Queen's University Management School.

MGT7177 Statistics For Business: Assignment 2

PREDICTING TERM DEPOSIT SUBSCRIPTION USING LOGISTIC REGRESSION

Semester 1, 2022/2023
Student: BHARAT GURUNATHAN HARI DOSS
Student ID: 40387258
Word Count: 2179

Contents

1.Introduction

Since term deposits are usually an important part of a bank's operations, a robust marketing strategy is imperative. One of the most widely used marketing techniques nowadays is telephone marketing, which consulting firms may find valuable for attracting new clients(Hung, Hanh et al. 2019). Through a marketing campaign that involves calling consumers personally, the bank hopes to choose the best set of customers(Moro, Cortez et al. 2014). A logistic regression model (LR) is used to assess the potential causes to successful deposit subscriptions and anticipate the success rate of bank telemarketing(Jiang 2018). The goal of the marketing campaign was to persuade customers to open a term deposit. The information may be used to understand the factors that affect a customer's choice to subscribe to the term deposit(Miguéis, Camanho et al. 2017).

Related work Literature:

In (Hosseini 2021)work, the logistic regression model is applied. It was primarily concerned with the association between telemarketing success (output variable) and several good characteristics. Furthermore, in (Hosseini 2021) study, the prediction is applied based on the created model to locate some possible consumers who would really subscribe to the deposits.

(Moro, Cortez et al. 2014) set out to predict the success of bank telemarketing. They used a comprehensive data set from 2008 to 2013 and a data collection with 150 characteristics. They examine the logistic regression (LR), decision trees, support vector machines, and neural networks models of data mining (NN). Le & Viviani (2017) carried out research to identify the risk factors for bank failure. They used discriminant analysis and logistic regression, two established statistical techniques, to analyse a sample of 3000 US banks.

Table1: Hypotheses

No of Relation	Variable X (Type)	Target Variable (Type)	Measures of association	Description
1	Job(categorical)	Subscribed(binary: 'yes','no')	Chi-squared test	Bank Client Variable
2	marital_status(categorical)	Subscribed(binary: 'yes','no')	Chi-squared test	Bank Client Variable
3	Education(categorical)	Subscribed(binary: 'yes','no')	Chi-squared test	Bank Client Variable
4	Month(categorical)	Subscribed(binary: 'yes','no')	Chi-squared test	The Most Recent Contact During the Marketing Campaign Variables
5	Poutcome(categorical)	Subscribed(binary: 'yes','no')	Chi-squared test	Other Variables

2. Methodology and Analysis:

The methodology includes filtering and processing of the raw datasets for anomalies like missing data,data standardization ,assessing the data and producing summary statistics, evaluating the training and testing datasets, testing multicollinearity among diverse features, and finding a collection of predictor variables to forecast the Term Deposit Subscriptions . The flowchart below depicts the fundamental processes taken(Jiang 2018).

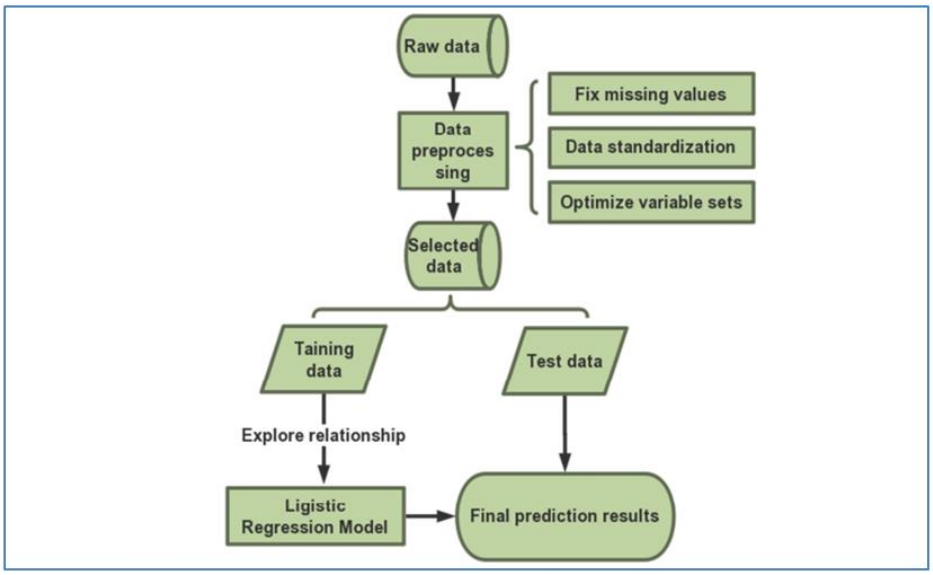

Figure 1: Methodology of LR model(Jiang 2018)

It is used to assess whether closely a sample matches the pattern of a defined population or to look into the relationship between two categorical variables' independence. (Franke, Ho et al. 2012).For visualisations ggplot2 is used.The ggplot2 functions and their outputs can be added to the plot as additional layers. For example, adding a title using ggplot2's labs() method (Ito and Murphy 2013).

To avoid overfitting, we divided the 41153 observations into training data (80%) and test data (20%). Because the answer variable "yes" only accepts a tiny amount of the data, data partitioning via the function createDataPartition is used to provide a balanced response in both training and test data(Lever, Krzywinski et al. 2016).

It is employed for measure of association because the target variable is categorical(goodness of fit).The logistic regression is evaluated using log likelihood, the deviance statistic, pseudo R squared, and the odds ratio (Smith and McKenna 2013).Predictors are evaluated using significance and the odds ratio, and models are compared using the Akaike information criteria (AIC)(Li and Nyholt 2001).

Dataset: Data set includes 41153 clients. Client traits and campaign features, and other social and economic qualities are among the 20 predictor variables.

The client's choice to purchase the long-term deposit, stated as "yes" or "no," is the response variable. The variables education, employment, marital, housing, loan, and default have a 'unknown' level. There are total 40 missing values in the dataset.

Addressing Data Quality Issues:

Age Column: To find outliers in Age data, using the Box Plot function and giving the mean of age to the outliers which are less than 5 and more than 100.

