**ANALYZING AND PREDICTING BANK MARKETING STRATEGY**

COURSE: ALY6015 Intermediate Analytics

INSTRUCTOR: Roy Wada

WINTER QUARTER\_B, 2019

GROUP-INFO

**Ankit Khatavkar**

**Jigar Panchal**

**Shivani Raina**

DATE: 31st MARCH 2019



**TABLE OF CONTENTS**

1. Introduction

2. Problem Statement

3. Data Information and Description

4. Method used

5. Results and Findings

6. Conclusion

7. References

**INTRODUCTION**

Marketing campaigns are specified by keeping in consideration of customer needs and their overall satisfaction. There are different variables that determine the success of marketing campaigns.

The Four P’s which will help to determine the marketing campaigns are:

Segment of the Population: Which section of the population marketing campaign is going to represent and why? and this will help to know which section of the population have higher chances of receiving the messages of marketing campaign.

Distribution channel to reach the customer's place: How to get most customers through the marketing campaign and which instrument will be helpful to get the message out

Price: What is the best price to offer to potential clients?

Promotional Strategy: In-depth analysis of the previous campaigns which will give insights on previous patterns and mistakes through this effective way for current campaign could be find out.

A term store is a fixed-term store of cash held at a financial institution. Term stores are typically transient stores with developments extending from one month to a couple of years. At the point when a term store is acquired, the customer comprehends that the cash must be pulled back after the term has finished or now and again, by giving a foreordained number of days take note.

Our principle challenge lies in accomplishing high precision in the expectation part of our modelling procedure. Since the information is delicate and needs the most extreme precision to be kept up while foreseeing. We at first begin with understanding the idea of the information and picturing its angles and the relationship it holds among itself so as to show better and eﬃciently. We at that point hold a near investigation of the displaying methods lastly select the most appropriate model which fits the given dataset with most extreme precision.

**PROBLEM STATEMENT**

Our key objective is to analyze the data attributes provided by the dataset we have currently chosen i.e. banking.csv, We are going to use data from a marketing campaign implemented by a major banking institution. The outcome, y, is whether or not a bank salesperson was able to get a client to sign up for a term deposit (and is labeled 0 for no, and 1 for yes). The objective is to utilize classification algorithms to help the bank management and sales team understand how to maximize clients signing up for a term deposit.

**DATA INFORMATION AND DESCRIPTION**

The data frame is pulled from the UCI machine learning repository. The banking data frame has a total 45211 obs. of 17 variables. The dataset gives the information about a marketing campaign of a financial institution which can be analyzed in order to find ways to look for future strategies and improve future marketing campaigns for the bank.

Number of observations and variable names:

|  |  |
| --- | --- |
| 'data.frame' : | 45211 obs. of 17 variables: |
| $ Age : | int 58 44 33 47 33 35 28 42 58 43 ... |
| $ Maritial.status : | Marital status, Factor w/ 3 levels "divorced","married" |
| $ Education : | Factor w/ 4 levels "primary","secondary" |
| $ Default : | Has credit in default?, Factor w/ 2 levels "no","yes" |
| $ Housing | Has housing loan?, Factor w/ 2 levels "no","yes" |
| $ Contact : | Contact communication type , Factor w/ 3 levels "cellular","telephone" |
| $ Day : | Last contact day of the week, int |
| $ Duration : | Last contact duration, in seconds numeric |
| $ Campaign : | Number of contacts performed during this campaign and for this client, Numeric (includes last contact) |
| $ Pday : | Number of days that passed by after the client was last contacted from a previous campaign numeric; 999 means clients were not previously contacted |
| $Previous: | Previously was part of campaign, int [1:41188] 0 0 0 0 |
| $y | Whether or not a bank salesperson was able to get a client to sign up for a term deposit (and is labeled 0 for no, and 1 for yes). |

**Table 1.1**

Therefore, in the above table named as Table 1.1, data is described clearly which describes the type of data. For example, marital status, education, month0, contact, housing is categorical in nature and class, previous, duration, and age.

We will now move on by first finding the descriptive statistics using the summary function.

