Project

Banking dataset Analysis and Prediction

Supervisor Name: Savita Seharawat
Submitted by: Manmeet Kaur (0776137)

Table of Contents

Final Report

Banking Dataset- Analysis and prediction

Abstract:

In today's world, where huge amount of data is generated in every field of day-to-day activities, banking sector is one of them. As an outcome of work, various machine learning concept are studied with respect to Bank marketing data classification. Banking is a provision of the services by bank to an individual customer. The dataset is originally collected from UCI Machine learning repository and Kaggle website. The data is related to bank marketing campaigns of banking institution based on phone call. In this work, Python is used as a coding language and Machine learning concept is used as statistical learning for data analysis. The main reason of using machine learning is to build a predictive model to produce the better prediction. The outcome of the result is analysed with supervised Random Forest algorithm for classification purpose. The Customer bank dataset is used for term deposit prediction. This dataset is publicly available at UCI machine learning repository. It contains customer information. This dataset contains 11163 records and 17 attribute. The dataset has two types of prediction either Yes or No. There are 16 input feature and 1 output. The facts are associated with the direct advertising and marketing campaigns of a Portuguese banking institution. The marketing campaigns have been primarily based on telephone calls. Often, a couple of contact to the identical client was required, in order to access if the product would be or now not subscribed by means of the customer or no longer. Term deposits are a major source of income for a bank. A term deposit is a coin's investment held at a financial organization. Your money is invested for an agreed rate of interest over a hard and fast amount of time, or time period.

The financial institution has numerous outreach plans to promote period deposits to their clients together with email marketing, advertisements, telephonic advertising and marketing, and digital advertising and marketing. Telephonic advertising campaigns continue to be one of the simplest ways to reach out to human beings. However, they require big funding as massive call centres are employed to really execute those campaigns. Hence, it's far important to become aware of the customers maximum probable to convert beforehand so they can be mainly target via calls. The data is related to direct marketing campaigns (phone calls) of a Portuguese banking organization. The category purpose is to predict if the patron will join a term deposit.

Goal: The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Theme:

We use theme like classification and regression, data visualization, data manipulation, importing data.

Keywords:

Banking, Descriptive analysis, predictive analysis, ggplot, logistic regression, train and test split.

Research questions:

Question **1**. To cumulate that how many customers taking loan according to their education level?

Question2. How many customers taking house loan from the bank?

Question 3. Use Box Plot to compare the age of customers for the top 5 of the most common employment forms

Question 4. To calculate which month customer deposit maximum amount of money?

Tools: -

All formulation and data visualization will be done in Python.

GitHub account link: -

Manmeetkaur137 (github.com)

INTRODUCTION

Bank is a financial institution, which provide various service to the customer which perform deposit and providing a loan at an interest rate to the various customer. Banks store massive amount of information about their customer to improve the banking strategies and to maintain good relationship between the customers. Customers are the main asset of the bank. Usually, the selected customer is contacted directly through mail, email, personal contact, telephone cellular or any other contact to advertise the new service this kind of marketing called direct marketing. The objective of marketing in banking is to attract the new customers. The collected data from UCI machine learning repository, is related to bank marketing campaign of banking institution the classification goal is to predict if the customer will subscribe the term deposit. The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be subscribed ('yes') or not ('no') subscribed. The test.csv which is the test data that consists of 11163 records and 16 features without the target feature. The dataset contains train and test data Features of educate information are Input variables(financial institution client facts): Age , job, marital, schooling, default, balance, housing, mortgage, contact, day, month, length, campaign, pdays, preceding, poutcome. Output variable (preferred target): y-has the purchaser subscribed a term deposit? (Binary: 'sure’, ‘no') and the test information have already been pre-processed. In this

paintings Python used as a programming language, high-stage, interpreter, and considerable well-known library are freely to be had sources for all predominant platform from the Python internet site and Machine gaining knowledge of method for records analysis technique and automates analytical constructing model to predict the accuracy of the financial institution patron records. Where every example in a dataset is defined through a fixed of attributes and category algorithm used along with Naive bayes classifier set of rules gave the first-rate performance degree accuracy of the statistics. The financial institution ought to target the ability patron who have spent extensive quantity of time responding the bank calls. The most important item of this work is to find a way to use system gaining knowledge of method, for evaluation and making the prediction using existing dataset in banking marketing for developing effective selection-making understanding and to construct a system learning version the use of class set of rules to expect the accuracy of the facts.

Objective:

The main objective of building the model is to describe whether the customer has opted for term deposit. The bank should target the potential customer with considerable amount of time responding to the phone calls. The work implemented resulted in measuring accuracy, precision, recall and F1 score, towards term deposit prediction.

