Bank Telemarketing

STAT – 530 PROJECT

Faiyaz Ahmad, Kruthika Gopinathan, Rutuja Magdum

Contributions:

1. Faiyaz Ahmad: EDA,Data Preprocessing, Feature Selection, Logistic Regression, Decision Tree, Report
2. Kruthika Gopinathan: EDA, Logistic Regression, Model Evaluation and Analysis, Report
3. Rutuja Magdum: EDA, Hypothesis Testing(Chi-square test), Report

1. Introduction

Marketing is one of the core strengths of any company operating in highly competitive markets in the current times. In today's fast-paced world, banks are constantly looking for new and innovative ways to reach out to their customers. Telemarketing is a common strategy for banks ,because bank products and services might be difficult for some people to understand. It is easier for customers or target users to grasp products or services when they are explained directly. It allows users to ask questions, making it easier for them to understand the products or services being offered. All banks care about their reputation and branding and need to identify target users who will not buy products or services through telemarketing to protect their reputation. However, the success of a bank telemarketing operation is dependent on a variety of factors, including the telemarketing team's quality, the accuracy of customer data, the effectiveness of the sales pitch, and overall customer experience. As a result, developing a comprehensive strategy is vital to the success of a bank telemarketing campaign.

This approach includes numerous elements, such as defining the target demographic, developing a tailored sales pitch, training the telemarketing workforce effectively, and evaluating the outcomes. With the help of effective bank marketing banks can stand out from their competitors by offering unique products, services, and experiences to their customers. A well-executed bank telemarketing project can bring significant benefits to a bank's customer acquisition and retention efforts, generate revenue, and help the bank stay competitive in the constantly evolving financial industry.

Example: Citibank utilized telemarketing to advertise its credit card offers to existing and prospective clients. Citibank's telemarketing workforce has been trained to provide customized product recommendations depending on the customer's purchasing habits and credit history. As a result, Citibank's telemarketing operations assisted the corporation in growing its client base, increasing sales, and generating substantial money.Banks can use telemarketing to expand their customer base, increase sales, and generate significant revenue. Banks, on the other hand, must carefully design their telemarketing strategy, ensure compliance with local norms and regulations, and train their telemarketing workers to deliver an excellent client experience.

2. Literature Review

The advent of toll-free phone numbers in the 1970s allowed clients to contact banks for free and inquire about their products and services, which was the first example of bank telemarketing. Banks began utilizing outbound telemarketing to promote credit cards and other financial goods to their clients in the 1980s. These attempts, however, were frequently met with opposition from customers who found telemarketing calls intrusive and disruptive. In the 1990s, banks began to focus more on inbound telemarketing, where customers would call in to inquire about products and services, rather than outbound telemarketing. Technology improvements enabled banks to acquire and analyze consumer data more efficiently, allowing them to better understand their customers' requirements and preferences.

Banks nowadays employ a range of telemarketing tactics to reach out to clients and produce leads, such as cold calling, warm calling, and lead nurturing. They also employ telemarketing as a customer support technique, allowing customers to contact them directly to address problems or obtain answers to queries.Several reasons have contributed to bank telemarketing's growth and success. The increasing competitiveness in the banking business, as banks attempt to recruit and maintain customers in a highly competitive market, is one of the key factors. As a result, many banks have turned to telemarketing to reach out to customers and provide a variety of financial products and services.

An additional driving factor for the growth of bank telemarketing is technological advancement, which has made it easier and less expensive for banks to reach out to their customers via telemarketing. Banks may now reach out to a large number of clients quickly and efficiently thanks to the availability of modern call center software and automated dialing technologies. Furthermore, customer demand for individualized and tailored financial goods and services has aided the expansion of bank telemarketing. Customers are increasingly seeking more personalized and customized financial goods and services that are tailored to their unique needs and preferences, and telemarketing enables banks to provide bespoke solutions to their customers. Overall, bank telemarketing has advanced significantly since its inception, and it remains a vital sales and customer service approach for banks of all sizes.

3. Data Source

Dataset

The Bank Marketing data set is a public dataset that comprises information from a Portuguese banking institution's direct telemarketing operations. The dataset contains information about clients, their demographic and economic features, and the outcomes of prior marketing efforts, as well as information regarding the current campaign's outcomes, such as whether or not the customer subscribed to the bank's term deposit. There are 41,188 occurrences in the dataset, with 20 input variables and one binary output variable indicating whether or not the client subscribed to the bank's term deposit. Information such as the customer's age, job, marital status, education level, home loan, personal loan, and previous marketing campaign results are among the input variables.

The Bank Marketing dataset has been widely used in the banking and financial industry for machine learning and data analysis research. It has been used to predict whether or not a customer will subscribe to a term deposit based on demographic and economic characteristics, as well as to identify which elements have the greatest influence on a customer's decision to subscribe. Overall, the Bank Marketing dataset is a

significant resource for banking and finance scholars and practitioners interested in studying customer behavior and increasing the effectiveness of bank telemarketing efforts.

There have been many research studies conducted using the Bank Telemarketing dataset. Following are a few examples of research findings:

- A study by Ahmed and Islam (2019) used the Bank Telemarketing dataset to develop a predictive model for identifying customers who are likely to subscribe to a term deposit. The study found that logistic regression was the most effective algorithm for predicting customer behavior.

- A study by Ferreira et al. (2019) used the Bank Telemarketing dataset to explore the impact of feature selection on the performance of machine learning models. The study found that feature selection improved the performance of some algorithms, but not all.

About the features:

In the dataset there are 10 Categorical Features and 7 Numerical Features.

- Demographic information: age, job, marital, education
- Financial Health: balance, default, housing, loan, campaign, pdays, previous
- Time Characteristics: day, month, duration, campaign, pdays, and previous
- Characteristics of Campaign: campaign, pdays, and previous

Description:

VARIABLE	DESCRIPTION
Age	Age of a person
Job	Job type of a person (entrepreneur, student, blue-collar etc)
Marital	Marital status of a person (single, married, divorced)
Education	Education level of a person (primary, secondary etc)
Default	Indicates if a person has credit in default
Balance	Average yearly balance
Housing, Personal	Indicates if a person has a housing or personal loan respectively (y/no)
Contact	Communication type (cellular,telephone)
Day	Last contact day of the month
Month	Last contact month

Duration	Last contact duration in seconds
Campaign	Number of contacts performed during the campaign and for a client
pdays	Number of days that passed by after the client was last contacted from a previous campaign
previous	Number of contacts performed before this campaign and for this client
poutcome	Outcome of the previous campaign (failure, success)
y	Response variable - Indicates if client has subscribed to a term deposit)

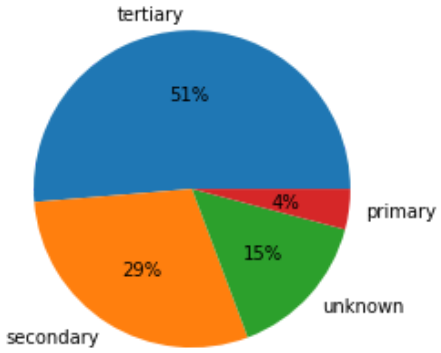
4. Data Statistics and Exploratory Data Analysis

SUMMARY TABLE

	age (Years)	balance (Euro)	day	duration (Seconds)	campaign (Nos)	pdays (days)	previous (days)
Minimum	18	-8019	1	0	1	-1	0
1st Quartile	33	72	8	103	1	-1	0
Median	39	448	16	180	2	-1	0
3rd Quartile	48	1428	21	319	3	-1	0
Maximum	95	102127	31	4918	63	871	275
Mean	40.9	1362.3	15.8	258.2	2.8	40.2	0.6
Std Deviation	10.6	3044.8	8.3	257.5	3.1	100.1	2.3

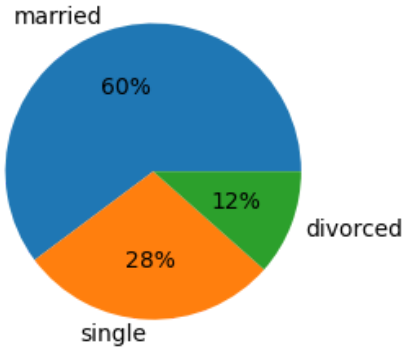
The above table shows the summary of the distribution of the continuous variables

DATA VISUALIZATION :



