

Final Report

Modelling and forecasting the interest rate changes.

Name: Viswanath Durairaj

DECLARATION

All sentences or passages quoted in this report, or computer code of any form whatsoever used and/or submitted at any stages, which are taken from other people's work have been specifically acknowledged by clear citation of the source, specifying author, work, date, and page(s). Any part of my own written work, or software coding, which is substantially based upon other people's work, is duly accompanied by clear citation of the source, specifying author, work, date, and page(s). I understand that failure to do these amounts to plagiarism and will be considered grounds for failure in this module and the degree examination.

Name: Viswanath Durairaj

Date: 01/06/2022

Table of contents

1. Introduction.....	7
2. Step of Data Analysis.....	8
2.1 Business Objective	8
Phases for Business Objective:	8
2.2 Data collection.....	9
2.3 Data Pre-processing	10
2.3.1 Data Cleaning	10
2.3.2. Data Integration	12
2.3.3 Data Transformation	12
2.3.4. Data Reduction	13
2.4 EDA Process	14
2.5 Model selection and building test model.....	15
2.5.1 Supervised Learning	15
2.5.2 Unsurprised learning.....	21
2.5.3 Reinforcement learning	21
2.6 Time Series Analysis	22
2.6.1 ARIMA Model	23
3. MySQL.....	24
4. AWS Instance.....	25
4.1 AMI - Selection	25
4.2 Choosing the instance	26
4.3 Configure instance.....	26
4.4 Storage Adding	26
4.5 Adding Tags	26
4.6 Configuring Security Group	26
4.7 Review	27
5. Variable selection	27
5.1 Inflation	27
5.2 Interest rate.....	27
5.3 Unemployment.....	28
5.4 Credit card Lending to household.....	28

5.5 Average weekly Income	28
5.6 GDP	28
6. Implementation.....	29
6.1 Plotting the Time series	30
6.2 Data and Correlational Matrix	32
6.3 Dimensionality Reduction	33
6.4 Multi linear Regression model Building	33
6.5 ADF Analysis	36
6.6 ARIMA model Forecasting	37
6.7 Forecasted data applied on PCA	40
6.8 MLR applied on Predicted PCA	41
7. Future Work.....	41
8. Conclusion	42
9. Bibliography.....	43
10. Appendix.....	44

Abstract

The whole world is under the control of "Economy". Working as a part of the corporate side, which is the bank sector. The banks are usually exposed with high risk economic factors, since it is involved with capital market, trade, and sale. The primary interest rate is set by bank of England or the federal bank. The central bank will increase the interest rate to reduce the demand, balancing the rate of economic growth. I have worked on this project regarding the prediction of Interest rates to obtain stability economically when inflation in interest rate affecting in the future outcomes. The sector of interest rate focused on this project is households interest rate. Initially data was collected from the "Bank of England website", "Office for National statistics", The classification of models are researched, and the suited requirements are model building and Time series forecasting has been done to predict future interest rate using the collected historical data. We use the python program to perform the data analysis. For database server used MYSQL, to make the remote connection possible, by using the AWS instance as a schema because it will be used as a retrieving tool like even though host system is shutdown it can be accessible. In our project, we performed back testing, check for a good strategy and model for selection. The desired output was predicted which tends the banking sector to take needful actions to maintain balance economically on the effect in increase of the interest rate.

Acknowledgment

There was immense support given in my academic journey at the university over my study year. I acknowledge the support given by my project supervisor, whom has been there for my consultation and given me guidelines from the start to end of my final dissertation. Finally thanking the university for the required facilities and fraternity for ensuring that I have acquired adequate skills and expertise to complete my project.

I'm really thankful to Jeremy Levesley, Academic Supervisor Alexander Gorban, Program Director Andrew Morozov (Dr.), for giving me the opportunity, constant support and encouragement throughout the journey.

1. Introduction

In this study we will analyse the data obtained from bank of England, this study is based on interest rate prediction of credit card lending households. Interest rate has a crucial role in the banking sector, by considering interest rate we can bring valuable insights into this project. The results obtained through analysis and model building can help banks across the United Kingdom.

Changes in Interest rate can affect other factors around it, such as currency value. Interest rate in the company have gone up from 0.1% to 0.25% in the year 2021. There are people around the country with questions such as “why have interest rate gone up?”, “how much have interest rates gone up?”, “how far will it keep increasing?”. People have other concerns on how rise in interest rate can help in inflation. The purpose of this project is to predict the rise in interest rate in coming future. The interest rate focused on this project is the interest rate of households around the country. This report will explain in detail on the data processing and implementation methodology used to build the model.

Interest rate and banks profitability are interlinked as banks make more money by taking advantage of the difference between the interest rate that banks pay to customer and interest the banks can earn by investing. Also, how interest rates reduce the associated loan spreads, thereby, increases banks risk appetite. Hence, there is a need to predict the interest rates for the coming years. Interest rate also affects the ability of an individual to pay their debts. Rise in inflation will increase the interest rate, this will reduce the mortgage approvals, the currency value and gross domestic product (GDP). All this will contribute to unemployment, which is a concerning factor in the country's economy. The aim is to provide a futural insight to banks on how the interest rate changes are going to be in the coming years. This can help financial sectors such as banks to see the future ups and downs and act accordingly.

This report will further expand on the steps taken in performing the analysis of data, model building and time series forecasting. This report will make the reader understand the implementation in detail and the underlying techniques used to make the prediction.

About Bank of England

The Bank of England is the United Kingdom's central bank and serves as a foundation for most modern central banks in the world. It is one of the world's oldest bank and placed 8th in the order, having been founded in 1694 to serve as the banker for the English government and still one of the bankers for the United Kingdom government. The Attlee ministry nationalised the bank in 1946.

In 1998, the Bank became a separate public institution, completely owned by the Treasury Solicitor on behalf of government, but with independent flexibility by creating monetary policy.

The Treasury Solicitor's Department is the old name for The Government Legal Department, which is the United Kingdom's Government Legal Service's largest in-house legal organisation. The Treasury Solicitor oversees the department [2]. The Treasury Solicitor Act of 1876, which established the Treasury Solicitor as a single organisation. It established the post in law, an office with perpetual succession. The employees of the department wield legal authorities that are solely vested in the organisation.

Monetary policy is the ability of a central bank of that country or the respective government to influence the amount of money in the economy and the cost of borrowing [2]. The central bank of The United Kingdom implements two primary monetary policy.

- The Bank Rate, which is the interest rate charged towards the banks for borrowing money from the central bank.
- Quantitative easing is used to acquire bonds to cut interest rates on savings and loans.

The Monetary Policy Committee makes decisions on monetary policy. The meeting will be eight times in a year, The policy will be structured and announced roughly six weeks once. The committee is consisting of nine individuals. They hold multiple meetings to examine the state of the economy before deciding what action to be taken in situations needed. The entire impact of monetary policy on the economy can take up to two years. As a result, the members must examine what inflation and economic growth are likely to be in the future years.

The Bank is one in all eight banks authorised to issue banknotes within the UK, features a monopoly on the difficulty of banknotes in England and Wales and regulates the problem of banknotes by commercial banks in Scotland and Northern Ireland. The Bank's headquarters are in London's main financial district, town of London, on Threadneedle Street, since 1734. it's sometimes called the old lady of Threadneedle Street, taken from a cartoon by James Gill ray in 1797.

As a regulator and financial organisation, the Bank of England has not offered consumer banking services for several years, but it still does manage some public-facing services like exchanging superseded bank notes. Until 2016 the bank provided personal banking services as a privilege for workers.

2. Step of Data Analysis

The main motive of designing the data analytics lifecycle is to solve the big data problems and module building project. The data analytics lifecycle is separated into variety of phases, which either move forward or backward at the phases in the lifecycle, while working on the real time project, analyst might be able to move forward phases and may return to easier stage as to gather the more information which is missed or uncovered. This allows the data analyst to iterative phase and move towards effective project result. Data lifecycles fall on several phases:

2.1 Business Objective

The understanding the business objective is important part in analysis, and this initially process. For achieving business, there should be a set of business goals. The main is to identify what is the process is going to done and what is the purpose of the company. We should discuss the project details for the respective management and executive level. The objective for the business is to be specific and measurable, which results in growth for companies.

Phases for Business Objective:

1. What are you achieving?

Like the Moscow technique, we should split the work under this technique. Moscow stands for must, should, could, and won't, where this case satisfies the necessary objective to proceed with it. The objective is judgment towards work. We should make a proper decision and note the conflict what is the growth for the long-time and profit for the short time. For, every phase we should set an objective for a business no matter large organization or a small organization. There are set of principles are individual objectives, gathering company targets, and a team objective.

2. Why are you trying to achieve those objectives?

There is more reason to get business success. There are two important business strategy objectives and goals. Either doing objectives or goals does not rely on successful understanding. Both conditions should be satisfied. Nowadays, the state of objective is to be clear, and it should give us clarity and should provide a clear format for workers. The focused output should be meaningful and to be measurable.

3. Make Objective Good?

SMART is a technique where to set an objective for a company. This approach follows some rules where the goal is to be minimal and should make a note of when the goal is to be achieved, and what will be the outcome for that goal. SMART stands for specific, measurable, achievable, relevant, and timely.

- **Specific** - This is the first step of the approach the business objective should be specific. The goal is to be precise. While fixing this we should maintain a set of questions that what the goal is to be achieved and in between we should be responsible for what we have done.
- **Measurable** - In this approach, it has a high and a low, gain or loss, uptrend or downtrend which satisfies profit margins, quality, and quantity. Often, this approach should be measured or else it drops to a low margin.
- **Achievable** - Here, the objective is to be an achievement, not dissatisfaction. It should be realistic if not the organization will fall towards achievement. It is not too easy to achieve business goals and the margin is set high.
- **Relevant** - The business approach is to be relevant to waste time on paper works and build efforts towards objectives. It measures relevant to achieving the target audience. Should identify the plan properly which should have, and which is not a necessity.
- **Timely** - The company should maintain timeframes then only it realizes the time to achieve business objectives. It must keep a set of timeframes for the short period and maintain some timeframe for the long run of the objective.

2.2 Data collection

Data sourcing and collection are second step in data analysis. Data collection is process of collecting the information, to find solution for research problem or about a specific problem statement, to evaluate outcomes and time series forecast trends. Data collection will fall on two methods:

Comparison	Primary Data	Secondary Data
Data	Real time data.	Historical Data.
Source	Experiment, questionnaires, observations, personal interview etc.	Website, Journal articles, Government Publications.
Process	Very Involved.	Quick and easy.
Collection Time	Take a long time.	Take a short time.

Reliability & suitability	It is collected for a particular purpose, so data will be more reliable and suitable.	Secondary Data is less reliable and suitable, because someone else has collected the data, it may not perfectly suit for our purpose.
--------------------------------------	---	---

While performing the data research, you'll require to evaluate their hypothesis, therefore it's important to grasp the major difference between quantitative and qualitative data.

	Quantitative Data	Qualitative Data
Data type	Observation, objects, picture, symbols.	Number and statistics.
Definition	The information can't be delivery as the numerical.	The information can be expressed as the numerical value.
Data analysis purpose	Pattern, interpret social interaction, explain and understanding.	Develop the prediction for the future, check cause and effect.

2.3 Data Pre-processing

The process of analysing the data, according to certain features in the data file, to be worked on as an input data. The methods performed in the data handling like making corrections to the corrupted or inaccurate data from the file used as input dataset. It is also used to clear or delete the data which are similar or duplicated and even mislabelled, this occurs when the data set is created using multiple sources.

The results and the output determined from the input data set given will not be trustworthy, in case of the data set is not suitable for the required output or it can be an erroneous data set. The entire process has different phases and time differs which tends to work on every data set in different methods according to the requirements, hence creating a template for the process will ensure the execution of all the phases in time [5]. The process completes the Data pre-processing will be as follows [5],

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction

2.3.1 Data Cleaning

Appropriate data cleaning will save you time and money while increasing the efficiency of your business. Data scientists believe that effective data is more crucial than the most sophisticated algorithms in machine learning. This is because the data used to train machine learning models is only as good as the data used to train them. In case, if you are using improper data to train your models, the eventual results will not only be untrustworthy, but also potentially damaging to your company [7].

If you train the machine learning model with duplicated data, probably the result will give weight for duplicate. Data will be collected from the multi-source/ website, may be use the scraped data from the data analysis, or had collected from the multi survey or client responses, it will frequently end up with duplicated data. Duplicate records/data will cause, slow down analysis and more storage will be required. So, deduplicate data need to remove for the well-balanced result [7].

Data cleaning Techniques fall with following steps.

- Handling missing value
- Noisy
- Remove Outlier

Handling missing value

In dataset some of variable will have the missing value, data may not store/ missed for some of the period, or human error, there can be multi reason for missing value from the dataset. The missing value are quite common occurrence in real world dataset.

