

**Predicting the Bank Customer Tendencies in Subscribing the Bank's
Product Through Logistic Regression Model**

**AIT 580
Data Analytics Research
IWAN
G01236399**

Professor: Harry J Foxwell, Phd

Abstract

The times' efficiencies in obtaining new customer or the existing customer to buy or use new product service or the existing product services is a challenge of all financial institutions especially for banks. the banks obviously need to campaign their services. The campaigns are conducted by their sales marketing [3]. By conducting the campaigns of their services product, the customer will get known well about the services which are provided by the banks. Other that, since most financial institution nowadays, such as banks, have more data about their prospective customers, they will able to establish and use the platform and automation tools which supported by the big data method [4]. Such platform can ease the banks by telling information about their customers who are saving up for a big purchase and most likely need banking products to solve the problem that they face [4]. this study will implement one of the data analytics methods which called logistic regression to determine whether the banks customers will subscribe the product prior to make a phone call to the customers.

1. Introduction

A bank is a financial institution that functions to provide services such as deposit and loan [1]. Specifically, Banks provide a safe place for their customers to store their cash [1]. Furthermore, the cash which is stored by the customer can be lent to their other customers since the function of banks is also to provide loans for the society [1].

Indeed, providing the services for customer should give benefits for the customer and for the banks as well. For instance, after the customer open their account in a bank and store their cash, the customer will also able to do routine banking transaction such as deposits, withdrawals, check writing, and bill payment [1]. Such facilitations are the examples of what customer will get through their deposit account. Yet, banks also receive benefits by providing such facilities. The benefits come from the fee of all the services which are provided to the customers [2]. For example, the monthly maintenance fee that should paid by the customers every month and such fee is charged once a month to their customers [2].

To provide such services, the banks obviously need to campaign their services in products form. The campaigns are conducted by their sales marketing [3]. By conducting the campaigns of their services product, the customer will get known well about the services which are provided by the banks. The campaign could be done in several ways. For instances, the campaign can be conducted through the Automation and Big Data method [4]. Since most financial institution nowadays, such as banks, have more data about their prospective customers, they will able to establish and use the platform and automation tools which supported by the big data method [4]. Such platform can ease the banks by telling information about their customers who are saving up for a big purchase and most likely need banking products to solve the problem that they face [4].

According to how the bank could ease the marketing product sales, this study will implement one of the data analytics methods which called logistic regression to determine whether the banks customers will subscribe the product prior to make a phone call to the customers. In addition, the customers' information dataset is derived through the UCI Machine Learning repository website [5].

2. Review of Literature

In this section, there are three components that will be discussed. First, the bank customer dataset that will be used to conduct the research. Second, the data preparation that

will be done prior to implement the method of the research. Third, the logistic regression as the method to conduct the research so that the answer whether the customer will subscribe the banks product or not can be obtained.

a. The Importance of Customers' Dataset

Dataset is collection of information which is organized into some type of structure [6]. The data collection might contain several information such as names, salaries, contact information, etc. [6].

Such dataset is important since it will be used to analyze then banking customer behavior [7]. For instance, the customer dataset can be used to build a framework called Intelligence Customer Analytics for Recognition [7]. The function of such framework is to provide an insight about customers so that some requirements and data environment can be satisfied [7].

Other than that, the utilization of the bank customer dataset can also provide the perception about digital banking to the society [8]. According to such study, the dataset was successfully helped to improve the financial performance, bank marketing services, and the digital banking solution through the multivariate factor analysis, structural equation modelling, and analysis of variance test [8].

The dataset in this research contains information of the banks' customer such as age, job, products that already owned, and so on. All such information will be used as the predictors and response to determine whether the customer will prefer to subscribe the banks product. However, such data should through the cleansing process that will be discussed in the further section of this chapter.

b. Data Preparation for Customer Banking Analytics

This section discusses the data preparation that will be used in the methodology chapter. The step of data preparation takes a lot of time based on the experience of this research. Other than that, generally, most of data scientist spend most of their time for doing this step [9].

The data preparation step contains two kinds of work such as finding and cleaning the data [9]. Finding the data might be the most critical issue in any data science project since there are several questions must be answered in order to obtain the data that can be used for the further analysis such as the follows [9]:

- Who might have the data?
- Whether the data is available for me?
- How can I get my hands on it?

Specifically, since the data will be used for the logistic regression model, the customer data should contain the predictors which fit to logistic regression rule. For instance, while conducting the statistical analysis for internet banking usage, the predictors contain the customer data such as age, gender, marital status, and education [10]. However, there is one variable that functions as binary respond that represent whether the customer is an internet banking user or not [10]. Therefore, in this research,

the customer dataset that will be used contains the predictors and one binary response variable, which is y, to predict whether the customer will subscribe the product.