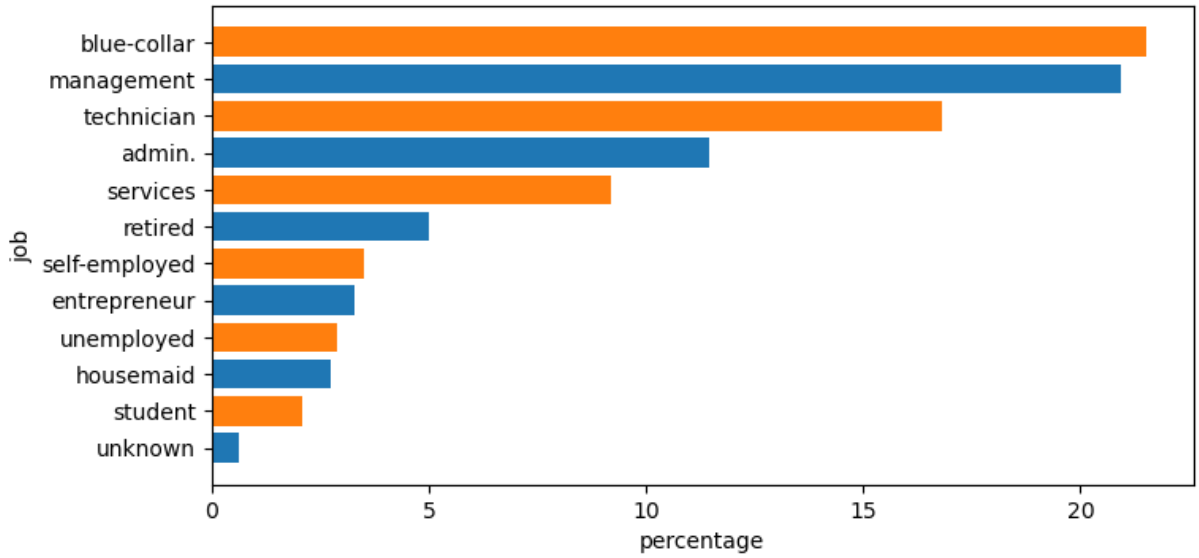
Pie chart for the feature: education

The above chart shows that maximum clients have tertiary type of education



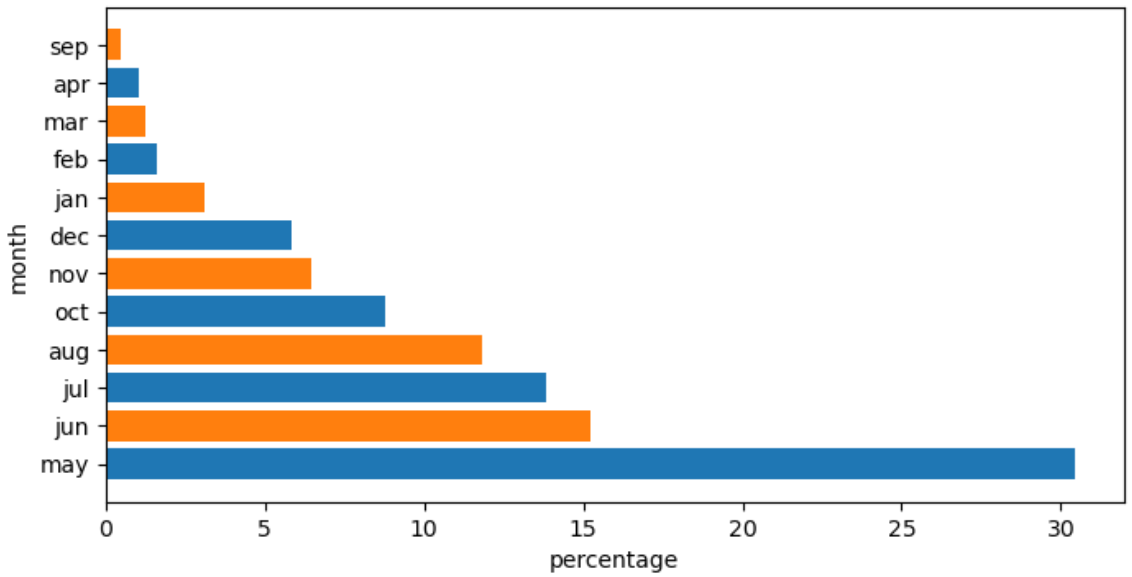
Pie chart for the feature: marital

The above chart shows that maximum clients are married



1

The above graph shows that maximum clients have blue-collar type of coworkers



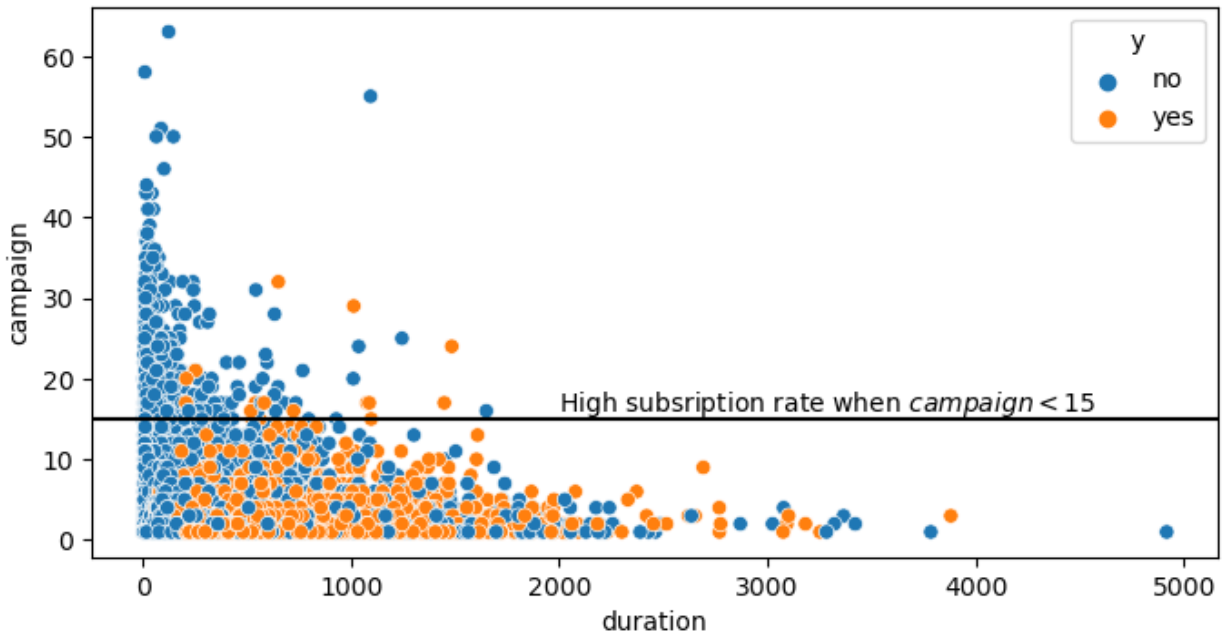
The above graph shows that maximum clients have been connected in the month of May

Features	Yes	No
Default	2%	98%
housing	56%	44%
loan	16%	84%
y	12%	88%

From the above table, we can see that

a) Only 2% of the clients have credit in default
b) 56% of the clients have housing loan ,and the rest 44% of the clients do not have a housing loan
c) 16% of the clients have a loan(personal or housing), whereas 84% of the clients do not have any loan
d) The response variable y can be interpreted as 12% of the clients subscribe to a term deposit and 88% of the clients do not subscribe. This shows that data is highly **imbalanced**.