While passing the data in the machine learning model missing value can skew the result and degrade the model accuracy. In the place of missing, it commonly denoted as “Not Available” or “NA”, this may be used as to refill the value. It can also fill manually, but it’s difficult for the large dataset. For dealing the large dataset, there some of the method want to apply likely:

- The attribute mean value of the variable can be used to refill the “NA” value when the data is normally distributed.
- For the Non – normally distribution, the attribute median value will be used to replace.
- In case of regression or decision tree algorithm, the probable value will be refilling the missing place.

Noisy Data

The data with contents the large amount of meaningless information is called as noisy data. Noisy data increases the amount of storage space required unnecessarily and might have a negative impact on model building. It can be generated because of data entry error/ human error, data collection from the various resource. There are some of way to handle the this

- Binning Method
- Regression
- Clustering

Removing the Outlier

The uncommon value in dataset is called as outlier, it might cause of affects the statistical result and increase the variability in the dataset. The outlier is like error value, it’s easy to identify, while analyst will be confronted with outlier and will make plan to handle the outlier [6]. Therefore, omitting the outlier can bring good result and more statistically significant. In the below Figure 1, you can be able see to outlier.

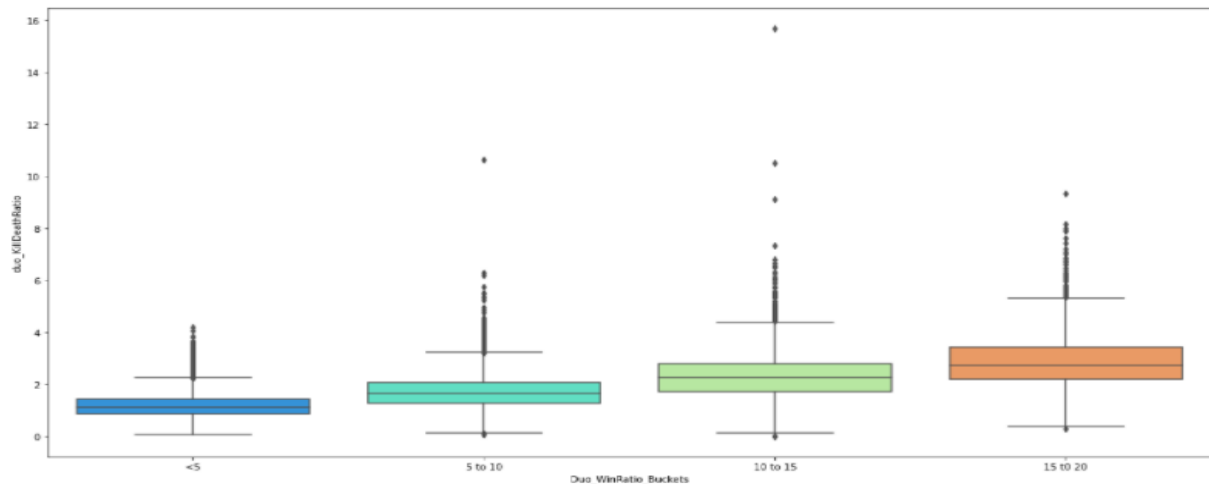


Figure 1- Outlier

2.3.2. Data Integration

The goal of the data integration is used to gain the valuable information to help for solve the problems and gain new insight, while organizations spend a lot of time and money collecting data, data is only valuable as it is used, the challenge is data soils, data is often stored across many departments in various formats that needs to transform.

Data Integration simply put combines data from the multi-source into the single large dataset likely data warehouse. This process is one of the major components in the data management. There are various problems to be consider during the data integration.

- *Object matching and Schema integration* - The format and attributes in data might cause struggle in this process.
- Removing redundant attributes.
- Detection and resolution of data value conflicts.

2.3.3 Data Transformation

After the data cleaning process had done, we need to change the structure, value, or format of data as the usable format by utilising the Data Transformation methodologies listed below to condense quality data into other formats.

1. Generalization
2. Normalization
3. Attribute Selection
4. Aggregation

1. Generalization

The idea which uses hierarchies to covert low-level or granular data to high-level information. To set an example changing the initial information data of the given address, naming the town to nation which is generally changing the data to a higher-level information.

2. Normalization

Data Normalization is the most important techniques in data transformation. The numerical value is fitted in certain range in which the way of scaled up or down is carried in this process. The data value will be impelled to a certain container to create correlation between multidata point. Data normalization can be determined by various methods such as Minimum – Maximum normalization, Z- score normalization and Decimal Normalization.

3. Attribute Selection

Attribute selection is the process to reduce the number input variable to apply in model build. The purpose to reducing the input variable is to reduce both cost of computational modelling and to increase the performance of the model. For reducing the input variable, we use the statistical based method for feature selection, the method which involve to evaluating the relationship between the individual input variable and the target variable by applying statistics, the variable which selected will have strong relationship with target variable.

The above-mentioned method will be faster and more effective, the reason of choosing the statistical measure is depends on both input variable and output variable, and its data type.

4. Aggregation

For achieve the effectively business objective, present data in summarized format for using the statistical analysis form the collected data is known as the data aggregation. This process helps us to provide the ability of future forecasting and prediction modelling, it's important in data warehousing as it help for decision making depend on the vast amount of dataset. Data aggregation techniques effectively help to minimize the performance problem, and provided tiding based on related clusters. There are some of the types of aggregation with mathematical function

- **Sum** – All the specific data are added together to get a total
- **Average** – Average value of the data.
- **Min** – Shows the lowest value of the variable.
- **Max** – Shows the highest value of the variable.
- **Count** – The total amount of data in individual category.

2.3.4. Data Reduction

In data analysis, difficult to handle the too large dataset, there is one option is to generate a simplified representation of the dataset that is substantially less in size but delivers similar analytical findings. There are various data reduction strategies.

1. Data cube aggregation
2. Dimensionality reduction
3. Data compression
4. Discretization
5. Numerosity reduction
6. Attribute subset selection

1. Data cube aggregation

Data cube aggregation store the multi-dimensional aggregation, this operation is applied for construction in data cube format. Generally, it is used for the interpret the data, and to aggregate the data in the summarized form.

2. Dimensionality reduction

The feature extraction is performed by the way of dimensionality reduction techniques, the attribute or individual factor are referred to the dimensionality of a dataset. This main purpose of dimensionality reduction technique is to reduce the redundant feature which consider in the ML algorithms. “Principal Component Analysis” method is one of the common methods used to reduce the dimensions [1].

3. Data compression

The size of the data reduced significantly, by using the encoding techniques. Due to this compressing of data might cause either lossy or non – lossy data.

- **Non-Lossy** – usually make to use of statistical redundancy to constitute data without losing any information, it can be possible because of real-world data, this makes the process is more reversible.
- **Lossy** – After reconstruction from compressed data, if the original data cannot be obtained, this is referred lossy.

4. Discretization

The technique used for dividing continuous natural qualities into data with interval is set as data discretization. The implementation discretization is majorly due to the continuous features, which have a very low possibilities of being correlated with the target or destined variable. As an outcome, the results that are determined will be more difficult to understand. Groups that correspond to the target variable will be understood after discretizing the variable of the data.

5. Numerosity reduction

In the data analysis, numerosity reduction will take place for this below reason,

- Instead of a model, this would eliminate the need to store large datasets and reducing cost.
- As part of data pre-processing- noise reduction.
- Like regression model, data can describe as equation.
- The data can be represented as a model or equation like a regression model.

6. Attribute Subset Selection

Selection of attribute is very important stage because it might cause to high dimensional data which reflect the issues during the Machine learning algorithm. The factors which are add more value to model training, that's the attribute only want to consider during model training and unfitted attributes will be dropped.

2.4 EDA Process

Exploratory Data Analysis process is an important step in Data analytics, the process helps as to understand the depth of data, analysing and investigating the data to learn the different data characteristics.

Before performing the data modelling, it helps as to check whether data make sense, whether selected variable play the significant role for the output and help as to build to data modelling with the variable which as more contribute for the output.

It eliminates irregularities and unnecessary value from data likely outlier. EDA process generating summary statistics and to creating the various graph like histograms, scatter plot and box plot which help as to understand data better for the numerical data. EDA will also perform like data understanding, handling the missing data, handling the outliers, analysis the relationship between the data.

2.5 Model selection and building test model

The statistical and computation are combined to make the ML model, which help for business decision making problems. The ML model can be able to make prediction, in way of training model and apply the model with new data. After importing the set of the data for training, the algorithm which provided will extract the pattern and learn themselves from those data, then the trained model can be used to predict the unseen new dataset.

While building the model you might face some complexity like data “overfitting” and “underfitting”, “generalization error”, and “validation of model selection”.

For example, consider the two different machine learning algorithms, in the first algorithm we train data and applied in the model, the model shows the accuracy of 0.24% and then the real-world data have been as the input then it gives the 0.25% accuracy. Then respected the things in the second algorithm same as first, for the training data model result the accuracy of 0.95%, and for real world data shows the 0.27% accuracy.

It means accuracy of first model manifest that model is underfitting. The underfitted might causes the low dimensional, heavily regularized, and bad modelling assumptions. For second model the training data has shown 0.95%, but the real-world data has the bad accuracy of 0.27% that means model manifest overfitting, it might cause high dimensional, weakly regularized, and not enough modelling assumption.

Before applying the Machine learning model, too important to check, “What kind of machine learning should we use”, in case of the selecting the inappropriate model, it might end up losing time, energy, and processing cost. The model selection will be based on problem statement. Machine learning model will vary depend on the business problem and the dataset. The machine learning is classified into three types:

2.5.1 Supervised Learning

Supervised learning is its basic ML model will train under the supervisor which is worked by using the training data makes result as an output, and it may as work on input – output pairs principle. Basically, function runs by train the model with the training dataset as the label, and then apply the function with unknown data to make predictive factors. One of the major things want to be consider, test data should in labelled. The supervised learning will be suitable for the real-life problem which help us for decision making, make economical stable in way of predicting. The supervised learning commonly used for “Classification” and “Regression” problems. One of the real time examples, you can predict the weather for a particular day based on the some of the macroeconomic variable like humidity, precipitation, pressure values and wind speed and supervised learning algorithm is mainly used for the forecast trends for next month or next quarter for the different products. As mentioned above the supervised learning classified into two types.

Classification

In supervised learning first method will be a classification. The classification algorithm performs to identify the new observation of category based on the training data. The model will learn from the given training dataset or observation, then model ready to classifies the new observation of

unknown data into several groups or classes. The classes may call as the categories or target/labels variable. This method is used, when the problem statement looking for output as “yes” or “no”, “true” or “false” etc. In the classification algorithm, input variable will be x and output/target variable will be y .

$$y = f(x), \text{ where } y \text{ is categorical output.}$$

The main process of the algorithm is to recognize the category of a data, and to predict the result for the categorical data in real time.

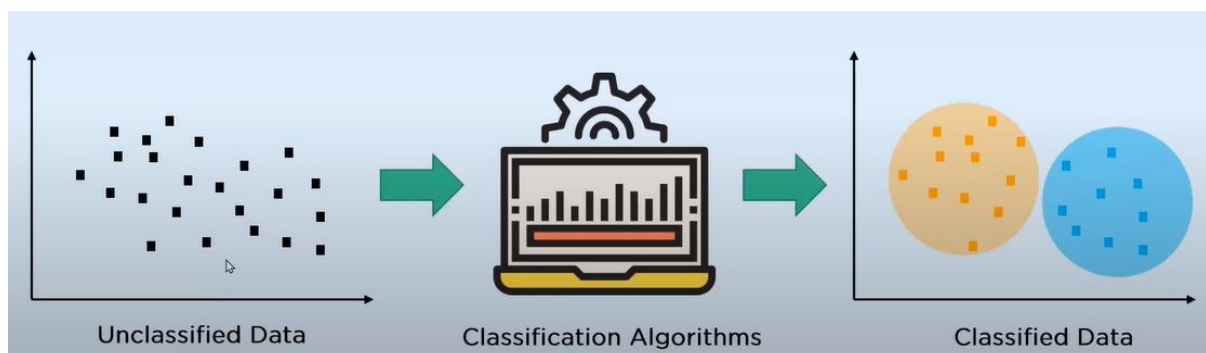


Figure 2- Classification algorithm

The Figure 2 diagram can help us to understand the classification algorithm. In left hand side the plotting shows that the data are unclassified. The unclassified data are applied in the classification algorithm, the model learn itself in way of split the data as train and test, then right hand side display that the unclassified data is separated into two classes based on it features. This algorithm applies the classification in a dataset is called as a classifier. “**Binary classifier**” and “**multi-class classifier**” are two types of classifiers which helps us to implement in data.

