Professor: Myriam Quispe-Agnoli

Students: Hien Thi Thao Do – 11012370

Khanh Kieu Dang – 11012515

TERM DEPOSIT

**Table of Contents**

[Executive Summary 2](#_Toc58446054)

[Data Dictionary 2](#_Toc58446055)

[Data Cleaning 3](#_Toc58446056)

[1. Understanding dataset 3](#_Toc58446057)

[2. Exploring data by visualization 3](#_Toc58446058)

[3. Data Cleaning 4](#_Toc58446059)

[Reducing data dimension 4](#_Toc58446060)

[1. The effect of predictors on term deposit decision 4](#_Toc58446061)

[2. Assessing the correlation of selected variables 5](#_Toc58446062)

[Data Mining Techniques 5](#_Toc58446063)

[1. Finding the best predictive model 5](#_Toc58446064)

[ Logistic Regression 5](#_Toc58446065)

[ K-Nearest Neighbors (k-NN) 7](#_Toc58446066)

[ Naive Bayes Classifier for categorical predictors 7](#_Toc58446067)

[2. Ranking goal: 7](#_Toc58446068)

[3. Profiling goal 8](#_Toc58446069)

[Conclusion 8](#_Toc58446070)

[Recommendation 9](#_Toc58446071)

[Appendix 9](#_Toc58446072)

# 

# **Executive Summary**

Term deposit has been growing in the recent times due to the important role in raising fund and lending capacity of the bank and under customer’s point of view, it also considered as the safe investment. Regarding to the highly potential growth of term deposit product, we have determined to perform the analysis to predict whether clients will decide to subscribe term deposit or not based on given factors. In this report, we will follow the data mining process to figure out the targeted aim. There are several steps in this process and each of them will be clearly explained in the following pages. After performing data mining process which contains three techniques as logistic regression, K-Nearest Neighbors (k-NN) and naive Bayes classifier, the logistic regression model has been selected as the best model to predict whether customer will determine to invest on term deposit or not due to the highest accurate rate. Apart from finding the best predictive model based on the predictor information. We have two other goals are finding a certain percentage of customers that are most likely to invest on term deposit (ranking) and profiling term deposit: finding out which factors are associated with term deposit. We will use regression model for these goals. The business application for this project is to figure out potential group of clients for term deposit products and increase cost efficiency for marketing campaign.

# **Data Dictionary**

There are 21 variables and 41,188 observations in the data set; particularly, 11 categorical variables in the data and 10 numerical variables in the data. There are two parties as company and clients in this project. In term of customer, the predictors are age, education, marital status, housing, etc.… Regarding to company, the regressors are campaign, contact type, duration…

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Position** | **Data Type** | **Description** |
| Age | 1 | Numeric | Age of customer |
| Job | 2 | Categorical | Type of job, including admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployment, unknown |
| Marital | 3 | Categorical | Marital status of customer including divorced, married, single, unknown |
| Education | 4 | Categorical | Education level of customer including: basic.4y, basic.6y, basic.9y, high school, illiterate, professional.course, university.degree, unknown |
| Default | 5 | Categorical | Whether the customer has credit default or not (no, yes and unknown) |
| Housing | 6 | Categorical | Whether the customer has housing loan or not (no, yes and unknown) |
| Loan | 7 | Categorical | Whether the customer has personal loan or not (no, yes and unknown) |
| Contact | 8 | Categorical | Contact communication type including cellular, telephone |
| Month | 9 | Categorical | Last contact month of year (from jan to dec) |
| Day\_of\_week | 10 | Categorical | Last contact day of week (mon, tue, wed, thu, fri) |
| Duration | 11 | Numeric | Last contact duration, in seconds |
| Campaign | 12 | Numeric | Number of contacts performed during this campaign and for this client |
| Pdays | 13 | Numeric | Number of days that passed by after the clients was last contacted from a previous campaign (numeric 999 means that clients was not contacted previously) |
| Previous | 14 | Numeric | Number of contacts performed before this campaign and for this client |
| Poutcome | 15 | Categorical | Outcome of the previous marketing campaign (“failure”, “nonexistent”, “success”) |
| Emp.var.rate | 16 | Numeric | Employment variation rate - quarterly indicator |
| Cons.price.idx | 17 | Numeric | Consumer price index - monthly indicator |
| Cons.cof.idx | 18 | Numeric | Consumer confidence index - monthly indicator |
| Euribor3m | 19 | Numeric | Euribor 3 months rate - daily indicator |
| Nr.employed | 20 | Numeric | Number of employees - quarterly indicator |
| y | 21 | Binary | Whether the client has subscribed a term deposit or not? |

# **Data Cleaning**

## **Understanding dataset**

As we can see from the summary statistics table (appendix 1.1) for numerical variables below, the median of duration is much lower than its mean which can indicate that the distribution of duration is left skewed. Regarding to variability, it seems that when comparing min, max of each variable, age, duration, campaign and pdays have a large range. Moreover, the standard deviation of duration, pdays and nr.employed are large so we can get conclusion that these variables have large variability. However, from data dictionary it appears that the value 999 of pdays means that clients was not contacted previously so we should treat these observations in different way. There are no missing values in the data based on the table is illustrated in the appendix 1.2.

## **Exploring data by visualization**

Regarding to bar chart of each variable, we are not only assessing on the missing values, inconsistent and invalid values but also on distribution of each variable. For instance, regarding to the chart in the appendix 1.3 there are 11 variables and small proportion of unknown value in customer’s job column and admin job has the highest percentage compare to other carriers. The rest of visualized graphs for other variables are presented in the appendix 1.4.

* The relation of each predictor on the other

In term of the relation of marital status and job (appendix 1.5), there is higher proportion of married customer in the portfolio compare to the other groups as single and divorced and most of them work as admin, blue-collar and technician. The other relations of customer job and customer education to personal loan and the summarizing data are presented in the appendix 1.6.

* Multidimensional visualization

The bar chart shows the effect of duration on term deposit (appendix 1.7), it shows that the average of duration is higher with term deposit yes than the average of duration with term deposit no. It means that the higher duration, the more probability that customer will subscribe term deposit. The second bar of the effect of number of contacts on term deposit, it shows that if customer subscribes the term deposit, the company needs only to contact them two times. The panel chart of previous contact on term deposit shows that most of our customers were not contacted in previous marketing campaign. Besides, it also appears that there were more likely that customer contacted in previous marketing campaign will subscribe the term deposit. The panel chart of previous outcome campaign shows that if the previous outcome was successful, there is more likely that customer will subscribe term deposit. The scatterplot (appendix 1.8) shows that there is relationship between number of contact and contact duration. As we can see, when the number of contacts is high, the contact duration is low, and customer didn’t subscribe the term deposit. The box plot of term deposit and euribor 3-month rate and number of employees heatmap (appendix 1.9) shows the correlation matrix between numerical variables are presented in the appendix 1.10, it appears that emp.var.rate vs euribor3m, emp.var.rate vs nr.emplyed are strongly correlated. Therefore, we will not include these variables at the same time in the model to avoid multicollinearity.

## **Data Cleaning**

Proportions of yes and no in Term Deposit in the dataset including all unknown observations are 88% and 12%, respectively. After removing all unknown observations, its proportions are 87% for “no” term deposit and 13% for “yes” term deposit and the dimension of dataset is 30,488 records which is large enough to build predictive model . Therefore, we can remove all unknown observations without affecting the results. For grouping the variables, there are 11 dummy variables in job, we grouped these dummy variables into 5 groups, include self-employed (self-employed, entrepreneur), white-collar (admin., management, services, technician), blue-collar (blue-collar, housemaid), retired, and unemployed (unemployed and student). After grouping all values, the job variable has only 5 values with white-collar customer accounting for the highest proportion. While retired customers occupy the lowest percentage. Besides, there are 7 dummy variables in education variable, we grouped these dummy variables into 4 groups, include below high school (illiterate, basic.4y, basic.6y, basic.9y), high school, professional course, and university degree. After grouping, we can see that there are only 4 values in the education variables with university accounting for the highest proportion. The month variable has been grouped into 4 seasons including spring, summer, fall and winter. “Pdays” variable: which is the number of days that passed by after the client was contacted from a previous campaign. From the histogram in the appendix 1.11, the range of number of days is from 0 to 27 and value “999” and there is no value in the middle 27 and 999. Keeping the “999” value which can be impact on the range. Therefore, we will change the value “999” in this column to “0”. It means the customers were not contacted from the previous campaign. Additionally, we have performed checking on the “Poutcome” variable and the result from last campaign for those “999” values is that whether “failure” or “nonexistent” customer which prove consistent outcome between “Pdays” and “Poutcome” variables.

# **Reducing data dimension**

## **The effect of predictors on term deposit decision**

From the figure in the appendix 2.1, the job type has impact on term deposit decision. People who are retired or unemployed have more tendency to invest money than other people. In the reality, those people tend to take the lower risk on investment and earn monthly interest from term deposit compare to who are working. Therefore, we will select this variable as predictor. In term of marital status, it appears that people who are single have more plan to invest money in term deposit than other people with divorced or married status. It can be explained that divorced or married people have more fees to expend such as fee for their children, they don’t have much saving money to invest on term deposit than single people so we will include marital status in the predictive model. The other effects of independent variables on term deposits are detailly illustrated in the appendix 2.2.To sum up, according to the analyzing above, the categorical variables that we include in our model are job, marital, education, contact type, month and poutcome. However, It is appear that job and education might be correlated so in the part below we will check its correlation.

## **Assessing the correlation of selected variables**

From the chart of customer job and education in the appendix 2.3, there is a little correlation between education and job. Most people who below high-school education have blue-collar jobs and only a few of people have university degrees work in this field. In the contrast, for the white-collar job, the highest proportion in this job is the people who have graduate from university and second and third place are the people who graduate from high school and have professional course certificate, respectively. Converting job categorical variable into ordinal categorical variable with 4 levels as: below-high school: 1, high.school:2, professional.course: 3, university.degree: 4 and job variable as: retired: 1, unemployed: 2, blue-collar: 3, self-employed: 4, and white-collar: 5. It also appears that there is little correlation between these two variables based on the correlation matrix table in the appendix 2.4. Therefore, we will include both job and education in the model. For the numeric variables as age the mentioned boxplot of age shows that the median and the range of term deposit “yes”, “no” is the same so we will not include in the model. Duration and campaign, from scatterplot between duration and campaign in part data visualization, the duration and campaign have impact on term deposit decision. In particular, when the number of contacts is high, the contact duration is low, and customer didn’t subscribe the term deposit. Therefore, we will include these two variables in the model.. For the variables which are euribor3m, nr.employed and em.var.rate: From the side-by-side boxplot in part visualization, we see that the medians of no and yes decision are different with both euribor3m and nr.employed. However, in the heat map, it shows that mp.var.rate vs euribor3m, emp.var.rate vs nr.emplyed are strongly correlated. Therefore, we will run 2 models, one model includes only emp.var.rate and one model includes euribor3m and nr.emplyed and then we will compare each model to select the best one. For the other variables that we decide to not include in the model, we have explained the reason on the appendix 2.5.

# **Data Mining Techniques**

## **Finding the best predictive model**

In fact, the records are imbalances, particularly there are 13% customers investing on term deposit, so we determine to oversampling data. The training data includes 50% of customer investing on term deposit and 50% of customers not investing on term deposit. The proportion of “yes” and “no” in term deposit in the validation data is the same as original data with 13% of yes and 87% of no.

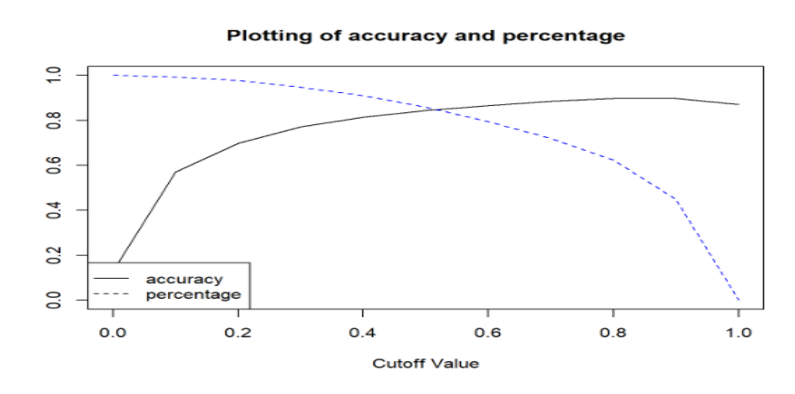
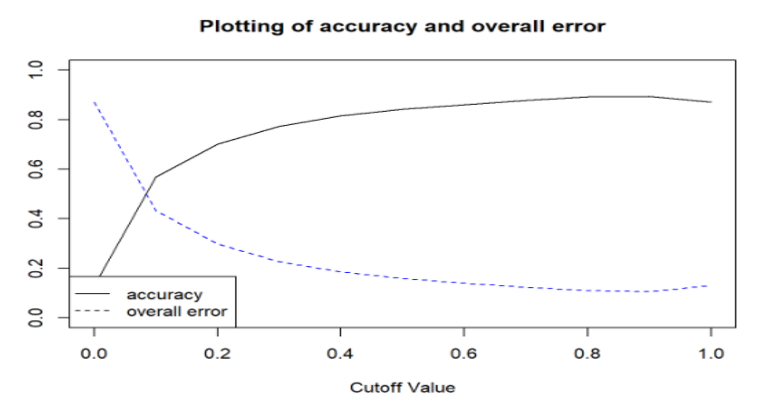
### **Logistic Regression**

We run two logistic regression models with selected predictors above and then used forward selection, backward elimination, and stepwise regression to select the best model basing on three our goals. The outcome from model (appendix 3.1). Basing on parsimony that higher insights with fewer predictors, we see that model 1 and model 2 with backward elimination seem to be the best model. Therefore, we will check the accuracy rate of these 2 models to select the best predictive one for our main goal.

Evaluation of accuracy rate:

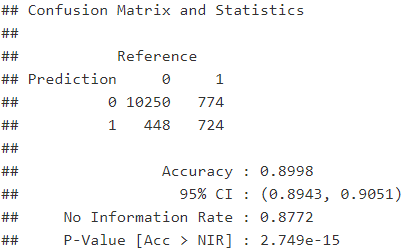
Firstly, we will run the prediction on each model to find the probability and then we will use different cutoff values to define the classification. Finally, we use the overall figure and confusion matrix to pick up the best model. From the plotting of accuracy rate and overall error based on cutoff value (appendix 3.2), we have the cutoff equals to 0.85 will brings the highest accuracy rate. For instance, the prediction table (appendix 3.3) represents six samples of clients who decide to accept the term deposit or not from the predicted model and if we choose the cut off value at 0.85 which means all customers will subscribe the term deposit. After evaluating the performance of two models, we can see that the accuracy rate of model 1 (appendix 3.4) is a little bit better than that of model 2 (0.898 compared to 0.894). Therefore, we will use model 1 as the best model with cutoff value at 0.85.

In case the aim is to predict the number of customers who will decide to invest on the term deposit, the model is good enough in general to classify customers with the accuracy rate at 89%. However, in the real word, the company will want to spend marketing cost to those who are more likely to purchase Term Deposit. Therefore, determining customers who are willing to purchase term deposit is more important. In the model 1 at cutoff value at 0.85, among customers who purchased term deposit, the percentage of customers correctly identified to purchase term deposit is 1050/(1050+880) = 1050/1930 = 54.4% which is quite low, so we have to reduce the cutoff value to capture more customers who are willing to buy the term deposit even the accuracy rate is lower. As we can see with the cutoff value at 0.5, the percentage of customers correctly identified to purchase term deposit increases: 1655/(1655+275) = 85.75% even the accuracy rate is lower at 84%. Overall, the accuracy rate at 84% is still good. For more illustration, I plotted the relationship between the accuracy rate and the percentage of customers correctly identified to purchase term deposit with cutoff values from 0 to 1.



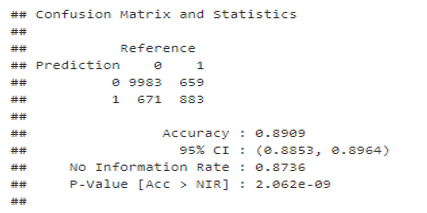
### **K-Nearest Neighbors (k-NN)**

Another algorithm for classifying customer’s propensity is KNN. After running KNN with different K values. We figure out that the value of K at 13 gives the best predictive model with accuracy rate of 0.900. This accuracy is a little bit higher than the accuracy of model 1 (appendix 3.8). However, in case the defining the customers who purchased the term deposit, the performance of KNN model is not good, with accuracy rate at 724/(724+774) = 48.33%.



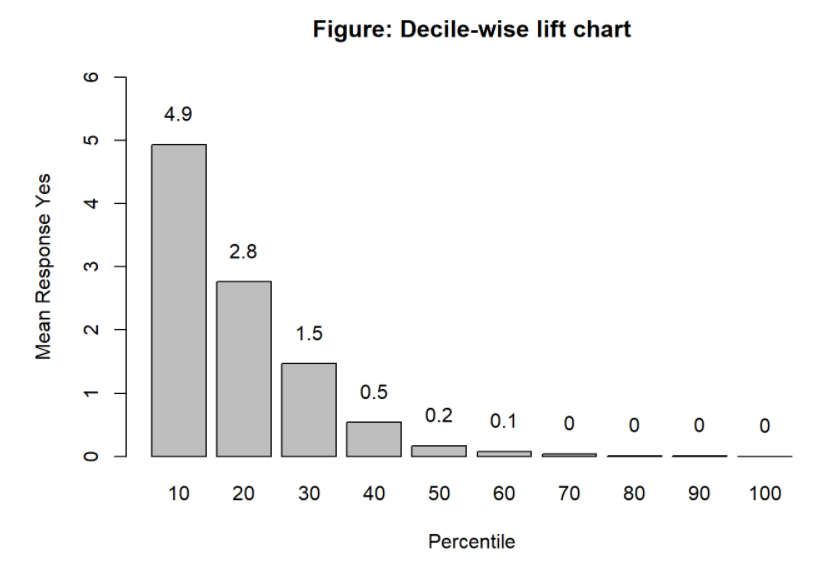
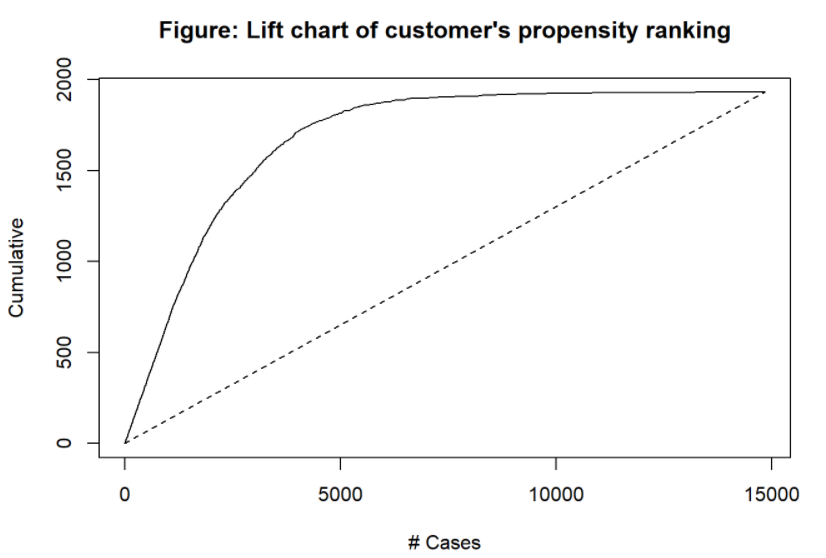
### **Naive Bayes Classifier for categorical predictors**

Using Naive Bayes algorithm with predictors of model 1.backward elimination (appendix 3.9): job, education, month, contact type, poutcome, duration, euribor3m, and nr.employed. For numerica variables such as duration, euribor3m, and nr.employed, I created a bin and changed these variables into factors. After running Naive Bayes with naiveBayes() function in R on training data, I evaluate the model on validation data by confusion matrix. From the confusion matrix (appendix 3.10), the accuracy rate of Naive Bayes model is 0.8909 which indicates this is also a good model for forecasting, however its accuracy rate is a little bit less than that of Model 1 using logistic regression and KNN model.



## **2. Ranking goal:**

From decile chart: the first column shows that for the first 10% of observation, we can get more than nearly 5 times in term of “yes” as well as just random selection. The second bar chart shows that the second 10% is giving us close to a 2.8 times in term of “yes decision” compare to random selection. We can see the numbers more clearly in the lift chart. From lift chart: if we select top 10% (1,484) customers with the highest probability of buying term deposit, the number of customers who will buy term deposit can be 945 customers (curve line). If we select 10% customers randomly, the number of customers who will buy term deposit is 192 customers (line-base). From that point, we can see that the model with lift chart and decile-chart can help the company reach as many as targeted customers in effective costs.



## Profiling goal

We are more interested in finding out which factors are associated with customer’s decision on buying term deposit, and for those factors we would like to quantify these factors. Thus, we will find the model that can fit the data best basing on Deviance. Regarding to the result (appendix 3.11) It appears that model 2.forward selection with predictors: job, marital status, education, contact type, month, poutcome, duration, campaign, and emp.var.rate is the best model with the lowest residual deviance (2717) and the lowest RMSE (0.3229). In particular, the coefficients of unemployed is 0.83, exp(0.83) = 2.29 (>1) are the odds of unemployment customers on buying term deposit. This means that the unemployment customers are more likely to purchase term deposit than blue-collar customers. Likewise, the coefficient of retired customers is positive (>0) that means the retired customers are more likely to purchase term deposit than blue-collar customers (holding other predictors constant). The p-values of self-employed customers and white-collar customers are insignificant that means the probability of self-employed and white-collar customers on buying term deposit is not statistically significant when compared to the probability of blue-collar customers on buying term deposit. The coefficient of customers who have university degree is 0.507 which is the highest and greater than 0, that means university customers have the highest probability of purchasing term deposit comparing to other different educational customers (holding other predictors constant). The coefficient of emp.var.rate is -0.941 (numerical variable), thus exp(-0.941) = 0.39 are the odds of employment variation rate of customers buying term deposit, that means e(-0.941) = 0.39 are the multiplicative factor by which odds of customers buying term deposit will decrease when the employment variation rate increases by 1 unit. In other words, the Beta of emp.var.rate < 0 (negative), thus the odds of buying term deposit less than 1, it indicates that the higher employment variation rate, the lower probability of buying term deposit. Likewise, to other predictors, we figure out that factors that are associated with buying term deposit is the customers who are retired, have the university degree, have successful outcome in previous campaign and the bank should contact them in spring, they are the most likely to buy term deposit.

