SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE
(AUTONOMOUS)
UJIRE - 574240

DEPARTMENT OF STATISTICS

CERTIFICATE

Certified that this is the bonafide record of project work done by Ms. Spoorthi K during the year 2023 as a part of her M.Sc (Statistics) fourth semester course work.

Reg. No.	**2**	**1**	**4**	**0**	**1**	**6**

Project Guide **Head of the Department**

Examiner

1.

2.

Place: Ujire
Date:

BANK CUSTOMER TERM DEPOSIT SUBSCRIPTION PREDICTION

Project Report submitted to the

SDM COLLEGE (Autonomous)

in partial fulfilment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

by

Spoorthi K

Under the supervision of

Asst. Prof. Mr. Pradeep K

Department of Post Graduate Studies

and Research in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE - 574240

Karnataka, INDIA

August 2023

DECLARATION

I, Spoorthi K, hereby declare that the matter embodied in this report entitled '**Bank Customer Term Deposit Subscription Prediction**' is a bonafide record of project work carried out by me under the guidance and supervision of **Asst. Prof. Mr. Pradeep K** , Department of Statistics, SDM College, Ujire - 574240, Karnataka, India. I further declare that no part of the work contained in the report has previously been formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or any other similar title or recognition of any other university.

Date:
Place: Ujire

(Spoorthi K)
E-mail: spoorthik093@gmail.com

CERTIFICATE

This is to certify that the project report entitled '**Bank Customer Term Deposit Subscription Prediction**' is a bonafide record of an authentic work carried out by **Spoorthi K**, under my guidance and supervision in the Department of Post Graduate Studies and Research in Statistics, SDM College, Ujire, in partial fulfilment of the requirements for the award of the degree of Master of Science in Statistics, under Mangalore University, Mangalagangothri. I further certify that this report or part thereof has not previously been presented or submitted elsewhere for the award of any Degree, Diploma, Associateship, Fellowhsip or any other similar title of any other institution or university.

Date:
Place: Ujire

(Pradeep K)
E-mail: pradeep.k@sdmcujire.in

ACKNOWLEDGEMENTS

Firstly, I would like to thank our Principal Dr. Kumar Hegde for providing the necessary facilities for the completion of this project work in our college.

I would also thank our Dean Dr. Vishwanath P for his support.

It is my privilege to thank our HOD Prof. Shanthiprakash for his suggestions and support.

I am very grateful to my Research Supervisor, Asst. Prof. Mr. Pradeep K , Department of Statistics, SDM College, Ujire, for his kind help and encouragement throughout my project work.

I gratefully acknowledge my teachers at the Department of Statistics, SDM College,Ujire, Asst. Prof. Ms. Shwetha Kumari and Asst. Prof. Ms. Supriya Shivadasan Padmavati for their support during my project work.

I am also thankful to all my family members and friends for their constant encouragement and help in each step.

My sincere thanks also goes to the students of SDM College, Ujire, who have helped me directly or indirectly during my project work.

Finally to all who helped me in many ways, I say, 'Thank You!'.

(Spoorthi K)

Contents

1 Chapter 1

Introduction

Figure 1: Bank Marketing

Bank marketing campaigns play a crucial role in the financial industry, helping banks reach their target customers, promote their products and services, and build strong customer relationships. In today's competitive banking landscape, effective marketing campaigns are essential for attracting new customers, retaining existing ones, and staying ahead of the competition.

Bank marketing campaigns encompass a wide range of initiatives, including product launches, branding efforts, customer acquisition strategies, cross-selling promotions, digital marketing tactics, community engagement activities, customer retention programs, and educational initiatives. Each campaign serves a specific purpose and utilizes various marketing channels to engage with customers effectively.

Deposits are the main source of revenue for banks. Many banks offer different types of accounts to attract customers willing to deposit their funds. The terms and conditions of depositing depend on the type of account. For instance, current accounts are held by customers willing to withdraw their funds at any time. On the other hand, fixed deposits are held by customers ready to lock their funds for a given period. The rate of interest is one of the motivating factors which encourage individuals to open fixed deposit accounts. A bank can increase the number of

subscribers to term deposits through effective marketing. Banks should have an effective marketing campaign strategy to reach their customers. Customer service is one of the marketing techniques that should be applied by banks. In this regard, the bank should ensure that the customers are treated fairly. The response team should assist customers within the shortest time possible. Video content campaigns are also used by various banks to attract customers. The primary objective of the video content campaigns is to ensure that the customers understand the products offered by the bank. If customers do not understand the terms under a fixed deposit account, they may not subscribe to it. Notably, customers are likely to subscribe to something that they know. The choice of a marketing strategy plays an important role in determining the level of subscription by banks. With the improvements in technology, banks can use Big Data to collect and analyze customer data. These data can be used to identify the likelihood of customers to subscribe to term deposits. If the bank realizes that many customers do not understand the terms and conditions of term deposits, the application of direct marketing strategies would be appropriate. The interaction between the bank officials and customers may increase the number of customers who are willing to subscribe to term deposits. Such communications improve customers' understanding of the bank's products.

Studies show that bank marketing campaigns focus on competitive strategies. Some of the strategies include demographic targeting, customer outreach, loyalty programs, and technology adoption. These strategies not only help the banks to reach many customers but to sell their products to the general public. By targeting a specific group of customers, banks can achieve their organizational objectives. One of the goals is an increase in the number of subscriptions to term deposits. The literature review tries to understand the findings of various researchers. The focus of the review is the factors that can increase customers' subscription to a term deposit.

Bank marketing plays a crucial role in attracting and retaining customers, promoting financial products and services, and driving business growth for banks. However, with a wide customer base and limited resources, it becomes essential for banks to target their marketing efforts effectively. This is where bank marketing classification comes into play.

Bank marketing classification refers to the process of categorizing customers or prospects into distinct groups based on various factors such as demographics, behavior patterns, psychographics, and predictive analytics. By segmenting customers, banks can better understand their needs, preferences, and behaviors, enabling them to tailor their marketing strategies accordingly.

Demographic segmentation allows banks to categorize customers based on characteristics such as age, gender, income, occupation, and education level. This segmentation helps banks create targeted marketing campaigns that resonate with specific customer groups.

Behavioral segmentation focuses on analyzing customer behavior patterns, including transaction history, purchase frequency, and response to previous marketing initiatives. By identifying customer segments that exhibit similar behaviors, banks can design personalized marketing campaigns that are more likely to generate positive responses.

Psychographic segmentation delves into customers' lifestyles, values, attitudes, and interests. By understanding the psychographic traits of different customer segments, banks can align their marketing messages with the preferences and aspirations of their target audience.

Predictive analytics takes advantage of advanced data analysis techniques to predict future customer behaviors. Using machine learning and predictive modeling, banks can classify customers based on their likelihood to respond to marketing campaigns, open new accounts, or engage in specific financial transactions.

Customer Lifetime Value (CLV) segmentation focuses on identifying high-value customers who have the potential for long-term profitability. By analyzing factors such as customer loyalty, account balances, and potential for cross-selling or upselling, banks can prioritize their marketing efforts on customers who can make a significant impact on their business growth.

In summary, bank marketing classification enables banks to optimize their marketing strategies, allocate resources efficiently, and enhance customer engagement, acquisition, and retention. By understanding their customers better and tailoring their marketing efforts accordingly, banks can build stronger relationships, increase customer satisfaction, and drive sustainable growth in a competitive financial landscape.

1.1 Motivation:

The primary motivation behind bank marketing classification is to achieve more targeted and personalized marketing campaigns. By segmenting customers into distinct groups based on demographics, behavior patterns, or psychographics, banks can tailor their messaging, offers, and promotions to specific customer segments. This approach ensures that marketing efforts reach the right audience with content that is relevant to their unique needs and preferences. By delivering messages that resonate with customers, banks increase the likelihood of capturing their attention and generating a positive response.

Effective allocation of marketing resources is crucial for banks to maximize their return on investment. By implementing bank marketing classification, banks can identify high-potential customer segments that are more likely to yield positive outcomes. This segmentation allows banks to allocate their marketing budgets, manpower, and time more efficiently by focusing on those segments that have a higher likelihood of conversion or engagement.

Understanding customer behaviors, preferences, and needs through classification enables banks to develop marketing strategies that engage customers more effectively. By analyzing customer data and segmenting them into groups with similar characteristics, banks can create targeted marketing campaigns that resonate with specific segments. These campaigns can deliver personalized messages, offers, and recommendations that capture customers' attention and establish a stronger emotional connection. Improved customer engagement leads to higher levels of customer satisfaction, loyalty, and advocacy, all of which contribute to long-term success for banks.

Effective bank marketing classification provides banks with a competitive edge in the market. By understanding their customers better and delivering personalized experiences, banks can differentiate themselves from their competitors. Targeted marketing campaigns based on customer segmentation enable banks to stand out by providing tailored solutions and meeting specific customer needs. This helps attract new customers, retain existing ones, and position the bank as a trusted financial partner.

In summary, bank marketing classification is motivated by the need for targeted marketing, resource optimization, improved customer engagement, enhanced cross-selling and upselling, and gaining a competitive advantage. By implementing classification techniques, banks can optimize their marketing strategies, deliver personalized experiences, and achieve better results in customer acquisition, retention, and overall business growth.

1.2 Literature Review

1. Mihova, Yana. (2019). Knowledge creation in banking marketing using machine learning techniques. 10.13140/RG.2.2.31756.16007 had used the machine learning techniques for analysis and making prediction using existing data in banking marketing. The success rate of banking marketing depend on the result and decision in order to make more accurate prediction statistical tool and methods are used. A different stage for data analysis and to find, how they can be used together in a process converting raw data to effective decision making knowledge and building the predictive model in this work used decision tree algorithm will help to predict the customer will subscribe the term deposit

2. Elsalamony et all. discussed all bank marketing campaign are depend on customer large data, the size of data source is impossible for human to analyst to come up with satisfying information that will help in decision making process. Data mining model are helping in the performance of the campaigns, in this work used most important data mining technique Multilayer Perception Neural Network(MLPNN),Naïve bayes, logistic regression, and decision tree, the purpose is increasing the campaign effectiveness and identifying the characteristics that effect a success

3. A data driven approach was suggested in archive.ics.uci.edu/ml/datasets/Bank +Marketing, 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 include the effect of financial crisis, analyzed large set of feature 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

4. Bank direct marketing is an interactive process Miguéis, V.L., Camanho, A.S. & Borges, J. Predicting direct marketing response in banking: comparison of class imbalance methods. Serv Bus 11, 831–849(2017).https://doi.org/10.1007/s11628-016-0332-3, for building the good relationship among customers, to study the customer characteristics and behavior use an effective multi-channel communication. A part from profit growth, which may raise customer positive response, the goal of bank marketing is to increase the customer response of direct marketing campaign.