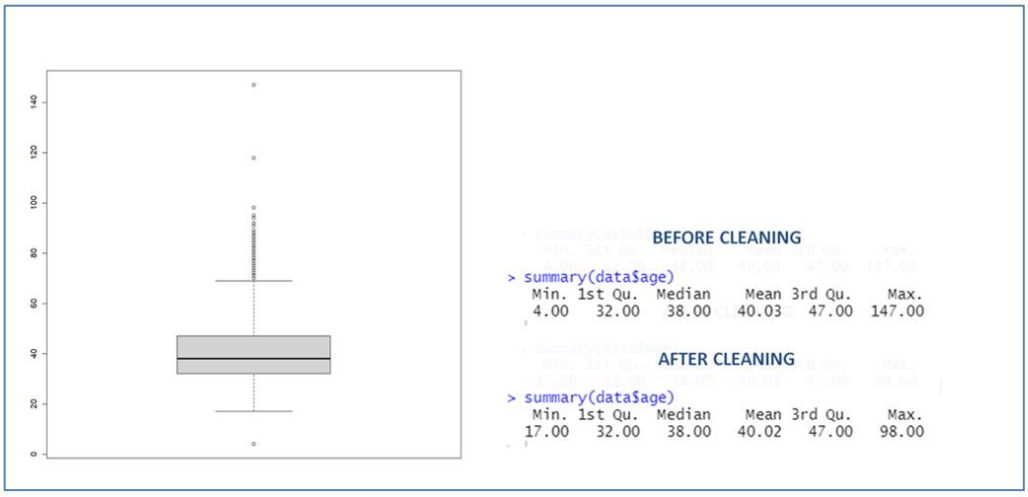

Figure 2: Addressing age column data issue

Default Column: The data quality issue in Default Column has been corrected, the "no" value is misspelt as 'n' per the data dictionary.

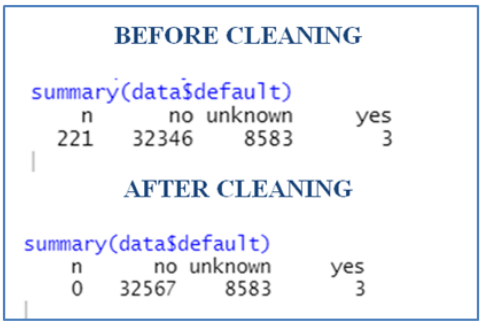

Figure 3: Addressing Default column data issue

Month Column: The data quality problem in the Month Column has been resolved; the "mar" value is misspelt as "march" according to the data dictionary.

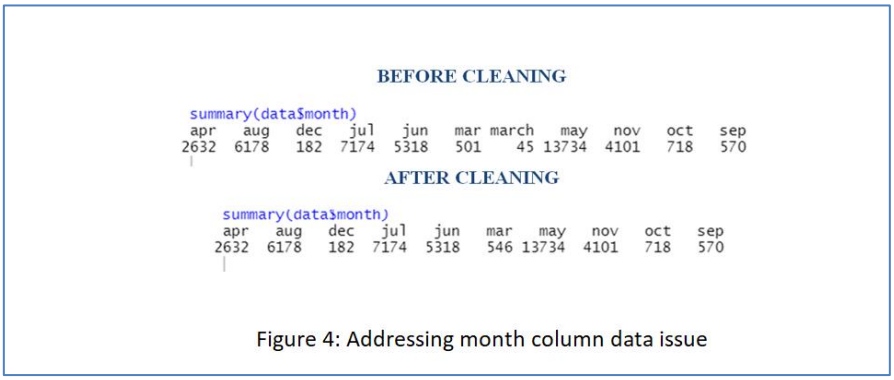

Figure 4: Addressing month column data issue

Pdays Column: The 999 code indicates that the consumer was not contacted earlier, which might be assigned as 0.

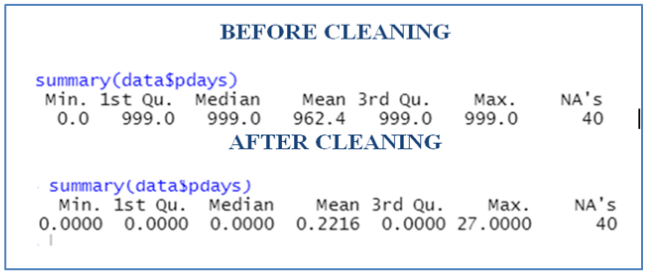

Figure 5: Addressing Pdays column data issue

3. Results and Discussion

Descriptive Statistics:

Descriptive statistics are used to describe data in an ordered manner by explaining the connection between variables within a sample or population. Before conducting inferential statistical comparisons, descriptive statistics should always be calculated as a crucial first step in conducting analysis. (Kaur, Stoltzfus et al. 2018).

```
> summary(data)
       ID              age                 job              marital                    education        default          housing          loan
 Min.   :    1   Min.   :17.00   admin.     :10416   divorced: 4609   university.degree  :12160   n      :    0   no     :18604   no     :33923
 1st Qu.:10289   1st Qu.:32.00   blue-collar: 9244   married :24905   high.school        : 9508   no     :32567   unknown:  990   unknown:  990
 Median :20577   Median :38.00   technician : 6739   single  :11559   basic.9y           : 6038   unknown: 8583   yes    :21559   yes    : 6240
 Mean   :20577   Mean   :40.02   services   : 3963   unknown :   80   professional.course: 5240   yes    :    3
 3rd Qu.:30865   3rd Qu.:47.00   management : 2920                    basic.4y           : 4176
 Max.   :41153   Max.   :98.00   retired    : 1719                    basic.6y           : 2286
                                 (Other)    : 6152                    (Other)            : 1745
     contact          month       day_of_week    duration         campaign          pdays            previous            poutcome
 cellular :26144   may    :13734   fri:7827    Min.   :   0.0   Min.   : 1.000   Min.   : 0.0000   Min.   :0.0000   failure    : 4252
 telephone:15009   jul    : 7174   mon:8479    1st Qu.: 102.0   1st Qu.: 1.000   1st Qu.: 0.0000   1st Qu.:0.0000   nonexistent:35528
                   aug    : 6178   thu:8623    Median : 180.0   Median : 2.000   Median : 0.0000   Median :0.0000   success    : 1373
                   jun    : 5318   tue:8090    Mean   : 258.2   Mean   : 2.568   Mean   : 0.2216   Mean   :0.1731
                   nov    : 4101   wed:8134    3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: 0.0000   3rd Qu.:0.0000
                   apr    : 2632               Max.   :4918.0   Max.   :56.000   Max.   :27.0000   Max.   :7.0000
                   (Other): 2016                                                  NA's   :40
  emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m      nr.employed   subscribed
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634   Min.   :4964   no :36513
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.344   1st Qu.:5099   yes: 4640
 Median : 1.10000   Median :93.75   Median :-41.80   Median :4.857   Median :5191
 Mean   : 0.08102   Mean   :93.58   Mean   :-40.51   Mean   :3.620   Mean   :5167
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045   Max.   :5228
```

Figure 6: Descriptive Statistics of dataset variables

Relationship between Client Age and Subscribed:

Clients phoned by the bank in its telemarketing efforts range in age from 17 to 98 years old.The average is 40 years old. From the figure below, its shows the majority of clients that called were in their 30s and 40s (32 to 47 years old fall within the 25th to 75th percentiles).