What is a Term Deposit?

A Term deposit is a deposit that a bank or a monetary group offers with a fixed price (often higher than just commencing deposit account) in which your cash can be back at a particular adulthood time. For more records on the subject of Term Deposits please click on in this hyperlink from Investopedia. Term deposits are a major source of earnings for a financial institution. A time period deposit is a coin’s investment held at a financial group. Your cash is invested for an agreed price of hobby over a hard and fast amount of time, or term. The financial

institution has numerous outreach plans to promote term deposits to their clients which includes email marketing, commercials, telephonic advertising and marketing, and virtual advertising

Literature Review:

The Customer financial institution dataset is used for term deposit prediction. This dataset is publicly available at UCI system studying repository. It consists of patron information. The dataset has types of prediction either Yes or No. There are sixteen enter function and 1 output. Attribute in the dataset with data type and description is following: **Bank client data:** [1] - age (numeric), [2] - job : type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services"), [3] - marital : marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed), [4] - education (categorical: "unknown", "secondary", "primary", "tertiary"), [5] - default: has credit in default? (Binary: "yes", "no"), [6] - balance: average yearly balance, in euros (numeric), [7] - housing: has housing loan? (Binary: "yes", "no"), [8] - loan: has personal loan? (Binary: "yes", "no"). **Related with the last contact of the current campaign: [**9] - contact: contact communication type (categorical: "unknown", "telephone", "cellular"), [10] - day: last contact day of the month (numeric), [11] - month: last contact month of year (categorical: "Jan", "feb", "mar", …, "nov", "dec"), [12] - duration: last contact duration, in seconds (numeric). **Other attributes: [**13] - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact), [14] - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted), [15] - previous: number of contacts performed before this campaign and for this client (numeric), [16] - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success"). **Output variable** (desired target) : [17] - deposit - has the client subscribed a term deposit? (Binary: "yes", "no"). The device gaining

knowledge of techniques for evaluation and making prediction the usage of current statistics in banking marketing. The fulfilment charge of banking advertising depends on the result and decision in order to make greater correct prediction statistical device and techniques are used. A specific level for facts analysis and to find, how they can be used together in a manner converting uncooked information to effective choice-making expertise and building the predictive version in this work used selection tree algorithm will help to are expecting the client will subscribe the term deposit. Esslemont et all. Discussed all financial institution advertising marketing campaign are depending on client big statistics, the size of facts source is not possible for human to analyst to give you fulfilling information with the intention to help in choice making method. Data mining model are helping in the performance of the campaigns, in this work used most important data mining technique Naive bayes, logistic regression, and decision tree, the purpose is increasing the campaign effectiveness and identifying the characteristics that effect a success. A data driven approach was suggested to predict the success of bank telemarketing used data mining approach to predict the success telemarketing call for term deposits, data related to Portuguese retail bank it includes the effect of financial crisis, analysed large set of features related to bank client, social and economic characteristics, and product. In the modelling phase a semi-automatic feature had selected, performed with the data prior, and reduce set of the feature. Compare data mining model super vector machine, decision tree, logistic regression, and neural network, using two metrics, the four models were tested, and neural network present the best result, decision tree is a knowledge extraction method were applied to neural network to predict the several key attribute. Finally, the selected model as credible and valuable for telemarketing campaign. Bank direct marketing is an interactive process, for building the good relationship among customers, to study the customer characteristics and behaviour use an effective multi-channel communication. Apart from income growth, which may also increase patron fantastic response, the intention of financial

institution advertising and marketing is to growth the client reaction of direct marketing campaign. Customer profiling in, using category approach for financial institution telemarketing, facts mining processes began by way of many agencies to restore the purchaser profiling. Decision tree, random wooded area, and Naive Bayes have been used, for predicting the client profiles and growing the telemarketing sales type is useful for measured accuracy percentage, precision, and consider quotes. Before evaluating the classifiers pre-processing and normalization were performed for conducting the experiments and evaluation system RapidMiner tool was used. Finally, end result show that selection tree is the exceptional classifier for predicting the patron profile and behaviour.

The information that banks receive from their clients, traders, companions, and contractors is dynamic and can be used for distinctive functions, relying on which parameters are used to examine them. Basically, the scope of AI for banking can be grouped into five huge businesses.

It needs to be pressured that due to internal opposition and gift-day financial catastrophe, there are big pressures for European banks to increase a economic asset. To treatment this

problem, one followed technique is provided appealing lengthy-time period deposit programs with suitable interest costs, specifically through the usage of directed advertising and marketing and marketing campaigns. Also, the same drivers are urgent for a reduction in costs and time. Thus, there may be a need for an development in efficiency: lesser contacts must be performed, but an approximately amount of successes (customers subscribing the deposit) ought to be saved.