- **Binary classifier** – The given data which has the two types of features or the possible outcome as only two feature/ classes. For an example: yes or no, male or female, etc.
- **Multi-class classifier** – If the given data has the multi feature, and outcome feature also expect more. For an example: classification type based on the groups like type of music and so.

Basically, there are two types of learners in the classification problem:

- **Lazy learners** - it initially run with training data and wait for test data, takes the more time for prediction and less time for train data. K_NN algorithm is one of the examples for lazy learner.
- **Eager learners** – it opposite to lazy learner, it takes more time for train data and less time for prediction. Decision tree, naïve bayes are example for eager learner.

Classification algorithms are classified into further types:

- *Linear model*
 - Logistic Regression
 - Support Vector Machine
- *Non-linear model*
 - K – Nearest Neighbours
 - Kernel SVM
 - Naïve Bayes
 - Decision Tree Classification
 - Random Forest Classification

Logistic Regression

In machine learning the logistic regression is one of the most popular algorithms. Logistic regression is used for predicting the target variable of categorical data with help of independent variables. Therefore, the outcome of the dependent variable must be in categorical or discrete value. It may be either true or false, 0 or 1, yes or no and so on. The logistic regression the outcome also gives as the probabilistic values which lie between the 0 and 1.

The linear regression and logistic regression are much similar, except in the terms of application. The linear regression will implement to solve the regression problem, but the logistic regression used to solve the classification. The equation of the logistics regression:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

There are some of the types of the logistic regression analysis like:

- **Binary logistic regression** - In case the predicting, the dependent variable has only two characteristics, e.g., high, or low, yes or no.
- **Ordinal logistic regression** – The predicting the factors has the three or more categories they are natural ordering level, e.g., patient condition (stable, critical, good), survey (agree, disagree, neutral).
- **Nominal Logistic regression** – This is same as the ordinal logistic regression, but the characteristics are non as natural ordering, e.g., subject (math, history, art)
- **Poisson logistic regression** – In Poisson logistic regression has the three or more categories, and when looking on the characteristic, they have the number of time event occurs, e.g., 0,1, 2, ..., etc.

In logistics regression, “S” shaped logistics function has been fit instead of fitting the regression line, it is the significant ML algorithm, since using continuous and discrete datasets, it may generate probabilities and categorise new data. The sigmoid function is used to plot the predicted value to probabilities, the logistics regression value must be in range of 0 and 1, so that it capable to form the “S” curve is called as the Sigmoid function or logistic function and utilize the concept of the threshold value to define whether probability of either 0 or 1, the value above the threshold tends to 1 and below the threshold trends to 0.

Regression

The regression analysis helps as to find out the relationship between the dependent variable and independent variable, it's like predictive modelling technique. The dependent variable means as target variable. The regression majorly helps us to understand the value of dependent factor/variable is changing related to independent variables/factors.

The regression helps to identify the correlation between the variable and permit to predict the continuous output depends on one or more variable. It majorly used for forecasting, time series modelling and determining the strength of prediction. Regression models are classified into several type:

Simple linear regression

The simple linear regression is statistical method to identify the relationship between the dependent variable and the one independent variable, because of there is only one independent variable, so the model is termed as simple linear regression. In case considering the two or more independent factors then it's called as multi linear regression. The equation of simple linear regression,

$$y = \beta_0 + \beta_1 X + \varepsilon$$

where, y - dependent variable mean target variable

X – independent variable means explanatory variable

β_0 - Intercept

β_1 – Gradient

In the simple linear regression, focus to predict the Y_i using the single independent variable X_i , the equation is display Figure 3:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

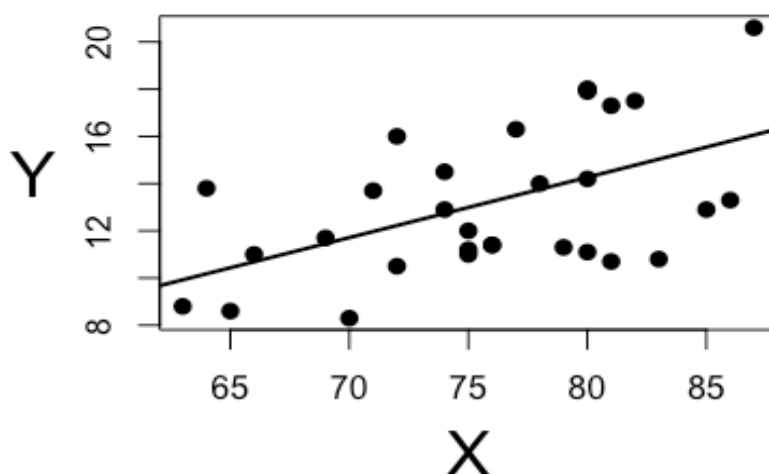


Figure 3- Simple linear regression

Multi linear regression

In multi linear regression, is statistical technique to identify the corresponding relationship between the target variable and explanatory variable. In multi linear regression, two or more independent factors will be considered to predict the continuous outcome of a variable.

Basically, it's like the make the prediction on one variable, based the information on the serval explanatory variables, explanatory variable is parameter, that help us to evacuate dependent variable

The multi linear regression model falls based on following assumption:

- If you might find any linear relationship between the target variable, and explanatory variable.
- There is no need of high correlation between explanatory variables.
- The observations of y_i are randomly selected from the population.
- In residuals, normal distribution should be normally distributed – mean of 0 and variance of σ

In the multi linear regression, we use the term R- square, it means determination of co-efficient, that used to calculate the variation of outcome, can be describe by the variation of the independent variable. In Multi linear regression R square make favour of increase the more predictors, even albeit predictors might be not related to target variables [13]. R – square value should be between 0 to 1, 0 is determine that the dependent variable cannot be predict by an independent variable, 1 indicates that dependent variable can be predicted, or independent variable perfectly related with the dependent variable.

The output of the multi linear regression, graph can be displayed as horizontally equation, or vertically in table form, the information from the multiple variables as independent factors creates the prediction accurately on the rate of effect have on the outcome variables [13].

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_p X_{ip} + \varepsilon_i$$

where, Y_i - dependent variable mean target variable

X_i – independent variable means explanatory variable

β_0 - Intercept

β_p – each explanatory variable slope coefficient

ε_i - model error term

Advantage of Multi linear regression

The various advantage of using the multi linear regression to analysis the data, the major two advantage are, first thing is capability to determine the relative effects of one or more independent variables to the benchmark value. The second advantage is ability to find the outliers. For an example, while reviewing the data of office management salaries, there are some of related variables like numbers of hours worked, budget mean average salary of each department, department size which are correlation with salaries, for scenario examined with multi linear regression, result displayed that one of manager have been overpaid.

Disadvantage of Multi linear regression:

The data utilised is generally the source of any disadvantages for multiple regression model. For an example, when analysing the prices of the home the community has the 100 houses, but dataset which used analysing only have data for 10 homes, if you move forward with incomplete data to analysis, it shows that the 70% percentage of homeowner are young parent which means school might shows the effect on sales prices of the home in the community. When analysing data without loss of any data mean dataset have the information of 100 homes, the result shows that only 7% school play the effect on the sale prices.

Polynomial Regression

Simple linear regression algorithm, perform only in the case of linear data, if in case of non – linear data linear regression can't be capable to plot the best-fit line, it fails in that condition, in below graph linear regression result shows that $y = 0_0 + 0_1x$, which means does not comes close the reality, polynomial regression implemented to overcome this problem. The polynomial regression which helps to find the curvilinear relation between the target variable and the explanatory variable.

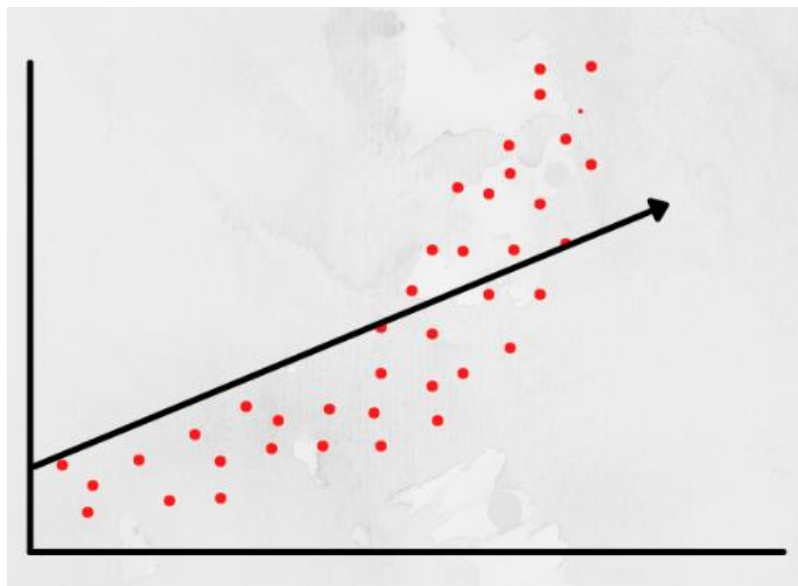


Figure 4- Polynomial regression algorithm

The polynomial regression is same as form of linear regression, it will perform to due to non-linear relation, so we are adding some of the polynomial terms to linear regression to convert into polynomial regression.

The equation of linear regression is converted to into polynomial regression like:

$$y = a_0 + a_1x_1 + a_2x^2 + \dots + a_nx^n$$

We are wisely applying the hyperparameter which is used for the degree of order and using the high degree of polynomial tires which helpful for overfit the data and smaller degrees, and the model tries to under, so want to find the optimum value of degree.

2.5.2 Unsurprised learning

The unlabelled are used to train the model, unlabelled means that there no fixed output variable. The model learns from the data discovers patterns and features in the data and returns the output. One of the examples, model that identify the image of the vehicles to classify if its car or bus, so the model identifies the parts of a vehicle such as length and width of the vehicle and so on, based on this feature the model classifies if the vehicle is bus or car. The unsupervised learning generally used for the "Clustering" and "Association" problems.

Clustering

The Clustering is grouping the data points and creating the partition based on their similarity, if the two things are similar in some ways, they often share other characteristic. The cluster is set of similar data point or set of point that are more like each other than to point in the other cluster. It classified under the unsupervised techniques, the key difference from the other machine learning techniques is that clustering does not have a response class.

After the grouping observation, manually need to look that the cluster and optionally associate meaning each other, the ultimate prediction, set of clusters themselves and this using this techniques performance only with data that in numeric form. There are many methods to perform prediction of clusters in way of calculating the similarity. The clusters are broken down into four methods:

- Centroid based clustering
- Connectivity based clustering
- Distribution based clustering
- Density based clustering

There are some of weakness in clustering. In most clustering method, we need to supply the number of clusters and use an approximation method to estimate the number of clusters called as elbow method and always remember that clustering algorithm are mostly sensitive to outlier.

Association

The association rule learning is come under the unsupervised techniques, which help to identify the dependency of data on another data item and plot accordingly, which make profitable. It makes an effect to find some associations/interesting relations among the factors in large dataset. It might happen based on the different rules to identify the interesting relation between factors among data. It forms the set of objects that occurs in conjunction on the dataset. Association rule helps marketing strategy more effective manner. For example, if people purchasing the bread and they also tends to buy the butter or jam. For this scenario typical applied the association rule as "Market Basket Analysis"

2.5.3 Reinforcement learning

Reinforcement learning drains a machine to do suitable accents and will increase rewards in a particular situation. The process will use an agent and an environment to produce action, rewards the agent has starts and end date, but there might be difference parts for reaching the end state like

maze, in this learning technique. In the learning there will not be any predefined target variable. The example for reinforcement learning is given, such as to train a machine that can identify the shape of an object given in the list with the help of determining the difference in the object with the help of the shape such as square, triangle, rectangle, or a circle in the given list. This shows the model tries to predict the shape of the objects.

The reinforcement learning is a kind of reward-based methods, so every task or every phase presented will be a reward received by the agent. When the given task is completed is not achieved correctly, which will lead to applied penalties. The Game industry widely use the model to build the game and used to train the machine-like robots to perform the human tasks.

There are two types of reinforcement learning are “positive” and “negative”, and wisely used two leaning model are “Markov Decision Process” and “Q learning”.

2.6 Time Series Analysis

Time series analysis means analysis the data which given in a set of order based on time, for example GDP data, oil price data and stock price data, are comes under the data of last one week or last year/ month, so basically these are manifest time series data. While plotting the time series data, x- axis there will be time in equal interval and y – axis have the magnitude of the data.

The important of time series analysis are listed below:

- **Business Forecasting** - time series analysis can use for any type of business sector because the pass defines what going to happen in future [3]. For example, trying to predict the price of stock market for tomorrow and retailers tries to know how many numbers of goods they are going to sell of next day and so on.
- **Understand past behaviour** – it helps us understand past data of every peak and dip. The business reason attached for trends up and down, so we can understand with will respect to time [3].
- **Plan future** – In the way of analysing the past and then we can be able to forecast the future using this time series analysis [3].
- **Evaluate current accomplishment** – for a scenario, we have predicted, that the salesman going to sale the hundred chocolates in a day, but salesman didn’t the achieve it.