Afterwards, the data separation process, which a part of the data preparation, will be done by splitting the into two category such as train and test dataset. For example, while conducting the credit scorecard through the logistic regression model, the train and test dataset was prepared by randomly selecting 9 subset as the training set and 1 set as testing [11]. However, in this research the dataset will be divided into 70% for data train and 30% for data test since the data that satisfies the model is limited.

c. Logistic Regression Model for Marketing

This model is powerful and common especially in marketing fields [12]. Such reason brings this research to use Logistic Regression as the modelling method since the dataset which is used by this research contains customers information. Other than that, the purpose of this research is to find out whether the model works well to predict the customers' inclination towards the product.

For instance, it has already been proved that the logistic regression works well for examining the impact of customer's previous transaction towards their tendency in loyalty and general cross-buying [13]. In such study, the logistic regression models were implemented by combining the incorporating panel data from a large bank [13]. The result of the research shows that the large variation of customers was observed, and shows which customer that have the tendency to the cross-category financial service purpose [13].

The output of this model is binomial [14]. In particular, the response of this model can be "true" or "false", "0" or "1", and so on [14]. Therefore, as mentioned in the previous section, y in the dataset functions as the predictor and the value is converted into 0 and 1 so that all the predictors including the can perfectly fit the logistic regression rule.

Since the modelling in this research will be conducted using R, the logistic regression formula will be such as the follows:

<code>glm(response ~ predictor 1 + predictor 2 + ... + predictor N)</code>
--

Once the all the coefficients are obtained and the r-square shows the closest number to 1, the testing will be done based on the confusion matrix [14]. The confusion matrix is created to see the correct and the incorrect proportion of the response [14]. Specifically, the confusion matrix contains four components such as true positives, true negatives, false positives, and false negatives [14]. Such components will be used to perform the proportion of the correct and the incorrect responses.

In addition, since of the limitation dataset to conduct the testing, this research will separate the data into two classifications, which are training and testing dataset. The separations' portion will be 30% for the training dataset and 70% for the testing dataset.

3. Methods

This methodology chapter contains more specific about how the data is prepared, how the data is modeled, and how the data is tested in this study. In addition, the dataset which conducted in this study is derived through the UCI Machine Learning Repository [5].

a. The Customer Dataset

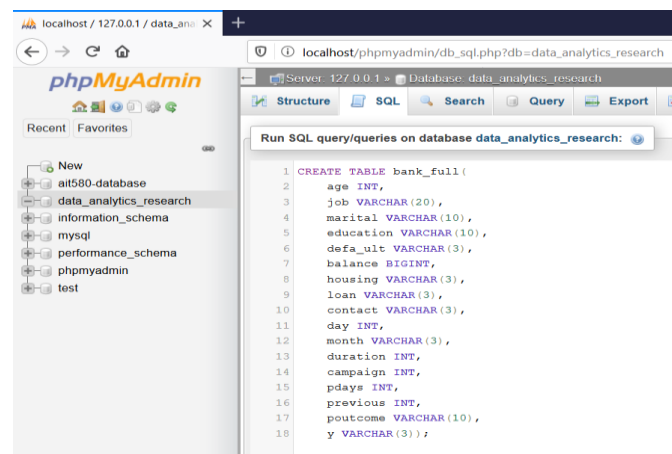
The dataset for this research contains information of the banks' customer. Specifically, the data contains the following information [5]:

- **Age:** the information of the customers' age. The data type of this information is numerical
- **Job:** the information of the customers' profession. Specifically, what do they do for living. The data type of this information is categorical
- **Marital:** it tells the status of the customer whether they are single, married, or divorced. The data type of this information is categorical
- **Education:** the information of the customers' education level. This information comprises four categories. The first category is primary which means elementary education [15]. The second category is secondary education which means junior/senior high school. This level is attended right after graduating the elementary school and before entering the university level [16]. The third one is tertiary which means the education which is attended right after graduating the senior high school such as diplomas, undergraduate and graduate certificates, bachelors, masters, and doctoral [17]. The last category is unknown which means the education levels of the customer is not identified. The data type of this information is categorical
- **Default:** this data tells about whether the customer owns any credit facility which functions to spend up money based on the given limit and should be paid again including the bank interest [18]. The data type of this information is categorical
- **Balance:** This information tells the current balance of the customer in the bank. The data type of this information is numerical
- **Housing:** It informs the whether the customer have any house loan facility. This facility is a sum of money given by a financial institution to purchase a house [19]. The data type of this information is categorical
- **Loan:** It tells the information about whether the customer have any personal loans facilities. Personal loan facility is a loan which have a low interest rates for consumers and usually the credit amount is smaller than the other types of loan [20]. The data type of this information is categorical
- **Contact:** This information tells about the communication types of the customer. Specifically, it tells whether the customer can be contacted via cellular or telephone. The data type of this information is categorical
- **Day:** It informs about the last day of contacted in a week. The data is supposed to be the name of days in week. However, they are already converted into

number based on the day sequence name in a week such 1 is for Monday, 2 is for Tuesday, and so on. The data type of this information is categorical

