STAT 828: DATA MINING – DIRECTED KNOWLEDGE DISCOVERY PROJECT

GROUP NAME: BAROONA

EXECUTIVE SUMMARY

Finance and banking are one of the most extensive and extremely competitive markets. Any organisation going face-to-face with the big players need to fortify that they understand precisely how buyers like to interact with their sales and marketing processes. Customers today lean on both online as well as offline means to shop before making a decision. Almost on every occasion, an offline phenomenon such as making a phone call or visiting a branch is a positive indicator of a possible conversion. In this report, a bank marketing dataset of a Portuguese bank is selected where the marketing campaigns were based on phone calls. The report is based on the sample of the entire dataset which is about 50% of the original dataset. Initially, the dataset is split into training data, testing data and evaluation data. The outcome of the campaign is predicted using various classification and prediction models. The models used for predicting the result are Decision trees (C5.0, CART, QUEST, CHAID, RANDOM TREES), Neural Networks, Bayesian network and Support Vector Machines (SVM). Each model is presented briefly in this report and the best model is selected for analysis. C5.0 produced the best results among all models.

INTRODUCTION

Banks and financial institutions exist to offer financial services to people and to make huge profits. Having said that, banks also devote remarkable resources and business intellect to gain capital. One of the most common ways for banks to do this is to engage in direct marketing campaigns like phone calls and face-to-face meetings to promote and provide services. Phone calls, i.e. Telemarketing is a conventional marketing technique that helps to soar profits for any given business. Moreover, it also offers a more interactive and personal medium of sale service which can initiate an instant rapport with the prospective customers. Furthermore, telemarketing can help an organization to reach out more customers than with in-person or by going door-to-door and it can benefit a company to sell a product to both existing and new customers. For banking industry, telemarketing can be useful to communicate with large number of customers and offer them with all the services that they have for them. This may include, information about loans, term deposits, mortgages, Overdraft facility, Credit cards etc.

For this project, a data set of a Portuguese Bank direct marketing campaign is used. This dataset is obtained from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The primary objective of this project is to find the best model to predict whether a customer will subscribe for a term deposit or not using various classification techniques. Our secondary objective is to determine what factors in this data set would contribute the most for the sale of term deposits to the potential customers. The target users for this project are the marketing team of a banking institution who are looking to increase their inflow of cash deposits. The following sections of this report includes the description of the dataset in detail, all the methods (classification techniques) that has been applied on the dataset to get the results and eventually the best one is described thoroughly. Moreover, the results for the best model is presented followed by conclusion which summarises the most important findings and the scope of future research is suggested.

DESCRIPTION OF THE DATASET

Original Data:

The original data has been extracted from the UCI Machine Learning Repository. The data is a result of a direct marketing campaign executed by a Portuguese banking institution to promote term deposits. The campaign was based on phone calls. It contains a total of 45211 instances. There are 17 attributes in total, out of which 16 are independent variables and 1 is dependent variable (outcome variable). The outcome variable is binary (yes/no), where yes means a customer will subscribe for a term deposit or no otherwise. The description of all the attributes is given in the table below.

Sr.#	ATTRIBUTES	TYPE	DESCRIPTION	VALUES
			CLIENT DATA	
1	age	Numeric	Age of the Client	Positive Integer
2	job	Categorical	Type of Job	admin., unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services
3	marital	Categorical	Marital status of the client	married, divorced, single
4	education	Categorical	Education level of the client	unknown, secondary, primary, tertiary
5	default	Binary	Whether a client has credit in default or not	yes, no
6	balance	Numeric	Average yearly balance in Euros	Integer
7	housing	Binary	Whether a client has housing loan or not	yes, no
8	loan	Binary	Whether a client has a personal loan or not	yes, no
			CONTACT RELATED DATA	
9	contact	Categorical	Contact communication type	unknown, telephone, cellular
10	day	Numeric	Last contact day of the month	Positive integer (1 – 31)
11	month	Categorical	Last contact month of the year	jan,feb, mar…,dec
12	duration	Numeric	Last contact duration in seconds	Positive integer

CAMPAIGN RELATED DATA				
13	campaign	Numeric	number of contacts performed during this campaign	Positive Integer
14	pdays	Numeric	number of days that passed by after the client was last contacted from a previous campaign	Integer (-1 means client was not contacted previously)
15	previous	Numeric	number of contacts performed before this campaign and for this client	Integer
16	poutcome	Categorical	outcome of the previous marketing campaign	unknown, other, failure, success
17	y	Binary	Whether a client has subscribed a term deposit or not	yes, no

Table 1: Description of the original data

Data Pre-processing:

The bank marketing data set that we have used is almost a clean dataset and does not contain any missing values. However, there are some changes such as renaming of variables, identifying the outliers etc that have been made in few attributes as mentioned below: (Table 13 contains the summary)

- Firstly, we have selected a sample data from the original data which contains 22606 instances (50% of the original data). This has been done to get accurate models with minimum misclassification.
- Dataset observations – 22606, predictors – 16, outcome variable - 1
- There are no missing values in the dataset.
- The variable "default" has been renamed as "default_credit" which means whether the client has any credit in default or not. The renaming has been done to make the variable more specific.
- The variable "y" (outcome variable) which indicates whether a client has subscribed a term deposit or not has been renamed as "term_deposit_susbscribed". The renaming variable gives a meaning to this variable which makes the outcome easy to understand.
- **Age:** The variable "age" had 180 outliers(Age >= 74 years) [min = 18, max = 95] , so the data was right-skewed. We couldn't remove the outliers as they add meaning to the data, and they showed nature of the age of people(right-skewed). Just to have better understanding of the distribution of the age, it has been converted to categorical type. Three distinct age groups are created which are: Youth (age between 18 and 35), Workforce (between 36 and 59) and Retired (above 60). (Appendix 5,6,7,8)
- **Balance:** This variable had 2400 outliers with some outliers (balance < -1884 euros & balance > 3412 euros) under debt. The maximum account balance was recorded as 98417 euros and maximum debt was 8019 euros, which are plausible amounts therefore no imputation was carried out. (Appendix 17 &18)
- **Duration:** This variable had 1600 outliers (duration >= 646 seconds) with minimum and maximum call duration of 2 and 3881 seconds. As, this was also reliable measure so here also no imputation was done. (Appendix 27& 28)

- **Campaign:** This variable is right-skewed and has around 1500 outliers (campaign >= 7)[min = 1, max = 58]. In order to maximize the Clients for subscription of term deposit 58 is quite reliable count so no change was made in this variable as well. (Appendix 33, 34 & 35)
- **Pdays:** This variable was also right skewed and had around 4000 outliers(pdays >= 1)[min = -1, max = 854]. Most values were -1 which means client was not previously contacted. Due to very high frequency of -1 the other values were treated as outliers. So, we didn't remove them. (Appendix 31&32)
- **Previous:** This variable was also right skewed and had around 4000 outliers(previous >= 1)[min = 0, max = 275]. Most values were 0 which means client was not contacted before campaign. Due to very high frequency of 0 the other values were treated as outliers. So, we didn't remove them.
- **Poutcome:** The observation "other" was merged with "unknown" as they belong to the same hierarchy according to the dataset and also there were total 900 records for 'others'. And also, they both don't convey any meaning. (Appendix 34,35,36)

Neither new variables nor any concept hierarchies were created. Neither any observations nor any variables were excluded from the analysis. Dataset after data preprocessing observation – 22606, predictors – 16, outcome variable – 1.

Descriptive Statistics:

- The bank should carry out their next marketing campaign targeting clients from the workforce as they have 47% chance of subscribing the term deposit. Followed by youth with 41% chance of subscribing. Bank should also come up with some strategy targeting the clients who are retired as they have 11% chance of subscribing the term deposit. [Appendix 8]
- Clients who are retired, who are at management position or who are technician have high chances of subscribing the term deposit. [Appendix 10]
- Clients with the zero or negative balance have no money to subscribe to term deposit. While clients with positive average balance or higher balance have money to spend and are more likely to subscribe the term deposit.
- Married clients have 52% chance of subscribing, followed by single clients with 36% chance of subscribing the term deposit. [Appendix 12]
- Clients who have completed secondary and tertiary education have respectively 46% and 38% chances of subscribing the term deposit. [Appendix 14]
- Client who don't have any housing loan and personal loan don't have financial burden, so they have respectively 61% and 91% chances of subscribing the term deposit. [Appendix 20]
- 82% of the clients who were contacted to their cellular did subscribe the term deposit. [Appendix 22]
- Clients who were contacted between 8th and 21st day of the month have 95% chances of subscribing the term deposit.
- Highest marketing activity was carried out in May and August and least carried out in December and January. Bank should consider marketing in December and January. [Appendix 26]
- The longer the duration of the contact higher the chances of clients to subscribe the term deposit. Same goes for number of times client were contacted during the campaign.

3. METHODOLOGY

This section describes numerous modelling methods and models to predict the outcome of the data. Eight models have been applied and explored briefly and the best model is illustrated in detail. Various factors which made an impact on the outcome are also analysed and explained.

3.1 Classification Analysis

3.1.1 Data Splitting (Training, Testing and Evaluation)

The original dataset was split randomly into three subsets, Training (60%), Evaluation (20%) and Testing (20%) to test the robustness of each model and get good results. The data split has been done in R. The information about the data after the split has been mentioned in the table below:

Target Value	Train Segment		Evaluation Segment		Test Segment		Sample Dataset
	Count	Percentage	Count	Percentage	Count	Percentage	Count
yes	1623	12.02	539	11.83	520	11.42	2682
no	11875	87.98	4016	88.17	4033	88.58	19924
Total	13498		4555		4553		22606

Table 2: Data Fragmentation

3.1.2 Models

The models and the methods used are explained in the sections below. They were all implemented on the training data, then the few of the best models were then evaluated and tested by running the models on the respective data subsets. For a complete summary, refer the appendix.

3.1.2.1 Support Vector Machines (SVM)

The goal of SVM is to find a hyperplane in N-dimensional space that classify the observations distinctively. To isolate categories, there may be many numbers of hyperplanes possible. But one with the maximum margin is chosen as it classifies the unknown records accurately. SVM uses kernel functions to transform the observations to higher dimension. There is also a cost parameter which determines how wide a margin is.