It is applicable to refer in short to the preceding research and research in the related regions of the problem to find out and to top off the research gaps. The following are the same

studies performed through the eminent authors and practitioners on the location of service great of banks.

(Dhandabani, 2010) Examined the character of linkage among carrier first-rate and clients loyalty in Indian retail banking. Study used confirmatory component analysis to discover the carrier great dimension. The effects dimensions are reliability, Responsiveness, Knowledge and restoration, and Tangibles. The service fine dimensions result in patron pride and the patron 'pleasure ends in consumer's loyalty. The structure equation model reveals that there's no significant direct linkage among provider excellent and customers loyalty. At the equal time, the provider satisfactory has a significant oblique effect on purchaser's loyalty particularly through purchaser's satisfaction.

(Maya Basant Lohani, 2012) Examined on provider exceptional in selected banks and measured in 5 dimensions through the use of SERVQUAL scale advanced via Parasuraman et al (1998 and found out that there exists a small perceptual difference regarding standard provider quality with respective banks. The study of located that bank have more awareness at the tangible element like a computerization, physical centers, and many others. To draw the customers.

(Hertz 113-114) Knowledge of industries in which the financial institution's customers operate. Often a financial institution's mortgage portfolio might be focused on mainly specialised industries along with real assets, delivery, and natural resources. Evaluating the nature of those portfolios can additionally require a know-how of the corporation and reporting practices of these industries. The shape of the banking enterprise and the Bank's role and reputation in the marketplace. For instance, if the financial group has a terrible rating, it is able to now not have got admission to better extraordinary loans, thereby taking extra credit score rating hazard.

(Jain, 2012) in their study "Customers Perception on Service Quality in Banking Sector: with Special references to Indian personal banks in Moradabad place" try to learn and apprehend the customer belief concerning carrier quality and to examine and apprehend the one-of-a-kind dimension of carrier pleasant in banks.

J. Joshua Selvakumar (2010) research the effect of carrier Quality on patron satisfaction in public region and personal region banks. The look at examines the impact of provider high-quality determinants at the diploma the perceived and real carrier high-quality, customer delight can be extremely progressed.

Dr. Manasa Nagabhushanam (2010) conducted a research study on service quality of banks in India. The study encompasses the service quality of all the banks i.e., public sector, private sector, and foreign banks and measures the attributes on SERVQUAL scale. The study was conducted to analyse the expected and perceived gap among customers and bankers.

Ms. Nisha Malik and Mr. Chand Prakash Saini (2011) studied Private sector banks quality and customers satisfactory by conducting an empirical study of two Private sector Banks. The aim of proposed study was to find out perception of HDFC and ICICI bank customers regarding to the service quality parameter and gap analysis of expected and acknowledged quality parameters. Ans also revels the relationship between psychographics factors and satisfaction levels of rural and urban customers.

DESIGN AND METHODOLOGY:

Firstly, our dataset is uncleaned like there are some duplicate values. First, we will clean our dataset to perform further operations. In methodology part we will use regression model to find out relationship between dependent and independent variables. We will also use descriptive and predictive analysis to find out how many customers are ready to subscribe the term deposit.

On this Banking related dataset, we will also use classification and clustering. By using these different models, we will collect different results for sales data. Below we create diagram of system design and methodology: -