5. Customer profiling in S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, "Customer profiling using classification approach for bank telemarketing," JOIV: International Journal on Informatics Visualization, vol. 1, no. 4-2, pp. 214–217, 2017., using classification approach for bank telemarketing, data mining approaches started by many companies to restore the customer profiling. Decision tree, random forest, and Naïve Bayes were used, for predicting the customer profiles and increasing the telemarketing sales classification is useful for measured accuracy percentage, precision and recall rates. Before evaluating the classifiers preprocessing and normalization were conducted for conducting the experiments and evaluation process RapidMiner tool was used. Finally, result show that decision tree is the best classifier for predicting the customer profile and behavior

6. From visualization aspects, K. Sagar and A. Saha, "A systematic review of software usability studies," International Journal of Information Technology, pp. 1–24,2017. explained several types of visualization techniques, such as radial, hierarchical, graph,and bar chart visualization, and presented the impact of human–computer interaction knowledge on opinion visualization systems. Prior domain knowledge yielded high understand ability , user-friendliness, usefulness, and informativeness. Age factor affected the usability metrics of other systems, such as visual appeal, comprehensiveness, and intuitiveness. These findings were projected to the visualization of the direct marketing industry because it is mainly aided by end users and customers

1.3 Objectives

1. To understand how bank clients responded to the marketing campaign and assess its effectiveness.

2. To understand characteristics of customer with respect to banking.

3. To develop classification models to predict subscription of term deposit and personal loan.

1.4 Scope of the study

Term deposit classification is particularly useful in the banking and financial services industry. Banks and financial institutions offer term deposits as a type of investment product to their customers. Term deposits are fixed-term investments where customers deposit a certain amount of money for a specific period, typically earning a fixed interest rate.
Ultimately, the study aims to guide financial institutions in formulating robust business strategies concerning term deposits, optimizing product offerings, and making data-driven decisions to enhance customer engagement and profitability in the dynamic and competitive banking and financial services industry.

2 Chapter 2

2.1 Materials and Methods

The dataset is selected from Kaggle. This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository.It has 11162 rows and 17 columns.The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

2.2 About the Data:

- age: (numeric)
- job:type of job (categorical:'admin.','bluecollar','entrepreneur','housemaid','management','retired','selfemployed','services','student','technician','unemployed','unknown')
- marital: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- education: (categorical: primary, secondary, tertiary and unknown)
- default: has credit in default? (categorical: 'no','yes','unknown')
- housing: has housing loan? (categorical: 'no','yes','unknown')
- loan: has personal loan? (categorical: 'no','yes','unknown')
- balance: Balance of the individual.

- contact: contact communication type (categorical: 'cellular','telephone')

- month: last contact month of year (categorical: 'jan','feb', 'mar', ..., 'nov', 'dec')

- day: last contact day of the week (categorical:'mon','tue','wed','thu','fri')

- duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

- previous: number of contacts performed before this campaign and for this client (numeric)

- poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

- deposit - has the client subscribed a term deposit? (binary: 'yes','no')

2.3 Statistical Methods:

The open source softwares like Python programming languages has been used to carry out the analysis.

2.3.1 Python:

It is the programming language used for analysis and building model.
Python libraries:
Numpy:
NumPy is the general-purpose array processing package tool for handling the n-dimensional arrays. It also has functions for working in domain of linear algebra, fourier transform and matrices.
Matplotlib:
Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It is a comprehensive library for creating static, animated and interactive visualizations in python.
Pandas:
Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting. Pandas mainly used for machine learning in from of dataframes. Pandas allows various data manipulation operations such as groupby, join, merge, data cleaning.
Seaborn:
Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library.

2.3.2 Graphical techniques:

A graph is basically an illustration of how two variables relate to one another. Some simple popular graphical techniques are pie chart, bar plot, count-plot. These are very informative, simple to understand and interpret.

- Bar-Graph:
  Bar graphs are one of the many techniques used to present data in a visual form so that the reader may readily recognize patterns or trends. Bar graphs usually present categorical and numeric variables grouped in class intervals.

- Pie chart:
  A pie chart is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data, and it is a type of pictorial representation of data. A pie chart requires a list of categorical variables and numerical variables. Here, the term "pie" represents the whole, and the "slices" represent the parts of the whole.

- Hist plot:
  A histogram is used to represent data provided in the form of some groups. It is an accurate method for the graphical representation of numerical data distribution. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency.

- Q-Q plot:
  A Q–Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point on the plot corresponds to one of the quantiles of the second distribution plotted against the same quantile of the first distribution.

2.3.3 Chi Square test for independence of attributes:

his test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables. Here the sample data are displayed in a contingency table. The hypothesis under consideration are given as follows :
H_0 : The two categorical variables are independent.

H_1 : The two categorical variables are dependent.
The test statistic is given as follows :

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$$

where O represents the observed frequency, E is the expected frequency under the null hypothesis and is computed as follows :

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

The expected frequency count for each cell of the table must be atleast 5.
The test procedure is to reject the null hypothesis H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$, where $\chi^2_{\alpha,(r-1)(c-1)}$ is the upper α^{th} percentile value of the central chi-square distribution with $(r-1)(c-1)$ degrees of freedom, r is the number of rows and c is the number of columns.

One can also use the $p-value$ to draw conclusion about the test. If the $p-value$ is less than 0.05, then reject the null hypothesis H_0 and accept the alternative hypothesis H_1.

2.3.4 Independent Samples T-Test

The independent t-test, also called the two sample t-test, independent-samples t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. Procedures for independent sample t-test:

1. Set up the hypothesis.
    - Null Hypothesis: It is assumed when the means of the two groups are not significantly different.
    - Alternative Hypothesis: Assumes that the means of the two groups are significantly different.
2. Calculate the standard deviation for the independent sample t-test by using this formula: S=Σ(X_1-$\bar{X}_1$)^2 + Σ(X_2-$\bar{X}_2$)^2 n_1+n_2-2

2.3.5 SMOTE

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

2.3.6 Box Cox Transformation

A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

2.3.7 Yeo Johnson Transformation

Both Box-Cox and Yeo-Johnson transform non-normal distribution into a normal distribution. However, Box-Cox requires all samples to be positive, while Yeo-Johnson has no restrictions.

2.3.8 Confusion Matrix

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data.

2.3.9 ROC Curve

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is the plot of the true positive rate against the false positive rate, at various threshold settings.

2.3.10 Logistic Regression

In statistics, logistic regression, or logit regression, or logit model, is a type of probabilistic statistical classification model. It can be seen as a special case of generalized linear model. It is used in estimating the parameters of a qualitative response model. It is used widely in many fields such as medical, social science, engineering, marketing, economics, etc.

Logistic regression is used to refer specifically to the problem in which the depen-

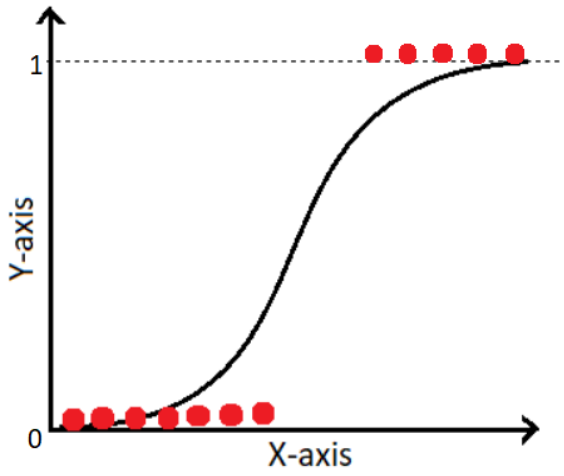


Figure 2: Logistic Regression

dent variable is binary, i.e., the number of available categories is two. It measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. Logistic regression is used for predicting binary outcomes of the dependent variable, treating the dependent variable as the outcome of a bernoulli trial. It is used to predict the odds of being a case based on the values of the independent variables or predictors. The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase. The logistic function is given as follows :

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

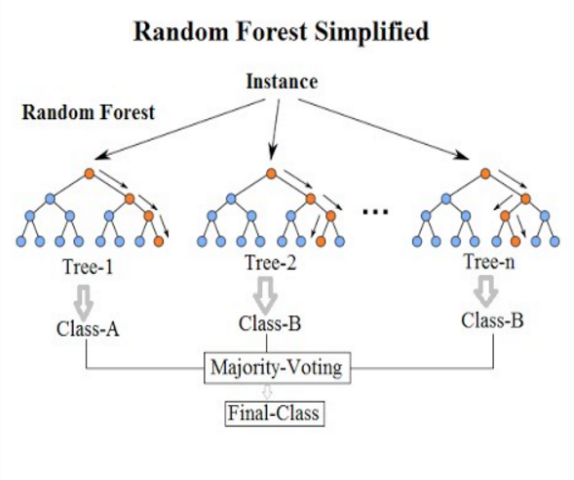
where $F(x)$ is the probability that the dependent variable equals a case, given some linear combination x of the predictors, β_0 is the intercept from the linear regression

equation, $\beta_1 x$ is the regression coefficient multiplied by some value of the predictor, base e denotes the exponential function.

2.3.11 Random Forest Classification

Random forests are a popular supervised machine learning algorithm.

- Random forests are for supervised machine learning, where there is a labeled target variable.
- Random forests can be used for solving regression (numeric target variable) and classification (categorical target variable) problems.
- Random forests are an ensemble method, meaning they combine predictions from other models.
- Each of the smaller models in the random forest ensemble is a decision tree.

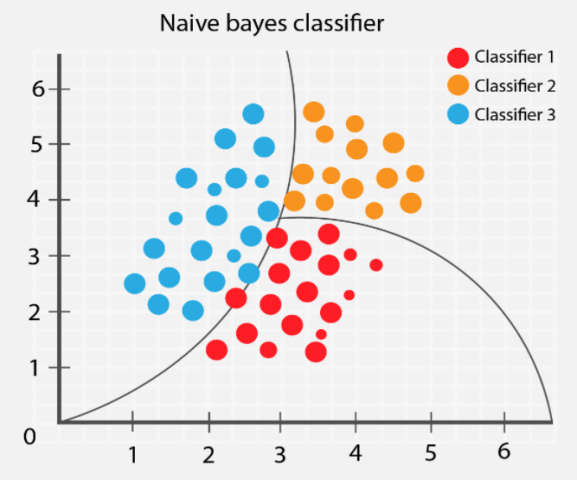


Random Forest

In a random forest classification, multiple decision trees are created using different random subsets of the data and features. Each decision tree is like an expert, providing its opinion on how to classify the data. Predictions are made by calculating the prediction for each decision tree, then taking the most popular result. (For regression, predictions use an averaging technique instead.)

2.3.12 Naive Bayes Classifiers

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

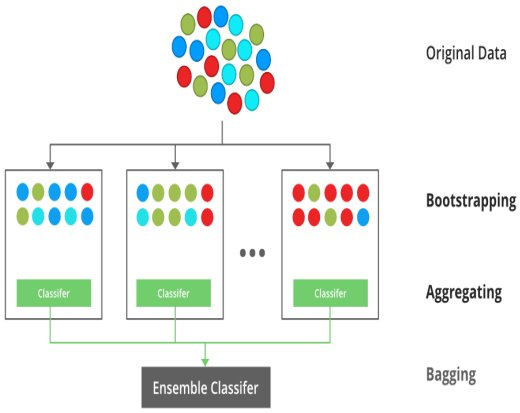


Naive Bayes

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.
$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$
Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

2.3.13 XG Boost Classifier

XGBoost, short for Extreme Gradient Boosting, is a highly popular machine learning algorithm known for its exceptional performance in classification and regression tasks. It is built on the gradient boosting framework, which involves combining multiple weak learners, typically decision trees, to form a robust and accurate predictive model.