- **Month:** This data informs the last time of the customer being contacted. The data are named based on the name of months. The data type of this information is categorical
- **Duration:** It informs the duration in second at the last time when the customer was contacted. The data type of this information is numerical
- **Campaign:** It shows the number of called that have been done to the customer including the last time contacted for this products' campaign. The data type of this information is numerical
- **Pdays:** It tells the number of days that have been passed since the last day the customer was contacted. The data type of this information is numerical
- **Previous:** It shows the number of called that have been done to the customer prior to this campaign. The data type of this information is numerical
- **Poutcome:** This is the result of the previous marketing's campaign. Specifically, it tells whether the campaign was successful or not. The data type of this information is categorical
- **Y:** it informs whether the result of this products' campaign is successful or not. It also functions as the response variable. The output of this data is yes or no, and it will be converted further to be 1 or 0 to fit the logistic regression method. The data type of this information is categorical

To give a better picture of the dataset, the dataset will be demonstrated through the SQL schema as shows in figure 1.



The screenshot shows the phpMyAdmin interface. On the left, a sidebar lists databases: 'ait580-database', 'data_analytics_research', 'information_schema', 'mysql', 'performance_schema', 'phpmyadmin', and 'test'. The 'data_analytics_research' database is selected. The main panel displays the 'SQL' tab with a 'Run SQL query/queries on database data_analytics_research:' prompt. Below this, a SQL query is entered to create a table named 'bank_full' with the following schema:

```
1 CREATE TABLE bank_full (
2   age INT,
3   job VARCHAR(20),
4   marital VARCHAR(10),
5   education VARCHAR(10),
6   default VARCHAR(3),
7   balance BIGINT,
8   housing VARCHAR(3),
9   loan VARCHAR(3),
10  contact VARCHAR(3),
11  day INT,
12  month VARCHAR(3),
13  duration INT,
14  campaign INT,
15  pdays INT,
16  previous INT,
17  poutcome VARCHAR(10),
18  y VARCHAR(3));
```

Figure 1. SQL Schema to Create a table for Customer Dataset.

The SQL Schema on figure 1 shows the table creation in MySQL. The table creation also shows the data type of each predictors.

Furthermore, to portrays the dataset inside the table, the selection query will be conducted. The SQL schema including the result selection is shown by figure 2.

Figure 2 shows the phpMyAdmin interface with the following SQL query and results:

```
SELECT * FROM bank_full ORDER BY `age` DESC LIMIT 5
```

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	outcome	y
95	retired	divorced	primary	no	2282	no	no	tel	21	apr	207	17	-1	0	unknown	yes
95	retired	married	secondary	no	0	no	no	tel	1	oct	215	1	-1	0	unknown	no
94	retired	divorced	secondary	no	1234	no	no	cel	3	mar	212	1	-1	0	unknown	no
93	retired	married	unknown	no	775	no	no	cel	22	jul	860	2	177	7	success	yes
93	retired	married	unknown	no	775	no	no	cel	4	aug	476	2	13	9	success	yes

Figure 2. SQL Schema to view the data for Customer Dataset.

The SQL in figure 2 shows the five oldest customer data. Specifically, the data is sorted and limited for only the top five based on the oldest to the youngest age.

Lastly, to know the total row of the dataset, the SQL for counting the total row will be done as shown on figure 3.

Figure 3 shows the phpMyAdmin interface with the following SQL query and result:

```
SELECT COUNT(*) FROM bank_full
```

COUNT(*)
45211

Figure 3. SQL Schema to view the total row of the customer dataset.

According to SQL which shown on figure 3, the total number of the customer dataset's row is 45211. Such total row will further be used to do the logistic regression modelling. However, the data transformation will be done first through the next section.

b. Data Preparation

the modelling method that will be used in this research is logistic regression, the default of y predictors values is converted to be 1 and 0. Also, the data preparation will be done using python. The illustration process of the data preparation in this research is shown by figure 2.



Figure 4. Data Preparation Process

Specifically, since the first dataset the response variable still does not fit the binary response, the dataset is rebuilt through the following code:

```

import os # for OS interface (to get/change directory)
import pandas as pd # for data frame creation
os.chdir('D:/George Mason University/Semester 2/Analytics Big Data to Information/Data
Analytics Research Project/Dataset')
os.getcwd()

bank = pd.read_csv("bank-full.csv", sep=";")
bank.head()
bank.info()

answers = []
i = 0
while (i < len(bank)):
    if (bank.y[i] == "no"):
        answers.append(0)
    else:
        answers.append(1)
    i = i + 1

bank['answer'] = answers

df_new_bank = pd.DataFrame({'age': bank['age'],
                             'job': bank['job'],
                             'marital': bank['marital'],
                             'education': bank['education'],
                             'default': bank['default'],
                             'balance': bank['balance'],
                             'housing': bank['housing'],
                             'loan': bank['loan'],
                             'contact': bank['contact'],
                             'day': bank['day'],
                             'month': bank['month'],
                             'duration': bank['duration'],
                             'campaign': bank['campaign'],
                             'pdays': bank['pdays'],
                             'previous': bank['previous'],
                             'poutcome': bank['poutcome'],
                             'y': bank['answer']})

df_new_bank.to_csv(r'bank_customer.csv', index = False, header=True)
  
```

According to the code above, the value of the response variable is changed from “yes” or “no” to 1 or 0. Afterwards, the dataset is exported to another excel file, which is called bank customer.csv, such as shown on the data preparation process in figure 4.