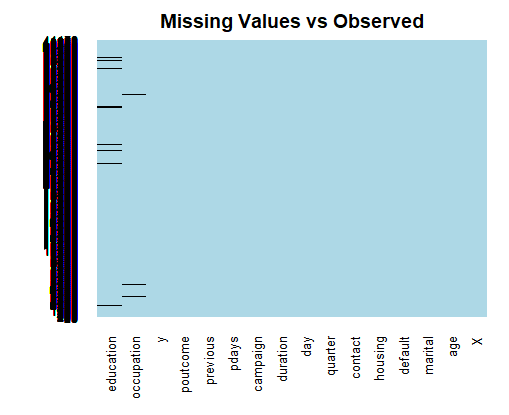
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Age | Marital | Education | Occupation | Default | Housing |
| Min: 1 | Min: 17 | Divorced: 4612 | Low: 12531 | Min: 0 | No: 32588 | No: 18622 |
| 1st Qu:10298 | 1st Qu: 32 | Married: 24928 | Intermediate: 9515 | 1st Qu: 0 | Unknown: 8597 | Unkown: 990 |
| Median: 20595 | Median: 38 | Single: 11568 | High: 17411 | Median: 1 | Yes: 3 | Yes: 21576 |
| Mean: 20595 | Mean: 40.02 | Unknown: 80 | Unknown: 1731 | Mean: 8.737 |  |  |
| 3rdQu: 30891 | 3rd Qu: 47 | 3rd Qu: 1 |  |  |  |  |
| Max: 41188 | Max: 98 | Max: 999 |  |  |  |  |

**Table1.2**

The above table named as Table 1.2 shows the descriptive statistics for some of the fields in the banking dataset who wish to enroll for the term deposit.

After, this the next step is to check the NA values, therefore with the help of missmap() function we were able to find the NA values.

The below plot shows the NA values for education and occupation.



**Fig 1.1**

**Interpretation:**

From the above plot named as Figure 1.1, we can summarize that education and occupation have missing values. It has been observed that there were 1731 missing values in education and 330 missing values in occupation. We aimed to replace the NA value by assigning 999 to those NA values as this would help to know the impact of these values in future classification. Now we proceed with exploring the data and summarizing the key indicators and their distribution.

Replaced NA with 999:

banking$education[is.na(banking$education)]<- 999 #Replace NA by introducing new integer value: 999

banking$occupation[is.na(banking$occupation)]<- 999 ##Replace NA by introducing new integer value: 999

**METHOD USED**

Next step is to generate the intercept-only model. It is a model that essentially speaks to a perfect circumstance where each client will agree to accept term store regardless of different factors. In other words, when no predictors are present (y~1), it will explain the trend component of the target variable. This model gives the mean estimation of the information. So, after obtaining the intercept value, next step is to convert this to probability which is computing odds and converting the odds to probability. The probability obtained will define the chances accepting the term deposit by customers.

We have use logistic regression to find out whether campaign was more successful among lower vs. higher educated clients, therefore we will get the answer for below questions:

What is the probability that the lowest education group (education=0) signed up for a term deposit (y=1) in response to this campaign?

What is the probability that the highest education group (education=2) signed up for a term deposit (y=1) in response to this campaign?

banking$education<-factor(banking$education, labels=c("LOW","INTERMEDIATE","HIGH","UNKNOWN"))

str(banking$education)

levels(banking$education)

education\_data<- glm(y~education, data= banking, family='binomial')

With the help of cross table we are able to find the probability for type of education to get the maximum number of people who would sign up for term deposit.

banking$y

banking$education no yes Total

--------------------------------------------------------------------------------------------------------------------

LOW 11438 1093 12531

9.133 71.936

0.913 0.087 0.304

0.313 0.236

0.278 0.027

INTERMEDIATE 8484 1031 9515

0.198 1.561

0.892 0.108 0.231

0.232 0.222

0.206 0.025

HIGH 15146 2265 17411

5.965 46.986

0.870 0.130 0.423

0.414 0.488

0.368 0.055

UNKNOWN 1480 251 1731

2.041 16.079

0.855 0.145 0.042

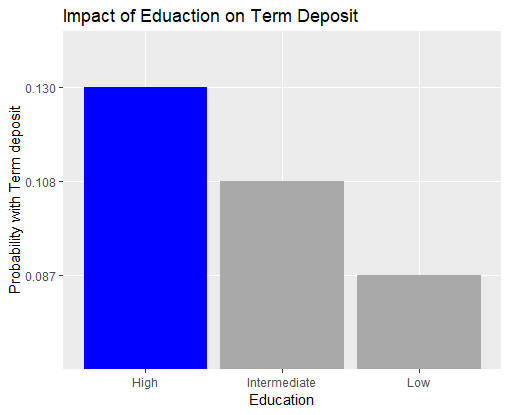
0.040 0.054

0.036 0.006

Total 36548 4640 41188

0.887 0.113

The above probability helps to know the impact of education on Term Deposit



**Fig 1.2**

**Interpretation:**

From the above figure labeled as Figure 1.2, we found out the probability of getting Yes to Term deposit by Highly and Lowly educated client we have used logit2prob function. Through the results, it can be concluded that ratio of highly educated people making up the term deposit is more than the lowly educated people.