XGBoost Classifier

The algorithm works in an iterative manner, where each subsequent weak learner is trained to correct the errors made by the previous ones. It employs a gradient-based optimization technique to minimize a specified loss function, resulting in a model that continuously improves its predictive capability.

XGBoost incorporates various techniques to enhance its performance, such as regularization methods like L1 and L2 regularization to prevent overfitting, and tree pruning to reduce complexity. It also includes advanced features like parallel processing and column block caching, which accelerate the training process and make it more efficient.

3 Chapter 3
Results and discussion

3.1 Exploratory Data Analysis:

3.1.1 Analysis of Marital Status Distribution Among Bank Clients

Table 1: Distribution of Marital Status

Marital Status	Percentage
single	31.8%
Married	56.9%
Divorced	11.2%

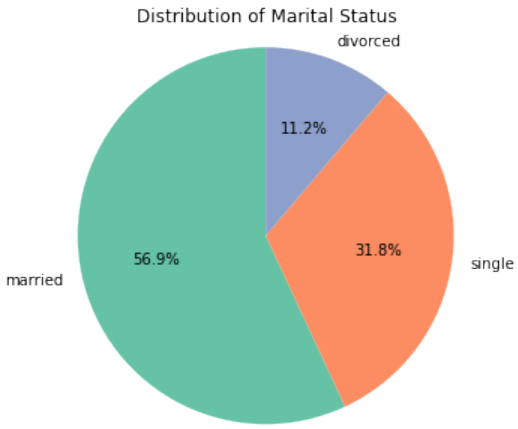


Figure 3: Pie Chart for the distribution of Marital Status

The pie chart provides a visual representation of the distribution of clients' marital status in the bank. It shows that 56.9% of the bank's clients are married. In other words, more than half of the clients are in a marital relationship.31.8% of the clients who are single. This means that just over a third of the clients are not married and are considered single. 11.2% of the bank's clients are divorced. This means that approximately 1 in 10 clients have been through a divorce.

3.1.2 Analysis of Education Level Distribution Among Bank Customers

Table 2: Education level

Education Level	Number of clients	Percentage of clients
Secondary	5067	48.5 %
Tertiary	3535	33.8%
Primary	1369	13.1%
Unknown	478	4.6%

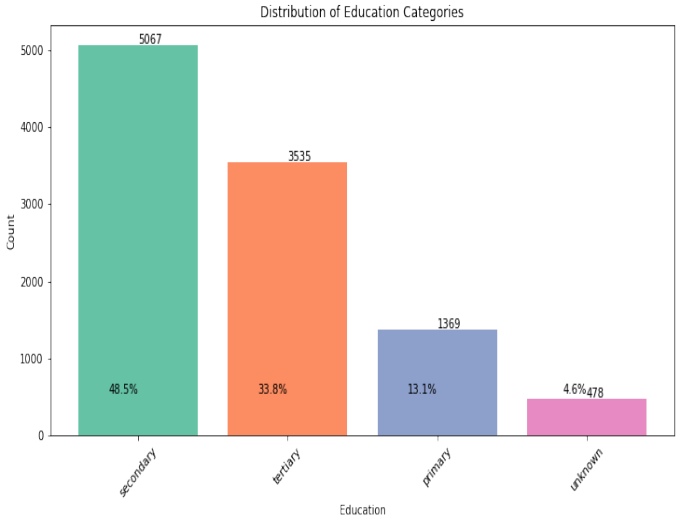


Figure 4: Bar graph representing education level of clients

48.5% of the bank's customers fall into the "Secondary" education category. This category likely includes individuals who have completed their secondary education (high school or equivalent).The second largest percentage in the data represents 33.8% of the customers who are classified under the "Tertiary" education category. This category typically includes individuals who have completed higher education, such as college or university degrees.The smallest percentage, which is 4.6%, represents the customers whose education category is listed as "Unknown." This category likely includes customers for whom the education information is missing or not recorded.

3.1.3 Analysis of Default, Housing Loan, and Personal Loan Distribution among Bank Customers

Table 3: Distribution of Credit in default

Default	Percentage
Yes	0.9%
No	99.1%

Table 4: Distribution of Housing Loan

Housing Loan	Percentage
Yes	45.5%
No	54.5%

Table 5: Distribution of Personal Loan

Personal Loan	Percentage
Yes	11.8%
No	88.2%

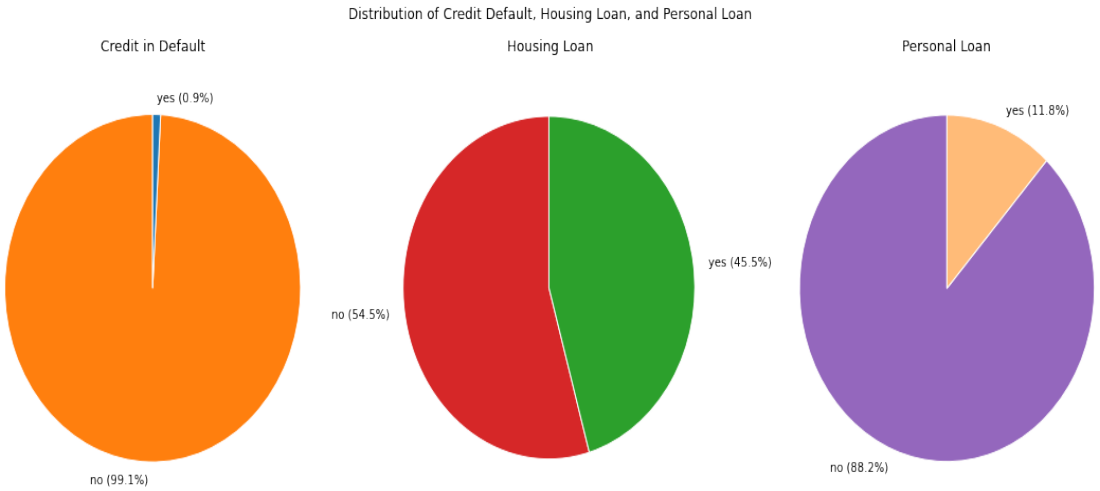


Figure 5: Pie chart showing distribution of Default, Housing Loan and Personal Loan

This slice of the pie indicates that 99.1% of the bank's customers do not have any credit in default. This means that the vast majority of customers have a good credit history and are not currently in default on any credit obligations.The second category represents 54.5% of customers who do not have a housing loan. This means that a little over half of the customers do not currently have a housing loan.The third category shows that 88.2% of customers do not have a personal loan. This means that the majority of customers have not taken out a personal loan.customers have subscribed to a Housing Loan (45.5%) compared to those who have taken a Personal Loan.

3.1.4 Monthly Customer Count Analysis

Table 6: Counts of customer by month:

Month	Counts
January	327
February	747
March	269
April	893
May	2522
June	1147
July	1355
August	1462
September	316
October	388
November	915
December	108

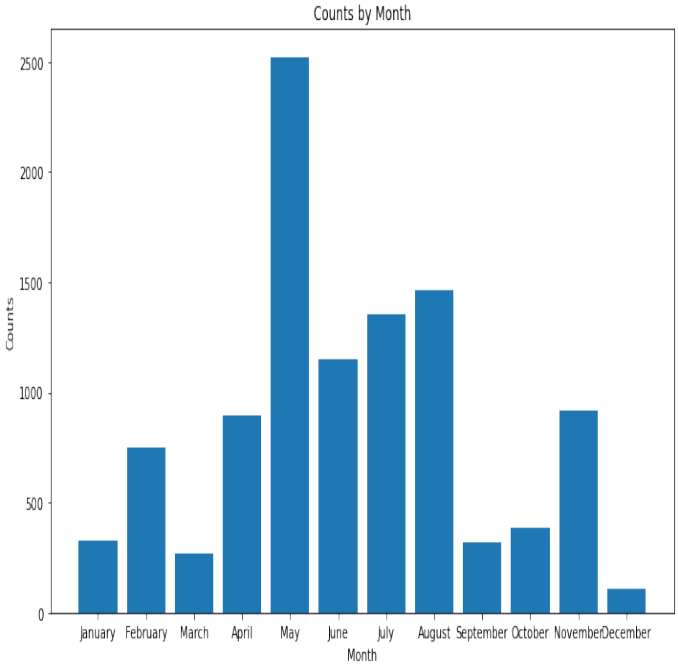


Figure 6: Bar chart showing monthly count of customers

May stands out as the month with the highest frequency of client contacts. This indicates that the bank actively reaches out to its clients during May, possibly for various reasons such as marketing campaigns, promotional offers, or other financial services.June and July are the next two months with a moderate frequency of client contacts. While they are not as actively contacted as May, they still receive a considerable number of interactions. This suggests that these months may also be important periods for the bank's marketing and customer engagement efforts.Similar to June and July, August also experiences a moderate level of client contacts. This might be due to the continuation of marketing activities or the introduction of new offers for the summer period.In contrast to May, March, and December have the fewest client contacts. These months might be relatively quieter in terms of marketing and customer outreach. December could be associated with reduced activity due to the holiday season, while March might lack major marketing initiatives.

3.1.5 Analysis of Distribution of Previous Outcome among Bank Clients

Table 7: Distribution of Previous outcome

Previous outcome	Percentage
Success	10.1%
Failure	11.2%
Other	4.9%
Unknown	73.8%

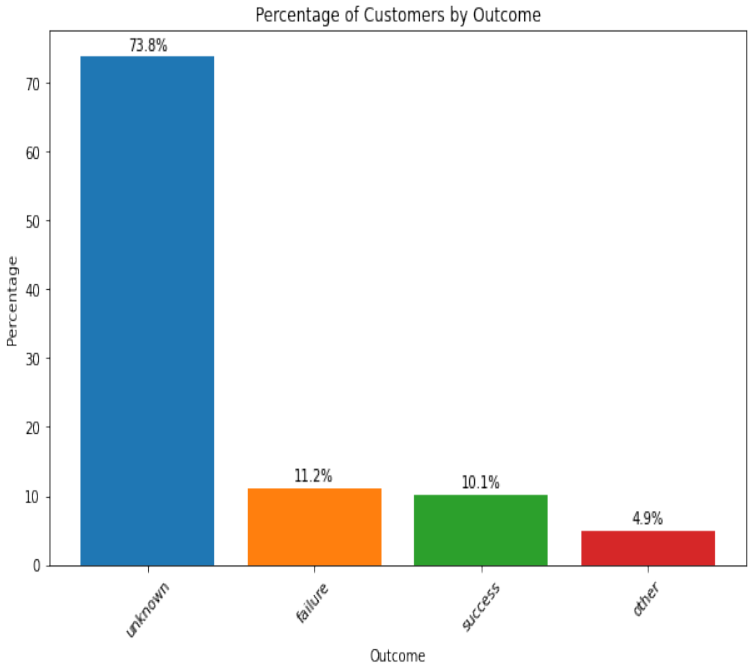


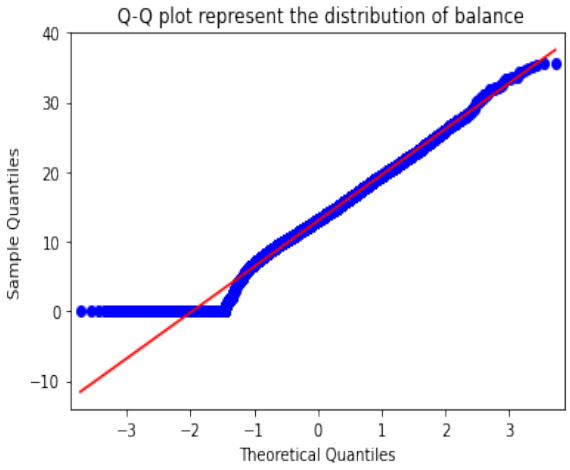
Figure 7: Bar graph showing the distribution of previous outcome

The majority of the outcomes from the previous marketing campaign are classified as "Unknown," accounting for 73.8% of the total outcomes. The "Unknown" category likely includes instances where the bank could not determine or did not have data on the specific outcome for a significant portion of the customers contacted during the campaign. The "Failure" category represents 11.2% of the outcomes. This means that around 11.2% of the customers contacted during the previous marketing campaign did not respond positively to the bank's offer or did not take the desired action (e.g., subscribing to a product or service)The "Success" category accounts for a relatively low percentage of outcomes, standing at 10.1%. This implies that only a small portion of the customers contacted during the campaign responded positively and took the desired action, such as subscribing to a product or service offered by the bank.