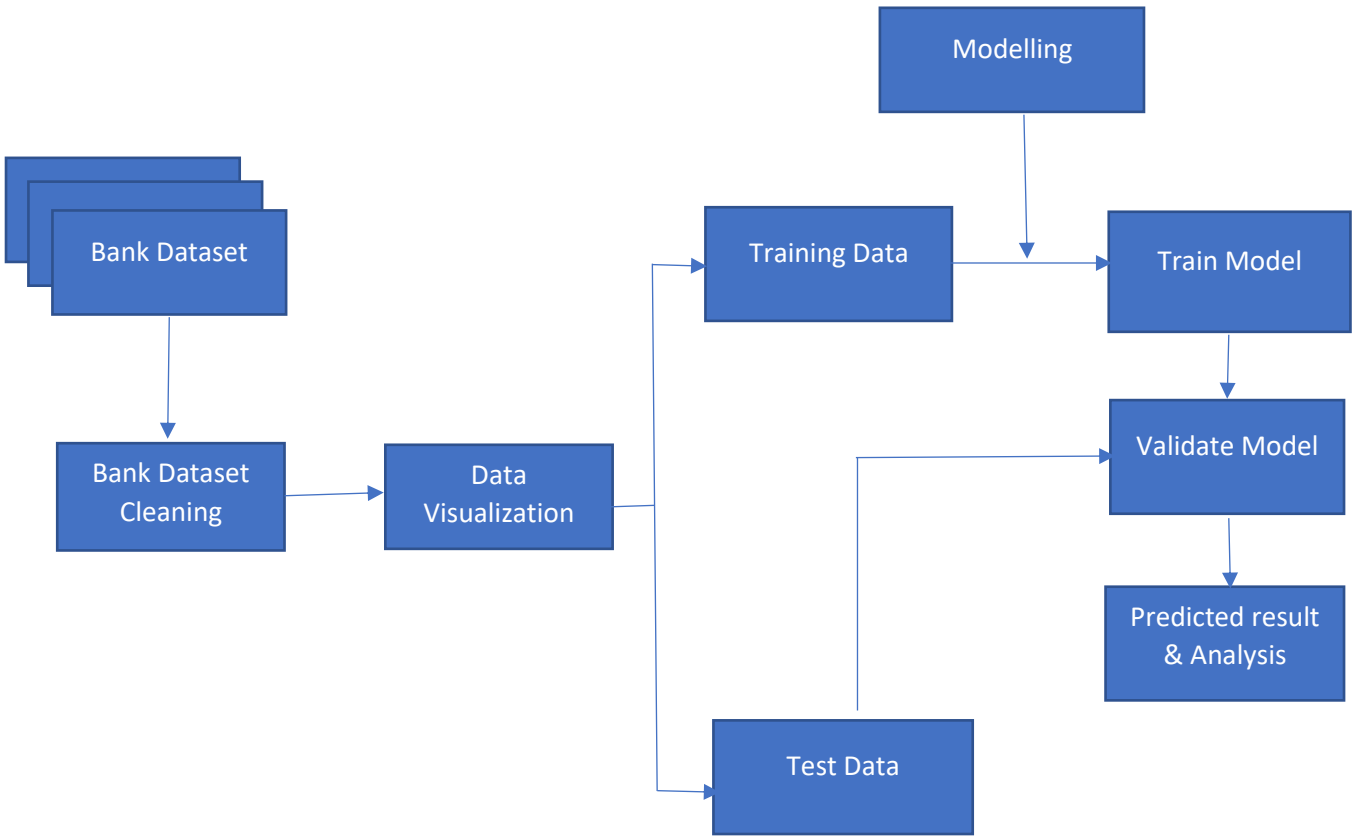

Fig. System Design and Methodology

The system design includes various stages like data collection, pre-processing, making training, testing data, implementing algorithm and last stage is predicted result. The accumulated uncooked statistics can be incomplete or noisy. The data should go through pre-processing section to smooth the information before the use of the facts forgetting to know, another step for schooling version is characteristic extraction. In next step we visualize the dataset and then split the dataset into training and testing. When we seeking to expect the output first, we want to teach the model the use of a dataset and model attempt to research the data to make correct prediction. Test information is unbiased of education records, if a model suit to the educate

facts, then it also suits to the test records, least overfitting has taken vicinity. Random forest model is most used set of rules in ML maximum of time this algorithm offers the satisfactory accuracy end result. Training the version using ML algorithm with the education records. The technique of trained version is evaluated with the test data is validate records, sooner or later, the model gives the first-rate predicted result.

A. **Python:** Python is a high-level, interpreted, and popular-cause programming language. Python is used by a software developer as a help language, it is easy to examine and its syntax may be very easy code and consists of lot of code library, easy to build fashions for machine gaining knowledge of. This program includes fewer traces of code than the alternative programming language. Various companies used Anaconda, is the most popular Python distribution broadly used for machine studying and records technological know-how.

B. **Machine Learning:** Machine mastering is a software of synthetic intelligence, is a method for facts evaluation and automates constructing model it learns from the previous facts based totally at the ideas it identifies facts sample and take selection on minimal human intervention. Machine mastering particularly concerned with sample and accuracy. Most industries working on gadget gaining knowledge of method to examine big number of facts which include financial service, Government, Healthcare, Retail, and Transportation.

C. **Supervised:** The majority of machine learning uses supervised learning . This work is licensed under a Creative Commons Attribution 4.0 International License collect the data and produce the output data based on previous experience. The task of learning function that maps an input and output variable and use an algorithm to learn the mapping function from the input to the output, the process of an algorithm learning from the training data. Supervised learning classified into two groups, classification, and regression.

