that really guided me to implement our ensemble model for prediction of forex. The literature review from Yao and Tan (2000), Chen and Leung (2004), Kamruzzaman and Sarker (2003),Czekalski et al. (2015) provided us that Artificial Neural Network(ANN) model has performed better than other model for predicting the Forex, thereby I have considered deep learning algorithm as one of our prediction model which act as the ANN for better accuracy. The author Yu et al. (2005), gave rigid foundation to our project that ensemble model can produce better accuracy than single model. Similarly, Lin et al. (2013) also supported our idea of using big data for prediction with his cloud computing model produced better accuracy. Especially,Urrutia et al. (2015) was the base idea to implement our project by considering the influencing factors for prediction. Thereby taking all these positive and supportive ideas I was able to complete my project successfully with great efficiency, accuracy and better performance.

3 Proposed Solution: Foreign Exchange Rate Prediction Model

To predict the foreign exchange rate accurately, we proposed a data mining ensemble model which is an ensemble of Artificial Neural Network, Regression based Random Forest, and Generalised Linear Model using Sparkling water. As per our analysis, the forex is one of the challenging and difficult markets to predict, several authors Yu et al. (2005), suggested that single model is not sufficient to predict the forex accurately hence in this research an ensemble of three different model is used. Due to the large volume of data in the real-time the data from 1995 to 2016 (In this research demonstration the dataset from 2000-2016 is only considered because all the research are done in same time frame and the modern era is only considered) which can affect the performance of our ensemble model. The data in this research also includes the dataset of factors (Interest rate,Inflation rate(Consumer prize index), Gross domestic value,Shares etc.) which mainly influences the forex. By taking this performance issue under consideration, in this research, big data technology of sparkling water is used. This project involves several processes of implementation from requirements gathering, Data collection, development of the architecture, pre-processing of data, data analysis and the evaluation are explained below. This proposed model of implementation as per our expectation produced the best results of accuracy.

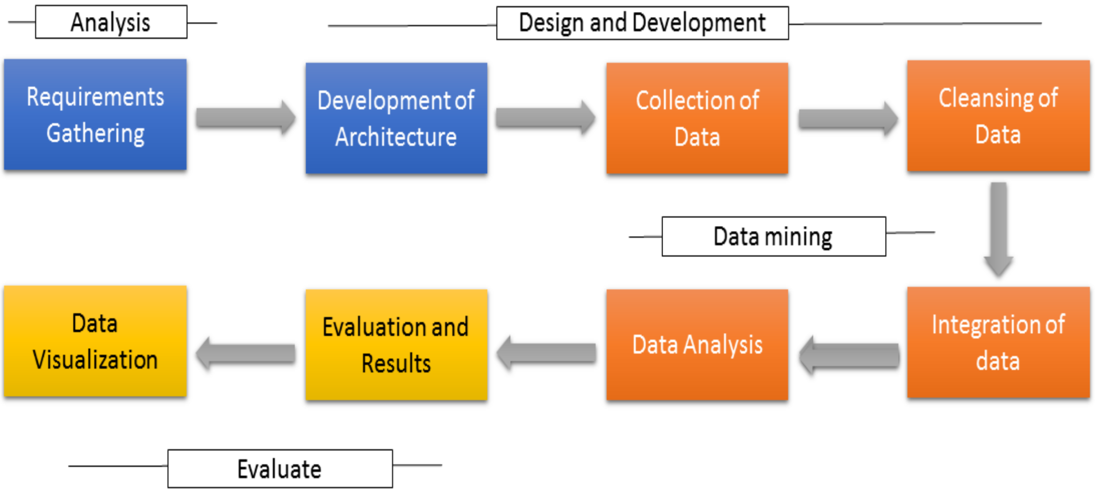
3.1 Introduction

Our goal is to predict the foreign currency exchange rate(Forex) more accurately by implementing the proposed ensemble prediction model using sparkling water. In this process of implementation several phases are involved, which are requirement gathering, development of proposed architecture and data mining models, data analysis, presentation of results, and comparison of the independent model with the ensemble model are performed successfully. The complete information about each phase is explained in detail below.

3.2 Implementation of Prediction model

Every implementation has to follow system development life cycle, that are Analysis, Design, Development and Evaluate. In the process of Analysis, we first analysed our main goal of predicting the forex accurately and efficiently. In analysis, we did several literature review which is discussed above, helped us to find out the best model for prediction,identifying factors that influences the forex rates, handling data for splitting test and train, and evaluating the results.

In the analysis, the requirements gathering are done to find the type of data, the availability of data, storage system, tools for processing, analysing the data and visualizing the results.In the process of design and development, the design of the architecture is developed, implementation of proposed architecture and model are conducted. Especially, in the development phase data exploration, pre-processing, and data analysis using proposed models are achieved.The results are presented for the evaluation and visualization.



Process Flow for Implementation of Ensemble Prediction Model

Figure 1: Process Workflow diagram

In order to make sure that the process flow implementation are conveyed properly, every process of the process flow are provided with detailed explanation in the following discussions.

3.2.1 Requirement Specification

The requirement specification is the most important part of the process flow of implementation, as we analyse what are the main requirements to complete this project successfully and to complete on time. If this process is not done, there might have been a lot of chance where the project might have struck in the middle of the process. The requirement specification has helped us to know, what are the data, what type of tools and technologies are required to complete this project. As this process is initially done which helped our project to complete successfully.

Data Collection:

The first process of the requirement specification is the collection of the dataset, which is the main and prioritised requirement. It is unimaginable to complete this project successfully without finding the proper dataset. Since that project is about predicting the Forex we first collected Forex datasets. In order to evaluate the model perfectly, the currencies exchange rate of Indian Rupees - INR against United State Dollars - USD, Canadian Dollars - CAD and Great Britain Pounds - GBP are used and the data collection are explained in detail below.

The historical data of foreign currencies exchange rate of Indian Rupees - INR against United State Dollars - USD are collected from http://www.investing.com, the Canadian Dollars - CAD and Great Britain Pounds - GBP are from https://www.oanda.com. As per the literature review suggestions, the factor that influenced the Forex rate are considered and their respective dataset are collected separately.

The factors that mainly influencing the Forex rate are, Interest rate,Gross Domestic Product, Inflation rate / Consumer prize index, Debt of the country,Producer price index, Share Prize, and Gold prize.

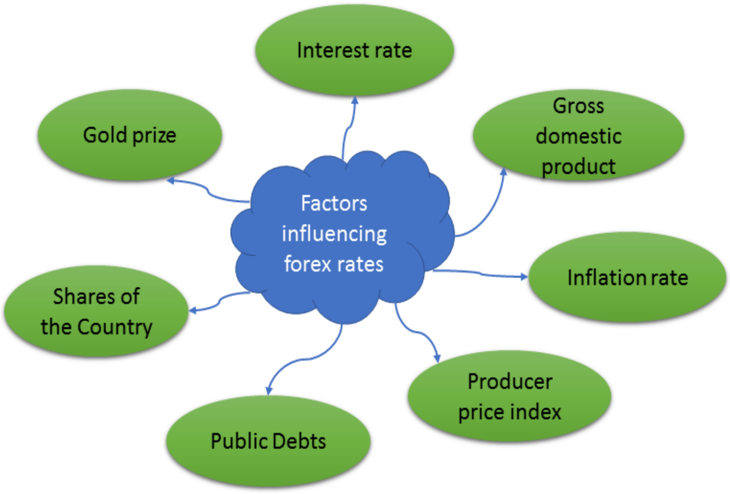


Figure 2: Factors influencying the forex

The interest rate is one of the major factor that influences the Forex rate, it is the rate of interest charge by the country for any financial transaction. According to compareremit (2016) and Patel et al. (2014), the interest rate is highly correlated with Forex rate. The country with high interest rate will cause severe impact on the local business of the country. The country with higher interest rate decreases the purchase power of consumers and also people who plan to take loans has to pay high interest which will ultimately reduce the investors. The data for the interest rate of United states, Canada and United kingdom are collected from http://www.federalreserve.gov, http://www.bankofcanada.ca/rates/ and http://www.bankofengland.co.uk respectively.

Gross Domestic Product is the measure of goods and services of one's countries, and it also signifies the economic health of the country. It is mainly based on complete expenses made by the country for their businesses, private activities and exports of goods. The country with higher GDP indicates that it has good economic growth which brings the focus of the investors Patel et al. (2014), Kayal (2010). The country with higher GDP has the lower foreign currency exchange rate. The dataset for GDP of all three countries(USA,Canada and U.K) is taken from https://data.oecd.org/gdp/.

The inflation rate determines the prize of the product in a country and it mainly depends on the export strength of the country. The country with low inflation rate will attract the investors to invest more in these countries therefore, inflation rate has the important role in the valuation of the foreign currency exchange rate investopedia (2013). In our project, the Consumer prize index highly correlate with the inflation rate and was finding hard to collect the data set of the inflation for all the three countries of the same duration. Hence, the dataset for consumer prize index is considered and collected from http://stats.oecd.org/ for all the three countries.

The country with more expenditure and also with low-income will cause debt to the country, this also reflects their economy. No countries in the world will get more attracted towards the country with high debt this makes no investors to invest in countries with high debts. So the country with high debts will have high exchange rate Uddin et al. (2013). The datasets for the debts of all the three countries are collected from http://stats.oecd.org/.

The producer prize index is a weighted index to represent the countries producing power from small scale to large scale. It shows the producing power of the wholesale,commodities market and manufacturing industries, this PPI helps in understanding the CPI of the country. Since CPI is highly correlated with exchanges the PPI is also considered for the prediction of Forex. The dataset for the PPI is collected from https://data.oecd.org/price/ for all three countries investopedia (2013).

The financial economy of the countries also determined by the financial shares by the countries. The dataset for the shares for all the three countries are collected from https://data.oecd.org/. The Gold price is also an another factor considered for the prediction of the Forex, The dataset for the Gold for all the three countries are collected from http://www.gold.org.

All these datasets consist of the historical values of their respective fields and the data within the range of 2000 to 2016 is only considered for our analysis.

Environmental Specification,

In order to develop any system, the analysis of the environmental specification is an important factor if not it will delay our work in progress. The environmental specification includes the hardware and the software specifications. The hardware specification includes the Laptop specification and the software specification includes Virtual machine specifications, operating system specification, and the software for developing, evaluating and visualising the model. In order to implement our project, the high configuration system is required as it has to handle a large amount of data using Big data and also to speed up the process of data analysis. If low configuration system is used it slows down our process by consuming lot of time while training and testing the model during its evaluation. The hardware configuration used this project are , the HP Envy Series Laptop with Intel i7 processor, 12 Gigabytes of RAM for high-performance computation, it also has Windows 10 operating system for interacting with the machine.