We used different kernels like Linear, RBF (Radial basis Function), sigmoid and Polynomial. Values for the regularization parameter (known as C/COST) were 10, 100, 1000. The higher the value of C, more overfitted model we get. For linear kernel results were consistent, for RBF kernel we found fluctuations over various value of RBF gamma [0.1, 0.01, 0.001]. Stopping criteria determines when to stop the optimization algorithm. Values which were considered were 1.0E–3, 1.0E-4, and 1.0E–5. Lower the value -> more accurate the model -> Longer it takes to train a model.

The best model was EXPERT MODE, STOPPING CRITERIA = 1.0E-3, REGULARIZATION PARAMETER(C) = 100, KERNEL TYPE = LINEAR

3.1.2.2 Naïve Bayes Classifier

Naïve Bayes uses Bayes theorem, which is based on conditional probability and uses the formula P(A | B) = P(A) * P(B | A) / P(B). We are carrying out this model in r using ‘e1071’ package. 50% of the total observations were randomly selected then divide into train[60%], evaluation[20%] and Test[20%] dataset.There is one problem with naïve bayes, if it encounters a new observation totally odd from training data[occurrence of it in training data = 0], then it will assign probability 0 to it, which is not appropriate. That’s

why Laplace smoothing is used where one small constant is added just to make sure occurrence of total odd event would not be zero but it will be equal to that constant.

bank_train_bayes <- naiveBayes(term_deposit_subscribed ~ ., laplace = 2, data = bank_notscaled_train

The above line was used to carry out the modelling for training data. The laplace parameters was changed[1, 2, 3] just to see whether there are any totally odd events in testing of evaluation data or not. But the results were same for all the three values. The Naïve Baysian Model with laplace = 2 was chosen. This classifier assumes normality for the continuous variable which is not holding here.

3.1.2.3 C5.0

The C5.0 is based on the concept of decision tree. This algorithm splits the data samples into two or more subsets so that the samples within each subset are more homogeneous than in the previous subset. This is a iterative process and the resulting two subsets are then split again, and the process repeats until the homogeneity measure is reached or until some other stopping, condition is satisfied.

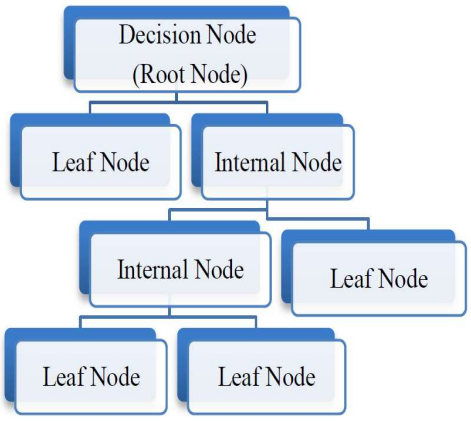


Fig1: Visual Representation of C5.0

This model is implemented in IBM SPSS Modeler on the training data. Pruning has not been done because it did not give good results and misclassification was more. A simple mode was executed. However, the misclassification ratio was set to 9:1.

3.1.2.4 Logistic Regression

The logistic regression model is appropriate to deal with issues of many types of data sets. It is provided sufficiently several and well-distributed samples. Moreover, it is suitable for explaining and testing hypotheses regarding the relationships between a categorical outcome variable and one or more categories of continuous predictor attributes. The logistic formulas are stated in terms of the probability that Y = 1 (or yes), which is referred to as P. The probability that Y is 0 (or no).

$$\ln\left(\frac{P}{1-P}\right) = w_{\circ} + w_1 A$$

3.2.1.5 Neural Networks:

The Neural Network is famous for its strong approximation of a function for classification as well as prediction situations. The dataset consisted of a combination of numeric and categorical variables. Hence, this model was utilized. The construction and linking procedure is similar to the brain. It is available as a node in SPSS Modeler. While building the model, overfit prevention was set to 30% and Multilayer Perceptron technique

was applied as it doesn't compromise on the model accuracy. Also, stopping criteria was set to 15 minutes. It generated very accurate results, but class specific error rates and the misclassification errors were unacceptable.

3.2.1.6 C&R Tree:

The Classification and Regression Tree Node in IBM SPSS constructs a decision tree with an impurity reduction technique (Gini index). The algorithm's performance is the same for continuous and symbolic input variables. Also, it can ignore less important predictor fields automatically. Misclassification costs in the ratio of 0.88 : 0.1 was applied to add weightage to the minority outcome. The model generated moderately accurate results but C5.0's was the best.

3.2.1.7 CHAID:

This model is quite useful in direct marketing field to select group of customers and their responses to given variables are used to predict other variables. This method is quite efficient for large sample sizes as it generates non-binary tree which is quite easy to interpret. The interaction among variables is predicted by comparing Bonferroni adjusted p-values which are computed using chi-square test. This node is in IBM SPSS Modeler which works quite effectively for both categorical and continuous inputs. To build the model, misclassification cost was set 1:1, significance level was 0.05 for both splitting and merging classes and in stopping rules minimum records in parent and child branch were set to 2% and 1% respectively.

3.2.1.8 QUEST:

The QUEST classifier works similar to CHAID with few differences as it generates binary tree and it doesn't compute tests for merging classes due to which this algorithm works much faster than C&R tree or CHAID which takes longer to compute the results. This method is highly effective for large datasets. It takes up only categorical variable as an outcome. The node in IBM SPSS Modeler was used with parameters similar to CHAID. (refer Appendix)

Model comparison using Lift Chart:

Client subscription to a term deposit is visualised in the form of an ensemble lift chart as shown below. The diagram on the bottom left illustrates the performance of various models when the outcome is no. Neural Network algorithm is slightly higher than the others. And, the lift chart on the right indicates the predictive performance of all the models when the client subscribes to the deposit. Even in this case, Neural network model produces the best lift. With the help of these charts, we can single out Neural networks but various other parameters need to be considered to build the perfect model.

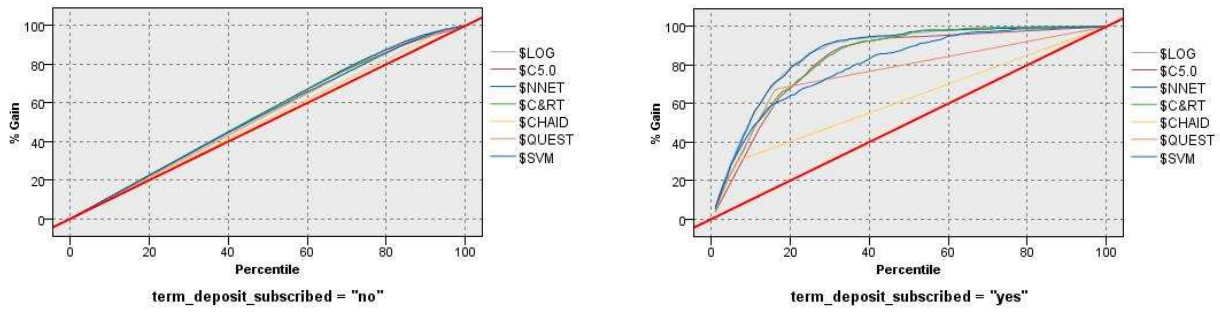


Fig2: Lift Charts of various models

The following table describes the class specific error rates of all the models built and their respective lift across all the three partitioned segments of the data.

	Train Dataset					Evaluation Dataset					Test Dataset				
Algorithm	Ac (y)	Ac(n)	Ac(o)	MCE%	Lift(y)	Ac (y)	Ac(n)	Ac(o)	MCE%	Lift(y)	Ac (y)	Ac(n)	Ac(o)	Lift(y)	MCE%
C5.0	89.71 %	73.61 %	74.50 %	10.29 %	2.63	88.86 %	73.78 %	74.50 %	11.13 %	2.64	86.73 %	72.99 %	74.50 %	2.56	13.26 %
Neural Net	45.23 %	95.81 %	89.72 %	54.78 %	4.95	45.08 %	95.34 %	89.40 %	54.92 %	4.78	40.96 %	95.49 %	89.26 %	4.72	59.04 %
C&R Tree	81.21 %	76.96 %	77.47 %	18.79 %	2.7	83.30 %	77.19 %	77.91 %	16.70 %	2.78	79.23 %	77.34 %	77.55 %	2.72	20.77 %
Naïve Baye's	92.25 %	53.48 %	88.93 %	46.52 %	4.24	93.17 %	53.32 %	88.93 %	46.68 %	4.24	93.29 %	56.02 %	88.93 %	4.49	43.98 %
SVM	92.25 %	53.11 %	87.96 %	46.89 %	4.05	92.25 %	54.80 %	87.96 %	45.20 %	4.11	92.47 %	53.95 %	87.96 %	4.16	46.05 %
CHAID	32.83 %	96.77 %	88.75 %	42.34 %	4.89	29.52 %	96.54 %	88.75 %	46.49 %	4.50	31.02 %	96.39 %	88.75 %	4.56	46.77 %
QUEST	27.62 %	96.90 %	88.95 %	45.61 %	1.03	25.83 %	96.41 %	88.95 %	50.70 %	1.02	30.08 %	96.47 %	88.95 %	1.03	45.02 %

Table 3: Model Characteristics

Elaborate analysis was carried out with the help of each and every model. While some models overfit the dataset, others had varying performances over the other two partitioned datasets. The highest overall accuracy was observed in the Neural Network model but the class specific accuracy of the value "yes" turned out to be a very small value (45.2%) and misclassification error was pretty high (54.78%) to adjudge it as a good model. Among all the other algorithms, C5.0 Decision Tree produced the best results in terms of minimizing the miscalculation error, its lift value is second to the neural networks and is easily interpretable. (Table #).

4. RESULTS

The best model is C5.0 with 30% pruning severity, 500 minimum nodes in child branch, with global pruning, with cost ratio of 9:1, with overall accuracy of 74.5%, with 73% accuracy and lift 1.11 for class 'no' and with 89% accuracy and lift 2.6 for class 'yes'. We wanted to correctly identify 'yes' labels so we compromised the overall accuracy and class 'no' accuracy. The accuracies and lift were consistent throughout all the dataset so we can rely on the model. We can see results in Fig *4.* [whole tree in Fig C5.0]

Many predictors were given as an input but we can see that the 3 most important of them are duration, poutcome and contact in Fig *3*. We can also see most important rules in the table below[For all rules Table *4*].