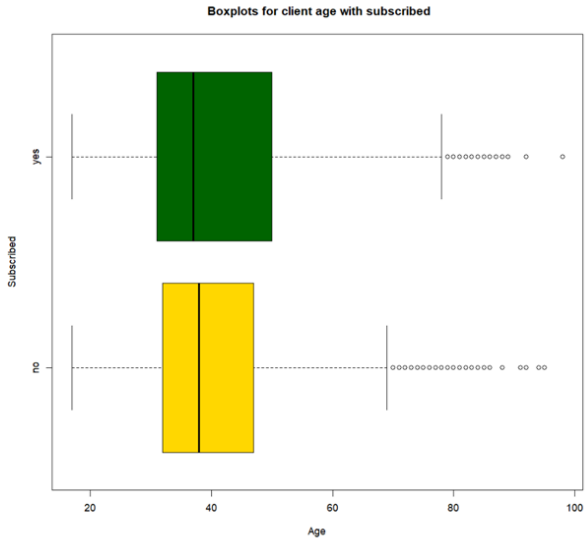

Figure 7: Boxplot of age with Subscription of customers

Relationship between Customer job details and Subscribed:

The chisq.test was used to assess the relationship. Because the p-value is less than 0.05, so ruling out the null hypothesis and concluding that there is a significant relationship.

```
        Pearson's Chi-squared test

data:  data$job and data$subscribed
X-squared = 960.03, df = 11, p-value < 0.00000000000000022
```

Figure 8: Chi-squared test between job and Subscription of customers

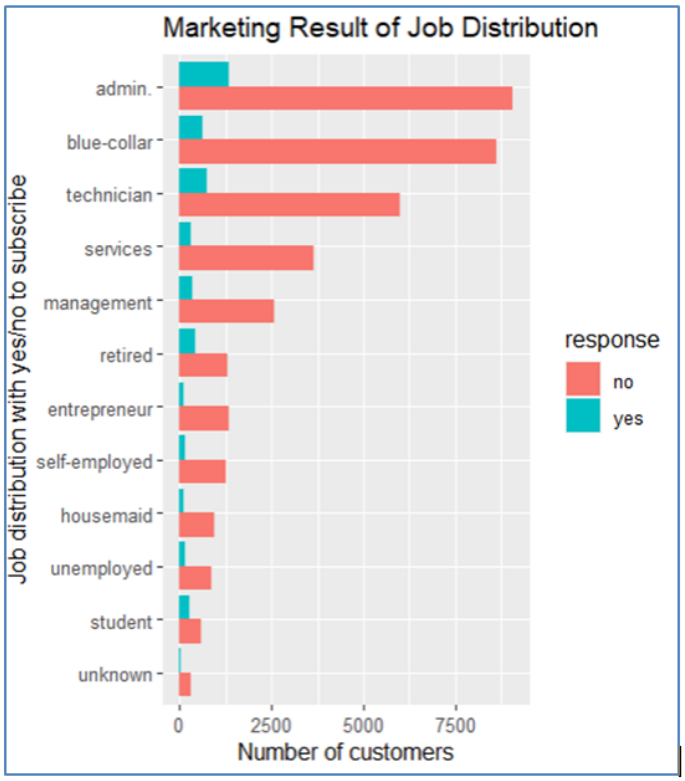

Figure 9: ggplot between relationship of job and Subscription of customers

From the descriptive statistical data(Figure 6), it shows that the majority of consumers work in admin and blue collar jobs. As per the ggplot , students and retired consumers account for more than half of all subscriptions. Blue-collar workers have the lowest conversion rate. Customers in the student, retired, administrative, and jobless groups are more likely to respond positively to the marketing.

Relationship between the marital status and number of customers subscribed:

If the p-value is less than 0.05, a relationship is considered to be significant(Di Leo and Sardanelli 2020).

```
        Pearson's Chi-squared test

data:  data$marital and data$subscribed
X-squared = 122.58, df = 3, p-value < 0.00000000000000022
```

Figure 10: Chi-squared test between marital status and Subscription of customers

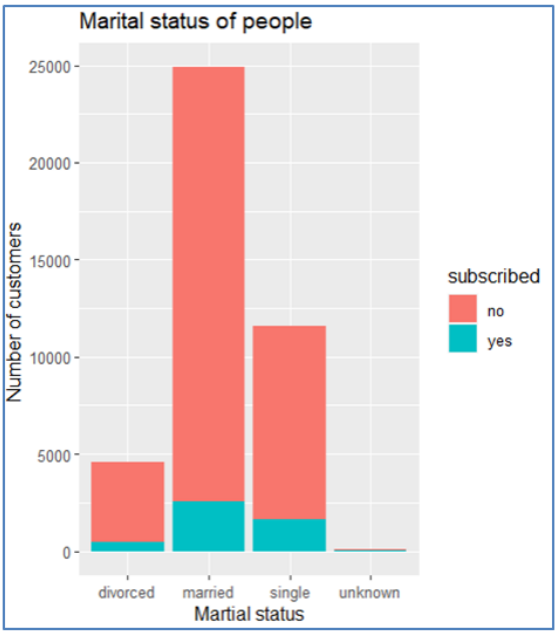

Figure 11: ggplot between marital status and Subscription of customers

Above chart shows that term deposit subscriptions from a "single" client are more likely than those from other groups (divorced and married). Most married customers make up the marketing target and married customers are more likely to respond better than divorced ones.

Relationship between the Education and subscribed clients:

```
        Pearson's Chi-squared test

data:  data$education and data$subscribed
X-squared = 192.91, df = 7, p-value < 0.00000000000000022
```

Figure 12: Chi-squared test between Education and Subscription

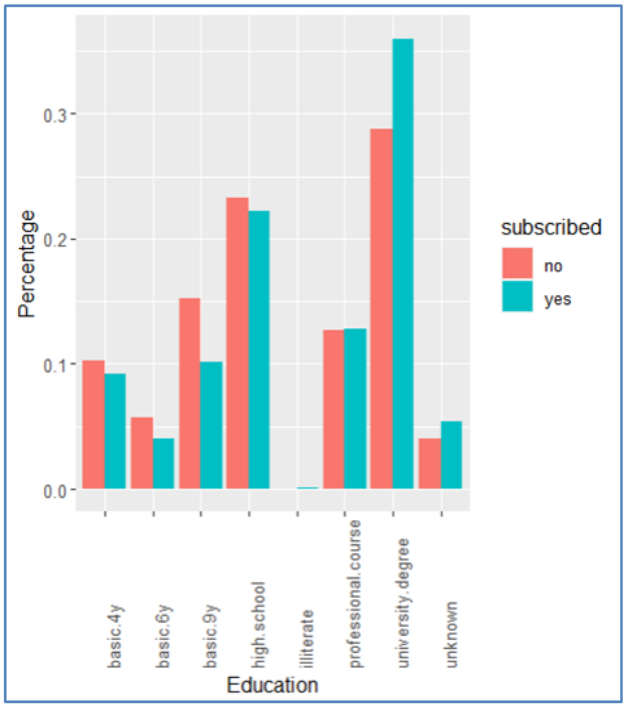

Figure 13: ggplot between Education and Subscription

The chance of enrolling in a term deposit appears to be positively connected with the number of years of education. Those with a university education are the most likely to sign up for a term deposit. High school, professional training, and university degrees should be the three target audiences for bank marketing. Because there are just 18 illiterate clients, the "illiterate" category has the highest subscription rate.