# **Conclusion**

In conclusion, the below table illustrate the accuracy rate and percentage of customers that is correctly identified to invest on term deposit. According to the comparing result from those models, there are two models that we consider the best for two common business applications in term deposit. Firstly, in case the main business target is figuring out which customers tend to subscribe on term deposit, which is considered as the most popular application, the logistic regression model at cutoff value 0.5 is the best model. In case, the aim is not emphasized on class of purchase term deposit, the K-NN model is the best.

|  |  |  |
| --- | --- | --- |
|  | **Accuracy Rate** | **Percentage of customer correctly identified to purchase term deposit** |
| **Logistic Regression Model at cutoff value 0.85** | 89.82% | 54.40% |
| **Logistic Regression Model at cutoff value 0.5** | 84.39% | 85.75% |
| **K-Nearest Neighbor** | 90% | 48.33% |
| **Naïve Bayes** | 89.09% | 57.26% |

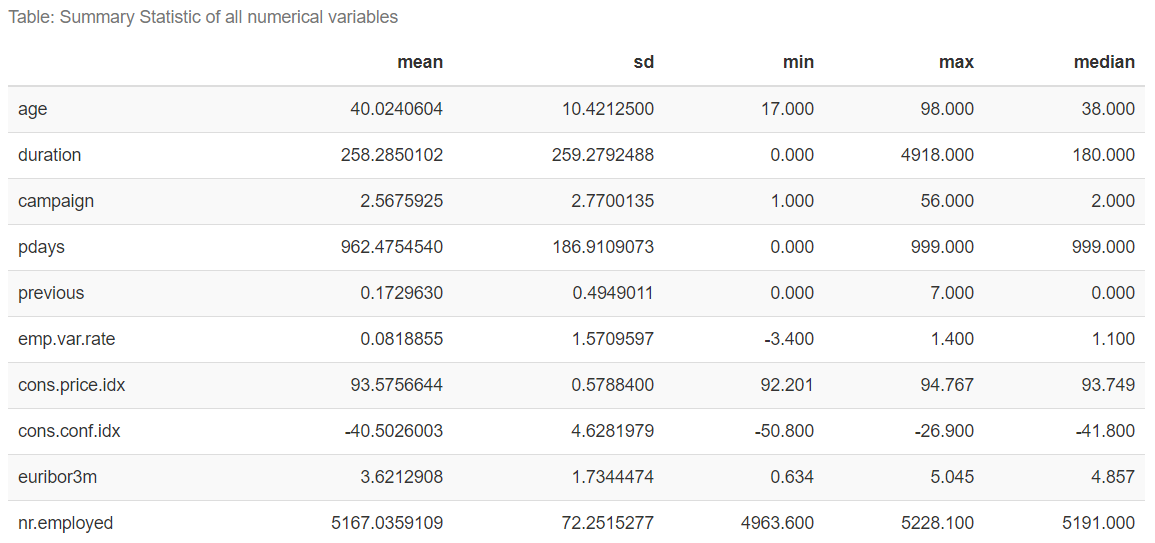
# **Recommendation**

From the analysis, we identify the factors which are associated with buying term deposit as customers who retired, having the university degree and spring is the best time to contact customer. In order to have effective plan for increasing sales of term deposit, the elements that have been identified should be take into consideration. For instance, company should run the marketing campaign from wintertime to springtime since customers usually want to get information before being contact or deciding to invest in spring season. Since potential customer are retired and having university degree, the campaign should be hold in the university during college sport event happen. The reason is that many people who graduate from university tend to support for the sport team from their college, thus, the campaign will be catch the attention from audience.

# **Appendix**

1. **Data Cleaning**

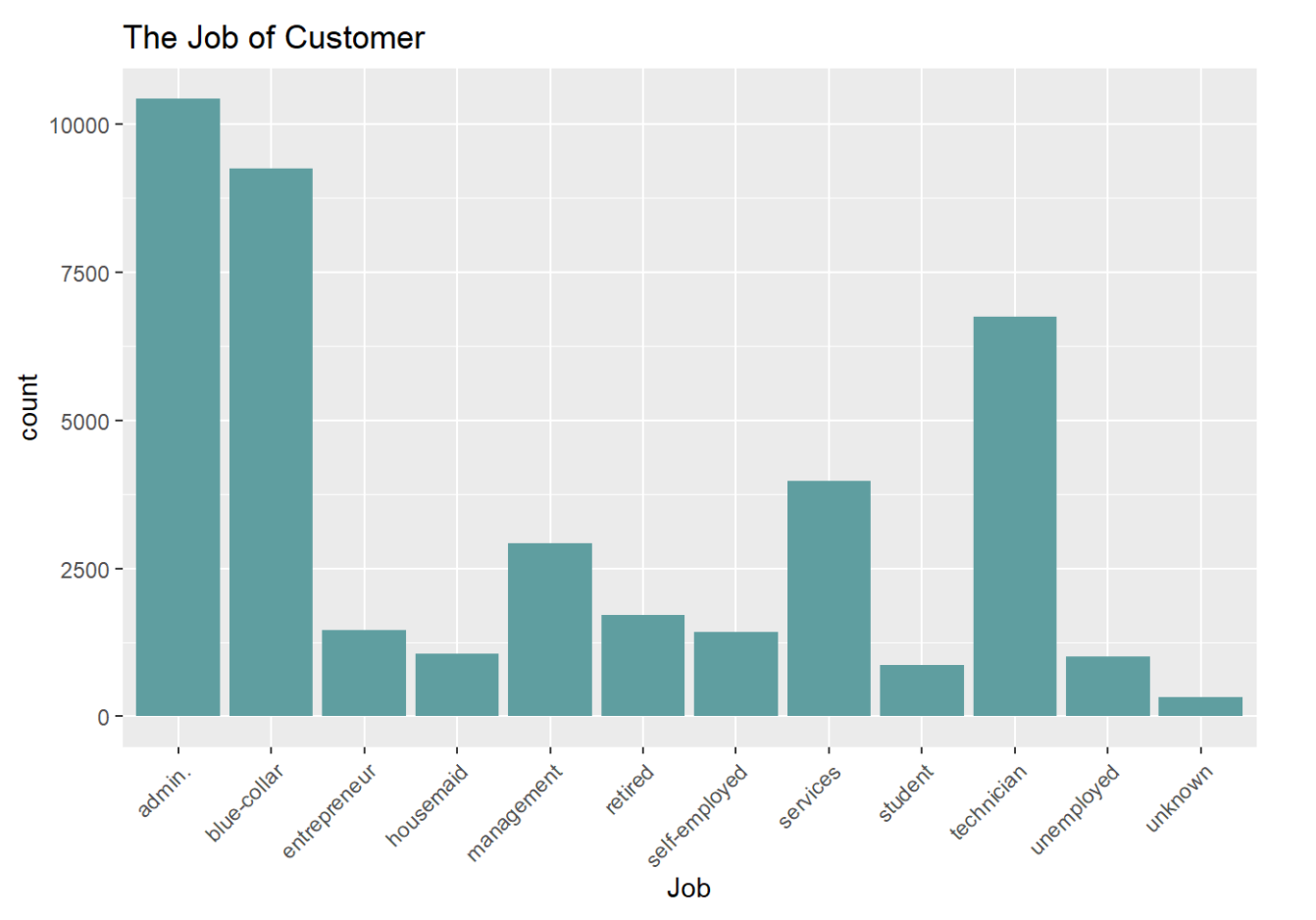
Appendix 1.1



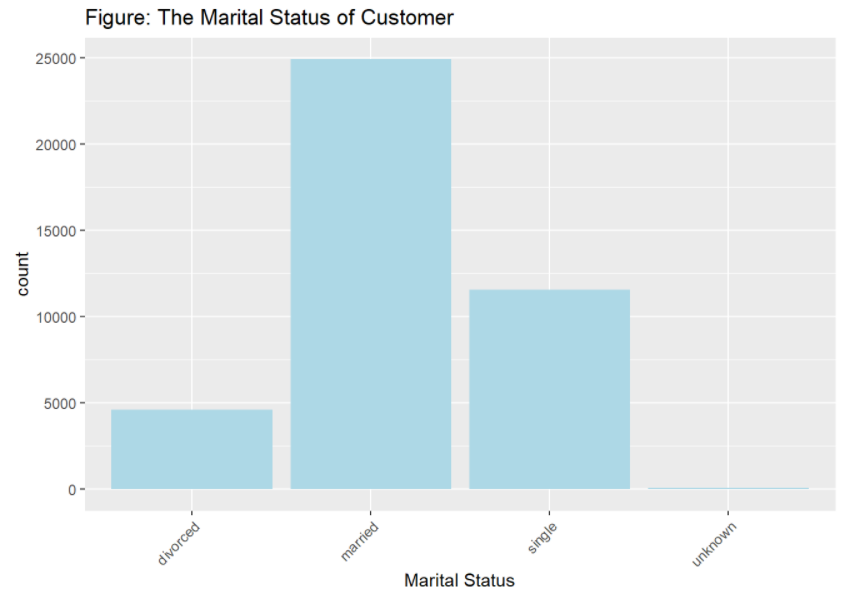
Appendix 1.2. Counting the missing values

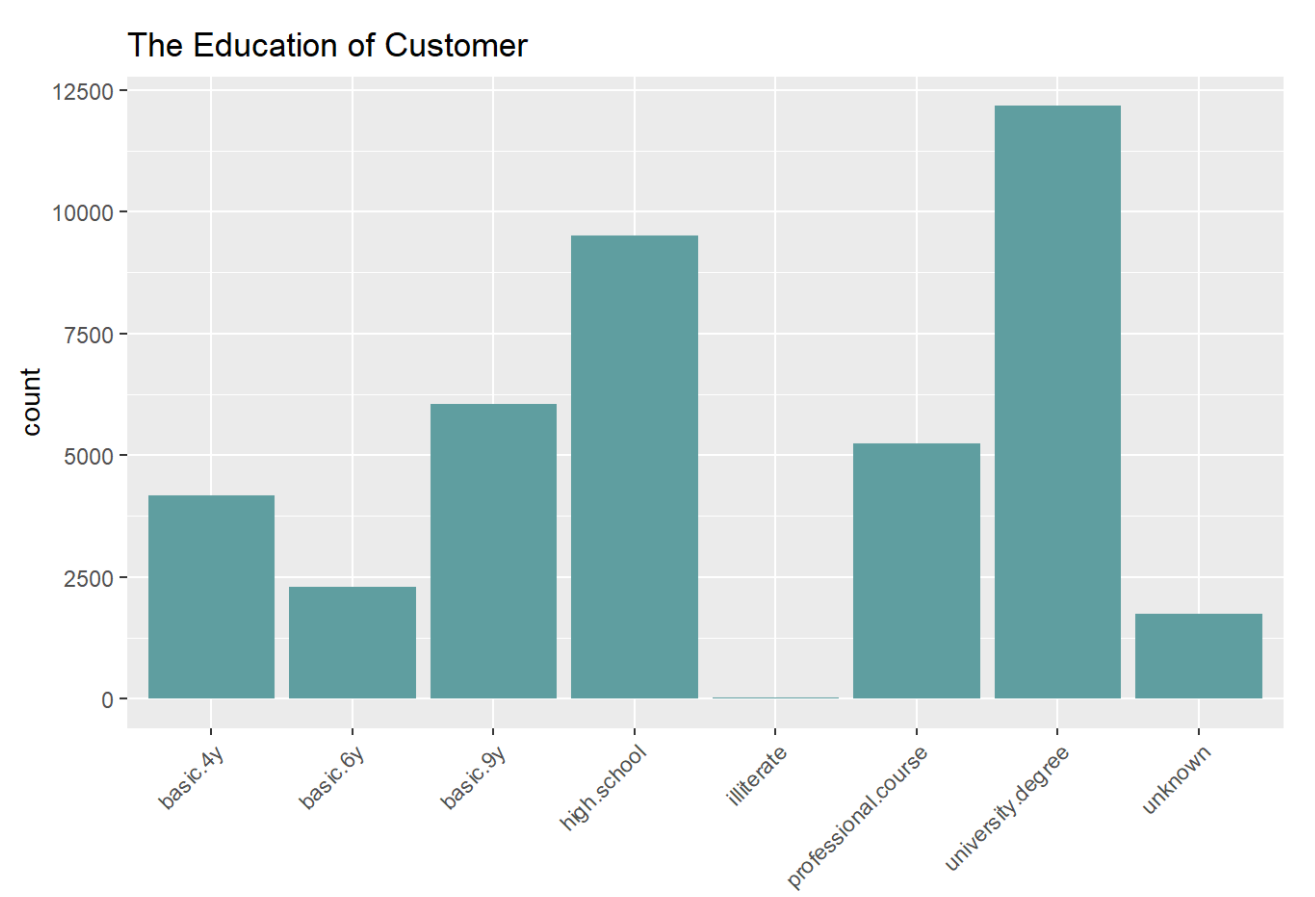


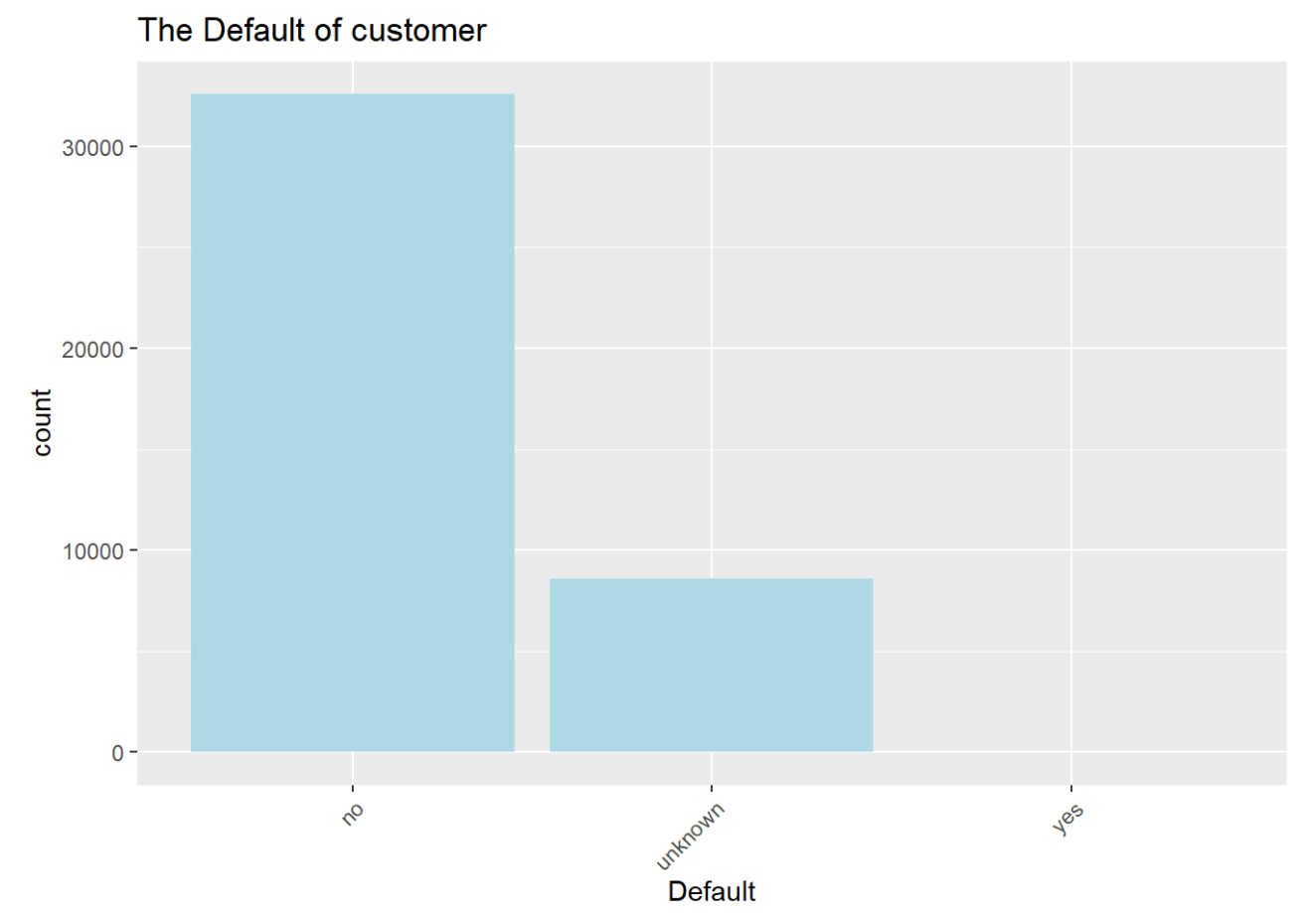
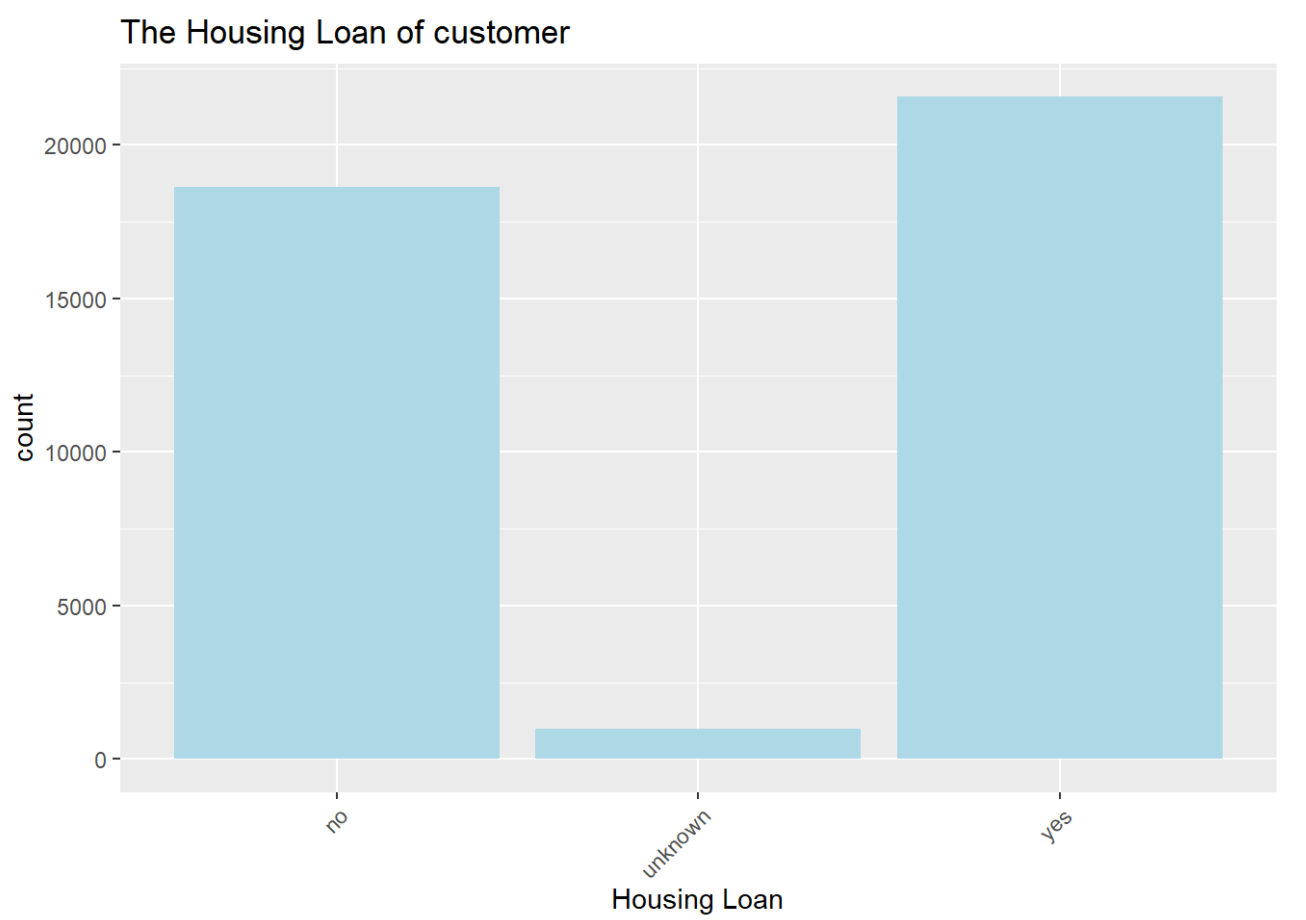
Appendix 1.3:the chart of customer’s job and marital status