D. **Unsupervised:** Unsupervised learning algorithm are used when the information used for training the machine that is neither classified nor labelled and algorithm allowed to act on the data without guidance, this algorithm mainly deals with hidden structure from unlabelled data and this algorithm does not give the right output. Unsupervised learning algorithm are less accurate compared to supervised learning, Unsupervised learning classified into two groups, clustering, and association problems.

E. **Random Forest:** Random Forest is a supervised, flexible, straightforward learning algorithm used for classification and regression. This random forest consists of multitude of decision tree and results are aggregated, random forest collect the classification and select the most voted prediction as the result, this algorithm reduce the risk of overfitting. Random forest algorithm is reduced overfitting, high accuracy and estimates missing data.

F. **Naive Bayes:** It is handiest class technique primarily based on Bayes theorem used for solving class hassle with an assumption of impartial among predictors and calculate the possibility of an event associated with preceding understanding. A Naive Bayes classifier assumes that the presence of a particular features in a category is unrelated to the presence of any other feature.

Data Dictionary:

For Categorical Attributes first I took the one variable at a time to check the datatype and assigned the correct datatype. Then I check the number of levels corresponding to each attribute and count of values corresponding to each level.

Table 1

Attribute	Description	Type of Attribute	Number of levels	Count
Job	Type of job	Character	12	admin. 1334 blue-collar 1944 admin. 1334 technician 1823 services 923
Marital	Marital status	Character	3	married 6351 single 3518 divorced 1293
Education	Customer education	Character	4	secondary 5476 unknown 497 tertiary 3689 primary 1500
Default	Has credit in default?	Binary	2	no 10994 yes 168
Housing	Has housing loan?	Binary	2	no 5881 yes 5281
Loan	Has personal loan?	Binary	2	no 9702 yes 1460
Contact	Contact communication type	Character	3	cellular 8042 unknown 2346 telephone 774

Month	Last contact month of year	Character	12	July 1514 August 1519 June 1222 April 923 February 776
Poutcome	Outcome of the previous marketing campaign	Character	4	unknown 8326 failure 1228 success 1071 other 537
Deposit	Has the client subscribed a term deposit?	Binary	2	no 5873 yes 5289

FOR NUMERICAL ATTRIBUTE

For numerical attributes I used seven number summary that is mean, minimum, maximum, count, standard deviation,25%,50%,75%

Table 2

Column1	count	mean	std	min	25%	50%	75%	max
age	11162	41	12	18	32	39	49	95
balance	11162	1529	3225	-6847	122	550	1708	81204
day	11162	16	8	1	8	15	22	31
duration	11162	372	347	2	138	255	496	3881
campaign	11162	3	3	1	1	2	3	63
pdays	11162	51	109	-1	-1	-1	21	854
previous	11162	1	2	0	0	0	1	58

Data Visualization:

1.Client with job profile Visualization

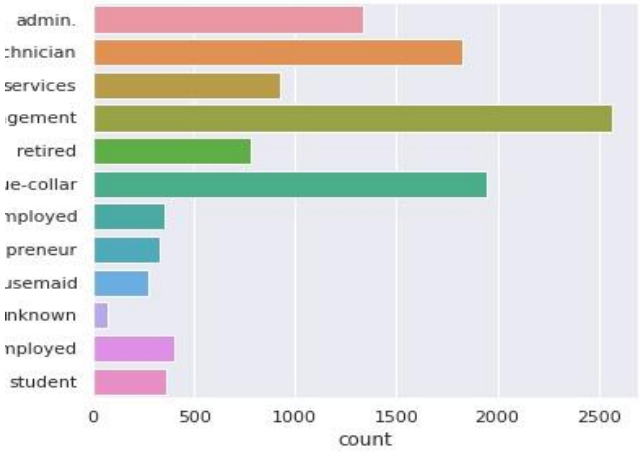

Figure 1

This horizontal bar graph shows that management job has the highest number of counts like 2500. And unknown is very low count near about 100 and unemployed and entrepreneur are same

2.Target variable Visualization

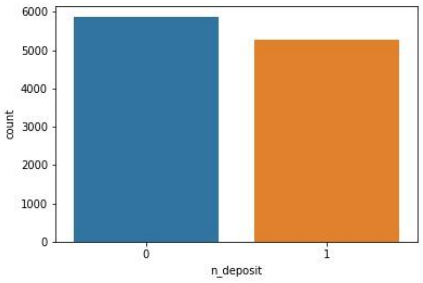

Figure 2

I can see the total number of count of the deposit (1) means has the client subscribed a term deposit or(0) not subscribed

The total number of counts 0 is5873 and count 1 is 5289.