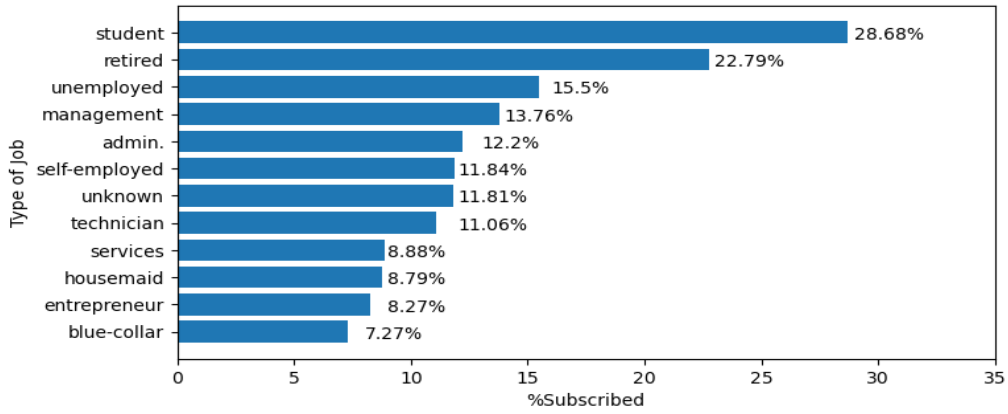
Relationship between Campaign vs Duration

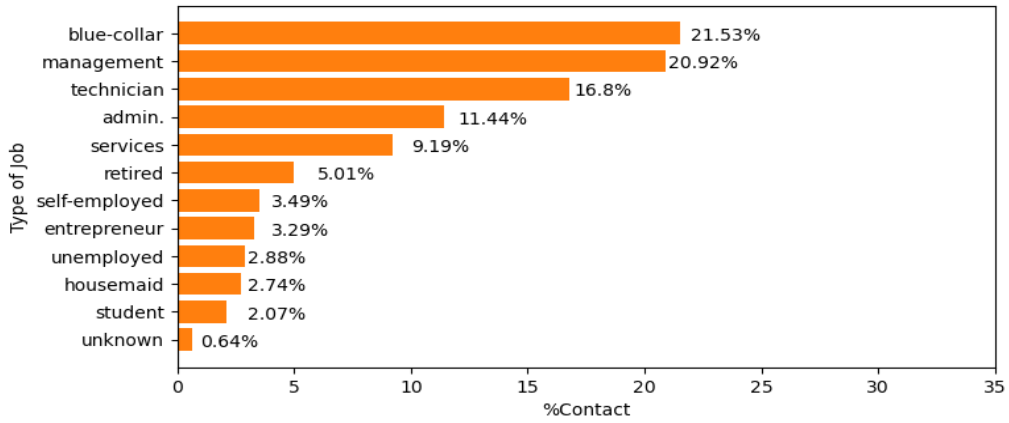


Observations:

a) When the number of campaigns increase, the duration decreases exponentially
b) Campaign is inversely proportional to the subscription rate
c) When the campaign just started then the average duration tends to be higher

Subscription rate vs type of jobs:

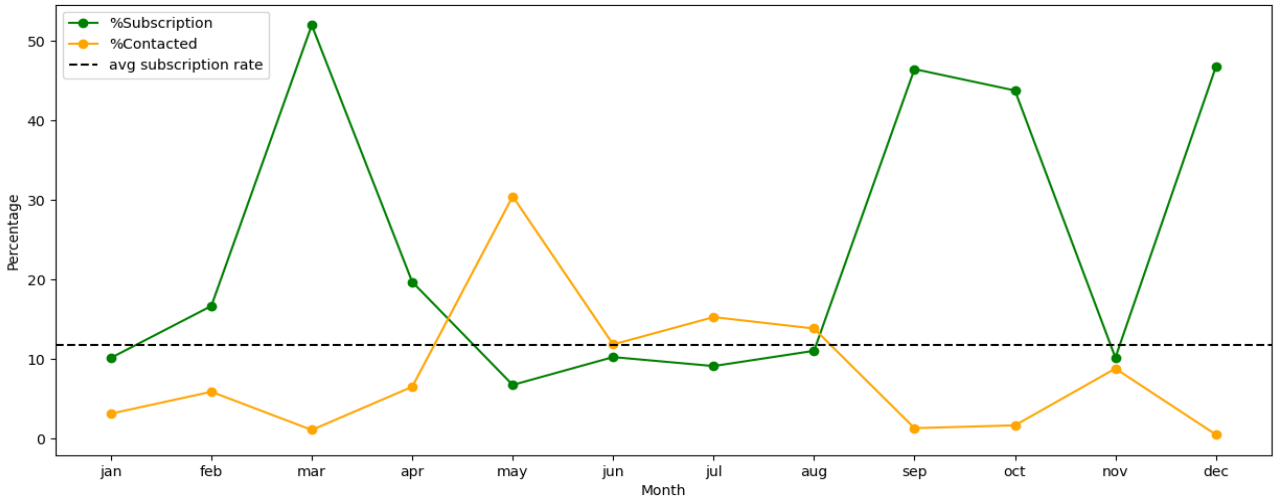




Observations:

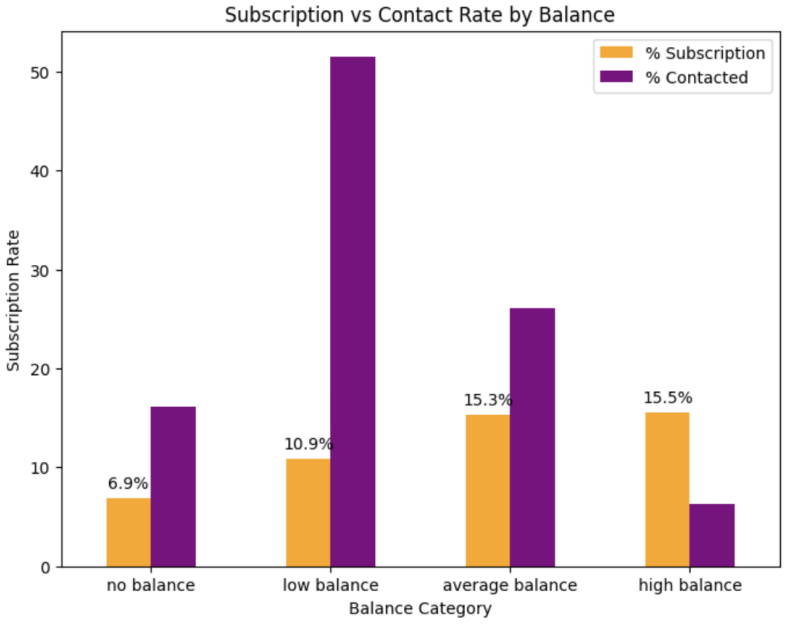
a) 50% of total subscription is contributed by management, technician, and blue collar and are contacted the most

Monthwise contact rate and subscription:



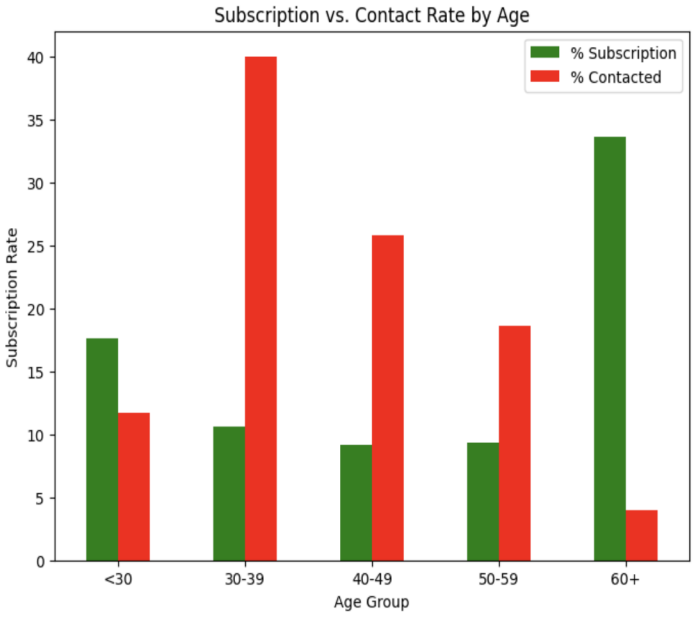
Observations:

a) Nearly 30% of the clients are contacted in the month of May
b) March and December had the highest number of subscriptions