3.2 Two Samples T-Test

3.2.1 To determine if there is a significant difference in the mean bank balance between the customers who open a term deposit and who does not open a term deposit.

checking for the normality using Q-Q plot for balance



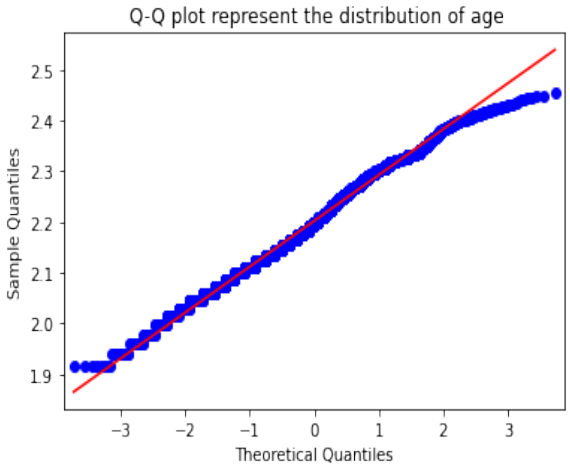
Testing the following hypothesis:

- H_0:There is no significant difference in the mean bank balance between the customers who open a term deposit and who does not open a term deposit.
- H_1: There is a significant difference in the mean bank balance between the customers who open a term deposit and who does not open a term deposit.

T-statistic: 13.7249
P-value: 1.680124928240775e-42
we observe that p-value is less than 0.05 ,we reject the null hypothesis.There is a significant difference in the mean bank balance between the customers who open a term deposit and who does not open a term deposit.

3.2.2 To determine if there is a significant difference in the mean age between the customers who open a term deposit and who does not open a term deposit.

checking for the normality using Q-Q plot for age



Testing the following hypothesis:

- H_0:There is no significant difference in the mean age between the customers who open a term deposit and who does not open a term deposit.
- H_1: There is a significant difference in the mean age between the customers who open a term deposit and who does not open a term deposit.

T-statistic: -0.70606
P-value: 0.4801
we observe that p-value is more than 0.05 ,we accept the null hypothesis.There is no a significant difference in the mean age between the customers who open a term deposit and who does not open a term deposit.In practical terms, this result suggests that the average age of customers who open a term deposit is likely to be similar to the average age of customers who do not open a term deposit. The lack of a significant difference implies that age may not be a significant factor in determining whether a customer opens a term deposit or not.

3.2.3 To determine if there is a significant difference in the mean bank balance between the customers who has a personal loan and who does not have a personal loan.

Testing the following hypothesis:

- H_0:There is no significant difference in the mean bank balance between the customers who has a personal loan and who does not have a personal loan.
- H_1: There is a significant difference in the mean bank balance between the customers who has a personal loan and who does not have a personal loan.

T-statistic: -9.23327
P-value: 3.133283488428764e-20
we observe that p-value is less than 0.05 ,we reject the null hypothesis.There is a significant difference in the mean bank balance between the customers who has a personal loan and who does not have a personal loan.

3.3 Bivariate analysis using Chi-Square Test

3.3.1 Testing the association between deposit and marital status

Chi-Square test is a non-parametric method used to compare the relationship between the two categorical variables in a contingency table.
The hypothesis under consideration are,

- H_0(Null hypothesis): There is no significant relationship between deposit and marital status.
- H_1(Alternative hypothesis):There is significant relationship between deposit and marital status.

	marital		
deposit	divorced	married	single
no	596	3290	1496
yes	577	2659	1831

The test statistic is given as follows,

$$\chi^2 = \frac{P(Observedvalue - Expectedvalue)^2}{Expectedvalue}$$

Expected value is calculated using the formula:

$$E = \frac{rowtotal \times columntotal}{grandtotal}$$

The obtained values are as follows: ˆ p-value = 1.3151555300064489e-20 ˆ Cramer V = 0.092582 Here we can observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis and we conclude that There is significant relationship between deposit and marital status.From the Cramer V value, the strength of the association between deposit and marital status is 0.09. Hence, we conclude that there is a weak association between deposit and marital status

3.3.2 Testing the association between deposit and contact

Chi-Square test is a non-parametric method used to compare the relationship between the two categorical variables in a contingency table.
The hypothesis under consideration are,

- H_0(Null hypothesis): There is no significant relationship between deposit and contacts performed.
- H_1(Alternative hypothesis):There is significant relationship between deposit and contacts performed.

	contact		
deposit	cellular	telephone	unknown
no	3400	359	1623
yes	4204	384	479

The test statistic is given as follows,

$$\chi^2 = \frac{P(Observedvalue - Expectedvalue)^2}{Expectedvalue}$$

Expected value is calculated using the formula:

$$E = \frac{rowtotal \times columntotal}{grandtotal}$$

p-value = 1.2094637110074326e-152 ˆ Cramer V = 0.258397 Here we can observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that There is significant relationship between deposit and contacts performed.From the Cramer V value, the strength of the association between deposit and contacts performed is 0.26. Hence, we conclude that there is a moderate association between deposit and contacts performed.

3.3.3 Testing the association between deposit and previous outcome

Chi-Square test is a non-parametric method used to compare the relationship between the two categorical variables in a contingency table.
The hypothesis under consideration are,

- H_0(Null hypothesis): There is no significant relationship between deposit and previous outcome.
- H_1(Alternative hypothesis): There is a significant relationship between deposit and previous outcome.

	poutcome			
deposit	failure	other	success	unknown
no	573	217	92	4500
yes	595	297	968	3207

The test statistic is given as follows,

$$\chi^2 = \frac{P(Observedvalue - Expectedvalue)^2}{Expectedvalue}$$

Expected value is calculated using the formula:

$$E = \frac{rowtotal \times columntotal}{grandtotal}$$

The obtained values are as follows:

- $p - value$ =1.4639958675420722e-204
- Cramer V = 0.32028

Here we can observe that the obtained $p - value$ is less than 0.05. Hence we reject the null hypothesis. There is a significant relationship between deposit and previous outcome. From the Cramer V value,the strength of the association between deposit and previous outcome is 0.32. Hence, we conclude that there is a relatively strong association between deposit and previous outcome.

3.3.4 Testing the association between loan and education

Chi-Square test is a non-parametric method used to compare the relationship between the two categorical variables in a contingency table.
The hypothesis under consideration are,

- H_0(Null hypothesis): There is no significant relationship between personal loan and education level.
- H_1(Alternative hypothesis): There is a significant relationship between personal loan and education level.

	education			
loan	primary	secondary	tertiary	unknown
no	1208	4334	3214	458
yes	161	733	320	20

The test statistic is given as follows,

$$\chi^2 = \frac{P(Observedvalue - Expectedvalue)^2}{Expectedvalue}$$

Expected value is calculated using the formula:

$$E = \frac{rowtotal \times columntotal}{grandtotal}$$

The obtained values are as follows:

- $p - value$ 1.0510217160327127e-18
- Cramer V = 0.020033199

Here we can observe that the obtained $p - value$ is less than 0.05. Hence we reject the null hypothesis. There is a significant relationship between personal loan and education level. From the Cramer V value, the strength of the association between personal loan and education level is 0.02. Hence, we conclude that there is a weak association between personal loan and education level.

3.3.5 Testing the association between loan and marital status

Chi-Square test is a non-parametric method used to compare the relationship between the two categorical variables in a contingency table.
The hypothesis under consideration are,

- H_0(Null hypothesis): There is no significant relationship between personal loan and marital status.
- H_1(Alternative hypothesis): There is a significant relationship between personal loan and marital status.

	marital		
loan	divorced	married	single
no	1010	5172	3033
yes	163	777	294

The test statistic is given as follows,

$$\chi^2 = \frac{P(Observedvalue - Expectedvalue)^2}{Expectedvalue}$$

Expected value is calculated using the formula:

$$E = \frac{rowtotal \times columntotal}{grandtotal}$$

The obtained values are as follows:

- $p - value$ =7.289681441660971e-10
- Cramer V = 0.0619

Here we can observe that the obtained $p-value$ is less than 0.05. Hence we do reject the null hypothesis. There is a significant relationship between personal loan and marital status. From the Cramer V value,the strength of the association between personal loan and marital status is 0.06. Hence, we conclude that there is a weak association between personal loan and marital status.

3.4 Multivaritae Analysis

3.4.1 Logistic Regression

Logistic regression model
Let the deposit be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing,loan, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.

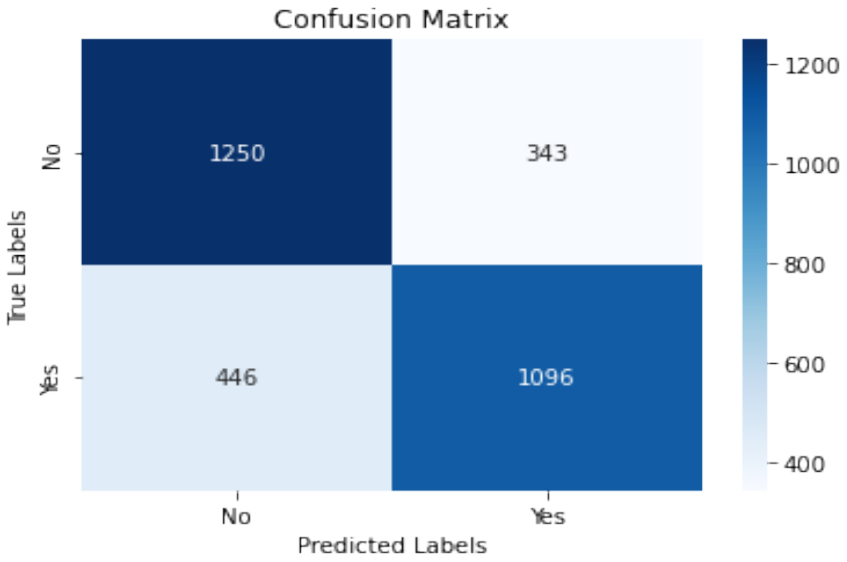
deposit	frequency
0	5382
1	5067

The performance of the model is as follows:

- Training set Accuracy: 0.7581
- Training set F1 score: 0.757775
- Training set Precision: 0.758328
- Training set Recall: 0.758135

- Test set Accuracy: 0.7483
- Test set Fl score: 0.7479
- Test set Precision: 0.7491
- Test set Recall: 0.7483
- Root Mean Square Error: 0.491797

Confusion matrix is as follows:



In conclusion, the Logistic Regression model shows decent performance on both the training and test sets, with the test set metrics being close to the training set metrics.

