UNIVERSITY OF LEICESTER

PREDICTION MARKET FOR MACRO ECONOMIC VARIABLE UNEMPLOYMENT RATE (HSBC-CANADA)

A Dissertation Report Submitted for Master's in Data Analysis for Business Intelligence

By

Mr. Venkata Satyanarayana Bondu

Under Supervisor: Prof Evgeny Mirkes

Prof Juxi Li

Department of Mathematics

England, United Kingdom

September 2022

ABSTRACT

Banking sector generally initiate their revenue based on some of the following characteristics and variables. They are, Interest rate, Deposits rate, Unemployment rate, GDP, Yield and so on. They collect the data from the public based on their usage, refine it, and update the database which is not stationary and they use the collected and updated data for future predictions. The following project is stimulated in the same way as the data is gathered and collected from various banking sector websites, cleaned the raw and unstable data, followed by some basic machine learning and statistical models for the future predictions. The prediction results could also generate a good result for the banking as their profits could be increased and their loss could be decreased based on the predicted values of the following banking sectors.

Here we try to estimate and analyse the customer behaviour based on the targeted variable unemployment rate as we tried to compare the unemployment data with the other macro-economic variables as mentioned in the above like Interests rate, Deposits rate, Unemployment rate, Yield, GDP.

There are several steps followed for the task accomplishment. The steps are involved calculatedly. Multi various models are used to analyse and get the data cleaned and gathered according to our database created.

Initially we need to create our own database so that the data collected could be on our own boundary limit we have chosen. The data should be ready for any kind of query execution based on the selected and used variables. The data should not exceed the boundary limit.

Next, we need to analyse the unemployment rate with respect to the other targeted variables used so far as at which point of theme can a standard customer can continue based on the interest rate, deposit rate, and income rate of the bank is followed and maintained using machine learning algorithms.

Here we use the regression models such as, Linear regression model, Auto Regressive Integrated Moving Average (ARIMA), including Seasonal Auto Regressive Integrated Moving Average eXogenous factors (SARIMAX) with respect to ADA Boost Regression model and Random Forest Regression model for better accuracy of the moving rate.

Observed results of the datasets and applied models of calculation for the factors and variables used explains us about the prediction of the particular banking sector we used to predict the future development rate for its profit and loss.

By using the data of more deposits, cash and income rate of the bank, we could estimate the unemployment rate and predict its future analysis of the bank and can assure their profits in the business by successfully implementing the following models. This could be the best and most appropriate model of analyzation of a customer behaviour and their change in unemployment rates with respect to other variables.

ACKNOWLEDGEMENT

Initially I would like to thank my academic supervisor Professor Evgeny Mirkes and my project director Andrew Morozov and best help and support from the supervisor Professor Juxi Li.

I would greatly thank the banking societies which had provided me a vast and valid information in creating my database with very appropriate and exact values of the database I needed.

I sincerely thank my best groupmates Harsha Reddy, Mayuri Rawat, Saiteja Reddy Vangumalla and Sreeja Ravella for being my fireside and helping me in all the aspects about the project when I'm in a very great need of help.

I would like to thank my beloved and respected University, University of Leicester for teaching me and bringing me up in the sector of Data Analysis for Business Intelligence as it helped me a lot in this project and in other many ways.

DECLARATION

The Project “Prediction Market for Macro Economic Variable Unemployment Rate” for master’s in Data Analysis for Business Intelligence in the University of Leicester is attested by myself Mr. Venkata Satyanarayana Bondu (209055863). I sincerely declare that the project and report is completely prepared and maintained by my own work and knowledge that there’s nothing external factor which had affected my report under the way of the plagiarism and similarity content.

Name: Venkata Satyanarayana Bondu.

ID Number: 209055863.

E-mail: vsb8@student.le.ac.uk

Signature: Venkata Satyanarayana Bondu.

TABLE OF CONTENTS: -

Contents

TABLE OF FIGURES

1. INTRODUCTION: -

The theme of the project is to analyse and produce the results of the macro-economic variable Unemployment Rate of the country CANADA. The unemployment rate in the country has been affected by various reasons and has been changing every year. As it is very hard to find and analyse all the data for all the years of unemployment rate, we here by choose specific years as dataset values, calculate them and analyse them in previous, present and future years. This is the way to work with unemployment rate as the targeted variable. There has been a very vast inflation of unemployment when compared to the other targeted variables such as GDP, Interests rate, Deposits and other variables. So, we here by compare our target variable unemployment rate with other variables to predict its future values about their profits and loss.

The main target macro-economic variable of the thesis is Unemployment. As the youth has been increasing day by day in each and every country, I have chosen unemployment and its affects as the macro-economic variable.

Canadian business levels out of the blue succumbed to a third consecutive month in August and the jobless rate bounced, a potential sign financing cost climbs have begun to cool the tight work market.

The economy shed 39,700 positions last month, Measurements Canada covered Friday in Ottawa, an unexpected negative perusing contrasted with the 15,000-increase expected by financial experts in a Bloomberg overview.

The Canadian dollar's benefit debilitated somewhat after the news, exchanging up 0.5% to $1.303 at 8:40 a.m. in Toronto after prior moving as much as 0.8%. Yields on Canadian 10-year securities broadened declines, falling 9.6 premise focuses to 3.1%.

The abatement in work and higher jobless rate might be proof the nation's workforce is re-adjusting as the Bank of Canada's forceful rate climbs begin to cool financial development and slow interest. The unexpected work searchers may similarly ease wage development as the work supply grows.

These are some of the examples of unemployment in Canada which could be a relevant reference of work to understand the further analysis done on the unemployment rate in Canada which had affected its banking sector.

As we also have some natural calamities and pandemic which are one of the main reasons to bring such a huge unemployment in the country Canada. The years 2019-2022 till today, there’s unemployment fluctuation in Canada as many youths were not employed. Also, there’s recission in between the period of 2007-2022, many students who were freshly passed out from the universities of Canada in search of jobs were not offered with a good placement as there’s job crisis and even working employees who were removed from the jobs are not able to find and work on the new ones. There were many protests by other

country immigrants based on the unemployment and study affairs. This caused a great lead in the unemployment.

The average hourly wage rate was up 5.4% from a year ago, compared to 5.2% from both June and July. That's the fastest increase in records up to 1997, outside the pandemic.

The unemployment has been increasing for several years since 2000 and it has continued for so on period till 2007. After a specific recission time, there were many opportunities for the new passed out youth and other non-residents of Canada. This has been a great opportunity for the student's worldwide as there were new opportunities.

The unemployment rate has been decreased to some extent and there's been a good income and GDP across the Canada. It was well and good till 2018 and there was the pandemic which is still showing its fluctuations. Even now, some parts of Canada were with mere unemployment which could be cleared by the following years.

More than 70% of Canada are employed completely as the rest are the new students and other standard people who are in the search of employment. We try to predict the updates and the decrease in the unemployment for the following and coming years through our analysis using various models of machine learning and regression models for forecasting.

The wages have been increased compared to previous years and so the employment has changed and increased with more new opportunities. The same way we have calculated the past data for future predictions.

The formulas used for each model has a specific definition and usage. They are completely independent and we could also boost them for better accuracy purposes.

As there was no proper employability in the country Canada, there would not be proper income, deposit and yield for the cash flow and production. They are inversely proportional to each other. If there's no employment, there would not be a good GDP for the country Canada. This could be a big problem and also a solution for the analysis of unemployment rate.

When we normalize the data, we could clearly observe the changes of GDP, Deposits, Interests, Income and Yield which are inverse to unemployment. If there's a proper employment, there will always be a good Income, there will be a good Deposits, there will be a good Interest rate, there will be a good cash flow and there will be a good GDP across the country.

We try to show and predict that the unemployment of the country could decrease and there will be an increase in the employment of the country.

We compare the values and datasets of the other targeted variables to our target variable unemployment rate just to make sure and correct that there will be increase in the employment rate and decrease in the unemployment rate.

Approximately 60,000 new jobs were offered to the students and most of them are well placed and rest of them were having an opportunity of jobs and employment in other

sectors which could provide them proper wage. And some of the specific jobs also offer the sponsorship to their visa and stay so that there could be more employability and reduce unemployment. In this way unemployment could be predicted with the calculation of unemployment to other variables such as GDP, Interests Rate, Deposits Rate, Yield, Income and other socio sectors.

2. LITERATURE REVIEW: -

Unemployment is one idea that explains how economies, production structures, sectorial developments and regional, national developments. Covid-19 affects the unemployment rate, GDP, Income rate, Interest rate, and Deposits rate. This is a systematic literature review, and I have determined the literature used in this study. It wiped out 1.7 million wages and salary jobs in just 12 months until January 2021. The pandemic caused and created long-lasting effects on employment. This phenomenon is also known as hysteresis employment. The covid-19 health crisis is a great shock that has made a change in the lives and livelihoods of the individuals across the globe. There are some variables that have mostly been affected by the pandemic in banking sector as mentioned above. Unemployment rate is one of the variables which has been affected very hardly by it. And so, we are here to calculate the data collected in a systematic way. It is explained in the following steps:

1) DATA COLLECTION: -
   The data of the variables is collected in several websites and ways so that the data could be little huge and easy to calculate. The collection of data is not so easy that it has to be maintained individual and unique from all other sources so that we could get more appropriate result.

2) DATA MANAGEMENT: -
   The collected data is stored and varied in the excel sheets with separated column names so that the data could not merge with other. The variables are named with independent column names and mentioned with individual values. So that the data could stay individual and reduce or increase if needed.

3) DATA CLEANING: -
   After managing the data in required format the unwanted data from the whole collected data can be cleaned by removing the unwanted rows and columns. So that we could decide the size of data based on our requirement for the calculations. This could show us the changes in the results and outputs of visualizations.

4) DATA SOURCES: -
   The data is mainly collected from the basic sources Fred, Kaggle, World Bank and other. As the data is so vast and huge to calculate, I chose a specific period of time to calculate the data and predict its affects in future analysis.

MACRO ECONOMIC VARIABLES: -

Macro-economics is the study of economic aggregates and their behaviour. Macro-economic variables are linked up to aggregates of economics of a country, of a region, the whole population of the area or region or a country, all the multinational organizations of a country. For instance, the approximate production of the country is formed with the yield of all its profits and loss, livelihoods, and standard individual personals and its tertiary sectors. The general and other commonly applied economic variables are Income, Deposits, Yield, Unemployment rate, GDP, Interest rates and so on. Macro economics generally uses many terms that are quite general and commonly used words but have more proper depth and meaningful subjected purpose within them in macro economics such as investments, and capitals. The macro-economics is also educated through econometric models that depend on the key variables of the macro-economics. Economic outputs are the aggregate outputs of the goods and services by any economy, in other ways GDP and Yield are defined. If the aggregate income is greater, then the economy is running a trading surplus, If the aggregate expenditure is greater, then the economy is running a trading deficit.

Economics also distinguishes between a flow variable and a stock variable. A flow variable is a specified as a quantity per unit of time; a stock variable is a specific quantity at a specified time.

Economics also specifies between both normal and real variables. The variables which we have choose as macro-economic variables are the mixture of both real and normal variables. Each variable has its own character and database. Each variable has its independent as well as dependent affects on other variables too. They are inversely proportional to each other. If one variable is affected, then we have the other variables also affected. This is the main and key study of the used technical four main macro-economic variables.

Here are some of the examples of the macro-economic variables and their datasets which contains the whole information, used in this process: -

1) GDP
2) Interest rate
3) Unemployment rate
4) Yield

And these are defined and described in the very understandable way as following definitions and descriptions below: -

GDP:

The term GDP is formerly defined as Gross Domestic Product. In other words, this is the total and annual gross percentage of the development as per the standard livelihoods of the nation, organization, region and so on. This is the total value of goods and services produced within them. It also measures the total income as the payments for the total production will be deposited and transferred and stored in the banking societies. The world averages about 10,000 USD. The GDP per capita for the United States was 56,115.70 USD.