The second factor which can be taken into consideration is which day of the week is best to contact clients, that means which day will result the highest probability of client term deposit sign ups. To get the results, first we have converted the day variable from integers to factors. The reason to do this is factors are more descriptive than integers, "Monday", "Tuesday", "Wednesday”, “Thursday”, “Friday” is more descriptive than 1, 2, 3.

Days <- c("Monday","Tuesday","Wedenesday","Thursday","Friday")

Probabilities <- c(0.1080874,0.1177998,0.1166708,0.1211875,0.1080874)

str(Days)

str(Probabilities)

days\_table<-data.frame(days=Days,Prob=Probabilities)

CrossTable(banking$day,banking$y)

With the help of cross table we are able to find the probability for each weekdays to get the best day in a week to sign up for term deposit.

banking$y

banking$day no yes Total

--------------------------------------------------------------------------------------------------------------------

Friday 6981 846 7827

0.184 1.449

0.892 0.108 0.190

0.191 0.182

0.169 0.021

Monday 7667 847 8514

1.664 13.111

0.901 0.099 0.207

0.210 0.183

0.186 0.021

Thursday 7578 1045 8623

0.708 5.574

0.879 0.121 0.209

0.207 0.225

0.184 0.025

Tuesday 7137 953 8090

0.241 1.901

0.882 0.118 0.196

0.195 0.205

0.173 0.023

Wednesday 7185 949 8134

0.148 1.165

0.883 0.117 0.197

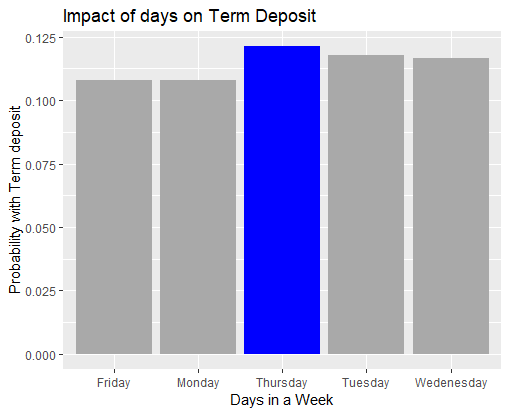
0.197 0.205

0.174 0.023

Total 36548 4640 41188

0.887 0.113

The above probabilities help to plot the below graph for finding the impact of days on Term Deposit.



**Fig 1.3**

**Interpretation:**

To get the intercept and slope of each levels in the “Day” variable logistic regression is implemented. After finding the probability of each coefficient, and with the help of above plot labeled as Fig 1.3. we estimated that Thursday has the most significant coefficient and it results the highest probability.

Now moving forward, we have also calculated the probability for the occupation category. Instead of specifying some specific occupation, we have labeled them as A, B and C. The below graph demonstrates the impact of occupation on Term Deposit.

With the help of cross table we are able to find the probability for each occupation to get the best occupation to interest customers to sign up for term deposit.

banking$y

banking$occupation no yes Total

---------------------------------------------------------------------------------------------------------------------

0 13216 1067 14283

23.182 182.598

0.925 0.075 0.347

0.362 0.230

0.321 0.026

1 20283 2683 22966

0.450 3.546

0.883 0.117 0.558

0.555 0.578

0.492 0.065

2 2756 853 3609

62.234 490.202

0.764 0.236 0.088

0.075 0.184

0.067 0.021

999 293 37 330

0.000 0.001

0.888 0.112 0.008

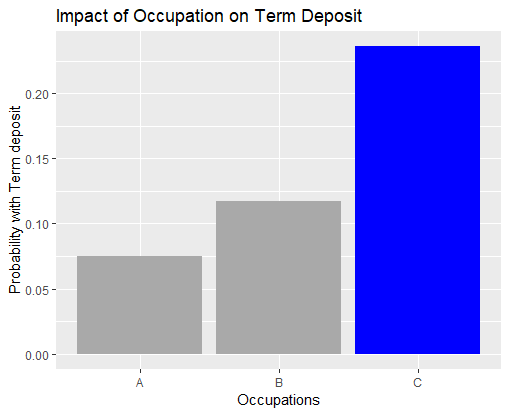
0.008 0.008

0.007 0.001

Total 36548 4640 41188

0.887 0.113

The above probabilities help to plot the below graph for finding the impact of occupation on Term Deposit.

  
**Fig 1.4**

**Interpretation:**

From the above graph labeled as Fig 1.4, we are able to predict the probability that occupation C has more impact on term deposit which means people from occupation C are more like to enroll for term deposit.