The Software specification are the software that are required and used in this project implementation because any improper usage of the software leads to inconsistency in the project. Our project uses big data so the proper information is important to conduct the analysis in big data using sparkling water. To achieve this, the model of the parallel processing system is setup using Hortonwork sandbox which is a portable Hadoop environment and also provided all required software for processing big data like Apache spark, hive, Hadoop distributed file system (HDFS). In order to use Hortonwork sandbox, the oracle virtual box is installed in the Windows 10 and then the sandbox image is loaded into it. It also required the google chrome browser to interact with the Hortonwork GUI. In this project, the putty software is used to interact with the sandbox through the terminal for loading files and starting other applications etc. Using the terminal, the Sparkling water software application is downloaded and installed in the Hortonwork sandbox. The other softwares that are used in our project are, R studio for data analysis without big data, Apache spark which comes along with Hortonwork sandbox for starting and executing Sparkling water, and tableau for visualisation.The other requirements are strong technical knowledge with R- programming,Mathematical stats knowledge, Spark, Hive, and H2o.

After acquiring all these requirement specifications, the project was successfully completed for predicting the Forex with better accuracy.

3.2.2 Architecture Design

In order to develop a strong and robust model for predicting the Forex, a better architecture is developed that delivers the best accuracy,efficiency and performance. Architecture diagram has been a great support for making this project successful, it has a dramatic representation of our strategy followed to make this project successful. In this architectural design, it discusses what kind of tools and techniques are used, how they are used, in what order they are used and the interaction between them in order to attain our goal of Forex prediction accuracy.

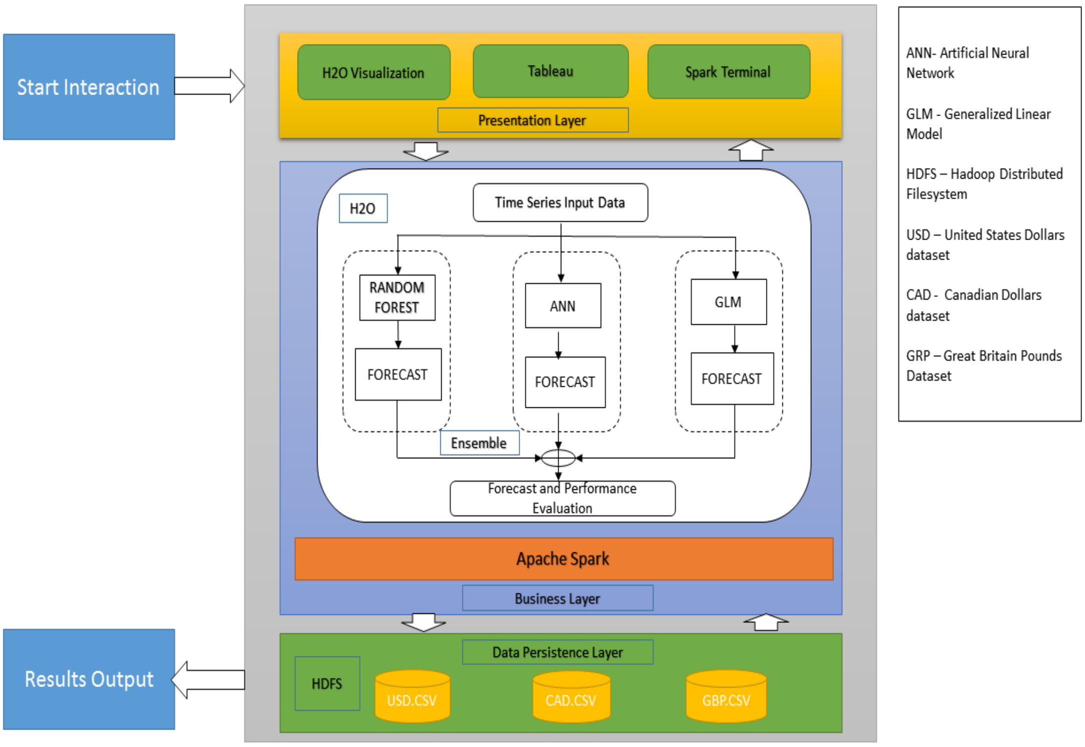


Figure 3: Architecture Diagram followed

Any architecture in the Information Technology like Software development, Data analytics, Data modelling system follows three tier -architecture, i.e Data persistence Layer, Business Layer, and Presentation Layer and the same type of architecture is followed in our system. The Data persistence Layer deals with the type of data is used, how it is used, where is stored and how other layers can interact with it are explained in this layer, the business Layer deals with all computation and the logic of our models are implemented in this layer and in the presentation layer, how the results are provided to the users are implement in this layer.

As discussed above, the data persistence layer has information about what storage system is used, why it is used, what type of data, are discussed here. In this project, the data storage system used is Hadoop and the reason choosing it is our model uses a large amount of historical data. It also supports the parallel processing which provides high computation facility by which we can produce high accuracy Wu et al. (2014) Deng et al. (2014). That is the reason, that the storage system used in this project is Hadoop distributed file system (HDFS). This entire HDFS system is available in the form of virtualized application using Hortonwork sandbox. The data collected for all the three countries including their respective factors datasets are integrated and stored in this Hdfs system and the type of data used is comma separated value file (.CSV) is used for our computation. There are three CSV files in the HDFS called USD.csv,CAD.csv, and GBP.csv which are integrated FOREX data of United States of America, Canada and the United Kingdom respectively.

The business layer, where all the computations are taking place, the computations includes processing of data, training and testing of the model, prediction of the data are taking place. In order to make use of the big data to its maximum, the sparkling water (Spark + H2o) is used for the spark cluster computing system with the deep learning of H2o. This helps in easier and faster data processing. This is the layer where our ensemble model of deep learning, Distributed Random forest, and deep generalized linear regression is used for predicting the FOREX.

The presentation layer, helps the users to interact with the business layer where the model resides and the output of the model are evaluated and presented to the users. The H2o has a specialised GUI, where the users can interact with it, where the flow is created and the performance of the model is shown. The system is also provided with Spark terminal to scale program for interacting with data and the model. In this project, the visualisation tool tableau is also used to compare the performance of the model by comparing the with the actual FOREX value with the predicted FOREX values.

3.2.3 Data Analysis

This phase of our project is a very important phase because the raw data collected from several websites are treated with several techniques to make them useful for the model to predict the foreign exchange rate effectively and accurately. This phase involves the complete data analysis which involves the integration of data where all different datasets are integrated to each other to make a single complete data set fit into the model, the integrated data is then cleansed so that data is free from noise and NAs, the cleansed data has to undergo some transformation like converting the date to several useful information like which day of the week, is the day is weekday or weekend, etc. , the transformed data is now completely fit to the model for prediction, then these data are transferred to data analysis using data mining techniques where the training and the testing are done and the last part of the analysis are presenting and evaluating the results.

3.2.4 Data Integration:

As discussed in the data collection process in this project, that is collected from several resources. The data for each feature are collected from different places, hence this data needs to integrate in order to make the data properly fits into our prediction model. The seven separate data set of FOREX rates dataset and its factors are integrated into a single complete dataset for each country.

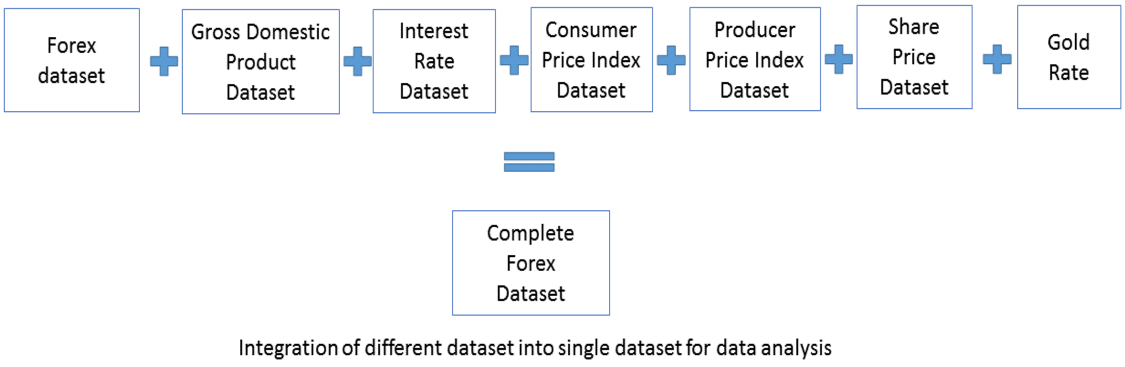
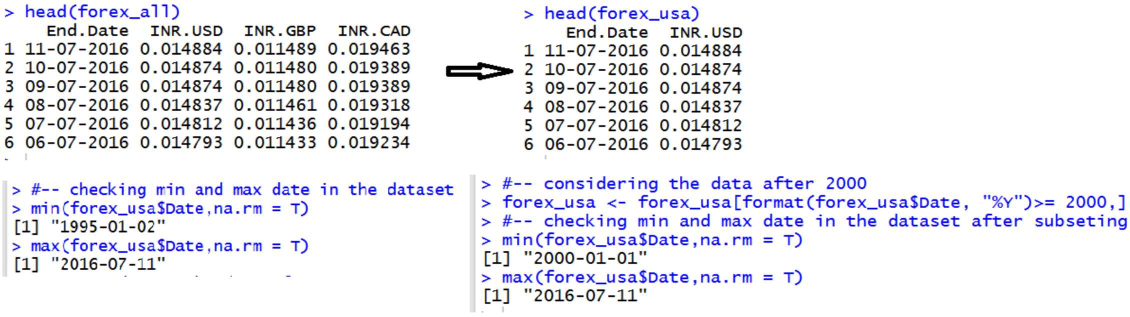

Forex dataset
Gross Domestic Product Dataset
Interest Rate Dataset
Consumer Price Index Dataset
Producer Price Index Dataset
Share Price Dataset
Gold Rate
Complete Forex Dataset
Integration of different dataset into single dataset for data analysis