Relationship between the month and subscribed clients:

```
        Pearson's Chi-squared test

data:  data$month and data$subscribed
X-squared = 3096.6, df = 10, p-value < 0.00000000000000022
```

Figure 14: Chi-squared test between month and Subscription

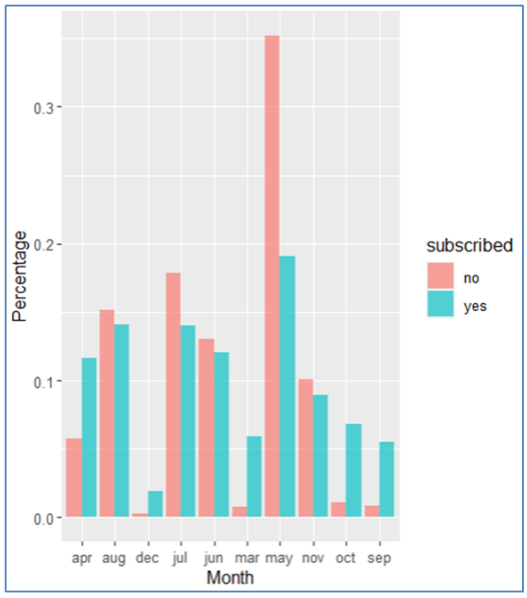

Figure 15: ggplot between month and Subscription

During the months of January and February, no communication was made. Between May and August, the bank contacted the majority of its customers. . The largest surge happens in May, yet the ratio of subscribers to people contacted is the poorest. The greatest subscription rate was in March. The months with a low contact frequency (March, September, October, and December) has excellent outcomes.

Relationship between the default and subscribed clients:

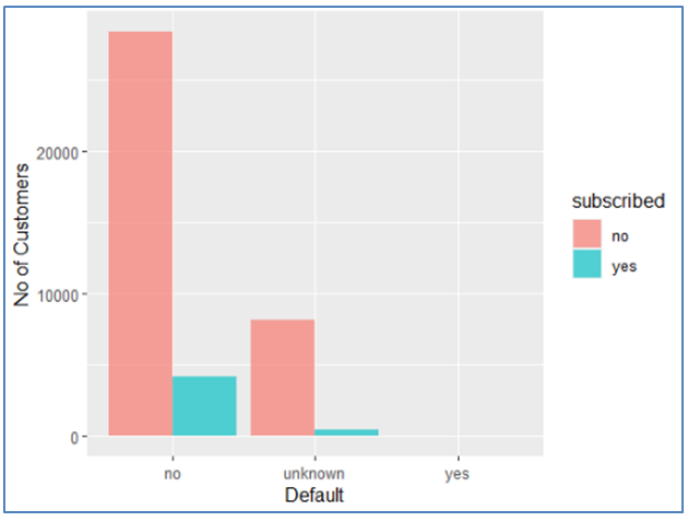

Figure 16: ggplot between default and Subscription

According to the figure above, the majority of the consumers in the marketing campaign had no default history, which is sensible given that defaulters are less likely to have idle funds.

Relationship between the Previous Campaign and subscribed clients:

```
        Pearson's Chi-squared test

data:  data$poutcome and data$subscribed
X-squared = 4225.7, df = 2, p-value < 0.00000000000000022
```

Figure 17: Chi-squared test between month and Subscription

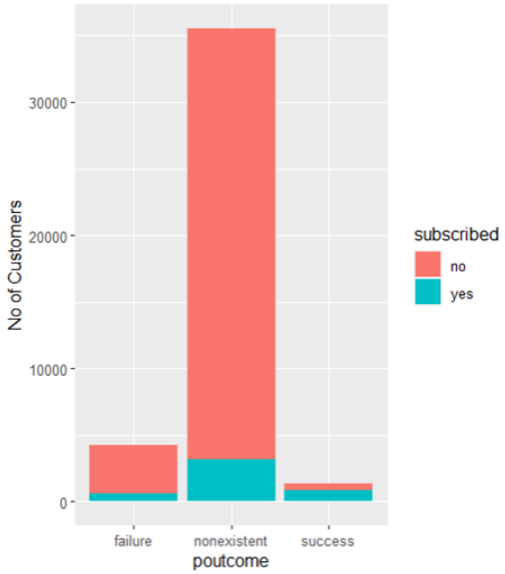

Figure 18: ggplot between poutcome and Subscription

The plot below shows that the majority of consumers were not contacted previous to this campaign. Customers are less likely to respond positively to a promotion if it has previously failed.

Customers are more likely to respond positively to the advertising if the preceding effort was successful(Brennan and Binney 2010).The outcome is not significant for no previous campaign, however the majority of consumers that offered favourable results were not approached before to this campaign.

The boxplot analysis of categorical characteristics is provided below. The binary values for the column "subscribed" are "yes" and "no" (subscribed to a term deposit).

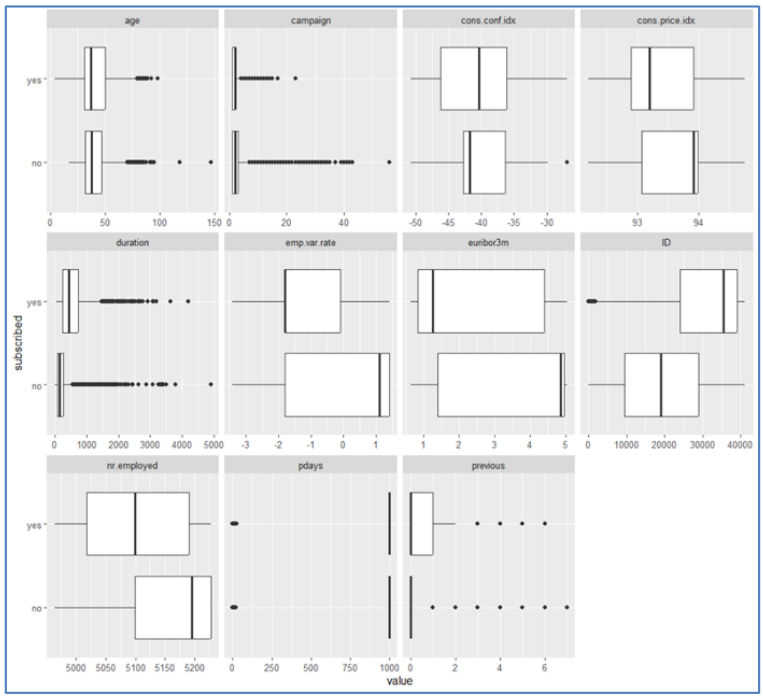

Figure 19: Boxplot of dataset variables and Subscriped customers

Regression Analysis Findings

However, before proceeding to the train-test split, it is necessary to analyse correlation further. A correlation matrix was constructed using continuous variables and discrete variables .