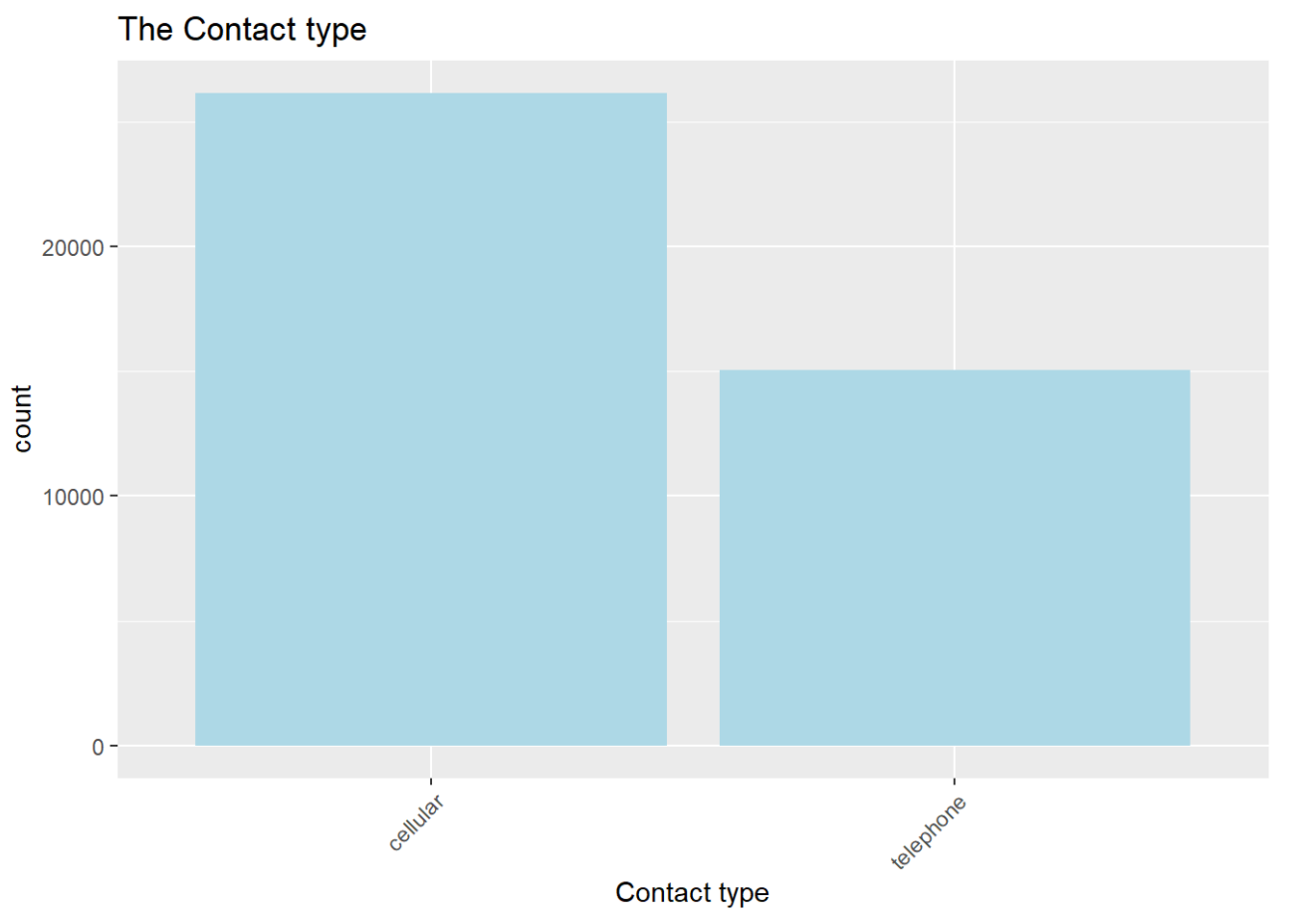


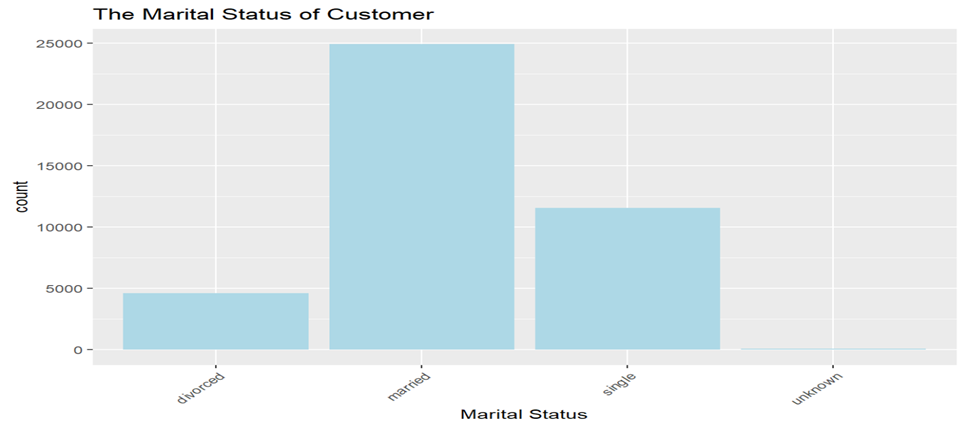
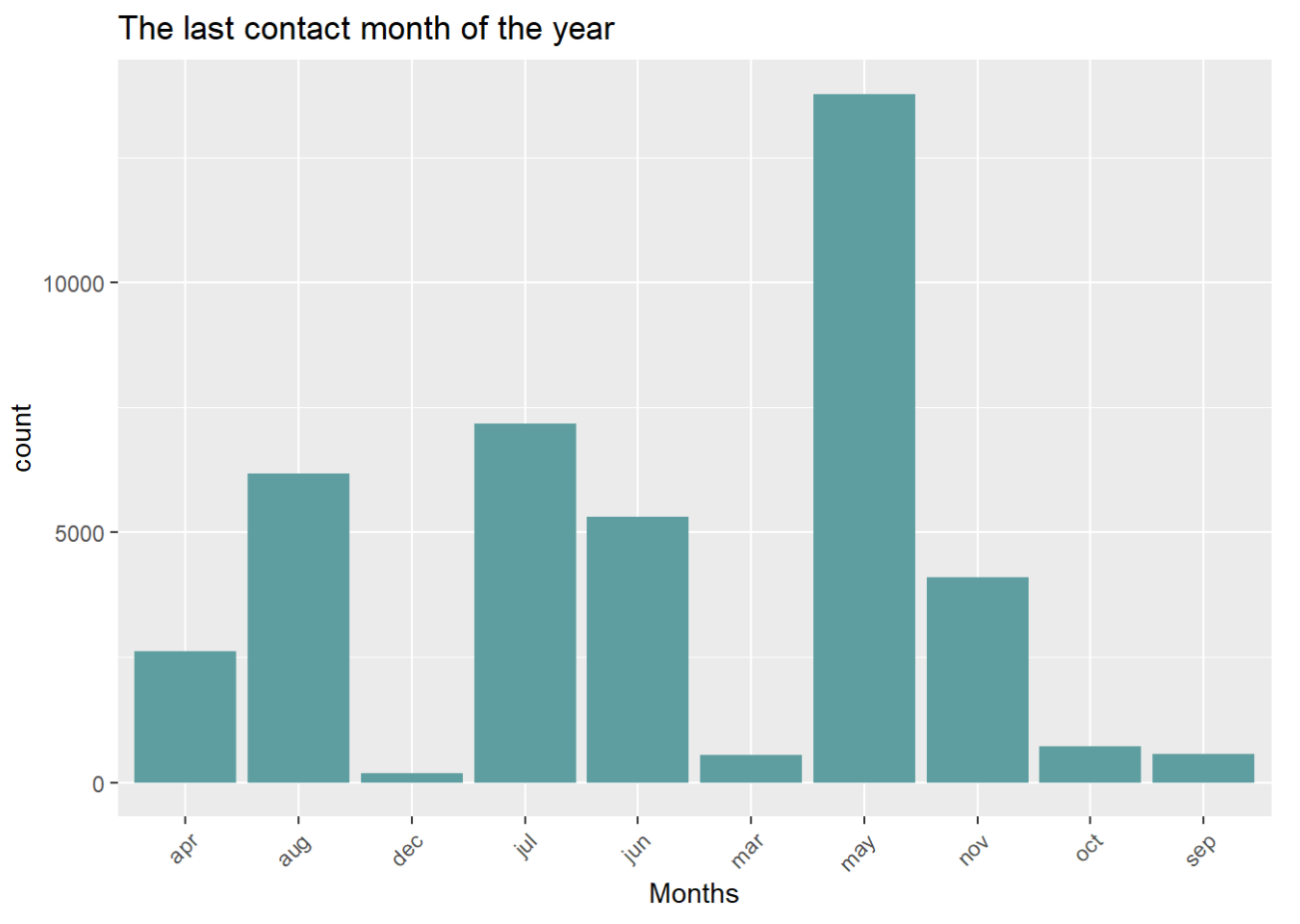
Appendix 1.4: Exploring data on the variables marital status, education, credit default, housing loan, personal loan



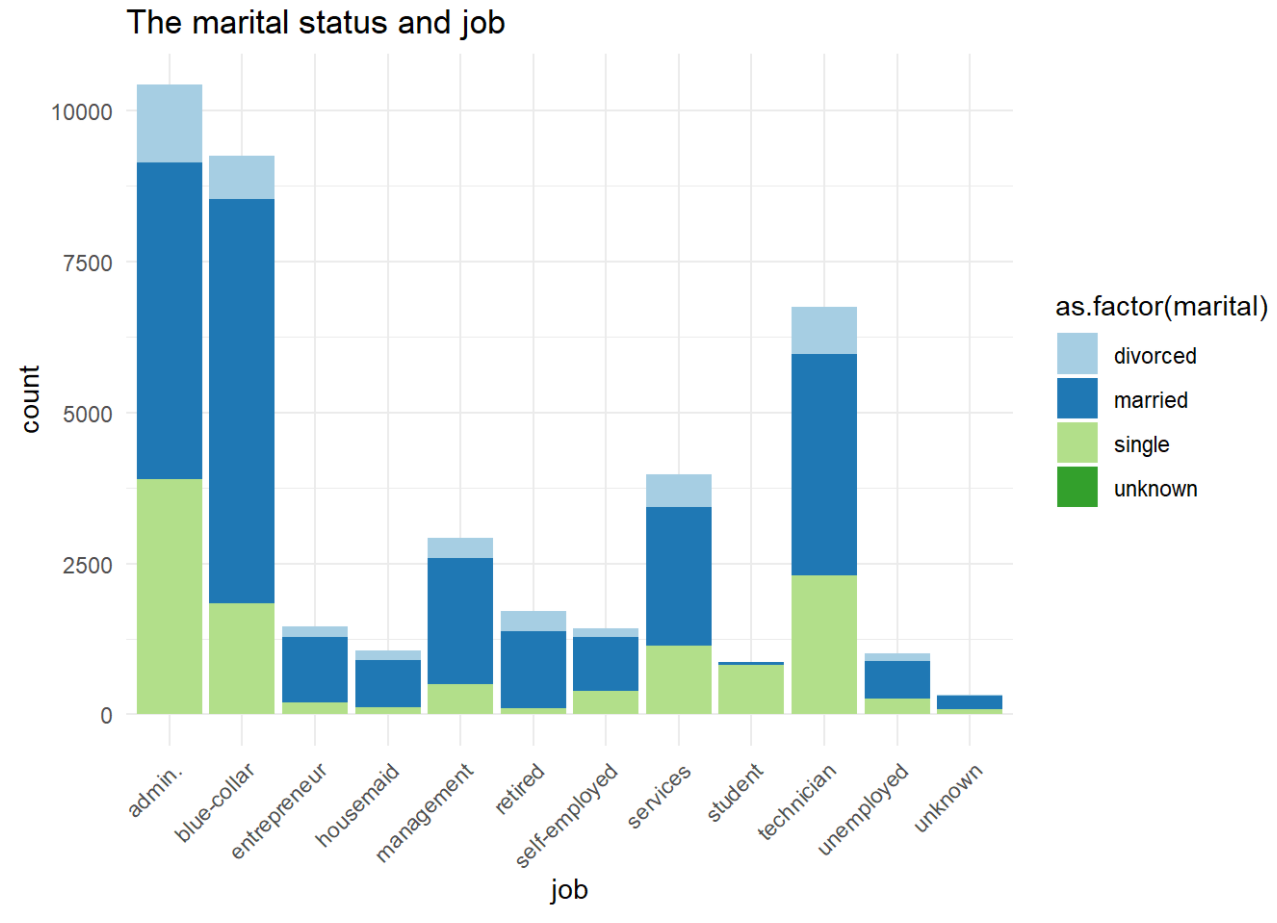




Appendix 1.5. The relation of each predictor on the other: marital status and job

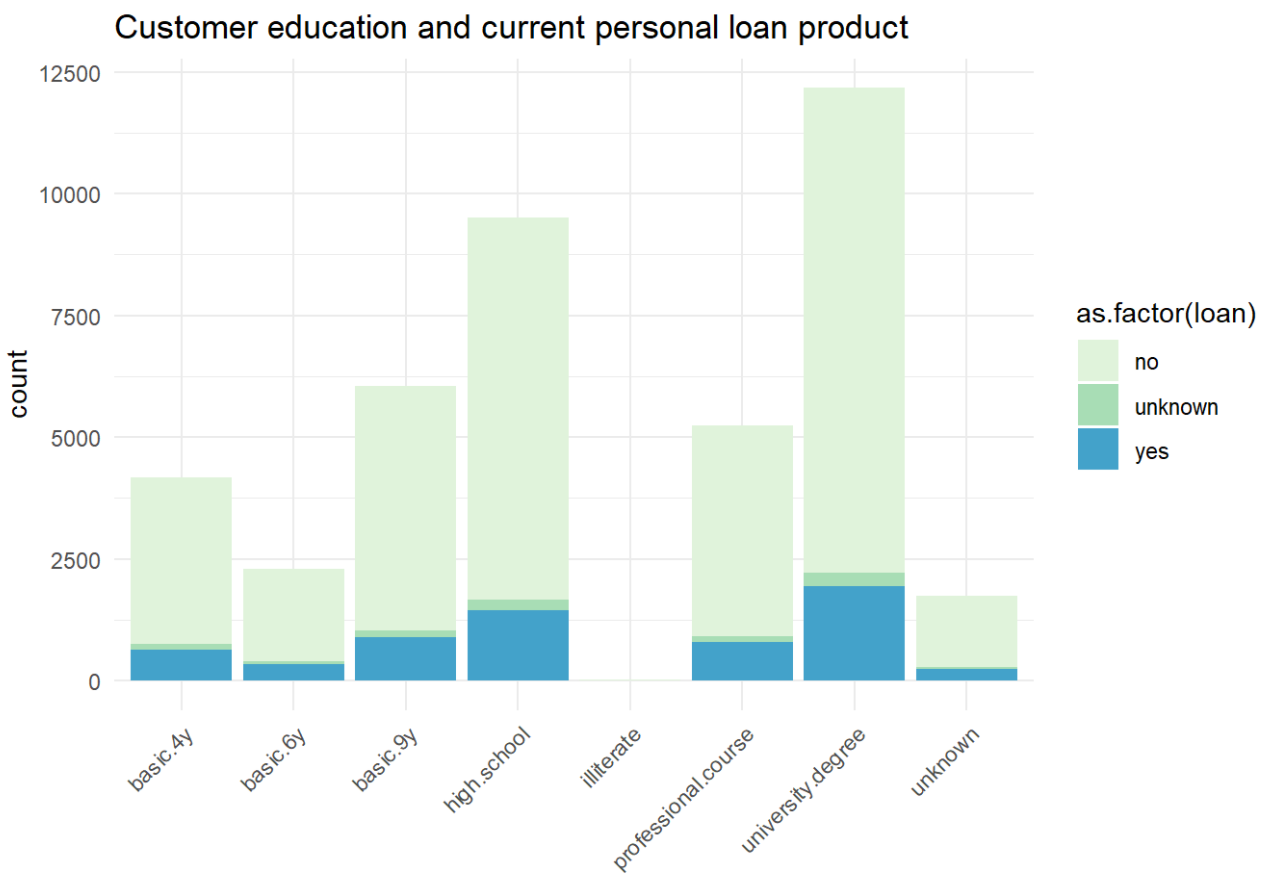


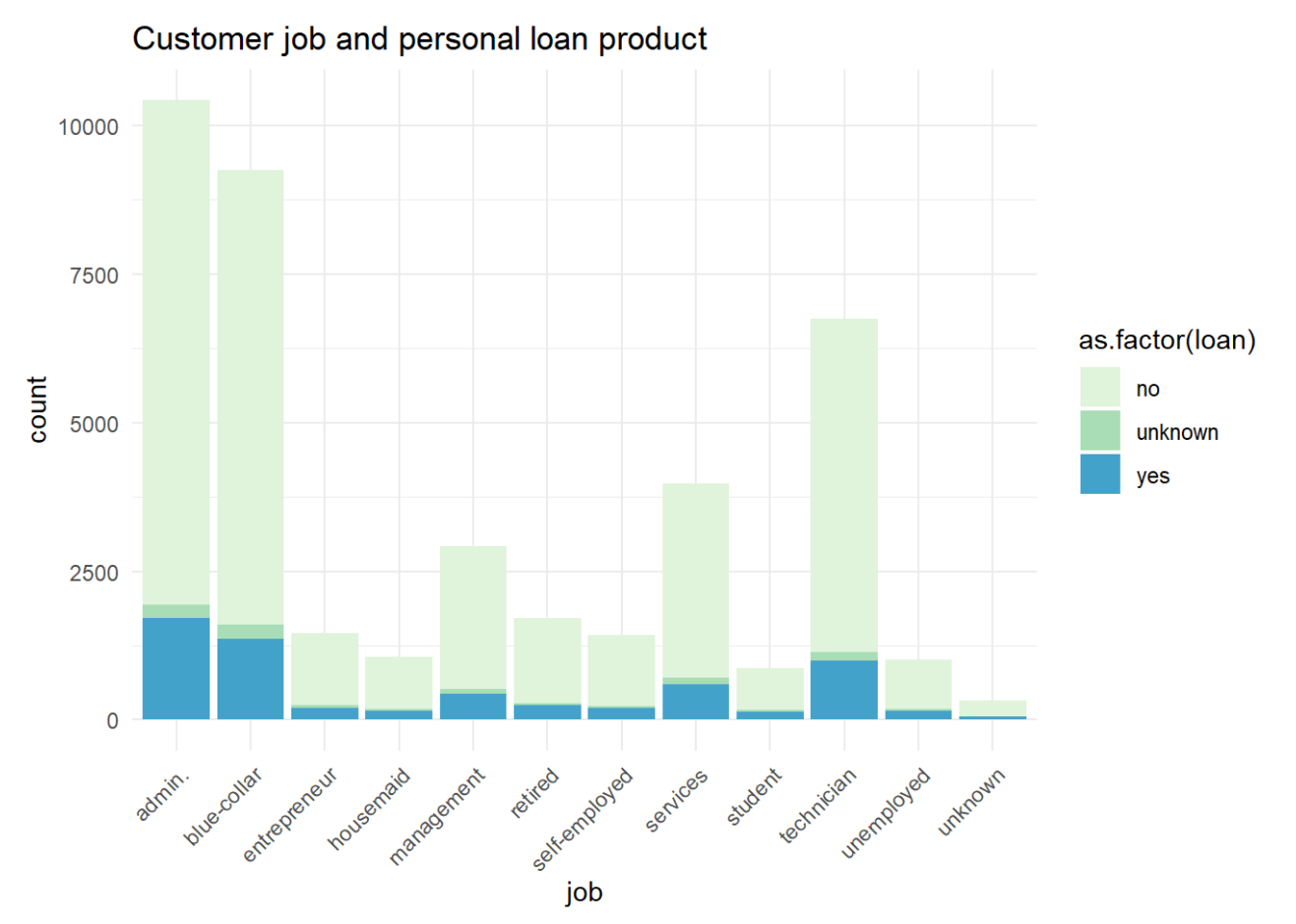
Appendix 1.6

The average and minimum age of customers who decide to choose the term deposit or not are similar as 37 years old and 31 years old, respectively. However, the age range of people who tend to apply term deposit is larger than who are not, whose age is up to 50 years old.

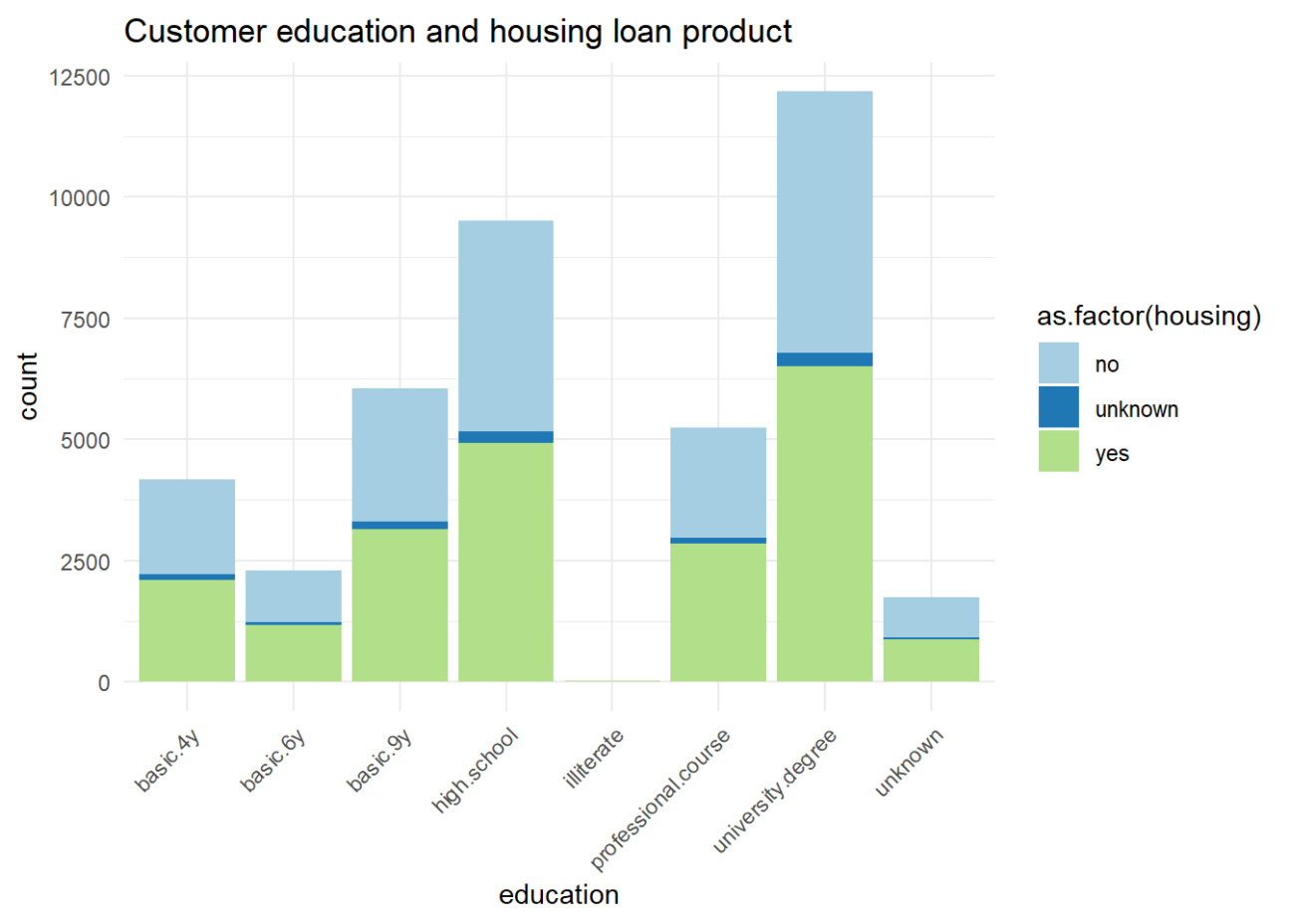


There is significant lower percentage of clients have applied for term deposit compare to those are not which across all groups of customers based on job and education level.