INTEREST RATE: -

One of the most important variables is interest rate which is also defined as the cost of the credit and also the cost of borrowing or receiving money. Though there are many types of the interest rates, prime rate is the interest rate that sounds business qualifies for. These interest rates are calculated and maintained by the GDP and Income to the banking societies that are stored in. The interests are put on the stored money which are placed in the banking societies. Interest rates affect not only how much consumers will borrow, but will also affect how much businesses will borrow, if they can invest the money for a higher expected return than the interest rate on the burrowed funds.

UNEMPLOYMENT RATE: -

The terms Employment and Unemployment are also the very significant macro-economic variables in the era. These two terms play a crucial role in terms of income and GDP. The terms employment and unemployment are inversely proportional to the other macro-economic variables. As the employment boosts and generates the income to any part of the country, the same way unemployment demotes the per capita income and GDP of the country or any organization. The unemployment rate falls when the economy of the country is increasing. If the economy grows fast, shortages may increase leading to the higher prices of the general goods and services in daily livelihood.

YIELD: -

The term Yield is also defined as the total annual production of the country or organization of any region. The total production should pay and proceed with the rules and regulations as it needs to pay tax to the reupdated officials as this could be taken and collected as the revenue of the profits from the particular organization, region or a country also. The currency exchanges are made as quickly as possible, before the currency falls even further in value, which can happen in hours. Since the maximum and real value can be received by immediately exchanging the hyperinflated currency, whether it be other currencies or for goods and services.

Businesses will keep borrowing if they have projects where they can earn a greater expected return than the interest rate on the loan, but as some point, additional projects will yield a diminishing marginal return, where the expected value of additional invested capital will fall to equal the interest rate reaching what macroeconomics call a capital stock equilibrium.

METHODOLOGY USED: -

The main methodologies used for this thesis are regression models using machine learning and fore casting analysis using time series. The complete methodology machine learning models and time series analysis are worked with the help of jupyter notebook workbench. Using python programming helped a lot for the better understanding of the visualizations of the graphs and their values using dataset compressions was a big task. In machine learning models I've specifically chose and opted some regression models to find the best and more appropriate accuracy rates and percentages of the dataset values.

The same way I have chosen the best and very apt models of time series analysis and forecasting methods to predict future for the non-stationary data. The most challenging part of the analysis is forecasting analysis and their results while predicting their future. Also, there are many more models to choose but these could be very appropriate to find the values of the data collected. The following are the regression models and forecasting models used for the thesis in addition to teste. They are: -

1) Linear Regression Model.
2) ADA Boost Regression Model.
3) Random Forest Regression Method.
4) Auto Regressive Moving Average Model.
5) Seasonal Auto Regressive Moving Average with eXogenous factors.

The above-mentioned methods are the fore most main and important methods used for this thesis analysis with future forecasting. Also, in addition to these, there are some tests which are performed to check whether the results are so accurate and are acceptable to proceed for further calculation part. They are defined as follows:

MACHINE LEARNING MODELS: -

LINEAR REGRESSION MODEL: -

The linear regression model is one of the best models used for the regression calculations. This is the standard model which is often used to find the best and appropriate results of the accuracy percentages. Linear regression method has separate formulas and its own calculations. Since it is very easy to understand and calculate, most of the people are more likely habituated to linear regression model. Let's have a quick look with its formulas and its working.

Formula: - The formula for linear regression is defined simply i.e., Y= mX + b, where Y is the response variable which is also called as dependent variable, X is the predictor variable which Is also called as independent variable, m is the estimated slope and b is the estimated intercept point.

Calculating Linear Regression: -

By using the formula Y = mx + b

- ➔ The linear regression interpretation of the slope co-efficient, m, is the estimated change in Y for a 1-unit increase of X
- ➔ The interpretation of the intercept parameter, b, is the estimated value of Y when X is equal to 0.

The initial part of linear regression contains the very appropriate fit to the values of the slope and Y exponent terms. They are estimated parameters that build the regression line of the best fit. Use the goodness of fit section to learn how appropriate and tight the relationship is at. R-square(accuracy) quantifies the percentage of variance in Y which can be determined by X.

Graphing the Linear Regression line: -

The linear Regression calculation provides a generic and natural graph of the data entered and the regressive line. Graphing is not only just for the purpose of visualization terminology, but also to recheck the outliers of the data entered for the calculation using regression model. If any two points are distant to all other outliers, there are little meanings to be understood. They could be untruly influencing points on the regression line equated or the outliers could also be the most important finding of themselves. Here's an example graph of the Linear regression line and its slope representation.

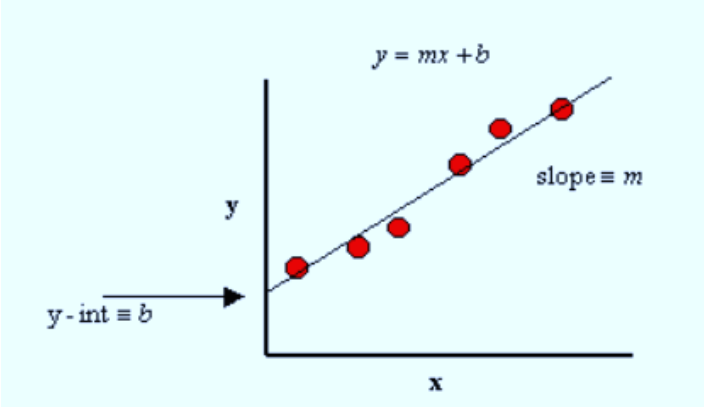


Figure 1: Linear regression example.

The above graph is the best example for the graph of a regression observation and its slope line. As we can observe the both x and y axis, the slope line m and the outliers pointed on the regression line. We can observe that there's also an intercepting point b on the y axis. Also, we could estimate and calculate the distance between two outliers on the slope line. As we know that linear regression is about the connection and the relationship between the dependent and independent variables, we have two types of regression lines and regression resulted graphs. They are,

- ➔ Positive Linear Relationship
- ➔ Negative Linear Relationship.

i) POSITIVE LINEAR RELATIONSHIP: -

The name itself defines that if the relationship between two variables dependent and independent are positive, then we got to say that the regression slope between them is positive and it is called as positive linear relationship.

Ex: -

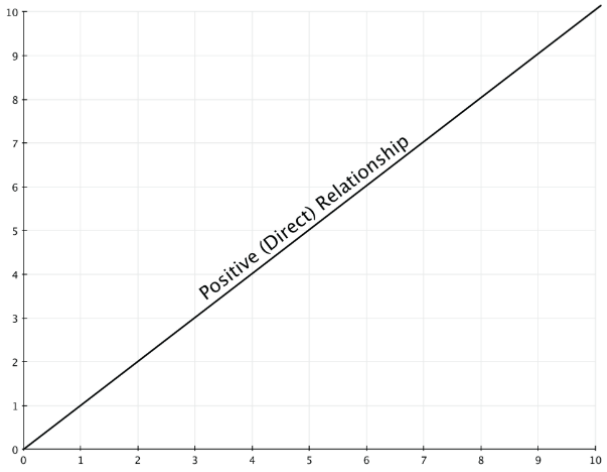


Figure 2: - Linear positive.

ii) NEGATIVE LINEAR RELATIONSHIP: -

The term negative states that if the relationship between two variables both dependent and independent are negative, then we call it as the negative linear relationship of the linear regression. There's a possibility of occurring negative results if the data is not stationary.

Ex: -

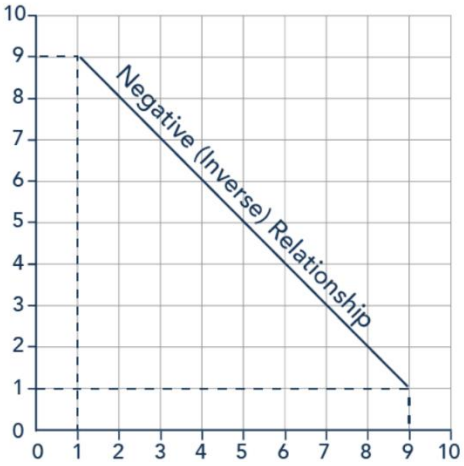


Figure 3:- Linear negative

ADA BOOST REGRESSION MODEL: -

As the same way as linear regression model, ada boost regression model is also a method which is mostly used in the process of boosting the accuracy of the observed regression model. This method is mainly used to observe if there are any changes in terms of positive sign after boosting the actual accuracy results. This method has a specific derived formula which is little difficult to understand but very easy to work if we are able to understand the formula procedure of the ada boost regression model. Ada boost regression model is basically followed by the decision tree of the values entered and boosted for better accuracy results or otherwise R-Square results.

This is also the second model for the previous model what ever model that have been used before using ada boost for the results of accuracy.

This is also a kind of regressor which is meta-estimator that starts the representation by fitting the regressor required on the original dataset classified and then fits the rest of the duplicates of the regressor on the same datasets according to the error of the current position.

RANDOM FOREST REGRESSION MODEL: -

The random forest regression model is also similar to both linear regression and ada boost regression model. This method is also used to improve and boost the accuracy results. The main theme of this model is to boost the accuracy result previously observed and check if the result obtained newly is so appropriate to work on the further time series forecasting or not. Random forests are essentially collection of decision trees. Random decision forests are the better examples of learning models, that uses multiple algorithms to observe better predictive result and performance out of the classification and solving all the types of regression questionnaires and problems.

These are a class and a kind of machine learning algorithms which are used to combine and concatenate multiple random decision tree results trained on a subset of the datasets collected. The usage of more decision trees algorithms shows a decrease and reduction in the variance. This regression algorithm is most commonly used to its ability to work for huge and large sized datasets of most kinds.

Working on random forest on the python workbench is so easy to create decision trees and obtain results using algorithms. We need to import the datasets required using csv or any other format by installing the essential libraries needed. And then we need to separate the features and target variables from the other. As the data is collected from multiple sources, the data gets mixed-up with both wanted and unwanted forms of data so, we need to separate the features and target variable a side from the other variables and their data. After loading the datasets, both the independent and dependent variables chose need to be separated in order to obtain both the results as well as concatenated results. Further we need to split the data into different sectors into a train set and a test set. As we perform several tests to verify if the results and R-Squared values are accurate or not, we need to split the data into two sets. Which, the train sets are used to train the data in a required format, and the test sets are the observations of the train sets used to test the datasets that are used for accuracy.

CORRELATION MODELS: -

The word correlation means defining a relationship between two variables o any two objects. The process of establishing a relationship or connection between two variables is called as correlation. The increasingly similar basis underlying national bank incomes and banking aggregates allows correlation to take place more easily. Correlation is a measure of the linear association between two variables. It has a value between -1 and 1 where:

=> -1 indicates a perfectly negative linear correlation between two variables.

=> 0 indicates no linear correlation between two variables.

=>1 indicates a perfectly positive linear correlation between two variables.

But in some cases, we want to understand the correlation between more than just one pair of variables.

In these cases, we can create a **correlation matrix**, which is a square table that shows the correlation coefficients between several variables.

As our data is seasonal, there are two more steps to forecast the data collected. However, we use correlation to forecast the data, these two methods could help much further.

The word correlation is a very commonly used word in our day-to-day life to represent a kind of association. However, in statistical terminology we use correlation to define the relationship and bond between two quantitative variables. We also calculate that the association is linear, as one variable increases or decreases is a fixed amount for the value increase or reduction in the order. The other method that is many times utilized in these conditions is relapse, which includes assessing the best straight line to sum up the affiliation.

The level of affiliation is estimated by a connection coefficient, signified by r. It is some of the time called Pearson's connection coefficient after its originator and is a proportion of direct affiliation. On the off chance that a bended line is expected to communicate the relationship, other and more confounded proportions of the connection should be utilized.

The connection coefficient is estimated on a scale that shifts from + 1 through 0 to - 1. Complete connection between two factors is communicated by either + 1 or - 1. At the point when one variable increment as different build the connection is positive; when one declines as different expands it is negative. Complete shortfall of relationship is addressed by 0.