3.4.2 Random Forest Classifier

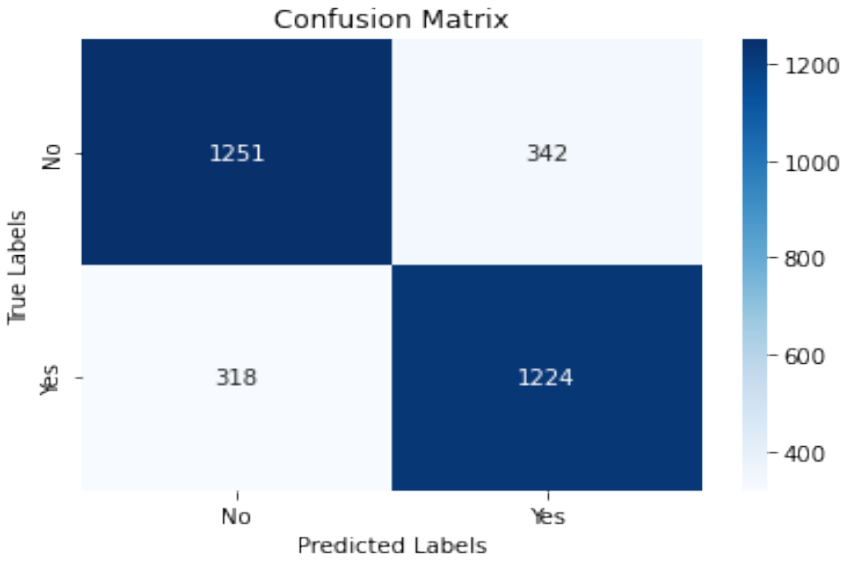
Let the deposit be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing,loan, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.

The performance of the model is as follows:

- Training set Accuracy: 0.8048
- Training set F1 score: 0.804788
- Training set Precision: 0.804857
- Training set Recall: 0.804758

- Test set Accuracy: 0.7901
- Test set Fl score: 0.7901
- Test set Precision: 0.7902
- Test set Recall: 0.7901
- Root Mean Square Error: 0.441861

Confusion matrix is as follows:



In conclusion, the Random Forest model shows reasonably good performance on both the training and test sets, with the test set metrics being slightly lower than the training set metrics.

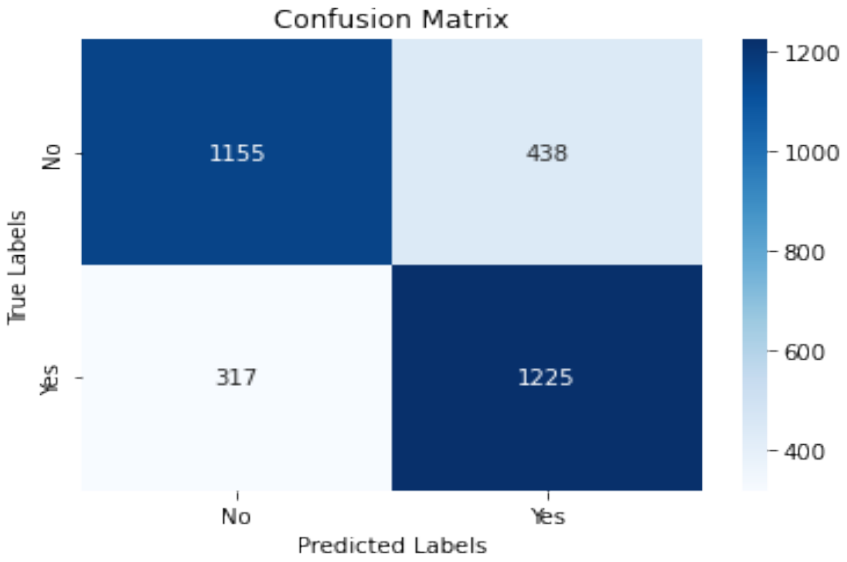
3.4.3 Naive Bayes Classifier

Let the deposit be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing,loan, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.
The performance of the model is as follows:

- Training set Accuracy: 0.7465
- Training set F1 score: 0.746399
- Training set Precision: 0.749444
- Training set Recall: 0.746514

- Test set Accuracy: 0.7592
- Test set Fl score: 0.7590
- Test set Precision: 0.7610
- Test set Recall: 0.7592
- Root Mean Square Error: 0.5034743

Confusion matrix is as follows:



In conclusion, the Naive Bayes model shows reasonably good performance on both the training and test sets, with the test set metrics being slightly higher than the training set metrics

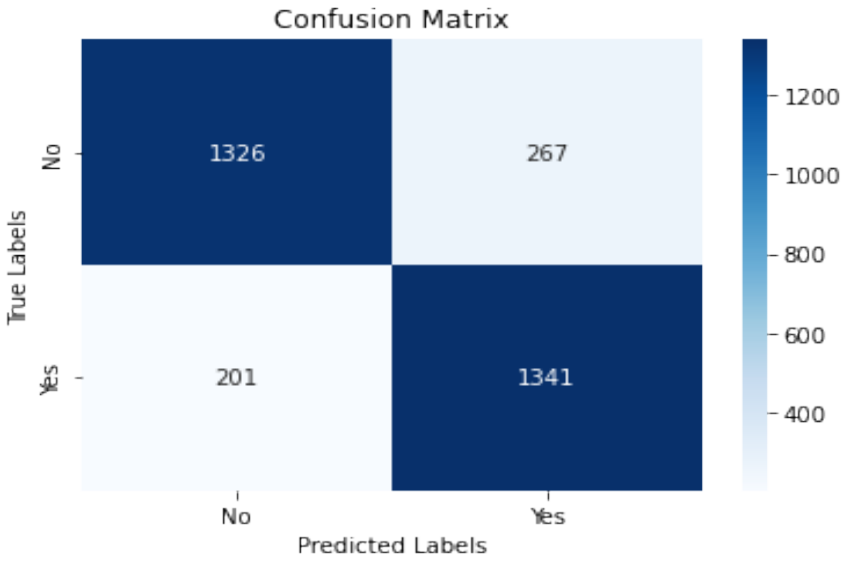
3.4.4 XG Boost Classifier

Let the deposit be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing,loan, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.
The performance of the model is as follows:

- Training set Accuracy: 0.8778
- Training set F1 score: 0.877792
- Training set Precision: 0.879375
- Training set Recall: 0.877769

- Test set Accuracy: 0.8507
- Test set Fl score: 0.8507
- Test set Precision: 0.8514
- Test set Recall: 0.8507
- Root Mean Square Error: 0.34961598

Confusion matrix is as follows:

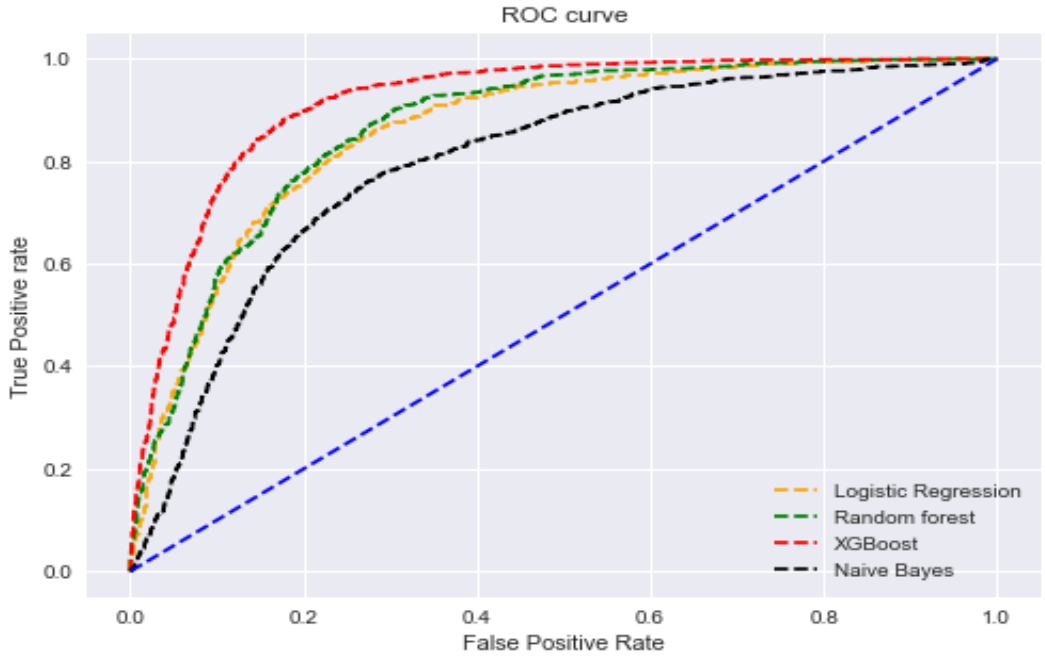


In conclusion, the XGBoost model demonstrates strong performance on both the training and test sets.

3.4.5 Model comparision

The below table shows the comparision of all the models.

models	Training Accuracy	Test Accuracy	RMSE	F1-Score
LogisticRegression	0.7581	0.7483	0.4917	0.7479
RandomForest	0.8048	0.7901	0.4418	0.8047
NaiveBayesClassifier	0.7465	0.7592	0.5034	0.7590
XGBoostClassifier	0.8778	0.8507	0.3496	0.8507



Overall, the XGBoost classifier achieved the highest accuracy and F1-score on both the training and test sets, indicating good generalization performance. Random Forest also performed well, followed by Logistic Regression and Naive Bayes Classifier.