#Rule	Rule statement	Classified as
1	duration <= 208.5 and poutcome = 'failure'	no
2	duration <= 208.5 and poutcome = 'success'	yes
3	duration <= 208.5 and duration > 521.5	yes
4	duration <= 208.5 and duration <= 521.5 and contact = 'telephone'	yes
5	duration <= 208.5 and duration <= 521.5 and contact = 'unknown'	no
6	duration <= 208.5 and duration <= 521.5 and contact = 'cellular' and poutcome = ['failure', 'success']	yes

Table 4: Results of the C5.0 Tree

First rule is obvious as duration and poutcome were the most important predictors. Also, if duration of the call is shorter, it seems that client is not interested in the campaign details and s/he is not looking forward to subscribing the term deposit. Same goes with the result of previous campaign(poutcome), if a client didn't invest in the previous campaign s/he is more likely not to subscribe the term deposit. So that's why those observations are labelled as 'no'. Second rule is the exact opposite of the former one and we can see that it is

classified as ‘yes’. Third rule states the same that higher the duration higher the interest of the client in the campaign and higher the chances of the client subscribing the term deposit. Now the third most important predictor comes into the picture which is contact. Forth rule indicates that if communication medium is telephone and duration is high then those observation can be classified as ‘yes’. The clients are using telephone that suggests that either clients are retired or housewives or someone who stays at home. As these clients have money and easy to convince, so they are labelled as ‘yes’. Fifth rule same as forth rule, different in the terms of only communication medium which is ‘unknown’. Unknown can mean many things that client was not contacted or contacted through mediums like word of mouth by bank employee, other platforms like bank’s mobile application or emails. They are labelled as ‘no’. Sixth rule is the combination of all the above rules except the communication medium is ‘cellular’ and poutcome is either ‘failure’ or ‘success’ and they are labelled as ‘yes’. These rules show that duration, poutcome and the contact is the most associated with the outcome whether client subscribes a term deposit or not.

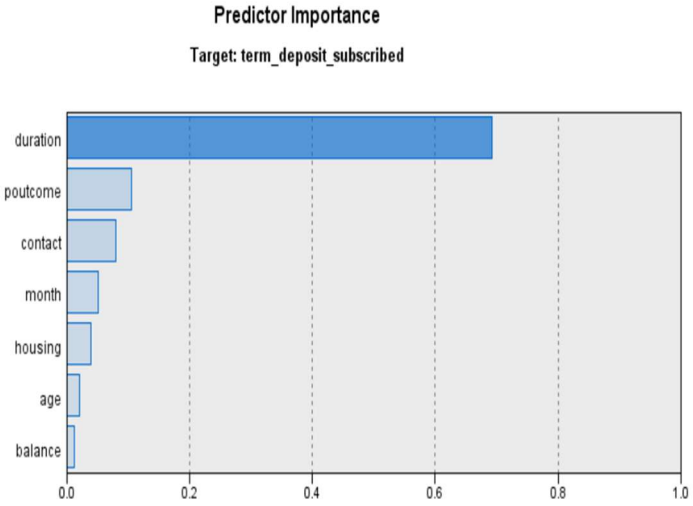


Fig4: Predictor Importance of C5.0 Tree

#Rule	Rule statement	Classified as
1	duration <= 208.5 and poutcome = ‘failure’	no
2	duration <= 208.5 and poutcome = ‘success’	yes
3	duration <= 208.5 and poutcome = ‘unknown’ and month = [‘apr’, ‘march’, ‘oct’, ‘sep’]	yes
4	duration <= 208.5 and poutcome = ‘unknown’ and month != [‘apr’, ‘march’, ‘oct’, ‘sep’]	no
5	duration <= 208.5 and duration > 521.5	yes
6	duration <= 208.5 and duration <= 521.5 and contact = ‘telephone’	yes
7	duration <= 208.5 and duration <= 521.5 and contact = ‘unknown’	no
8	duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = [‘failure’, ‘success’]	yes
9	duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘no’ and balance <= 106.5	no
10	duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘no’ and balance > 106.5	yes
11	duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘yes’ and duration > 381.5	yes
12	duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘yes’ and duration <= 381.5 and age = ‘Retired’	yes

13	duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘yes’ and duration <= 381.5 and age = [‘Workforce’, ‘Youth’]	no

Table5: Rules of C5.0 Tree

Fig4: The C5.0 Tree with frequencies

5. CONCLUSION

This project is focused on data mining techniques through which an ideal model is produced with the goal of enhancing marketing approach for a bank.

The key findings from this project are:

- C5.0 is the most suitable model for this data. The best predictive factor to attract a client is the duration of the call. More the duration means it is very likely that the client will subscribe a term deposit.
- The outcome of the previous campaign creates a massive impact on the result. If the previous outcome is success, a client is 90% likely to opt for the term deposit again.
- The ideal months for contacting a client is during the mid-year (May to August). During these months, 71% of the people have subscribed for term deposit.

We believe that the bank campaign management can attract a potential client by initiating a friendly rapport with them and convince them to opt for term deposits by engaging them in conversation. Furthermore, the bank needs to focus during the mid-year by offering new schemes to the client. For future research, we suggest that latest data should be used to yield a better model, ideally, by adding several characteristics which are not associated to the contact execution such as demographics of the client (eg: place of residence, annual income etc).

REFERENCES

1. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
2. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems.
3. Witten, I. and Frank, E., "Data Mining – Pratical Machine Learning Tools and Techniques", 3rd edition, Elsevier, USA, 2005
4. https://medium.com/optima-blog/using-polar-coordinates-for-better-visualization-1d337b6c9dec
5. http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

6. APPENDIX

Variable	Type	Min	Max	Median	Mean	Std.Dev
age	Factor	Retired: 883 Youth: 8385 Workforce: 13338				
job	Factor	admin: 2538 blue-collar: 4831 entrepreneur: 743 housemaid: 634 student: 461 management: 4815 retired: 1149 self-employed: 770 services: 2102 technician: 3779 unemployed: 643 unknown: 141				
marital	Factor	married: 13615 single: 6390 divorced: 6390				
education	Factor	primary: 3430 secondary: 11541 tertiary: 6732 unknown: 903				
default_credit	Binary	no: 22198 yes: 408				
balance	Numeric	-8019	98417	442	1349	3013.346
housing	Binary	yes: 12588 no: 10018				
loan	Binary	yes: 3598 no: 19008				
contact	Factor	cellular: 14577 telephone: 1482 unknown: 6547				
day	Numeric	1	31	16	15.85	8.29
month	Factor	jan: 668 feb: 1293 mar: 246 apr: 1427 may: 6917 jun: 2691 jul: 3464 aug: 3121 sep: 288 oct: 351 nov: 2040 dec: 100				
duration	Numeric	2	3881	181	258.6	257.1873
campaign	Numeric	1	58	2	2.756	3.0656
pdays	Numeric	-1	854	-1	39.4	98.998
previous	Numeric	0	275	0	0.5768	2.599
poutcome	Factor	failure: 2389 success: 756 unknown: 19461				
term_deposit_subscribed	Factor	no: 19939 yes: 2667				

Table 13: Summary of all the variables

APPENDIX 1 (ID 45539510):

50%(22606 obs) of the total observations(45211 obs) were randomly selected then divided into train(13498 obs)[60%], evaluation(4555 obs)[20%] and Test(4553 obs)[20%] dataset. [obs - observations][16 input variables][1 response variable]

SUPPORT VECTOR MACHINE

Methodology: SVM(SUPPORT VECTOR MACHINE) is a robust regression and classification tool. SVM maximizes the estimation accuracy of a model avoiding the overfitting of the training data. SVM is suitable for analyzing a dataset with a very high number of independent variables(Predictor variables). SVM works by converting the data to high-dimensional feature space so that observations can be classified even observations are not otherwise linearly isolated. The line separating the categories is found then observations are transformed in such a way that the separating line could be represented or drawn as a hyperplane. The function which is used to transform the observations is known as the kernel function. There are many kernel functions available[Linear, Radial basis function, Sigmoid, Polynomial] and can be used to transform the data. Other than separating line between the different categories, a classification SVM also computes the marginal line the defined space between the two categories. The Observations lying on the margins are also known as support vectors. Wider the margin -> better the model will be at predicting the new observations. Narrow margin -> overfitted model.

(A)Why SVM?: Convex Optimization is one the characteristics of SVM, because of that Outcome is always a global minimum and not a local minimum(Guaranteed Optimality). It is used for both linearly and non-linearly separable data. Used in a dataset where some observations are labeled, and some are not. There are many kernel options available for feature mapping.

(B) and (C): The SVM was performed on IBM SPSS MODELER. No other data preprocessing was carried out especially for SVM. First default settings[simple] of the SVM node was used. Then for the experimental purposes and finding the best performance settings were altered. Kernels like Linear, RBF(Radial basis function), Sigmoid, and Polynomial were tried. As we didn't have a deeper knowledge of how sigmoid and polynomial works, so they were not carried out furthermore. We also tried to change stopping criteria[1.0E-3 -- 1.0E-6]. Stopping criteria 1.0E-6 was taking so much of computation time and at the end giving the same results as other values, so we decided not to take into consideration. The regularization parameter (known as C/COST) took values like 10, 100, and 1000. Cost even > 100 was taking too much time and was overfitting the training model. RBF gamma took values like 0.1,0.01,0.001[We can't go lower than this in IBM]. Regression epsilon setting was not considered as it is for regression purposes.**[Different values of parameters in table 1] [Further node setting table 10]**

(D) & (E): The results of all the models were then compares based on class accuracies[Table 1]. The best model had the setting Expert Mode, Stopping Criteria = 1.0E-3, Regularization Parameter(c) = 100, kernel type = Linear. There were other combinations of settings which produced the same result, but this was chosen because of faster results. The results from linear kernels seem consistent with different parameter[unlike RBF kernel], that's why linear kernel with highlighted parameters mentioned above was chosen.

NAÏVE BAYES CLASSIFICATION

(A)Why NAÏVE BAYES?: It is a common-sense(logic) based technique which is simplest among all the algorithms you would encounter in the field of data science, yet it is so powerful that sometimes it overcomes the other complex algorithms. This modelling technique is all about the probability. It is easy to calculate the probability of an event[which is equal to the number of cases following an event divided by a total number of cases]. Naïve Bayes uses Bayes theorem, which is based on conditional probability and uses the formula P(A | B) = P(A) * P(B | A) / P(B).