After that, to give a little illustration of the customers’ dataset, some visualization of the customers dataset is performed. The visualization is performed through python. The visualizations are such the follows:

- **Customers Marital Status**

This visualization shows the Customers Marital Status summary. As explained before, the marital status is divided into three category such

as married, single, and divorced. The visualization of those three categories according to the dataset is shown by figure 5.

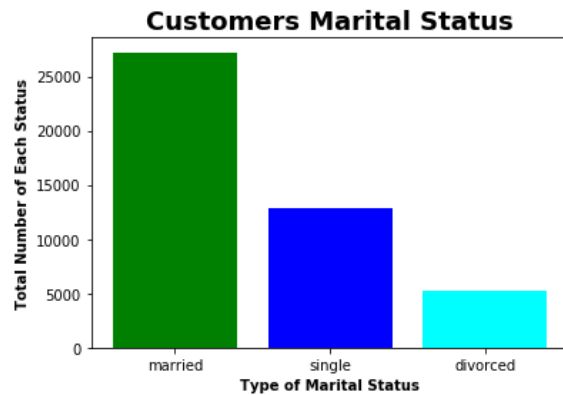


Figure 5. The Visualization of Customers Marital Status

According to Figure 5, the visualization concludes that the married customers has the highest number in the dataset.

- **Customers Education Level**

This visualization shows the Customers Educations Level summary. As explained before, the Educations' Level is divided into four category such as secondary, tertiary, primary, and unknown. The visualization of those four categories according to the dataset is shown by figure 6.

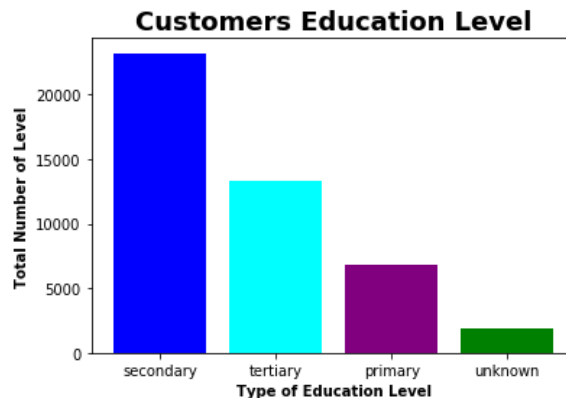


Figure 6. The Visualization of Customers Education Level

According to Figure 6, the visualization concludes that the customers that own the secondary education level, has the highest number in the dataset.

- **Customers Age Summary**

This visualization shows the Customers Age summary. The visualization is shown by figure 7.

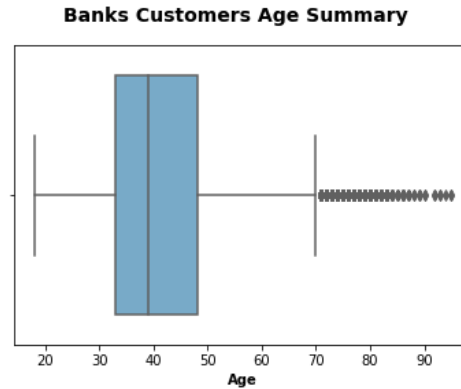


Figure 7. The Visualization of Customers Age

According to figure 7, the visualization concludes that the mean of the customers age is about 40-year-old. Other than that, the oldest customers in the dataset surpasses the age of 90.

Furthermore, after the data is successfully read through R studio, the data exploration will be done in R as well. The following code shows will result the scatterplot and the correlation of each predictor in the customer dataset.

```
library(tidyverse)
library(GGally)
ggscatmat(select(bank, -y)) + labs(title = "scatterplots and correlations of customer detail")
```

Through the code above, the scatterplots and correlation will be obtained. The result after running the code above is shown through figure 8.