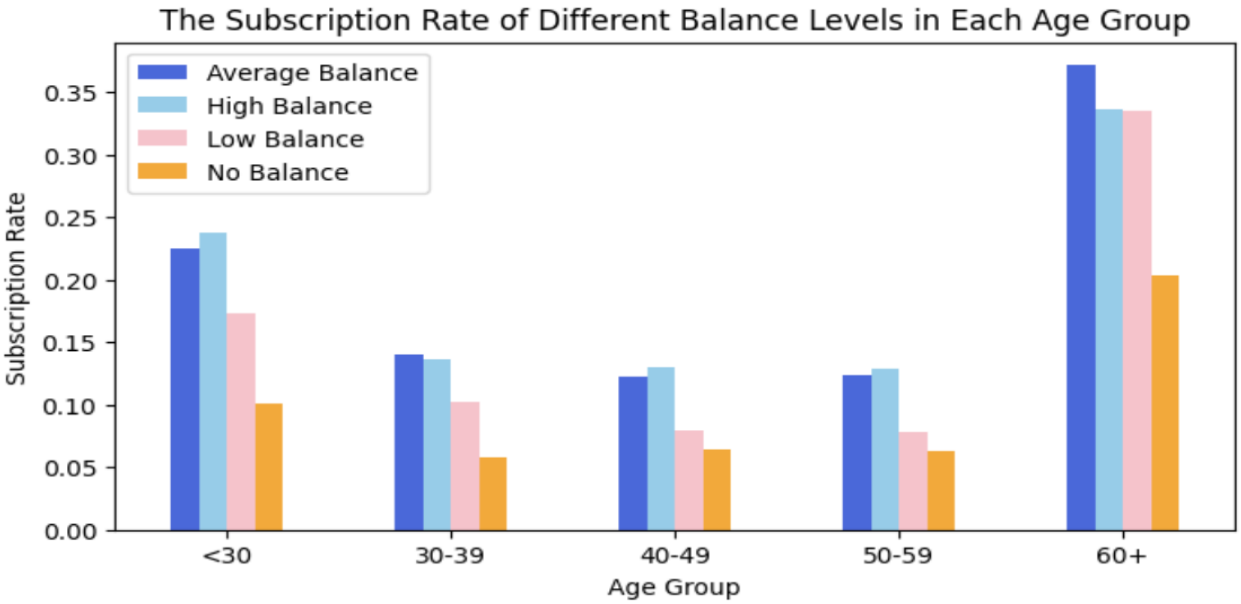
We can see that the young people are contacted more compared to the adults (60+)
The subscription rate is more in case of elderly people. It could possibly be because elderly are retired and it is not risky for them to invest. On the other hand, young people would not have enough balance to invest

Question : Does the demography of a person (age in this case) affect the subscription rate ?Do older people subscribe more ?



We can see that the people with more balance tend to subscribe more.
But the bank is trying to contact the clients with low balance , and their subscription rate is less.

Question: Does the financial health status of a client depend on their subscription ?



- People aged above 60 and those below 30 show a higher tendency to subscribe compared to other age groups, with the former exhibiting an exceptionally high willingness to subscribe (acceptance rate of 35%)
- The bank should concentrate on young customers who have a positive balance and have demonstrated a subscription rate ranging from 15% to 20%

5. Methodology

5.1. Data Preprocessing

5.1.1. Splitting data into training and testing set (80%-20% split)

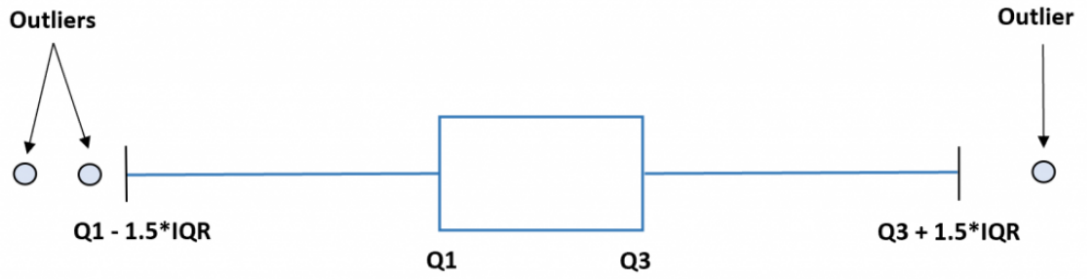
In machine learning, it is essential to prepare the data before building a model. Three important data preparation steps are splitting the data into training and testing sets, treating outliers, and converting categorical variables to factors.

Firstly, splitting the data into training and testing sets is crucial to build a robust and reliable model. The training set is used to train the model, while the testing set is used to evaluate the model's performance. This separation ensures that the model is not overfitting to the training data, and it can generalize well to new, and unseen data. A common practice is to split the data into 80% of the data used for training and the rest for testing.

5.1.2. Treating Outlier

Secondly, treating outliers is important in machine learning for several reasons. Outliers are data points that are significantly different from the rest of the data and can lead to inaccurate predictions. Outliers can also introduce biases into the data, impacting the results of our analysis. By identifying and treating outliers, we can improve the performance, understanding, and interpretability of our models. Methods to treat outliers include removing them, replacing them with mean or median values, or transforming the data. In our case, we are replacing the outliers with median values. In our case, we have chosen to replace the outliers with median values for the balance, campaign, and duration features. This approach is commonly used when the outliers are few in number and have a low impact on the overall data distribution. To implement this approach, we have used the interquartile range (IQR) to identify the outliers. The IQR is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data. If a data point's value falls beyond 1.5 times the IQR from Q1 or Q3, we have replaced the value

with the median. This helps to reduce the impact of outliers and ensures that the model is not skewed by extreme values. Overall, treating outliers in this manner can lead to more reliable and accurate machine learning models.



5.1.3. Converting the categorical variables as factor

Thirdly, converting categorical variables to factors is important in machine learning. Categorical variables are variables that represent categories or groups, such as age, job, or education. Factors are a way to represent categorical variables in a numerical format that can be used in machine learning models. Converting categorical variables to factors allows the model to use them in calculations, and it avoids the model treating them as continuous variables. This conversion also helps to reduce the dimensionality of the data, making it easier to process and analyze.

In conclusion, data preparation is an essential step in machine learning, and splitting the data into training and testing sets, treating outliers, and converting categorical variables to factors are three important steps to consider. By performing these steps, we can build a robust and reliable model that can generalize well to new, unseen data.

5.2. Feature Selection

Feature selection is a crucial step in machine learning that involves selecting the most relevant features to include in our models. In this case, two methods have been employed: the chi-square test of independence for categorical variables and correlation analysis for numerical variables.

5.2.1. Categorical Feature Selection

The chi-square test of independence is a statistical method used to determine if there is a significant association between two categorical variables. In this study, this method has been applied to identify which of the categorical variables have the strongest relationship with the target variable. The results of the chi-square test can be used to eliminate irrelevant features that do not contribute much to the predictive power of the model.

By using the chi-square test of independence, we are going to investigate the association between the credit default, personal loan, housing loan, education and type of jobs.

With regard to hypothesis:

Null Hypothesis: Both the features are independent.

Alternate Hypothesis: Both the features are associated with each other.

Chisquare test of Independence					
S.No.	**Variable 1**	**Variable 2**	**df**	**X-squared**	**P-value (<)**
1	default	y	1	15.758	7.20E-05
2	loan	y	1	167.47	2.20E-16
3	housing	y	1	714.1	2.20E-16
4	education	y	3	189.07	2.20E-16
5	job	y	11	706.09	2.20E-16
6	month	y	11	2410.8	2.20E-16
7	default	loan	1	206.4	2.20E-16
8	housing	loan	1	76.975	2.20E-16
9	job	education	33	23045	2.20E-16

From the above table , it is evident that the default, loan, housing,education, job, and month are associated with each other because the p-value for each case is less than 0.05 significance level. We also checked the association within dependent variables. It is found that credit default is associated with personal loan, housing loan is associated with personal loan, and type of job is associated with level of education. We are going to use housing, loan, job, and month features for model training.