(B) and (C): We are carrying out this model in r using 'e1071' package. There is one problem with naïve Bayes, if it encounters a new observation totally odd from training data[occurrence of it in training data = 0], then it will assign probability 0 to it, which is not appropriate. That's why Laplace smoothing is used where one small constant is added just to make sure the occurrence of a total odd event would not be 0 but it will be equal to that constant.

```
bank_train_bayes <- naiveBayes(term_deposit_subscribed ~ ., laplace = 2, data = bank_notscaled_train)
```

The laplace parameter was changed[1, 2, 3] just to see whether there are any totally odd events in the testing of evaluation data or not. But the results were the same for all the three values.

(D) and (E): The results of all the models were then compares based on class accuracies[Table 1]. The best model had the setting The Naïve Bayesian Model with Laplace = 2 was chosen. There were other combinations of settings which produced the same result.

The reason that it is called Naive is not because it is simple or easy to implement. It is because the modelling technique follows a strong assumption about the observations having features independent of each other while in real-time data, it may be not independent. It assumes that the occurrence of one feature in a category is not related to the occurrence of all other features. If this assumption is satisfied, it performs extremely well and better than other complex models. We know that Naïve Bayes performs well when there are categorical variables. When there are numerical data, we can use binning or discretization to convert them to a categorical variable but for this is assumes that continuous variable is coming from a normal distribution.

Why the above model is not considered: As we know it assumes normality for numerical variables. Here we have a bunch of numerical variables[balance, day, duration, campaign, pdays, previous]. We tried to apply different transformation like a log, sqrt, box-cox, min-max normalisation. But data was still right skewed. We also tried to use discretize function under 'arules' package to carry out binning. The outcome variables with type factor were not making any sense. We were not able to explain the different levels of the factor variables and the assumption of normality was not holding. So, we can't rely on the integrity of this model. (**Fig 4**)

Compare two models [You can look at the whole information(table 4, 5, 6, 7, 8, 9)]

MODEL	Overall Accuracy	Class Accuracy(n-'no'/y-'yes')			Lift(n-'no'/y-'yes')		
		Train	Evaluation	Test	Train	Evaluation	Test
SVM	87.964	53.107/92.247	54.797/92.250	53.947/92.465	1.06/4.05	1.06/4.11	1.06/4.16
Naïve Bayes	88.930	53.484/92.860	53.321/93.172	56.015/93.285	1.06/4.24	1.06/4.24	1.07/4.49

Table 6 – comparison b/w models

Result: (A) We will choose SVM as it is better at estimating both classes as well as there is a less variation in accuracies compared to Naïve Bayes. And also, for above-mentioned reasons that question the integrity of Naïve Bayes. You can check the gain chart(**Fig 6**) and also predictor importance(**Fig 5**) for SVM.

(B) & (C) The question was whether the client subscribes to the term deposit or not. And the problem was to accurately identify the less frequent category which is 'yes'[Client subscribes term deposit?]. As we can see at the above table that SVM can clearly identify whether the client will subscribe for approx. 53% of the time with the misclassification error of approx. 47%. SVM was able to identify that the client will not subscribe accurately for approx. 92% of the time with misclassification error of approx. 8%. We can also look at the predictor importance(**Fig 5**), it says that 'poutcome' is the most importance variable, which makes sense as 'poutcome' represent the results of the previous campaign. Other predictors have low importance amongst all 'balance' has the least importance. When we look at the gain chart it says that we can identify class 'no' 1.1 times more accurately with the model[**Not that important**] and class 'yes' 4.2 times more accurately with the model compared to without the model.

Conclusion: (A) The most important finding is that most important predictor is 'poutcome', which stands for whether after the previous campaign client subscribed the term deposit or not? Which means the outcome of a campaign is crucial for us. We need to carry out the campaign and make sure to reach more clients. Also, the second important feature is 'duration' which indicates how long the client talked to the bank about the related campaign, higher the duration more chance that the client will subscribe. And also, last but not the least important feature 'month' which means the month of the year the bank carries out campaign also play an important role in getting more clients.

(B) Our model can be used by any financial institute to make sure how they lure can clients in subscribing the term deposit for their scheme or not. For example, which communication mode would be useful, which profession mostly subscribe to the term deposit. They can know their target audience like is it females who are in the management field or retired males. They can also know which month of the year would be appropriate to carry out campaigns. They can also target the audience based on the economic characteristics of the clients.

(C) The features that explain the financial status of the clients like income, average debit/credit from a bank account, pattern in payment transactions can also be included. More demographics like the area they live in, the number of children can also be added. The features regarding asset information other than bank balance can also be added. The continuous data is heavily right skewed as well as data has categories like 'unknown', which diminishes the model accuracy.

(D) Further analysis in what were the reasons for clients not subscribing the term deposit in the previous campaign can be carried out by getting feedback from them. Also using those feedbacks creating a dedicated questionnaire for different target audience can be done.

Table 1 Result of SVM

MODEL	STOPPING CRITERIA	REGULARIZATION PARAMETER(C)	RBF GAMMA	KERNEL TYPE	TRAIN_ACCURACY class(yes-y/no-n)	EVAL_ACCURACY class(yes-y/no-n)	TEST_ACCURACY class(yes-y/no-n)
Simple	1.0E-3	10	0.1	RBF	59.448/84.889	51.476/83.952	51.880/83.785
Expert	1.0E-3	10	-	Linear	51.161/92.751	52.768/93.122	51.692/92.912
		100			53.107/92.247	54.797/92.250	53.947/92.465
		1000			52.982/92.205	54.613/92.250	53.759/92.514
	1.0E-4	10			51.161/92.743	52.768/93.122	51.692/92.887
		100			52.982/92.272	54.613/92.225	53.759/92.465
	1.0E-5	10			51.224/92.743	52.768/93.122	51.692/92.887
		100			52.982/92.272	54.797/92.225	53.571/92.465
	1.0E-3	10	0.1	Sigmoid(bias = 0.0, gamma = 1.0)	6.403/93.314	6.089/92.923	7.143/93.211
	1.0E-3	10	0.1	Polynomial(bias = 0.0, gamma = 1.0, degree = 3)	64.093/77.925	49.077/76.452	50.188/75.230
	1.0E-3	10	0.1	RBF	59.448/84.889	51.476/83.952	51.880/83.785
		100			68.675/79.168	54.613/77.523	59.398/76.946
		1000			71.877/75.674	59.225/74.134	62.218/73.191
		10	0.01		18.267/98.597	18.635/98.904	18.985/98.607
		100			52.982/90.743	52.030/90.057	51.128/90.251
		1000			57.878/85.754	52.030/84.999	53.571/85.053
		10	0.001		18.267/98.597	18.635/98.879	18.797/98.632
		100					
		1000					
	1.0E-4	10	0.1	RBF	59.448/84.880	51.292/83.952	51.880/83.785
		100			68.675/79.168	54.613/77.523	59.398/76.946
		1000			71.877/75.674	59.225/74.134	62.218/73.191
		10	0.01		18.267/98.597	18.635/98.904	18.985/98.607
		100					
		1000					
		10	0.001		18.079/98.614	18.450/98.904	18.797/98.657
		100					
		1000					
		10	0.1		59.448/84.880	51.292/83.952	51.880/83.785
		100			68.675/79.168	54.613/77.523	59.398/76.946
		1000			71.877/75.674	59.225/74.134	62.218/73.191

	1.0E-5	10	0.01	RBF	18.267/98.597	18.635/98.904	18.985/98.607
		100			52.982/90.743	52.030/90.057	51.128/90.251
		1000			57.878/85.754	52.030/84.999	53.571/85.053
		10	0.001		57.878/85.737	52.030/85.024	53.571/85.053
		100					
		1000					

Table 2 Naïve Bayes Results

MODEL	LAPLACE	TRAIN_ACCURACY class(yes-y/no-n)	EVAL_ACCURACY class(yes-y/no-n)	TEST_ACCURACY class(yes-y/no-n)
Naïve Bayes Classification	1	53.484/92.860	53.321/93.172	56.015/93.285
	2			
	3			

Table 4 **Training table for SVM**

term_deposit_subscribed		no	yes	Total	Lift
no	Count	10982	923	11905	1.061599
	Row %	92.247	7.753	100	
	Column %	93.631	52.176	88.198	
	Total %	81.36	6.838	88.198	
yes	Count	747	846	1593	4.05225
	Row %	46.893	53.107	100	
	Column %	6.369	47.824	11.802	
	Total %	5.534	6.268	11.802	
Total	Count	11729	1769	13498	
	Row %	86.894	13.106	100	
	Column %	100	100	100	
	Total %	86.894	13.106	100	

Table 5 **Evaluation table for SVM**

term_deposit_subscribed		no	yes	Total	Lift
no	Count	3702	311	4013	1.064605
	Row %	92.25	7.75	100	
	Column %	93.793	51.151	88.101	
	Total %	81.273	6.828	88.101	
yes	Count	245	297	542	4.105272
	Row %	45.203	54.797	100	
	Column %	6.207	48.849	11.899	
	Total %	5.379	6.52	11.899	
Total	Count	3947	608	4555	
	Row %	86.652	13.348	100	
	Column %	100	100	100	
	Total %	86.652	13.348	100	

Table 6 **Testing table for SVM**

term_deposit_subscribed		no	yes	Total	Lift
no	Count	3718	303	4021	1.062304
	Row %	92.465	7.535	100	
	Column %	93.818	51.356	88.315	
	Total %	81.66	6.655	88.315	
yes	Count	245	287	532	4.163091
	Row %	46.053	53.947	100	
	Column %	6.182	48.644	11.685	
	Total %	5.381	6.304	11.685	
Total	Count	3963	590	4553	
	Row %	87.042	12.958	100	
	Column %	100	100	100	
	Total %	87.042	12.958	100	

Table 7 **Training table for Naïve Bayes**

term_deposit_subscribed		no	yes	Total	Lift
no	Count	11055	850	11905	1.062586
	Row %	92.86014	7.139857	100	
	Column %	93.71821	49.94125	88.19825	
	Total %	81.90102	6.297229	88.19825	
yes	Count	741	852	1593	4.241639
	Row %	46.51601	53.48399	100	
	Column %	6.28179	50.05875	11.80175	
	Total %	1.428	10.257	11.80175	
Total	Count	11796	1702	13498	
	Row %	87.39072	12.60928	100	
	Column %	100	100	100	
	Total %	87.39072	12.60928	100	