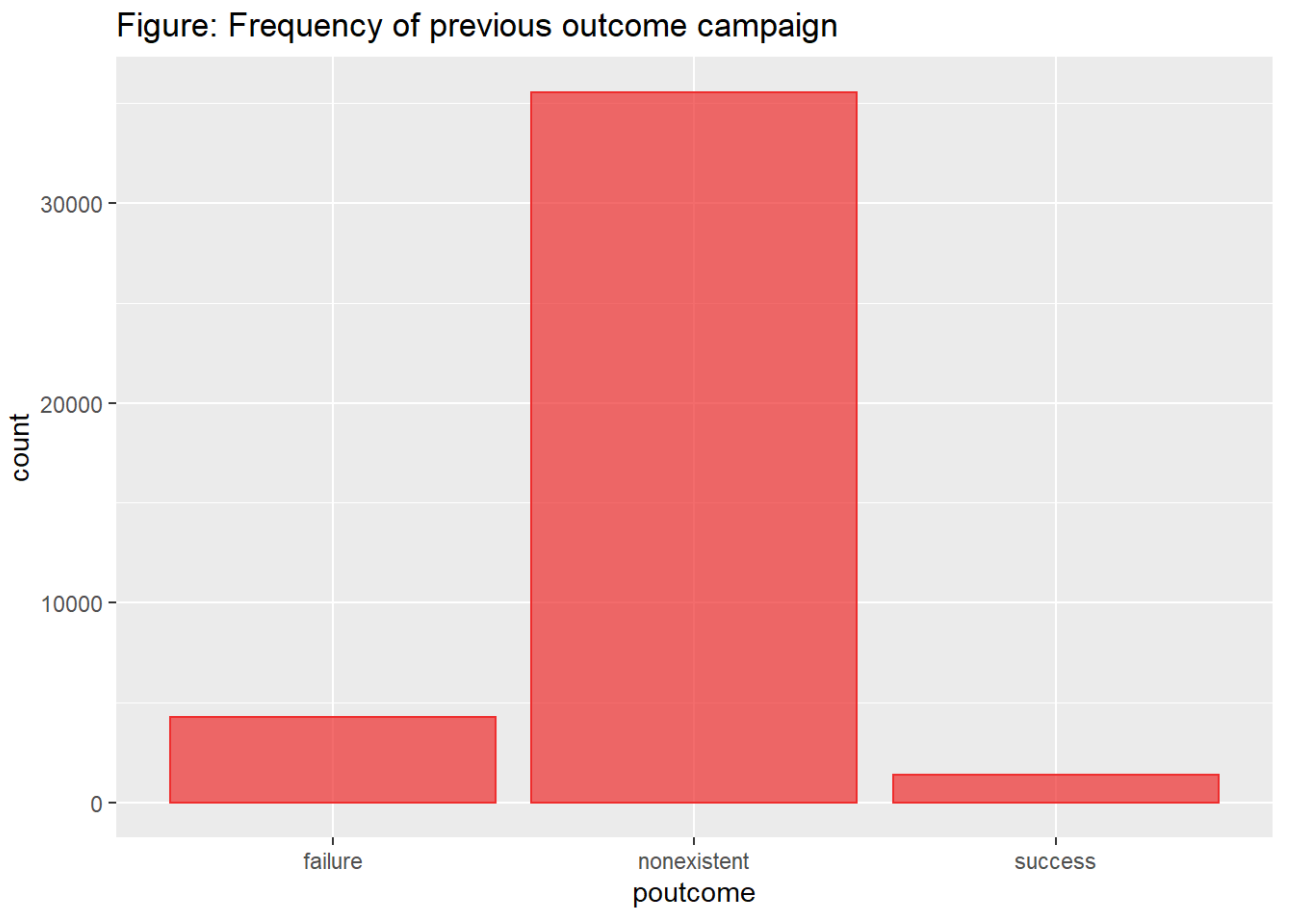
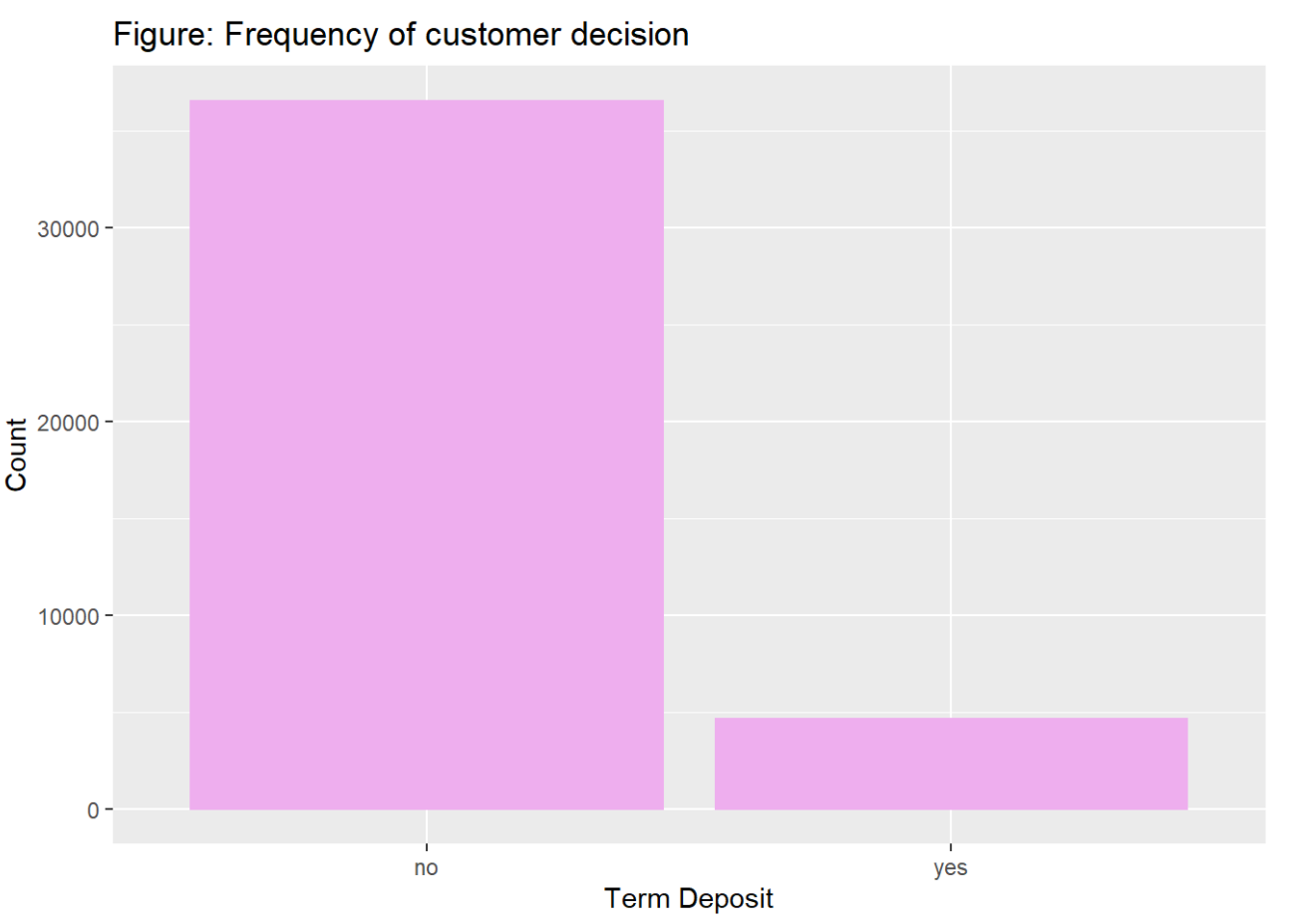


There is equalized in the number of customers who have the housing loan and those are not regardless the education level.

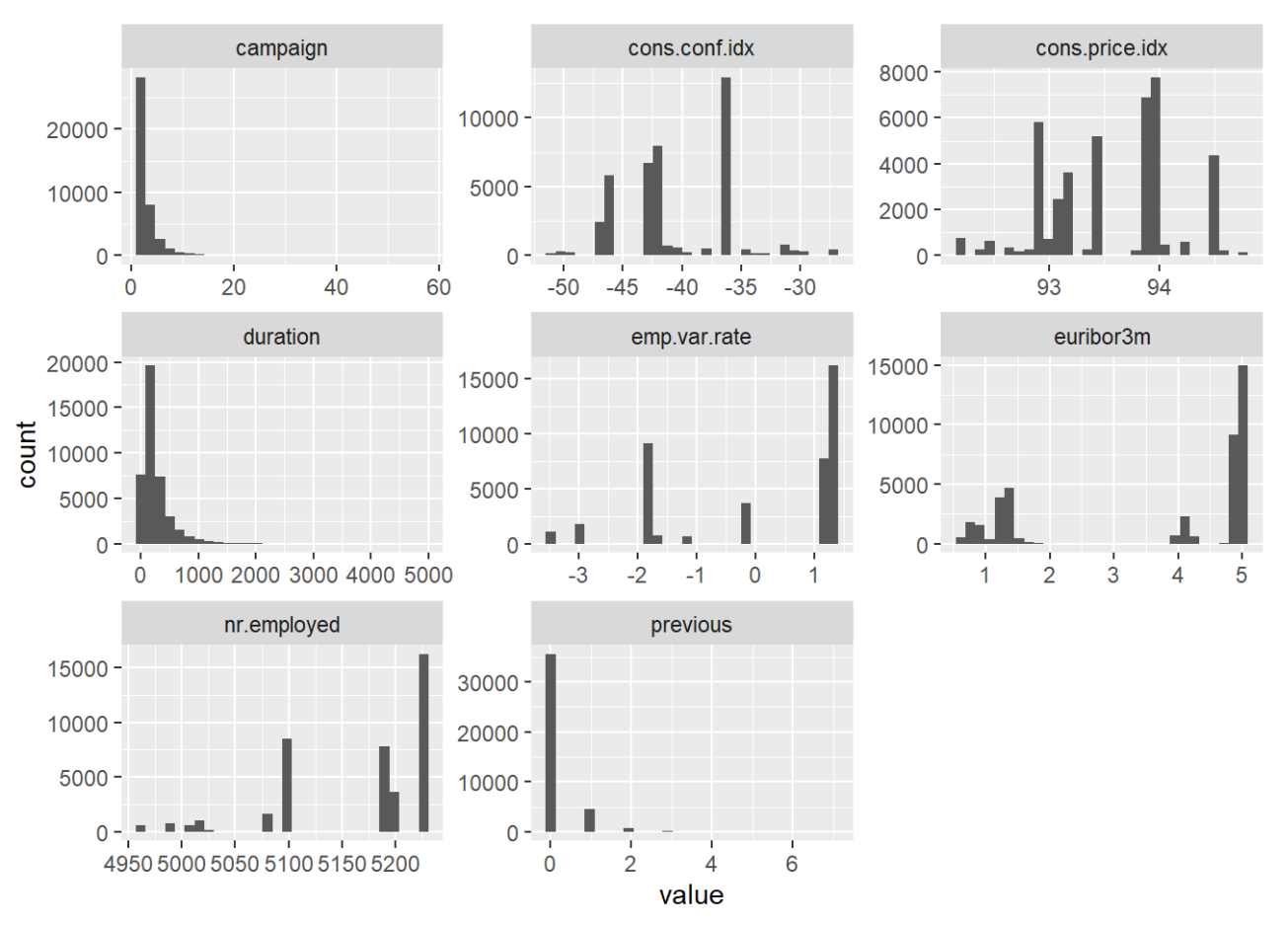


* Summarizing data

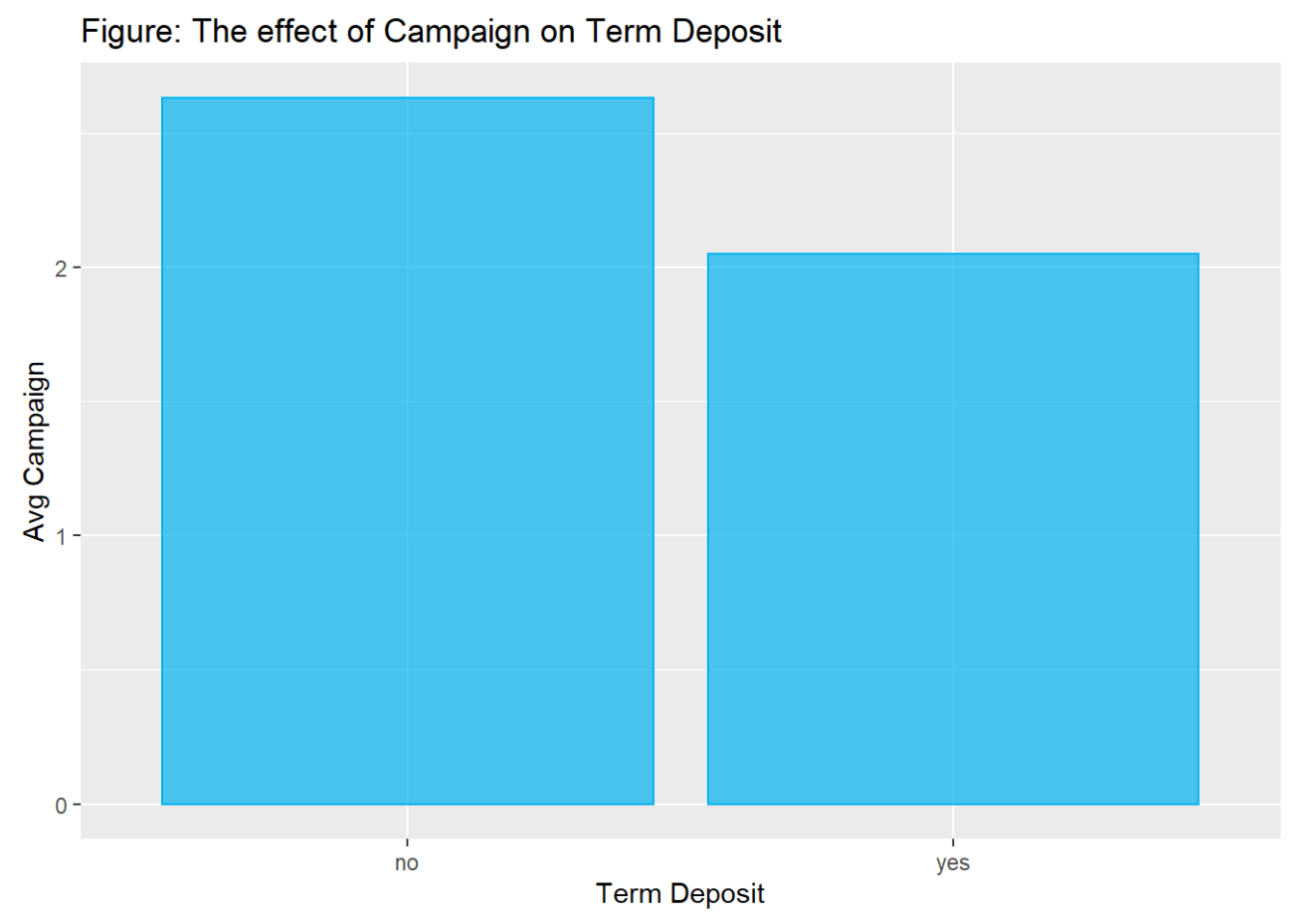
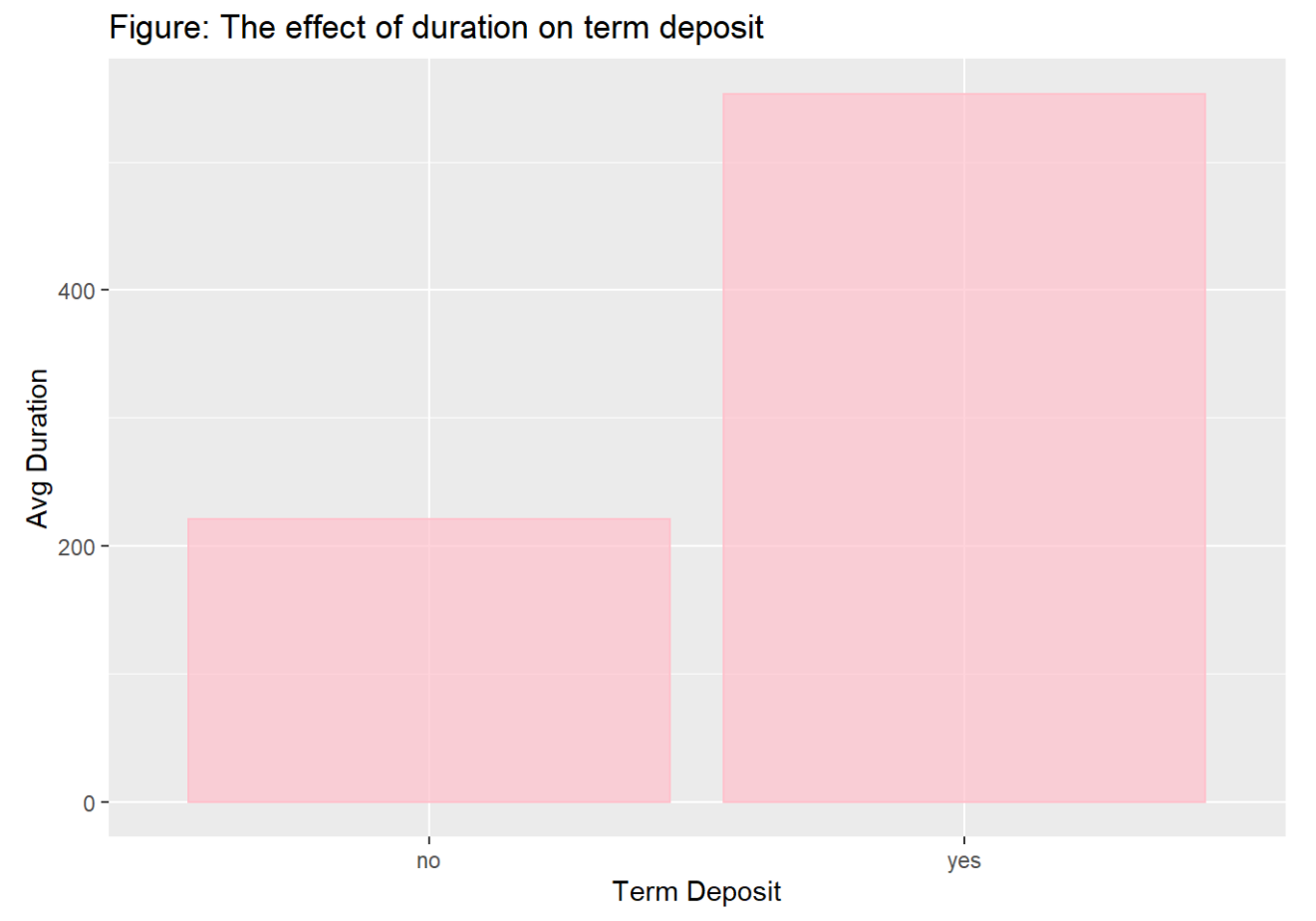
From the bar charts below, it shows that there are nearly 5,000 customers subscribed for term deposit compare to more than 35,000 customers did not subscribed. Besides, most of our customer that company contacted are people who we did not apply from previous campaign.



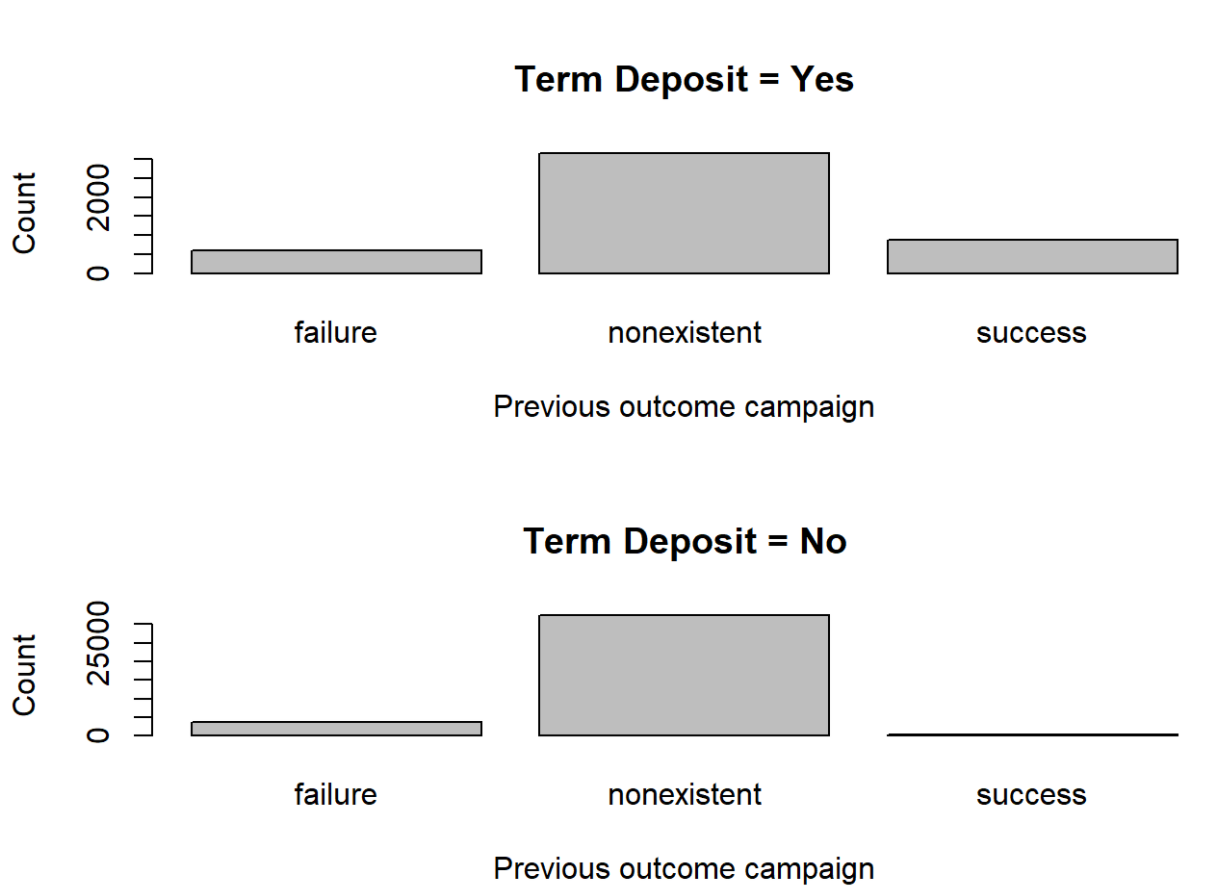
The histogram shows the distribution of each numerical variable, it appears that there are outliers in the campaign, duration.



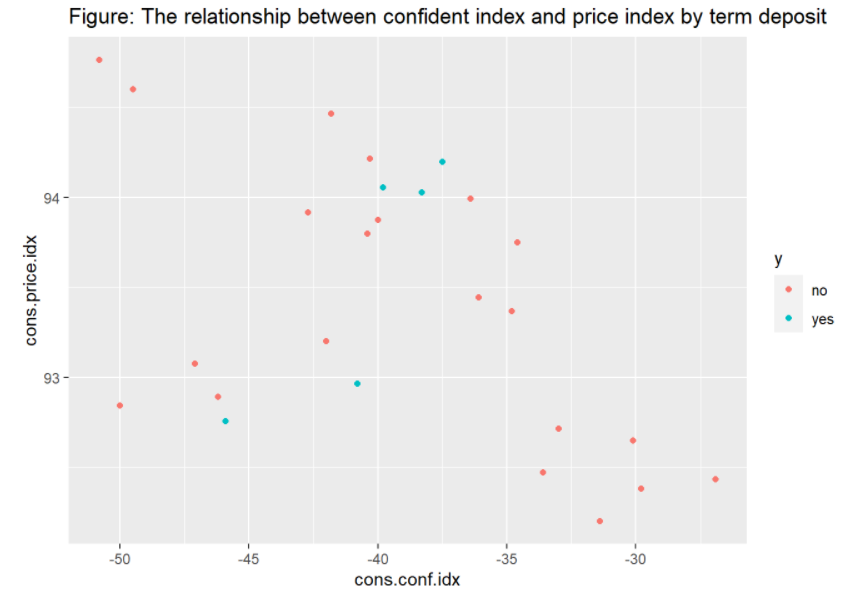
Appendix 1.7. Multidimensional visualization: The bar chart of effects of duration and campaign on term deposit.