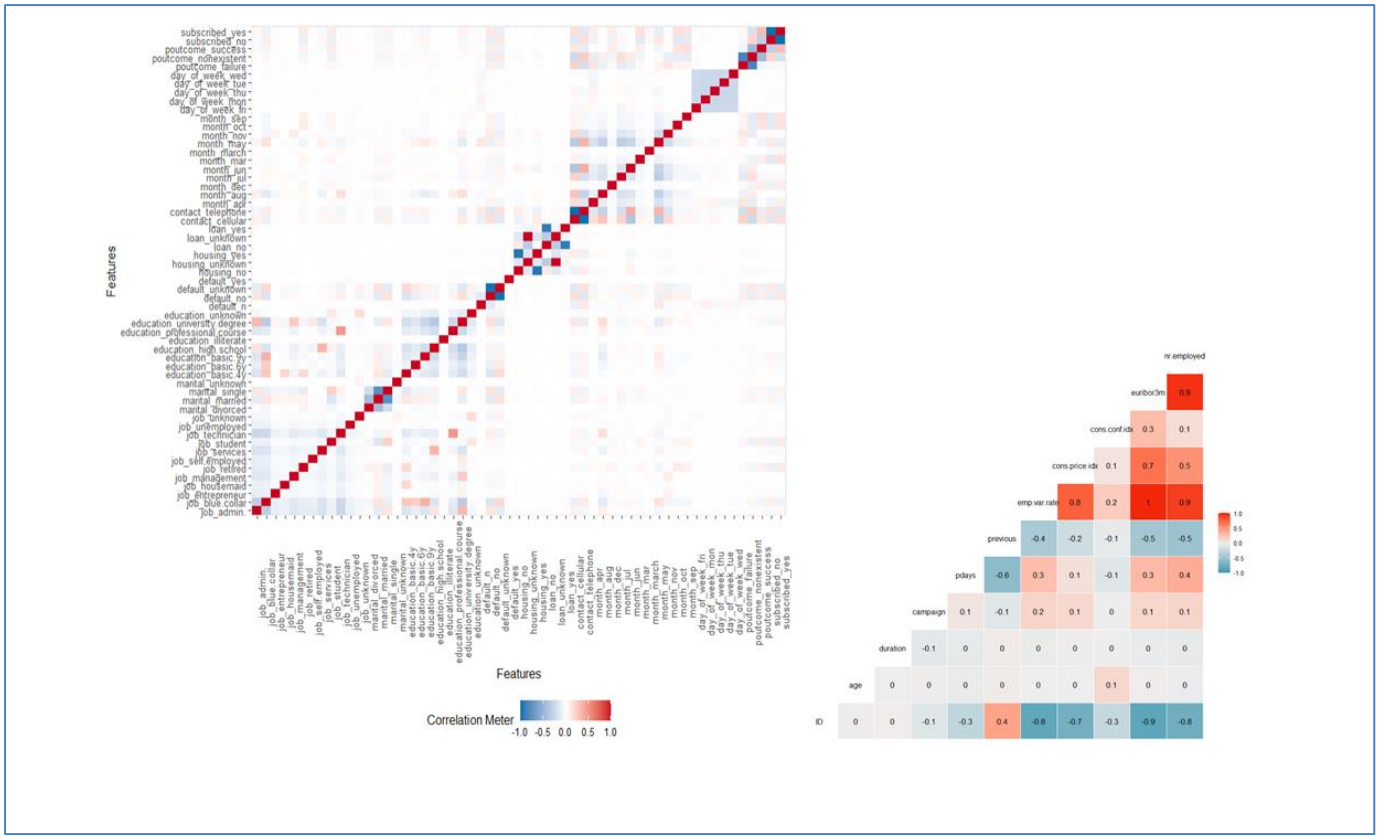

Figure 20: Correlation Matrix

Clearly, subscription of customers to term deposit shows a substantial association with "duration", a moderate link with "previous contacts", and a weak correlation with "month " and "number of campaigns". In the regression models, their effects on campaign outcomes will be studied further.

To obtain an accurate assessment of the model's accuracy, we can divide the data into a training and test set (Refaeilzadeh, Tang et al. 2009) and create the model fit, assess it, and analyse the model coefficients based on the training data. Then, by making predictions on an unidentified test dataset, the model's predictive accuracy can be assessed (Harrell Jr, Lee et al. 1996). Built the model after setting the formula using the predictors (contact + month + poutcome + cons.price.idx + cons.conf.idx + nr.employed). A logistic regression may be created in R using the glm() function (Calcagno and de Mazancourt 2010).

```
Call:
glm(formula = formula4, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0498  -0.3658  -0.3431  -0.2478   2.7237

Coefficients:
                      Estimate Std. Error z value             Pr(>|z|)
(Intercept)         63.073137   3.979114  15.851 < 0.0000000000000002 ***
monthaug            -0.124650   0.098036  -1.271              0.203563
monthdec             0.088125   0.192691   0.457              0.647427
monthjul             0.173437   0.090030   1.926              0.054050 .
monthjun             0.227740   0.088545   2.572              0.010110 *
monthmar             0.733486   0.124112   5.910   0.0000000034233573 ***
monthmarch           1.406449   0.346480   4.059   0.0000492314070458 ***
monthmay            -0.757372   0.071340 -10.616 < 0.0000000000000002 ***
monthnov            -0.346160   0.093489  -3.703              0.000213 ***
monthoct            -0.208126   0.124337  -1.674              0.094153 .
monthsep            -0.528714   0.132967  -3.976   0.0000700049348298 ***
poutcomenonexistent  0.481184   0.061070   7.879   0.0000000000000033 ***
poutcomesuccess      1.786514   0.086851  20.570 < 0.0000000000000002 ***
contacttelephone    -0.438387   0.066619  -6.581   0.0000000000468803 ***
cons.price.idx      -0.042890   0.041764  -1.027              0.304431
cons.conf.idx        0.022112   0.005131   4.310   0.0000163440336917 ***
nr.employed         -0.011749   0.000323 -36.375 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 18377  on 32906  degrees of freedom
AIC: 18411

Number of Fisher Scoring iterations: 6
```