**RESULTS AND FINDINGS**

To build the predictive model data is split into two parts training data and test data. The 80% of the data is in training and 20% of the data is in test data. logistic regression model is used. Logistic Regression is one of the prediction algorithms used to conduct regression analysis when the dependent variable is dichotomous (binary). In our case y (Term deposit) is the dependent variable and the variables in the training data is independent variable. The graph obtained from the regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent variables.

split<- sample(seq\_len(nrow(banking)),size= floor(0.80\*nrow(banking)))

train\_data<- banking[split, ]

test\_data<- banking[-split, ]

head(train\_data)

head(test\_data)

banking\_model<- glm(y~., data= train\_data,family=binomial(link='logit'))

After finding the logistic regression of the training data, predict function is used to predict the outcome for the test dataset.

test\_data$predicted\_data<- predict(banking\_model,test\_data,type='response')

head(test\_data$predicted\_data, 10)

test\_data$predicted\_Y\_value<- ifelse(test\_data$predicted\_data>0.5, 1,0)

head(test\_data$predicted\_Y\_value, 20)

We have then used a confusion matrix for proceeding with the following results. A confusion matrix is a synopsis of expectation results on a classification problem. The quantity of right and erroneous predictions is described with count values and separated by each class. This is the way to the confusion matrix. The confusion matrix demonstrates the manners by which our classification model is befuddled when it makes forecasts. It gives us knowledge not just into the errors being made by a classifier however more significantly the kinds of mistakes that are being made. Therefore, we have used confusion matrix to measure the number of correct and incorrect predictions.

confusion\_matrix\_table<- table(test\_data$predicted\_Y\_value, test\_data$y)

(7085+381)/sum(confusion\_matrix\_table)

library(InformationValue)

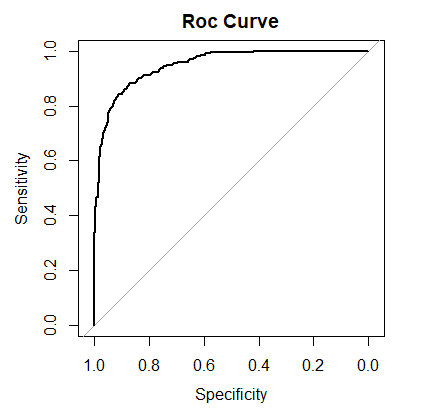
mis\_clasification\_error<- mean(test\_data$predicted\_Y\_value != test\_data$y)

model\_accuracy<- 1-mis\_clasification\_error

miss\_class\_value<- misClassError(as.integer(test\_data$y), as.integer(test\_data$predicted\_Y\_value), threshold = 0.5)

accuracy<- 1-miss\_class\_value # 91 %

With the help of predicted values, we will then find the accuracy of the model using Receiver Operating Characteristics (ROC) Curve. ROC stands for Receiver Operating Characteristic and is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity and false-positive rate is known as (1 − specificity).



**Fig 1.5**

The above plot named as Fig 1.5 is the ROC curve, which defines the accuracy of the model which is 91% in our case. Therefore, it can be said as our model is 91% accurate. The ROC Curve follows the level of genuine positives precisely anticipated by a given logit display as the prediction probability cutoff is brought from 1 down to 0. For a decent model, as the cutoff is brought down, it should check a greater amount of real 1's as positives and lesser of real 0's as 1's. So, for a decent model, there should be steep rise in curve, and the estimation of ROC ought to be nearer to 1.0

**CONCLUSION**

Through results we can determine that the highest level of marketing activity was on Thursday which means the probability of getting yes to Term deposit was on Thursday. For the next marketing campaign, it would be advisable to focus the marketing campaign on Thursday of every week. The level of education has huge difference on the impact of education on term deposit. The customers who are more educated or aware are more inclined to accept the term deposit. The probability of signing up for term deposit is also based on the occupation of the customer. The customer who is well earned is more likely to sign up for the term deposit. By observing the results of effect of different independent variable on the term deposit, strategies for the next campaign can be made and it is likely that the next campaign of the bank will be more powerful than the present one.

**REFERENCES**

1. Alice, M. (2018, June 24). How to Perform a Logistic Regression in R. Retrieved from <https://datascienceplus.com/perform-logistic-regression-in-r/>
2. Logistic Regression in R Tutorial. (n.d.). Retrieved from <https://www.datacamp.com/community/tutorials/logistic-regression-R>
3. Sachs, M. C. (2018, June 23). Retrieved from <https://cran.r-project.org/web/packages/plotROC/vignettes/examples.html>
4. Bachmann, J. M. (2017, November 12). Bank Marketing Dataset. Retrieved from <https://www.kaggle.com/janiobachmann/bank-marketing-dataset>
5. Amelia. (n.d.). Retrieved from <https://www.rdocumentation.org/packages/Amelia/versions/1.7.5/topics/missmap>
6. What is a Term Deposit and how does it work?. (2016, July 27). Retrieved from <https://www.canstar.com.au/term-deposits/what-is-a-term-deposit>
7. Confusion Matrix in Machine Learning. (2018, February 07). Retrieved from <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>