Component of Time series analysis:

- **Trends** – The movement of trends relatively get higher or lower over a period, when the time series analysis shows the general pattern upwards, which means trends gets increases over along a time then it is called as uptrend as same time when the trends get lower or trends gets decrease then it called as downtrend, if there is no trend then it called as horizontal trend [3].
- **Seasonality**- which basically upwards or downward swings but this is quite different it’s a respecting pattern within a time.

- **Irregularity** – it basically called as noise so these are erratic in nature, or you can say unsystematic it is also called as residual, this basically happen for short duration, and it is nonrepeating.
- **Cyclic** – Cyclic basically repeating up and down movement and they don't have a fixed pattern, this will happen at any time.

Stationarity

In time series analysis the data need to be stationary and for any kind of statistical model applying. Commonly most the model work as by assuming the data as stationary, if the time series has as particular behaviour over the time there is very high probability that will follow in future and the in theoretically and formula which are related to stationary are more mature and easier to implement as compared to non – stationary series. There some of reason of non – stationary time series, the first thing is trend which basically wearing mean over the time, secondly having the seasonality so this is variation of a specific time frame. Stationarity has some of criteria:

- **Constant mean** – mean should be constant according to time series.
- **Constant Variance** – variance means distance for mean; it should be equal at different time interval.
- Autocovariance that does not depend on time.

There are two types of stationarity to check in python:

- **Rolling statistics** – In way of plotting the moving average or variance to check how varies with the time.
- **ADCF test** – Augmented dickey fuller test is another statistical test to check the stationarity, the test result will comprise of p vale which mean test statistic and some critical value.

2.6.1 ARIMA Model

ARIMA is one of best model to work with time series analysis. It is basically combination two model AR model and MA model. AR model stands for Auto regression, it uses own lagged values to shows a changing variable. MA model stands for Moving average, this model applies the lagged observations to the dependency in middle of an observation and residual error. Integration binds it together which indicated by "I", differentiate of raw observation which allows for the time series to make stationary. ARIMA model use as statistical analysis method, which used to better understand for the data or forecast the trends, the prediction is based on past value. For smoothing series of data, the model use lagged moving average. Statistical method will help to identify the lags numbers and identify the stationarity [12].

This model has some of parameters of:

- p - the number of autoregressive lag or lag order
- q - moving average size window or moving average order
- d – order of differentiation

3. MySQL

In general, a database is used to store data in a structured manner. We are using MySQL as a database to store the dataset, once the dataset is stored it can be accessed using the MySQL client.

MySQL is a database management system, using which we can add, process and access data stored. MySQL server is responsible for performing all the necessary actions such as hosting a database and processing the queries. MySQL in general is a relational database, these types of databases split the data into tables instead of putting everything together. This helps in accessing files quickly and efficiently. These tables consist of other characteristics such as the views, columns and rows which provides flexibility when trying to access data through a programming environment.

There are rules involved when using a MySQL, which are listed below.

- One to One
- One - to Many
- Unique
- Required or optional
- Pointers

MySQL is an open-source package, we can simply download the MySQL package which consists of both the MySQL server and the MySQL client. The MySQL server is capable of hosting multiple databases in a system, and the MySQL client is used to access the MySQL server and process the required queries.

Advantages of MySQL

- Fast
- Scalable
- Easy to use
- Reliable

As mentioned, MySQL server can run without any issues on any of the computers and laptops. In our project an AWS instance is created, which is a dedicated instance to run the MySQL server. In general, MySQL is also capable of running in cluster of machines that are essentially networked together.

Security in MySQL

Datasets need to be stored and kept secured as they contain crucial business-related information. Storing and accessing datasets on the internet can be a concern when it comes to security, with MySQL we can access data in a perfectly secured manner. The system is very secure when it comes to providing privileges using password, the verification is based on the host. With host-based verification we can have both flexibility and security simultaneously.

Other Features of MySQL

- Modules are independent and each module uses multiple layered server design
- With the help of kernel threads, the design is completely multithreaded
- Using MySQL, we can provide an interface for in house database

- Server is provided as a separate program
- Testing is done using different range of compilers

In our project, the dataset is stored in a database and the database is hosted in an AWS instance. The database is hosted in the AWS instance using the MySQL server. So, the dataset can now be accessed using MySQL client from any other system. The connection is secured which makes it easy and efficient to access and the connection is achieved using the TCP/IP socket.

There are different MySQL clients which are available to use namely, MySQL admin, MySQL dump and graphical user interface (GUI) known as MySQL Workbench.

4. AWS Instance

The AWS offers the plenty of services under the different domain like storage, compute, database, and management tools and so on. Out of this EC2 instance fall on compute. EC2 is a web service focus to make life easier for the developer for initiating a secure and resettable compute capacity in the cloud, with help of instance it is easy to scale up or scale down our infrastructure based on the business demand.

EC2 can be combined well with almost all the services required in Amazon, while comparing of all the outcomes and the best thing is cost depends on the usage. For accessing the AWS instance, we required to create the AWS account and then for compute capability we would require the EC2 instance.

- Initially, to create the AMI which mean Amazon machine image - there are software and application packages, which the users are required and needed to run our application.
- Secondly, the hardware should be set by the user, it is an instance type depending on the workload and the size of the hardware will be chosen accordingly.
- The instance should be configured and while this process the user should remember about certain terms like stop, terminal, and patches.

The above steps are about the OS volume and the hardware which required, after the selection add the additional storage to EC2 instance depending on the requirement, and then add the tags or configure tags, which makes easily to identify EC2 at the later use in way of giving the meaning name. As the next steps, we need to configure the firewall which also called a configure security group for the EC2 instance, here we can be able to allow or deny connection from the external source to access the instance. It may work in both ways from external and internally, the firewall blocks the connection by considering the IP address and port number. Finally, we will review all the configuration. Step are given below to create the EC2.

4.1 AMI - Selection

AMI is basically just like a templet, that's used to create a new instance or new VM, which based on the user requirements. It also contains the software information about OS, access permission, and volume. AMI fall in two types as "Predefined AMIs" and "Custom AMIs". The predefined AMI are called as Amazon provided AMI, this type AMIs don't need to create they are already predefined, we get this type of instance in AMI marketplace., but Custom AMI are customized depend on the business needs.

4.2 Choosing the instance

The instance types are basically hardware specification that require for the machine, which trying to build. The instance type is separated into five family.

- *Compute optimized* – it gives the lot of processing power
- *Memory optimized* – helps in memory cache
- *GPU optimized* – this type of application helps in gaming, which required the high-power graphic.
- *Storage optimized* – it mostly helps in storage servers
- *General purpose* – instance equally balanced, which shows balance between the virtual CPU, memory, storage, and the network performance.

4.3 Configure instance

The third thing, we need to configure instance, we have lot of option for purchasing

- Different billing option
- Assigning the IP address
- Kind of authentication
- Shutdown behaviour

In way of using the different billing option, user can pick the instance for the normal price. And there is also option like

- *Reserved instance* - where the user can pay for an upfront before a year/month by using this option user can pay less.
- *Spot instance* – bidding for those instances, whoever bids high they will get instance for that period.

4.4 Storage Adding

In this fourth step, we will add the storage to instance. For storage we have the bunch storage option like:

- *Pomeral storage* – these types of storage are free to use.
- *Elastic block storage* – this also called as EBS, which is paid.
- The free subscription users they get uses of 30 gigabit of SSD or magnetic storage for the hole year.

4.5 Adding Tags

The adding the tags, helpful to identify the VM machine in an environment, where we have seen the seven hundred or thousand instance.

4.6 Configuring Security Group

Security is the actual firewall that sits Infront of the EC2 instance and it protects that ec2 instance from unintended inbound and outbound traffic and can be able to find the tune at access the EC2 instance based on the port numbers and IP address through this we can access the instance.

4.7 Review

In this phase which helps as to review the configuration that we have made or the configuration which we change, to make sure the build configurations satisfy the requirement, after reviewing click the submit button to launch the EC2 instance. Before launch the amazon console, there will be option to create the key pair, the key pair is about two things means public and private key.

The private key is download and keep manually by the user's; the key is download in the format of dot PEM file. The public key is used by the amazon to confirm the identity of the users.

The next step will be to access the EC2 instance, for example we have launched the Linux instance, for accessing this instance we required the tool named as putty through windows. Putty tool didn't accept the dot PEM file we should convert the file format as PPK using generator, then upload the PPK file in putty and using the IP address, in way of doing this step we can be able to access the EC2 instance.

5. Variable selection

5.1 Inflation

The price of the good and services tend is increased over the time is known as Inflation, which is expressed by the inflation rate risen over the period. The good are tangible items buy in shop like food, clothes etc., The services are the intangible things, which provided by another person, e.g., transport. Usually, the government select the good and services that they feel broadly represent the economy bundles them together and take their current prices. The rate of inflation is calculated by comparing the prices today to the prices of the same good and services of the past. For example, a person purchasing the milk, butter, and bread it cost come round 5 pounds, a same person buys the same thing after a week, but the cost come round 7 pounds which shows that inflation is increased 2%.

The rising of inflation rate is mostly affected all of us, in way of reducing our daily live, majorly interest rate on loan, credit card lending, increases in social security benefits, and tax allowances etc., In low inflation the value of our money will be stable. While in low inflation scenario, the business can invest, consumer can spend or save depends on certainty, certainty is nothing, but the money holds its value. It helps us to balance the more stable the economy, and to handle the large swings in inflation and interest rate [8].

5.2 Interest rate

Usually, the interest rate gets increase to back down the inflation rate down. There two perceptions behind the interest rate, first is from the borrower, interest rate is cost of borrowing money, when the interest rate are high, we have pay more to borrow money, when it low we have pay less for borrow money. Second thing is from lender, it opposite to borrower when the interest rate is high lending money makes the more profitable and when the interest rate gets low, then it makes less profitable. In the world everyone borrows the money, which means government, business, and consumer and so on. There are some of big contributes that changes the interest rate, one is central bank in way of monetary policy and second is loanable funds in a supply and demand. The central bank usually lends the money directly to private financial sector or financial institution. The central bank is monitoring the monetary policy and try to stable the economy. Interest rate will be high to control the inflation and risk, and the interest will be low to make the consumer encourage to spending and increase the economy. They are different type of interest:

- Simple interest
- Component interest
- Fixed interest rate
- Variable interest rate
- Annual percentage rate (APR)
- Annual equivalent rate (AER)

5.3 Unemployment

Unemployment is measured using the unemployment rate. The unemployment rate is calculated by considering the number of workers, who cannot be able to find the job is divided by the total number of workers in the labor force. For an example, let see there is small country has the labor force of 10 million population, in that 1 million people is suffering to find the job, which shows that 10% of people are hard to find the job. Here, full time student, people below age of 16 years old, active-duty military, retired people and so on, this categorizes of people are should not be considered as the labor force. There are type of unemployment:

- *Frictional unemployment* – The frictional unemployment is also called as the seasonal unemployment. This type of unemployment are people, who are temporarily without work, in between the jobs like workers could have been fired or searching for the new role, this type of peoples is qualified and skilled but are currently they are unemployment.
- *Structure unemployment* – The worker who lost the job due to skilled to become obsolete due to some structure change in the labor force, because of this reason unfortunately these worker skills are non- transferable, these workers are required to learn the new skill to meet the requirement. This type of unemployment happens periodically.

5.4 Credit card Lending to household

Credit card helps us, when we use the credit card for purchases borrowing the money from that bank, then the bank will pay for store, and we will pay the bank back later. We are allowed to borrow up to certain amount at a time depends on credit limit. In case once we reached our credit limit after that we cannot borrow any more until we paid off some of the balance. The borrowing money applied some interest. When we spend the company on the credit card, there will interest on the amount we own in our balance. The interest rate will not be applied immediately, there will be grace period on credit card purchases. If we pay back the entire balance before the bills due, interest rate will be not applied.

5.5 Average weekly Income

In average weekly income in Great Britain, there might estimate 588 pounds in total pay, and 550 pounds in regular pay at NOV 2021. Our income has slightly increased while comparing with COVID pandemic. In our project we have used the average weekly income in Great Britain. Due the rising of inflation, real pay growth is got decrease.

5.6 GDP

The GDP is stands for Gross Domestic Product, which means the value of market for all finished goods and services which is considered within the country in a consent period. For baker purchase the eggs, butter, and things with used for making the cakes, we don't calculate this in our GDP

because these goods are not finished, it like intermediate goods. If the households buying the eggs, better and some other thing is calculated in the GDP because it finished goods.