Figure 21: Summary of Logistic Regression accurate model

To determine if the predictor differs substantially from zero, the z-statistic, which has a normally distributed distribution, is utilised(Krackhardt 1988).The estimated nr.employed is -0.011749, while the z value for emp.var.rate is -36.375. As the number of employees increases, there is a decreasing possibility of subscribing to a term deposit since the value is negative. The marketing effort in March has a higher chance of generating subscriptions. Given that the estimate value is 0.022112 there is a positive correlation between the customer subscribed and the Consumer confidence index. A mathematical test called the Akaike information criterion(AIC) value is 18411, It evaluates the degree to which a model matches the facts it is meant to explain. It penalises models that include more independent variables to prevent over-fitting(Cama, Cristi Nicu et al. 2016).

```
> logisticPseudoR2s(model4)
Pseudo R^2 for logistic regression
Hosmer and Lemeshow R^2   0.208
Cox and Snell R^2         0.136
Nagelkerke R^2            0.269
```

Figure 22: Pseudo R^2 for logistic regression

A adjusted version of Cox and Snell R Squared is Nagelkerke R Squared. Between 0 and 1 is the range of values for Nagelkerke(Krackhardt 1988). The Nagelkerke R^2 value is 0.269 which is more than zero.

```
> exp(model4$coefficients)
                  (Intercept)                      monthaug                      monthdec                      monthjul
2467830534239354346406226244.0000000             0.8828057             1.0921247             1.1893863
                     monthjun                      monthmar                   monthmarch                      monthmay
             1.2557593             2.0823279             4.0814356             0.4688973
                     monthnov                      monthoct                      monthsep       poutcomenonexistent
             0.7073993             0.8121050             0.5893624             1.6179891
           poutcomesuccess          contacttelephone             cons.price.idx             cons.conf.idx
             5.9686113             0.6450761             0.9580166             1.0223583
                nr.employed
             0.9883195
```

Figure 23: Odds Ratios

While calculating Odds Ratios,the values of the predictors (poutcome, cons.conf.idx) are more than 1, the odds of the result rises. The likelihood of the remaining variables occurring are fewer than 1, and as the predictor gets bigger, those odds go less(Nick and Campbell 2007).

	GVIF	Df	GVIF^(1/(2*Df))
month	4.725106	10	1.080738
poutcome	1.301824	2	1.068164
contact	1.895528	1	1.376782
cons.price.idx	1.873955	1	1.368925
cons.conf.idx	2.327407	1	1.525584
nr.employed	1.980194	1	1.407194

Figure 24: VIF for LR model

The popular method for identifying multicollinearity is the variance inflation factor (VIF).The GVIF values greater than 10 are not considered in further model(Aheto 2019). There is a no problem with multicollinearity in this situation, which makes it likely that the odd ratio and the coefficient for these variables will be accurate.

When assessing model fit, the standardised residuals are the main area of attention. On average, only 5% should fall outside of 1.96(Tennant and Conaghan 2007). Standardized Residuals(>1.96) add up to 1330, which is less than 5% of 32923. (train data). 295 values (less than 1%) are outside the range of 2.58. Therefore, there is no reason to worry.

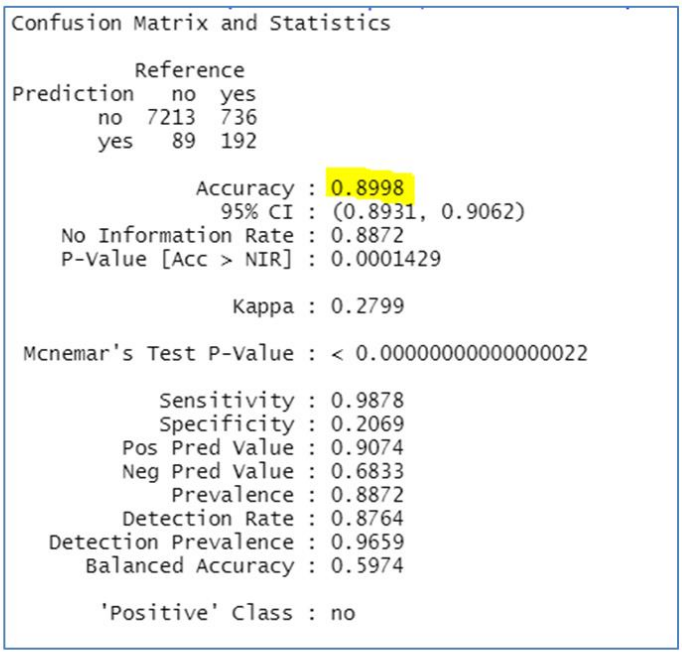

Figure 25: Confusion Matrix and Statistics of the model

There are 197 "yes" entries that were accurately predicted as "yes" and 7213 "no" entries that were correctly forecasted as "no" in the confusion matrix. 89 "no" predictions that were mistakenly made as "yes" and 736 "yes" predictions that were mistakenly made as "no."

Given all the defined customer features, the Logisitic Regression model has an accuracy of 89.9%. indicating a high level of strength in this model's ability to classify the customer response. The comparison of the other LR models is shown in the table below.

Table 2: LR Models

LR Models	Predictors	Residual deviance	AIC	Accuracy
Model 4	month + poutcome+contact + cons.price.idx + cons.conf.idx + nr.employed	18377	18411	89.9%
Model 3	age+ poutcome+month+ emp.var.rate +campaign +previous + euribor3m + cons.conf.idx + nr.employed	18350	18390	89.8%
Model 2	nr.employed + euribor3m + emp.var.rate+ cons.price.idx +poutcome + age + job	18795	18833	89.8%

Model 1	age + job + marital + education + poutcome + month	19812	19882	89.6%

4. Conclusion

In this paper, an Logistic Regression classification model was created to evaluate the association between telemarketing effectiveness and other parameters like clients' information, social-economic conditions and so on. The perception of the campaign's investment would have been influenced by higher education. Most of the customers who responded favourably to the campaign were first-time phone callers(Chang and Chen 1998). Clients are more likely to be people in their 20s to 30s, college students, and seniors. The months of March, December, September, and October should be used for the majority of calls (campaigns), according to future planning. Keep the conversation going as long as you can because long phone calls are more effective.

5. Reflective commentary

Because of the principles I've learned and my prior experience applying them practically in the last assignment, I now have a better understanding of the subject. Regarding the information I have learned from the term's classes, I consider my experience as a beginner in statistical analysis to be quite satisfying. My understanding of its practical applications, which are essential to the analytical topics in which I am interested in specialising, is growing.

6. References

Aheto, J. M. K. (2019). "Predictive model and determinants of under-five child mortality: evidence from the 2014 Ghana demographic and health survey." BMC public health **19**(1): 1-10.

Brennan, L. and W. Binney (2010). "Fear, guilt, and shame appeals in social marketing." Journal of business research **63**(2): 140-146.

Calcagno, V. and C. de Mazancourt (2010). "glmulti: an R package for easy automated model selection with (generalized) linear models." Journal of statistical software **34**(12): 1-29.