3.Client with marital status and housing loan

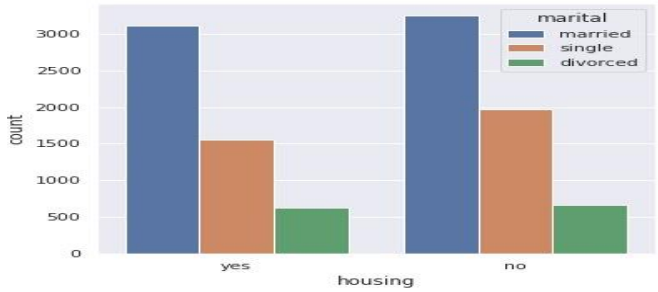

Figure 3

This bar graph shows the information approximately how many customers take the housing mortgage from the bank in step with their marital reputation. For ‘Yes’ constitute the customers who take the housing loan from banking and ‘No’ constitute the who don’t take any housing mortgage. The married customers have almost equal variety of counts similar to ‘Yes’ or ‘No’ that is records as above 3000 and approximately 3800 respectively.

4.Clients with education background

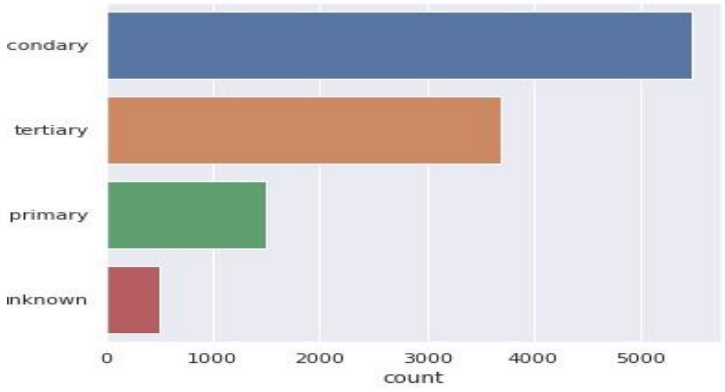

Figure 4

This graph represents the information according to their education background. Client whose education background is secondary are in high numbers. The second highest number of

educations with tertiary education, which is approximately 3800. And the unknown education is very low near about 500.

5.Client with education and month

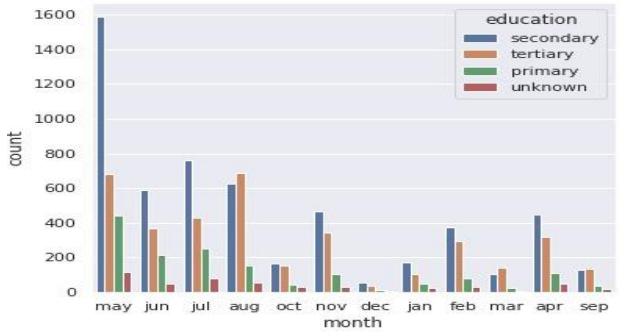

Figure 5

In May month, customer takes highest number of secondary educations. And the dec has the very lowest all level of education. In the month of jan and sep both are same education level.

6.Plotting pie chart

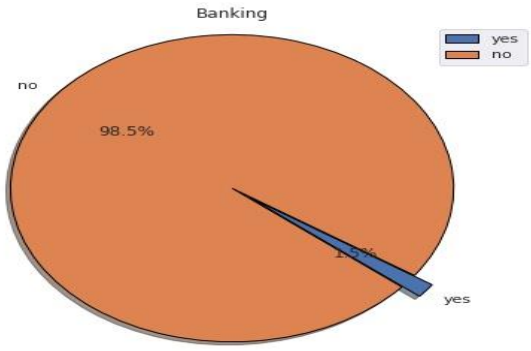

Figure 6

Client with housing loan, only 168 records corresponding to 'Yes' and 10994 to 'No'.

7.To cumulate that how many customers taking loan according to their education level?

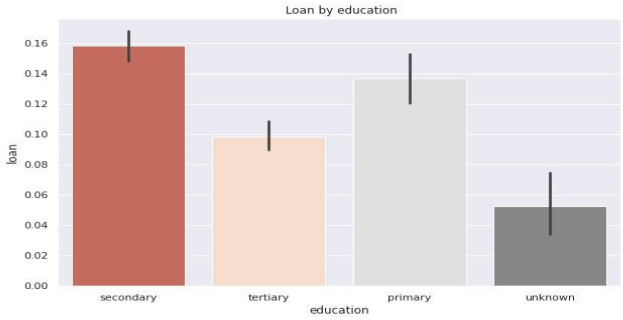

Figure 7

The person who takes the secondary education has taking the highest number of loans. Primary education level is 2nd highest number.