The goods which used to make the other goods, but still it is consider as the finished goods, this is called as the capital goods. The GDP counts only the new production, basically like old house sales this year it doesn't add to the GDP, since the house did not produce this year. Only the sales of new house will be added in GDP. It only considers the goods and services produced within the country; it will not consider the imported things, but in case the product produce in our country and export to country that will be added in our GDP.

6. Implementation

Initial exacted the data for the analysis, from the “*Bank of England*” and “*Office for National statistics*” website. Our macroeconomics variables are inflation, GDP, unemployment, average weekly income, interest rate, credit card lending to households. The data imported in the “Jupyter notebook” “with using of panda’s library, file has been in CSV format. Data have been taken as the quarterly and some of the dataset didn’t consist of the quarterly data, so we take the mean method to convert into quarterly in way of using the “Numpy”.

```
#importing the library:

import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from scipy.stats import gmean

#importing the dataset:

GDP = pd.read_csv('GDP.csv', sep = ',')
unemployment = pd.read_csv('unemployment.csv', sep = ',')
Inflation = pd.read_csv('inflation.csv', sep = ',')
credit_card = pd.read_csv('Credit card.csv', sep = ',')
interest = pd.read_csv('CFMHSDG Bank of England Database.csv', sep=',')
Avg_wkly_earnings = pd.read_csv('Average weekly earnings.csv', sep=',')
```

Figure 6- Importing Data

Firstly, we have taken the GDP data. The dataset has the two columns as the “Title” and “GDP Quarter growth on percentage”. The data are in quarterly from 1955 quarter 2 to 2021 quarter 4. We have renamed the column name as Title as Year, and second column name as GDP, it makes easy to access the data. Then secondly, we are taking the unemployment data, the data which contain both monthly and years data from 1971 to 2021, for unemployment data initial drop the yearly rows, and filter the monthly rows to convert into quarterly format using the mean method. Then inflation, credit card lending dataset and interest rate have the data in monthly as same as unemployment, here we are using the mean method to convert the data for quarterly. Finally for the average weekly income dataset, have three columns as period, total pay, regular pay, for our analysis we are using the total pay column, and data are in monthly, here also we are applying the mean method to covert the data into quarterly without losing the any information.

After changing the data into quarterly, we have merged all six datasets into one single data frame, we are using the period for 2000 quarter 1 to 2021 quarter 4 for model building and concatenate the data using the year column. We have performed the data pre- processing method to our dataset, and checking the missing value, outlier and noisy data, there are no missing value and outlier in our

dataset. We have converted the datatype of each column into float because some of the columns are in object datatype it will makes the error during the plotting.

	Year	GDP	Unemployment	Inflation	Interest_rate	Income	Credit_card
0	2000 Q1	1.0	5.8	1.1	16.00	306.0	7787.0
1	2000 Q2	0.6	5.5	1.0	16.00	311.0	7949.0
2	2000 Q3	0.4	5.3	1.2	15.66	315.0	8241.0
3	2000 Q4	0.3	5.2	1.4	15.00	321.0	8344.0
4	2001 Q1	0.9	5.1	1.3	15.00	325.0	8437.0
...
83	2020 Q4	1.5	5.2	0.8	17.00	566.0	14235.0
84	2021 Q1	-1.2	4.9	0.9	18.00	568.0	13492.0
85	2021 Q2	5.6	4.7	2.1	17.33	575.0	16363.0
86	2021 Q3	1.0	4.3	2.7	17.00	582.0	17530.0
87	2021 Q4	1.0	4.1	4.4	17.00	592.0	18268.0

88 rows × 7 columns

Figure 7. Structured Data

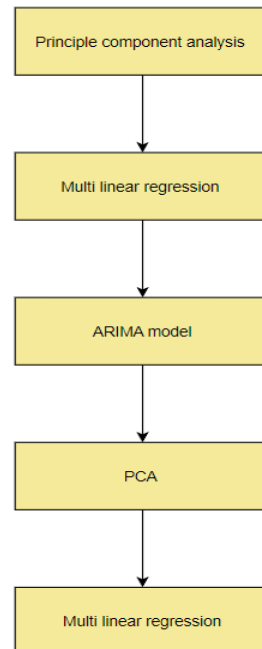


Figure 8- Project workflow

6.1 Plotting the Time series

We have plotted the line graph, to represent the time series data, because data analysis will not complete without visualization. It is simplest way to help us to view to understand the dependent

and independent variables change over the time. In below graph shows clear that, when the inflation gets increase then the interest rate gets changes in its nature. In credit card lending plot figure 10, we can be able to see that there is huge dip in 2020 because of covid.

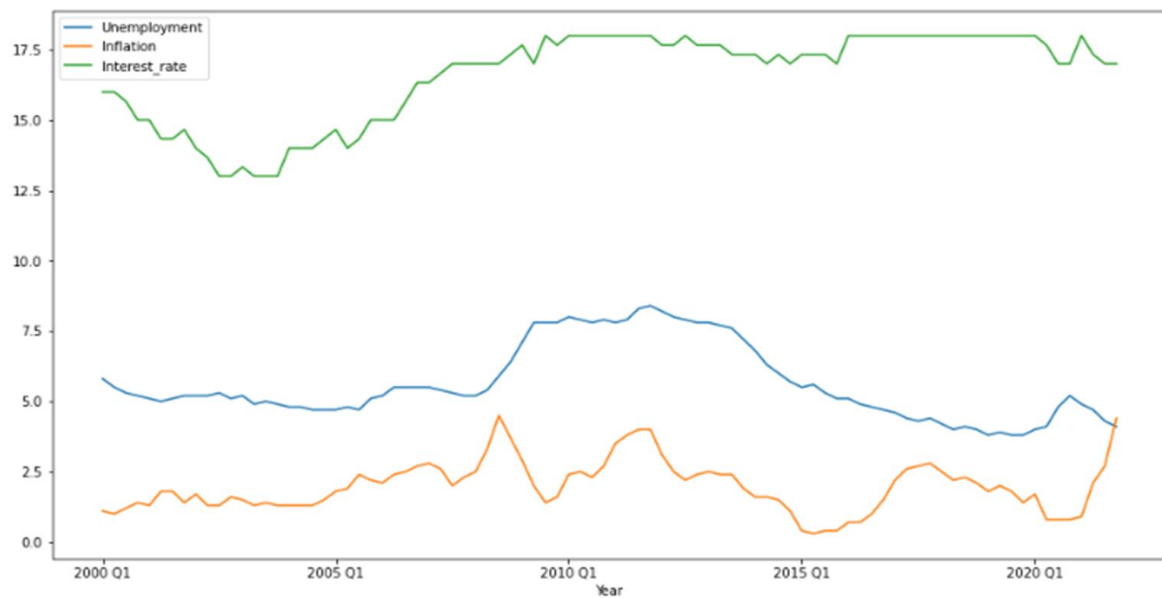


Figure 9- Time series for Unemployment, Inflation, Interest rate for credit card lending

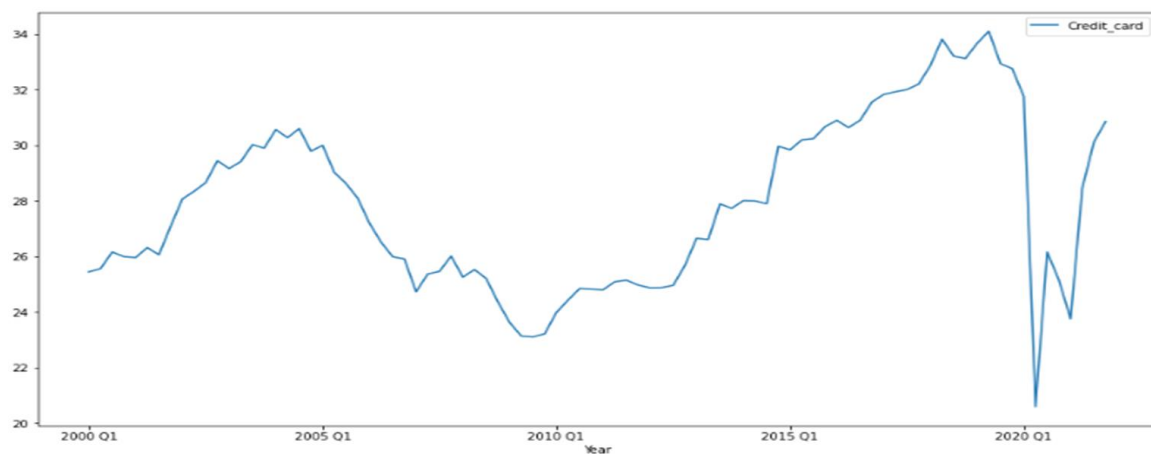


Figure 10- Figure 9- Time series for Credit card

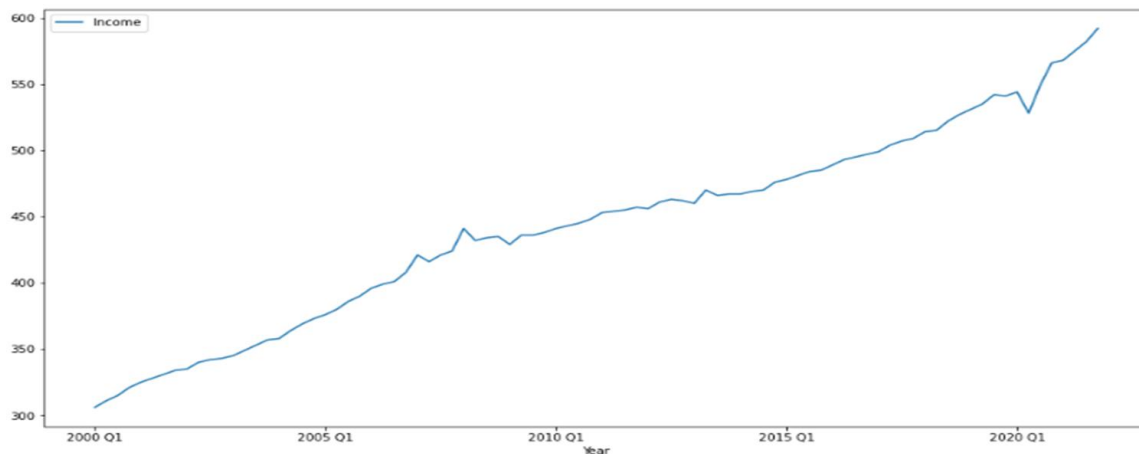


Figure 11- Time series for Income

6.2 Data and Correlational Matrix

In our dataset, we have applied the heat map to find the correlation matrix, which help us to identify how dependent variable correlated with explanatory variable and clarify how the macroeconomic variable related with each other [5]. The correlation matrix display for all possible variable in a data frame. This approach helps us to summarize the large dataset and makes visualization for our data. The correlation matrix is also commonly used in combination with other forms of statistical analysis.

And this helps us in multiple linear regression models, which contain the various explanatory variable, in this case correlation matrix determines the correlation efficient between the explanatory variables for the building model. As mentioned above, we are using the MLR model, which help us to known how the values of target value changes related to the independent variable.



Figure 12- Correlation matrix Heat map

6.3 Dimensionality Reduction

After finding the Correlation, we come to know that how interest rate is influence by our independent variable like interest rate as 0.77 correlation with income, and as the strong correlation with another variable too. So, we use principal component analysis method, with help of standard scaler which import from “sklearn. pre-processing”. Initially we convert the n – component as 2, after that we changed as n – component as 3, because while covert the component as 2 we got the total variance as 70% which we loss 30 % for information here, it might to affect the result. So, when reduce the component as 3, we got the total variance as 97%. While importing the variable to PCA, I have dropped the interest rate and GDP, because interest rate is our target variable and by using the backward method dropped the GDP. Here x1, x2 and x3 denoted the PCA 1, PCA 2, PCA 3 as three n-component.

	x1	x2	x3
0	-2.290723	-1.397850	-0.072189
1	-2.117262	-1.584688	-0.017531
2	-1.966178	-1.478124	0.211032
3	-1.872390	-1.324867	0.382370
4	-1.781410	-1.431322	0.340384
...
83	1.687653	-0.564946	-1.221957
84	1.604812	-0.624905	-0.993539
85	2.351145	0.406635	-0.072521
86	2.774956	0.799659	0.529225
87	3.001851	2.198792	1.836617

88 rows × 3 columns

Figure 13 – Principal component analysis for Normal Data

6.4 Multi linear Regression model Building

In multi linear regression, we require the independent variable and dependent variable. In our scenario, our target variable will be interest rate and independent variable will be output of PCA, which our three n – component variable. I have the build multi linear regression and make it as function, which can be able to access using X and Y variable as shown below graph.