Table 8 **Evaluation table for Naïve Bayes**

term_deposit_subscribed		no	yes	Total	Lift
no	Count	3739	274	4013	1.063125
	Row %	93.17219	6.82781	100	
	Column %	93.66232	48.66785	88.10099	
	Total %	82.08562	6.015368	88.10099	
yes	Count	253	289	542	4.313984
	Row %	46.67897	53.32103	100	
	Column %	6.337675	51.33215	11.89901	
	Total %	1.428	10.257	11.89901	
Total	Count	3992	563	4555	
	Row %	87.63996	12.36004	100	
	Column %	100	100	100	
	Total %	87.63996	12.36004	100	

Table 9 **Testing table for Naïve Bayes**

term_deposit_subscribed		no	yes	Total	Lift
no	Count	3751	270	4021	1.065816
	Row %	93.28525	6.714748	100	
	Column %	94.12798	47.53521	88.3154	
	Total %	82.38524	5.930156	88.3154	
yes	Count	234	298	532	4.490079
	Row %	43.98496	56.01504	100	
	Column %	5.87202	52.46479	11.6846	
	Total %	1.428	10.257	11.6846	
Total	Count	3985	568	4553	
	Row %	87.52471	12.47529	100	
	Column %	100	100	100	
	Total %	87.52471	12.47529	100	

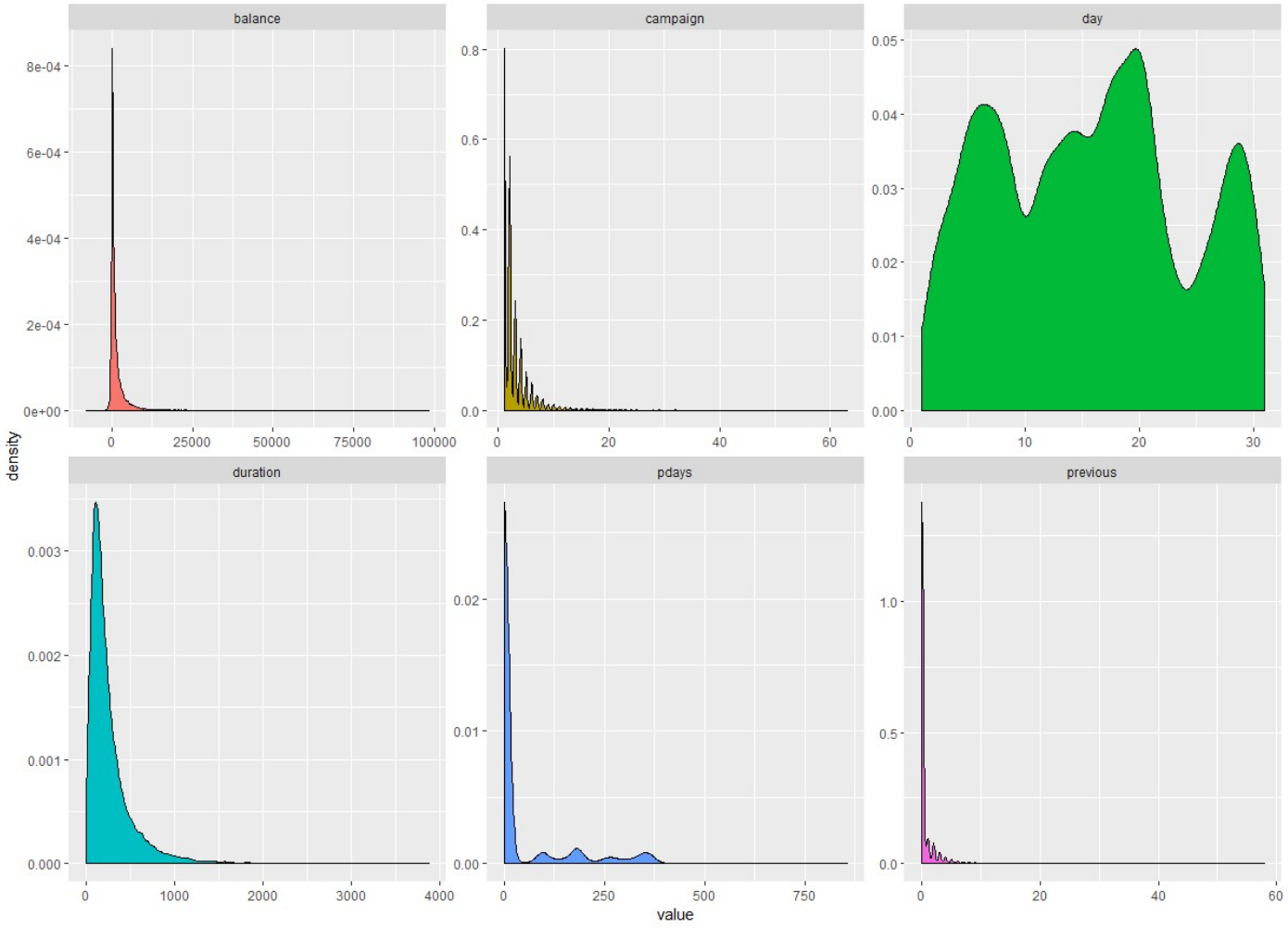


Fig 4 Density plot of continuous variables [Not normal distribution]

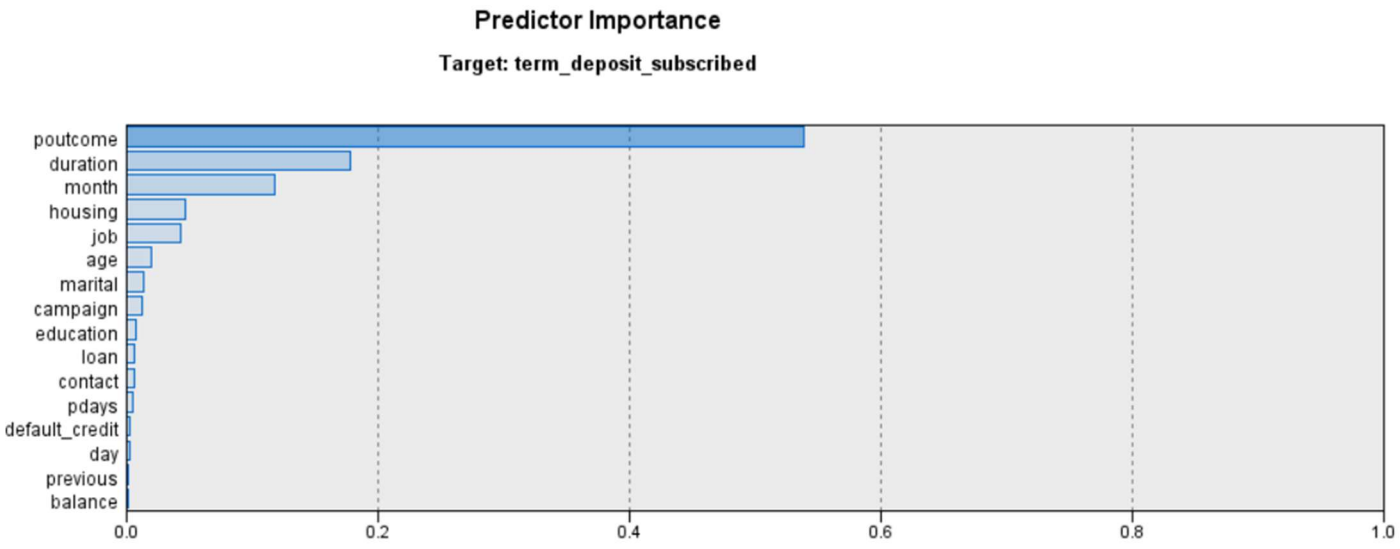


Fig 5 Predictor importance for SVM

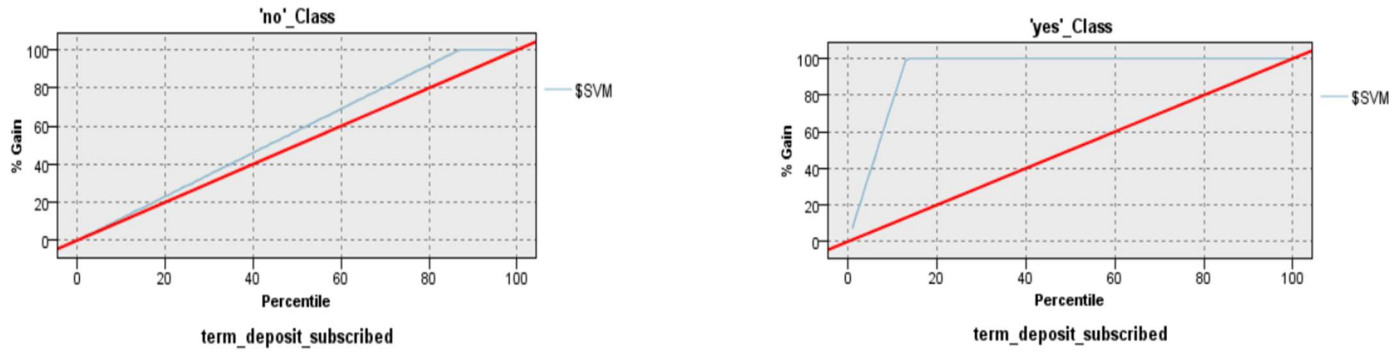


Fig 6: Lift Charts for SVM

Appendix 2 (ID 45665761) : CHAID and QUEST

Methodology

Brief description of CHAID and QUEST methods:

CHAID stands for Chi-squared Automatic Interaction Detection which is one of the classification algorithms that uses chi-square statistics to find an optimal split to build a decision tree. Under CHAID chi-square independence test is carried out between all the predictor variables and outcome variable. Then the results of p-values are compared to get the most significant variable, the variable with least p-value is chosen to make a split. Also, the predictor variable with more than 2 categories are compared and the categories with least difference are merged together, this process continue until we get a significant difference between the categories. In case of merging the nominal input variable any category can be merged (makes no difference as there is no specific order) but in case of ordinal variables only adjacent categories are merged. The best thing about this method is that it takes up continuous as well as categorical variables for input and outcome variables. Unlike C&R tree, it can create non-binary trees but the limitation for ordinal inputs is that it must have numerical input.

QUEST works similar to CHAID but there is one difference that it doesn't compute tests for category combinations which enhances the processing of tree as a smaller number of computations are being carried out. For this method inputs could be numerical or categorical but outcome should be categorical and it generates only binary splits.