Figure 4: Integration of the data

In most of the past implementation, the monthly and quarterly based prediction are only done but in this project daily basis prediction of Forex are achieved. To predict the daily foreign currency exchange rate, the Forex data collected consist of daily Forex rates of all the three countries from 1995 - 2016 are taken from which only 2000 - 2016 are only considered for demonstration.



```
> head(forex_all)
    End.Date  INR.USD  INR.GBP  INR.CAD
1 11-07-2016 0.014884 0.011489 0.019463
2 10-07-2016 0.014874 0.011480 0.019389
3 09-07-2016 0.014874 0.011480 0.019389
4 08-07-2016 0.014837 0.011461 0.019318
5 07-07-2016 0.014812 0.011436 0.019194
6 06-07-2016 0.014793 0.011433 0.019234
```

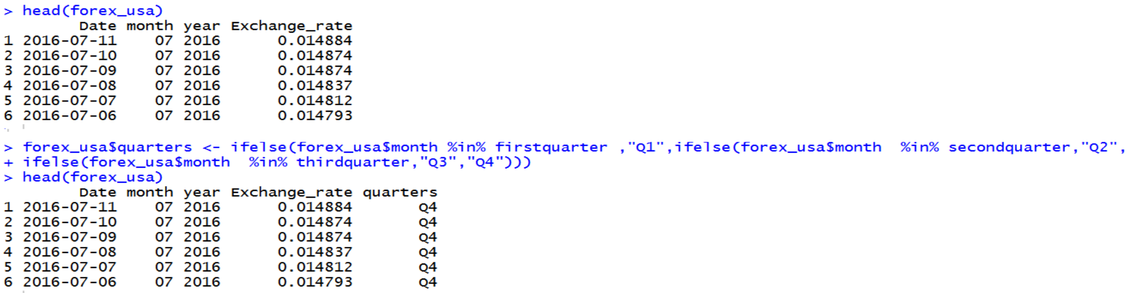
```
> head(forex_usa)
    End.Date  INR.USD
1 11-07-2016 0.014884
2 10-07-2016 0.014874
3 09-07-2016 0.014874
4 08-07-2016 0.014837
5 07-07-2016 0.014812
6 06-07-2016 0.014793
```

```
> #-- checking min and max date in the dataset
> min(forex_usa$Date,na.rm = T)
[1] "1995-01-02"
> max(forex_usa$Date,na.rm = T)
[1] "2016-07-11"
```

```
> #-- considering the data after 2000
> forex_usa <- forex_usa[format(forex_usa$Date, "%Y")>= 2000,]
> #-- checking min and max date in the dataset after subseting
> min(forex_usa$Date,na.rm = T)
[1] "2000-01-01"
> max(forex_usa$Date,na.rm = T)
[1] "2016-07-11"
```

Figure 5: Integration process 1

The first step of the integration process is integrating the Forex dataset with the respective to their GDP of the country. The dataset collected for GDP - Gross Domestic Product is not the daily basis it is the quarterly basis data hence we need to merge that dataset based on the quarter the date falls, for example, 02/01/2016 falls in the quarter Q1.

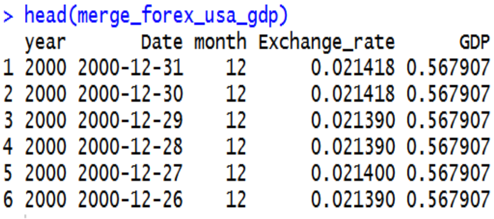


```
> head(forex_usa)
        Date month year Exchange_rate
1 2016-07-11    07 2016      0.014884
2 2016-07-10    07 2016      0.014874
3 2016-07-09    07 2016      0.014874
4 2016-07-08    07 2016      0.014837
5 2016-07-07    07 2016      0.014812
6 2016-07-06    07 2016      0.014793

> forex_usa$quarters <- ifelse(forex_usa$month %in% firstquarter ,"Q1",ifelse(forex_usa$month  %in% secondquarter,"Q2",
+ ifelse(forex_usa$month  %in% thirdquarter,"Q3","Q4")))
> head(forex_usa)
        Date month year Exchange_rate quarters
1 2016-07-11    07 2016      0.014884       Q4
2 2016-07-10    07 2016      0.014874       Q4
3 2016-07-09    07 2016      0.014874       Q4
4 2016-07-08    07 2016      0.014837       Q4
5 2016-07-07    07 2016      0.014812       Q4
6 2016-07-06    07 2016      0.014793       Q4
```

Figure 6: Integration process 2

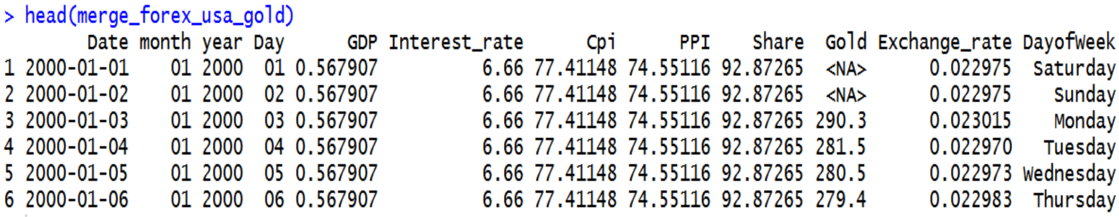
After the merging, the dataset, the combined two datasets are shown below,



```
> head(merge_forex_usa_gdp)
  year       Date month Exchange_rate      GDP
1 2000 2000-12-31    12      0.021418 0.567907
2 2000 2000-12-30    12      0.021418 0.567907
3 2000 2000-12-29    12      0.021390 0.567907
4 2000 2000-12-28    12      0.021390 0.567907
5 2000 2000-12-27    12      0.021400 0.567907
6 2000 2000-12-26    12      0.021390 0.567907
```

Figure 7: Integration process 3

Similarly, the other factors such as Interest Rate and Gold price are data present in the daily term. Therefore the merging using the date column are done on the dataset containing GDP, Interest rate and gold rate. The factor CPI- Consumer Price Index dataset, the values are available for the months, but the challenges faced is the date column has to date in the format "Jan 2012" which I have merged it with the date. Hence I have to convert that date format to dd/MM/yyyy for example "01/01/2012". Then both the dataset are merged based on month and year, similar to CPI the PPI dataset - Producer Price Index, Share dataset also of the same format hence the same process is followed to merge the dataset. After the integration of all the dataset, the complete dataset consists of Date, Month, Year, Day, GDP, Interest Rate, CPI,PPI, Share, Exchange rate and gold price which is shown below.



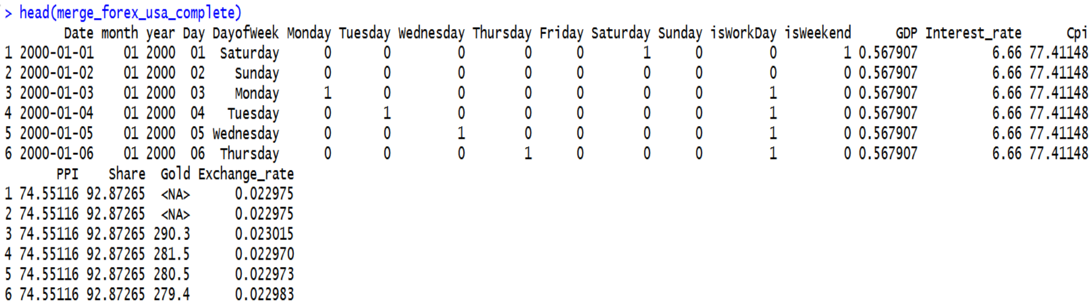
```
> head(merge_forex_usa_gold)
        Date month year Day      GDP Interest_rate      Cpi      PPI    Share  Gold Exchange_rate DayofWeek
1 2000-01-01    01 2000  01 0.567907          6.66 77.41148 74.55116 92.87265  <NA>      0.022975  Saturday
2 2000-01-02    01 2000  02 0.567907          6.66 77.41148 74.55116 92.87265  <NA>      0.022975    Sunday
3 2000-01-03    01 2000  03 0.567907          6.66 77.41148 74.55116 92.87265 290.3      0.023015    Monday
4 2000-01-04    01 2000  04 0.567907          6.66 77.41148 74.55116 92.87265 281.5      0.022970   Tuesday
5 2000-01-05    01 2000  05 0.567907          6.66 77.41148 74.55116 92.87265 280.5      0.022973 Wednesday
6 2000-01-06    01 2000  06 0.567907          6.66 77.41148 74.55116 92.87265 279.4      0.022983  Thursday
```

Figure 8: Integration process 4

The same process of integration is followed by the other two countries thereby our project at the end has three complete dataset for USD,CAD and GBP respectively.

3.2.5 Data Transformation:

In order to understand the data in detail and also help the model to understand the data easily, the further transformation is used. The transformation that is done in our dataset are bringing the day of the week from the date column, similarly, the column with binomial value for each day in the week for example if the day is Monday, the value in the column Monday has 1 and the other columns for rest of the day in a week is 0. Similarly, the other transformation used in our dataset is to check whether the day is working day or the weekend. All these transformations are done to the dataset so that it helps the model to predict the Forex accurately.The dataset at the end looks as shown in the below figure.