Association between credit default and personal loan

```
        loan
default          no         yes         Sum
    no  0.828844507 0.153006823 0.981851330
    yes 0.011546642 0.006602028 0.018148670
    Sum 0.840391149 0.159608851 1.000000000

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  tab
X-squared = 206.4, df = 1, p-value < 2.2e-16
```

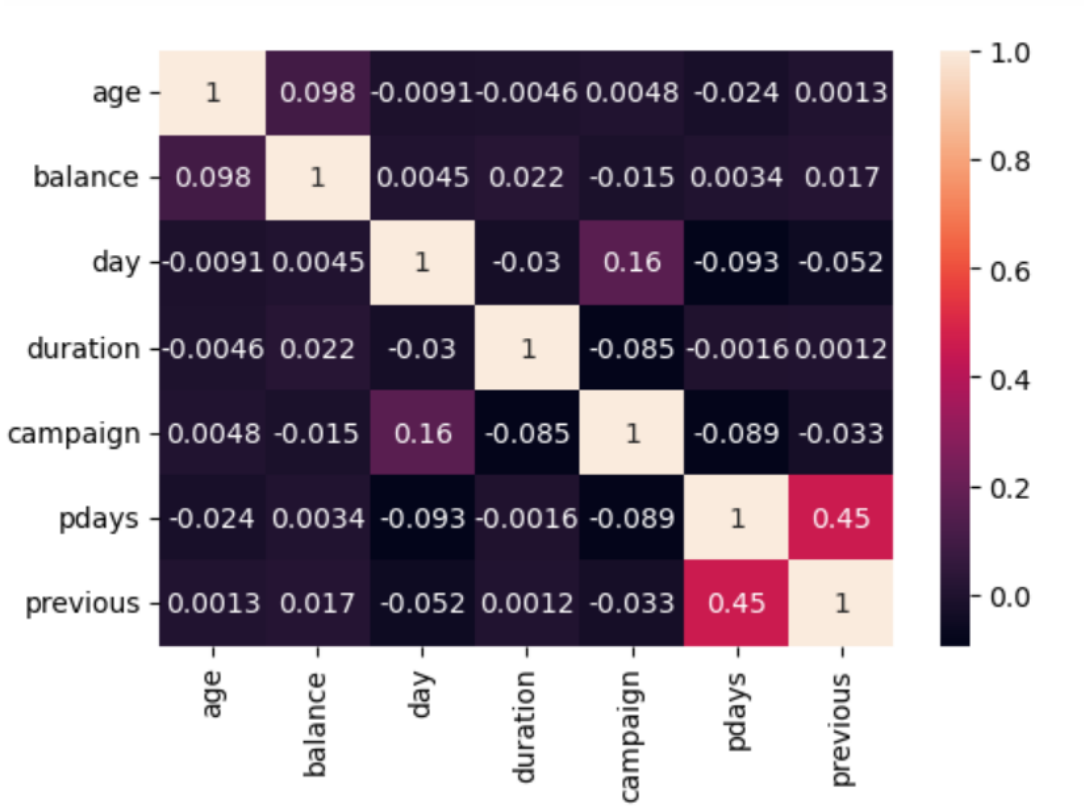
From above it is evident that, p-value is less than 0.05 significance level. We can reject the null hypothesis and conclude that credit default is associated with personal loan.

5.2.2. Numerical Feature selection

For numerical variables, correlation analysis has been used to identify the strength of the linear relationship between each numerical variable and the target variable. Correlation analysis measures the degree to which two variables are related, with a correlation coefficient ranging from -1 to 1. A high correlation coefficient indicates a strong positive relationship, while a low coefficient indicates a weak or negative relationship.

In this study, the correlation coefficients have been calculated using the Pearson correlation coefficient. Numerical features with a weak correlation with the target variable have been eliminated from the dataset, while those with a strong correlation have been retained for further analysis.

Overall, the feature selection process helps to identify the most important variables that contribute to the model's predictive power while eliminating irrelevant or redundant variables.



Above figure is correlation between the dependent variable. From the above chart, it is evident that most of the dependent variables are not strongly correlated with each other. Previous and pdays features are showing strong correlation having correlation coefficient of 0.45. Instead of using pdays and previous, we can use one of them for model training.

5.3. Model Training

In this section we are going to discuss the methods applied for training the model.

a) Logistic Regression and Decision Tree Classifier

Model training is a critical stage in machine learning, where we build and train models using the selected features to predict the target variable. In this study, two classification models have been trained and evaluated, namely Logistic Regression and Decision Tree Classifier.

The Logistic Regression model is a widely used statistical method for predicting binary outcomes, while Decision Tree Classifier is a machine learning algorithm that builds decision trees to model and predict the target variable. Both models were trained and evaluated using the bank telemarketing subscription prediction dataset.

Demographic information, customer financial health, and campaign characteristics were also analyzed during model training. These features were identified as being critical in predicting whether a customer would subscribe to the bank's telemarketing campaign.

In conclusion, the model training stage involves building and evaluating classification models, and identifying the critical features that contribute to the model's predictive power. By performing these steps,

we can develop more accurate and reliable machine learning models that can predict the likelihood of a customer subscribing to the bank's telemarketing campaign.

i) Customer Demography: Variables: age, job,marital status

Logistic regression Model:

```
Call:
glm(formula = y ~ age + job + marital, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9516  -0.5238  -0.4542  -0.3847   2.4183

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.512657   0.108645 -23.127  < 2e-16 ***
age               0.010919   0.001953   5.591 2.26e-08 ***
jobblue-collar   -0.507696   0.065134  -7.795 6.46e-15 ***
jobentrepreneur  -0.324338   0.114046  -2.844 0.004456 **
jobhousemaid     -0.322807   0.122984  -2.625 0.008670 **
jobmanagement     0.176137   0.058258   3.023 0.002500 **
jobretired        0.676357   0.083843   8.067 7.21e-16 ***
jobself-employed -0.011202   0.099448  -0.113 0.910315
jobservices      -0.335990   0.078174  -4.298 1.72e-05 ***
jobstudent        0.956036   0.097046   9.851  < 2e-16 ***
jobtechnician    -0.148539   0.063577  -2.336 0.019473 *
jobunemployed     0.352020   0.096342   3.654 0.000258 ***
jobunknown       -0.030076   0.204776  -0.147 0.883232
maritalmarried   -0.077945   0.053752  -1.450 0.147036
maritalsingle     0.372490   0.060799   6.127 8.98e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26140  on 36200  degrees of freedom
Residual deviance: 25390  on 36186  degrees of freedom
AIC: 25420

Number of Fisher Scoring iterations: 5
```

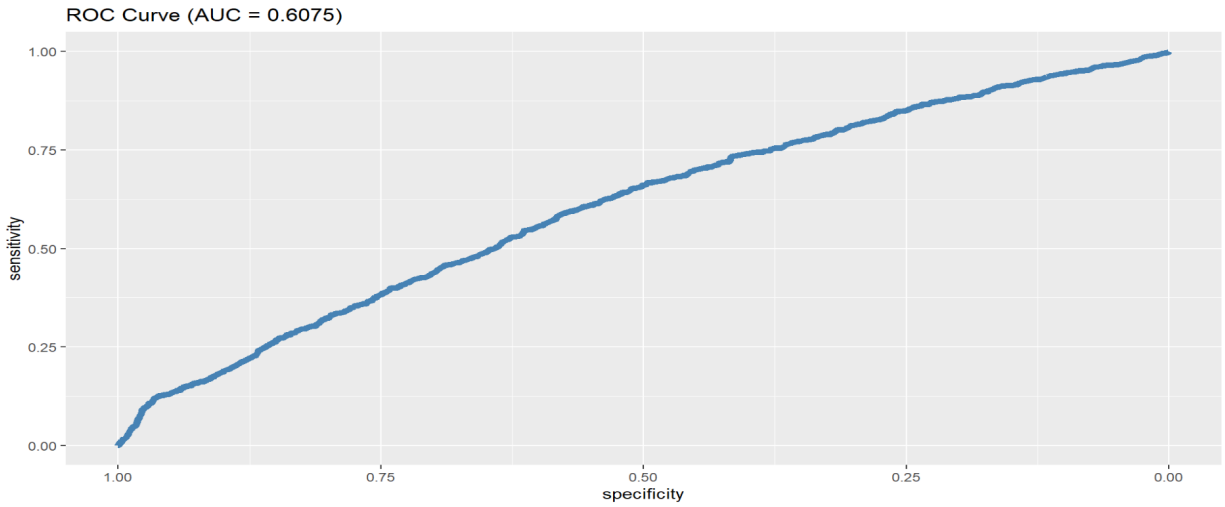
log(p/1-p) = -2.51 + 0.011*age -0.50*job_blue_collar -0.324338*job_entrepreneur -0.322807*job_housemaid +0.176137*job_management + 0.676357* job_retired -0.011202 *job_self-employed -0.335990*job_services+0.956036*job_student -0.148539*job_technician + 0.352020*job_unemployed -0.030076*job_unknown -0.077945*marital_married+ 0.372490* marital_single

Based on the model, we can explain the following related to different features

i) From above, it is evident that, If age in increase by 10 year than the odds of subscription will be increase by exp(10*0.011)-1)*100% i.e 11.6%.

ii) The odds of subscribing for a retired person is 96% higher than a person having other jobs.