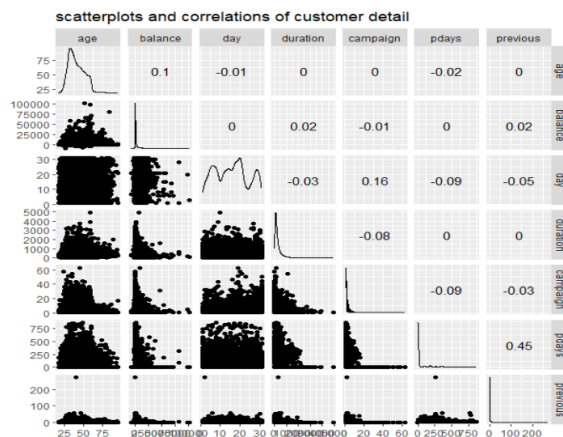


Figure 8. The scatterplots and correlation of the customer dataset variables.

According to figure 8, the highest correlation is shown by the correlation between previous and pdays predictors. To look closely the plot of correlation between those two predictors, the scatterplot of those two predictors will be generated through the following code.

```
# Use ggplot to plot pdays vs previous
ggplot(bank,aes(x=pdays ,y=previous)) + geom_point() +
labs(title = "correlations of customer detail between previous and pdyas")
```

After the code above is generated, the correlation plot will be performed as shown by figure 6.

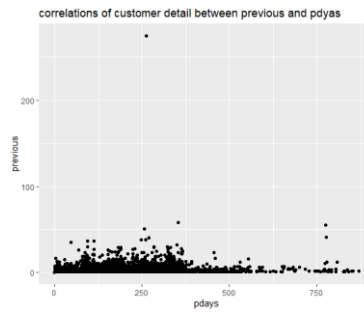


Figure 9. The highest correlation predictors plot in customer dataset.

Furthermore, as discussed before, the dataset will be classified into the train and test. The portion of the classification is 70% for train dataset and 30% for test dataset. The detail proportion of each dataset is shown by figure 10. The following code shows how the dataset is classified into those two datasets.

```
#Divide the data in training and test sets
set.seed(1)
train = sample(1:nrow(bank),0.7*nrow(bank))
bankTest = bank$y[-train]

length(train)
length(bankTest)

#assigning train dataset
bankDataTrain = bank[train, ]
head(bankDataTrain)
dim(bankDataTrain)

#assigning test dataset
bankDataTest = bank[-train, ]
head(bankDataTest)
dim(bankDataTest)
```

```
Console Terminal x
D:/George Mason University/Semester 2/Analytics Big Data to Information/Data Analytics Research Project/Dataset/ >
> dim(bankDataTrain)
[1] 31647 17
> dim(bankDataTest)
[1] 13564 17
>
```

Figure 10. The proportion of test and train dataset.

c. Logistic Regression Model

Since the data has already classified into train and test dataset, also the response variable has already converted to binary, the dataset is ready to be used

for the modelling. The following code shows how the logistic regression model is implemented to the train dataset.

```
glm.fit=glm(y~., data=bankDataTrain, family=binomial)
```

The summary of the model is shown through the following figure.

```

Console Terminal
D:\George Mason University\Semester 2\Analytics Big Data to Information/Data Analytics Research Project\Dataset\
Call:
glm(formula = y ~ ., family = binomial, data = bankDataTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6961  -0.3751  -0.2541  -0.1519   3.2000

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.511e+00  2.179e-01 -11.525 < 2e-16 ***
age          -8.768e-04  2.628e-03  -0.334  0.73865
jobblue-collar -3.907e-01  8.743e-02  -4.469 7.87e-06 ***
jobentrepreneur -2.592e-01  1.457e-01  -1.779 0.07526 .
jobhousemaid  -3.982e-01  1.579e-01  -2.522 0.01167 *
jobmanagement -1.586e-01  8.775e-02  -1.808 0.07066 .
jobretired     1.868e-01  1.155e-01  1.617 0.10584
jobself-employed -2.607e-01  1.316e-01  -1.981 0.04756 *
jobservices    -1.885e-01  9.914e-02  -1.902 0.05722 .
jobstudent     4.273e-01  1.305e-01  3.275 0.00106 **
jobtechnician  -1.739e-01  8.216e-02  -2.116 0.03434 *
jobunemployed  -2.579e-01  1.375e-01  -1.875 0.06074 .
jobunknown     -4.779e-01  2.896e-01  -1.650 0.09896 .
maritalmarried -2.233e-01  6.927e-02  -3.224 0.00126 **
maritalsingle  1.718e-02  7.941e-02  0.216 0.82876
educationsecondary 1.833e-01  7.793e-02  2.352 0.01865 *
educationtertiary 3.715e-01  9.044e-02  4.108 3.99e-05 ***
educationunknown 1.696e-01  1.254e-01  1.353 0.17612
defaultyes     -7.739e-03  1.908e-01  -0.041 0.96765
balance        1.148e-05  6.631e-06  1.731 0.08345
housingyes     -6.987e-01  5.227e-02 -13.367 < 2e-16 ***
loanyes        -5.009e-01  7.238e-02  -6.921 4.49e-12 ***
contacttelephone -1.049e-01  8.916e-02  -1.176 0.23950
contactunknown -1.578e+00  8.648e-02 -18.246 < 2e-16 ***
day           1.191e-02  2.968e-03  4.013 5.98e-05 ***
monthaug       -7.161e-01  9.469e-02  -7.562 3.97e-14 ***
monthdec       4.637e-01  2.161e-01  2.146 0.03186 *
monthfeb      -3.481e-02  1.075e-01  -0.324 0.74602
monthjan      -1.229e+00  1.475e-01  -8.333 < 2e-16 ***
monthjul      -7.362e-01  9.135e-02  -8.059 7.69e-16 ***
monthjun      5.152e-01  1.117e-01  4.614 3.94e-06 ***
monthmar      1.569e+00  1.431e-01  10.970 < 2e-16 ***
monthmay      -3.389e-01  8.632e-02  -3.927 8.62e-05 ***
monthnov      -8.333e-01  1.014e-01  -8.414 < 2e-16 ***
monthoct      9.077e-01  1.262e-01  7.194 6.27e-13 ***
monthsep      9.487e-01  1.421e-01  6.678 2.43e-11 ***
duration       4.153e-03  7.651e-05  54.284 < 2e-16 ***
campaign      -9.050e-02  1.196e-02  -7.567 3.81e-14 ***
pdays        2.501e-04  3.579e-04  0.699 0.48458
previous       6.616e-03  6.411e-03  1.032 0.30215
poutcomeother  1.057e-01  1.085e-01  0.973 0.33032
poutcomesuccess 2.300e+00  9.837e-02  23.379 < 2e-16 ***
poutcomeunknown -5.939e-02  1.097e-01  -0.541 0.58840
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22889  on 31646  degrees of freedom
Residual deviance: 15175  on 31604  degrees of freedom
AIC: 15261