Exactly when an inspector has assembled two series of discernments and wishes to see whether there is an association between them, the individual should at first form a disperse chart. The vertical scale tends to one pack of assessments and the even scale the other. If one bundle of insights includes exploratory results and the other involves a period scale or saw portrayal or something like that, putting the preliminary outcomes on the vertical hub is typical. This location what is known as the "dependent variable". The "independent variable", like time or level or one more seen gathering, is assessed along the even centre, or standard.

The estimation of the relationship coefficient is as per the following, with x addressing the upsides of the free factor (for this situation level) and y addressing the upsides of the reliant variable (for this situation physical dead space). The equation to be utilized is:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n-1)SD\,(x)SD(y)}$$

. Find the mean and standard deviation of x.

. Find the mean and standard deviation of y.

. Subtract 1 from n and multiply by SD(x) and SD(y), (n-1) SD(x)SD(y)

However, in deciphering relationship it is essential to recollect that connection isn't causation. There could possibly be a causative association between the two connected factors. Also, in the event that there is an association it could be roundabout.

Rho estimates the part due to the reliance of one variable on the other. For these data, Rho= 0.716, implying that the level of the child accounts for 72% of the variation in the amount of the physical dead space between children.

SIGNIFICANCE TEST: -

To test whether the affiliation is simply clear, and could have emerged by chance utilize the t test in the accompanying computation:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The suppositions guiding this test are as follows:

- ➔ That the two components are perhaps commonly diffused.
- ➔ That there is a direct link between them.
- ➔ The incorrect hypothesis is that they have no relationship.

It is worth noting that the trial of significance for the incline yields the same value of P as the trial of significance for the connection coefficient. Despite the fact that the two tests are inferred in an unusual method, they are arithmetic equivalent, which appears to be valid.

Here's a clear view and example of graphical representations of correlation points and correlation coefficients: -

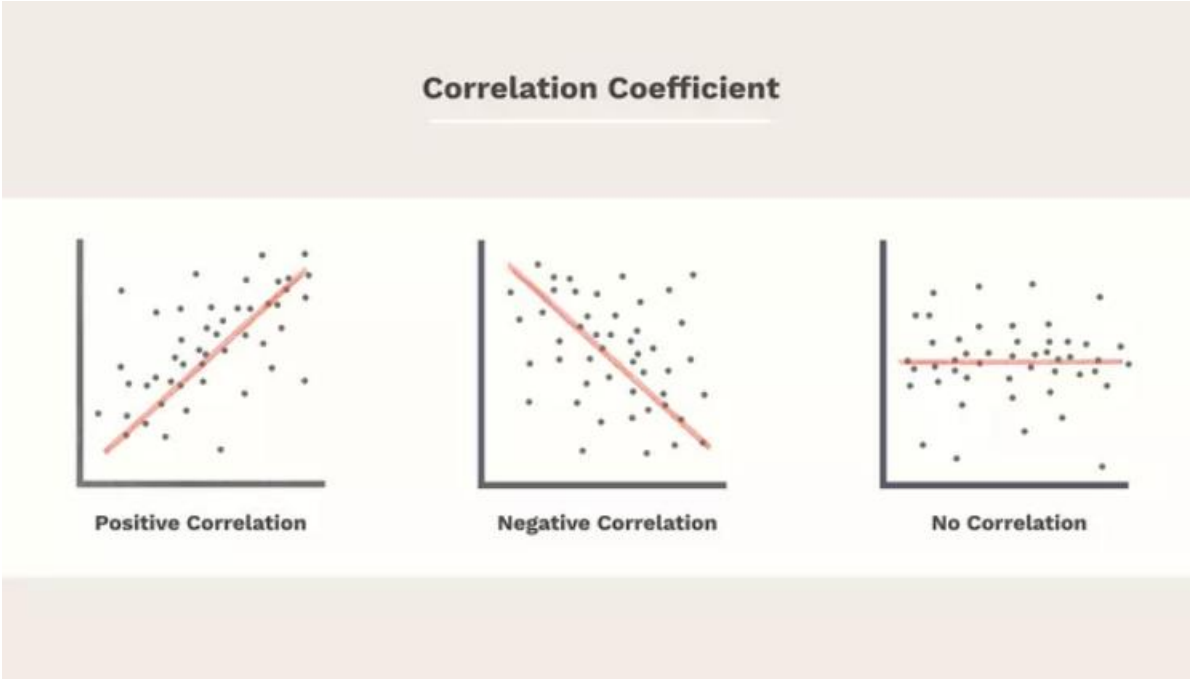


Figure 4: correlation examole.

PRINCIPAL COMPONENT ANALYSIS: -

Principal Component Analysis or PCA is one of a reduction method which is used for dimensional reductions. This method is often used for the reduction of dimensionally larger datasets, by changing a large set of variables into a smaller dataset but contains most of the information of the original datasets.

Normally, reducing the number of variables in an informative index means sacrificing precision, but the trick with dimensionality reduction is to sacrifice a little precision for simplicity. Because smaller informative groupings are easier to analyse and visualise, they make dissecting material much simpler and faster for AI computations without accidental factors to process.

Here's the step-by-step explanation of PCA analysis: -

- ➔ Standardization.
- ➔ Covariance Matrix computation.
- ➔ Compute the Eigen Vectors and Eigen Values of the covariance matrix to identify the principal components.
- ➔ Featuring a Vector.
- ➔ Recast the data along the principal component axes.

More simply, the reason why it is essential to apply normalisation before PCA is that the latter is extremely sensitive to variations in the underlying components. That is, if there are large differences between the scopes of starting factors, those with larger ranges will rule over those with smaller ranges (for example, a variable that reaches somewhere between 0 and 100 will rule over a variable that reaches somewhere between 0 and 1), resulting in one-sided results. Changing the information to equal scales can so avoid this problem.

The goal of this stage is to understand how the components of the informational collection differ from the mean in relation to one another, or, in the end, to determine whether there is any connection between them. Because some components are so closely related that they contain repeating data. To distinguish these correlations, we process the covariance framework in this manner.

Because a variable's covariance with itself is its fluctuation (Cov(a,a)=Var(a), we have the changes of each underlying variable in the principal slanting (upper left to base right). Furthermore, because the covariance is commutative (Cov(a,b)=Cov(b,a)), the passes of the covariance lattice are symmetric with respect to the basic inclining, implying that the upper and lower three-sided segments are equal.

Eigenvectors and eigenvalues are straight polynomial arithmetic concepts that we wish to record from the covariance lattice in order to determine the key elements of the information. Before we go into the specifics of these concepts, we should first define what we mean by head components.

In PCA, we will alco calculate the auto correlation and differencing methods just to check if the p value is <0.05 or not.

Here's an example of graphical representations of principal analysis components:

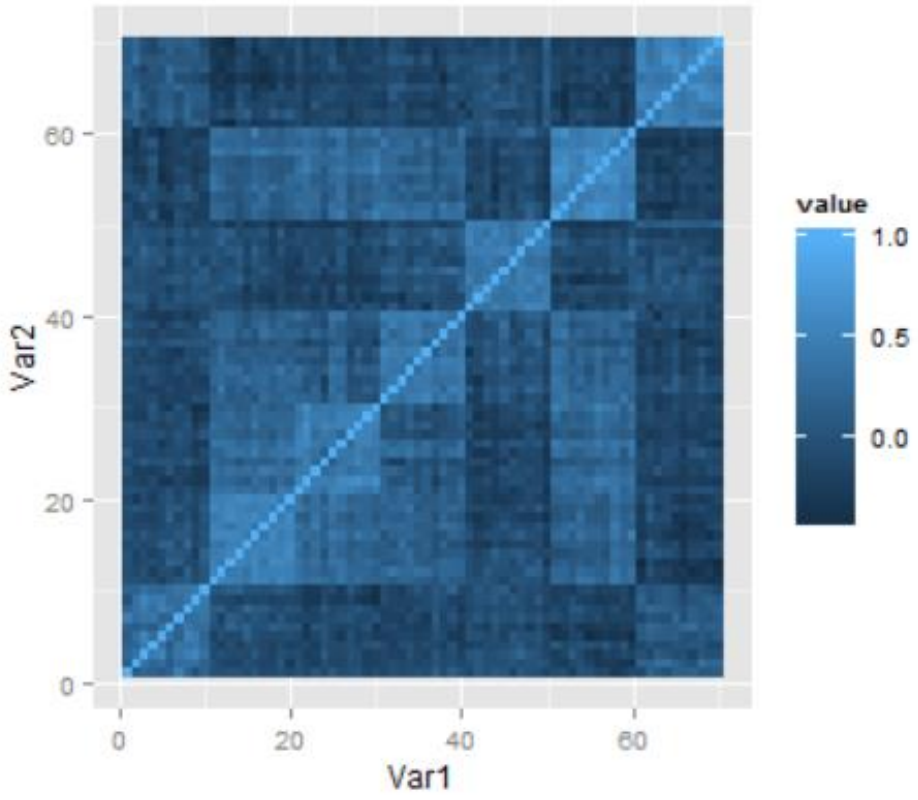


Figure 5: PCA example heatmap

OLS REGRESSION METHOD: -

OLS Regression method is also defined as Ordinary Least Squares regression method. This method is all about calculating the sum of least squares of all the variables. As we get different accuracy rates, and accuracy percentages we get them too many based on the datasets how large they are. This regression method helps us to calculate and understand the least squares of the datasets calculated and proceed to forecasting analysis. In statistics, ordinary least squares are a type of least squares method for estimating the unknown parameters in a linear regression model. The sum of the squares of the differences between the observed dependent variable in the given dataset and those predicted by the linear function of the independent variables.

Ordinary least squares are best in the very well-known regression techniques. This could also be a starting point of all the regression analysis. This might be the most commonly used statistic in the present social sciences. This is used to estimate and calculate the relationships between two or more variables and aggregates.

The results obtained from OLS regression methods includes an output feature class symbolized using OLS residuals, statistical results and diagnostics in the comments window as well as several choosable.

To use the OLS tool, you must supply an Input Feature Class with a Unique ID Field, the Dependent Variable to model, explain, or predict, and a list of Explanatory Variables. You must also specify a path for the Output Feature Class and, if desired, paths for the Output Report File, Coefficient Output Table, and Diagnostic Output Table.

Model performance is measured by the Multiple R-Squared and Adjusted R-Squared values. The possible values are 0.0 to 1.0. Because it incorporates model complexity (the number of variables) as it pertains to the data, the Adjusted R-Squared value is always somewhat lower than the Multiple R-Squared value and hence provides a more accurate estimate of model performance.

Variance Inflation Factor, Probability or Robust Probability, and Coefficient (VIF). Each explanatory variable's coefficient represents both the strength and type of association the explanatory variable has with the dependent variable. When the coefficient's sign is negative, the relationship is negative (for example, the larger the distance from the urban core, the smaller the number of residential burglaries). When the symbol is positive, so is the relationship (for example, the larger the population, the larger the number of residential burglaries). Coefficients are expressed in the same units as their corresponding explanatory variables (a coefficient of 0.005 associated with a variable representing population counts may be interpreted as 0.005 people).

The Joint F-Statistic and Joint Wald Statistic are both statistical significance indicators for the overall model. Only when the Koenker (BP) statistic is not statistically significant is the Joint F-Statistic reliable.

TIME SERIES FORECASTING MODELS: -

Time series analysis is basically used for forecasting the future results of the current data we have on the present day. Forecasting future analysis will be very useful to all organizations, regions and countries too. By calculating future analysis, we could stop from being many things happen. We could use this forecasting analysis in a very useful way. There are many models to perform forecasting analysis. But we choose specific models of time series analysis model based on our datasets. Time series are mainly performed after the performance of regression models and obtainance of the regression results and also the accuracy results. Baser on the values of regression or R-Square values, we will be calculating the future forecasting analysis with the help of the accurate percentages and aggregates. The point values help us to find the future results of the current datasets.

Time series data frequently evolve while observing the industrial processes or tracking business metrics. When watching modern cycles or following corporate business measurements, time series information typically emerges. The primary distinction between showing information using time series tactics and using interaction observation techniques I have chose two specific models and methodology for forecasting analysis. They are named as follows:

- ⇨ Auto Regressive Integrated Moving Average (ARIMA).
- ⇨ Seasonal Auto Regressive Integrated Moving Average with eXogenous factors (SARIMAX).