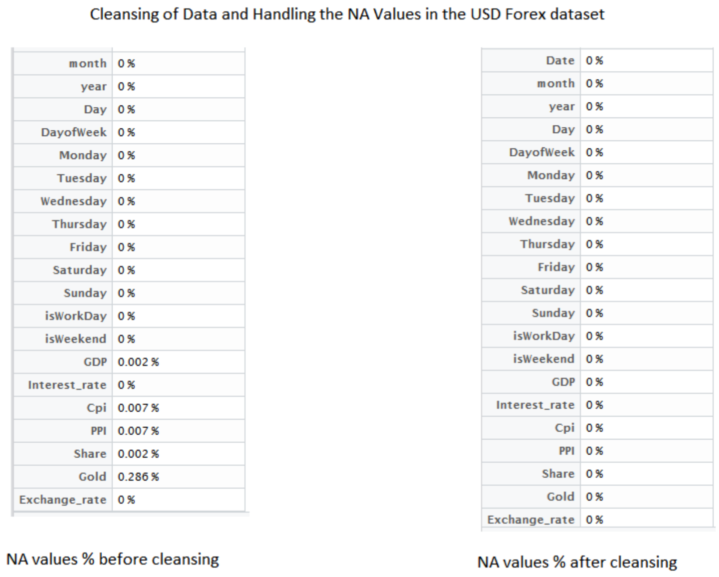
```
> head(merge_forex_usa_complete)
        Date month year Day DayofWeek Monday Tuesday Wednesday Thursday Friday Saturday Sunday isWorkDay isWeekend      GDP Interest_rate      Cpi
1 2000-01-01    01 2000  01  Saturday      0       0         0        0      0        1      0         0         1 0.567907          6.66 77.41148
2 2000-01-02    01 2000  02    Sunday      0       0         0        0      0        0      0         0         0 0.567907          6.66 77.41148
3 2000-01-03    01 2000  03    Monday      1       0         0        0      0        0      0         1         0 0.567907          6.66 77.41148
4 2000-01-04    01 2000  04   Tuesday      0       1         0        0      0        0      0         1         0 0.567907          6.66 77.41148
5 2000-01-05    01 2000  05 Wednesday      0       0         1        0      0        0      0         1         0 0.567907          6.66 77.41148
6 2000-01-06    01 2000  06  Thursday      0       0         0        1      0        0      0         1         0 0.567907          6.66 77.41148
       PPI    Share  Gold Exchange_rate
1 74.55116 92.87265  <NA>      0.022975
2 74.55116 92.87265  <NA>      0.022975
3 74.55116 92.87265 290.3      0.023015
4 74.55116 92.87265 281.5      0.022970
5 74.55116 92.87265 280.5      0.022973
6 74.55116 92.87265 279.4      0.022983
```

Figure 9: Data Transformation

3.2.6 Data Cleansing:

The cleansing of the dataset is very important that any noise in the data may lead to improper results which are not good for the research. Hence the dataset that was integrated and transformed are now cleansed in this step. In this process of cleansing, several cleansing processes like the handling of the NA values, blank values, outliers are done in order to make sure that our data is out of the noise. For this, our first step is finding the NA values in the dataset which is shown in the below figure.

It is clearly shown in the Left figure that among all 21columns only 5columns contains NA values and hence a proper way of handling these NA values is important otherwise it also might have led to inconsistency in the data. Therefore these 5 columns with NA values are initially considered for cleansing. The NA values in the column GDP available for all rows where the



Cleansing of Data and Handling the NA Values in the USD Forex dataset

month	0 %
year	0 %
Day	0 %
DayofWeek	0 %
Monday	0 %
Tuesday	0 %
Wednesday	0 %
Thursday	0 %
Friday	0 %
Saturday	0 %
Sunday	0 %
isWorkDay	0 %
isWeekend	0 %
GDP	0.002 %
Interest_rate	0 %
Cpi	0.007 %
PPI	0.007 %
Share	0.002 %
Gold	0.286 %
Exchange_rate	0 %

NA values % before cleansing

Date	0 %
month	0 %
year	0 %
Day	0 %
DayofWeek	0 %
Monday	0 %
Tuesday	0 %
Wednesday	0 %
Thursday	0 %
Friday	0 %
Saturday	0 %
Sunday	0 %
isWorkDay	0 %
isWeekend	0 %
GDP	0 %
Interest_rate	0 %
Cpi	0 %
PPI	0 %
Share	0 %
Gold	0 %
Exchange_rate	0 %

NA values % after cleansing

Figure 10: Cleansing of the data

month is July and year of 2016 which is obvious that GDP is available for two quarters of this year hence it is replaced with the last GDP value of the previous month because there could be slight change from past value so the recent value itself is taken as the current value for GDP and GDP doesn't change every month it changes quarterly.Similarly, all the other columns except Gold prize the NA values are replaced with the recent values and the Gold prize column the NA values are replaced with the Average of the Gold prize in that week.After treating NA values, the above-shown right figure shows that our data free from inconsistency with 0% NA values.

3.2.7 Implemented algorithm components for the prediction model:

The data mining technique is used for gathering the useful information from the dataset, and then further studies are done on top of it. Our main goal of this project is to predict the Foreign currency exchange rate (Forex) effectively and accurately using ensemble model using sparkling water. Not only predicting the Forex but also identifying and comparing the performance of the independent models with ensemble model. And so finding out, how effective that deep learning can predict the Forex when compared to ANN and other models.

1. Artificial Neural Network

   The Artificial Neural Network (ANN) in one of our ensemble model, the reason for considering this is because ANN is one of the famous and successful models for handling and predicting the time series data. It has its application in several fields like Recognition of Handwriting,Signature classification, Image recognition, Facial recognition etc. as it is the best model for predicting the time series data Eng et al. (2008), Freeman and Skapura (1992). ANN has its several types among them Perceptron and Multilayer perceptron (MLP), are the famous and most used type of the ANN.The Perceptron, consist of single input and output layer, where it consist of a Specialised function called Linear Threshold unit (LTU). It takes a number of inputs to predict the single output Gan and Ng (1995). In Perceptron, all the inputs are represented as x1,x2,x3,... Xn and the weights on the arc are represented as w1,w2,w3,.....wn. The LTU is a specialised unit where the summation of the weights and threshold are calculated. After the LTU, then the output from LTU is passed to Step Linear summation and threshold unit, where the output is predicted.

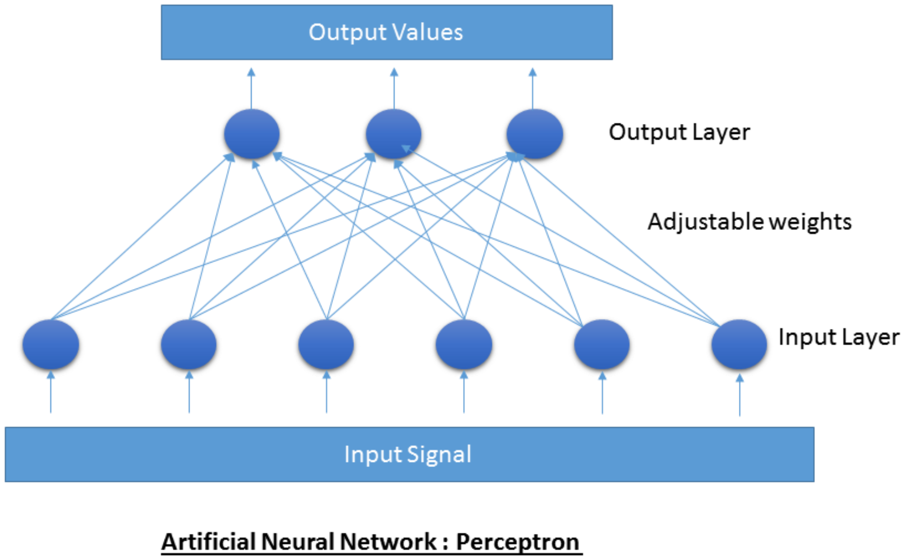


Figure 11: Artificical Neural Network

The above-shown diagram is an example of perceptron, in which all the circle are denoted as node, the first 6 neurons forming the first layer called input layer, these neurons are connected to 3 other neurons which forming an another layer called output layer and the connection between the neurons are also called as Adjustable weights or Arcs which has its own weights.

The algorithm followed in the perceptron are explained below,

(a) The N inputs provided by the user for computation are represented as

$$i_p = x_1, x_2, x_3, ..., x_n$$

and the weight across these inputs arc are represented as

$$W_i = w_1, w_2, w_3, ..., w_n$$

.

(b) Let us assume that, **T** is our expected output for all those N inputs of X and also consider that output obtained is represented as **O**. The core concept of the perceptron is to make sure that expected **T** and the obtained **O** results are same i.e

$$T = O$$

.

(c) The weight of the arc determines the output of the model. The weight of the arc calculated using the formula.

$$W_i = w_i + \Delta w_i$$

where

$$\Delta w_i = \eta(T - O)x_i$$

where

$$\Delta w_i$$

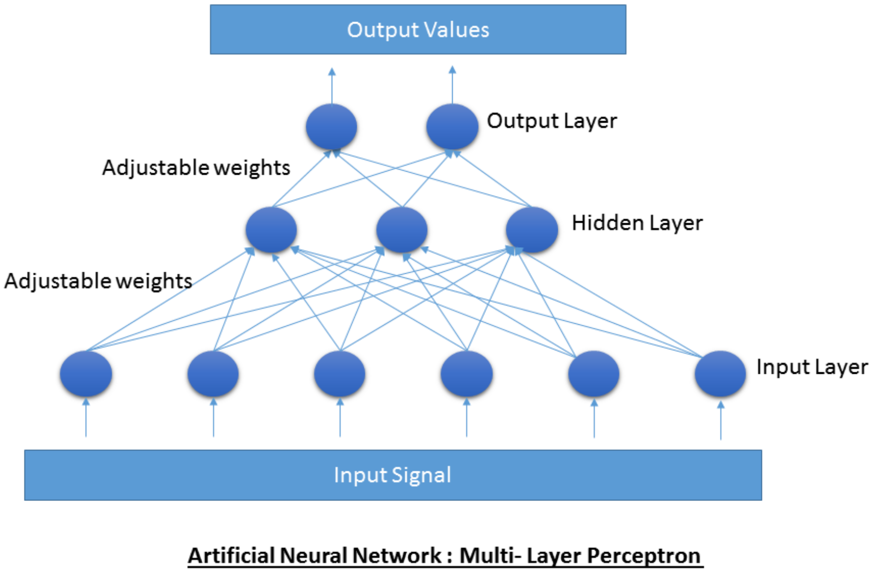


Figure 12: Artificical Neural Network - Multilayer Perceptron

is also called as the incremental factor and

$$\eta$$

is called as the learning rate i.e. the rate at which the model learns to predict the output value.

(d) The step three is repeated until the expected and obtained are same. In the ANN, epoch also represents the learning rate of the model

The Multilayer perceptron (MLP),is a more specialised perceptron where it consists of several layers, in which the first layer is called as an input layer, the last layer is called output layer, and all the other intermediate layer as called as hidden layers Gan and Ng (1995).

The above-shown figure is an example of Multilayer perceptron, where it consists of more than two layers i.e three layers. The first layer is the one where the inputs are provided to it and it is called as an input layer, the neurons in these input layers are connected to neurons in another intermediate layer called hidden layer, and the last layer where the output are presented is called output layer. In our project, all the independent attributes like date, gross domestic product, interest rate, consumer price index, etc. are fed to the input layer and the output of the exchange rate are obtained at the output layer Chandar et al. (2014),Eng et al. (2008).