```

Figure 11. The summary of the logistic regression model

According to figure 11, there are several predictors which are converted to a factor such as job, marital, education, and month. Those are because such predictors were stored as characters. Also, for modelling purposes, data which contains characters must be represented by using indicator variables, and that is how factors are treated through this modelling.

Afterwards, the model, which is stored inside the glm.fit, will be used to predict the probability of all the outcome responses in the test dataset. The following code and figure show how the model is used to predict the dataset and the result of prediction.

```
glm.probs=predict(glm.fit, bankDataTest, type="response")
```

```

D:/George Mason University/Semester 2/Analytics Big Data to Information/Data Analytics Research Project/Dataset/
> glm.probs
2      3      13      14      16      19      23      25      28      32      34
0.0099315076 0.0032333669 0.0345317862 0.0054846031 0.0212440061 0.0121699694 0.0042553566 0.0064422424 0.0039323613 0.0132367314 0.0084808062
37      45      49      54      55      59      67      72      73      75      80
0.0154522238 0.0706981394 0.0199748552 0.1435194646 0.0059887319 0.0091214815 0.1148073584 0.0051705062 0.0048609364 0.0090562304 0.0119511247
81      83      85      88      89      92      96      97      99      101      102
0.0140271982 0.0058423499 0.0077818856 0.5782771449 0.0064087651 0.0056929688 0.0181612050 0.0218501572 0.0097319083 0.0096866994 0.0093468127
104      108      110      111      112      113      116      119      129      134      138
0.0029181918 0.0066818925 0.1387075358 0.0116334514 0.0188769938 0.0262006648 0.0061718006 0.0053133797 0.0076465034 0.0094320212 0.0139929980
141      142      143      146      154      156      159      160      165      170      172
0.0068652235 0.0169979193 0.0111285394 0.0055968928 0.0060768316 0.0060650810 0.0041533889 0.0052780741 0.0127601127 0.0122480503 0.0071620127
174      178      180      188      189      190      198      203      206      215      216
0.0225883581 0.0041841822 0.0779962768 0.0157265334 0.0109767549 0.0641225799 0.0114648839 0.0241441419 0.0192133305 0.0091632012 0.0102005668
227      228      229      232      237      242      247      248      249      252      257
0.0084512806 0.0073629745 0.0165441184 0.0739416993 0.0119392758 0.0658867433 0.0083743765 0.0060674714 0.0109243315 0.0043037342 0.0054174686
266      267      268      273      275      280      282      283      290      291      292
0.0155426078 0.0123485938 0.0675169981 0.0033380837 0.0067177147 0.0185545574 0.0034799998 0.0273058023 0.0069982207 0.0070806386 0.0948481673
295      298      299      305      312      313      314      315      316      320      323
0.0143556547 0.0207713491 0.0049544602 0.0087851576 0.0059128654 0.0085514937 0.0070104026 0.0241466619 0.0197287417 0.0152633186 0.0089494854
325      327      332      333      336      339      340      345      348      350      351
0.0131299728 0.0093823757 0.0175735293 0.0071324913 0.0044884514 0.0530346720 0.0116793214 0.0046472196 0.0417498933 0.0087774966 0.0095002599
352      354      355      357      359      363      364      367      369      372      374
0.0815210926 0.0069605861 0.0261241175 0.0172990327 0.0063345107 0.0051003774 0.0133376656 0.0096747973 0.0186860270 0.0068984790 0.0210842548
378      379      384      385      390      392      393      397      399      402      403
0.0075374102 0.0172820550 0.0584440231 0.0179871754 0.0118144643 0.0105110600 0.0102578174 0.0166612684 0.0095781071 0.0147736627 0.0072149941
406      410      412      414      415      423      425      429      430      431      432
0.0388913286 0.0101305991 0.0444976824 0.0096736543 0.0087697247 0.0261418051 0.0116961311 0.0098228260 0.0076976967 0.0995208078 0.1133881572
434      435      443      443      444      446      449      457      467      468      472
0.0046787989 0.0071820991 0.0092959257 0.0112780097 0.0157910679 0.0109910294 0.0097551403 0.0078732572 0.0138703611 0.0453478866 0.0522693869
475      478      480      482      483      484      488      491      492      494      499
0.9166592680 0.0077911093 0.0120315710 0.0078195103 0.0060576326 0.0147927227 0.0115451678 0.0081048759 0.0084183996 0.0307980441 0.0049392186
507      509      512      519      524      529      530      535      545      548      551
0.0639186654 0.0164946664 0.0111998878 0.0042819641 0.0159107424 0.0305575159 0.0139096496 0.0140458567 0.0059436671 0.0095973389 0.0048095831
553      556      567      572      573      577      578      585      589      590      593
0.0385907427 0.0079306168 0.0264988260 0.0199623334 0.0047709372 0.0103254961 0.0059704323 0.0037167913 0.0247744499 0.0025170754 0.0089778578