The aim of the project is to predict whether the customer subscribe for the term deposit (i.e. yes or no) which is a binary classification problem, so two of the binary classification methods CHAID and QUEST were applied to the sample (50%) of original data using IBM SPSS MODELER as the input and outcome variables satisfy the requirements of these classifiers. The outcomes of both the methods are compared to get the best model by comparing lift, class specific and overall accuracies which are goodness of fit measures. Higher the values of these measures better the model. Both of the methods were suitable for the data as it takes up both categorical and continuous inputs. The classifiers were used with three different combinations of misclassification costs but results were quite similar. The parameters to generate the best model are given below:

Classification Algorithm	Parameters
CHAID	Maximum tree depth = 5 Min. records per parent branch =2% Min. records per child branch =1% Misclassification cost = 1:1(Yes: no) Significance level for splitting and merging =0.05 Remaining parameters as default
QUEST	Maximum tree depth =5 Pruning =yes Min. records per parent branch =2% Min. records per child branch =1% Significance level for splitting = 0.05 Remaining parameters as default

Table 6: Parameters for CHAID and QUEST

To get the best model significance level of 5% is used which is standard level used in test statistics that decides which variable to choose for best split and to avoid overfitting, records per branch are specified. Neither of

the model is robust to unbalance data as significance values were imputed using Bonferroni method. On comparison, Chaid model was found better than Quest as it generates slightly higher accuracies and results are consistent between two among class specific accuracies.

Model	Overall Accuracy	Class Specific Accuracy			Lift		
		Training	Evaluation	Test	Training	Evaluation	Test
CHAID	88.75%	96.774%(no) 32.831%(yes)	96.536%(no) 29.520%(yes)	96.394%(no) 31.015%(yes)	1.03 4.89	1.03 4.5	1.03 4.56
QUEST	88.95%	96.9%(no) 27.621%(yes)	96.412%(no) 25.83%(yes)	96.472%(no) 30.075%(yes)	1.03 4.6	1.02 4.14	1.03 4.7

Table7: Comparison of CHAID and QUEST Model

Results

The confusion matrix is displayed of Chaid classifier for training. evaluation and test data sets below:

Chaid										
Training	Term deposit subscribed	No	Yes	Total	#Good pred. of A	#pred.	#A	n	Lift	Error rate%
	No	11521	384	11905	11521	12591	11905	13498	1.03	8.5
	Yes	1070	523	1593	523	907	1593	13498	4.89	42.34
	Total	12591	907	13498						
Evaluation										
	No	3874	139	4013	3874	4256	4013	4555	1.03	8.97
	Yes	382	160	542	160	299	542	4555	4.5	46.49
	Total	4256	299	4555						
Test										
	No	3876	145	4021	3876	4243	4021	4553	1.03	8.65
	Yes	367	165	532	165	310	532	4553	4.56	46.77
	Total	4243	310	4553						
Quest										
Training										
	No	11536	369	11905	11536	12689	11905	13498	1.03	9.09
	Yes	1153	440	1593	440	809	1593	13498	4.6	45.61
	Total	12689	809	13498						
Evaluation										
	No	3869	144	4013	3869	4271	4013	4555	1.02	9.41
	Yes	402	140	542	140	284	542	4555	4.14	50.7
	Total	4271	284	4555						
Test										
	No	3890	131	4021	3890	4262	4021	4553	1.03	8.73
	yes	372	160	532	160	291	532	4553	4.7	45.02
	Total	4262	291	4553						

Table 8: Class specific error rate and lift calculation

From the above output it is clear that class specific error rate is consistent through all data sets, so therefore, it can prove to be the best model than the other. The predictor importance graphic (below) shows that "duration" is most significant predictor followed by month, pdays, contact and so on. In terms of decision tree, the full tree is displayed below and a brief description of the tree is as follows:

The first split is made using "duration", then month and then the pdays and so on. The first leaf of the output is described as - if the call duration of client was less than 88sec and month of contact was among April, December, Jan, May then he will not subscribe for term deposit as that falls under "no" category.

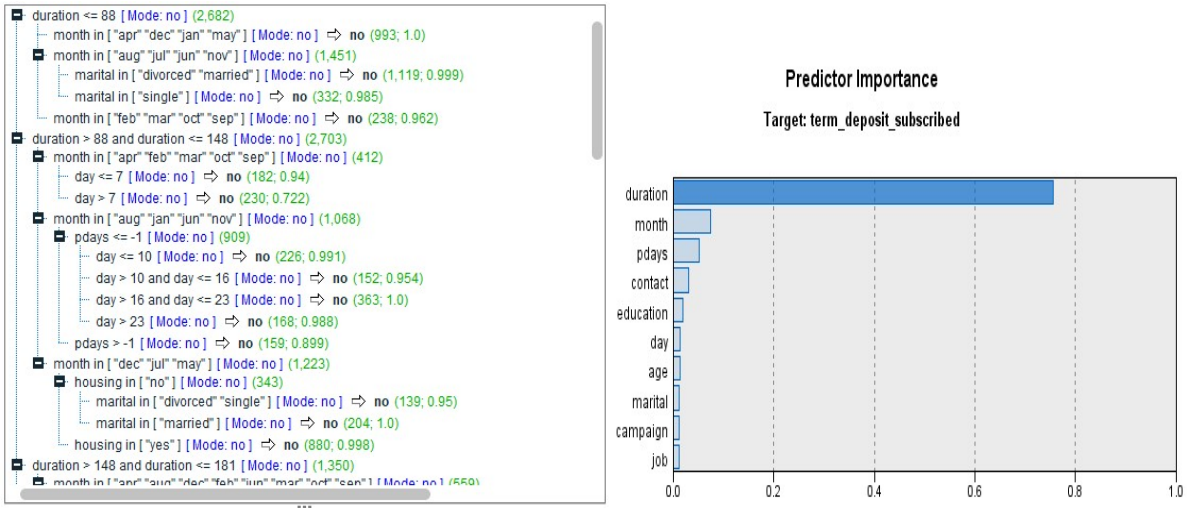


Figure 7: Decision tree of Chaid Classifier

Conclusion

To conclude, the call duration plays a significant role in predicting client's term deposit subscription besides other factors such as month of contact, days passed after previous contact etc. The call duration results are quite effective in terms of generating success in future. Other predictors such as month and number of times contacted automatically becomes reliable to forecast the maximum profit for the bank as more the person is in contact it is highly likely that client would subscribe term deposit. Therefore, bank should spend more time on conversation and client should be contacted on frequent basis to approach towards maximum subscriptions. The other important features of client such as the financial conditions of the client needs to be taken care of for better outcome as the client with large amount loan and low income would probably refuse to subscribe therefore predictors such as income could be clubbed with loan to make sure if that client could subscribe in future or not before contacting that client. The limitation of the analysis is that important factors such as demographics and loan were least considered as these factors would also generate more success if taken into account. In terms of dataset most of the predictors were categorical and some of the variables were right skewed. For future analysis, the factors responsible for-why the clients didn't subscribe for term deposit must be considered whether more data such as demographics of client needs to be collected or some improvisation among team members of the organisation should be conducted.

Appendix 3 : C&R Tree and Neural Networks (Student ID: 45534632)

Methodology:

The Bank Marketing Dataset consists of sixteen input variables and one output variable. The following models were built on the dataset and analysed.

Classification and Regression Tree (C&R Tree): It is a tree-based classification model which is quite analogous to C5.0 i.e., it splits the observations under study into segments having identical target field values. They build decision trees to generate simple rules, which can be interpreted easily. The process begins by inspecting the input variables to discover the most appropriate split by impurity reduction (techniques like Gini and Entropy reduction) resulting from the split. This procedure continues until it encounters some terminating criteria. The biggest strength of this Algorithm is that it can put up with numeric and categorical target fields. Also, this model has the ability to ignore the less important predictor fields automatically. The C&R tree was one of the models which generated outputs in no time as it could easily handle various datatypes and the text type target field.

For this project, the above method was utilized in the form of a classification node in IBM-SPSS Modeler. A C&R Tree node was connected to the training segment of our dataset (randomly split in R). A new model was generated by allocating a customized value for the maximum depth of the tree, which was 4. A default depth of 5 was also implemented but the misclassification error was slightly higher, so it was excluded. Pruning has been done to avoid overfitting of the model with 5 surrogates. The terminating rules were provided in the form of percentages where the minimum number of records in the parent branch were set to 2% and 1% in the child branch respectively. Two model outputs were generated: by considering misclassification costs, by not considering them. In the full dataset, the target variable, "term_deposit_subscribed" had a majority of the values as "no" (approximately 88%) and "yes" as the minority (~12%). Misclassification costs assign a weight to reduce the number of "yes" values being wrongly classified as "no". Hence, the cost matrix was refurbished with yes->no value as 0.88 and no->yes value of 0.12 (to hold the data integrity). The prior probability values were automatically adjusted based on the frequency of input data (set to default in IBM SPSS). As bagging and boosting techniques weren't used, ensemble settings were unchanged. The tree's output matrix was not altered when subject to changes in the impurity reduction technique. Hence, Gini was chosen as the preferred technique with a sensitivity of 0.0001. About 30% of the training data is internally utilized to detect errors during the process and to prevent overfitting. As the value was taken past 51%, the model generated higher misclassification error rates. Therefore, Overfit Prevention Set was chosen as 30%. The output node of this model was utilized to verify the performance on the other two datasets i.e., evaluation and test. It generated fruitful results without compromising on the class specific accuracy.

Neural Networks: As the name suggests, the structure and functioning of this model is quite similar to the human nervous system. It consists of three major layers and can have multiple sub layers depending on the complexity. The first layer, called input layer consists of all the input variables used in the analysis. The second layer, called the hidden layer acts as a channel of communication between the first and third layers. Multiple hidden layers can be used. The third layer is called the output layer, which contains the corresponding target variables. This method links the input units and output units by assigning weights to each one of them and backpropagation process reduces the error values in each step until a threshold value is reached.