3.4.6 Logistic Regression

Let the loan be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing, deposit, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.

loan	frequency
0	9215
1	1234

The data is imbalanced , applying smote function

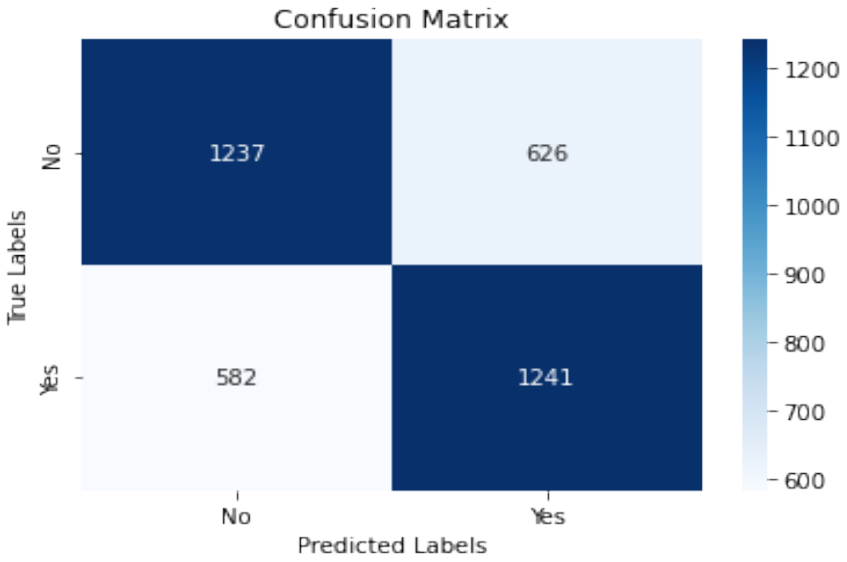
loan	frequency
0	9215
1	9215

The performance of the model is as follows:

- Training set Accuracy: 0.7581
- Training set F1 score: 0.757775
- Training set Precision: 0.758328
- Training set Recall: 0.758135

- Test set Accuracy: 0.7483
- Test set Fl score: 0.7479
- Test set Precision: 0.7491
- Test set Recall: 0.7483
- Root Mean Square Error: 0.4917976

Confusion matrix is as follows:



In conclusion, the Logistic Regression model shows decent performance on both the training and test sets, with the test set metrics being close to the training set metrics.

3.4.7 Random Forest Classifier

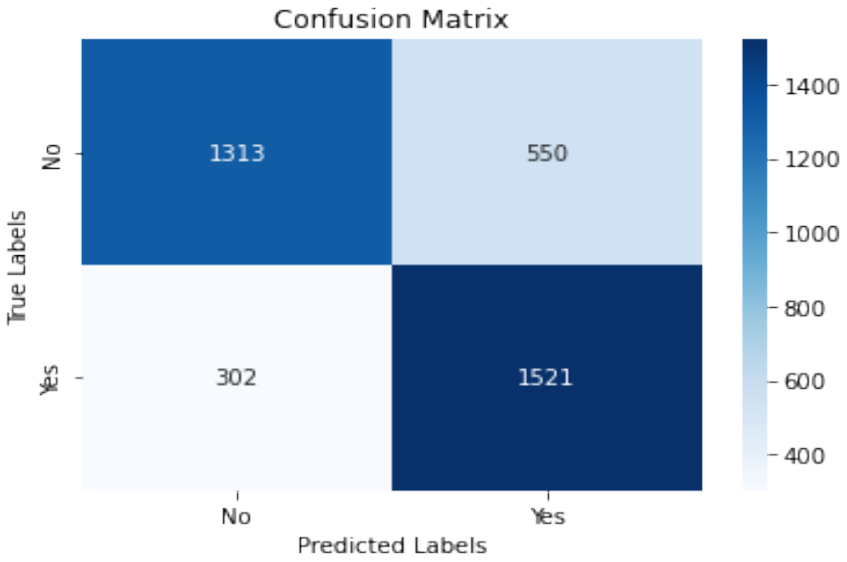
Let the loan be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing,deposit, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.

The performance of the model is as follows:

- Training set Accuracy: 0.8033
- Training set F1 score: 0.803300
- Training set Precision: 0.803457
- Training set Recall: 0.803254

- Test set Accuracy: 0.7907
- Test set Fl score: 0.7908
- Test set Precision: 0.7908
- Test set Recall: 0.7907
- Root Mean Square Error: 0.4435605

Confusion matrix is as follows:



In conclusion, the Random Forest model shows reasonably good performance on both the training and test sets, with the test set metrics being slightly lower than the training set metrics.

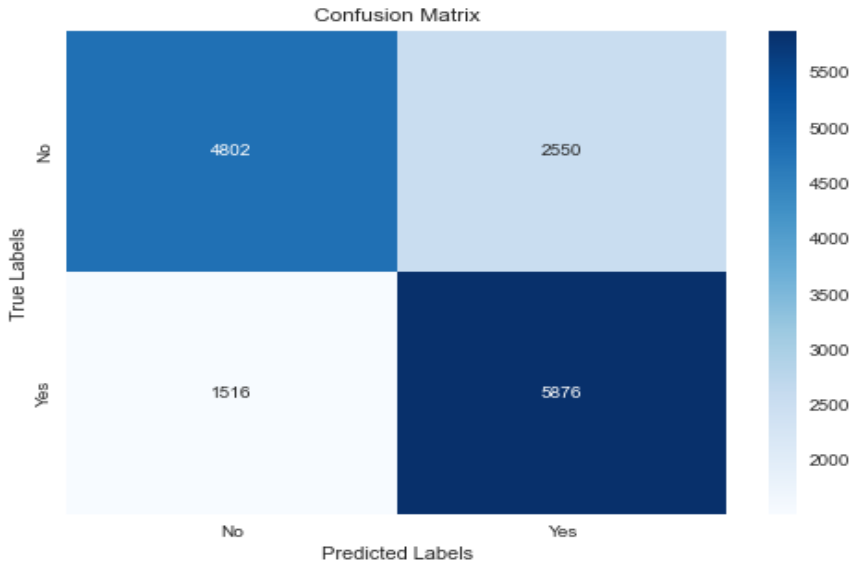
3.4.8 Naive Bayes Classifier

Let the loan be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing, deposit, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.
The performance of the model is as follows:

- Training set Accuracy: 0.7465
- Training set F1 score: 0.746399
- Training set Precision: 0.749444
- Training set Recall: 0.746514

- Test set Accuracy: 0.7592
- Test set Fl score: 0.7590
- Test set Precision: 0.7610
- Test set Recall: 0.7592
- Root Mean Square Error: 0.5034743

Confusion matrix is as follows:



In conclusion, the Naive Bayes model shows reasonably good performance on both the training and test sets, with the test set metrics being slightly higher than the training set metrics

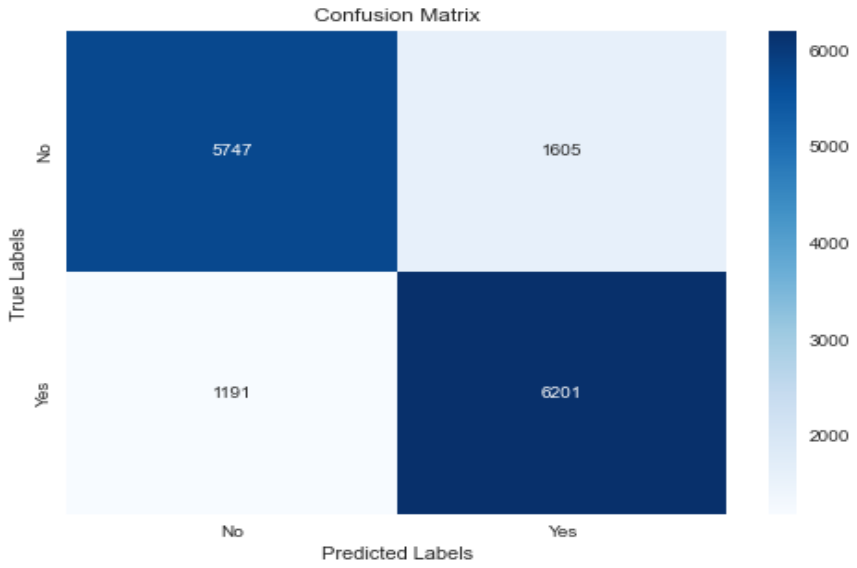
3.4.9 XG Boost Classifier

Let the loan be considered as the response variable. Here, the predictor variables age, job, marital, education, default, balance, housing, deposit, contact, day, month, duration,campaign, pdays,previous, poutcome of the campaign are taken into consideration to build the model.
The performance of the model is as follows:

- Training set Accuracy: 0.8574
- Training set F1 score: 0.857424
- Training set Precision: 0.858961
- Training set Recall: 0.857397

- Test set Accuracy: 0.8376
- Test set Fl score: 0.8376
- Test set Precision: 0.8384
- Test set Recall: 0.8376
- Root Mean Square Error: 0.3776284

Confusion matrix is as follows:

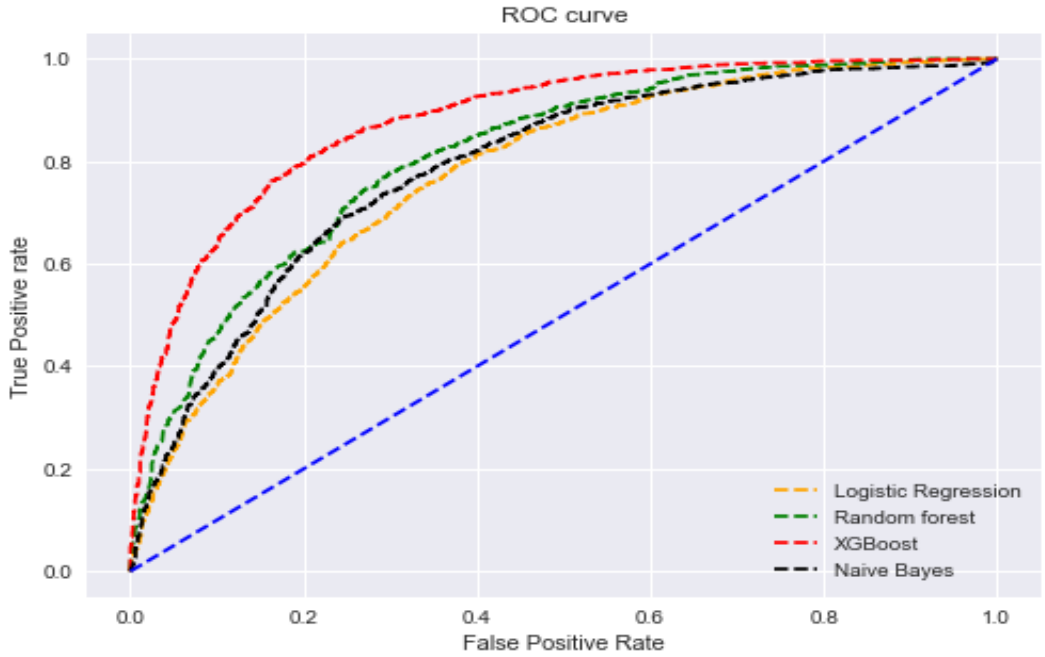


In conclusion, the XGBoost model demonstrates strong performance on both the training and test sets

3.4.10 Model comparision

The below table shows the comparision of all the models.

models	Training Accuracy	Test Accuracy	RMSE	F1-Score
LogisticRegression	0.7581	0.7483	0.4917	0.7577
RandomForest	0.8033	0.7907	0.4435	0.7908
NaiveBayesClassifier	0.7465	0.7592	0.5034	0.746399
XGBoostClassifier	0.8574	0.8376	0.3776	0.857424



Overall, the XGBoost classifier achieved the highest accuracy and F1-score on both the training and test sets, indicating good generalization performance. Random Forest also performed well, followed by Logistic Regression and Naive Bayes Classifier.