The above mentioned two models are the best time series forecasting analysis models for this thesis which I have opted for. We have the data for the past few years and decade and we are going to calculate future analysis for the data of banking sector. ARIMA model is one of the best fitting models to the any kind of datasets. As the data I've chose is not stationary, I have chose this model for better results and Sarimax to check if there are any better results rather than Arima model.

ARIMA: -

Arima is formerly described as the auto regressive integrated moving average. It calculated the integrated regressions of the moving averages which are not stationary of the datasets. This model is one of the best fits for the analysis of time series forecasting model. S moving average model uses a regression-like model on past forecast errors. Here, ε is time t, c is a constant and θ are parameters. It only requires the prior data of a time to generalize the forecast. Performs well on short term forecasts. This is a model for non-stationary time series.

Auto Regressive - AR(p) is a relapse model with slack upsides of y as indicators till the p-th time prior. In this equation, p = the number of slacked perceptions in the model, is repeating sound time t, c is a constant, and s are boundaries.

$$\hat{y}_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

Coordinated I(d) - The important item is repeated d times until the initial series becomes fixed. A fixed time series is one whose attributes are independent on the time at which the series is observed.

$$y'_t = (1 - B)^d y_t$$

Moving typical MA(q) - A moving typical model seeks a relapse-like model on prior estimation errors. Here, t is the time of background noise, c is a steady, and s are the bounds.

$$\hat{y}_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

As a result, ARIMA models can be efficiently and precisely used for momentary determination with only time series data, but it may require some experience and trial and error to identify the best set of bounds for each application scenario.

- ⇨ P = The number of autoregressive terms.
- ⇨ D = The number of non-seasonal differences needed for stationarity.
- ⇨ Q = The number of lagged forecast errors in the prediction equation.

SARIMAX: -

In this section, we introduced ARIMA models and their variants: Occasional ARIMA (SARIMA) and ARIMAX, which uses outside information (exogenous contributions) to improve the ARIMA model's presentation.

We used the Crate Jenkins technique to discover the best model for a portion of our dataset (time series of deals from Walmart's store 2). As a first step, we identified crucial properties within recent memory sequences such as stationarity and irregularity. This model is as same as arima which is used for forecasting analysis. But there are exogenous factors which are included in this model for future forecasting analysis. This model is a kind of boost up for the better results of the ARIMA model results. There are some basic steps to follow in this process. They are as following:

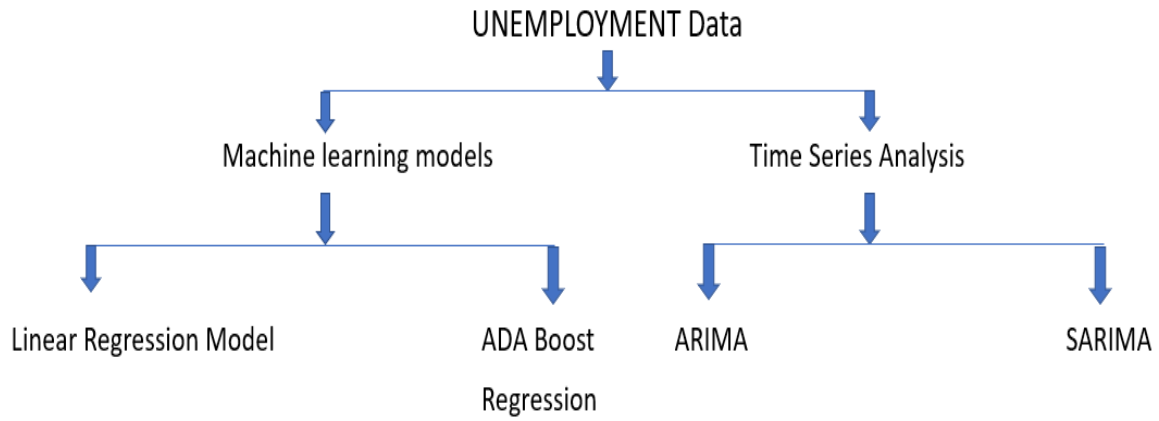
- ➔ Augmented Dickey-Fuller test.
- ➔ ACF and PACF analysis.
- ➔ Exploring model summary statistics.
- ➔ Analyse plots obtained using the stats model method.

We knew from the start that a SARIMA model would be the best fit because we were dealing with sporadic time series. In any event, working with the two models let us to observe the nuances in using the Container Jenkins for these two types of models. Furthermore, we can see clearly that if irregularity is not considered, we are not utilising all data and so are not producing the best forecasts.

ARIMAX and SARIMAX models primarily examine exogenous factors - that is, factors estimated at time t that influence the value of recent memory series at time t but are not auto regressed on. To accomplish this, we just insert the phrases on the right-hand side of our ARIMA and SARIMA conditions. SARIMA models take irregularity into account by essentially applying an ARIMA model to slacks that are number products of irregularity. When the irregularity is exhibited, an ARIMA model is used to the data to detect non-occasional design. In this way, we can predict and estimate the forecasting values and perform the model time series analysis for future forecasting of the current data.

IMPLEMENTATION OF MODELS AND ANALYSIS: -

The thesis is followed and designed in a systematic way for better understanding. As we have got many procedures in the thesis, this step could give a brief explanation about the models implemented. Primarily the data is collected, cleaned as needed for the procedures and then proceeded further. The data is collected from various websites and maintained under the workbench excel. The data is individually collected according to the variables required and opted for. The data is maintained and worked under group of machine learning models and time series analysis models. As mentioned in the above explanations, each macro-economic variable and our main target variable went under each machine learning model and time series analysis forecasting model. They are descripted as follows:



As shown in the above flowchart, our target variable is Unemployment, and its process is followed by machine learning models and forecasting analysis with the help of additional tests. I primarily gathered data from various websites which is raw data and stored in excel and then it is uploaded on to jupyter notebook which is python workbench. As I've performed machine learning algorithms and programming, I chose python programming which is so reliable to understand and work on. By understanding the required libraries and packages, I've performed various programming algorithms for the machine learning models. The steps to analyse the thesis are

- DATA COLLECTION AND CLEANING.
- DATA ANALYSATION.
- COMPARISON OF MACRO ECONOMIC VARIABLES.
- NORMALIZING RAW DATA.
- CORRELATION MATRIX AND TYPES.
- VISUALIZED PCA AND ITS EXPLAINED VARIANCE.
- LINEAR REGRESSION MODEL & EXAMPLE.
- FORECASTING ANALYSIS AND THEIR RESULTS.

- ARIMA RESULTS.
- SARIMAX RESULTS.
- CONCLUSION AND SUMMARY.

The complete thesis is followed in the above-mentioned way. I found it as a systematic procedure to justify the data.

DATA COLLECTION AND CLEANING: -

The data is collected and cleaned by using tools of machine learning. The data is collected from various websites such as fred, Kaggle, bank of Canada and other. The collected data is stored in Microsoft excel by separating them with their own variable names and databases.

By uploading the data into jupyter notebook which is the workbench of python programming, we could call the data and utilize them for the functions and calculations needed.

By installing required library functions, we could perform the programming activities. We need to import some libraries from pre-defined and installed library packages. As we upload the data and call it in the jupyter notebook, we could observe the output of data called with the total number of rows and columns too. It is shown as,

	DATE	GDP	interest_rate	yield	unmp_rate
0	01/01/2007	100.991274	4.5	4.121818	6.3
1	01/02/2007	101.103685	4.5	4.117500	6.2
2	01/03/2007	101.230039	4.5	4.045909	6.2
3	01/04/2007	101.355662	4.5	4.171500	6.2
4	01/05/2007	101.464284	4.5	4.292727	6.0

Figure 6: Unemployment data

As shown in the above image, the data will be displayed with the variables selected for datasets as column names, and their values are followed by the following rows including date.

We here by start the process of analysation of the collected and cleaned data for the further calculations and working models.

DATA ANALYZATION: -

The data called is now analysed as our target variable is unemployment rate, we calculate the data from selected period of time. As we have collected data from the years 2007 to the current year, we now analyse and visualize the graphical representation of unemployment rate in the country Canada.

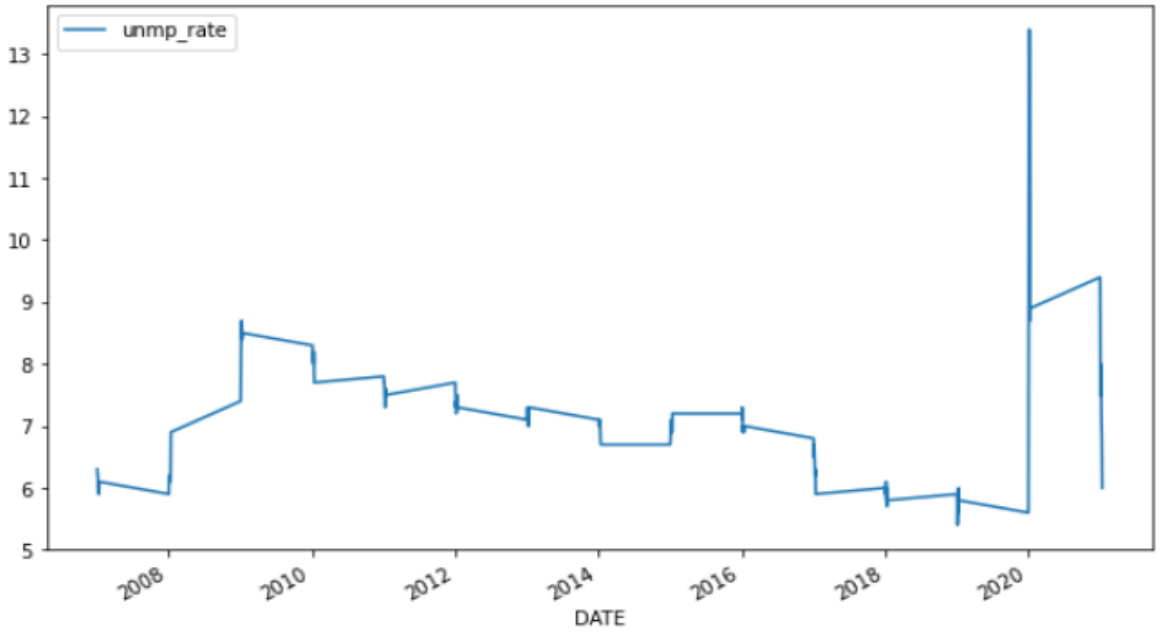


Figure 7: Unemployment analysis

The above picture shows us a clear information about the unemployment rate in the country Canada and its fluctuations. There are many external factors that has caused these ups and downs to the unemployment rate in the country Canada. There was recission in Canada at a particular point of time which led to unemployment in the years 2010 and 2011. And later on the unemployment rate has been decreased gradually as there was a good employment rate in the country. Later on, due to pandemic, in the year 2020, the unemployment has been increased very rapidly beyond expectations.

As unemployment is linked to the other macro-economic variables, even they were also affected by this very vast change. We could more clearly understand and observe the graphical representations when we compare our target variable to the other variables and their representations.