The above algorithm was utilized in the form of a classification node in IBM-SPSS Modeler. The "Neural Net" node was connected to the bank_marketing_train dataset. The performance of neural networks is on par with the C&R Trees as they too can handle numeric and categorical variables comfortably. Though it has its strengths, it is commonly referred to as 'Black Box' as its interpretation is more complex when compared to other models. All the input variables were used as predictors and the variable "term_deposit_subscribed" was set as target. A new standard model was built using the multilayer perceptron technique as it produced

relatively better results when compared to the Radial Basis Function technique, but the processing time had to be compromised with. (Appendix). The number of units in the hidden layer was automatically selected by the algorithm to optimize it. The only stopping parameter used was the execution time i.e., the upper limit of training time was chosen as 15 minutes. No minimum accuracy and number of max. training cycles was explicitly specified. And, as in the C&R Tree, the overfit prevention was set to 30%. Bagging and Boosting weren't used to build the model as they took up a lot of time for processing (Appendix).

Algorithm	Training Dataset					
	Accuracy of yes	Accuracy of no	Overall accuracy	Lift of yes	Lift of no	Misclassification Error
Neural Networks	45.23%	95.81%	89.72%	4.95	1.05	54.78%
C&R Tree	81.21%	76.96%	77.47%	2.70	1.10	18.79%

Table 9: Training set parameters for NN and C&R Tree

Algorithm	Evaluation Dataset					
	Accuracy of yes	Accuracy of no	Overall accuracy	Lift of yes	Lift of no	Misclassification Error
Neural Networks	45.08%	95.34%	89.40%	4.78	1.05	54.92%
C&R Tree	83.30%	77.19%	77.91%	2.78	1.10	16.70%

Table 10: Evaluation set parameters for NN and C&R Tree

Algorithm	Test Dataset					
	Accuracy of yes	Accuracy of no	Overall accuracy	Lift of yes	Lift of no	Misclassification Error
Neural Networks	40.96%	95.49%	89.26%	4.72	1.05	59.04%
C&R Tree	79.23%	77.34%	77.55%	2.72	1.09	20.77%

Table 11: Testing set parameters for NN and C&R Tree

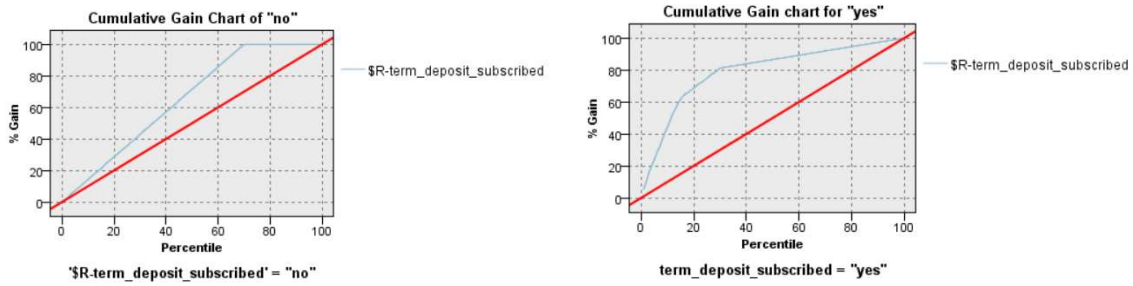


Figure 8: Lift Charts for NN and C&R Tree

he C&R Tree model is superior as misclassification errors are consistent and quite smaller than the Neural Networks and accuracy of yes (minority value) is higher. The overall accuracy of Neural Networks is higher than the tree, but the class specific accuracy of yes is quite low.

Results:

The IBM SPSS C&R Tree Node displayed the following results:

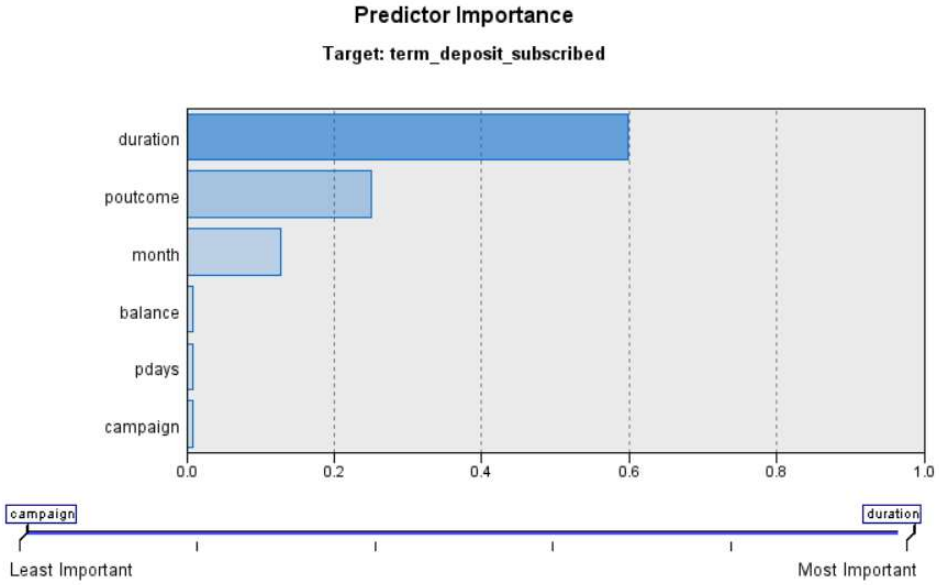


Figure 9: Predictor Importance

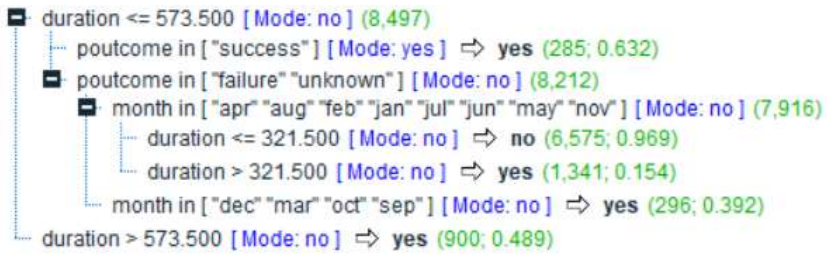


Figure 10 : The Classification and Regression Tree

The inputs: duration (60%), poutcome (25%) , balance (13%), pdays (1%) and campaign (1%) in decreasing importance are the major contributors to the construction of the C&R Tree. According to the parameters tuned above, the following four splits were observed based on the reduction of importance:

- The duration > 573.5 seconds was classified as yes and duration <= 573.5 seconds was further split into two branches based on the poutcome variable.
- The poutcome input that held the value “success” , was classified as yes and the poutcome input which held the values “failure” and “unknown” was further split into two branches based on the month variable.
- The month predictor variable which had the observations “dec”, “mar”, “oct”, “sep” was forecasted as yes and the observations with the rest of the months were further split into two branches based on the duration of call.
- The duration > 321.5 seconds was classified as yes and duration <= 321.5 seconds as no.

A closer look at the tree displayed above reveals that the factor "duration" contributes majorly in the classification process i.e., the client subscription to the term deposit is heavily reliant on the duration of the call, outcome of the previous marketing campaign and the account balance of the lead. Hence, the scale of impact of each predictor input on the subscription is clearly evident. Class specific accuracies of the test dataset are 79.23% and 77.34% for yes and no output values respectively, which effectively states that a successful client subscription can be predicted through this model. The lift of 2.72 for yes indicates that this model performs twice as better than a normal prediction.

Conclusion:

In a nutshell, Classification and Regression Tree algorithm is quite effective and a better model to forecast the term deposit subscription if its related banking factors are available. The variables duration, poutcome and balance are the top 3 inputs (among 16) that are pertinent to predict the success of a term deposit by a client. According to the results, the most important factor is the duration of a call. Also practically, duration plays a pivotal role especially on unseen data as higher call duration could mean that the customer is properly engaged in the conversation which increases the likelihood of a subscription. Streamlining the client-employee interactions at a contact centre with well-trained vocal staff would simplify the functional management of campaigns. And, the previous outcome's success almost guarantees a success to the term deposit subscription and vice-versa. The main limitation of the bank marketing dataset is that most of the data is right skewed. Additional predictors of the potential client demographics (like Annual Income, Region) could be very helpful. Majority of the variables were categorical in nature and one of them was converted into categorical for better results (age). There's ample scope for research as the data isn't totally relevant to the current market trends viz, telephone campaigning has taken a slight back step when compared to digital media marketing. It could be rectified by adding additional categories to the variable "contact" like digital, television etc. It could also be used to analyse how the economic and socio-cultural factors have impacted the subscription decision.

Appendix 4 : C5.0 Tree and Logistic Regression (Student ID: 45533865)

METHODOLOGY

Two models of decision tree classification are applied on the dataset:

1. **C5.0** 2. **Logistic Regression**

1.) C5.0: This model is executed with the C5.0 node in IBM SPSS Modeler. It is an improved version of C4.5 and ID3 algorithms. Our target is to predict term deposit subscribes that is yes or no. Since the target field is required to be categorical, the dataset would precisely suit this node. The C5.0 node generates either a decision tree or a rule set. This model is implemented by splitting the sample based on the attributes that provides the maximum information gain (Entropy) at each level.

C5.0 Decision Tree Settings:

Firstly, in IBM SPSS Modeler, the C5.0 decision node was applied on the training dataset. Predefined role is used for the fields tab, partitioned data build model foe each split is not selected under the model tab because we already split the data randomly into training, evaluation and testing. The use of misclassification costs is selected to identify the cost that has occurred due to misclassification so that a better model is predicted. Pruning has not been done because it did not produce good results and the misclassification was much higher. Lastly, in the analyse tab, calculate predictor importance option is selected to identify which factor has affected the most in to predict the outcome.

The model is initiated by splitting the sample according to the attribute that offers the maximum information gain. Each sub sample described by the first split is then split again, generally based on a different attribute, and the process continues until the subsamples cannot be split any further. Ultimately, the low-level splits are re-examined, and those attributes that do not impact significantly to the value of the model are removed or pruned. C5.0 model is quite robust in the existence of difficulties such as missing data or large number of input fields. Normally it does not require long training times to estimate the result. Moreover, C5.0 models are easy to perceive than some other model types, since the rules derived from this model have a very specific interpretation.

2.) Logistic Regression: The logistic regression model is appropriate to deal with issues of many types of data sets. It is provided sufficiently several and well-distributed samples. Moreover, it is suitable for explaining and testing hypotheses regarding the relationships between a categorical outcome variable and one or more categories of continuous predictor attributes. It is a statistical technique used for classifying records based on values of input attributes. It is like linear regression but takes a categorical target field instead of a numeric range. Logistic Regression uses highest probability estimation rather than the least squares estimation that is used in linear multiple regression. Initial values of the predicted parameters are used and the probability of the sample that came from a population with those parameters is calculated. The values of the estimated parameters are altered iteratively until the greatest probability value out of them is obtained.