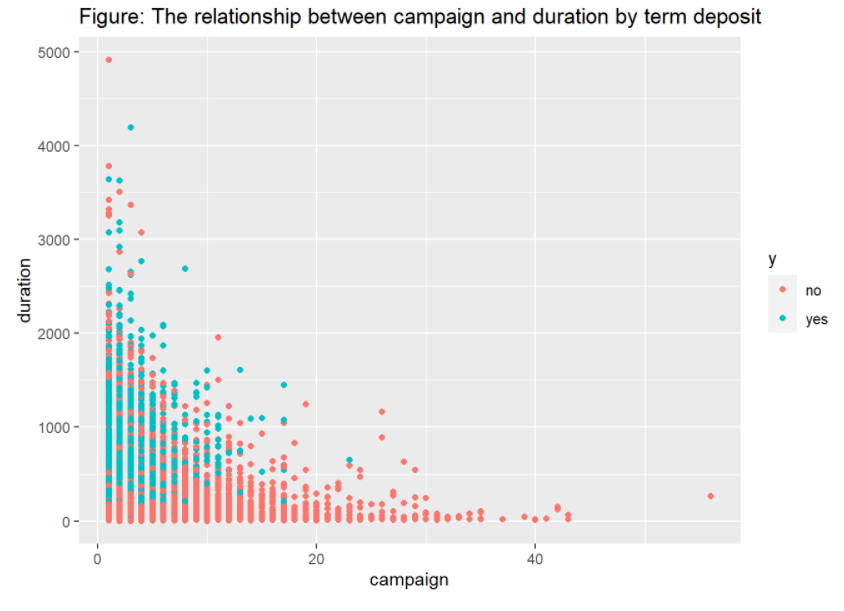




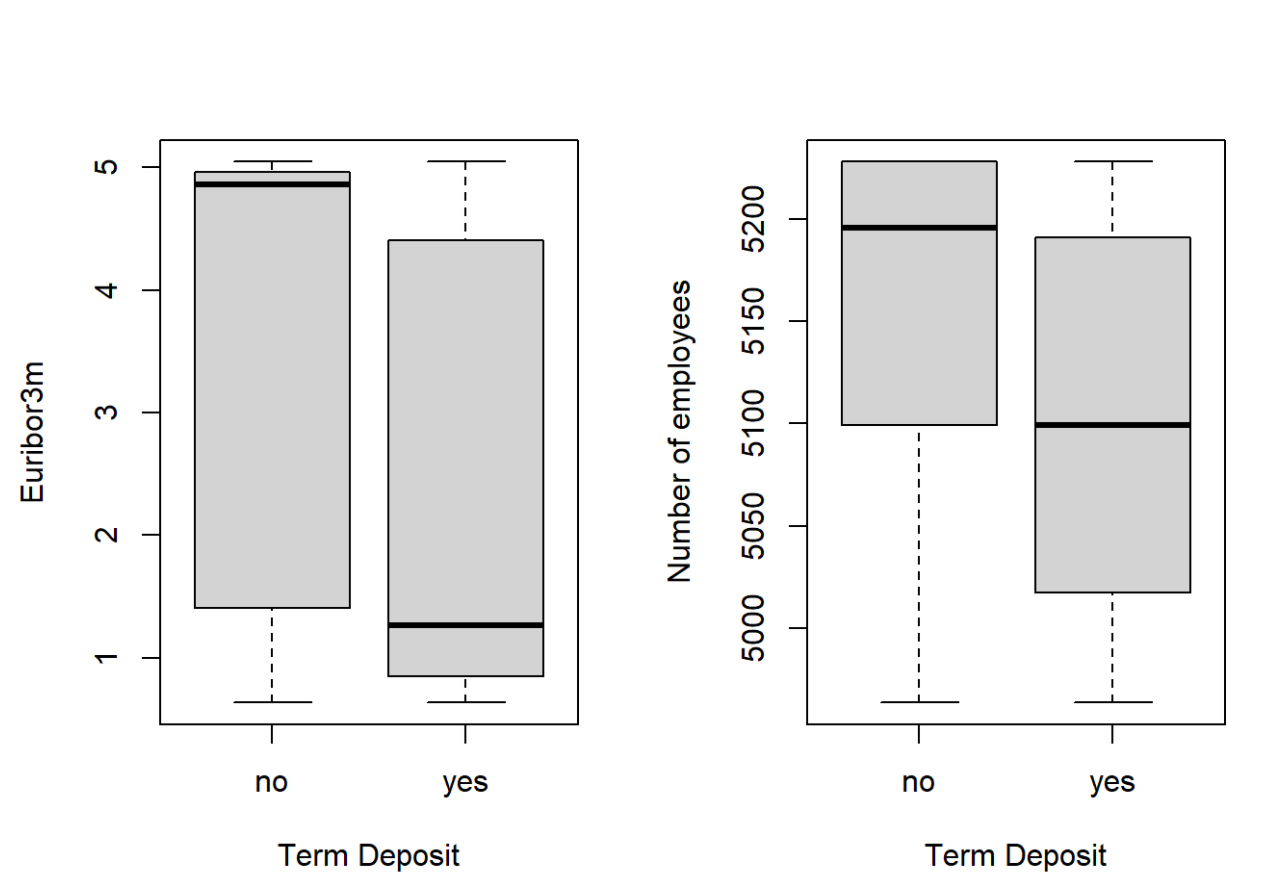


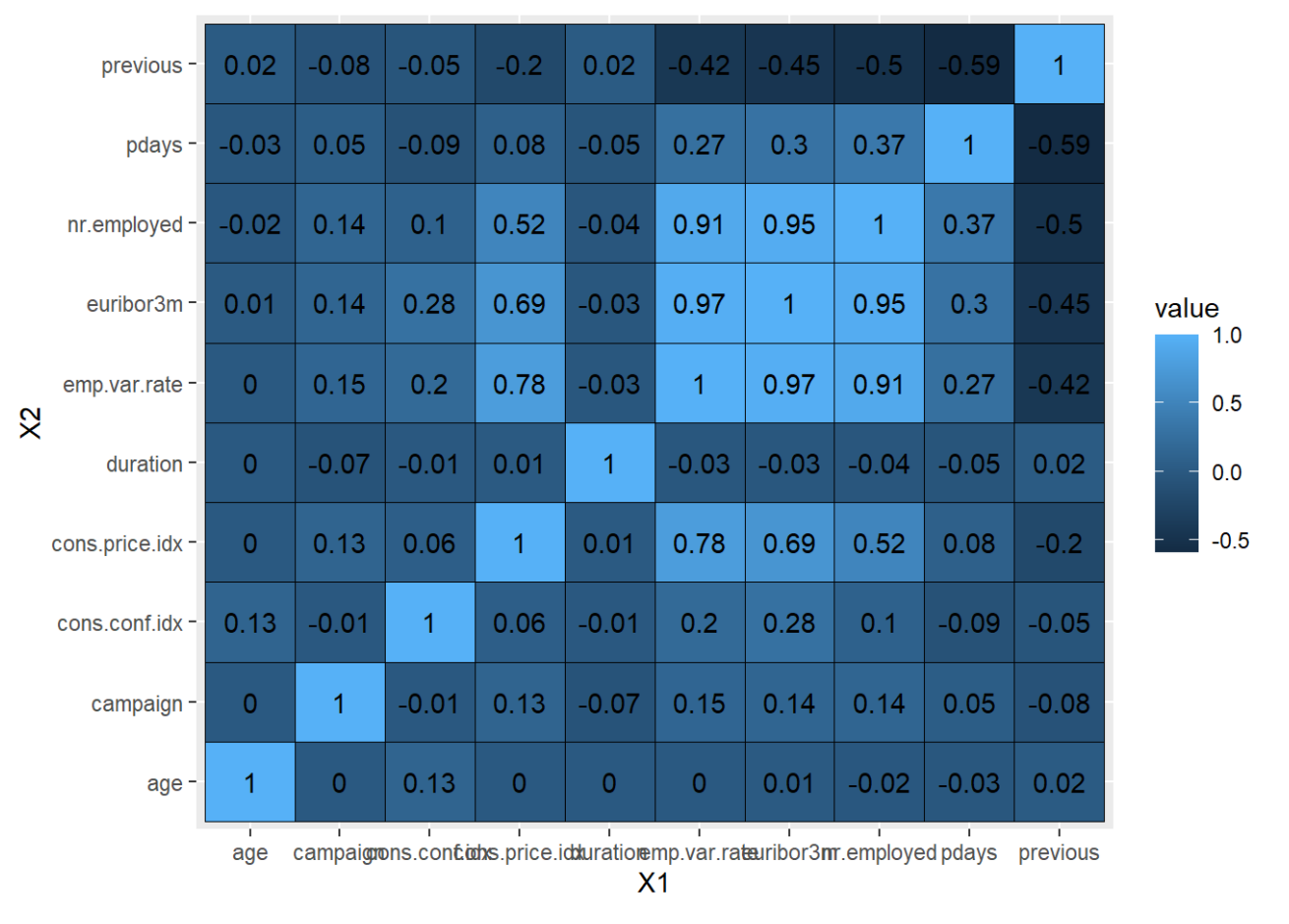
Appendix 1.8:





Appendix 1.9: The boxplot and heatmap

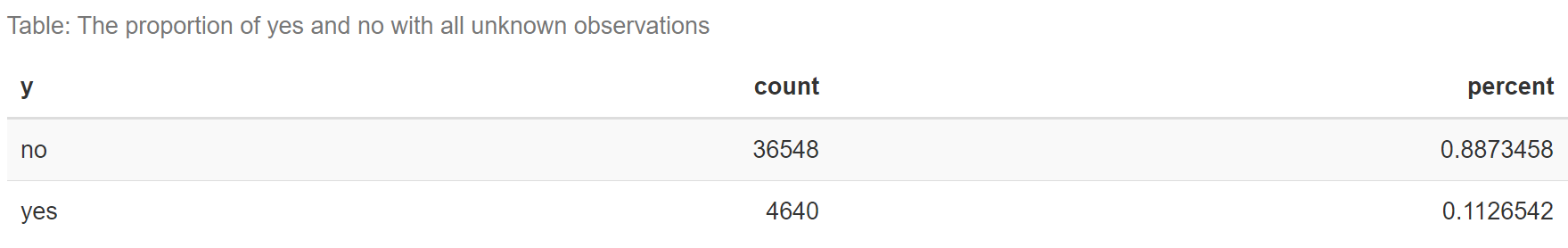


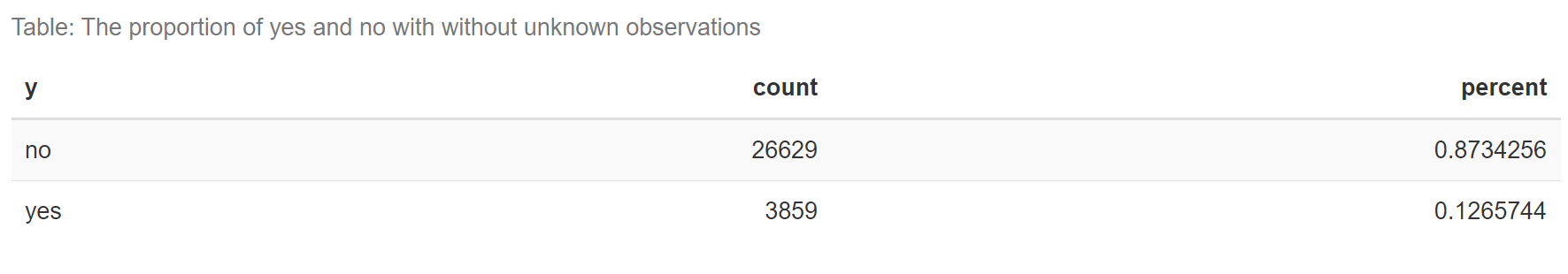


Appendix 1.10. Correlation matrix



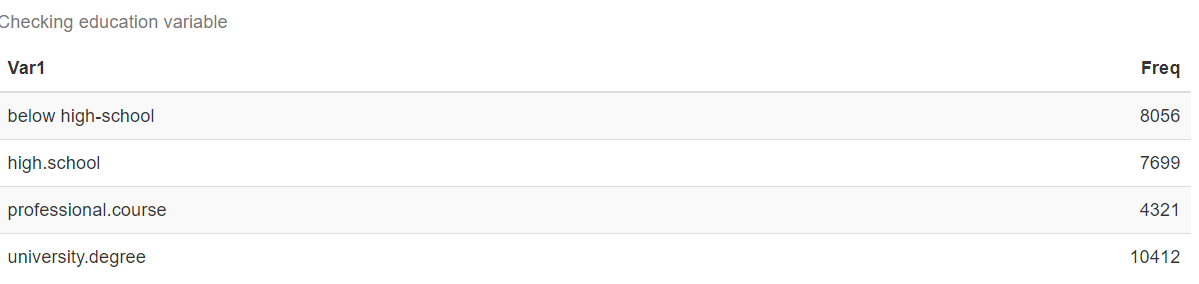
The proportion of the outcome after removing unknown variables



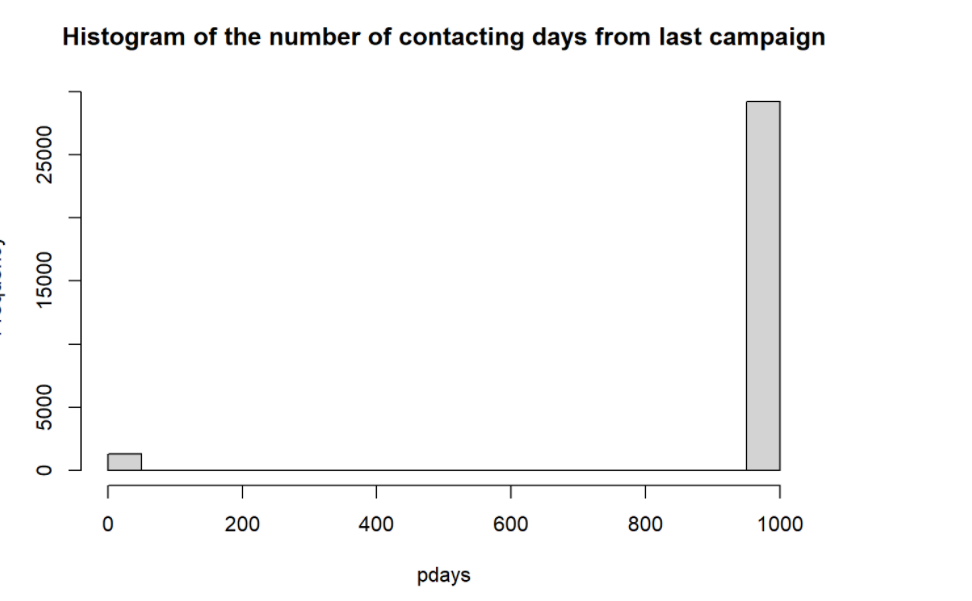
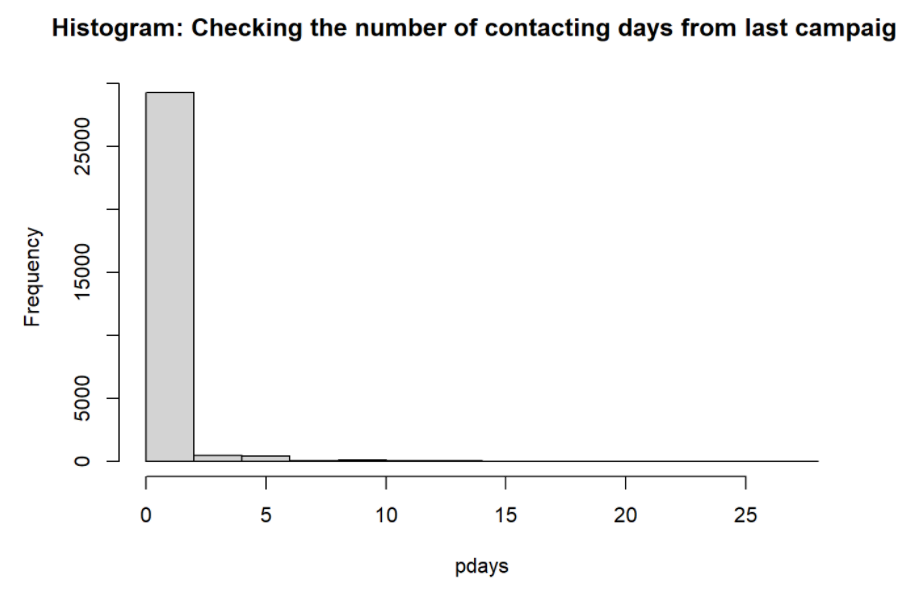


Checking the proportion after group variables



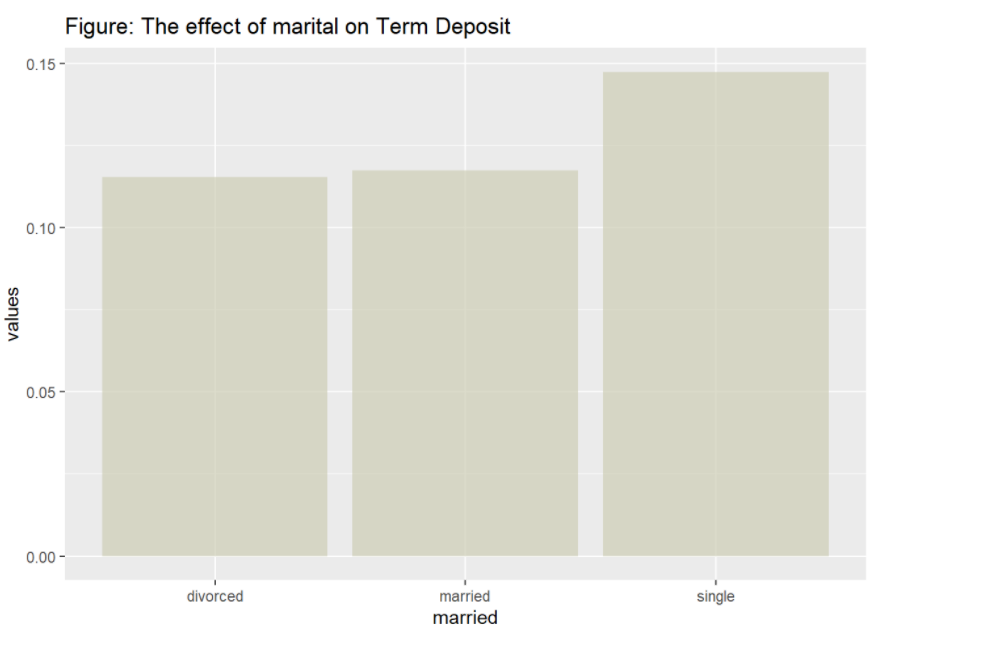
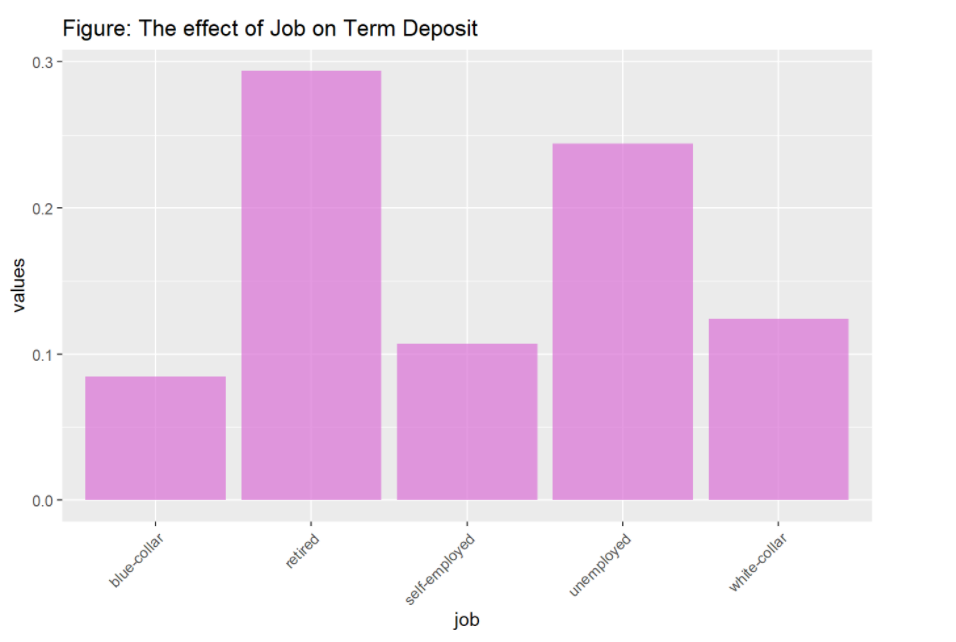


Appendix 1.11.