COMPARISON OF MACRO ECONOMIC VARIABLES: -

We here compare our macro-economic variable with other variables for the better graphical representations and understanding. Here I tried to compare our main target variable with GDP and interest rate. As each economic variable is inversely proportional to each other, this could be a best appropriate way to estimate and analyse the data collected. It is shown as follows: -

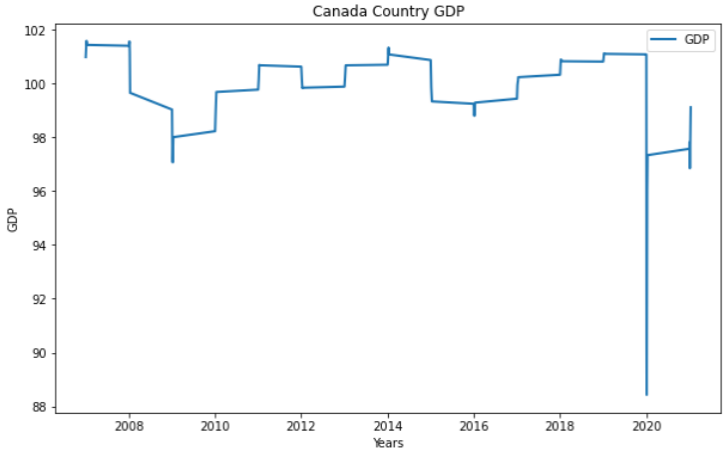


Figure 8: Canada GDP

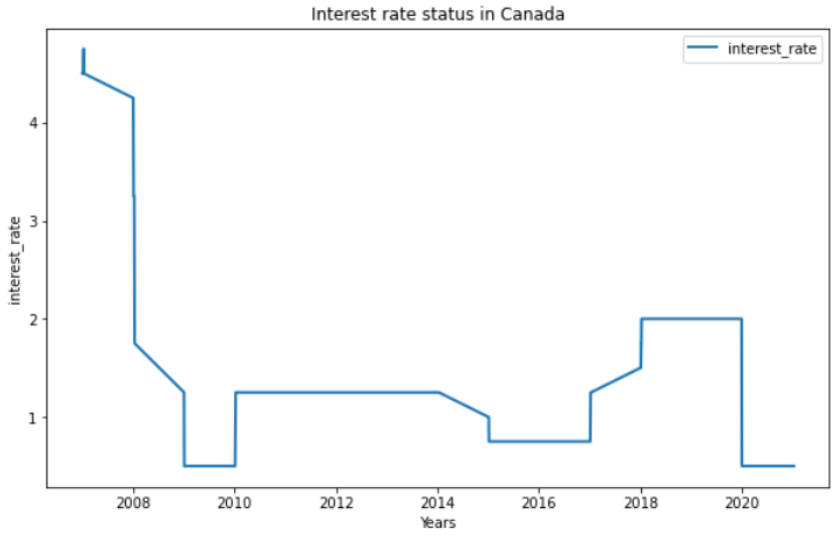


Figure 9: Canada interest rate

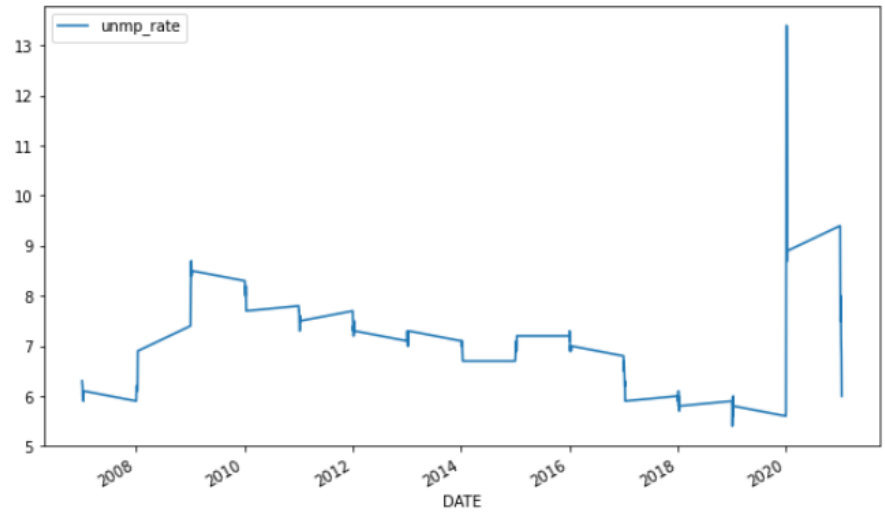


Figure 10: Canada unemployment

- The above displayed three graphs are the observations and graphical representations of GDP, Interest rate and Unemployment rate of the country Canada. As we can see that the target variable is inversely proportional to the other variables GDP and Interest rate. If the unemployment increases, then the rest two variables decrease. If the two variables increase, then the target variable decreases. As shown in the above pictures, each macro-economic variable has been visualized.
- Each macro-economic variable has affected the applications of Canada banking on its own.
- Every macro-economic variable graph has its own characteristics. As they affect the bank in their own way.
- Each year in the unemployment, interest rate and GDP has a very different record compared to each other.
- The same way we can calculate and analyse each and every economic variable and predict their results for future by using auto correlation and partial auto correlation methods that helps for forecasting.
- These are the graphical representations of the collected raw data. We further normalize them using normalization method.

CORRELATION MATRIX: -

- Correlation is a measure of the linear association between two variables. It has a value between -1 and 1 where:

  => -1 indicates a perfectly negative linear correlation between two variables.

  => 0 indicates no linear correlation between two variables.

  =>1 indicates a perfectly positive linear correlation between two variables.

- But in some cases, we want to understand the correlation between more than just one pair of variables.
- In these cases, we can create a **correlation matrix**, which is a square table that shows the correlation coefficients between several variables.

The main theme of correlation is to analyse and represent the relationship between two variables dependent and independent. The graphical representation of correlation could be in anyway. For better understanding, I've chosed heatmap representation. It is as follows: -

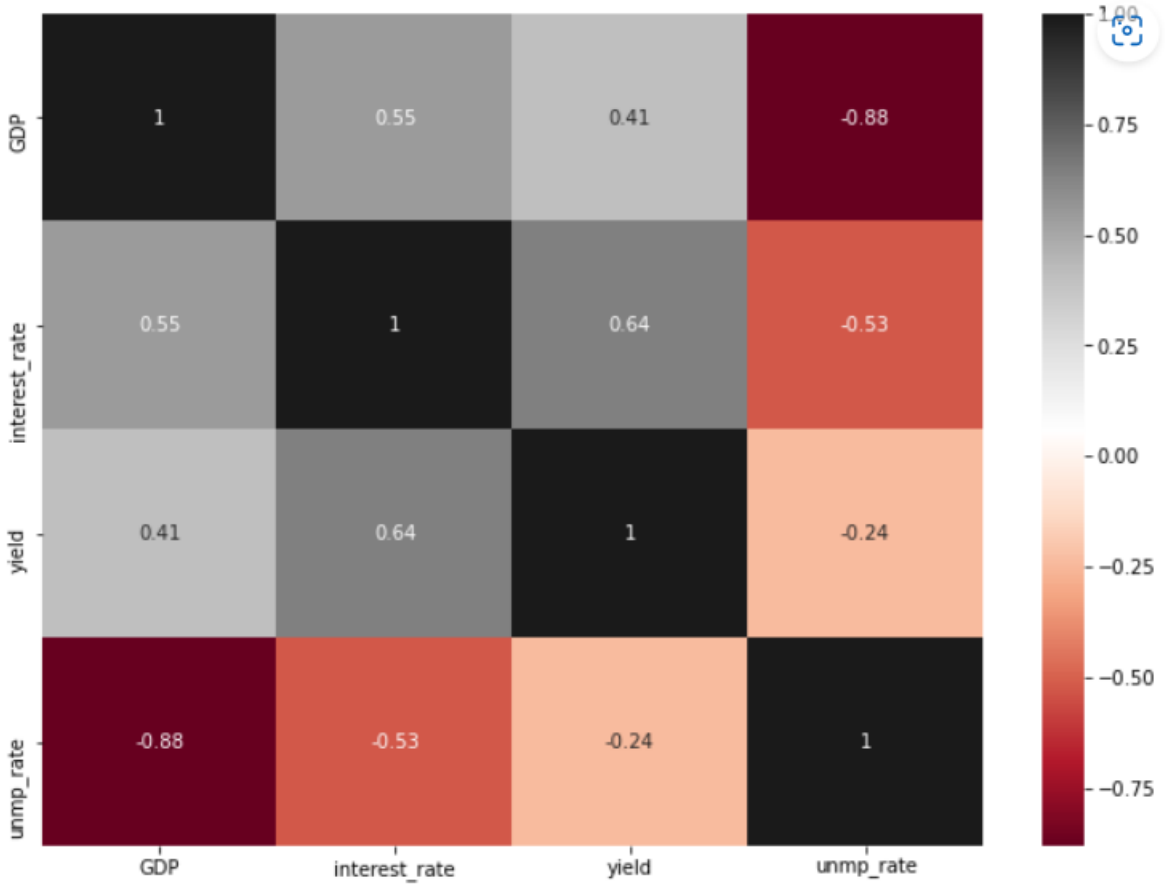


Figure 11: Unemployment correlation to variables

As we observe the above data, the target variable unemployment rate is too near to the negative correlation -0.88 of GDP. we could clearly say that the target variable unemployment rate Is inversely proportional to the rest of the variables as we have got the results in negative and hence, we say that the target variable is negatively correlated to the other macro-economic variables. We could also represent the correlated data in the tabular format as that could display the values and results of the correlation matrix. It is as follows: -

	GDP	interest_rate	yield	unmp_rate
GDP	1.000000	0.547884	0.406780	-0.880736
interest_rate	0.547884	1.000000	0.642623	-0.526349
yield	0.406780	0.642623	1.000000	-0.238474
unmp_rate	-0.880736	-0.526349	-0.238474	1.000000

Figure 12: unemployment correlation results

```
unmp_rate        1.000000
yield           -0.238474
interest_rate   -0.526349
GDP             -0.880736
```

PRINCIPAL COMPONENT ANALYSIS: -

Generally orincipal component analysis is used to store the wanted and delete the unwanted data from the datasets so that we could use the data which can be required and the rest is not needed. In this case, as our data is little huge, we use PCA to delete the unwanted data and opt out the data which is useful for our target variable unemployment rate. We use PCA to show the difference between threshold value and the explained variance. It is shown as follows: -

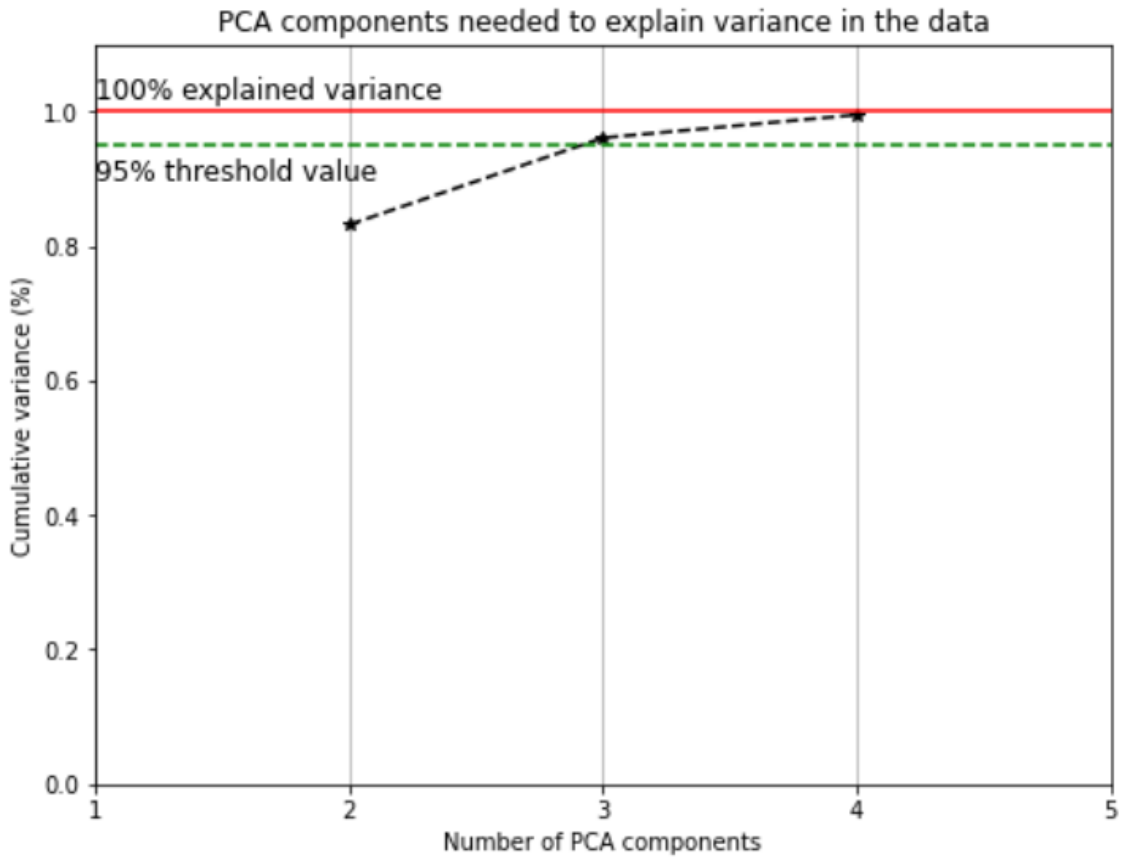


Figure 13: PCA actual

After this we use the library function sklearn and import the packages required for PCA. The calculated and analised data can be observed in the tabular format as well as heatmaps. For the initial understanding, lets have a look at tabular format:

	Pca_0	Pca_1	Pca_2	unmp_rate
Pca_0	1.098531e+00	-1.488567e-16	9.923781e-18	1.113561
Pca_1	-1.488567e-16	1.704593e-01	-2.791063e-17	0.095435
Pca_2	9.923781e-18	-2.791063e-17	4.523888e-02	-0.047526
unmp_rate	1.113561e+00	9.543525e-02	-4.752558e-02	1.473006

Figure 14: Unemployment PCA results

The above tabular format is the basic and example view of the pca analysis for the target variable. As we displayed all the rows and columns, it looks like follows: -

	Pca_0	Pca_1	Pca_2	unmp_rate
0	-1.014513	0.899995	0.365465	6.3
1	-1.071138	0.882228	0.362659	6.2
2	-1.126275	0.839669	0.369956	6.2
3	-1.197671	0.868660	0.337570	6.2
4	-1.265609	0.895467	0.312349	6.0
...	...	...	...	...
175	1.040272	-0.279566	0.071443	7.1
176	0.833327	-0.309113	0.036674	7.0
177	0.603899	-0.244201	-0.048138	6.8
178	0.388543	-0.273573	-0.069485	6.1
179	0.349347	-0.375754	-0.025913	6.0

As shown in the above graph, this is how it looks like the tabular formal of principal component analysis for the target variable unemployment rate. For better understanding, lets have a quick look at the heatmap of pca analysis.

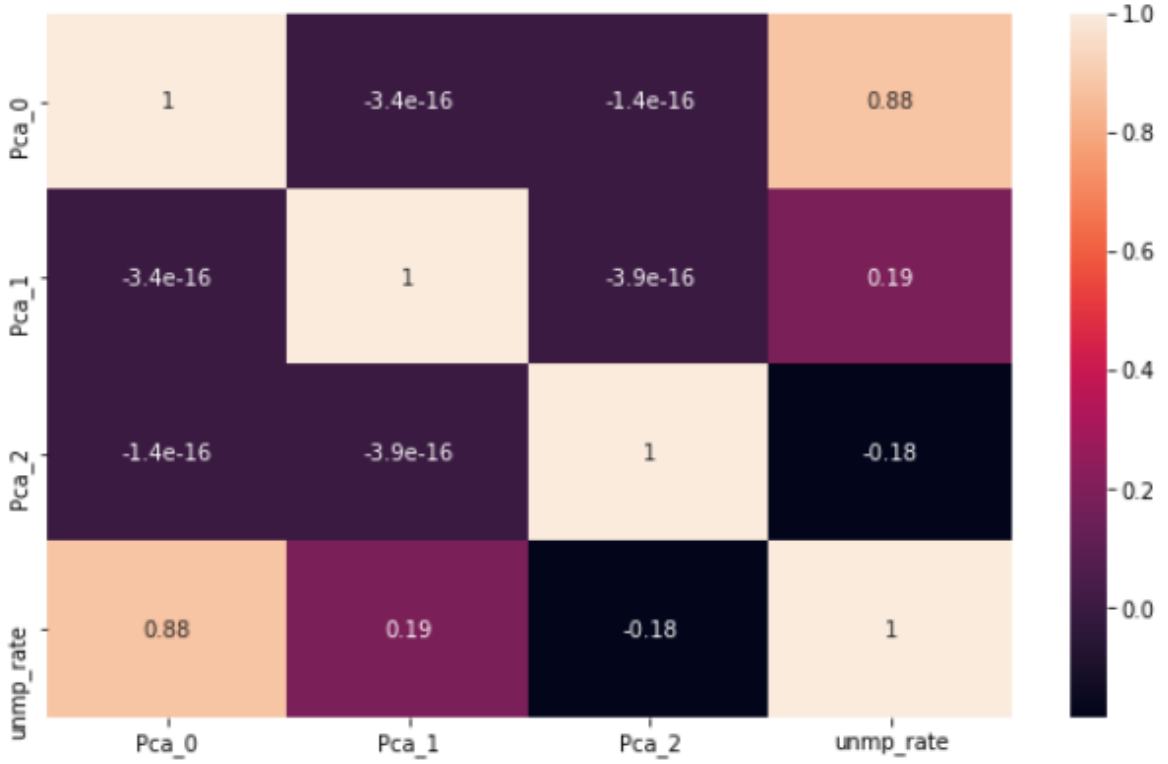


Figure 15: Unemployment rate PCA

The above graph is the heat map of the principal component analysis. As we can observe and compare both correlation and PCA, we can observe that it is negatively correlated to

the target variable in correlation as -0.88 and it is now positively correlated to target variable in this principal component analysis that is 0.88. To check whether it is true or false, we perform a covariance test so that the eigen vectors of the pca analysis states the appropriate results. The covariance graph is displayed as follows: -

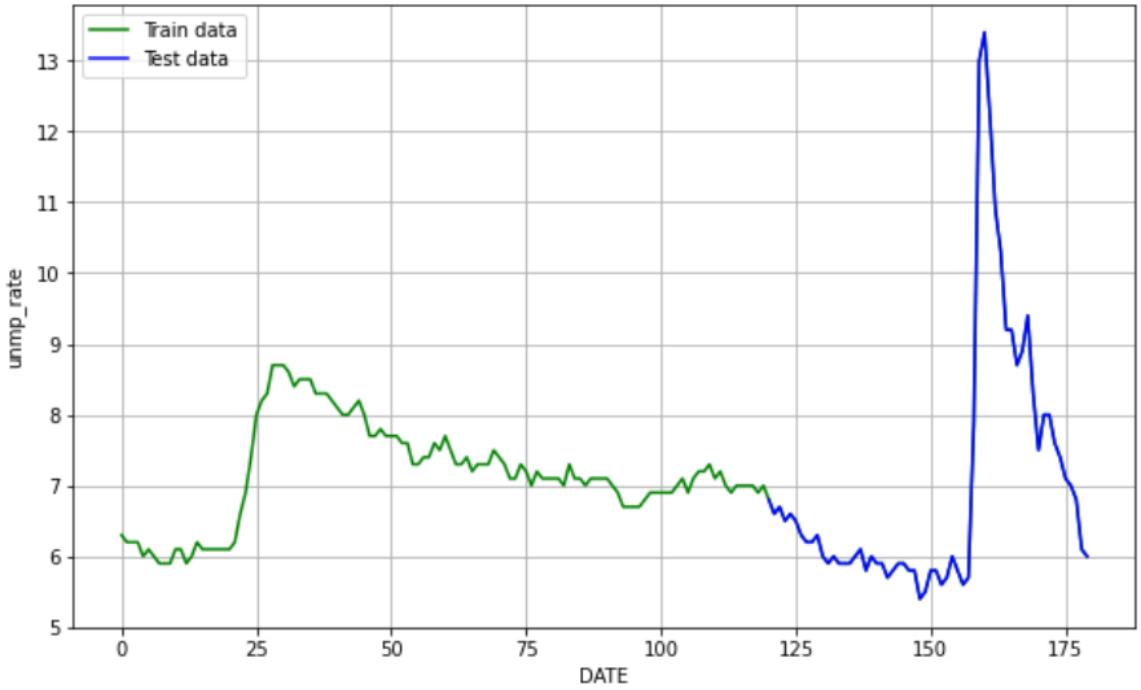


Figure 16: - Time series visualization

In the above graph, we performed the covariance test and plotted in the graphical representation as we can observe that the data of unemployment which is our target variable is most reliable and true with its results, as both the unemployment analysis and co variance analysis are more similar compared to each other. The eigen vectors and eigen values are as follows: -

```
Eigenvectors
[[ 0.64457788 -0.53915843 -0.49282165 -0.22573074]
 [ 0.0324299   0.72023329 -0.67627388 -0.15121489]
 [-0.01529736 -0.11076807 -0.32841442 0.93789146]
 [ 0.76369735 0.42225895 0.43809181 0.21572968]]

Eigenvalues
[2.4178806 0.22641116 0.10863612 0.03430726]
```

NORMALIZATION OF RAW DATA: -

When you have no idea how your information will be dispersed or when you understand the dispersion isn't Gaussian, standardisation is a good strategy to use (a ringer bend). Standardization is useful when your data has different dimensions and the computation, you're using doesn't make assumptions about the transmission of your data, for example, k-nearest neighbours and counterfeit brain organisations.

As we work on jupyter notebook, we first need to install the required packages and libraries to normalize the raw data collected. We need to install the library PANDAS, to work on normalizing the data. After normalizing the data, it is shown as follows: -

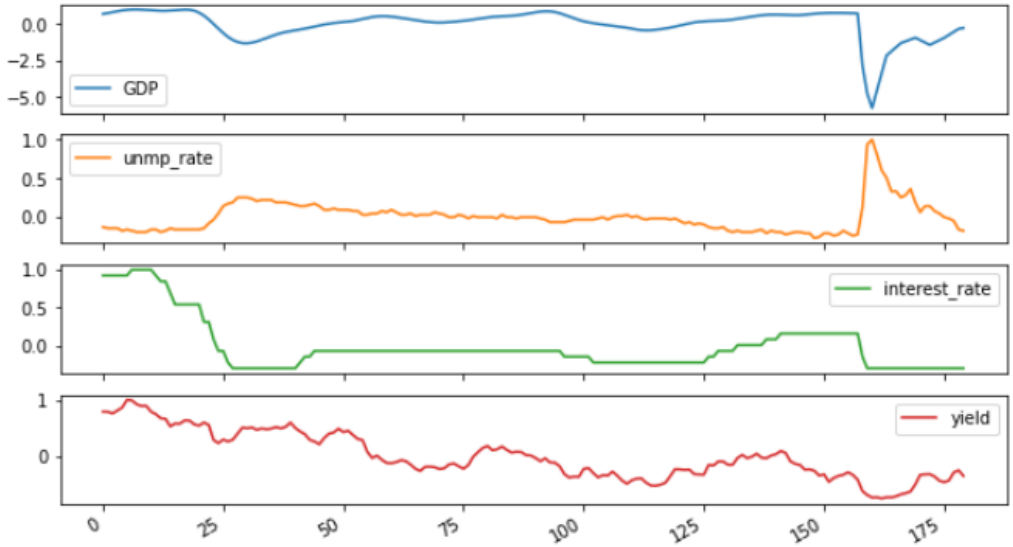


Figure 17: Normalized result visualization

The above graphical representation is the result of the normalization of the macro-economic variables selected. The target variable is normalized with the other rest of the variables. This could affect the result of the total outputs of the graphical representations. We generally use normalizing the data for the principal component analysis process as they are inter linked to one another. The results and outputs of the normalized data are further utilized in the principal component analysis before the step of forecasting. Also, from the results of principal component analysis, we normalize the data and hence it is shown as the graphical representation above. In the next step we split the data and perform regression analysis for the predictions and further time series forecasting.

LINEAR REGRESSION MODEL: -

As we performed basic checks and tests above, we nor perform regression model to check the accuracy aggregate of the target variable unemployment rate.