```

Figure 12. The prediction result of the train dataset response.

The predictions' result contains 13564 probabilities since such number is total row of the test dataset.

4. Results

As discussed through the previous section, the number of probabilities which are generated for the test dataset is 13564. Furthermore, confusion matrix should be performed first. The following code is the step of how the confusion matrix was built in this research.

```

# Initialize bankDataTest prediction to "0" which means no
glm.pred=rep(0, 13564)

# Change prediction to "1" which means "yes" if probabability > 0.5
glm.pred[glm.probs>.5]=1

# Display confusion matrix
table(glm.pred, bankDataTest$y)

```

According to code above, firstly, all the responses are assumed to be 0. In other words, the customer responses are considered to “no”. After that, according to the probabilities result, the one that have probabilities above 0,5 will be replaced by 1, which means a “yes” response. Finally, the confusion matrix can be performed, and the proportion of correctness and incorrectness of the model can be obtained. Hence, the conclusion of whether the model works well for the dataset can be seen. The following figure shows the result confusion matrix result.

```

D:/George Mason University/Semester 2/Analytics Big Data to Information/Data Analytics Research Project/Dataset/
> table(glm.pred, bankDataTest$y)
glm.pred      0      1
0      11692    1035
1       297      540

```

Figure 13. The confusion matrix of the logistic regression model.

Based on figure 13, the calculation of the correctness and the incorrectness of the model towards the dataset can be performed afterwards.

Specifically, through the confusion matrix as shown in figure 13, there are 4 components that compose the matrix. Each component is such the follows:

- **True Positive**

This component represents the number of matches between the prediction result and actual result inside the test dataset [14]. Specifically, if the customer is predicted to say yes, which means the prediction of the customer that wanted to subscribe the product are match to the actual response in the test dataset, that will be counted as the true positive. The total number of true positive in this research according to the figure 13 is 540.

- **False Positive**

This component represents the same thing as true positives. However, the values which are counted into this component is only the when the prediction is no and if it matches the actual value in the dataset, then it will be counted as False Positive. In this research, the number of the False Positive is 11692 as shown in figure 13.

- **True Negative**

This component will contribute the incorrect proportion later [14]. True Negative component contains the total number of mismatch prediction between the prediction result and the actual response [14]. Specifically, when predictions' result says yes, the actual value is no. In other words, when the customer is predicted wanting to subscribe the product, the actual value in the test dataset is vice versa. The number of True Negative result in this research according to figure 13 is 1035.

- **False Negative**

This component is almost the same such as the true negative since it also contributes the incorrect proportion of the model [14]. In particular, the number of False Negatives represents the count of case where the customer is predicted to say no; however, the actual value in the test dataset says yes. According to figure 13, the number of false negatives of this research is 297

According to the all components of the confusion matrix, it can be concluded that the proportion of the correctness and the incorrectness of the results by implementing the model to the test dataset is such the follows:

- ✓ The proportion of the correctness: **90%** out of 100%

$$(540 + 11692) / 13564 = 0.9017989$$

- ✓ The proportion of the incorrectness: **9.82%** out of 100%

$$(297 + 1035) / 13564 = 0.09820112$$

5. Discussion

According to the test result of in this research, it shows that the accuracy of the model is 90%. However, the proportion of the incorrect result is still existing. Therefore, some measurement still needed to be done for the further study.