iii) The odds of subscribing for a student is 160% higher than the person having other jobs.




```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    36200      26140
age      1    32.13     36199      26108 1.441e-08 ***
job     11   601.94     36188      25506 < 2.2e-16 ***
marital  2   116.18     36186      25390 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ROC-AUC score of a binary classifier is a measure of its ability to distinguish between positive and negative cases, with a value of 0.6075 indicating moderate performance. In addition, we conducted an ANOVA to determine the significance of the coefficients of the model's features. The results suggest that the coefficients for age, job, and marital status are statistically significant, with a p-value less than the standard significance level of 0.05.

ii) **Financial Health** : balance, housing, loan

Logistic Regression Model

```
Call:
glm(formula = y ~ balance + housing + loan, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8512  -0.5763  -0.4149  -0.3836   2.6209

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.719e+00  2.670e-02  -64.39   <2e-16 ***
balance      2.596e-04  1.893e-05   13.72   <2e-16 ***
housingyes  -8.532e-01  3.405e-02  -25.06   <2e-16 ***
loanyes     -6.113e-01  5.603e-02  -10.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26140  on 36200  degrees of freedom
Residual deviance: 25084  on 36197  degrees of freedom
AIC: 25092

Number of Fisher Scoring iterations: 5
```

Equation:

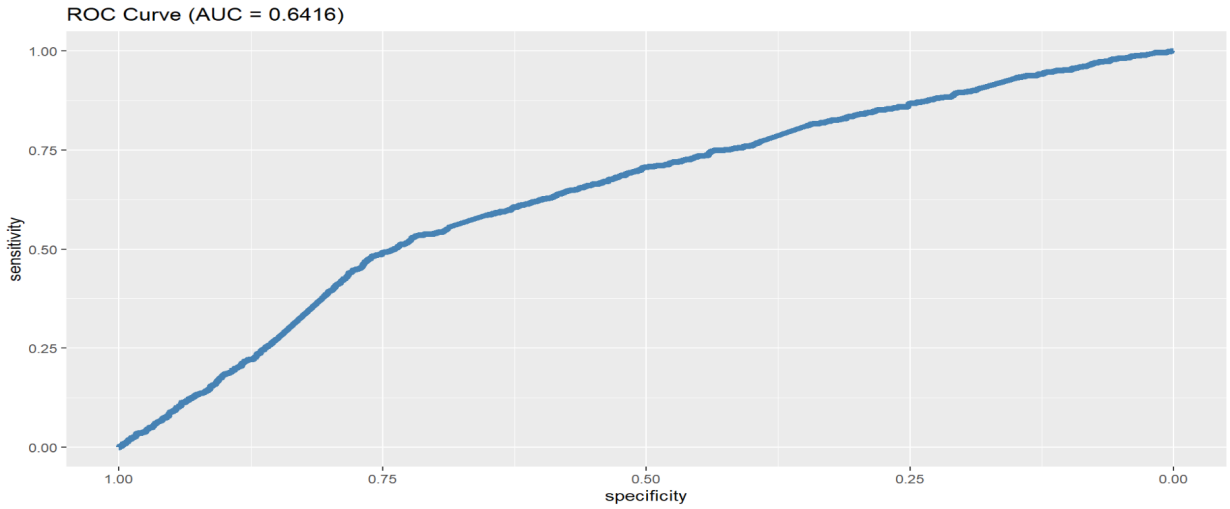
log(p/1-p) = -1.719 + 2.596e-05*balance -0.8532*housing_yes-0.6113*loan_yes

<u>Based on the model, we can explain the impact of different features</u>

i) Odds of subscription for person having housing loan is 57% lesser than the person not having housing loan.

ii) Odds of subscription for person having personal loan is 45% lesser than the person doesn’t have personal loan.

iii) The coefficient of balance is very less, it indicates that the people with higher balances tends have lesser odds of subscriptions.




```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    36200      26140
balance  1   245.39     36199      25895 < 2.2e-16 ***
housing  1   675.17     36198      25220 < 2.2e-16 ***
loan     1   135.90     36197      25084 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iii) **Campaign characteristics**: duration, campaign, day, month

Logistic Regression Model

```
Call:
glm(formula = y ~ duration + campaign + day + month, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0203  -0.4817  -0.3719  -0.2945   2.6395

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.2659330  0.0785817 -28.835  < 2e-16 ***
duration     0.0042783  0.0001122  38.144  < 2e-16 ***
campaign    -0.1298589  0.0156978  -8.272  < 2e-16 ***
day         -0.0007434  0.0023657  -0.314    0.753
monthaug    -0.3821556  0.0723401  -5.283 1.27e-07 ***
monthdec     1.3250660  0.1663189   7.967 1.63e-15 ***
monthfeb    -0.0390787  0.0853410  -0.458    0.647
monthjan    -0.7650684  0.1169555  -6.542 6.09e-11 ***
monthjul    -0.7842069  0.0730203 -10.740  < 2e-16 ***
monthjun    -0.5638121  0.0762615  -7.393 1.43e-13 ***
monthmar     1.6653224  0.1205127  13.819  < 2e-16 ***
monthmay    -1.1623814  0.0669558 -17.360  < 2e-16 ***
monthnov    -0.6303403  0.0806412  -7.817 5.43e-15 ***
monthoct     1.2736502  0.1039319  12.255  < 2e-16 ***
monthsep     1.4177111  0.1102343  12.861  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26140  on 36200  degrees of freedom
Residual deviance: 22880  on 36186  degrees of freedom
AIC: 22910

Number of Fisher Scoring iterations: 5
```

Equation :

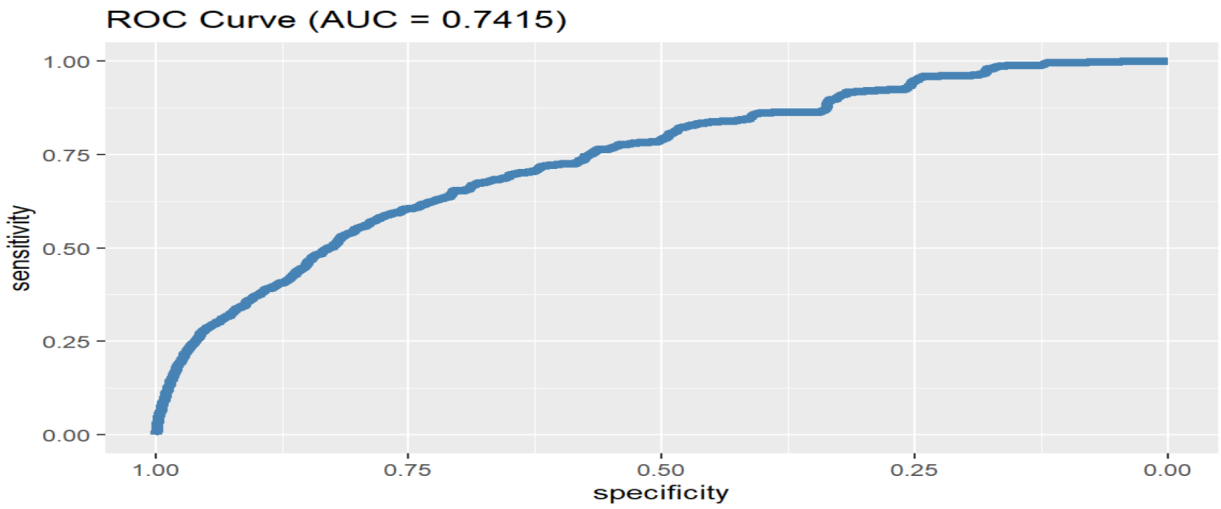
log(p/1-p) = -2.2659330 + 0.0042783*duration-0.1298589*campaign-0.0007434*day
-0.3821556*month_aug + 1.3250660*month_dec -0.0390787*month_feb
-0.7650684*jan-0.7842069*july-0.5638121*jun+1.6653224*month_mar-1.1623814*month_may
-0.6303403*month_nov+1.2736502*month_oct+1.4177111*month_sep

Based on the model, we can explain the impact of different features

i) If call duration is increased by 100 seconds, then Odds of subscription will be increase by 53%

ii) If number of call increases in campaign is increased by 1 unit then odds of subscription is decreased by 13%

iii) The coefficients of month december, march, october, and september is greater than 1. We can say that, during these month if activity of campaign is increased, then odds of subscription tends to increase.