Cama, M., et al. (2016). The role of multicollinearity in landslide susceptibility assessment by means of Binary Logistic Regression: comparison between VIF and AIC stepwise selection. EGU General Assembly Conference Abstracts.

Chang, T. Z. and S. J. Chen (1998). "Market orientation, service quality and business profitability: a conceptual model and empirical evidence." Journal of services marketing.

Di Leo, G. and F. Sardanelli (2020). "Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach." European radiology experimental **4**(1): 1-8.

Franke, T. M., et al. (2012). "The chi-square test: Often used and more often misinterpreted." American journal of evaluation **33**(3): 448-458.

Harrell Jr, F. E., et al. (1996). "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." Statistics in medicine **15**(4): 361-387.

Hosseini, S. (2021). "A decision support system based on machined learned Bayesian network for predicting successful direct sales marketing." Journal of Management Analytics **8**(2): 295-315.

Hung, P. D., et al. (2019). Term deposit subscription prediction using spark MLlib and ML packages. Proceedings of the 2019 5th International Conference on E-Business and Applications.

Ito, K. and D. Murphy (2013). "Application of ggplot2 to pharmacometric graphics." CPT: pharmacometrics & systems pharmacology **2**(10): 1-16.

Jiang, Y. (2018). "Using logistic regression model to predict the success of bank telemarketing." International Journal on Data Science and Technology **4**(1): 35.

Kaur, P., et al. (2018). "Descriptive statistics." International Journal of Academic Medicine **4**(1): 60.

Krackhardt, D. (1988). "Predicting with networks: Nonparametric multiple regression analysis of dyadic data." Social networks **10**(4): 359-381.

Lever, J., et al. (2016). "Points of significance: model selection and overfitting." Nature methods **13**(9): 703-705.

Li, W. and D. R. Nyholt (2001). "Marker selection by Akaike information criterion and Bayesian information criterion." Genetic Epidemiology **21**(S1): S272-S277.

Miguéis, V. L., et al. (2017). "Predicting direct marketing response in banking: comparison of class imbalance methods." Service Business **11**(4): 831-849.

Moro, S., et al. (2014). "A data-driven approach to predict the success of bank telemarketing." Decision Support Systems **62**: 22-31.

Nick, T. G. and K. M. Campbell (2007). "Logistic regression." Topics in biostatistics: 273-301.

Refaeilzadeh, P., et al. (2009). "Cross-validation." Encyclopedia of database systems **5**: 532-538.

Smith, T. J. and C. M. McKenna (2013). "A comparison of logistic regression pseudo R2 indices." Multiple Linear Regression Viewpoints **39**(2): 17-26.

Tennant, A. and P. G. Conaghan (2007). "The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?" Arthritis Care & Research **57**(8): 1358-1362.

7. Appendix 1: R Code

```
getwd()

setwd("c:/BHARATGURUNATHANHARIDOSS/Assignment1")

dir()

#Load the Library

library(tidyverse)

library(corrplot)

library(psych)

library(dplyr)

library(ggplot2)

library(gtools)

library(caret)

library(readr)

library(readxl)

library(car)

library(GGally)

library(DataExplorer)

library(explore)

library(ltm)

data<-read_excel("banksv.xlsx")

view(data)

summary(data)

describe(data)


#to Check for NA values

anyNA(data)
```

```
sum(is.na(data))


#outputs will be more readablity:
options(scipen = 150)
# converted to factors if it is character
data<- data %>% mutate_if(is.character, as.factor)
#Addressing Data quality issues
#Data quality issue of age
summary(data$age)
boxplot(data$age)
data$age[data$age>100]<- mean(data$age)
data$age[data$age < 5] <- mean(data$age)
boxplot(data$age) # After cleaning data
summary(data$age)

summary(data$job)
summary(data$marital)

#Data quality issue of default
summary(data$default)
data$default[data$default == 'n' ] <- 'no'
data$default<-as.character(data$default)
data$default<- as.factor(data$default)
summary(data$default)

summary(data$housing)
summary(data$loan)
summary(data$contact)
```

```
#data quality issue of month
summary(data$month)
data$month[data$month == 'march' ] <- 'mar'
data$month<-as.character(data$month)
data$month<- as.factor(data$month)
summary(data$month)

summary(data$day_of_week)
summary(data$duration)
summary(data$housing)
#data quality issue of pdays
summary(data$pdays)
data$pdays[data$pdays == 999] <- 0 #as 999 means not contacted
boxplot(data$pdays)

summary(data$subscribed)
data$subscribed <- relevel(data$subscribed, "yes")

#Data Analysis and Data Visualisation


# plot with target variable:
ggplot(data, aes(x = subscribed)) + geom_bar(colour="black", fill="darkblue") +
  ggtitle(" Target variable (subscribed)") +
  xlab("Did consumer subscribed (yes/no)") + ylab("Density")
plot_boxplot(data,by="subscribed")
#Relationship between the age and subscribed
chisq.test(data$age,data$subscribed)
```

```
boxplot(age~subscribed,
        data=data,
        col=(c("gold","darkgreen")),
        main=" Boxplots for client age with subscribed",
        ylab="Subscribed",
        xlab="Age",
        horizontal = TRUE)
#Relationship between the job and no of people subscribed
chisq.test(data$job, data$subscribed, correct = FALSE)
#ggplot plot of job of all customers
ggplot(data, aes(x = job)) +
  geom_bar(colour="white", fill="lightblue") +
  ggtitle("Jobs of Overall people") +
  xlab("jobs") + ylab("total number of people")
#ggplot plot of job with only people who subscribed
people_yes <- data %>% group_by(job) %>% filter(subscribed == 'yes')
view(people_yes)

ggplot(people_yes,aes(x= job))+ geom_bar(colour="white", fill="darkblue") +
  ggtitle("Jobs of people that subscribed") +
  xlab("Types of Jobs") + ylab("Number of people") +
  theme(
    plot.title = element_text(color="black", size=15, face="bold")
  )
#ggplot plot of job
job_tab <- data.frame(table(data$job, data$subscribed))
colnames(job_tab) <- c("job","response","count")
ggplot(data=job_y_tab, aes(x=count,y=reorder(job,count), fill=response))+
```

```
  geom_bar(stat = 'identity', position = 'dodge')+
  xlab("Number of customers")+ylab("Job distribution with yes/no to subscribe") +
      ggtitle("Marketing Result of Job Distribution")

#Relationship between the marital status and no of people subscribed
chisq.test(data$marital, data$subscribed, correct = FALSE)
#ggplot plot of  marital status of all customers
ggplot(data, aes(x = marital)) +
  geom_bar(colour="black", fill="lightblue") +
  ggtitle("Marital status of people") +
  xlab("jobs") + ylab("Number of customers")
#ggplot plot of marital status of all customers and subscribed rate
ggplot(data = data, aes(x = marital, fill = subscribed)) +
  geom_bar() + ggtitle("Marital status of people") +
  xlab("Martial status") + ylab("Number of customers ")


#Relationship between the education and no of people subscribed
chisq.test(data$education, data$subscribed, correct = FALSE)

table(data$education)
edutable <- table(data$education, data$subscribed)
tab <- as.data.frame(prop.table(edutable, 2))
colnames(tab) <-  c("education", "subscribed", "perc")
ggplot(data = tab, aes(x = education, y = perc, fill = subscribed)) +
  geom_bar(stat = 'identity', position = 'dodge')  +
  xlab("Education") +
  ylab("Percentage") + theme(axis.text.x=element_text(angle = 90, hjust = 0))
```

```
#Relationship between the month and no of people subscribed
chisq.test(data$month, data$subscribed, correct = FALSE)

monthtable <- table(data$month, data$subscribed)
tab <- as.data.frame(prop.table(monthtable, 2))
colnames(tab) <-  c("month", "subscribed", "perc")
ggplot(data = tab, aes(x = month, y = perc, fill =subscribed )) +
  geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) +
  xlab("Month")+
  ylab("Percentage")
#Relationship between the default and no of people subscribed
chisq.test(data$default, data$subscribed, correct = FALSE)

def_tab <- data.frame(table(data$default,data$subscribed))
colnames(def_tab) <- c("default","subscribed","number_of_customers")
ggplot(data=def_tab, aes(x=default,y=number_of_customers, fill=subscribed))+
  geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3)+ xlab("Default")+
  ylab("No of Customers")
#Relationship between the poutcome and no of people subscribed
chisq.test(data$poutcome, data$subscribed, correct = FALSE)
ggplot(data=data, aes(x=poutcome, fill=subscribed))+
  geom_bar(position = 'stack')+
  labs(X="Number of customers",
       y="No of Customers")
# Correlation
plot_correlation(data, type = 'continuous')# Graphical representation of correlation for continuous variables
plot_correlation(data, type = 'discrete') # Graphical representation of correlation for discrete variables
```

```
ggcorr(data, label=TRUE)
#Split data into train and test dataset
set.seed(40387258)
index <- createDataPartition(data$subscribed, p= 0.8, list=FALSE)
train <- data[index,]
test <- data[-index,]
#model1
# To produce a logistic regression
formula <- subscribed ~ age + job + marital + education + poutcome + month
model1 <- glm(formula, data = train, family = "binomial")
summary(model1)
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <-  1 -  dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}
logisticPseudoR2s(model1)
vif(model1)
predictions<- predict(model1, test, type = "response")
class_pred <- as.factor(ifelse(predictions > .5, "yes", "no"))
postResample(class_pred, test$subscribed)
```

```
confusionMatrix(data=class_pred, test$subscribed)
```