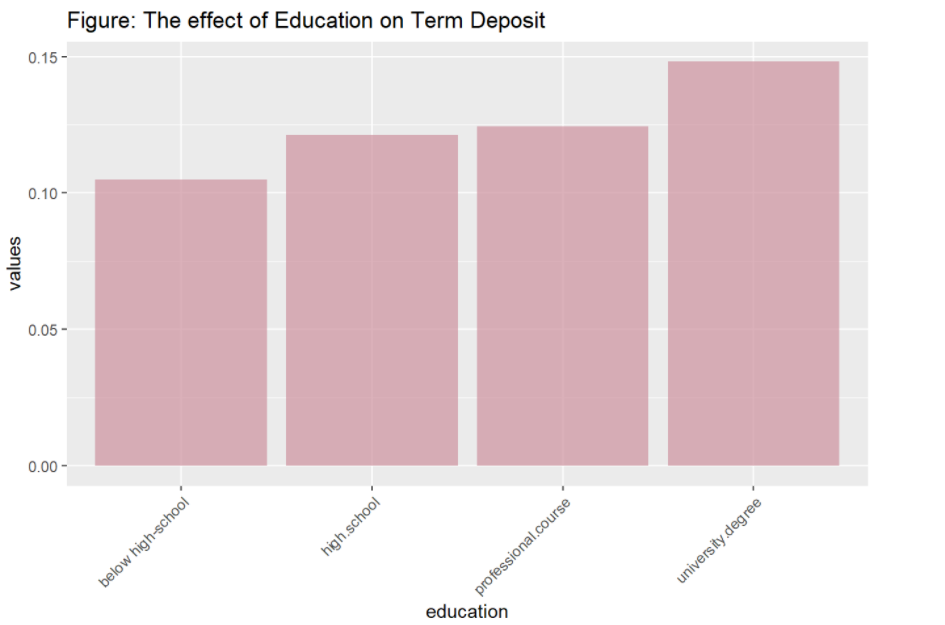
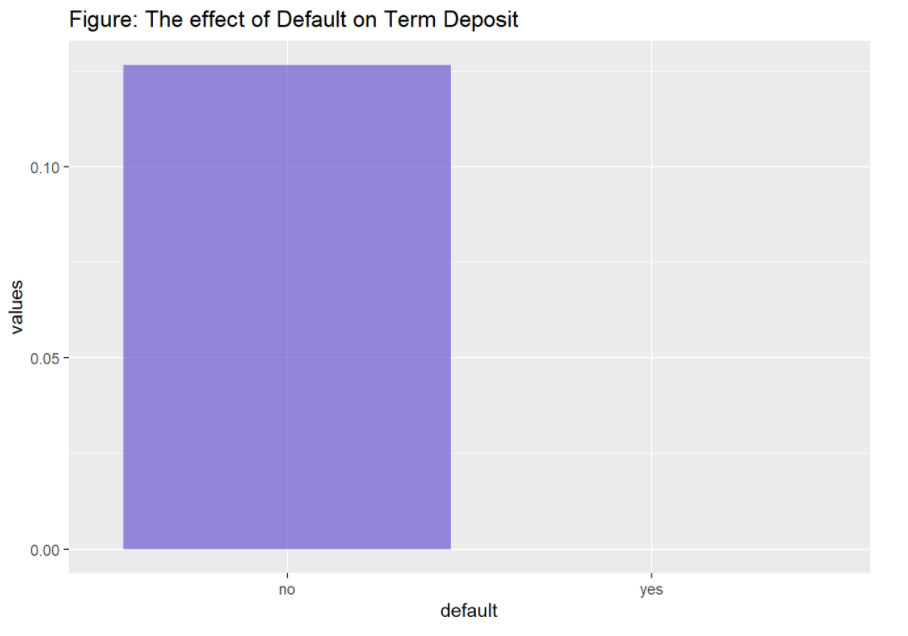
**Reducing data dimension**

Appendix 2.1. The figure of the effect of job/ marital status on term deposit

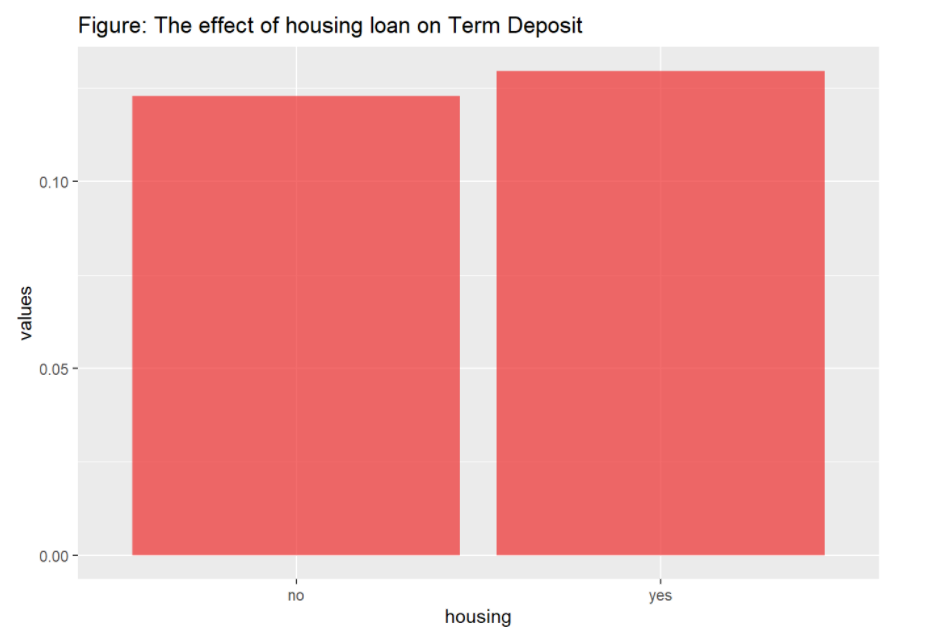
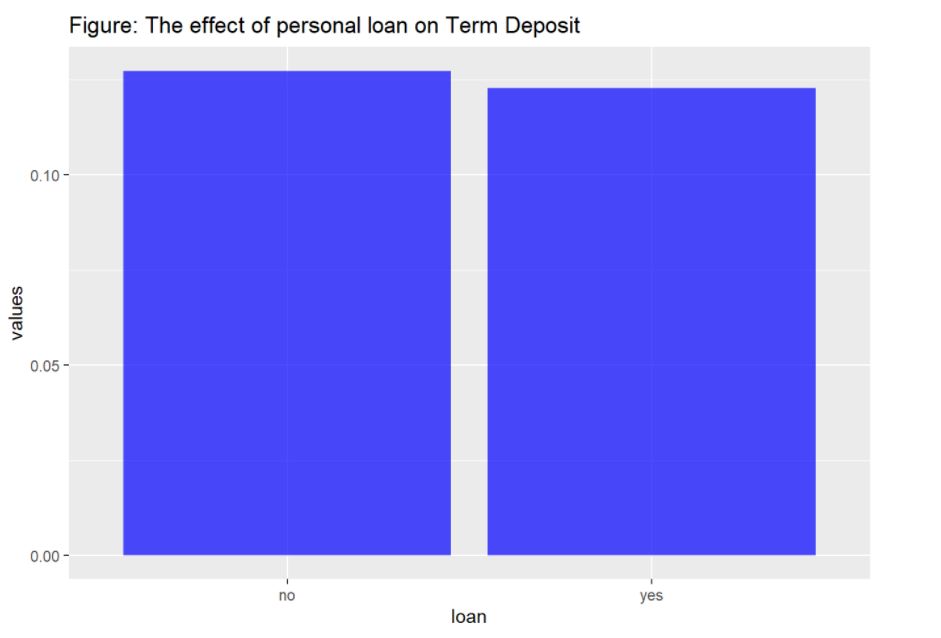


Appendix 2.2 : The effects of independent variables as education, credit default, housing loan and personal loan on term deposits

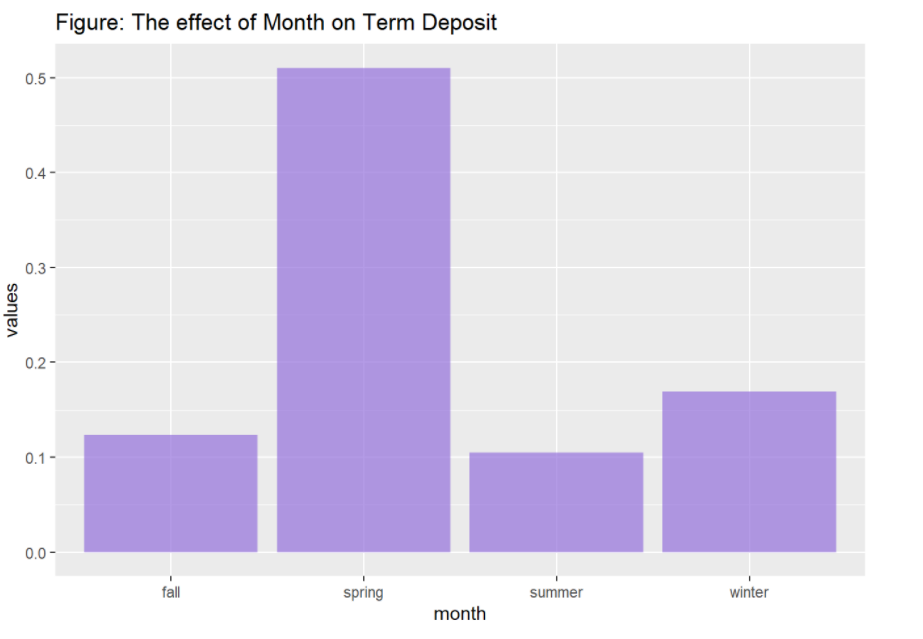
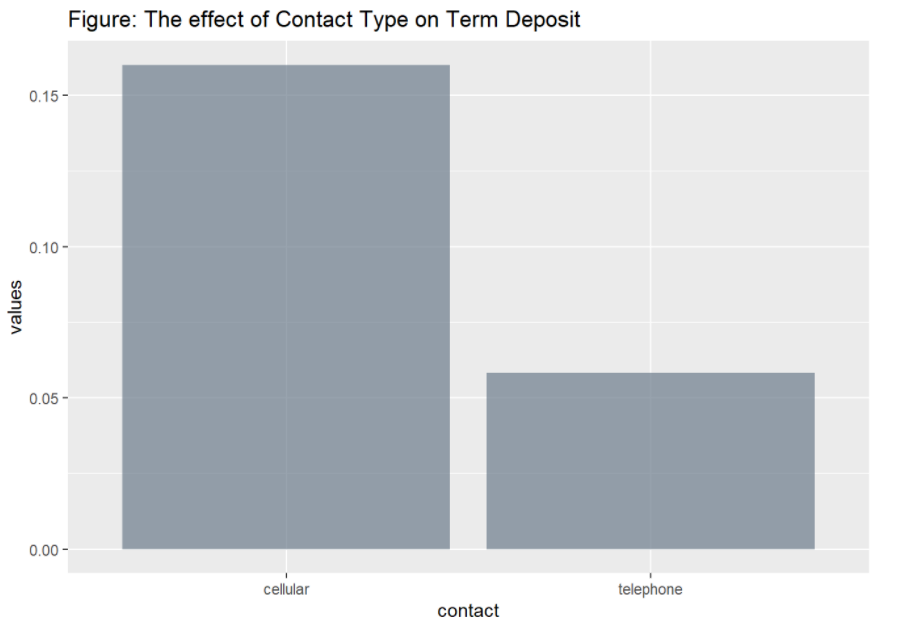
The higher education level, the higher investment on term deposit. The possible reason is that usually people with higher degree tend to have more income thus they have more options on investment. We will include education in the predictive model. As we can see from the bar chart below, it shows that customers have no default will invest money on term deposit which is true in real world because only people who are not in debt can invest their money for the bank. However, as we can see from the visualization above, the proportion of term deposit, the portfolio of customers in the dataset who are default is very small (just 3 customers) so we can assume that all of our customers don’t have default. That means we should not include this variable in the predictive model.

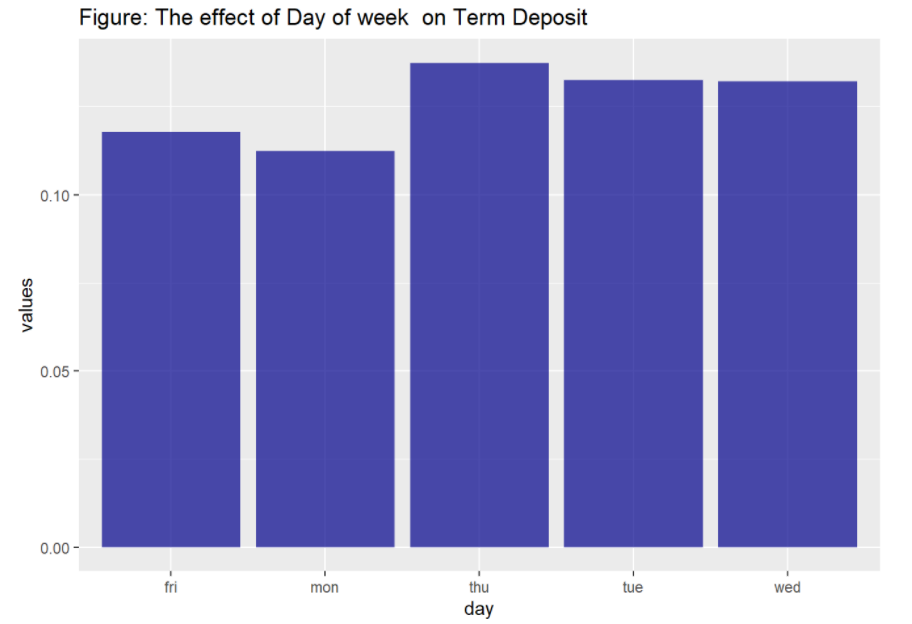
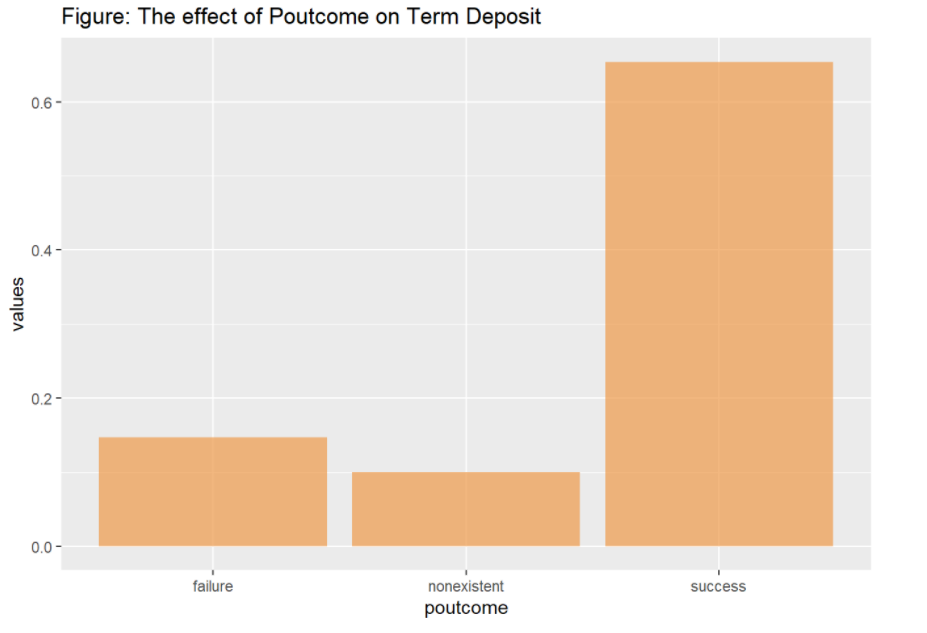
The bar chart appears that housing loan does not have impact on term deposit decision. Even customer who have loan housing or not, they have the same interest on term deposit. Therefore, we exclude housing loan variable in our model. Likewise, the bar chart of personal loan on term deposit also shows that customers who have the personal loan or not, they have the same interest on term deposit. Therefore, we do not include this variable as predictor in our model.

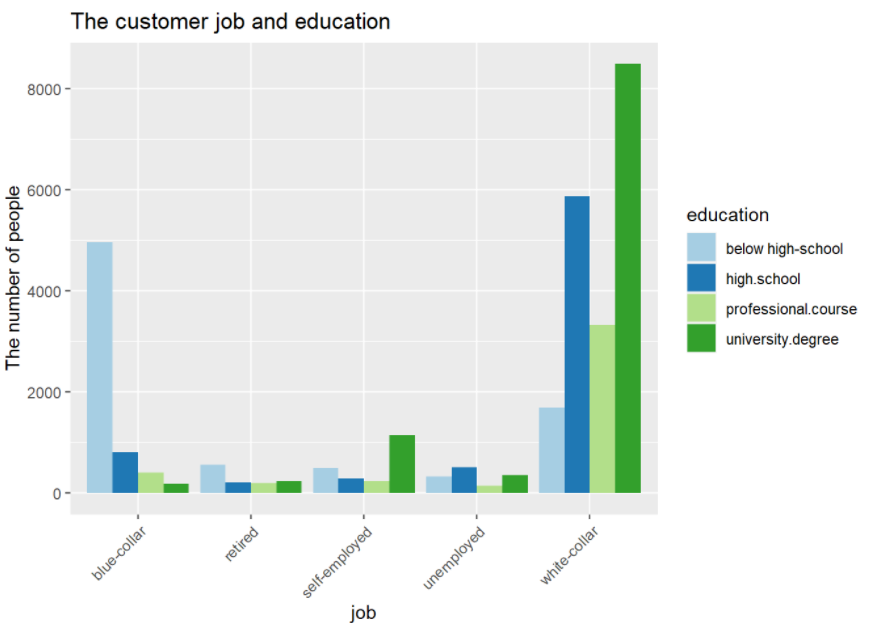
The bar chart shows that the company can increase significantly the number of customers who invest term depos by calling them by cellular.Therefore, we include the contact type on the predictive model. In the spring time people tend to invest more on term deposit compare to other periods in the year. It can be explained that after expenditure in holiday season in the winter, people tend to use the money that invest and term deposit is one of the investment.



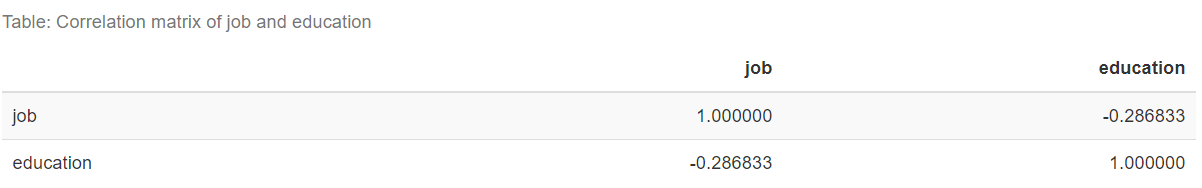
There are not much differences in the term deposit decision when company contacting customers from monday to friday. Therefore, we remove the day variable in our model. It appears that customers who the company contacted in previous campaign and succeed in getting customer have more interest in investing term deposit with this campaign even though the number of this customers are small. Therefore, we will include this variable in the model.

Appendix 2.3. The chart of customer job and education

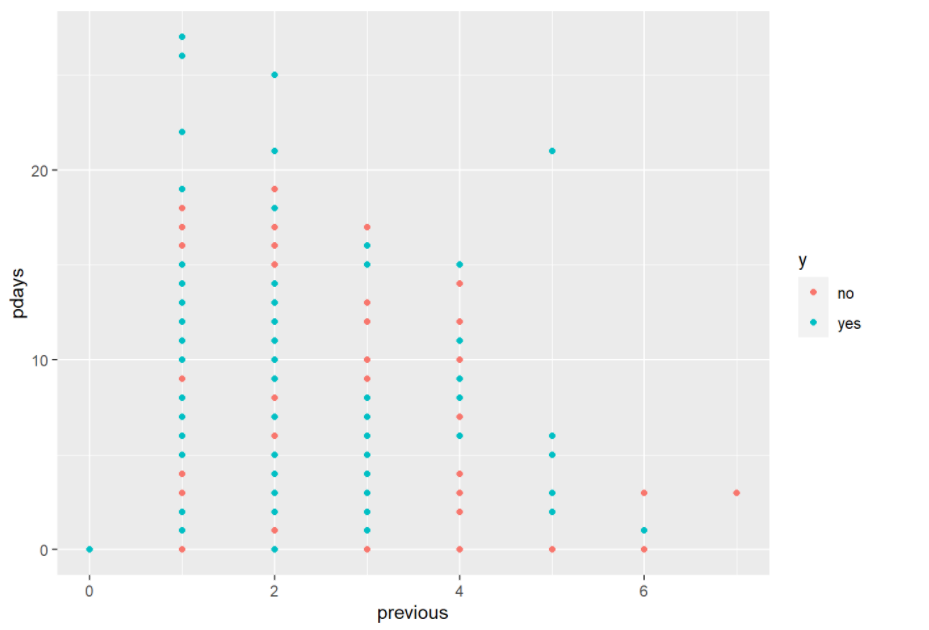


Appendix 2.4. Correlation matrix of job and education



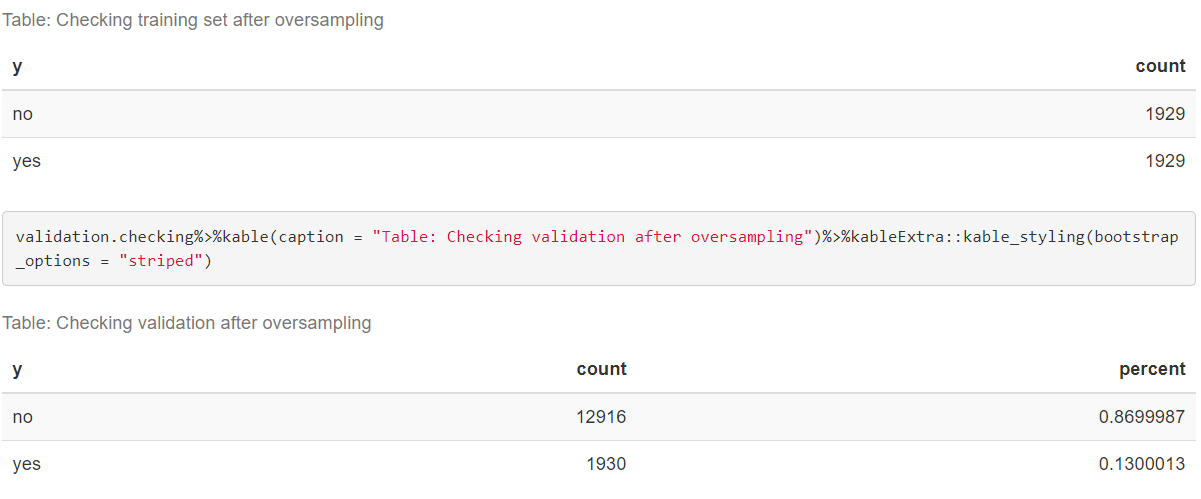
Appendix 2.5.

For “cons.price.idx” and “cons.conf.idx”, the mentioned scatterplot between cons.cof.idx and cons.price.idx shows that there is no differences in term deposit decision therefore, we will not include the model. For “pdays” variable, the scatterplot shows that the pdays and previous does not affect on term deposit decision so we will not include these variables in our model.