4 Chapter 4

4.1 Conclusion and Summary

4.1.1 Conclusion

Here are the conclusions from the project

- The majority of the bank's clients are married, followed by single clients and those who are divorced.
- The largest group of bank customers has completed secondary education .The second largest group comprises individuals with tertiary education, such as college or university degrees, while a small percentage falls under the "Unknown" category with missing or unrecorded education information.
- The vast majority of the bank's customers do not have any credit in default, indicating a good credit history for most clients.More than half of the customers do not have a housing loan, and the majority of customers have not taken out a personal loan.
- May has the highest frequency of client contacts, suggesting active outreach and engagement during that month. June, July, and August also receive moderate levels of client contacts.
- The majority of outcomes from the previous marketing campaign are classified as "Unknown," indicating that specific outcomes for a significant portion of customers couldn't be determined, while a relatively small portion of outcomes were classified as "Success," indicating positive responses and subscriptions to offered products or services.
- There is a significant difference in mean bank balance between customers who open a term deposit and those who don't.But there is no significant difference in mean age between customers who open a term deposit and those who don't. and also there is a significant difference in mean bank balance between customers with a personal loan and those without.
- There is a weak association between deposit and marital status, a moderate association between deposit and contacts performed, and a relatively strong association between deposit and previous outcome.There is a weak association between personal loan and education level, and a weak association between personal loan and marital status.

- In predicting whether the blank clients will subscribe to a term deposit or not, the XGBoost classifier achieved the highest accuracy and F1-score on both training and test sets, followed by Random Forest, Logistic Regression, and Naive Bayes Classifier.
- In predicting whether the blank clients will have personal loan or not, the XGBoost classifier achieved the highest accuracy and F1-score on both training and test sets, followed by Random Forest, Logistic Regression, and Naive Bayes Classifier.

4.1.2 Summary

The "Bank Customer Term Deposit Subscription Prediction" project aimed to assess the effectiveness of a financial institution's marketing campaign and predict the likelihood of bank clients subscribing to a term deposit and taking a personal loan. Using a dataset with 11,162 rows and 17 columns from Kaggle, the project conducted univariate and exploratory data analysis to understand customer demographics and behaviors.

The analysis revealed that the majority of the bank's clients were married, followed by single and divorced individuals. Moreover, most clients had completed secondary education, with a significant number holding tertiary degrees. Additionally, the dataset indicated that the majority of customers had a positive credit history, and only a small portion had credit in default. Furthermore, a considerable number of clients did not have housing or personal loans.

Evaluating the marketing campaign, it was observed that the bank made the most client contacts in May, suggesting active outreach during that month. Notably, a significant portion of the previous campaign outcomes were labeled as "Unknown," indicating that the bank couldn't determine the specific response of many customers. Meanwhile, smaller proportions were classified as "Failure" and "Success," providing insights into customer responses.

The project also explored associations and differences related to term deposit subscriptions and personal loans. It discovered a significant difference in mean bank balance between customers who opened a term deposit and those who didn't, while no significant difference in mean age was found. Weak associations were observed between term deposit subscriptions and marital status, as well as between personal loans and education level and marital status.

To predict customer behavior, the project built predictive models using Logistic Regression, Random Forest, Naive Bayes, and XGBoost classifiers. The results showed that the XGBoost classifier performed exceptionally well, achieving the highest accuracy and F1-score on both training and test sets for predicting both term deposit subscription and personal loan uptake.

In conclusion, the project provided valuable insights into customer behavior, preferences, and the effectiveness of the marketing campaign. By leveraging machine learning algorithms and predictive modeling, the financial institution can identify potential clients more likely to subscribe to term deposits and take personal loans, leading to improved future marketing strategies and targeted efforts. These findings pave the way for enhancing customer engagement and optimizing product uptake in future marketing campaigns for the bank.

5 Chapter 5

5.1 References

1. Elsalamony, H.A. (2014) 'Bank direct marketing analysis of data mining techniques',International Journal of Computer Applications, 85(7), pp.12-22.

2. Gregory, P. (2014) 'CRISP-DM, still the top methodology for analytics, data mining, or data science projects', [Online] Available:https://www.kdnuggets.com /2014/10/crisp-dmtopmethodology-analytics-data-mining-data-scienceprojects.html.

3. Mihova, Yana. (2019). Knowledge creation in banking marketing using machine learning techniques. 10.13140/RG.2.2.31756.16007

4. Akshansh Sharma et all, May(2020). Python: The Programming Language of Future, IJIRT — Volume 6 Issue 12 — ISSN: 2349-6002

5. https://towardsdatascience.com/anova-t-test-and-other-statistical-tests-with-python-e7a36a2fdc0c

6. https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

7. https://www.geeksforgeeks.org/machine-learning-with-python/

8. https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

9. https://www.javatpoint.com/machine-learning-random-forest-algorithm

10. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

11. https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

12. https://towardsdatascience.com/smote-fdce2f605729

13. https://www.geeksforgeeks.org/data-transformation-in-data-mining/

14. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

15. https://www.geeksforgeeks.org/auc-roc-curve/

6 Chapter 6

6.1 Appendix

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns

data=pd.read_csv(r"D:\SPOO\Documents\Project\bank.csv")
data.head()
data.shape
data.columns
data.describe()
data.shape

#check for missing values
data.isnull().sum()

#checking for outliers
cat=data.select_dtypes(include="O")
num=data.select_dtypes(exclude="O")

df_num = data[['age', 'balance', 'day', 'duration', 'campaign', 'pdays','previous'
col = ['age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous']
plt.figure(figsize=(15,18))
for i,v in enumerate(col):
    plt.subplot(4,2,i+1)
    sns.boxplot(x=v,data=df_num)
    plt.title("Boxplot of {}".format(v),size=20,color="black")
    plt.xlabel("{}".format(v),size=15)
plt.tight_layout()
plt.show()

#handling outliers
#1.negative balance,high balance removed
data.drop(data[(data['balance']>40000)|(data['balance']<0)].index,inplace=True,axi
```

```
#2.duration,campaign,previous
data.drop(data[data['duration']>3000].index,inplace=True,axis=0)
data.drop(data[data['campaign']>30].index,inplace=True,axis=0)
data.drop(data[data['previous']>30].index,inplace=True,axis=0)
data.drop(['pdays'],axis=1,inplace=True)

#univariate analysis
#EDA

marital_counts = data['marital'].value_counts()
colors = sns.color_palette('Set2', len(marital_counts))
plt.figure(figsize=(5, 5))
plt.pie(marital_counts, labels=marital_counts.index, colors=colors, autopct='%1.1f
plt.title('Distribution of Marital Status')
plt.axis('equal')
plt.show()

education_counts = data['education'].value_counts()
total_count = education_counts.sum()
percentages = (education_counts / total_count) * 100
colors = sns.color_palette('Set2', len(education_counts))
plt.figure(figsize=(12, 6))
plt.bar(education_counts.index, education_counts.values, color=colors)
for i, count in enumerate(education_counts):
    plt.text(i, count + 20, f'{count}', ha='left')
for i, percent in enumerate(percentages):
    plt.text(i, count + 60, f'{percent:.1f}%', ha='right')
plt.xlabel('Education')
plt.ylabel('Count')
plt.title('Distribution of Education Categories')
plt.xticks(rotation=45)
plt.show()

credit_default_counts = data['default'].value_counts()
housing_loan_counts = data['housing'].value_counts()
personal_loan_counts = data['loan'].value_counts()
credit_default_percentages = credit_default_counts / credit_default_counts.sum() *
housing_loan_percentages = housing_loan_counts / housing_loan_counts.sum() * 100
```

```
personal_loan_percentages = personal_loan_counts / personal_loan_counts.sum() * 10
colors_default = ['#FF7F0E', '#1F77B4']
colors_housing = ['#D62728', '#2CA02C']
colors_personal = ['#9467BD', '#FFBB78']
fig, ax = plt.subplots(1, 3, figsize=(15, 6))
ax[0].pie(credit_default_counts, labels=[f'{k} ({v:.1f}%)' for k, v in zip(credit_
ax[0].set_title('Credit in Default')
ax[1].pie(housing_loan_counts, labels=[f'{k} ({v:.1f}%)' for k, v in zip(housing_l
ax[1].set_title('Housing Loan')
ax[2].pie(personal_loan_counts, labels=[f'{k} ({v:.1f}%)' for k, v in zip(personal
ax[2].set_title('Personal Loan')
fig.suptitle('Distribution of Credit Default, Housing Loan, and Personal Loan')
ax[0].axis('equal')
ax[1].axis('equal')
ax[2].axis('equal')
plt.tight_layout()
plt.show()

month_counts = data['month'].value_counts()
plt.figure(figsize=(10, 6))
plt.bar(month_counts.index, month_counts.values)
for i, count in enumerate(month_counts):
    plt.text(i, count + 20, f'{count}', ha='center')
plt.xlabel('Month')
plt.ylabel('Count')
plt.title('Counts of Customers by Month')
plt.xticks(rotation=45)
plt.show()

poutcome_counts = data['poutcome'].value_counts()
percentages = (poutcome_counts / poutcome_counts.sum()) * 100
colors = ['#1F77B4', '#FF7F0E', '#2CA02C', '#D62728']
plt.figure(figsize=(10, 6))
plt.bar(percentages.index, percentages.values, color=colors)
for i, percent in enumerate(percentages):
    plt.text(i, percent + 1, f'{percent:.1f}%', ha='center')
plt.xlabel('Outcome')
plt.ylabel('Percentage')
```

```
plt.title('Percentage of Customers by Outcome')
plt.xticks(rotation=45)
plt.show()

#statistical analysis
#bivariate analysis
data['duration'].hist()
data['age'].hist()
data['balance'].hist()

from scipy.stats import boxcox
tdata,lambda_val=boxcox(data['duration'])
data['duration']=tdata
data['duration'].hist()

tdata,lambda_val=boxcox(data['age'])
data['age']=tdata
data['age'].hist()

from scipy.stats import yeojohnson
tdata,lambda_val=yeojohnson(data['balance'])
data['balance']=tdata
data['balance'].hist()

import statsmodels.graphics.gofplots as sm
sm.qqplot(data["age"],line="s")
plt.title(" Q-Q plot represent the distribution of age")
import statsmodels.graphics.gofplots as sm
sm.qqplot(data["balance"],line="s")
plt.title(" Q-Q plot represent the distribution of balance")
import statsmodels.graphics.gofplots as sm
sm.qqplot(data["duration"],line="s")
plt.title(" Q-Q plot represent the distribution of duration")

#statistical tests
#Testing the significant difference between the deposit and balance
from scipy.stats import ttest_ind
accepted_balance = data[data['deposit'] == 'yes']['balance']
```

```
rejected_balance = data[data['deposit'] =='no']['balance']
t_statistic, p_value = ttest_ind(accepted_balance, rejected_balance)
print(t_statistic, p_value )
if p_value < 0.05:
    print("There is a significant difference in mean bank balance between the cust
else:
    print("There is no significant difference in mean bank balance between the cus

#Testing the association between deposit and marital status

from scipy.stats import chi2_contingency
education_marital_crosstab = pd.crosstab(data['deposit'], data['marital'])
chi2_stat, p_value, _, _ = chi2_contingency(education_marital_crosstab)
print(p_value)
print(education_marital_crosstab)
if p_value < 0.05:
    print("There is a significant relationship between deposit and marital status.
else:
    print("There is no significant relationship between deposit and marital status

import scipy.stats as ss

def cramers_v(x, y):
    confusion_matrix = pd.crosstab(x, y)
    chi2 = ss.chi2_contingency(confusion_matrix)[0]
    n = confusion_matrix.sum().sum()
    phi2 = chi2 / n
    r, k = confusion_matrix.shape
    phi2corr = max(0, phi2 - ((k-1)*(r-1)) / (n-1))
    r_corr = r - ((r-1)**2) / (n-1)
    k_corr = k - ((k-1)**2) / (n-1)
    return np.sqrt(phi2corr / min((k_corr-1), (r_corr-1)))