#Model2

```
formula2 <- subscribed ~ poutcome + age + job + nr.employed + euribor3m  + emp.var.rate+ cons.price.idx

model2<- glm(formula2, data = train, family = "binomial"(link="logit"))

summary(model2)

logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <-  1 -  dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}

logisticPseudoR2s(model2)
```

#To get Odds Ratios

```
exp(model1$coefficients)
```

#For Confidence intervals

```
exp(confint(model2))
```

#Predicted Probabilities

```
train$predictedProbabilities <- fitted(model2)

head(data.frame(train$predictedProbabilities, train$subscribed))
```

#Analysing the Residuals

```
resid(model2)
```

```
train$standardisedResiduals <- rstandard(model2)

train$studentisedResiduals <- rstudent(model2)

sum(train$standardisedResiduals > 1.96)
```

#Examining Influential Cases

```
train$cook <- cooks.distance(model2)

sum(train$cook > 1)
```

#Examining leverage

```
train$leverage <- hatvalues(model2)

sum(train$leverage > 0.0009)

vif(model2)

predictions<- predict(model2, test, type = "response")

class_pred <- as.factor(ifelse(predictions > .5, "yes", "no"))

postResample(class_pred, test$subscribed)

confusionMatrix(data=class_pred, test$subscribed)
```

#Model3

```
formula3 <- subscribed ~ age+ poutcome+month+ emp.var.rate +campaign +previous + euribor3m +
cons.conf.idx + nr.employed

model3 = glm(formula3 ,data = train,family = "binomial")

summary(model3)

logisticPseudoR2s <- function(LogModel) {

  dev <- LogModel$deviance

  nullDev <- LogModel$null.deviance

  modelN <- length(LogModel$fitted.values)

  R.l <-  1 -  dev / nullDev

  R.cs <- 1- exp ( -(nullDev - dev) / modelN)

  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))

  cat("Pseudo R^2 for logistic regression\n")

  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")

  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
```

```
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}
logisticPseudoR2s(model3)
#To get Odds Ratios
exp(model3$coefficients)
#For Confidence intervals
exp(confint(model3))
#Predicted Probabilities
train$predictedProbabilities <- fitted(model3)
head(data.frame(train$predictedProbabilities, train$subscribed))
#Analysing the Residuals
train$standardisedResiduals <- rstandard(model3)
train$studentisedResiduals <- rstudent(model3)
sum(train$standardisedResiduals > 1.96)
#Examining Influential Cases
train$cook <- cooks.distance(model3)
sum(train$cook > 1)
#Examining leverage
train$leverage <- hatvalues(model3)
sum(train$leverage > 0.0009)
vif(model3)
predictions<- predict(model3, test, type = "response")
class_pred <- as.factor(ifelse(predictions > .5, "yes", "no"))
postResample(class_pred, test$subscribed)
confusionMatrix(data=class_pred, test$subscribed)
#model4
formula4 <- subscribed ~   month + poutcome+contact   + cons.price.idx + cons.conf.idx + nr.employed
model4 = glm(formula4 ,data = train,family = "binomial")
```

```
summary(model4)
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <-  1 -  dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}
logisticPseudoR2s(model4)
#To get Odds Ratios
exp(model4$coefficients)
#For Confidence intervals
exp(confint(model4))
#Predicted Probabilities
train$predictedProbabilities <- fitted(model4)
head(data.frame(train$predictedProbabilities, train$subscribed))
#Analysing the Residuals
resid(model4)
train$standardisedResiduals <- rstandard(model4)
train$studentisedResiduals <- rstudent(model4)
sum(train$standardisedResiduals > 1.96)
sum(train$standardisedResiduals > 2.58)
 #Examining Influential Cases
```

```
train$cook <- cooks.distance(model4)

sum(train$cook > 1)

#Examining leverage

train$leverage <- hatvalues(model4)

sum(train$leverage > 0.0009)

# Examining VIF

vif(model4)

predictions<- predict(model4, test, type = "response")

class_pred <- as.factor(ifelse(predictions > .5, "yes", "no"))

postResample(class_pred, test$subscribed)

confusionMatrix(data=class_pred, test$subscribed)
```