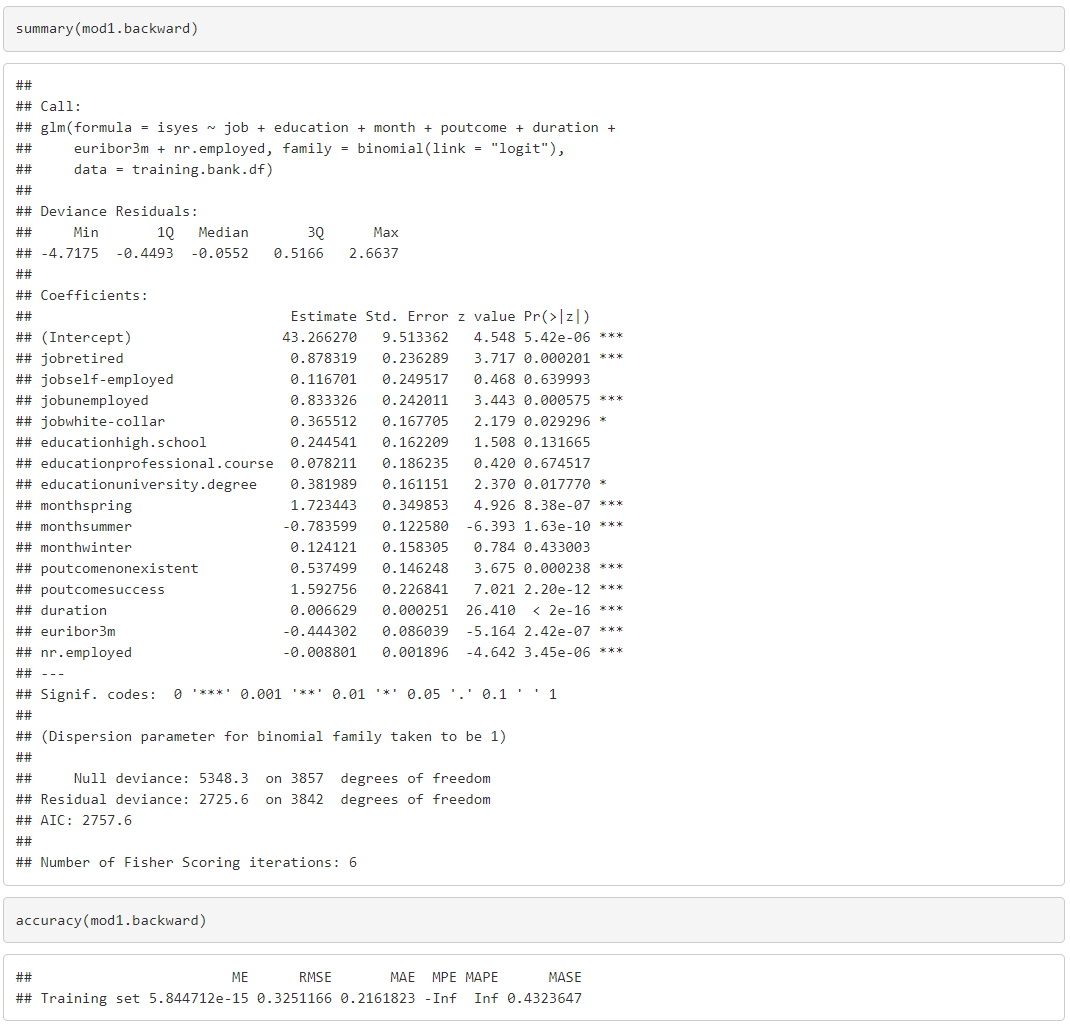
**Data Mining Techniques**

Checking the proportion after oversampling



Appendix 3.1. The result of logistic regression model 1 and model 2

Model 1: Backward elimination



**Model 2: Backward elimination**

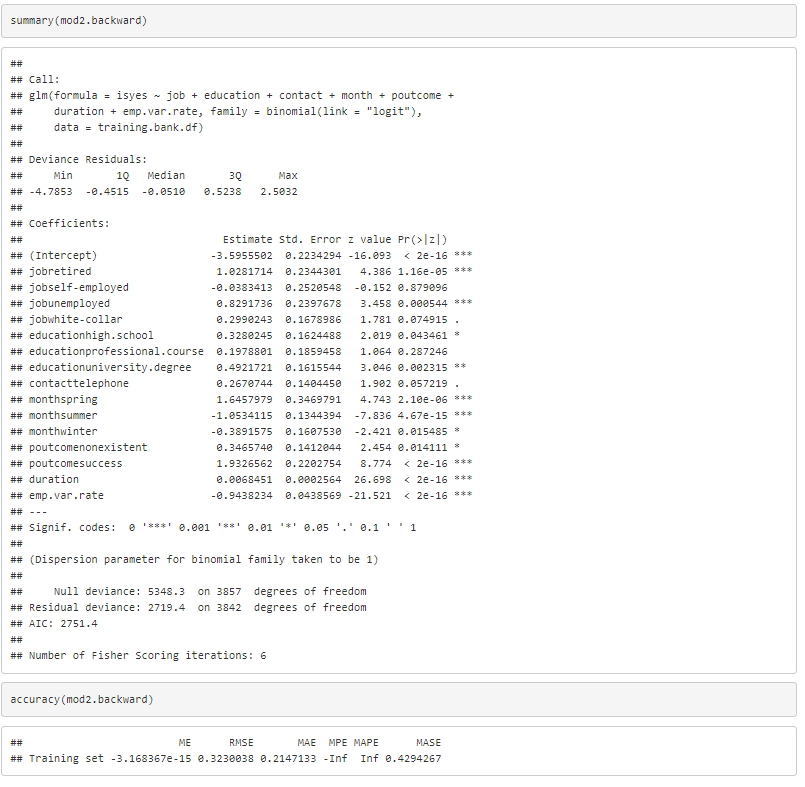
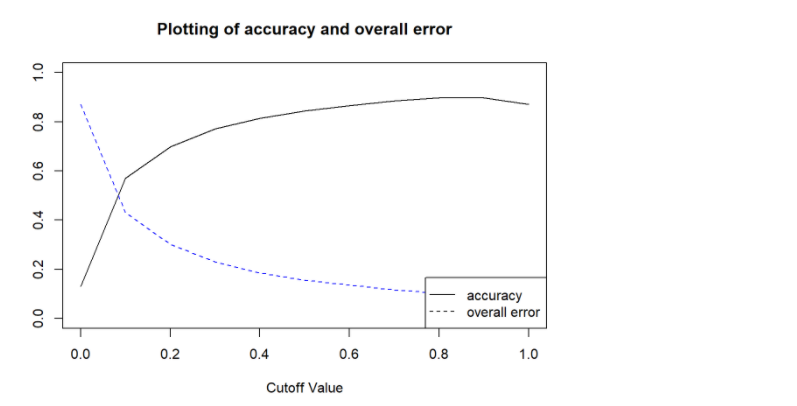


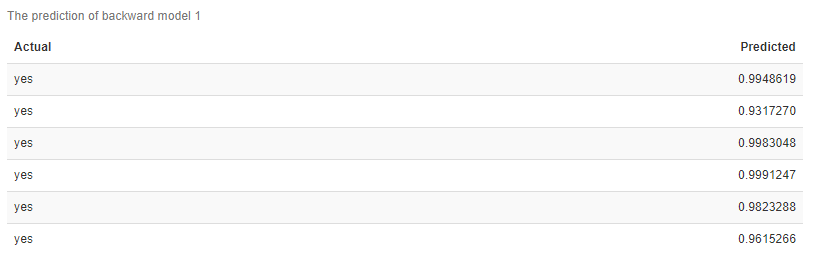
Table of customer propensity on term deposit



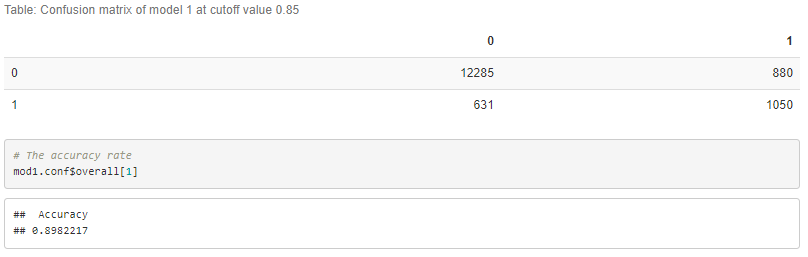
Appendix 3.2. Plotting of accuracy and overall error model 1

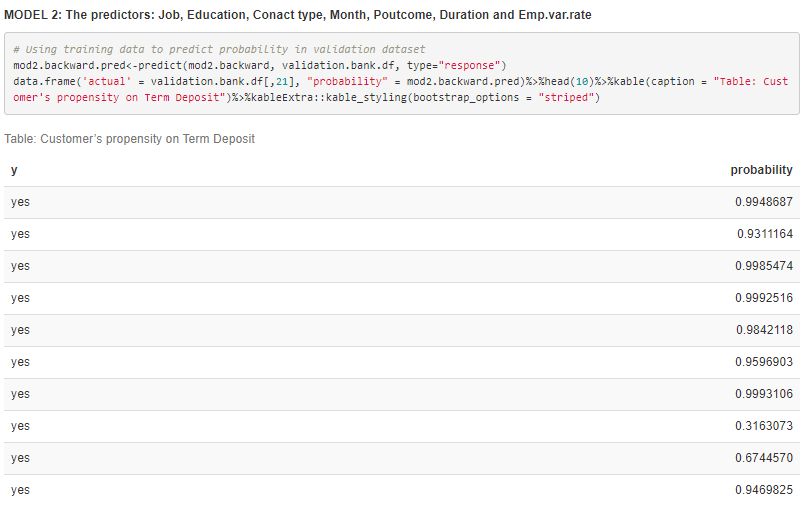


Appendix 3.3. The six samples of clients from prediction model 1

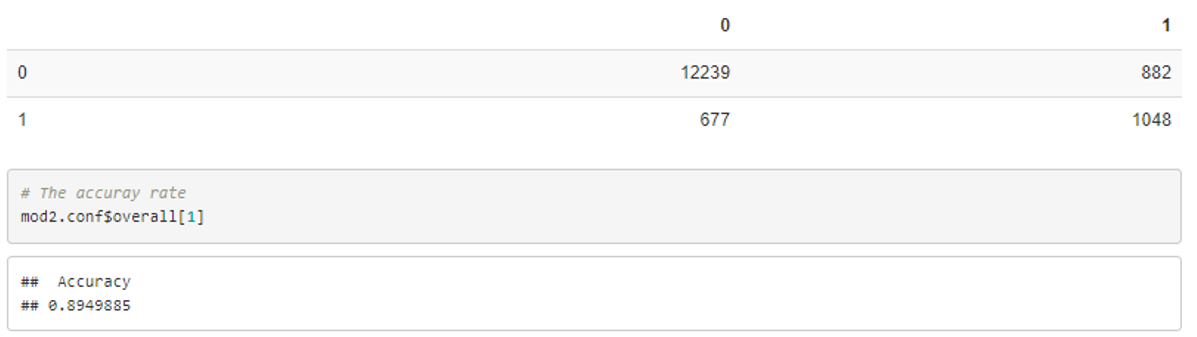


Appendix 3.4. Accuracy rate of model 1





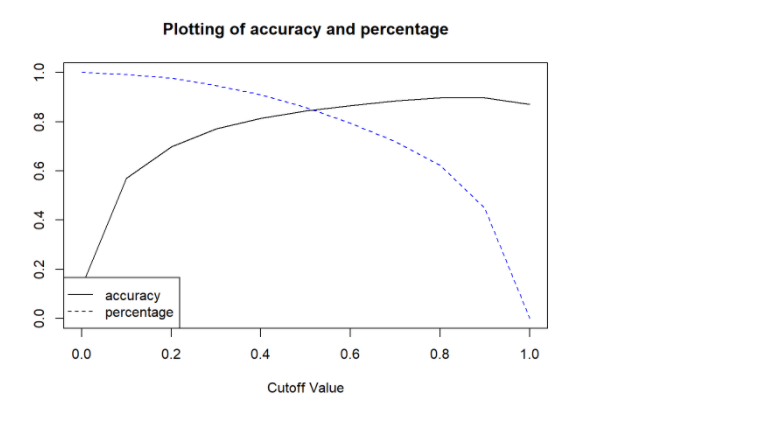
Appendix 3.6. Accuracy of model 2



The six samples of clients from prediction model 2

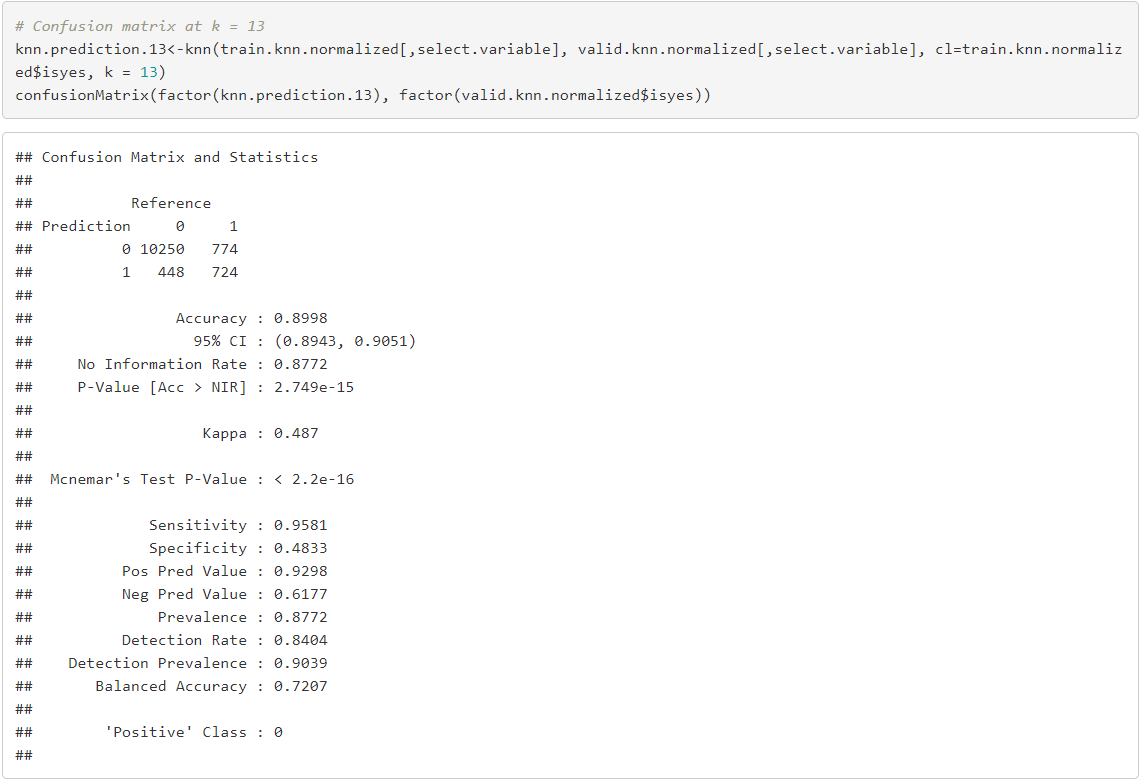


Appendix 3.7. Plotting of accuarcy and percentage of customers



Appendix 3.8. Accuracy of K-NN

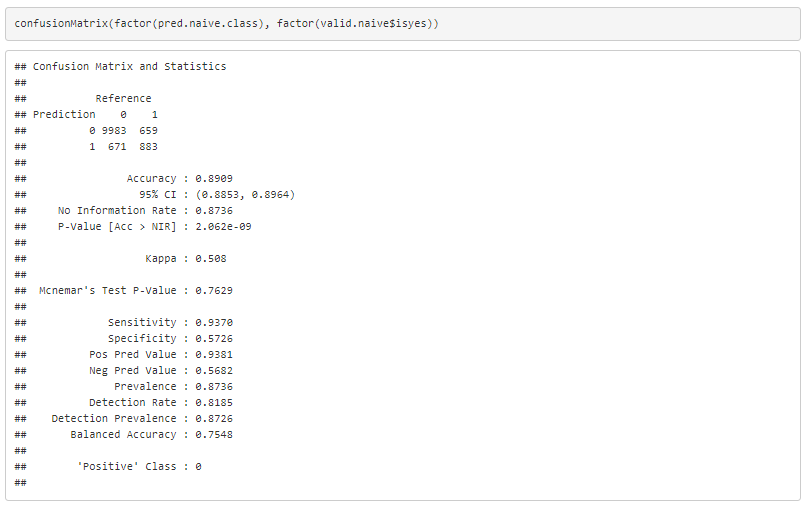




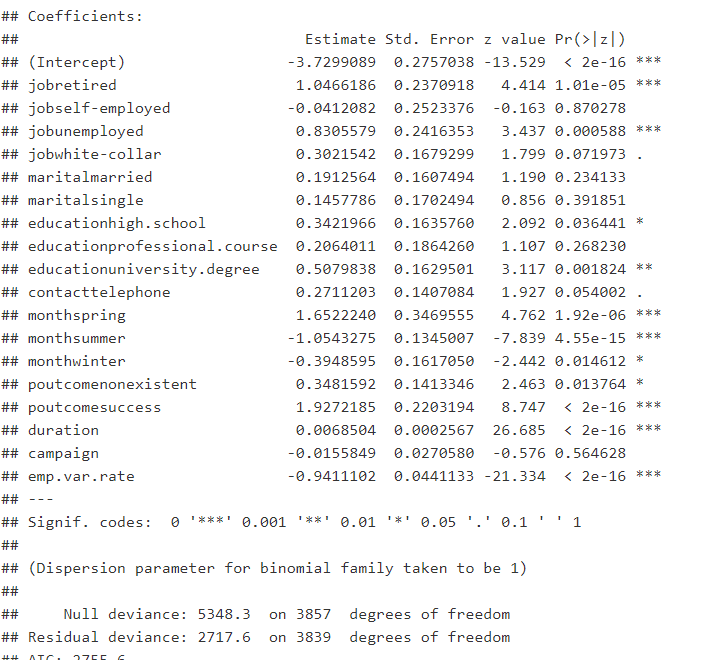
Appendix 3.9. The result of Naive Bayes Classifier for categorical predictors



Appendix 3.10. Confusion matrix of Naïve Bayes classifier



Appendix 3.11. Profiling goal



R codes that have been used in this project

1. Looking at the data

kbl(dim(bank.df))

| x |
| --- |
| 41188 |
| 21 |

head(bank.df, 10)%>%kable(caption = "Table: Ten fisrt rows of bank data")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

kbl(str(bank.df))%>%

kable\_styling()

|  |
| --- |
|  |

2.Summary Statistics of numerical variables and missing values

numerical.variables<-c(1,11,12,13,14,16,17,18,19,20)

table.summary<-data.frame(mean=sapply(bank.df[,numerical.variables], mean),sd=sapply(bank.df[,numerical.variables], sd), min=sapply(bank.df[,numerical.variables],min), max=sapply(bank.df[,numerical.variables],max), median=sapply(bank.df[,numerical.variables],median))

table.missing<-data.frame(miss.val=sapply(bank.df, function(x)

sum(length(which(is.na(x))))))

table.summary%>%kable(caption = "Table: Summary Statistic of all numerical variables")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

table.missing%>%kable(caption = "Table: Number of missing values")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

3. Understanding data by visualization

*Categorized variables*

par(mfrow=c(1,3))

ggplot(bank.df, aes(x=bank.df$job))+

geom\_histogram(fill = "cadetblue", stat = "count")+

ggtitle("Figure: The Job of Customer")+

labs(x = "Job")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(bank.df, aes(x=bank.df$marital))+

geom\_histogram(fill = "lightblue", stat = "count")+

ggtitle("Figure: The Marital Status of Customer")+

labs(x = "Marital Status")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(bank.df, aes(x=bank.df$education))+

geom\_histogram(fill = "cadetblue", stat = "count")+

ggtitle("Figure: The Education of Customer")+

labs(x = "Education Status")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

par(mfrow=c(1,2))

ggplot(bank.df, aes(x=bank.df$default))+

geom\_histogram(fill = "lightblue", stat = "count")+

ggtitle("Figure: The Default of customer")+

labs(x = "Default")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(bank.df, aes(x=bank.df$housing))+

geom\_histogram(fill = "lightblue", stat = "count")+

ggtitle("Figure: The Housing Loan of customer")+

labs(x = "Housing Loan")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

par(mfrow=c(1,3))

ggplot(bank.df, aes(x=bank.df$loan))+

geom\_histogram(fill = "cadetblue", stat = "count")+

ggtitle("Figure: The Loan of Banking File")+

labs(x = "Loan")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(bank.df, aes(x=bank.df$contact))+

geom\_histogram(fill = "lightblue", stat = "count")+

ggtitle("Figure: The Contact type")+

labs(x = "Contact type")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(bank.df, aes(x=bank.df$month))+

geom\_histogram(fill = "cadetblue", stat = "count")+

ggtitle("Figure: The last contact month of the year")+

labs(x = "Months")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

*Numeric Variables*

ggplot(bank.df, aes(y, age))+

geom\_boxplot() +

ggtitle("Figure: Term Deposit Selection based on the age")+

labs(x = "Term Deposit Selection", y="Age")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

The relation of each predictor on the other

ggplot(data=bank.df)+geom\_bar(aes(x=job, fill= as.factor(marital)))+

theme\_minimal()+

ggtitle(label = "Figure: The marital status and job")+

scale\_fill\_brewer(palette="Paired")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(data=bank.df)+geom\_bar(aes(x=education, fill= as.factor(loan)))+

theme\_minimal()+

ggtitle(label = "Figure: Customer education and current personal loan product")+

scale\_fill\_brewer(palette="GnBu")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(data=bank.df)+geom\_bar(aes(x=job, fill= as.factor(loan)))+

theme\_minimal()+

ggtitle(label = "Figure: Customer job and personal loan product")+

scale\_fill\_brewer(palette="GnBu")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

ggplot(data=bank.df)+geom\_bar(aes(x=education, fill= as.factor(housing)))+

ggtitle(label = "Figure: Customer education and housing loan product")+

theme\_minimal()+

scale\_fill\_brewer(palette="Paired")+

theme(axis.text.x=element\_text(angle=45, hjust =1))

Bar charts and distribution plots to summarize data

*# Compute number of customer decision in the data*

count.decision<-aggregate(bank.df$y, by=list(bank.df$y), FUN=count)

names(count.decision)<-c("Term Deposit", "Count")

ggplot(data=count.decision, mapping = aes(x=`Term Deposit`, y=`Count`))+geom\_col(color="plum2", fill="plum2") + ggtitle("Figure: Frequency of customer decision")

*# Number of previous outcome*

ggplot(data=bank.df, mapping=aes(x=poutcome))+geom\_bar(colour="firebrick2", alpha=0.7, fill="firebrick2") + ggtitle("Figure: Frequency of previous outcome campaign")

bank.df[,-c(1,13)]%>%keep(is.numeric)%>%gather()%>%ggplot(aes(value))+facet\_wrap(~key, scales = "free")+geom\_histogram(bins=30)

Understand data by multidimensional visualization

*# Compute mean of Duration per Y = Yes, No*

data.duration<-aggregate(bank.df$duration, by=list(bank.df$y), FUN=mean)

names(data.duration)<-c("Term Deposit","Avg Duration")

ggplot(data=data.duration, mapping=aes(x=`Term Deposit`, y=`Avg Duration`))+geom\_col(color="pink", fill="pink", alpha=0.7) + ggtitle("Figure: The effect of duration on term deposit")

*# Compute mean of campaign per Y =Yes, No*

data.campaign<-aggregate(bank.df$campaign, by=list(bank.df$y), FUN=mean)

names(data.campaign)<-c("Term Deposit", "Avg Campaign")

ggplot(data=data.campaign, mapping=aes(x=`Term Deposit`, y=`Avg Campaign`)) + geom\_col(color="deepskyblue2", fill="deepskyblue2", alpha=0.7) + ggtitle("Figure: The effect of Campaign on Term Deposit")