2. Deep Learning

It can also be called as modernised and specialised neural network that provides stability while training the model, and also provides high scalability while handling the large data. Because of it, the deep learning is implemented in all fields where the high accuracy of prediction are expected Candel et al. (2015).It is also similar to machine learning Artificial neural network, the only difference is DL can handle complex non - linear data. It is mainly designed based on the feed-forward network on ANN, but there are several researches delivered than even the recurrent neural networks has got more applications. The deep learning neural network is more flexible and comfortable in training with back propagation algorithm.

The training of the deep learning neural network uses back propagation, where the weights are calculated using the formula given below,

$$w_ij(t+1) = w_ij(t) + \frac{\eta\partial C}{\partial w_ij} + \xi(t)$$

In the given equation, the , $\eta is the learning rate, C is the cost function and \xi(t) a stochastic term$

3. Random Forest

   Random forest is the cluster of trees, is also an ensemble learning model used for the classification and regression based problems. The prediction in the random forest is the voting where the majority of the falls based on each tree. In our project, the random forest is used as the regression based in order to predict the Forex. The random forest takes three inputs, the dataset, the number of trees and the depth of the tree. Based on the number of trees, for each and every tree it creates the binary tree based on the depth of the tree. During training, the predictors are added at the end of the tree as leaves as a binomial tree. The K-means are done on top of the tree, where the voting of the predictors on individual trees are done and then the predicted values are provided Liaw and Wiener (2002) .

4. Generalised Linear model (GLM):

   It is an another form of Linear regression model, where it uses the continuous independent variable over the continuous or categorical dependent variable. It uses three important statistical model including ANOVA, ANCOVA. The model is represented by the formula

   $$y_i = N(x_i^{\beta}, \sigma^2)$$

   where the xi is co-variates and b is coefficients, the model has a hypothesis of Yi is exponential distribution and $\mu i i s assumed to be non-linear$.

   The GLM has three main components , they are Random, Systematic, and Link Function. The Random components comprises of distribution based on the probability of all those dependent variable. It uses linear regression for continuous variable and logistic regression for the categorical variables. The Systematic component mainly contributes the independent variable, and the Link function informs the relation between the independent and dependent variable Nelder and Baker (1972) .

   Based on all these three models the ensemble in done on the outputs to bring the better accuracy of the prediction in the Forex rate.The model flow diagram are shown below in (Fig.13).

5. Data mining in Sparkling Water:

   The data that are integrated, cleansed, transformed, is now ready for prediction hence that data is split into three datasets, one for training, validating, and testing. The training dataset consists of all the historical data from 2000 to 2015 by which all the three models are getting trained,the validating dataset consists of data randomly picked 1500 rows from the complete dataset so that for every iteration the model verifies the dependent column of the exchange rate. and the test data consists of data only for 2016 which we are really trying to forecast. Once the data is split, it is moved to the Hadoop distributed file system (HDFS) for analysis. The same procedure is followed for all the three countries so that the consistency is maintained in the data and the results.

   The data that are successfully split is now loaded in the Hadoop distributed file system (HDFS) is shown in the below diagram. The figure (Fig.14) shows there is a separate

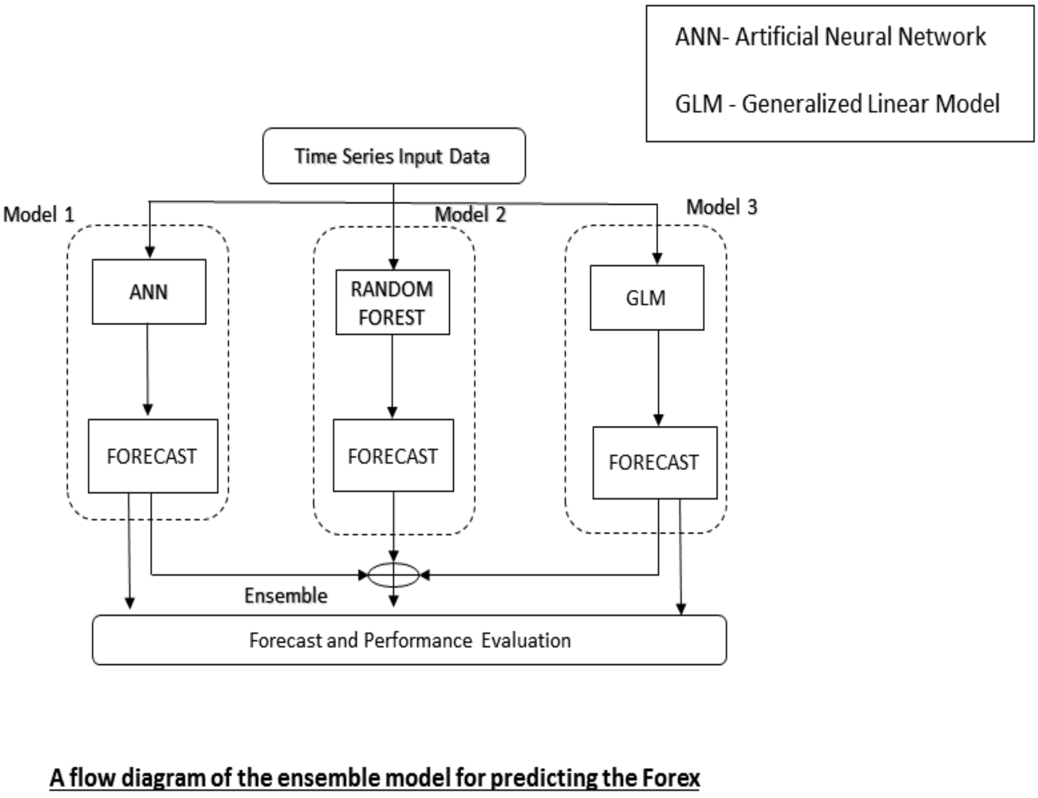
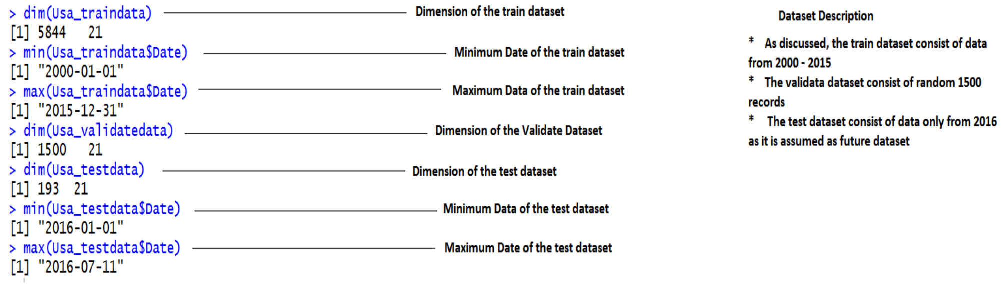

ANN- Artificial Neural Network
GLM - Generalized Linear Model
Time Series Input Data
Model 1
Model 2
Model 3
ANN
RANDOM FOREST
GLM
FORECAST
FORECAST
FORECAST
Ensemble
Forecast and Performance Evaluation
A flow diagram of the ensemble model for predicting the Forex

Figure 13: Model Flow Diagram



```
> dim(Usa_traindata)          ———————————— Dimension of the train dataset
[1] 5844   21
> min(Usa_traindata$Date)     ———————————— Minimum Date of the train dataset
[1] "2000-01-01"
> max(Usa_traindata$Date)     ———————————— Maximum Data of the train dataset
[1] "2015-12-31"
> dim(Usa_validatedata)       ———————————— Dimension of the Validate Dataset
[1] 1500   21
> dim(Usa_testdata)           ———————————— Dimension of the test dataset
[1] 193   21
> min(Usa_testdata$Date)      ———————————— Minimum Data of the test dataset
[1] "2016-01-01"
> max(Usa_testdata$Date)      ———————————— Maximum Date of the test dataset
[1] "2016-07-11"
```

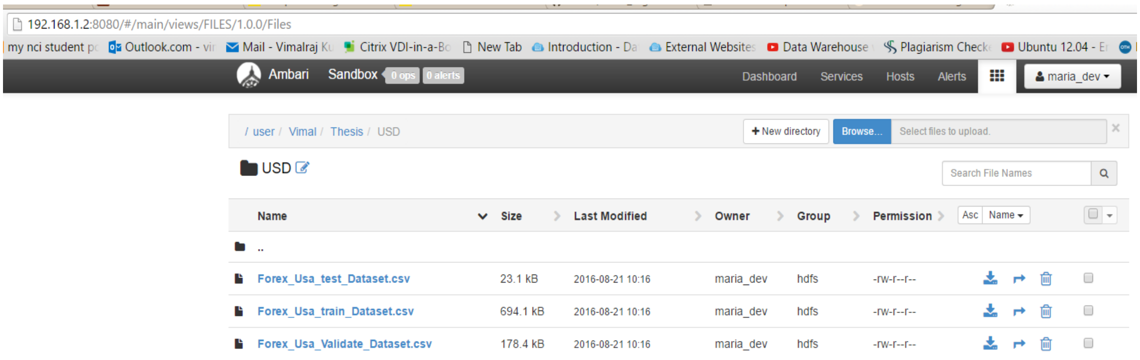
Dataset Description

* As discussed, the train dataset consist of data from 2000 - 2015

* The validata dataset consist of random 1500 records

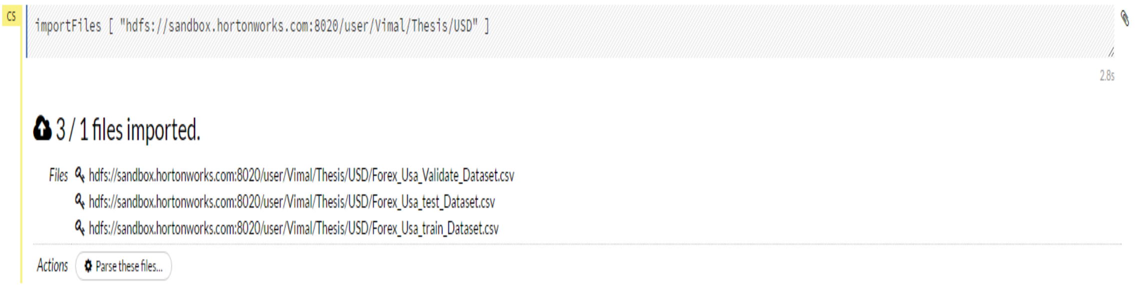
* The test dataset consist of data only from 2016 as it is assumed as future dataset

Figure 14: Distribution of Dataset



Name	Size	Last Modified	Owner	Group	Permission
..					
Forex_Usa_test_Dataset.csv	23.1 kB	2016-08-21 10:16	maria_dev	hdfs	-rw-r--r--
Forex_Usa_train_Dataset.csv	694.1 kB	2016-08-21 10:16	maria_dev	hdfs	-rw-r--r--
Forex_Usa_Validate_Dataset.csv	178.4 kB	2016-08-21 10:16	maria_dev	hdfs	-rw-r--r--

Figure 15: Load data to Hdfs


```
importFiles [ "hdfs://sandbox.hortonworks.com:8020/user/Vimal/Thesis/USD" ]
```

3 / 1 files imported.