We got the accuracy (R-Square) result as 80.81%. The graphical representation of linear regression is as follows: -

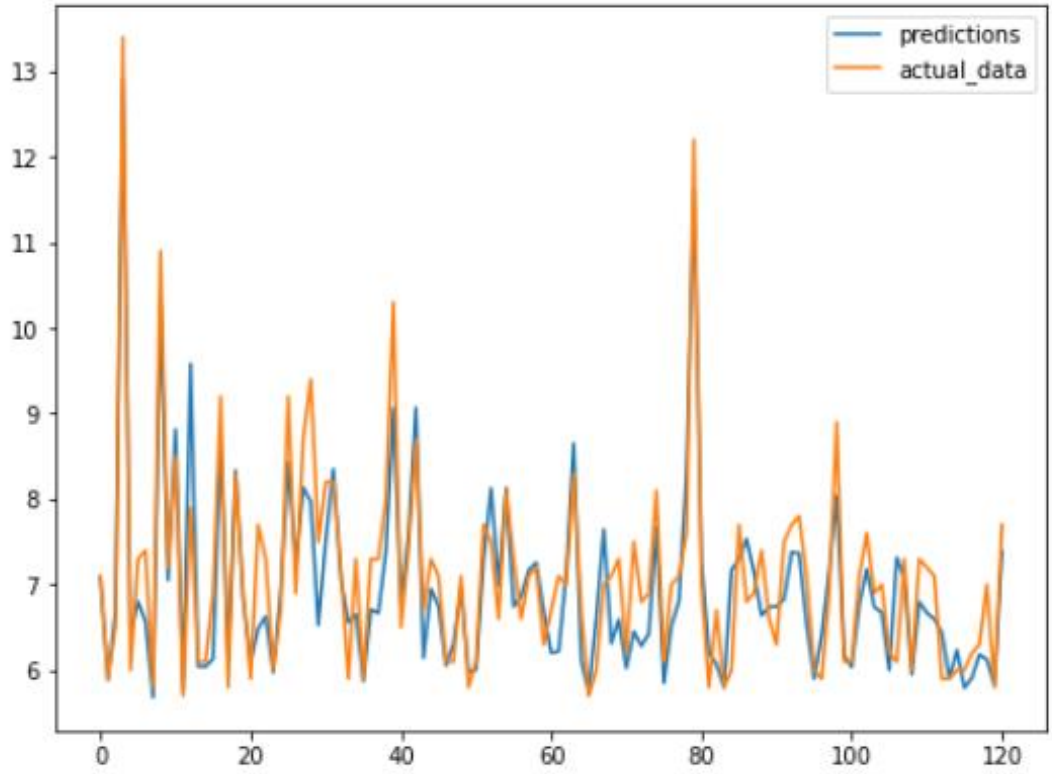


Figure 18: - Linear regression results

The above graph displays us the actual data and predicted data for the target variable unemployment rate using linear regression method. We have also calculated the mean absolute error, mean square error; root mean square error and accuracy values for the target variable unemployment rate. They are as follows: -

```
Mean Absolute Error: 0.4197200601080828
Mean Squared Error: 0.28175339963405455
Root Mean Squared Error: 0.5308044834343947
Accuracy 0.8081044615798683
```

OLS REGRESSION ANALYSIS: -

The ols regression is a method which is generally used to compare the regression values and perform multi regression at a single time. Here we compared the ols regression values of our target variable to the other variables. The results are as follows: -

OLS Regression Results

Dep. Variable:	unmp_rate	R-squared:	0.778
Model:	OLS	Adj. R-squared:	0.776
Method:	Least Squares	F-statistic:	310.9
Date:	Tue, 13 Sep 2022	Prob (F-statistic):	1.19e-58
Time:	20:36:29	Log-Likelihood:	-154.13
No. Observations:	180	AIC:	314.3
Df Residuals:	177	BIC:	323.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	59.7028	2.586	23.087	0.000	54.599	64.806
GDP	-0.5267	0.026	-20.014	0.000	-0.579	-0.475
interest_rate	-0.0691	0.047	-1.480	0.141	-0.161	0.023

Omnibus:	6.595	Durbin-Watson:	0.328
Prob(Omnibus):	0.037	Jarque-Bera (JB):	6.550
Skew:	-0.467	Prob(JB):	0.0378
Kurtosis:	3.046	Cond. No.	6.02e+03

Figure 19: OLS Regression results

The F-test provides information about the model's overall significance.

At the end of the day, it tells you whether the autonomous elements that you used in the model are significant.

The essential assumption of the flawed conjecture is that your model will fit the data just as well as one with no autonomous elements.

A model, on the other hand, only considers the block.

According to the elective speculation, your model is more genuine for investigating information than the model that primarily considers the capture. The P test is a measured tool for examining the significance of the relationship between the autonomous and reliant variables.

In this case, Gross Domestic Product, Joblessness, Loan Cost, and Yield are all subjected to discrete significance tests with Pay to see whether they are significant.

The underlying assumption of the erroneous hypothesis is that the free component has no influence on the value of the reliant variable.

TIME SERIES MODELS: -

Time series analysis is used to predict the future of the present data of the target variable unemployment rate. To perform forecasting analysis, we first need to test the data we have if the data is stationary or non-stationary. If the data is non stationary, then we need to perform auto correlation and partial differentiation si that the data gets stabilized into stationary form.

Here, we discuss a more comprehensive test to confirm whether or not a model is fixed.

The P-value should be less than 0.05.

If the p-value is more than 0.05, it is considered weak evidence against the erroneous theory that information isn't fixed.

Otherwise, it is regarded as convincing evidence against the incorrect speculation. Reject the hypothesis, and the information is considered fixed.

If information is not fixed, we will perform differencing to fix it.

Differing is simply changing one value and then treating it as invalid; at that point, we will attempt to determine whether the information is fixed. The following differences are the adf tests performed: -

```
ADF Test Statistic: -7.265888873543857
p-value: 1.6354917068214632e-10
#Lags Used: 10
Number of Observations: 167
strong evidence against the null hypothesis (Ho), reject the null hypot
hesis. Data is non-stationary
```

As the result above is not stationary, we need to partialize the data and make it into stationary data to perform forecasting analysis and obtain their results.

```
ADF Test Statistic: -4.262037309782892
p-value: 0.0005164781036158478
#Lags Used: 10
Number of Observations: 166
strong evidence against the null hypothesis (Ho), reject the null hypot
hesis. Data is stationary

As the p value is <0.05, we now consider that the data is stationary.
```

Following differentiation, we will plot the relationship and fractional autocorrelation plots to establish a P and Q value, and we will simply confirm the connection and autocorrelation by evaluating whether our data is irregular.
The plots depicting connection and autocorrelation charts are shown below: -

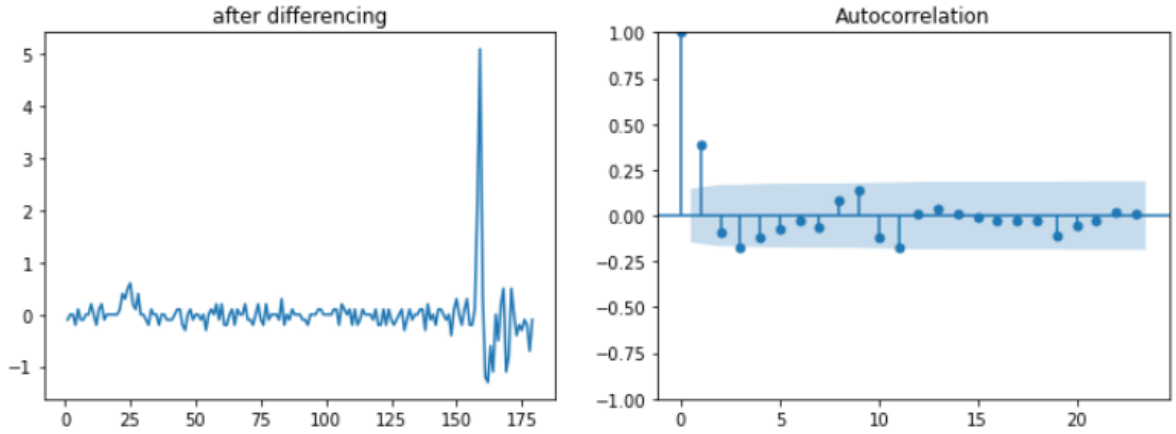


Figure 20: - Auto correlation visualization

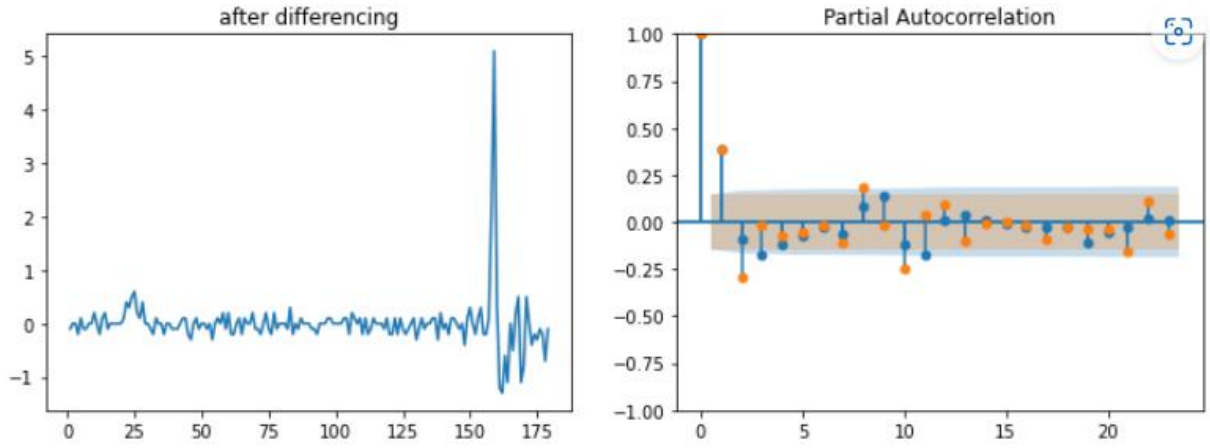


Figure 21: - Partial Auto correlation visualization

ARIMA MODEL: -

We shall use a certainty stretch while performing estimate values. The standard certainty span is 95%.

It is used to assess the correctness of expected information values in the centre of certainty stretches.

Certainty Spans are used to determine whether a stretch has a 95 percent chance of being obvious. As a result, we design a 95% confidence stretch esteem, which is depicted as a hazy region in the diagram and will encompass all clear advantages of the objective variable. We now try to plot the graph and observe the predicted graph of the auto regressive model: -

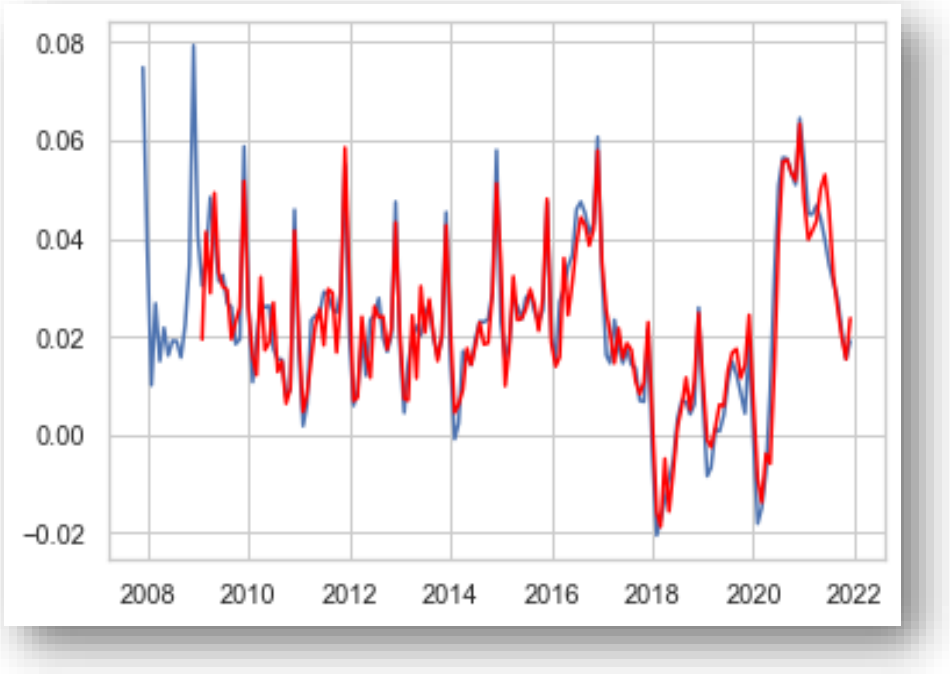


Figure 22: - ARIMA visualization

In the above graph, we could observe the actual and predicted line graphs as shown. As we have collected the data from 2007 to 2021, we have predicted the future forecasting for the year 2022. And its results are too easy to understand more in the tabular format. The results of ARIMA are as follows: -

Dep. Variable:	unmp_rate	**No. Observations:**	117
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	59.024
Date:	Tue, 13 Sep 2022	**AIC**	-112.048
Time:	19:46:34	**BIC**	-103.787
Sample:	0	**HQIC**	-108.695
	- 117		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7807	0.154	5.064	0.000	0.479	1.083
ma.L1	-0.6014	0.193	-3.122	0.002	-0.979	-0.224
sigma2	0.0211	0.003	6.578	0.000	0.015	0.027

Ljung-Box (L1) (Q):	0.18	**Jarque-Bera (JB):**	3.51
Prob(Q):	0.67	**Prob(JB):**	0.17
Heteroskedasticity (H):	0.52	**Skew:**	0.40
Prob(H) (two-sided):	0.05	**Kurtosis:**	3.28

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 23: - ARIMA Tabular results

The (p,q,d) values for this ARIMA model stood at (1,1,1) and also the p-value is lesser than 0.05 for all the errors so we consider this model as best fit for forecasting analysis.