Initial I have split my data into train and test, with size of 70% and 30%. Then we have created the instance for using the linear regression inbuild function to train our model, then fit the x_train and y_train in model.

```

# multi Linear Regression :
def multi_linear_regression(X,Y):
    X_train,X_test,y_train,y_test= train_test_split(X,Y, random_state=0)
    regression_model = LinearRegression()
    regression_model.fit(X_train, y_train)
    intercept = regression_model.intercept_
    coefficient = regression_model.coef_
    print('-'*100)
    print("The intercept for our model is {}".format(intercept))
    print('-'*100)
    r2 = regression_model.score(X_test, y_test)
    print('Predictive Accuracy(R-Squared):',r2)
    print('-'*100)
    for coef in zip(X.columns, regression_model.coef_):
        print("The coefficient for {} is {}".format(coef[0],coef[1]))
        print('-'*100)
    y_predict = regression_model.predict(X_test)
    pred_y_df = pd.DataFrame({'Actual value':y_test,'Predict value':y_predict, 'Difference':y_test-y_predict})
    print(pred_y_df)
    sns.regplot(x="Actual value", y="Predict value", data=pred_y_df);
    print('-'*100)
    from sklearn.metrics import mean_squared_error
    rmse=mean_squared_error(y_predict,y_test)
    print('MLR value:',rmse)

```

Figure 14- Multi linear regression

When all the explanatory variables are equal to zero, the intercept term is the value of the dependent variable. The estimated change in the target variable for a one unit change in that independent variable, held constant for the other explanatory variables, is the slope coefficient. Once the train model has be ready, we will fit the x_{test} to predict y_{test} . After fitting the prediction, we find R- square value, in way of using the python inbuild function `r2_score` as shown Figure 14. The R-square value represented to provide the measure of how good our model predicts, high r^2 value represent the good prediction, low r^2 value represent poor prediction basically accuracy of model prediction, for our model accuracy value displayed the 75%, which represented the good prediction. Normally the value above the 70% is say as good prediction. RMSE value, which means the root mean square error, which represent the variation between the test vs predict, we got the 0.66 as rmse value, the value arrange is depends on the scaler value.

The intercept for our model is 16.57804271750663

Predictive Accuracy(R-Squared): 0.7523701884364308

The Coefficient for x3 is -0.7060309004158339

The Coefficient for x1 is 0.6417779305089909

The Coefficient for x2 is 0.8319957087847302

	Actual value	Predict value	Difference
2	15.66	13.937405	1.722595
13	13.00	14.538330	-1.538330
53	17.66	18.107718	-0.447718
41	18.00	17.576667	0.423333
66	18.00	17.194852	0.805148
30	17.00	15.820302	1.179698
45	18.00	17.996699	0.003301
43	18.00	17.721933	0.278067
77	18.00	17.898826	0.101174
87	17.00	19.037241	-2.037241
7	14.66	14.297506	0.362494
26	15.66	15.693341	-0.033341
33	17.00	16.244736	0.755264
63	17.00	17.071747	-0.071747
8	14.00	14.418012	-0.418012
16	14.00	14.723982	-0.723982
24	15.00	15.482799	-0.482799
56	17.33	17.586635	-0.256635
76	18.00	17.698199	0.301801
42	18.00	17.565111	0.434889
22	14.33	15.184485	-0.854485
6	14.33	14.181232	0.148768

Figure 15- Outcome of MLR Model

We have performed the back testing for MLR model, the back testing which means start date of period will be same, but the end period will changed, like train the model using the 2000 quarter 1 to 2011 quarter 4 data and then predicting the 2012 all four quarters. Similarly, we have performed for different type of end period and checking the model efficiency, and our model performed pretty good in back testing and accuracy of the model is also good in most of the cases.

```

-----
The intercept for our model is 16.535878693245913
-----
Predictive Accuracy(R-Squared): 0.7656729941798157
-----
The Coefficient for x2 is 1.1717981846808935
-----
The Coefficient for x1 is 0.13322524917336784
-----
The Coefficient for x3 is 0.8445939584216847
-----

```

	Actual value	Predict value	Difference
2	15.66	13.948208	1.711800
13	13.00	14.330602	-1.330602
53	17.66	17.890248	-0.230248
41	18.00	17.565210	0.434790
66	18.00	16.924296	1.075704
30	17.00	16.137324	0.862676
45	18.00	18.001375	-0.001375
43	18.00	17.681702	0.318298
77	18.00	17.433940	0.566060
87	17.00	19.197565	-2.197565
7	14.66	14.271566	0.388434
26	15.66	15.902688	-0.242688
33	17.00	16.633408	0.366592
63	17.00	16.726915	0.273085
8	14.00	14.306305	-0.306305
16	14.00	14.415400	-0.415400
24	15.00	15.574591	-0.574591
56	17.33	17.312118	0.017882
76	18.00	17.305139	0.694861
42	18.00	17.511045	0.488955
22	14.33	15.215140	-0.885140
6	14.33	14.306937	0.023063
61	17.33	16.819822	0.510178
48	17.66	18.009833	-0.349833
79	18.00	17.454600	0.545400
54	17.33	17.724283	-0.394283
72	18.00	17.292756	0.707244
78	18.00	17.545159	0.454841
81	17.66	17.733928	-0.073928
3	15.00	14.075228	0.924772

```

-----
Predict the Interest_rate value 2012.1 : [18.01]
-----
Predict the Interest_rate value 2012.2 : [17.92]
-----
Predict the Interest_rate value 2012.3 : [17.86]
-----
Predict the Interest_rate value 2012.4 : [17.81]
-----

```

Figure 16- Back testing MLR outcome

6.5 ADF Analysis

In time series analysis, initially need to determine the number of differencing required, in case to make the time series stationary. The ADF test and KPSS test is perform the stationarity testing activity in AR model. In other word, it is called as statistical significance test. Here, we required to check the null hypothesis and alternative hypothesis, in way of using “adfuller” the build function in python with report the p- value. The p- value should be obtained less than the 0.05 as significance level to reject the hull hypothesis to make the evidence as series is stationary.

In order of analysing the ADF test it return the following as the outcome:

- Statistic test value
- P- value
- N_ lag value
- Critical value cut-offs

In below Figure 18, you can see that, we have tested the stationary for interest rate, result displayed that p – value is less than the 0.05, and as same side we have applied for all variable in our dataset and result shows that p – value is less than 0.05. So, all factor in our dataset is get stationary.

```
def ad_fuller(dataset):
    ad_fuller=adfuller(dataset.diff()[1:])
    print("ADF analysis\n")
    print("Statistic Test : " , ad_fuller[0].round(2))
    print("p-value : " , ad_fuller[1])
    print("# n_lags : " , ad_fuller[2])
    print("No of observation: " , ad_fuller[3])
    for key,value in ad_fuller[4].items():s
        print(f" critical value {key} : {value}")
```

Figure 17 – Augmented Dickey Fuller Test

```
ADF analysis

Statistic Test : -3.28
p-value : 0.015866047152376536
# n_lags : 3
No of observation: 83
critical value 1% : -3.5117123057187376
critical value 5% : -2.8970475206326833
critical value 10% : -2.5857126912469153
```

Figure 18 – Outcome of ADF test

6.6 ARIMA model Forecasting

We are using the ARIMA model for forecasting the data. In python to create the ARIMA model we have stats model library. In this model, p, d, and q are passing parameters while fitting the model. We have applied ARIMA model for every variable except the GDP in our dataset.

```
#ARIMA model building
from statsmodels.tsa.arima_model import ARIMA
Final = ARIMA(Var_Data['Inflation'], order=(1,0,1)).fit()
pred1= Final.predict(start = len(Var_Data)-6, end = (len(Var_Data)+8))
Var_Data['Inflation'].plot(legend= True,label = 'Inflation', figsize=(10,6))
pred1.plot(legend= True,label = 'Predicted Inflation')
AIRMA_Dataframe = pd.DataFrame(pred1,columns = ['Inflation_pred'])
AIRMA_Dataframe
```

Figure 19 – ARIMA Model

In above Figure 19, you can see that, we have called the ARIMA using the stats model and the fitted the model by passing the parameter of (1, 0, 1). This set the 1 as the lag value for autoregression, 1 uses as difference order which help the time series stationary, 0 as moving average and create the instance named as pred1. Then we have plotted the train data and predicted data. As same as we have followed for another factor and visualized below:

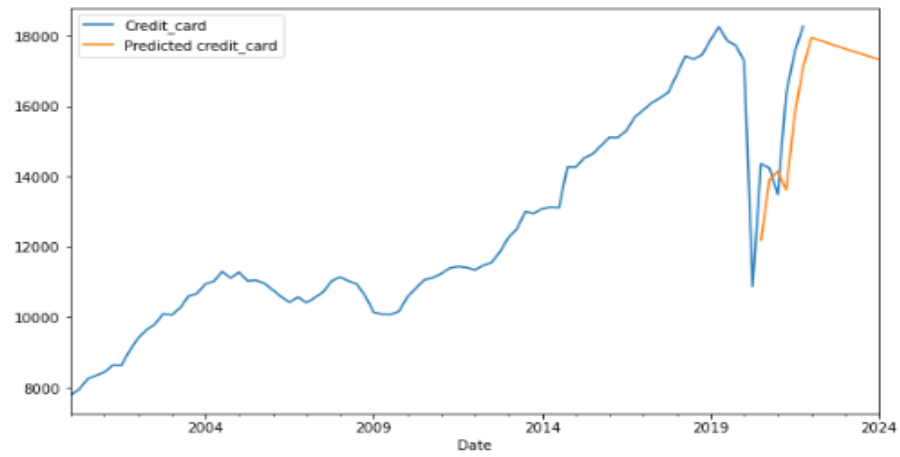


Figure 20- ARIMA model for Credit card

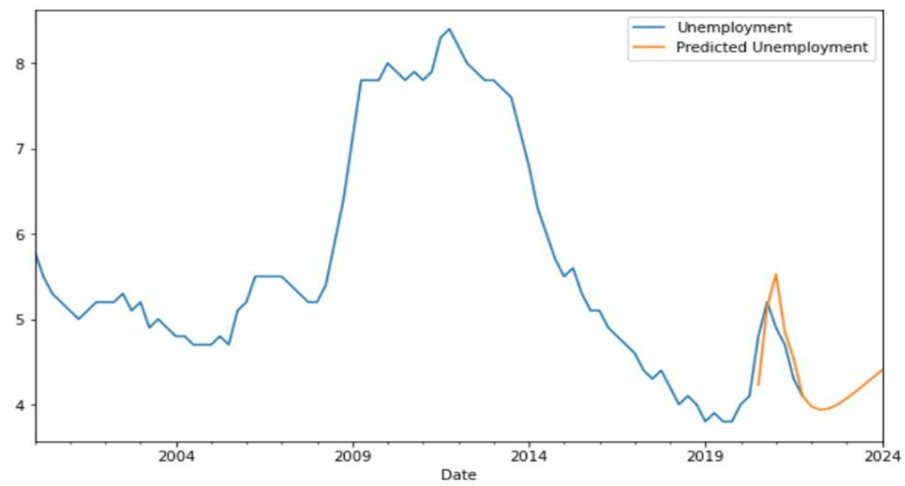


Figure 21- ARIMA model for unemployment

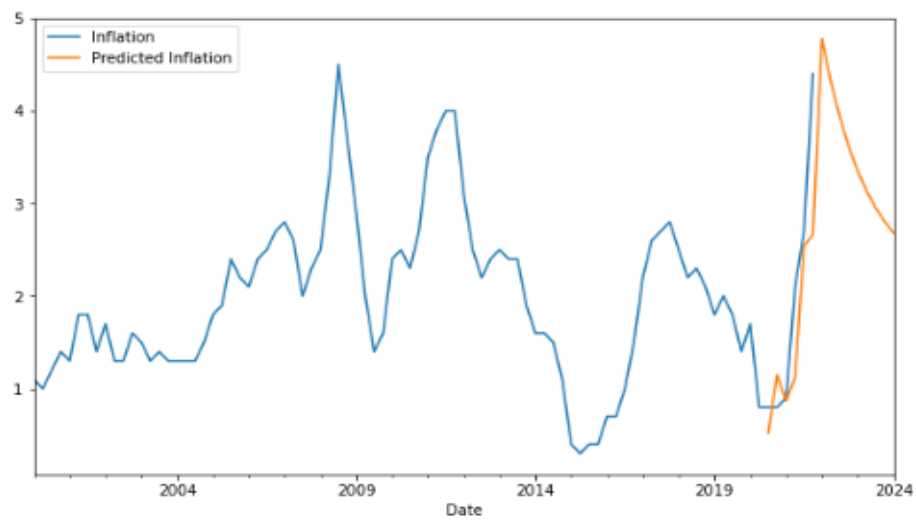


Figure 22- ARIMA model for Inflation

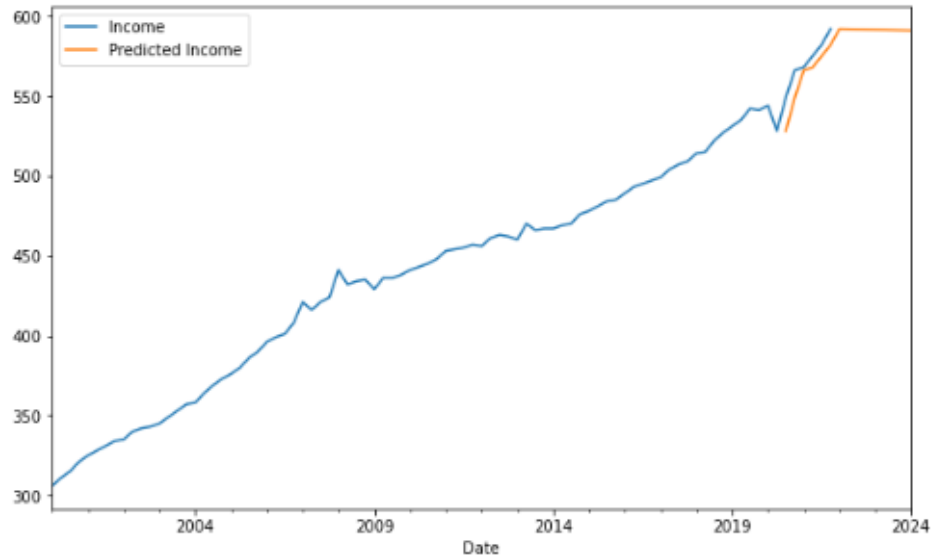


Figure 23- ARIMA model for Income

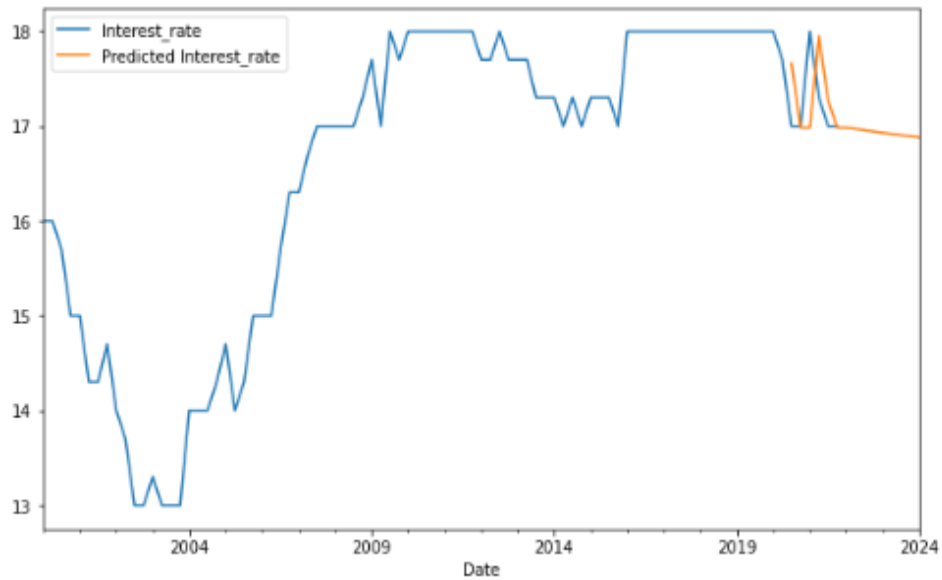


Figure 24- ARIMA model for Interest rate

In ARIMA predicted visualization shows that the prediction line has merged with train data, due to scalar value is small, variation may show the same difference in graph.

	Inflation_pred	Credit_card_pred	Unemployment_pred	Income_pred	Interest_rate_prediction
2020-09-30	0.525589	12188.603872	4.135714	527.932313	17.661873
2020-12-31	1.148113	13901.275238	5.112495	548.914365	16.984981
2021-03-31	0.868426	14145.167385	5.243315	565.899836	16.984981
2021-06-30	1.121671	13611.829453	4.759198	567.898126	17.951970
2021-09-30	2.538941	15754.707205	4.691878	574.892144	17.275078
2021-12-31	2.667715	17098.886719	4.154466	581.886161	16.984981
2022-03-31	4.778843	17946.736467	4.110972	591.877614	16.984981
2022-06-30	4.313712	17864.988183	4.145963	591.755333	16.970458
2022-09-30	3.929238	17784.535632	4.180004	591.633156	16.956414
2022-12-31	3.611435	17705.358275	4.213062	591.511084	16.942833
2023-03-31	3.348740	17627.435901	4.245184	591.389116	16.929701
2023-06-30	3.131598	17550.748618	4.276398	591.267252	16.917003
2023-09-30	2.952111	17475.276848	4.306728	591.145493	16.904724
2023-12-31	2.803747	17401.001327	4.336200	591.023837	16.892850
2024-03-31	2.681110	17327.903093	4.364838	590.902286	16.881368

Figure 25- Forecasted Data using ARIMA model

6.7 Forecasted data applied on PCA

As we seen brief explanation about the PCA in above topic, as same as that we applied the PCA for the predicted data, the data frame which displayed above. Initially, I have tried with two components, the variance shows 75%, reason of the losing the information I have moved forward with the three components, then the total variance shows 97%, with pretty goods, which helps to make better decision making [1]. Here x1 denotes PCA 1, x2 denotes PCA 2 and x3 denotes PCA 3.

	x1	x2	x3
0	3.805222	2.052745	-1.161760
1	2.632618	-1.578748	-0.586417
2	2.364866	-2.035199	0.074466
3	3.109056	0.998659	1.642974
4	0.851962	-0.170801	0.585810
5	-0.637734	0.322730	-0.435345
6	-2.013480	0.111436	0.136289
7	-1.786881	0.420242	0.079917
8	-1.593098	0.274988	0.034007
9	-1.426553	0.140788	-0.003278
10	-1.282633	0.016154	-0.033459
11	-1.157522	-0.100157	-0.057792
12	-1.048063	-0.209186	-0.077314
13	-0.951647	-0.311804	-0.092885
14	-0.866112	-0.408744	-0.105213

Figure 26 – PCA applied for forecasted data

6.8 MLR applied on Predicted PCA

Multi Linear Regression was applied for forecasted Interest rate for credit card lending and output from PCA Analysis from ARIMA model.

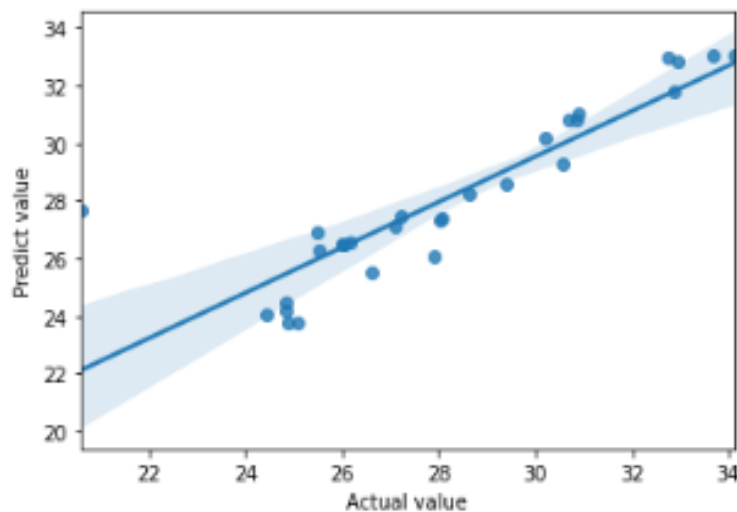


Figure 27 – MLR model outcome for predicted data PCA

Here, dependent variable is forecasted interest rate data and independent variable is Predicted data of PCA outcome. In above Figure 27, you can see that data plotted between the confidence interval, and the r square value is 78% and RMSE value is 0.68, it also good while comparing with above MLR which we did applied using the normal data [1]

7. Future Work

In the way of adding more independent variable, which correlated to interest's rate will help model to predict more accurately. Due the bank security purpose, more factors are not available in public, in case if bank use our model in by adding more factor which corelate with interest rate and explanatory might result in more accuracy rate. Also, we implemented the PCA method, which help to handle a large dataset. hence, adding the factor will not affect the model's efficiency.

8. Conclusion

In conclusion, by building the MLR model with accuracy of 76% we can predict the interest rate of credit card lending households. With the help of building the ARIMA model we can forecast the future interest rate, inflation, and other macro-economic variables. In this project, we were able to see the correlation between the interest rate and other explanatory variables. By using the PCA method, the dimension of the dataset has been reduced for analysis which has helped in handling the large datasets.

When applying the MLR model to the ARIMA predicted dataset, it was able provide more accuracy compared to the actual dataset. we have analysed the data obtained from the Bank of England, by using the interest rate as target variable the results obtained through analysis and model building can be helpful to banks across the country. We have seen how the interest rate can get affected by the inflation rate, not only inflation it can also affect the currency value. The project aim was to predict the rise in interest rate of credit card households which was implemented successfully with the help of the models used in this project. It can help individuals in seeing the future raises which can help in managing their debts. To conclude, the project has set out to achieve the aim of helping banks by providing a future insight on how the interest rate changes are going to be the upcoming years. This can help all the financial sectors to see the future ups and downs and act accordingly.

9. Bibliography

- [1] H. a. W. L. Abdi, "Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4),," pp. pp.433-459., 2010.
- [2] A. Budd, "The role and operations of the Bank of England Monetary Policy Committee. The Economic Journal, 108(451),," pp. pp.1783-1794., 1998.,.
- [3] G. J. G. R. G. a. L. G. Box, "Time series analysis: forecasting and control. John Wiley & Sons.,," 2015. .
- [4] C. a. S. E. Dziuban, "When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychological bulletin,," pp. 81(6),p.358., 1974. .
- [5] A. S. W. W. R. a. S. E. Famili, "Data preprocessing and intelligent data analysis. Intelligent data analysis, 1(1),," pp. pp.3-23., 1997..
- [6] J. Frost, "Guidelines for Removing and Handling Outliers in Data. [online] Statistics By Jim. Available at:," 2022.. [Online]. Available: <<https://statisticsbyjim.com/basics/remove-outliers/>>.
- [7] J. Hellerstein, "Quantitative data cleaning for large databases. United Nations Economic Commission for Europe (UNECE), 25,," pp. pp.1-42., 2008. .
- [8] T. L. K. S. L. a. S. S. Mehmood, "A review of variable selection methods in partial least squares regression. Chemometrics and intelligent laboratory systems,," pp. 118, pp.62-69., 2012..
- [9] M. S. A. M. S. A. C. a. P. R. Sadiku, "Data visualization. International Journal of Engineering Research And Advanced Technology (IJERAT), 2(12),," pp. pp.11-16., 2016.
- [10] K. M. M. M. T. a. Y. J. Ross, "Multi-source data analysis and evaluation of machine learning techniques for SQL injection detection. In Proceedings of the ACMSE 2018 Conference," pp. (pp. 1-8)., 2018, March. .
- [11] P. Z. W. Z. C. L. Y. a. Z. Z. Wang, "A dynamic programming-based approach for cloud instance type selection and optimisation. International Journal of Information Technology and Management, 19(4),," pp. pp.358-375., 2020..
- [12] G. a. A. Y. Weisang, "Vagaries of the Euro: an Introduction to ARIMA Modeling. Case Studies In Business, Industry And Government Statistics, 2(1),," pp. pp.45-55., 2008..
- [13] G. a. G. N. Uyanik, "A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106,," pp. pp.234-240., 2013..