8. How many customers taking house loan from the bank?

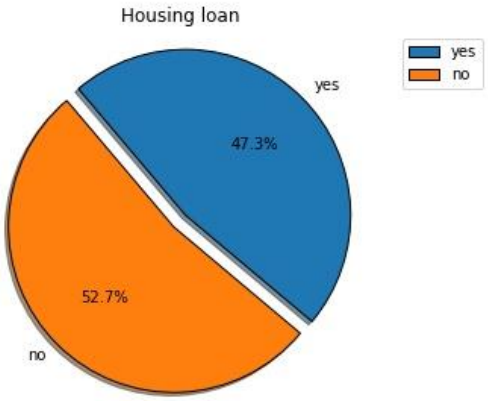

Figure 8

In this pie chart, we can see that 47.3% customers take less house loan and 52.7% took more house loan.

9.Use Box Plot to compare the age of customers for the top 5 of the most common employment forms

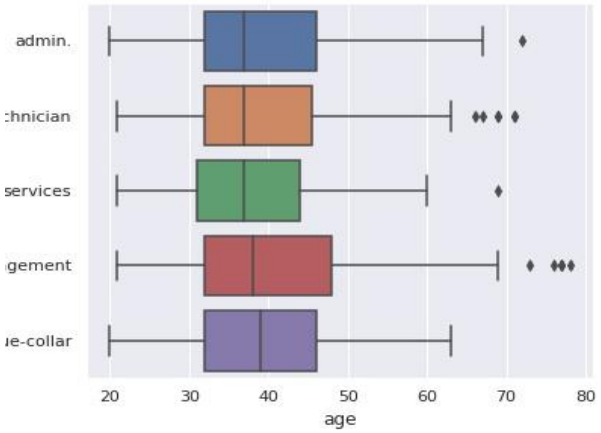

Figure 9

The plot shows that among the top-5 client categories, the most senior customers represent the management, and the largest number of outliers is among the management and technician.

10.To calculate which month customer deposit maximum amount of money?

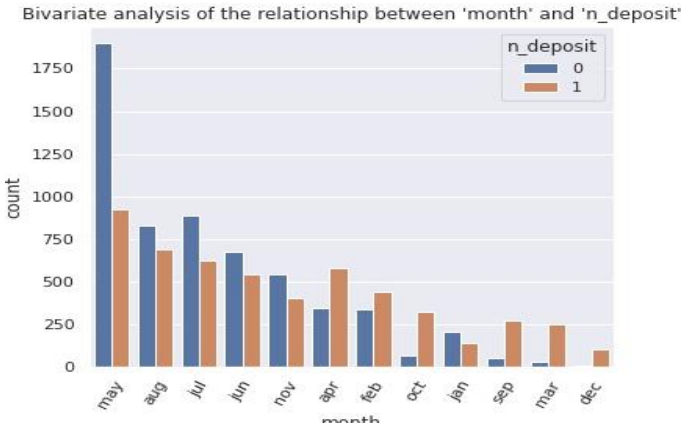

Figure 10

May got a slightly more subscribers than the other months. Regardless how many people is contacted the subscription average is almost the same with the exception of December and January. These months were got the fewest subscriptions. One possible reason could be the fact the people go for holidays. (In the Americas people are used to take holidays in this period of the year)

11.Correlation Matrix

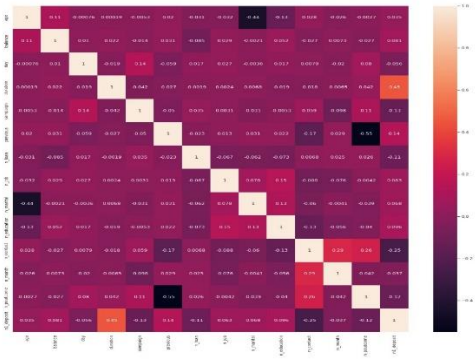

Figure 11

It is square matrix = each row represents a variable, and all the columns represent the same variables as rows, hence the number of rows = number of columns

Duration (0.45) is highly and more positive corelated with n1_deposit.

"campaign" is the most negative correlated feature to n1_deposit

"n_jobs" and "n_material" and "education" are positive correlated to each other

Train-Test-Split Method

In train-test-split method the entire dataset is partitioned into training and testing sets. The training set contain 70% and in the testing set it contain 30% . Then I apply the over-sampling technique to make our target variable balance.

Modelling

In the modelling part there are four methods Logistic Regression, KNN, Naive Bayes, random forest that I have applied on my dataset.