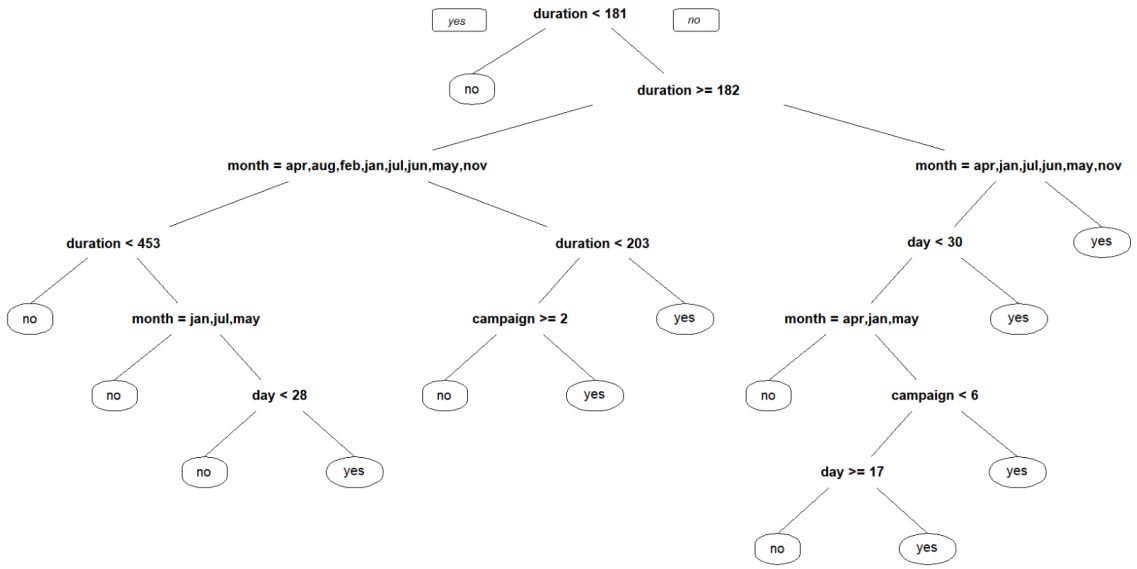



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     36200      26140
duration  1  1433.91     36199      24706 < 2.2e-16 ***
campaign  1   124.41     36198      24582 < 2.2e-16 ***
day       1    12.58     36197      24569  0.000389 ***
month    11  1688.99     36186      22880 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In the case of a 0.74 ROC-AUC score, the model is able to distinguish between positive and negative cases with an accuracy that is significantly better than random chance. It is a better predictor than the model with demographic and financial health variables. From anova, we can say the duration, campaign, day and months coefficient is significant.

Decision Tree based classifier:

We built a decision tree classifier for with campaign characteristics, it gives better way to visualize the impact of different variables.



<u>Based on the model, we can explain the impact of different features</u>

i) If call duration is less than 181 seconds, then the customer is not going to subscribe to the term deposit plan.

Evaluation

row.names <chr>	training_set <dbl>	test_set <dbl>
Accuracy	0.6842905	0.6826859
Precision	0.2563199	0.2190802
Recall	0.8924021	0.6707897
F_1	0.3982520	0.3302881

To verify the performance of the decision tree classifier, we have used the F-1, Precision and Recall. For this model we are getting 0.33 F-1 which is very less, precision 0.25 means 75% of time it will misclassify the subscription prediction.

iv) Mix Model:

Logistic Regression Model:

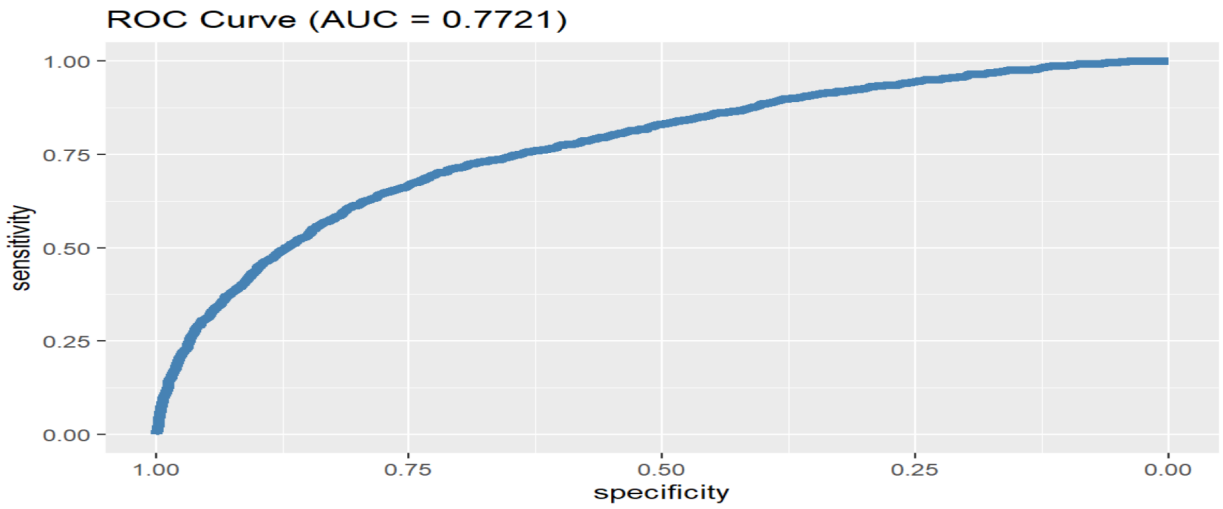
```
Call:
glm(formula = y ~ age + job + balance + marital + housing + loan +
    duration + campaign + day + month + previous, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.7949  -0.4769  -0.3489  -0.2537   2.8926

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.972e+00  1.460e-01 -13.505  < 2e-16 ***
age               9.963e-04  2.103e-03   0.474  0.63572
jobblue-collar   -3.237e-01  6.942e-02  -4.662 3.13e-06 ***
jobentrepreneur  -2.542e-01  1.206e-01  -2.107  0.03509 *
jobhousemaid     -3.909e-01  1.312e-01  -2.979  0.00290 **
jobmanagement     1.005e-01  6.328e-02   1.588  0.11222
jobretired        2.913e-01  9.355e-02   3.113  0.00185 **
jobself-employed -7.223e-02  1.068e-01  -0.676  0.49881
jobservices      -1.865e-01  8.286e-02  -2.251  0.02438 *
jobstudent        4.784e-01  1.082e-01   4.423 9.74e-06 ***
jobtechnician    -9.811e-02  6.850e-02  -1.432  0.15206
jobunemployed     1.027e-01  1.054e-01   0.975  0.32977
jobunknown       -2.880e-01  2.178e-01  -1.322  0.18604
balance           2.000e-04  2.093e-05   9.555  < 2e-16 ***
maritalmarried   -1.707e-01  5.762e-02  -2.962  0.00306 **
maritalsingle     1.250e-01  6.586e-02   1.898  0.05775 .
housingyes       -6.393e-01  4.254e-02 -15.027  < 2e-16 ***
loanyes          -4.614e-01  5.937e-02  -7.771 7.79e-15 ***
duration          4.270e-03  1.150e-04  37.118  < 2e-16 ***
campaign         -1.208e-01  1.585e-02  -7.622 2.50e-14 ***
day              -2.316e-03  2.382e-03  -0.972  0.33096

monthaug         -6.743e-01  7.762e-02  -8.688  < 2e-16 ***
monthdec          8.064e-01  1.722e-01   4.684 2.82e-06 ***
monthfeb         -2.857e-01  8.854e-02  -3.227  0.00125 **
monthjan         -1.024e+00  1.199e-01  -8.538  < 2e-16 ***
monthjul         -7.078e-01  7.629e-02  -9.278  < 2e-16 ***
monthjun         -6.624e-01  7.930e-02  -8.352  < 2e-16 ***
monthmar          1.201e+00  1.255e-01   9.566  < 2e-16 ***
monthmay         -9.313e-01  6.891e-02 -13.514  < 2e-16 ***
monthnov         -7.274e-01  8.311e-02  -8.752  < 2e-16 ***
monthoct          8.210e-01  1.087e-01   7.555 4.19e-14 ***
monthsep          8.655e-01  1.152e-01   7.510 5.93e-14 ***
previous          8.967e-02  7.537e-03  11.897  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26140  on 36200  degrees of freedom
Residual deviance: 22000  on 36168  degrees of freedom
AIC: 22066

Number of Fisher Scoring iterations: 5
```