# Example usage:
variable1 = data['deposit']
variable2 = data['marital']

cramer_v = cramers_v(variable1, variable2)
```

```
print("Cramer's V:", cramer_v)

#model building
#Label Encoding
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
data['job']=le.fit_transform(data['job'])
data['marital']=le.fit_transform(data['marital'])
data['education']=le.fit_transform(data['education'])
data['housing']=le.fit_transform(data['housing'])
data['loan']=le.fit_transform(data['loan'])
data['contact']=le.fit_transform(data['contact'])
data['month']=le.fit_transform(data['month'])
data['poutcome']=le.fit_transform(data['poutcome'])
data['deposit']=le.fit_transform(data['deposit'])
data['default']=le.fit_transform(data['default'])

#scaling
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
scaled_features=scaler.fit_transform(data)

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
import math
from sklearn.metrics import accuracy_score , classification_report, ConfusionMatri

#logistic regression
#target is deposit

#split data
from sklearn.model_selection import train_test_split
X = data.drop('deposit', axis = 1)
y = data['deposit']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)

from sklearn.linear_model import LogisticRegression
model1= LogisticRegression()
model1.fit(X_train,y_train)

y_train_pred = model1.predict(X_train)
y_test_pred =  model1.predict(X_test)
# Test set performance
model_test_accuracy = accuracy_score(y_test, y_test_pred)
model_test_f1 = f1_score(y_test, y_test_pred, average='weighted')
model_test_precision = precision_score(y_test, y_test_pred , average='weighted')
model_test_recall  = recall_score(y_test, y_test_pred,average='weighted')

  # Training set performance
model_train_accuracy = accuracy_score(y_train, y_train_pred)
model_train_f1 = f1_score(y_train, y_train_pred, average= 'weighted')
model_train_precision = precision_score(y_train, y_train_pred,average='weighted')
model_train_recall = recall_score(y_train, y_train_pred,average='weighted')
print('Model performance for Training set')
print("- Accuracy: {:.4f}".format(model_train_accuracy))
print('- F1 score: {:4f}'.format(model_train_f1))
print('- Precision: {:4f}'.format(model_train_precision))
print('- Recall: {:4f}'.format(model_train_recall))

print('Model performance for Test set')
print('- Accuracy: {:.4f}'.format(model_test_accuracy) )
print('- Fl score: {:.4f}'.format(model_test_f1))
print('- Precision: {:.4f}'.format(model_test_precision))
print('- Recall: {:.4f}'.format(model_test_recall))
from sklearn.metrics import mean_squared_error
# Calculate MSE
mse = mean_squared_error(y_train,y_train_pred)
```

```
# Calculate RMSE
rmse = np.sqrt(mse)

print("Root Mean Square Error (RMSE):", rmse)

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_test_pred)
class_names = ['No', 'Yes']
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=class_names, ytickl
# Add labels, title, and axis ticks
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.title('Confusion Matrix')
plt.show()

#Random forest
import pandas as pd
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Define the parameter grid for grid search
param_grid = {
    'n_estimators': [50, 70, 90],
    'max_depth': [1, 2, 3],
    'min_samples_split': [5, 10, 15],
    'min_samples_leaf': [1, 2, 4]
}

# Create the Random Forest classifier
classifier = RandomForestClassifier(random_state=42)

#Perform grid search
grid_search = RandomizedSearchCV(classifier, param_grid, cv=5)
grid_search.fit(X_train,y_train)

# Use the best model for prediction
```

```
best_model_rf = grid_search.best_estimator_

y_train_pred = best_model_rf.predict(X_train)
y_test_pred =  best_model_rf.predict(X_test)
# Test set performance
model_test_accuracy = accuracy_score(y_test, y_test_pred)
model_test_f1 = f1_score(y_test, y_test_pred, average='weighted')
model_test_precision = precision_score(y_test, y_test_pred , average='weighted')
model_test_recall  = recall_score(y_test, y_test_pred,average='weighted')

  # Training set performance
model_train_accuracy = accuracy_score(y_train, y_train_pred)
model_train_f1 = f1_score(y_train, y_train_pred, average= 'weighted')
model_train_precision = precision_score(y_train, y_train_pred,average='weighted')
model_train_recall = recall_score(y_train, y_train_pred,average='weighted')
print('Model performance for Training set')
print("- Accuracy: {:.4f}".format(model_train_accuracy))
print('- F1 score: {:4f}'.format(model_train_f1))
print('- Precision: {:4f}'.format(model_train_precision))
print('- Recall: {:4f}'.format(model_train_recall))

print('Model performance for Test set')
print('- Accuracy: {:.4f}'.format(model_test_accuracy) )
print('- Fl score: {:.4f}'.format(model_test_f1))
print('- Precision: {:.4f}'.format(model_test_precision))
print('- Recall: {:.4f}'.format(model_test_recall))
mse = mean_squared_error(y_train,y_train_pred)

# Calculate RMSE
rmse = np.sqrt(mse)

print("Root Mean Square Error (RMSE):", rmse)

#naive bayes
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, classification_report

# Create an instance of the Naive Bayes classifier
```

```
classifier = GaussianNB()

# Train the classifier
classifier.fit(X_train, y_train)
y_train_pred = classifier.predict(X_train)
y_test_pred =  classifier.predict(X_test)
# Test set performance
model_test_accuracy = accuracy_score(y_test, y_test_pred)
model_test_f1 = f1_score(y_test, y_test_pred, average='weighted')
model_test_precision = precision_score(y_test, y_test_pred , average='weighted')
model_test_recall  = recall_score(y_test, y_test_pred,average='weighted')

  # Training set performance
model_train_accuracy = accuracy_score(y_train, y_train_pred)
model_train_f1 = f1_score(y_train, y_train_pred, average= 'weighted')
model_train_precision = precision_score(y_train, y_train_pred,average='weighted')
model_train_recall = recall_score(y_train, y_train_pred,average='weighted')
print('Model performance for Training set')
print("- Accuracy: {:.4f}".format(model_train_accuracy))
print('- F1 score: {:4f}'.format(model_train_f1))
print('- Precision: {:4f}'.format(model_train_precision))
print('- Recall: {:4f}'.format(model_train_recall))

print('Model performance for Test set')
print('- Accuracy: {:.4f}'.format(model_test_accuracy) )
print('- Fl score: {:.4f}'.format(model_test_f1))
print('- Precision: {:.4f}'.format(model_test_precision))
print('- Recall: {:.4f}'.format(model_test_recall))
mse = mean_squared_error(y_train,y_train_pred)

# Calculate RMSE
rmse = np.sqrt(mse)

print("Root Mean Square Error (RMSE):", rmse)
import xgboost as xgb
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV
```

```
# Define the parameter grid for hyperparameter tuning
param_grid = {
    'learning_rate': [0.1, 0.01, 0.001],
    'max_depth': [1,2,3],
    'n_estimators': [100, 200, 300]
}

# Perform grid search for hyperparameter tuning
grid_search = GridSearchCV(model, param_grid, scoring='accuracy', cv=5)
grid_search.fit(X_train, y_train)

# Get the best parameters and the best model
best_params = grid_search.best_params_
best_model_xg = grid_search.best_estimator_

# Train the best model
best_model_xg.fit(X_train, y_train)


y_train_pred = best_model_xg.predict(X_train)
y_test_pred =  best_model_xg.predict(X_test)
# Test set performance
model_test_accuracy = accuracy_score(y_test, y_test_pred)
model_test_f1 = f1_score(y_test, y_test_pred, average='weighted')
model_test_precision = precision_score(y_test, y_test_pred , average='weighted')
model_test_recall  = recall_score(y_test, y_test_pred,average='weighted')

  # Training set performance
model_train_accuracy = accuracy_score(y_train, y_train_pred)
model_train_f1 = f1_score(y_train, y_train_pred, average= 'weighted')
model_train_precision = precision_score(y_train, y_train_pred,average='weighted')
model_train_recall = recall_score(y_train, y_train_pred,average='weighted')
print('Model performance for Training set')
print("- Accuracy: {:.4f}".format(model_train_accuracy))
print('- F1 score: {:4f}'.format(model_train_f1))
print('- Precision: {:4f}'.format(model_train_precision))
print('- Recall: {:4f}'.format(model_train_recall))
```

```
print('Model performance for Test set')
print('- Accuracy: {:.4f}'.format(model_test_accuracy) )
print('- Fl score: {:.4f}'.format(model_test_f1))
print('- Precision: {:.4f}'.format(model_test_precision))
print('- Recall: {:.4f}'.format(model_test_recall))
mse = mean_squared_error(y_train,y_train_pred)

# Calculate RMSE
rmse = np.sqrt(mse)

print("Root Mean Square Error (RMSE):", rmse)

# predict probabilities
pred_prob1 = model1.predict_proba(X_test)
pred_prob2 = best_model_rf.predict_proba(X_test)
pred_prob3 = best_model_xg.predict_proba(X_test)
pred_prob4 = classifier.predict_proba(X_test)

from sklearn.metrics import roc_curve

# roc curve for models
fpr1, tpr1, thresh1 = roc_curve(y_test, pred_prob1[:,1], pos_label=1)
fpr2, tpr2, thresh2 = roc_curve(y_test, pred_prob2[:,1], pos_label=1)
fpr3, tpr3, thresh3 = roc_curve(y_test, pred_prob3[:,1], pos_label=1)
fpr4, tpr4, thresh4 = roc_curve(y_test, pred_prob4[:,1], pos_label=1)
# roc curve for tpr = fpr
random_probs = [0 for i in range(len(y_test))]
p_fpr, p_tpr, _ = roc_curve(y_test, random_probs, pos_label=1)

from sklearn.metrics import roc_auc_score

# auc scores
auc_score1 = roc_auc_score(y_test, pred_prob1[:,1])
auc_score2 = roc_auc_score(y_test, pred_prob2[:,1])
auc_score3 = roc_auc_score(y_test, pred_prob3[:,1])
auc_score4 = roc_auc_score(y_test, pred_prob4[:,1])
# matplotlib
import matplotlib.pyplot as plt
```

```
plt.style.use('seaborn')

# plot roc curves
plt.plot(fpr1, tpr1, linestyle='--',color='orange', label='Logistic Regression')
plt.plot(fpr2, tpr2, linestyle='--',color='green', label='Random forest')
plt.plot(fpr3, tpr3, linestyle='--',color='red', label='XGBoost')
plt.plot(fpr4, tpr4, linestyle='--',color='black', label='Naive Bayes')
plt.plot(p_fpr, p_tpr, linestyle='--', color='blue')
# title
plt.title('ROC curve')
# x label
plt.xlabel('False Positive Rate')
# y label
plt.ylabel('True Positive rate')

plt.legend(loc='best')
plt.savefig('ROC',dpi=300)
plt.show();
```