Files hdfs://sandbox.hortonworks.com:8020/user/Vimal/Thesis/USD/Forex_Usa_Validate_Dataset.csv
hdfs://sandbox.hortonworks.com:8020/user/Vimal/Thesis/USD/Forex_Usa_test_Dataset.csv
hdfs://sandbox.hortonworks.com:8020/user/Vimal/Thesis/USD/Forex_Usa_train_Dataset.csv

Figure 16: Loading data to H2o

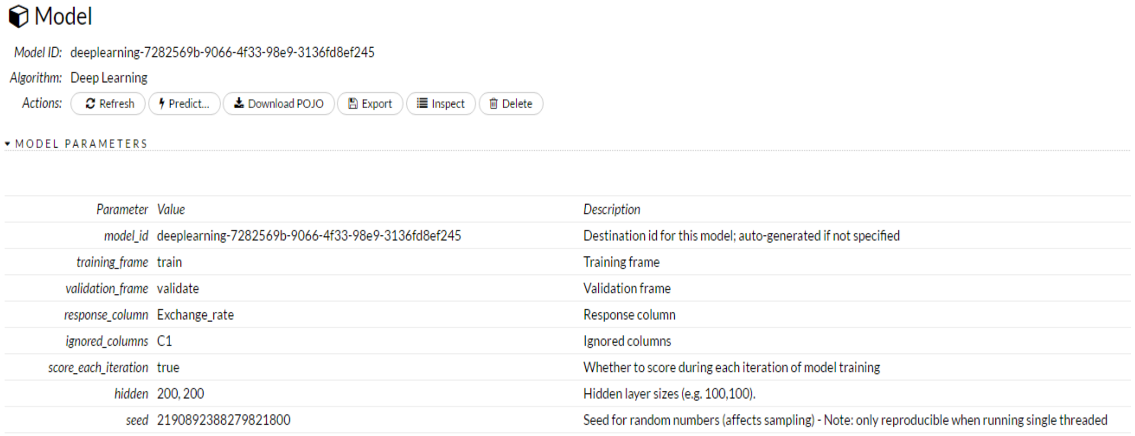
folder is created for each country like USD, CAD, and GBP but in this we can see only USD as I have showed only USD Dataset to minimize the space consumed by screen shots and conveying through this report. All the three currencies dataset are put in their respective folder, with three dataset one for train, validate and the last one for test.

In this screen-shot, the USD's all the three dataset of train, validate and test are loaded in Hadoop distributed file system(HDFS) successfully using Hortonwork are shown.

After the loading of dataset in HDFS, all three dataset is again loaded to H2o and then parsing of the all the three datasets are obtained, so that it is easy for the model to understand the data and allow the model to predict accurately. The loading and paring of the dataset is achieved using Sparkling water which are shown below.

All the three independent models are implemented in the sparkling water and their details are provided below,

(a) Deep Learning Algorithm

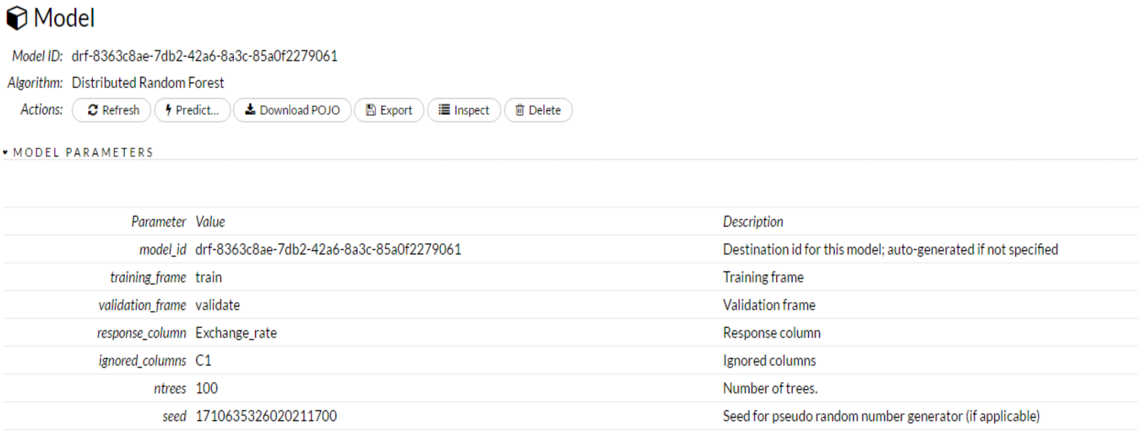


Model ID: deeplearning-7282569b-9066-4f33-98e9-3136fd8ef245

Algorithm: Deep Learning

Parameter	Value	Description
model_id	deeplearning-7282569b-9066-4f33-98e9-3136fd8ef245	Destination id for this model; auto-generated if not specified
training_frame	train	Training frame
validation_frame	validate	Validation frame
response_column	Exchange_rate	Response column
ignored_columns	C1	Ignored columns
score_each_iteration	true	Whether to score during each iteration of model training
hidden	200, 200	Hidden layer sizes (e.g. 100,100).
seed	2190892388279821800	Seed for random numbers (affects sampling) - Note: only reproducible when running single threaded

(b) Distributed Random forest

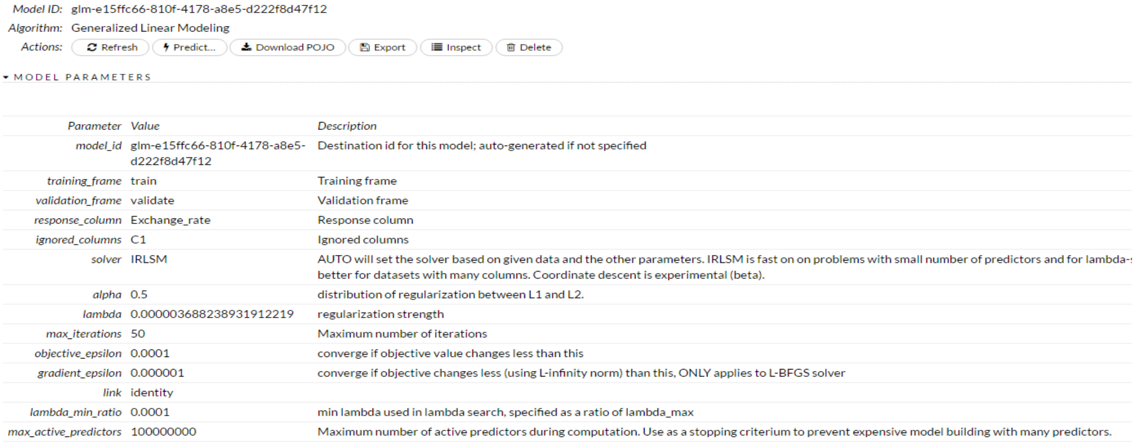


Model ID: drf-8363c8ae-7db2-42a6-8a3c-85a0f2279061

Algorithm: Distributed Random Forest

Parameter	Value	Description
model_id	drf-8363c8ae-7db2-42a6-8a3c-85a0f2279061	Destination id for this model; auto-generated if not specified
training_frame	train	Training frame
validation_frame	validate	Validation frame
response_column	Exchange_rate	Response column
ignored_columns	C1	Ignored columns
ntrees	100	Number of trees.
seed	1710635326020211700	Seed for pseudo random number generator (if applicable)

(c) Generalized Linear Model



Model ID: glm-e15ffc66-810f-4178-a8e5-d222f8d47f12

Algorithm: Generalized Linear Modeling

Parameter	Value	Description
model_id	glm-e15ffc66-810f-4178-a8e5-d222f8d47f12	Destination id for this model; auto-generated if not specified
training_frame	train	Training frame
validation_frame	validate	Validation frame
response_column	Exchange_rate	Response column
ignored_columns	C1	Ignored columns
solver	IRLSM	AUTO will set the solver based on given data and the other parameters. IRLSM is fast on on problems with small number of predictors and for lambda-s better for datasets with many columns. Coordinate descent is experimental (beta).
alpha	0.5	distribution of regularization between L1 and L2.
lambda	0.000003688238931912219	regularization strength
max_iterations	50	Maximum number of iterations
objective_epsilon	0.0001	converge if objective value changes less than this
gradient_epsilon	0.000001	converge if objective changes less (using L-infinity norm) than this, ONLY applies to L-BFGS solver
link	identity	
lambda_min_ratio	0.0001	min lambda used in lambda search, specified as a ratio of lambda_max
max_active_predictors	100000000	Maximum number of active predictors during computation. Use as a stopping criterium to prevent expensive model building with many predictors.

3.2.8 Presentation of Results

Variable Importance

It is very important to find out that which are all the factors that mainly influence the Forex are understood with the help of variable importance parameter. The below provided diagram shows the how much of factors CPI, Interest rate, PPI influences the Forex rate.