Logistic Regression: This model is supervised learning classification algorithm. It is used to predict the value of target variable.

KNN: The KNN stands for "K-Nearest Neighbour". It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

Naive Bayes: Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.

Random Forest: It is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Table 3

Column1	Model	Train_test_split
0	Logistic Regression	77.69483428
1	KNN	76.29143028
2	Naive Bayes	75.18662287
3	Random Forest	81.45715139

To check the accuracy of my dataset corresponding to each model. To find the best fit model for the dataset. I have found that the random forest is the best fit model for my dataset with highest accuracy 81% as compared to other models.

Confusion matrix corresponding to Random Forest Classifier Mode

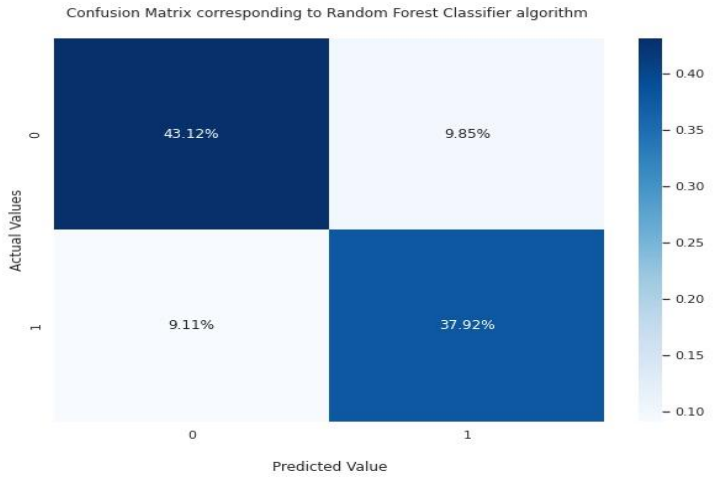

Figure 12

It is a tabular summary of the number of correct and incorrect predictions made by a classifier. The actual values correspond to 0 and predicted is 0 is 43.12% is right and the actual value correspond to 0 and predicted correspond to 1 is false (9.85%)

ROC Curve

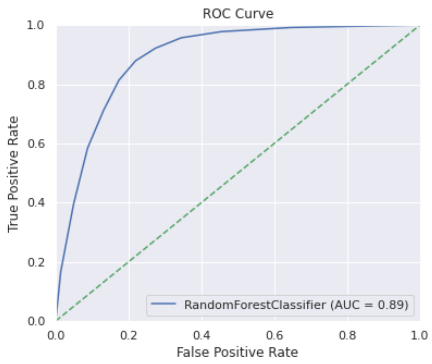

Figure 13

Receiver operating characteristic - it is metric to evaluate classifier output quality. The accuracy is 0.89. The area under curve is right. The main purpose of this connection between the clinical sensitivity and speciality for every possible cut off for test and connection of test.

Classification Report

Table 4

Column1	precision	recall	f1 score	support
0	0.82	0.83	0.83	1774
1	0.81	0.81	0.81	1575

Precision = how many predicted values is right is 0.82

Recall= how many actual values is right is 0.83

F1 score = It is positive class in binary classification or weighted average of the f1scores of each class for the multiclass task.

Support = It is occurrence class

CONCLUSION

In banking dataset big quantity of statistics is generated continuously and this information can be used to extract significant records. The bank dataset is used for term deposit prediction. It contains purchaser information. The dataset has styles of prediction both Yes and No. There are 16 input function and 1 output. After imposing (Supervised set of rules) Random Forest used for classification reason, the algorithm offers 81% accuracy for dataset by means of measuring accuracy, precision, recall, f1-score, support. The main goal of this work to predict whether a customer will subscribe to a time period deposit. The paintings on this paper have

used financial institution dataset from Kaggle internet site to make classification. After implementing, the result acquired become best.

REFERNCES

[1]. UCI Machine Learning Repository: Bank Marking Data Set

[2]. Dhandabani, (2010), " linkage between service quality and customers loyalty in Indian Retail banking*"* in the year 2010.

[3]. (Maya Basant Lohani, 2012) Focus on banking service quality "*Banks have more concentration on the tangible factor like a computerization, physical facilities, etc. to attract the customers*".

[4]. Jain, in the year 2012 "*Customers Perception on Service Quality in Banking Sector*".

[5]. Hertz 113-114, "*Bank's role and reputation within the marketplace*".

[6]. Broderick, A. and Vachirapornpuk, S. (2002). "*Service Quality in Internet Banking: The Importance of Customer Role. Marketing Intelligence and Planning*".