SARIMAX MODEL: -

SARIMAX model is possibly thought about when information exhibits irregularity, here our information isn't irregular anyway after differencing it shows a few qualities of irregular, thus we only attempted to carry out SARIMAX model. Though we have stated that ARIMA is the best fit fore time series forecasting, I've also tried of doing sarimax model to find if we get any better results. And the results are displayed as follows: -

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	180
Model:	SARIMAX(1, 0, 1)x(1, 1, [], 24)	**Log Likelihood**	-104.340
Date:	Tue, 13 Sep 2022	**AIC**	216.680
Time:	19:53:38	**BIC**	228.880
Sample:	0	**HQIC**	221.635
	- 180		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8886	0.048	18.446	0.000	0.794	0.983
ma.L1	0.5327	0.086	6.199	0.000	0.364	0.701
ar.S.L24	-0.4441	0.173	-2.563	0.010	-0.784	-0.104
sigma2	0.2121	0.013	16.053	0.000	0.186	0.238

Ljung-Box (L1) (Q):	0.21	**Jarque-Bera (JB):**	10528.84
Prob(Q):	0.65	**Prob(JB):**	0.00
Heteroskedasticity (H):	9.64	**Skew:**	4.85
Prob(H) (two-sided):	0.00	**Kurtosis:**	42.06

Figure 24: - SARIMAX Tabular results

The results of SARIMAX are as follows. As we can observe, the p,q,d values are not stationary and they kept on changing, not even a single set of p,q,d values satisfied the p value to be less than 0.05 as we have the p value 0.10. We do not consider SARIMAX as the best fit model to the thesis.

CONCLUSION AND SUMMARY: -

- We collected the data from many sources, cleaned it and used for analysation.
- We compared each macro-economic variable to one another.
- Used machine learning models and Time series models to find the best accuracy and fitted model to predict the data for future.
- We have also stated example for each and every model.
- We have performed PCA analysis and Correlation matrix to the collected data. Found eigen vectors and values for the target variable.

  ⇨ PCA results: - 0.88 % positively correlated.
  ⇨Correlation Results: -

```
unmp_rate        1.000000
yield           -0.238474
interest_rate   -0.526349
GDP             -0.880736
```

- We have found the mean absolute error, mean squared error, root mean squared error and accuracy values for all the models.
- Linear regression model is best to calculate regression. Its result is 80.81% accuracy.
- We have also used OLS regression method just to test if linear regression method satisfies the least squares or not.
- Out of all used models, Linear regression model in machine learning models, and ARIMA model in forecasting analysis are the best models that fit the both accuracy and predicted data for the macro-economic variable unemployment rate.

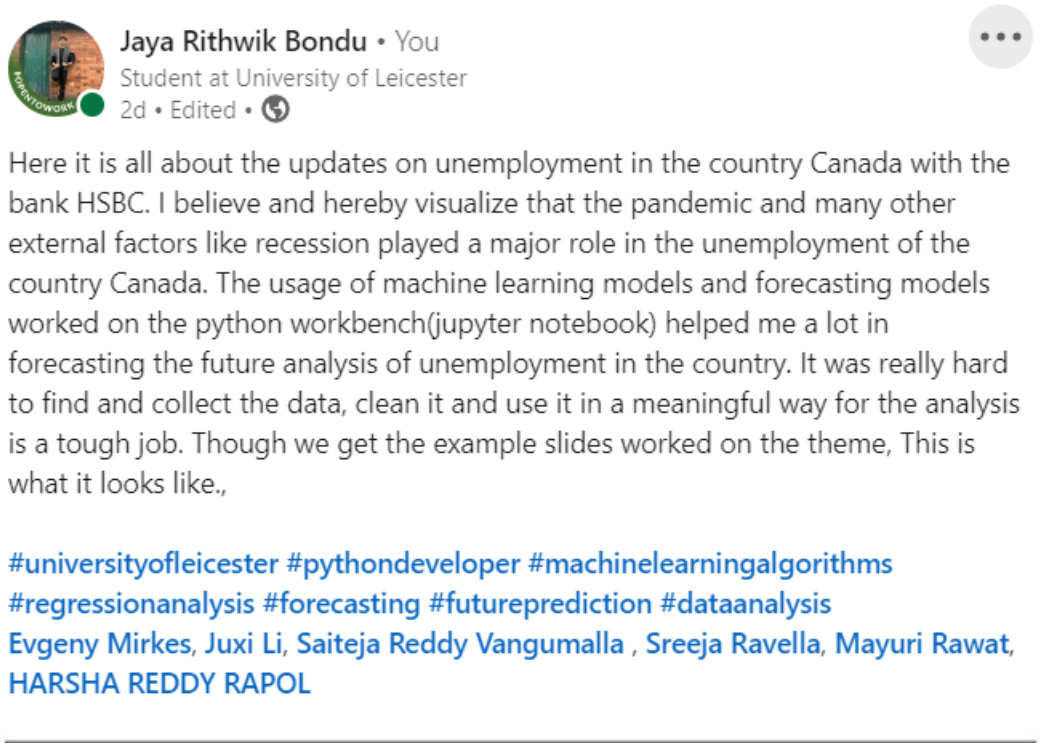
LINKEDIN POSTS: -

POST-1: -

https://www.linkedin.com/posts/jaya-rithwik-bondu-389037243_universityofleicester-hsbccanada-bankofcanada-activity-6979107963755995136-YyDw?utm_source=share&utm_medium=member_desktop

POST 2: -

https://www.linkedin.com/posts/jaya-rithwik-bondu-389037243_universityofleicester-pythondeveloper-machinelearningalgorithms-activity-6979112760517263360-_J4O?utm_source=share&utm_medium=member_desktop



Jaya Rithwik Bondu • You
Student at University of Leicester
2d • Edited •

Here it is all about the updates on unemployment in the country Canada with the bank HSBC. I believe and hereby visualize that the pandemic and many other external factors like recession played a major role in the unemployment of the country Canada. The usage of machine learning models and forecasting models worked on the python workbench(jupyter notebook) helped me a lot in forecasting the future analysis of unemployment in the country. It was really hard to find and collect the data, clean it and use it in a meaningful way for the analysis is a tough job. Though we get the example slides worked on the theme, This is what it looks like.,

#universityofleicester #pythondeveloper #machinelearningalgorithms #regressionanalysis #forecasting #futureprediction #dataanalysis
Evgeny Mirkes, Juxi Li, Saiteja Reddy Vangumalla , Sreeja Ravella, Mayuri Rawat, HARSHA REDDY RAPOL

POST 3: -

https://www.linkedin.com/posts/jaya-rithwik-bondu-389037243_universityofleicester-dataanalysis-pythonprogramming-activity-6979176259322843136-TOTV?utm_source=share&utm_medium=member_desktop

Jaya Rithwik Bondu • You
Student at University of Leicester
2d • Edited •

Hey everyone.. here's the final update of my project Prediction Market for Macro-Economic Variable Unemployment Rate (HSBC-CANADA).

I just wanted to mention that the project has been completed in time and here ar its highlights.

I've gathered data from various websites in the banking sector. Fred, kaggle, and bank of Canada helped me a lot in finding the datasets.

All the macro economic variables used are Interests rate, yield, unemployment rate and GDP. My target variable is Unemployment rate.

Unemployment rate is mainly increased In the very previous years due to pandemic and i thought of analysing the unemployment affects in the banking sectors as they are the cheapest aswell as the richest forms of income. So i have predicted the future forecasting of unemployment for the further years.

I've used few machine learning models and forecasting models in this process to find their best match. They are,

REFERENCES: -

1) Statista. (2019). *Unemployment rate Canada 2019 | Statista*. [online] Available at: https://www.statista.com/statistics/578362/unemployment-rate-canada/.

2) Koehrsen, W. (2017). *Random Forest in Python*. [online] Medium. Available at: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.

3) Dutta, A. (2019). *Random Forest Regression in Python - GeeksforGeeks*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/random-forest-regression-in-python/.

4) www.bloomberg.com. (n.d.). *Bloomberg - Are you a robot?* [online] Available at: https://www.bloomberg.com/news/articles/2022-09-09/canada-sheds-jobs-for-third-month-unemployment-jumps-to-5-4

5) Economicpoint.com. (2019). *Macroeconomic Variables*. [online] Available at: https://economicpoint.com/macroeconomics/variables.

6) Jaadi, Z. (2019). *A Step by Step Explanation of Principal Component Analysis*. [online] Built In. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

7) Nist.gov. (2019). *6.4. Introduction to Time Series Analysis*. [online] Available at: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm.

8) pro.arcgis.com. (n.d.). *How OLS regression works—ArcGIS Pro | Documentation*. [online] Available at: https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/how-ols-regression-works.htm

9) thismatter.com. (n.d.). *Macroeconomic Terms and Variables*. [online] Available at: https://thismatter.com/economics/macroeconomic-terms-variables.html.

10) Capital One. (n.d.). *Understanding ARIMA Models for Machine Learning*. [online] Available at: https://www.capitalone.com/tech/machine-learning/understanding-arima-models/.

11) phosgene89.github.io. (n.d.). *From AR to SARIMAX: Mathematical Definitions of Time Series Models*. [online] Available at: https://phosgene89.github.io/sarima.html.

12) Team, T.A. and Team, T.A. (n.d.). *How, When, and Why Should You Normalize / Standardize / Rescale Your Data? – Towards AI — The Best of Tech, Science, and Engineering*. [online] Available at: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.

13) Dela Cruz, P., Abante, M.V. and Garcia-Vigonte, F. (2022). *Systematic Literature Review: Unemployment Rate as factors affecting the Gross Domestic Product, Inflation Rate, and Population*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4121167.

14) gocardless.com. (n.d.). *Introduction to Macroeconomics*. [online] Available at: https://gocardless.com/guides/posts/introduction-to-macroeconomics/.

15) Google Developers. (n.d.). *Normalization*. [online] Available at: https://developers.google.com/machine-learning/data-prep/transform/normalization.

16) Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), p.20150202. doi:10.1098/rsta.2015.0202.

17)TheBMJ (2019). *11. Correlation and regression | The BMJ*. [online] Bmj.com. Available at: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression.

18)Duke.edu. (2019). *Introduction to ARIMA models*. [online] Available at: https://people.duke.edu/~rnau/411arim.htm.

19)How, When, and Why Should You Normalize / Standardize / Rescale https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

20)GitHub - Adit14/Time-Sereies-analysis-of-Foreign-Exchange-Data-using https://github.com/Adit14/Time-Sereies-analysis-of-Foreign-Exchange-Data-using-CRISPDM