10. Appendix

```
#importing the library:
```

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from scipy.stats import gmean
```

```
#importing the dataset:
```

```
GDP = pd.read_csv('GDP.csv', sep = ',')
unemployment = pd.read_csv('unemployment.csv', sep = ',')
Inflation = pd.read_csv('Wage inflection.csv', sep = ',')
credit_card = pd.read_csv('Credit card.csv', sep = ',')
Demo = pd.read_csv('DEMO.csv', sep=',')
interest = pd.read_csv('CFMHSDG Bank of England Database.csv', sep=',')
Avg_wkly_earnings = pd.read_csv('Average weekly earnings.csv', sep=',')
New = pd.read_csv('New_Dataframe.csv', sep=',')
ARIMA_PCA = pd.read_csv('PCA_AIRMA.csv', sep=',')
ARIMA_PCA
```

```
Avg_wkly_earnings
```

Importing library and dataset

```
#filtering the wage_inflection, change the column name and resetting the index:
```

```
Inflation = Inflation[Inflation['Title'].str.contains('20')]
Inflation = Inflation[Inflation['Title'].str.contains('Q')]
Inflation = Inflation.rename(columns={'CPIH ANNUAL RATE 00: ALL ITEMS 2015=100': 'Inflation'})
Inflation = Inflation.rename(columns={'Title': 'Year'})
Inflation = Inflation.reset_index(drop=True)
Inflation
```

```
#filtering the unemployment, changing the column and resetting the index:
```

```
unemployment = unemployment[unemployment['Title'].str.contains('20')]
unemployment = unemployment[unemployment['Title'].str.contains('Q')]
unemployment = unemployment.rename(columns={'Unemployment rate (aged 16 and over, seasonally adjusted)': 'Unemployment'})
unemployment = unemployment.rename(columns={'Title': 'Year'})
unemployment = unemployment.reset_index(drop=True)
#unemployment = unemployment['Unemployment'].astype("float")
unemployment
```

```
#filtering the unemployment, changing the column and resetting the index:
```

```
GDP = GDP[GDP['Title'].str.contains('20')]
GDP = GDP[GDP['Title'].str.contains('Q')]
GDP = GDP.rename(columns={'Gross Domestic Product: Quarter on Quarter growth: CVM SA %': 'GDP'})
GDP = GDP.rename(columns={'Title': 'Year'})
GDP = GDP.reset_index(drop=True)
GDP
```

```
##Merging the all dataset into one:
```

```
New_Dataframe = pd.merge(GDP, unemployment)
New_Dataframe = pd.merge(New_Dataframe, Inflation)
New_Dataframe
```

Filtering the required period for analysis

```
##structuring the interest dataset and append into Merged dataframe:
```

```
counter = 0
val = 0
Inter = []
intrset = []
for i in range(0,len(interest['Interest_rate'])-1,1):
    counter +=1;
    intrset.append(interest['Interest_rate'][i])
    if counter == 3:
        val = (gmean(intrset)).round(2)
        Inter.append(val)
        counter = 0
        val = 0
        intrset = []
New_Dataframe['Interest_rate']=Inter
New_Dataframe
```

```
#structuring the Avg_weekly earning dataset and append into Merged dataframe:
```

```
counter = 0
val = 0
avg_wkly = []
for i in range(len(Avg_wkly_earnings['Total pay'])):
    counter +=1;
    val += Avg_wkly_earnings['Total pay'][i]
    if counter == 3:
        val = (val/3).round()
        avg_wkly.append(val)
        counter = 0
        val = 0
New_Dataframe['Income']=avg_wkly
New_Dataframe
```

Structuring the columns

```
|: New_Dataframe = New_Dataframe.astype({'Unemployment': 'float', 'Inflation': 'float', 'GDP': 'float'})
```

Converting the datatype

```
# Correlation:
cormat = Head_Map_Data.corr()
cormat
```

```
import seaborn as sns
sns.heatmap(cormat);
```

```
# trying to find the correlation:

from pandas.plotting import scatter_matrix

plot_cols = [ "Interest_rate",
              "Credit_card",
              "Income",
              "GDP",
              "Unemployment",
              "Inflation"
            ]
## Create a scatter plot matrix --- a pair-wise scatter plots

def auto_pairs(plot_cols,New_Dataframe):
    fig = plt.figure(1, figsize=(12, 12))
    fig.clf()
    ax = fig.gca()
    scatter_matrix(New_Dataframe[plot_cols], diagonal='hist', ax = ax)
    plt.show()
    return('Done')
auto_pairs(plot_cols, New_Dataframe)
```

Correlation matrix using heat map

```

scaler = StandardScaler()
scaler.fit(New_Dataframe.drop(['Year', 'Interest_rate', 'GDP'], axis=1))

scaler_data = scaler.transform(New_Dataframe.drop(['Year', 'Interest_rate', 'GDP'], axis=1))
scaler_data

from sklearn.decomposition import PCA

pca = PCA(n_components = 3)

pca.fit(scaler_data)

x_pca = pca.transform(scaler_data)

x_pca

print(pca.explained_variance_ratio_)

second_PCA= pd.DataFrame(data = pca.fit_transform(x_pca), columns=['x1','x2','x3'])
second_PCA

```

Principal component analysis

```

# multi Linear Regression :

def multi_linear_regression(X,Y):
    X_train,X_test,y_train,y_test= train_test_split(X,Y, random_state=0)
    regression_model = LinearRegression()
    regression_model.fit(X_train, y_train)
    intercept = regression_model.intercept_
    coefficient = regression_model.coef_
    print('-'*100)
    print("The intercept for our model is {}".format(intercept))
    print('-'*100)
    r2 = regression_model.score(X_test, y_test)
    print('Predictive Accuracy(R-Squared):',r2)
    print('-'*100)
    for coef in zip(X.columns, regression_model.coef_):
        print("The Coefficient for {} is {}".format(coef[0],coef[1]))
        print('-'*100)
    y_predict = regression_model.predict(X_test)
    pred_y_df = pd.DataFrame({'Actual value':y_test,'Predict value':y_predict, 'Difference':y_test-y_predict})
    print(pred_y_df)
    sns.regplot(x="Actual value", y="Predict value", data=pred_y_df);
    print('-'*100)
    from sklearn.metrics import mean_squared_error
    rmse=mean_squared_error(y_predict,y_test)
    print('MLR value:',rmse)

```

Multi linear regression

```

# Back testing:

X_main= second_PCA[{'x1','x2','x3','Years'}]
Y_main= second_PCA[{'Interest_rate','Years'}]

def X_Y(X,Y,my_list):
    for i in my_list:
        X = X_main[X_main['Years'] <= i][{'x1','x2','x3'}]
        Y = Y_main[Y_main['Years'] <=i][{'Interest_rate'}]
        multi_linear_regression(X,Y)

my_list =[2011]
X_Y(X,Y,my_list)

```

Back testing

```
def ad_fuller(dataset):
    ad_fuller=adfuller(dataset.diff()[1:])
    print("ADF analysis\n")
    print("Statistic Test : " , ad_fuller[0].round(2))
    print("p-value : " , ad_fuller[1])
    print("# n_lags : " , ad_fuller[2])
    print("No of observation: " , ad_fuller[3])
    for key,value in ad_fuller[4].items():s
        print(f" critical value {key} : {value}")
```

ADF analysis

```
#ARIMA model building
from statsmodels.tsa.arima_model import ARIMA
Final = ARIMA(Var_Data['Inflation'], order=(1,0,1)).fit()
pred1= Final.predict(start = len(Var_Data)-6, end = (len(Var_Data)+8))
Var_Data['Inflation'].plot(legend= True,label = 'Inflation', figsize=(10,6))
pred1.plot(legend= True,label = 'Predicted Inflation')
AIRMA_Dataframe = pd.DataFrame(pred1,columns = ['Inflation_pred'])
AIRMA_Dataframe
```

```
#ARIMA model building
from statsmodels.tsa.arima_model import ARIMA
Final = ARIMA(Var_Data['Credit_card'], order=(1,0,1)).fit()
pred1= Final.predict(start = len(Var_Data)-6, end = (len(Var_Data)+8))
Var_Data['Credit_card'].plot(legend= True,label = 'Credit_card', figsize=(10,6))
pred1.plot(legend= True,label = 'Predicted credit_card')
AIRMA_Dataframe['Credit_card_pred']=pred1
AIRMA_Dataframe
```

```
#ARIMA model building
from statsmodels.tsa.arima_model import ARIMA
Final = ARIMA(Var_Data['Unemployment'], order=(1,0,1)).fit()
pred1= Final.predict(start = len(Var_Data)-6, end = (len(Var_Data)+8))
Var_Data['Unemployment'].plot(legend= True,label = 'unemployment', figsize=(10,6))
pred1.plot(legend= True,label = 'Predicted Unemployment')
AIRMA_Dataframe['Unemployment_pred']=pred1
AIRMA_Dataframe
```


Applying the ARIMA model to individual variable

```
#Predicted data applied in MLR module
X= ARIMA_PCA[{'x1','x2','x3'}]
Y= ARIMA_PCA['Interest_rate_prediction']
multi_linear_regression(X,Y)
```

MLR model applied on Predicated data

LinkedIn

Post-1




Viswanath Durairaj • You
Aspiring Graduate Data analyst-- seeking for full time opportunity | Student ...
3w • 🌐

...





I am excited to post that I have been working with multinational bank in analysing the Customer Behaviour when interest changes by using the various macroeconomic variables. This opportunity came across through [University of Leicester](#). The main goal of the project is to do the predictive analysis and Time series forecasting with respect to various macro factors like Interest rate, Average weekly Income, GDP etc. by using various machine learning algorithms and Time series models.


I'm really thankful to [Jeremy Levesley](#), academic supervisor [Alexander Gorban](#), program director Andrew Morozov(Dr.), for giving me the opportunity and my fellow teammates [Prathibha Mandhalapu Ramamoorthy](#) and [Rangaswamy Rachamadugu](#)

[#algorithms](#) [#dataanalysis](#) [#machinelearning](#) [#timeseries](#) [#project](#)
[#predictiveanalytics](#) [#algorithms](#)


 Rakunanthan Krishnasamy Subramanian and 33 others

1 comment

 Like  Comment  Share  Send

 2,606 views of your post in the feed

Post-2



Viswanath Durairaj • You
Aspiring Graduate Data analyst-- seeking for full time opportunity | Student ...
3w • 🌐


...

This post is to update my progress in the project so far, it has been an adventure and fun working alongside my teammates. As mentioned in my previous post I have been working on analysing the Customer Behaviour when interest rate changes by using the various macroeconomic variables. I'm working on building a regression model that would have target variables as Interest rate. We are trying to reduce the dimension of the data by using the Principal Component Analysis method. With the available data I have built a regression model which can predict the interest rates. For the model built we have performed back testing and stress testing, and the results were promising. We have also performed time series forecasting using the ARIMA Model. This work will help banking customers to predict the future interest rates as it increases. Working on this project has given me exposure to skills and knowledge which will help me in building more advanced data science models.

The experience has been exciting so far, I'm really looking forward to ending this project on a positive note.


I'm really thankful to [Jeremy Levesley](#), academic supervisor [Alexander Gorban](#), program director Andrew Morozov(Dr.), for giving me the opportunity, constant support and encouragement throughout the journey, and my fellow teammates [Prathibha Mandhalapu Ramamoorthy](#) and [Rangaswamy Rachamadugu](#)

[#datascience](#) [#experience](#) [#banking](#) [#testing](#) [#dataanalysis](#) [#predictivemodeling](#)
[#project](#) [#uol](#) [#backtesting](#)

 Rakunanthan Krishnasamy Subramanian and 21 others

2 comments

Post 3



Viswanath Durairaj • You

Aspiring Graduate Data analyst-- seeking for full time opportunity | Student ...

2w •

Project update part 3


I'm really thankful to [Jeremy Levesley](#), academic supervisor [Alexander Gorban](#), program director Andrew Morozov(Dr.), for giving me the opportunity, constant support and encouragement throughout the journey. and my fellow teammates [Prathibha Mandhalapu Ramamoorthy](#) and [Rangaswamy Rachamadugu](#)

#opportunity2022 #dataanalysis #datascience #projects #businessintelligence #exploratorydataanalysis #timeseriesanalysis

UNDERSTANDING CUSTOMER BEHAVIOUR UNDER INTEREST RATE CHANGES

Project update

Data Segmentation
Algorithms
Predictive Model
Time Series Forecasting



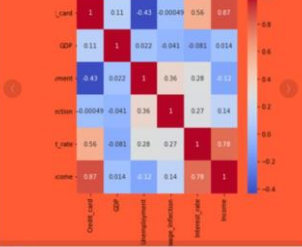
The aim objective of the project is to predicate the quarterly interest rate changes using quarterly data of macro-economic variables like gross domestic product, average weekly income, credit card spending, unemployment.

The collecting data for this project was the major challenge. Quarterly data from 2000 to 2021 was collected for the study, some of the variable didn't have the quarterly data like credit card and average weekly income, so we have converted the monthly data into quarterly data.

EDA Process

EDA (Exploratory Data Analysis) process helps us to understand more about the data and identify trends in data. Then we applied the PCA (Principal Component Analysis) on the macro-economic variables to reduce the dimensionality of data and reduce the n-component as two and got total variance as 80.49%.

Correlation



Database setup

We are using AWS RDS (Relational data schema) from Amazon to create an instance, so that we can access data from any system using instance Id and details.

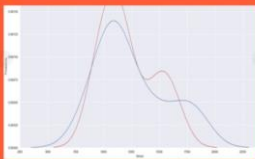
We Use MySQL as our database server and AWS instance as schema to make remote connection possible and use although server-side system is shutdown.


Model build

We have builded model for prediction and forecasting:

- ARIMA, Multiple Linear Regression, VAR, SARIMA.

Prediction





Viswanath Durairaj
Aspiring Data Analyst
Student at University of Leicester