Among all the factors, the date factor plays majority of contribution to the Forex, the next factor is Consumer price Index, then Year, Share, Interest rate, PPI and the exploratory factors

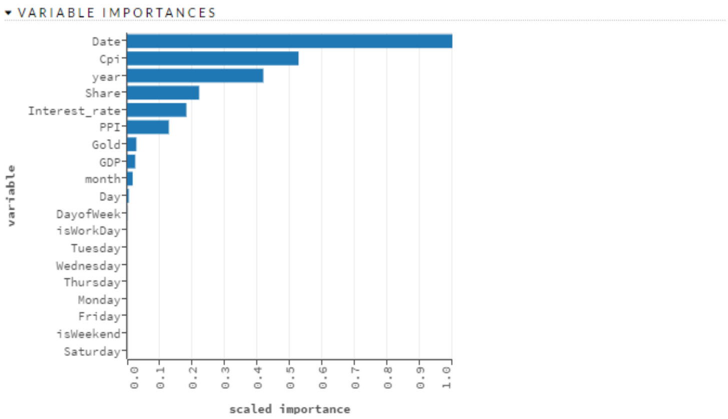
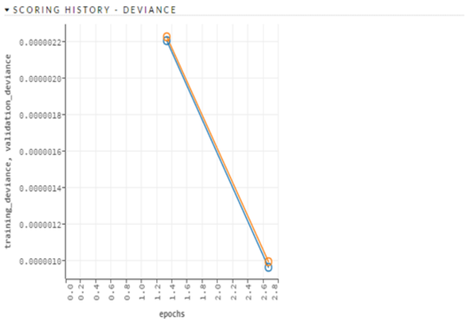


Figure 17: Variable Importance

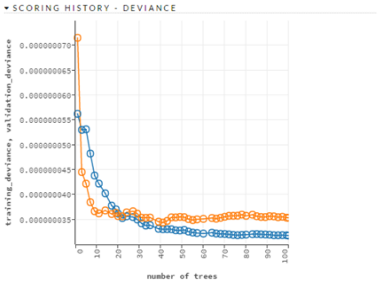
like Day, which day of the week, IsWeekEnd, IsWorking Day doesnt contribute to the Forex rate at all. Hence as a conclusion from the Variable importance it better to remove those fields from the dataset and then analysis is done

Scoring Rate of Deep learning and Distributed Random forest

The Scoring rate, delivers the time taken by the model to predict the efficiently by considering and testing against the validate dataset. The Less time scoring the model doesnt mean that the model is good enough to predict the value accurately.



Scoring Rate of the Deep learning model



Scoring Rate of the distributed random forest

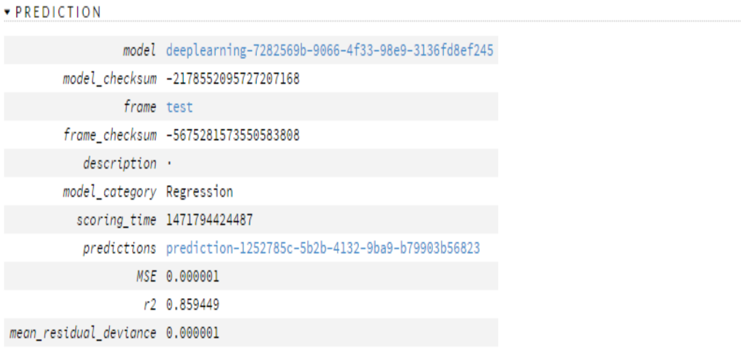
The above figures shows the scoring rate of deep learning and the distributed random forest. It very clearly understand that the Deep learning took a time and it was learning the model fast but the distributed random forest learning rate very gradual to reach its best accuracy for predicting the model.

Prediction accuracy

The model accuracy and the performance are evaluated using two important function called Mean Squared Error (MSE) and R-Square. The predicted accuracy of all the model implemented in our project are shown below,

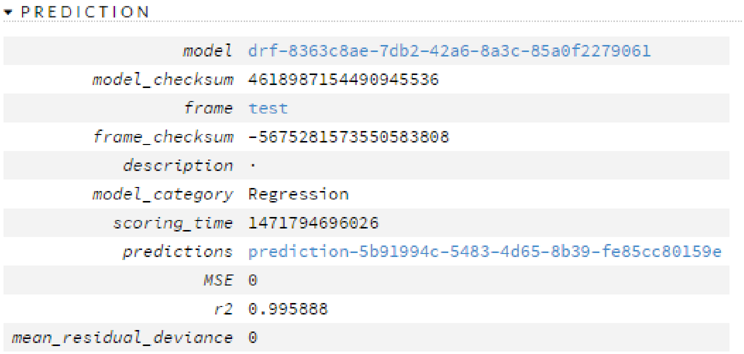
Prediction accuracy of Deep learning

The prediction matrics of Deep learning model for the Forex rate USD/INR date is done, the deep learning model was able to produce the accuracy of 86% with the R - Squared value



```
▾ PREDICTION
                model  deeplearning-7282569b-9066-4f33-98e9-3136fd8ef245
       model_checksum  -2178552095727207168
                frame  test
       frame_checksum  -5675281573550583808
          description  ·
       model_category  Regression
         scoring_time  1471794424487
          predictions  prediction-1252785c-5b2b-4132-9ba9-b79903b56823
                  MSE  0.000001
                   r2  0.859449
mean_residual_deviance  0.000001
```

Figure 18: Prediction accuracy of Deep learning



```
▾ PREDICTION
                model  drf-8363c8ae-7db2-42a6-8a3c-85a0f2279061
       model_checksum  4618987154490945536
                frame  test
       frame_checksum  -5675281573550583808
          description  ·
       model_category  Regression
         scoring_time  1471794696026
          predictions  prediction-5b91994c-5483-4d65-8b39-fe85cc80159e
                  MSE  0
                   r2  0.995888
mean_residual_deviance  0
```

Figure 19: Prediction accuracy of Deep Random Forest

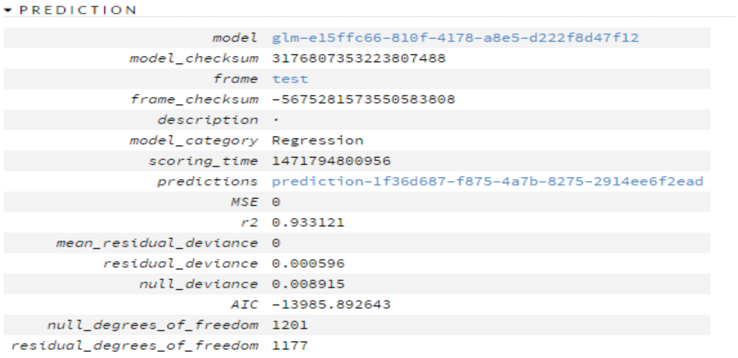
of 0.859449 and the MSE of 0.00001, it allmost the same to the other datasets like CAD/INR and GBP/INR dataset.

Prediction accuracy of Distributed Random forest

The below provided figure shows the prediction matrics of Deep Random forest model for the Forex rate USD/INR date is done and shown below, the deep learning model was able to produce the accuracy of 99% with the R - Squared value of 0.99958 and the MSE of 0, for the datasets CAD/INR has 94% with the R - Squared value of 0.9428 and the MSE of 0 and GBP/INR dataset with 92% accuracy with the R - Squared value of 0.92453 and the MSE of 0.

Prediction accuracy of Generalized Linear model

The above figure shows the prediction matrics of Generalized Linear model (GLM) for the

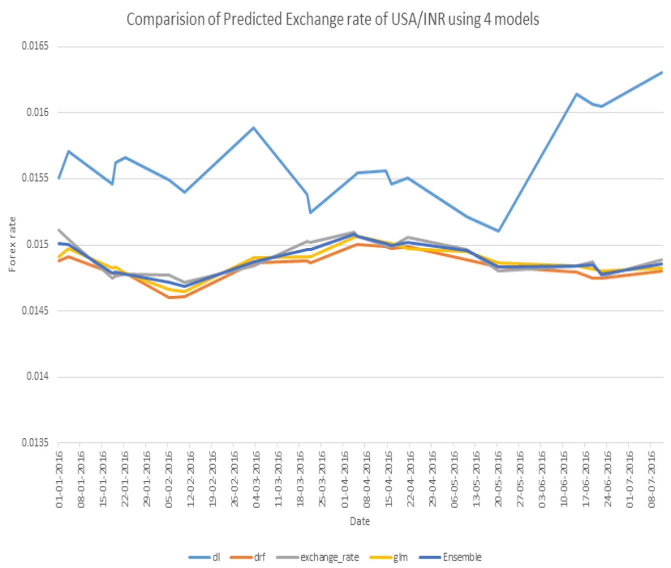


```
▾ PREDICTION
                   model  glm-e15ffc66-810f-4178-a8e5-d222f8d47f12
          model_checksum  3176807353223807488
                   frame  test
          frame_checksum  -5675281573550583808
             description  ·
          model_category  Regression
            scoring_time  1471794800956
             predictions  prediction-1f36d687-f875-4a7b-8275-2914ee6f2ead
                     MSE  0
                      r2  0.933121
  mean_residual_deviance  0
       residual_deviance  0.000596
           null_deviance  0.008915
                     AIC  -13985.892643
   null_degrees_of_freedom  1201
residual_degrees_of_freedom  1177
```

Figure 20: Prediction accuracy of Generalized Linear model

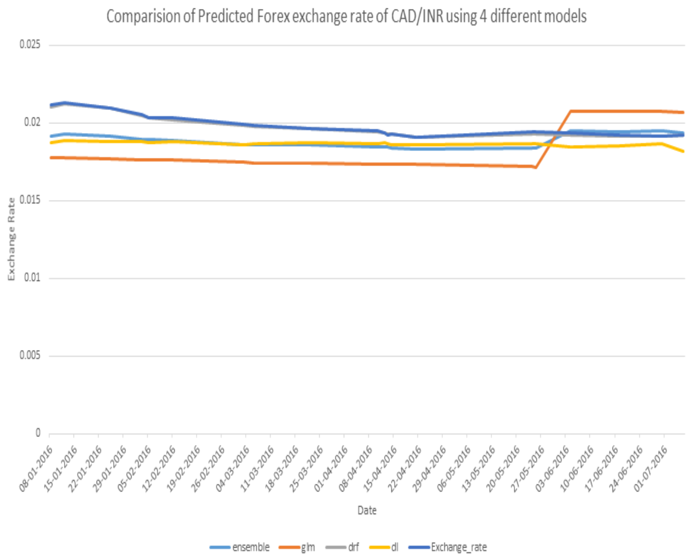
Forex rate USD/INR date is done, ane it found that the deep learning model was able to predict the Forex with the accuracy of 94% with the R - Squared value of 0.933121 and the MSE of 0, it allmost the same to the other datasets like CAD/INR and GBP/INR dataset.