*# The effect of number of contacts performed before this campaign and this client by panel plot*

data.previous<-aggregate(bank.df$previous, by=list(bank.df$y, bank.df$previous), FUN=count)

names(data.previous)<-c("Term Deposit", "Previous", "Count")

par(mfcol=c(2,1))

barplot(height = data.previous$Count[data.previous$`Term Deposit`=="yes"], names.arg=data.previous$Previous[data.previous$`Term Deposit`=="yes"], xlab="Previous", ylab="Count", main = "Term Deposit = Yes")

barplot(height = data.previous$Count[data.previous$`Term Deposit`=="no"], names.arg=data.previous$Previous[data.previous$`Term Deposit`=="no"], xlab="Previous", ylab="Count", main = "Term Deposit = No")

*# The effect of previous outcome marketing campaign on term deposit*

data.poutcome<-aggregate(bank.df$poutcome, by=list(bank.df$y, bank.df$poutcome), FUN = count)

names(data.poutcome)<-c("Term Deposit", "Previous Marketing Campaign" ,"Count")

par(mfrow=c(2,1))

barplot(height = data.poutcome$Count[data.poutcome$`Term Deposit`=="yes"], names.arg = data.poutcome$`Previous Marketing Campaign`[data.poutcome$`Term Deposit`=="yes"], xlab="Previous outcome campaign", ylab="Count", main="Term Deposit = Yes")

barplot(height = data.poutcome$Count[data.poutcome$`Term Deposit`=="no"], names.arg = data.poutcome$`Previous Marketing Campaign`[data.poutcome$`Term Deposit`=="no"], xlab="Previous outcome campaign", ylab="Count", main="Term Deposit = No")

*# Scatter plot of confident indext and price index of each customer*

ggplot(data=bank.df, aes(y=cons.price.idx,x=cons.conf.idx, color=y)) + geom\_point() +ggtitle("Figure: The relationship between confident index and price index by term deposit")

*# Scatter plot of duration and campaign of each customer*

ggplot(data=bank.df, aes(x=campaign, y=duration, color=y)) + geom\_point() + ggtitle ("Figure: The relationship between campaign and duration by term deposit")

Sid-by-Side Boxplot of euribor3m and nr.employed on term deposit

par(mfcol=c(1,2))

boxplot(bank.df$euribor3m~bank.df$y, xlab="Term Deposit", ylab="Euribor3m")

boxplot(bank.df$nr.employed~bank.df$y, xlab = "Term Deposit", ylab="Number of employees")

Heatmap and correlation matrix for correlations

cor.mat<-round(cor(bank.df[,numerical.variables]),2)

cor.mat%>%kable(caption = "Table: Correlation matrix of numerical variables")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

melted.cor.mat<-melt(cor.mat)

names(melted.cor.mat)<-c("X1", "X2", "value")

ggplot(melted.cor.mat, aes(x=X1, y=X2, fill=value)) + geom\_tile(col="black") + geom\_text(aes(x=X1, y=X2, label=value))

4. Cleaning data

Unknow variable

*# The proportion of Term Deposit with all unknown records*

unknown.term<-group\_by(bank.df, y)

unknown.term<-summarise(unknown.term,count=n())

unknown.term$count<-as.numeric(unknown.term$count)

unknown.term<-mutate(unknown.term, percent=count/sum(count))

unknown.term%>%kable(caption="Table: The proportion of yes and no with all unknown observations")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# The proportion of Term Deposit without unknown observations*

bank.df2<-filter(bank.df, job!="unknown", marital!="unknown",education!="unknown", default!="unknown", housing!="unknown")

nonunknown.term<-group\_by(bank.df2,y)

nonunknown.term<-summarise(nonunknown.term,count=n())

nonunknown.term$count<-as.numeric(nonunknown.term$count)

nonunknown.term<-mutate(nonunknown.term, percent=count/sum(count))

nonunknown.term%>%kable(caption="Table: The proportion of yes and no with without unknown observations")%>%kableExtra::kable\_styling(bootstrap\_options = "striped", full\_width = FALSE)

dim(bank.df2)

## [1] 30488 21

head(bank.df2, 10)%>%kable(caption = "Table: Bank dataset without unknown observations")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

Job variable

bank.df2$job[bank.df2$job=="entrepreneur"] <- "self-employed"

bank.df2$job[bank.df2$job=="housemaid"] <-"blue-collar"

bank.df2$job[bank.df2$job=="admin."] <-"white-collar"

bank.df2$job[bank.df2$job=="management"] <-"white-collar"

bank.df2$job[bank.df2$job=="services"] <-"white-collar"

bank.df2$job[bank.df2$job=="technician"] <-"white-collar"

bank.df2$job[bank.df2$job=="housemaid"] <-"blue-collar"

bank.df2$job[bank.df2$job=="student"]<- "unemployed"

table(bank.df2$job)%>%kable(caption="Table: Checking job variable")%>%kableExtra::kable\_styling(bootstrap\_options = "striped", full\_width = FALSE)

Education variable

bank.df2$education[bank.df2$education=="illiterate"]<- "below high-school"

bank.df2$education[bank.df2$education=="basic.4y"]<- "below high-school"

bank.df2$education[bank.df2$education=="basic.6y"]<- "below high-school"

bank.df2$education[bank.df2$education=="basic.9y"]<- "below high-school"

table(bank.df2$education)%>%kable(caption= "Table: Checking education variable")%>%kableExtra::kable\_styling(bootstrap\_options = "striped", full\_width = FALSE)

Month varible

bank.df2$month<-ifelse(bank.df2$month=="mar", "spring", ifelse(bank.df2$month=="apr"|bank.df2$month=="may"|bank.df2$month=="jun", "summer", ifelse(bank.df2$month=="jul"|bank.df2$month=="aug"|bank.df2$month=="sep", "fall", "winter")))

table(bank.df2$month)%>%kable(caption = "Table: Checking month")%>%kableExtra::kable\_styling(bootstrap\_options = "striped", full\_width = FALSE)

pdays variable

hist(bank.df2$pdays,main = "Figure: Histogram of the number of contacting days from last campaign", xlab="pdays")

bank.df2$pdays[bank.df2$pdays==999]<- 0

bank.df2$pdays<-as.numeric(bank.df2$pdays)

hist(bank.df2$pdays,main = "Figure: Checking the number of contacting days from last campaign", xlab="pdays")

5. Data Dimension and varibles selection

Categorical varibles

The effect of job type on Term deposit decision

job.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$job), mean)

names(job.agg)<-c("job", "values")

ggplot(data = job.agg, mapping = aes(x=job, y=values)) + geom\_col(fill="orchid", alpha=0.7) + theme (axis.text.x=element\_text(angle = 45, hjust = 1)) + ggtitle("Figure: The effect of Job on Term Deposit")

The effect of marital on Term deposit decision

married.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$marital), mean)

names(married.agg)<-c("married", "values")

ggplot(data = married.agg, mapping = aes(x=married, y=values)) + geom\_col(fill="lightyellow3", alpha=0.7) + ggtitle("Figure: The effect of marital on Term Deposit")

The effect of education on term deposit

education.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$education), mean)

names(education.agg)<-c("education", "values")

ggplot(data = education.agg, mapping = aes(x=education, y=values)) + geom\_col(fill="pink3", alpha=0.7) + theme (axis.text.x=element\_text(angle = 45, hjust = 1)) + ggtitle("Figure: The effect of Education on Term Deposit")

The effect of default on term deposit

default.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$default), mean)

names(default.agg)<-c("default", "values")

ggplot(data = default.agg, mapping = aes(x=default, y=values)) + geom\_col(fill="slateblue3", alpha=0.7) + ggtitle("Figure: The effect of Default on Term Deposit")

The effect of housing loan on Term deposit decision

housingloan.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$housing), mean)

names(housingloan.agg)<-c("housing", "values")

ggplot(data = housingloan.agg, mapping = aes(x=housing, y=values)) + geom\_col(fill="firebrick2", alpha=0.7) + ggtitle("Figure: The effect of housing loan on Term Deposit")

The effect of personal loan on Term deposit decision

loan.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$loan), mean)

names(loan.agg)<-c("loan", "values")

ggplot(data = loan.agg, mapping = aes(x=loan, y=values)) + geom\_col(fill="blue", alpha=0.7) + ggtitle("Figure: The effect of personal loan on Term Deposit")

The effect of contact type on Term deposit decision

contact.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$contact), mean)

names(contact.agg)<-c("contact", "values")

ggplot(data = contact.agg, mapping = aes(x=contact, y=values)) + geom\_col(fill="slategray4", alpha=0.7) + ggtitle("Figure: The effect of Contact Type on Term Deposit")

The effect of month on Term deposit decision

month.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$month), mean)

names(month.agg)<-c("month", "values")

ggplot(data=month.agg, mapping = aes(x=month, y=values)) + geom\_col(fill="mediumpurple", alpha=0.7) + ggtitle("Figure: The effect of Month on Term Deposit")

The effect of day of week on Term deposit decision

day.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$day\_of\_week), mean)

names(day.agg)<-c("day", "values")

ggplot(data=day.agg, mapping = aes(x=day, y=values)) + geom\_col(fill="blue4", alpha=0.7) + ggtitle("Figure: The effect of Day of week on Term Deposit")

The effect of poutcome on Term deposit decision

poutcome.agg<-aggregate(bank.df2$y=="yes", by=list(bank.df2$poutcome), mean)

names(poutcome.agg)<-c("poutcome", "values")

ggplot(data=poutcome.agg, mapping = aes(x=poutcome, y=values)) + geom\_col(fill="tan2", alpha=0.7) + ggtitle("Figure: The effect of Poutcome on Term Deposit")

Checking correlation of job and education

*By visualization*

job.edu<-group\_by(bank.df2, job, education)

job.edu<-summarise(job.edu, n=n())

ggplot(job.edu, aes(x=job, y=n))+geom\_col(aes(fill=education), position = "dodge")+theme(axis.text.x=element\_text(angle=45, hjust =1))+

scale\_fill\_brewer(palette="Paired")+ylab("The number of people")+ggtitle("Figure: The customer job and education")

*By correlation matrix*

edu.job2<-bank.df2[,c(2,4)]

edu.job2$job<-ifelse(edu.job2$job=="retired",1,ifelse(edu.job2$job=="unemployed",2,ifelse(edu.job2$job=="blue-collar",3,ifelse(edu.job2$job=="self-employed",4,5))))

edu.job2$education<-ifelse(edu.job2$education=="below high school",1,ifelse(edu.job2$education=="high.school",2,ifelse(edu.job2$education=="university.degree",3,4)))

cor(edu.job2)%>%kable(caption = "Table: Correlation matrix of job and education")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

Numerical variables

ggplot(data=bank.df2, aes(y=pdays,x=previous, color=y)) + geom\_point() + ggtitle("Figure: effect of pdays and previous on term deposit")

Export bank.df2 for project update 4 and for cleaned data to submit

write\_excel\_csv(bank.df2, "C:/Users/kieuk/OneDrive/Documents/BA636 Data Mining/PROJECT/Project update/Cleaned banking data/cleaneddata.csv")

6. Oversampling and partitioning data into training and validation dataset

*# Transforming term deposit into 0,1 y creating new variable*

bank.df2<-mutate(bank.df2, isyes=ifelse(bank.df2$y=="yes",1,0))

*#separate data*

term.yes<-filter(bank.df2, y=="yes")

term.no<-filter(bank.df2, y=="no")

*# partitioning data*

set.seed(2)

train.yes.row<-sample(c(1:3859), dim(term.yes)\*0.5)

train.yes<-term.yes[train.yes.row,]

valid.yes<-term.yes[-train.yes.row,]

train.no.row<-sample(c(1:26629), dim(term.yes)\*0.5)

train.no<-term.no[train.no.row,]

training.bank.df<-rbind(train.yes, train.no)

valid.no.row<-sample(setdiff(rownames(term.no), train.no.row), 0.87\*1930/0.13)

valid.no<-term.no[valid.no.row,]

validation.bank.df<-rbind(valid.yes, valid.no)

*# Checking again*

training.checking<-group\_by(training.bank.df, y)

training.checking<-summarise(training.checking, count=n())

validation.checking<-group\_by(validation.bank.df, y)

validation.checking<-summarise(validation.checking, count=n())

validation.checking<-mutate(validation.checking, percent=count/sum(count))

training.checking%>%kable(caption="Table: Checking training set after oversampling")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

validation.checking%>%kable(caption = "Table: Checking validation after oversampling")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

7. Building logistic regression model on training dataset

Model 1: All predictors as I mentioned above except for emp.var.rate to avoid multicollinearity

Original model

mod1<-glm(isyes~job + marital + education + contact+ month + poutcome + duration + campaign+ euribor3m+nr.employed, data = training.bank.df, family = binomial(link="logit"))

summary(mod1)

Model with forward selection

mod1.forward<-step(mod1,direction = "forward")

## Start: AIC=2763.13

## isyes ~ job + marital + education + contact + month + poutcome +

## duration + campaign + euribor3m + nr.employed

summary(mod1.forward)

Model with backward elimination

mod1.backward<-step(mod1, direction="backward")

Model with stepwise regression

mod1.step<-step(mod1, direction = "both")

Model 2: All predictors as I mentioned above except for euribor3m and nr.employed to void multicollinearity

Original model

mod2<-glm(isyes~job + marital + education + contact+ month + poutcome + duration + campaign+ emp.var.rate, data = training.bank.df, family = binomial(link="logit"))

summary(mod2)

Model with forward selection

mod2.forward<-step(mod2,direction = "forward")

Model with backward elimination

mod2.backward<-step(mod2, direction="backward")

Model with stepwise regression

mod2.step<-step(mod2, direction = "both")

8. Evaluating model on validation dataset

MODEL 1: The predictors: Job, Education, Month, Poutcome, Duration, Euribor3m and nr.employed

*# Using training data to predict the probability in validation dataset*

mod1.backward.pred<-predict(mod1.backward, validation.bank.df, type = "response")

data.frame('actual' = validation.bank.df[,21], "probability" = mod1.backward.pred)%>%head(10)%>%kable(caption = "Table: Customer's propensity on Term Deposit")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# Using cutoff figure to select the best cutoff value*

accT = c()

for (cut in seq(0,1,0.1)){

cm <- confusionMatrix(factor(ifelse(mod1.backward.pred> cut,1,0)), factor(validation.bank.df$isyes))

accT = c(accT, cm$overall[1])

}

plot(accT ~ seq(0,1,0.1),main="Figure: Plotting of accuracy and overall error", xlab = "Cutoff Value", ylab = "", type = "l", ylim = c(0, 1))

lines(1-accT ~ seq(0,1,0.1), type = "l", lty = 2, col="blue")

legend("bottomright", c("accuracy", "overall error"), lty = c(1, 2), merge = TRUE)

kbl(data.frame(Actual=validation.bank.df$y[1:6], Predicted= mod1.backward.pred[1:6]), caption = "Table: The prediction of backward model 1")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# Confusion matrix with cutoff value of 0.85*

mod1.conf<-confusionMatrix(factor(ifelse(mod1.backward.pred >0.85,1,0)), factor(validation.bank.df$isyes))

mod1.conf$table%>%kable(caption="Table: Confusion matrix of model 1 at cutoff value 0.85")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

MODEL 2: The predictors: Job, Education, Conact type, Month, Poutcome, Duration and Emp.var.rate

*# Using training data to predict probability in validation dataset*

mod2.backward.pred<-predict(mod2.backward, validation.bank.df, type="response")

data.frame('actual' = validation.bank.df[,21], "probability" = mod2.backward.pred)%>%head(10)%>%kable(caption = "Table: Customer's propensity on Term Deposit")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# Using cutoff figure to select the best cutoff value*

accT2 = c()

for (cut in seq(0,1,0.1)){

cm <- confusionMatrix(factor(ifelse(mod2.backward.pred> cut,1,0)), factor(validation.bank.df$isyes))

accT2 = c(accT2, cm$overall[1])

}

plot(accT2 ~ seq(0,1,0.1),main="Figure: Plotting of accuracy and overall error", xlab = "Cutoff Value", ylab = "", type = "l", ylim = c(0, 1))

lines(1-accT2 ~ seq(0,1,0.1), type = "l", lty = 2, col="blue")

legend("bottomleft", c("accuracy", "overall error"), lty = c(1, 2), merge = TRUE)

kbl(data.frame(Actual=validation.bank.df$y[1:6], Predicted= mod2.backward.pred[1:6]), caption = "Table: The prediction of backward model 2")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# Confusion matrix with cutoff value of 0.85*

mod2.conf<-confusionMatrix(factor(ifelse(mod2.backward.pred >0.85,1,0)), factor(validation.bank.df$isyes))

mod2.conf$table%>%kable(caption="Table: Confusion matrix of model 1 at cutoff value 0.85")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

9. Judging ranking performance of model 1

Lift chart

gain<-gains(validation.bank.df$isyes, mod1.backward.pred, groups = length(mod1.backward.pred))

plot(c(0,gain$cume.pct.of.total\*sum(validation.bank.df$isyes))~c(0, gain$cume.obs), xlab="# Cases", ylab="Cumulative", main = "Figure: Lift chart of customer's propensity ranking", type="l")

lines(c(0,sum(validation.bank.df$isyes))~c(0,dim(validation.bank.df)[1]), lty=2)

Decile chart

decile.data<-data.frame("actual" = validation.bank.df$isyes, "probability" = mod1.backward.pred)

gain.mod1<-gains(decile.data$actual, decile.data$probability)

midpoints.mod1 <-barplot(gain.mod1$mean.resp/mean(decile.data$actual), names.arg= gain.mod1$depth, ylim = c(0,6),

xlab = "Percentile", ylab = "Mean Response Yes", main = "Figure: Decile-wise lift chart")

text(midpoints.mod1, gain.mod1$mean.resp/mean(decile.data$actual)+0.5, labels=round(gain.mod1$mean.resp/mean(decile.data$actual), 1), cex = 1)

Case: Predicting the number of customers who will purchase term deposit is more important

*# At cutoff value: 0.5*

mod1.conf1<-confusionMatrix(factor(ifelse(mod1.backward.pred >0.5,1,0)), factor(validation.bank.df$isyes))

mod1.conf1$table%>%kable(caption="Table: Confusion matrix of model 1 at cutoff value 0.85")%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# The accuracy rate*

mod1.conf1$overall[1]

## Accuracy

## 0.843931

*# Plotting the relationship between sensitivity and cutoff value*

accuracy1 = c()

sensitivity1 = c ()

for (cut in seq(0,1,0.1)){

mod1.conf.matrix <- confusionMatrix(factor(ifelse(mod1.backward.pred >cut,1,0)), factor(validation.bank.df$isyes))

accuracy1 = c(accuracy1, mod1.conf.matrix$overall[1])

sensitivity1=c(sensitivity1, mod1.conf.matrix$byClass[2])

}

plot(accuracy1 ~ seq(0,1,0.1),main="Figure: Plotting of accuracy and percentage", xlab = "Cutoff Value", ylab = "", type = "l", ylim = c(0, 1))

lines(sensitivity1 ~ seq(0,1,0.1), type = "l", lty = 2, col="blue")

legend("bottomleft", c("accuracy", "percentage"), lty = c(1, 2), merge = TRUE)

10. K-Nearest Neighbors (k-NN)

*# Creating knn dataset with selected predictors: job, eduation, month, contact type, poutcome, duration, euribor3m, and nr.employed*

knn.df<-bank.df2[,c(2,4,9,11,15,19,20,21,22)]

*# Converting Categorical Variables to Binary Dummies*

knn.df<-fastDummies::dummy\_cols(knn.df, select\_columns = "job")

knn.df<-fastDummies::dummy\_cols(knn.df, select\_columns = "education")

knn.df<-fastDummies::dummy\_cols(knn.df, select\_columns = "month")

knn.df<-fastDummies::dummy\_cols(knn.df, select\_columns = "poutcome")

*# Partitioning data into 60% training set and 40% validation set*

set.seed(5)

train.knn.index<-sample(row.names(knn.df), 0.6\*dim(knn.df)[1])

valid.knn.index<-setdiff(row.names(knn.df), train.knn.index)

train.knn<-knn.df[train.knn.index,]

valid.knn<-knn.df[valid.knn.index,]

*# Normalizing duration, euribor3m, nr.employed*

train.knn.normalized<-train.knn

valid.knn.normalized<-valid.knn

normalized.values<-preProcess(knn.df[,c(4,6,7)], method=c("center", "scale"))

train.knn.normalized[,c(4,6,7)]<-predict(normalized.values, train.knn[,c(4,6,7)])

valid.knn.normalized[,c(4,6,7)]<-predict(normalized.values, valid.knn[,c(4,6,7)])

Running KNN algorithm with different K values

accuracy.df<-data.frame(k=seq(1,14,1), accuracy=rep(0,14))

select.variable<-c(4,6,7,10:25)

*# Computing knn for different k on validation dataset*

for (i in 1:14) {

knn.prediction<-knn(train.knn.normalized[,select.variable], valid.knn.normalized[,select.variable], cl=train.knn.normalized$isyes, k = i)

accuracy.df[i,2]<-confusionMatrix(factor(knn.prediction), factor(valid.knn.normalized$isyes))$overall[1]

}

accuracy.df%>%kable(caption="Table: accuracy rate at different K values", digits=3)%>%kableExtra::kable\_styling(bootstrap\_options = "striped")

*# Confusion matrix at k = 13*

knn.prediction.13<-knn(train.knn.normalized[,select.variable], valid.knn.normalized[,select.variable], cl=train.knn.normalized$isyes, k = 13)

confusionMatrix(factor(knn.prediction.13), factor(valid.knn.normalized$isyes))

11. Naive Bayes Classifier for categorical predictors

*# Creating Naivee Bayes dataset*

naive.df<-bank.df2[,c(2,4,9,11,15,19,20,21,22)]

*# Creating bin and Changing numerical variables to factor*

naive.df$isyes<-factor(naive.df$isyes)

naive.df$duration<-factor(round(naive.df$duration/100))

naive.df$euribor3m<-factor(round(naive.df$euribor3m))

naive.df$nr.employed<-factor(round(naive.df$nr.employed))

*# Partitioning data into 60% training data and 40% validation data*

set.seed(3)

train.naive.index<-sample(row.names(naive.df), 0.6\*dim(naive.df)[1])

valid.naive.index<-setdiff(row.names(naive.df), train.naive.index)

train.naive<-naive.df[train.naive.index,]

valid.naive<-naive.df[valid.naive.index,]

Running Naive Bayes algorithm

isyes.naive<-naiveBayes(isyes~.-y, data=train.naive)

isyes.naive

Predicting probabilities and class membership on validation data

*# Predicting probabilities*

pred.naive.prob<-predict(isyes.naive, newdata = valid.naive[,-8], type="raw")

*# Predicting class membership*

pred.naive.class<-predict(isyes.naive, newdata = valid.naive[,-8])

Evaluating Naive Bayes model by confusion matrix

confusionMatrix(factor(pred.naive.class), factor(valid.naive$isyes))