Equation:

log(p/1-p) = -1.972e+00 + 9.963e-04*Age -3.237e-01*jobblue-collar - 2.542e-01*jobentrepreneur - 3.909e-01* jobhousemaid + 1.005e-01* jobmanagement + 2.913e-01* jobretired - 7.223e-02* jobself-employed - 1.865e-01*jobservices + 4.784e-01*jobstudent - 9.881e-02* jobtechnician + 1.027e-01* jobunemployed - 2.880e-01*jobunknown + 2.000e-04*balance -1.707e-01* maritalmarried + 1.250e-01*maritalsingle - 6.393e-01*housingyes - 4.164e-01*loanyes + 4.270e-0.3* duration - 1.208e-01*campaign - 2.316e-03*day - 6.743e-01*monthaug + 8.064e-01*monthdec - 2.857e-01*monthfeb - 1.024e+00* monthjan - 7.078e-01*monthjul - 6.624e-01*monthjun + 1.201e+00*monthmar - 9.313e-01*monthmay - 7.274e-01*monthnov + 8.210e-01*monthoct + 8.655e-01*monthsep + 8.967-02*previous

Based on the model, we can explain the impact of different features

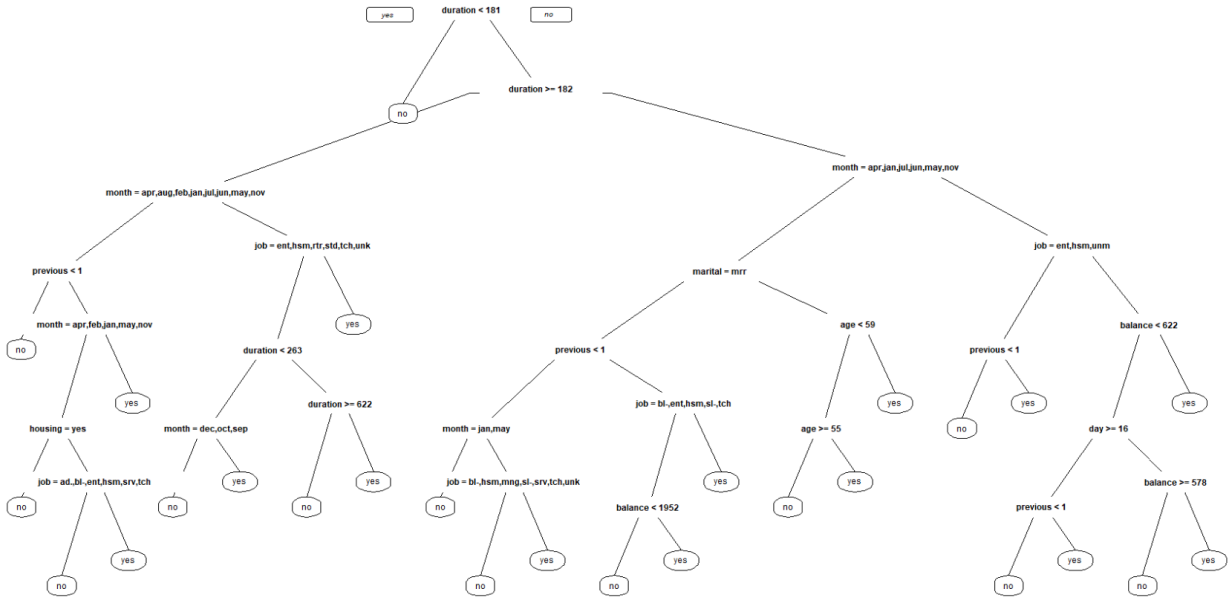
i) If age is increased by 20 years then Odds of subscription will increase by 2% .
ii) If day is increased by 10 days then Odds of subscription will decrease by 2.2%
iii) If the customer is a student, the Odds of subscription will be 61% higher than the non-student customer.
iv) If the customer is a retired from job then the Odds of subscription will be 33% higher than the non-retired customer.



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      36200      26140
age        1    32.13     36199      26108 1.441e-08 ***
job       11   601.94     36188      25506 < 2.2e-16 ***
balance    1   196.95     36187      25309 < 2.2e-16 ***
marital    2   110.51     36185      25199 < 2.2e-16 ***
housing    1   428.36     36184      24770 < 2.2e-16 ***
loan       1   104.42     36183      24666 < 2.2e-16 ***
duration   1  1419.02     36182      23247 < 2.2e-16 ***
campaign   1   114.20     36181      23133 < 2.2e-16 ***
day        1    15.59     36180      23117 7.861e-05 ***
month     11   983.56     36169      22134 < 2.2e-16 ***
previous   1   133.43     36168      22000 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ROC-AUC is 0.77, which is better than the other three models. This will really do well to separate the positive and negative responses, In turn it will give higher Precision and Recall.

Decision Tree Classifier:



row.names <chr>	training_set <dbl>	test_set <dbl>
Accuracy	0.9014944	0.8952275
Precision	0.6119254	0.5710491
Recall	0.4334592	0.4091342
F_1	0.5074586	0.4767184

The decision tree model with mixed variables giving good performance, with 0.47 F-1 Score, Precision and Recall is also decent in terms of performance. Model is also generalizing well.

5.4. Model Evaluation and Analysis

i) Model Evaluation: ROC-AUC

Model evaluation is an essential step in machine learning that involves measuring the performance of the trained models. In this study, two evaluation metrics have been used to evaluate the performance of the trained models: F-1 Score and ROC-AUC.

The F-1 Score is a measure of the harmonic mean of precision and recall. It provides a single score that summarizes the model's overall performance in predicting both positive and negative instances. The ROC-AUC is a measure of the model's ability to discriminate between positive and negative instances. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

Characteristics	Features	Logistic Regression Model
		ROC-AUC
Customer Demography	Age, Education, Marital, Job	0.6075
Customer Financial Health	balance, housing, loan	0.6416
Campaign	duration, campaign, day, month	0.7415
Mixed Model	age, job, balance, marital, housing, loan, duration, campaign, day, month, previous	0.774

From the above table, we can see the results obtained for different models developed and the best model is chosen.

a) Customer demography
   (Research Question: Does demography of a client affect the subscription rate ?)
   - From the above results, we can see that , demography of a person does not give good prediction results for subscribing to a term deposit. Age and marital status have larger p values which gives bad results

b) Customer Financial Health
   (Research Question: Does financial health of a person affect the subscription rate ?)
   - This model has a better ROC-AUC score, compared to the previous, but not great. All the 3 variables have lesser p-value, which enables us to to use the model for predictions

c) Campaign
   (Research Question: Do the time characteristics affect the subscription rate ?)
   - This model has a better score increased by 10%. Variables duration,day,month,campaign have better p-values and the model seems to be a good fit with ROC-AUC score of 74.15%

d) Mix Model
   (Research Question: Do all the properties, demography, financial health and campaign affect subscription rate ?)
   - With most of the variables having a lesser p-value, this model seems to give us the best fit with ROC-AUC score of 77.4%

6. References

[1] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

[2] Dua, D., & Graff, C. (2017). UCI machine learning repository [https://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Cortez, P., Moro, S., & Rita, P. (2018). Bank marketing data set. UCI machine learning repository [https://archive.ics.uci.edu/ml/datasets/bank+marketing] Irvine, CA: University of California, School of Information and Computer Science.

[4] Moro, S., Cortez, P., & Rita, P. (2018). Bank marketing dataset analysis: Are we predicting bank term deposit subscription? Journal of Data Science, 16(2), 219-234.

[5] Malaquias, R. F., Lima, R. S., & Meira Jr., W. (2020). Exploring imbalance in the bank marketing dataset using the Empirical Rule of Oversampling. Expert Systems with Applications, 141, 112967.

[6] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA, 1984.

[7] David S. Coppock, Why lift? Data modeling and mining, Information Management (2002) 5329-1 2002, (Online; accessed 19-July-2013).

[8] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (3) (1995) 273–297.

[9] Paulo Cortez, Data mining with neural networks and support vector machines using the r/rminer tool, Advances in Data Mining. Applications and Theoretical Aspects, 6171, Springer, 2010, pp. 572–583