Probably, by including the other predictors into the model, the accuracy of the model could be increased. Specifically, the dataset does not contain any information such as the salary of the customer. Such information could probably increase the accuracy of the model since the fix income of the customer could determine whether the customer need the banks' product or vice versa [21].

References

- [1] A. Barone, "Bank," Investopedia, 21 April 2020. [Online]. Available: <https://www.investopedia.com/terms/b/bank.asp>. [Accessed 7 May 2020].
- [2] R. Lake, "How Often Should You Monitor Your Checking Account?," Investopedia, 4 March 2020. [Online]. Available: <https://www.investopedia.com/how-often-should-you-monitor-your-checking-account-4798537>. [Accessed 7 May 2020].
- [3] Aspiring Minds Team, "Sales Manager in Banking Service Fresher entry Level," Aspiring Minds, 17 April 2019. [Online]. Available: <https://www.aspiringminds.com/hr-insights/featured-profiles/sales-manager-in-banking-service-fresher-entry-level>. [Accessed 7 May 2020].
- [4] D. Goodman, "The 5 Most Effective Marketing Strategies for Financial Services," everfi, 2020. [Online]. Available: <https://everfi.com/insights/blog/5-effective-financial-services-marketing-strategies/>. [Accessed 7 May 2020].
- [5] UCI Machine Learning Repository, "Bank Marketing Data Set," UCI, 2014. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>. [Accessed 7 May 2020].
- [6] M. Rouse, "data set," TechTarget, March 2016. [Online]. Available: <https://whatis.techtarget.com/definition/data-set>. [Accessed 7 May 2020].
- [7] N. Sun, J. G. Morris, J. Xu, X. Zhu and M. Xie, "iCARE: A framework for big data-based banking customer analytics," *IBM Journal of Research and Development*, vol. 58, no. 5/6, pp. 4:1 - 4:9, 2014.
- [8] C. L. Mbama and P. O. Ezepue, "Digital banking, customer experience and bank financial performance: UK customers' perceptions," *International Journal of Bank Marketing*, vol. 36, no. 2, pp. 230-255, 2018.
- [9] S. S. Skiena, *The Data Science Design Manual*, Stony Brook, New York: Springer, 2017.
- [10] B. Serener, "Statistical Analysis of Internet Banking Usage with Logistic Regression," *Procedia Computer Science*, vol. 102, pp. 648-653, 2016.
- [11] G. Dong, K. K. Lai and J. Yen, "Credit scorecard based on logistic regression with random coefficients," *Procedia Computer Science*, vol. 1, no. 1, pp. 2463-2468, 2010.
- [12] J. P. Lander, *R for Everyone*, Addison-Wesley, 2017.
- [13] A. Estrella-Ramon, "Explaining customers' financial service choice with loyalty and cross-buying behaviour," *The Journal of Service Marketing*, vol. 31, no. 6, pp. 539-555, 2017.

- [14] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning with Application in R, New York: Springer, 2013.
- [15] Learn.org, "What Is Primary Education?," Learn.org, 2003-2020. [Online]. Available: https://learn.org/articles/What_is_Primary_Education.html. [Accessed 7 May 2020].
- [16] A. Corsi-Bunker, "GUIDE TO THE EDUCATION SYSTEM IN THE UNITED STATES," [Online]. Available: <https://issn.umn.edu/publications/USEducation/4.pdf>. [Accessed 7 May 2020].
- [17] Learn.org, "What Is Tertiary Education?," Learn.org, 2003 - 2020. [Online]. Available: https://learn.org/articles/What_is_Tertiary_Education.html. [Accessed 7 May 2020].
- [18] J. Song, "Loan vs. Line of Credit: What's the Difference?," valuepenguin, 24 September 2019. [Online]. Available: <https://www.valuepenguin.com/loans/loan-vs-line-of-credit>. [Accessed 7 May 2020].
- [19] timesofindia, "What is a house loan?," timesofindia, 21 December 2018. [Online]. Available: <https://timesofindia.indiatimes.com/business/faqs/home-loan-faqs/what-is-a-house-loan/articleshow/60479745.cms>. [Accessed 7 May 2020].
- [20] J. Brozic, "6 things you should know about personal loans," creditkarma, 7 July 2019. [Online]. Available: <https://www.creditkarma.com/personal-loans/i/what-you-should-know-about-personal-loans/>. [Accessed 7 May 2020].
- [21] H. WAGNER, "Analyzing a bank's financial statements," investopedia, 3 May 2020. [Online]. Available: <https://www.investopedia.com/articles/stocks/07/bankfinancials.asp>. [Accessed 9 April 2020].