3.2.9 Comparison of The Proposed Solution With Independent Models



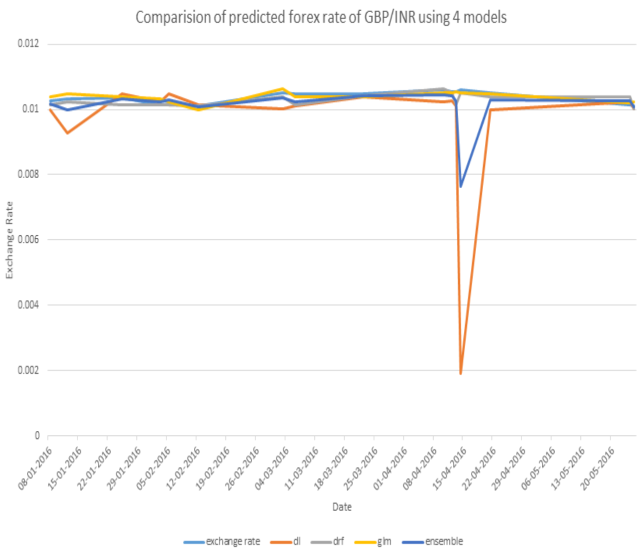
dl - Deep Learning
drf - Distributed Random forest
Glm – Generalised Linear model

The above graph is plotted between Forex rate and Date which represents the prediction exchange rate comparison of USD against INR. Here the exchange rate has been predicted using four other models such as deep learning, distributed random forest, generalised linear model and ensemble model. The results shows that except deep learning, the prediction of exchange rate from other models are very accurate, where the results of deep learning has slightly deviated from the other values. Also the output of ensemble model and predicted exchange rate is more close to one another.



dl - Deep Learning
drf - Distributed Random forest
Glm – Generalised Linear model

The above graph is plotted between Forex rate and Date which represents the prediction exchange rate comparison of CAD against INR. Here the exchange rate has been predicted using four other models such as deep learning, distributed random forest, generalised linear model and ensemble model. The results shows that the prediction of exchange rate from all the models are very close to accurate whereas the prediction of ensemble model is the most accurate when compared to other models.



dl - Deep Learning
drf - Distributed
Random forest
Glm – Generalised
Linear model

The above graph is plotted between Forex rate and Date which represents the prediction exchange rate comparison of GBP against INR. The results shows that except deep learning, the prediction of exchange rate from other models are very accurate. Deviation represents the presence of some noise in the input data even after cleaning it. Whereas the prediction of ensemble model is the most accurate when compared to other models.

3.3 Conclusion

As the implementation of the ensemble data mining model, with the proper requirement gathering, architectural design, data analysis, we were able to predict the Forex accurately, As our main goal of predicting the Forex with the maximum accuracy of 96% which is good accuracy, our ensemble model performed all most best of all the other model.

4 Evaluation of Proposed solution and overall project

4.1 Introduction

The Evaluation of Prediction Model as we can understand able to understand the performance of the model as it is the goal that we have to predict the Forex accurately. In the Evaluation process several techniques and tools are used to evaluate the model

4.2 Testing and Evaluation of Prediction Model

In order to test how well the model fit and able to predict the data, the test followed is called Goodness of Fit, which provided the accuracy of the model. This technique is used to test the accuracy of each and every model used in this project. The models under the sparking water for our prediction of Forex uses two important fitness function namely, R-Square and Mean Squared Error Willmott and Matsuura (2005).

R-Square, R-Square, shows the strength of the model and how well the model can predict the data and also explains the correlation between the actual value and the predicted values of the model. R-Square is also defined as the correlation between independent and dependent column.

$$R - square = 1 - \frac{\sum_{i=1}^{N} w_i (y_i - f_i)^2}{\sum_{i=1}^{N} w_i (y_i - y_{av})^2} = 1 - \frac{SSE}{SST}$$

In the above representation, the values fi are the values got from the model after prediction, Yav can also be represented as means of the observed data yi and wi are the weights added to every point of the data most of the time the value of wi=1.

SSE can also be defined as the sum of squares error,

SST is defined as the total made for the sum of squares.

The value of R-square always between 0 and 1, the value that are nearer to 1 represent that a higher chance of variance handled by the model.

Mean Squared Error:

Mean Squared Error can be explained as the error that the model handles or can be also called as standard error. Mean Squared Error is attained by the S.D of the random component in the data.

Mean Squared Error can be represented as

RMSE = MSE

in this equation, MSE is called as mean square error.

mean square error

MSE=SSE/v

It is similar to SSE, the value of it nearer to 0 indicates that the model is best used for the prediction.

As discussed above, the above mentioned parameters are used for evaluating all our independed and our ensemble model, the Deep Learning Algorithm which is considered to best for time series as usual produced one of the best accuracy, it has R-square value of 0.864 and with the RMSE of 0.00001, the Generalised Linear model produced the R-square value of 0.93333 and with the RMSE of 0.00059,the Distributed random forest produced our best results of all the independent model with the R-square value of 0.9958 and with the RMSE of 0, and our ensemble model produced the R-square value of 0.9645 and with the RMSE of 0.0002

4.3 Evaluation of Overall Project

As a result, all the deep learning model was predicting the Forex accurately, among all the three model the Distributed Random forest was predicting the Forex accurately. The Ensemble model and the random forest all most close to each other in the prediction. And also there is any need for further exploratory analysis as it doesn't contribute to the prediction of the Forex.

5 Conclusion and Future work

Our main goal is predicting the forex accurately, which we have successfully achieved by developing the ensemble model using Deep learning algorithm, Distributed Random forest, Generalised Linear model in sparkling water. We were able to predict the model accurately at a fast pace which is the industry really aiming for. Our research, also provided the contribution of factors influencing the forex and also made us clear that any exploratory analysis is not required for Deep learning. As a future analysis, several other factors like political stability is unidentified and finding how the political stability influences the forex rate. Also other factors like day today news, Fuel price, country's president speech can also be considered of how it influencing the forex rates.

References

Candel, A., Parmar, V., LeDell, E. and Arora, A. (2015). Deep learning with h2o.

Cărbureanu, M. (2011). The analysis of currency exchange rate evolution using a data mining technique., *Petroleum-Gas University of Ploiesti Bulletin, Economic Sciences Series* **63**(3).

Chandar, S. K., Sumathi, M. and Sivanandam, S. (2014). Neural network based forecasting of foreign currency exchange rates, *International Journal on Computer Science and Engineering* **6**(6): 202.

Chen, A.-S. and Leung, M. T. (2004). Regression neural network for error correction in foreign exchange forecasting and trading, *Computers & Operations Research* **31**(7): 1049–1068.

compareremit (2016). money-transfer-guide.
URL: *http://www.compareremit.com/money-transfer-guide/key-factors-affecting-currency-exchange-rates/*

Czekalski, P., Niezabitowski, M. and Styblinski, R. (2015). Ann for forex forecasting and trading, *Control Systems and Computer Science (CSCS), 2015 20th International Conference on*, IEEE, pp. 322–328.

Deng, J., Qu, Z., Zhu, Y., Muntean, G. M. and Wang, X. (2014). Towards efficient and scalable data mining using spark, *Information and Communications Technologies (ICT 2014), 2014 International Conference on*, pp. 1–6.

Eng, M. H., Li, Y., Wang, Q.-G. and Lee, T. H. (2008). Forecast forex with ann using fundamental data, *Information Management, Innovation Management and Industrial Engineering, 2008. ICIII'08. International Conference on*, Vol. 1, IEEE, pp. 279–282.

Freeman, J. A. and Skapura, D. M. (1992). Neural networks: Algorithms, applications and programming techniques, *JOURNAL-OPERATIONAL RESEARCH SOCIETY* **43**: 1106–1106.

Gan, W.-S. and Ng, K.-H. (1995). Multivariate forex forecasting using artificial neural networks, *Neural Networks, 1995. Proceedings., IEEE International Conference on*, Vol. 2, pp. 1018–1022 vol.2.

investopedia (2013). Forex tutorial: Economic theories, models, feeds data available.
URL: *http://www.investopedia.com/university/forexmarket/forex5.asp*

Kamruzzaman, J. and Sarker, R. (2003). Comparing ann based models with arima for prediction of forex rates, *Asor Bulletin* **22**(2): 2–11.

Kayal, A. (2010). A neural networks filtering mechanism for foreign exchange trading signals, *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, Vol. 3, pp. 159–167.

Kimata, J. D., Khan, M. and Paul, M. T. (2015). Forecasting exchange rate of solomon islands dollar against euro using artificial neural network, *2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, IEEE, pp. 1–12.

Leung, M. T., Chen, A.-S. and Daouk, H. (2000). Forecasting exchange rates using general regression neural networks, *Computers & Operations Research* **27**(11): 1093–1110.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Lin, S.-Y., Chen, C.-H. and Lo, C.-C. (2013). Currency exchange rates prediction based on linear regression analysis using cloud computing, *system* **6**(2).

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models, *Encyclopedia of statistical sciences* .

Patel, P. J., Patel, N. J. and Patel, A. R. (2014). Factors affecting currency exchange rate, economical formulas and prediction models, *International Journal of Application or Innovation in Engineering & Management (IJAIEM). Retrieved from: http://www. ijaiem. org/volume3issue3/IJAIEM-2014-03-05-013. pdf* .

Segal, M. R. (2004). Machine learning benchmarks and random forest regression, *Center for Bioinformatics & Molecular Biostatistics* .

Uddin, K. M. K., Quaosar, G. A. A. and Nandi, D. C. (2013). Factors affecting the fluctuation in exchange rate of the bangladesh: A co-integration approach, *The International Journal of Social Science. Retrieved from: http://www. tijoss. com/TIJOSS% 2018th% 20Folder/1Kamal. pdf* .

Urrutia, J. D., Olfindo, M. L. T. and Tampis, R. (2015). Modelling and forecasting the exchange rate of the philippines: A time series analysis, *AMERICAN RESEARCH THOUGHTS* .

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate research* **30**(1): 79–82.

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014). Data mining with big data, *IEEE transactions on knowledge and data engineering* **26**(1): 97–107.

Yao, J. and Tan, C. L. (2000). A case study on using neural networks to perform technical forecasting of forex, *Neurocomputing* **34**(1): 79–98.

Yu, L., Wang, S. and Lai, K. (2005). A novel nonlinear ensemble forecasting model incorporating {GLAR} and {ANN} for foreign exchange rates, *Computers Operations Research* **32**(10): 2523 – 2541. Applications of Neural Networks.
URL: *http://www.sciencedirect.com/science/article/pii/S030505480400156X*

Zhang, G. (2003). Time series forecasting using a hybrid {ARIMA} and neural network model, *Neurocomputing* **50**: 159 – 175.
URL: *http://www.sciencedirect.com/science/article/pii/S0925231201007020*