The above algorithm was executed in the form of the classification node in IBM SPSS Modeler. The 'Logistic' node was connected to the training dataset. The multinomial procedure is used instead of binary because the latter could not handle long data. The predictor importance is also computed to identify the most and the least contributing attributes. The step summary, parameter estimates, and case processing summary is also given (Appendix).

MODEL	Overall Accuracy	Class Accuracy (n- 'no'/ y-'yes')			Lift (n- 'no'/y-'yes')		
		Train	Evaluation	Test	Train	Evaluation	Test
C5.0	74.5%	73.608/89.71	73.78/88.86	72.99/86.731	1.11/2.63	1.11/2.64	1.10/2.56
Logistic Regression	90.45%	97.65/35.736	97.361/38.033	97.62/34.808	1.04/5.59	1.04/5.57	1.03/5.72

Table12: Comparison of two models

RESULTS

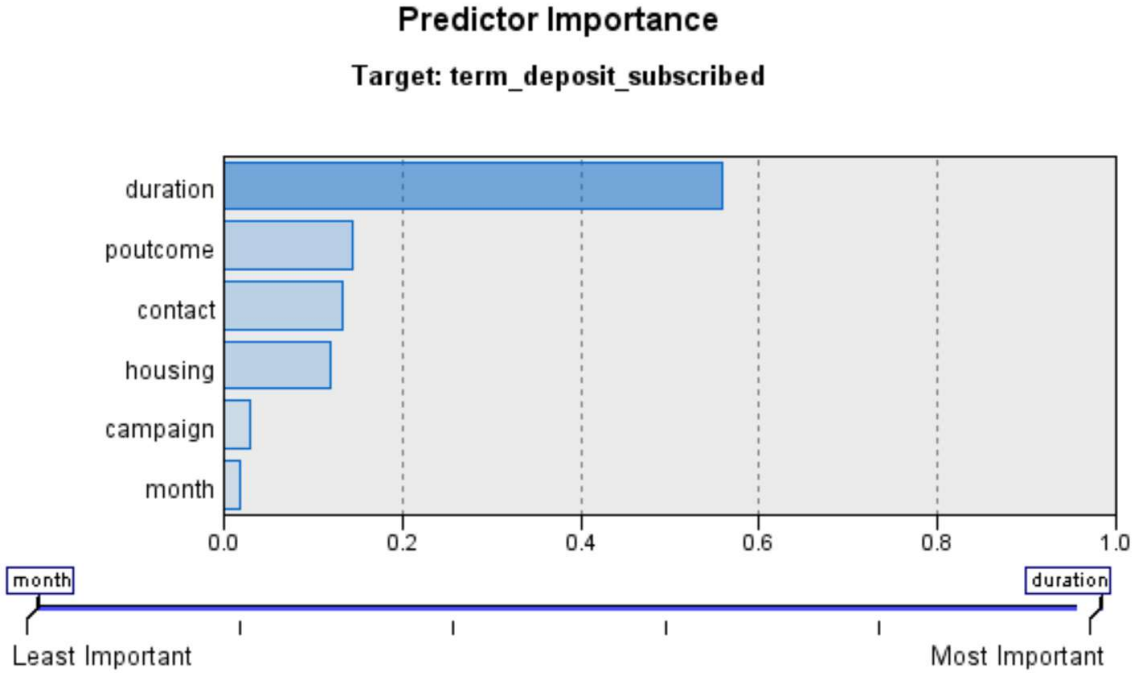


Figure 11: Predictor importance for C5.0 Model

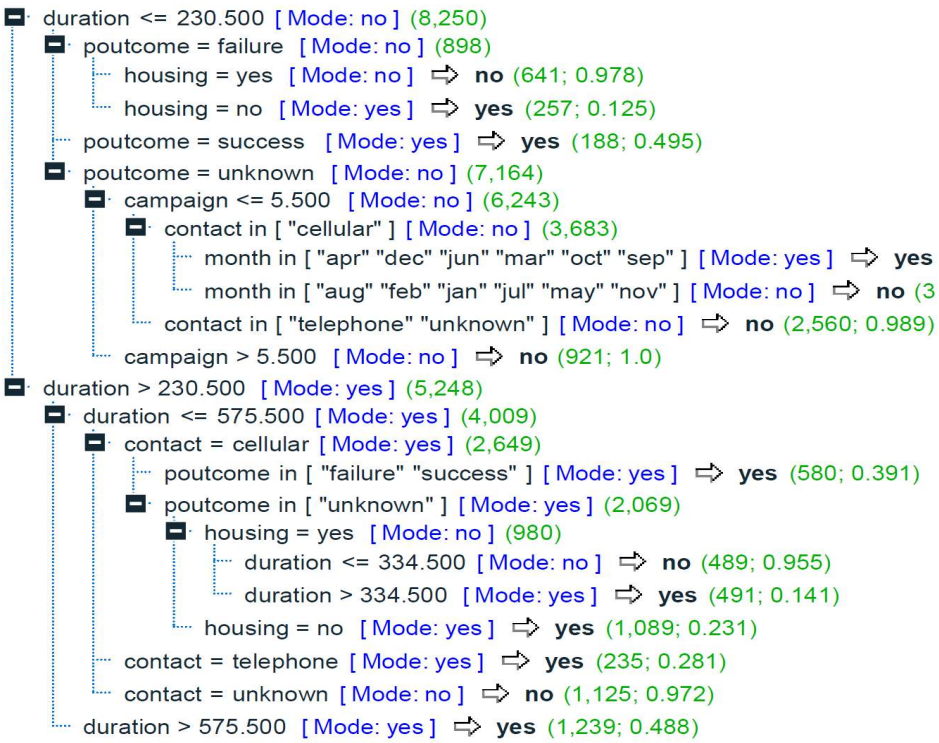


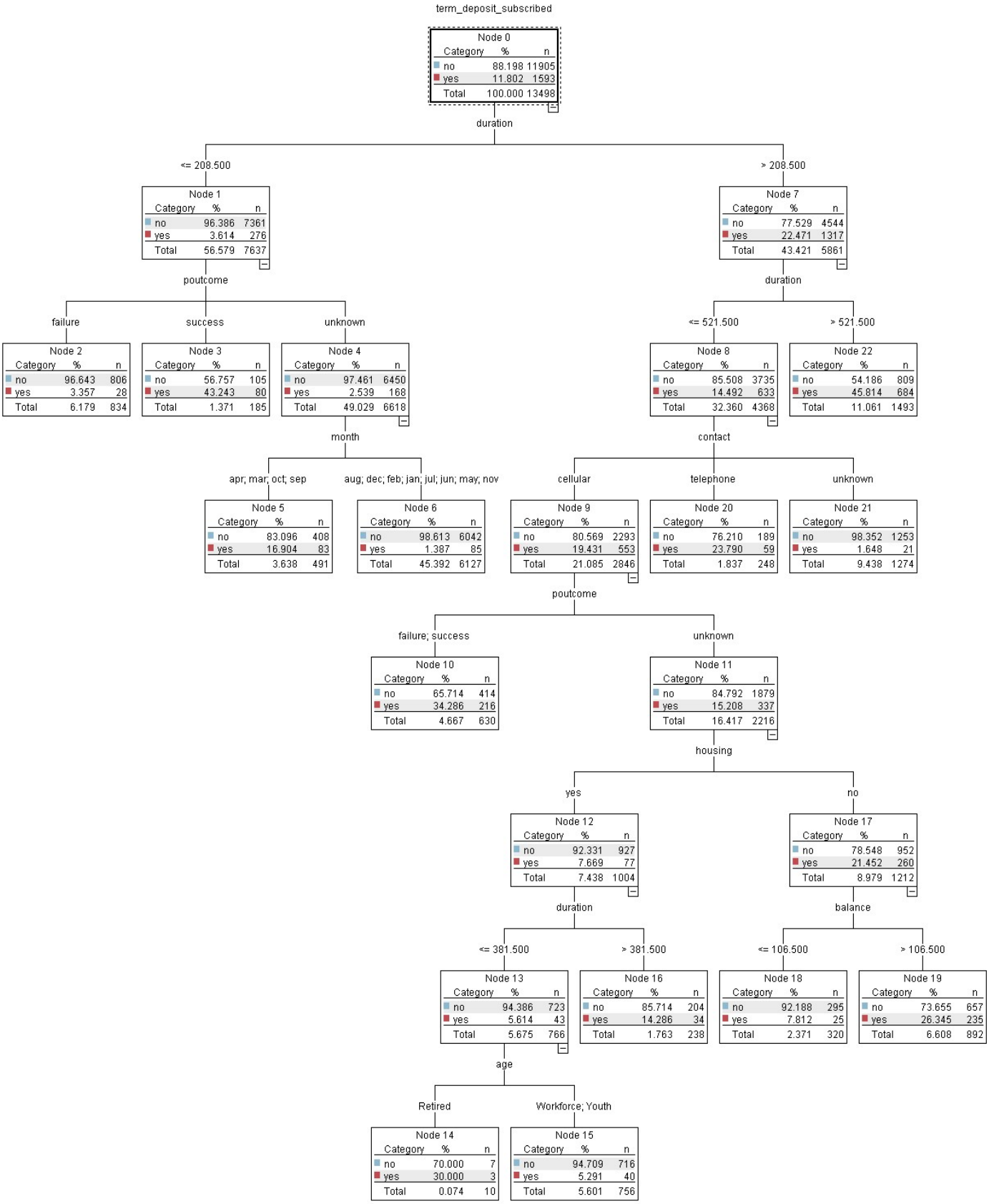
Figure 12: Decision Tree for C5.0 Model

It is evident from the table that, C5.0 is superior than Logistic Regression. Although, the overall accuracy of Logistic Regression is better than that of C5.0, the class accuracy for both the classes is much better in C5.0. Moreover, in C5.0 the output mostly depends on six predictor variables and it produces a decision tree with depth equal to six. The predictor importance chart is given (figure 1). The first split in the tree is done with the variable 'duration'. It has split he data with duration less than or equal to 230.5 and more than 230.5. It can be observed from the tree that if the duration is less than 230.5 the outcome is 'no' and it is 'yes' otherwise. It is further split based on poutcome, contact, housing, campaign and month.

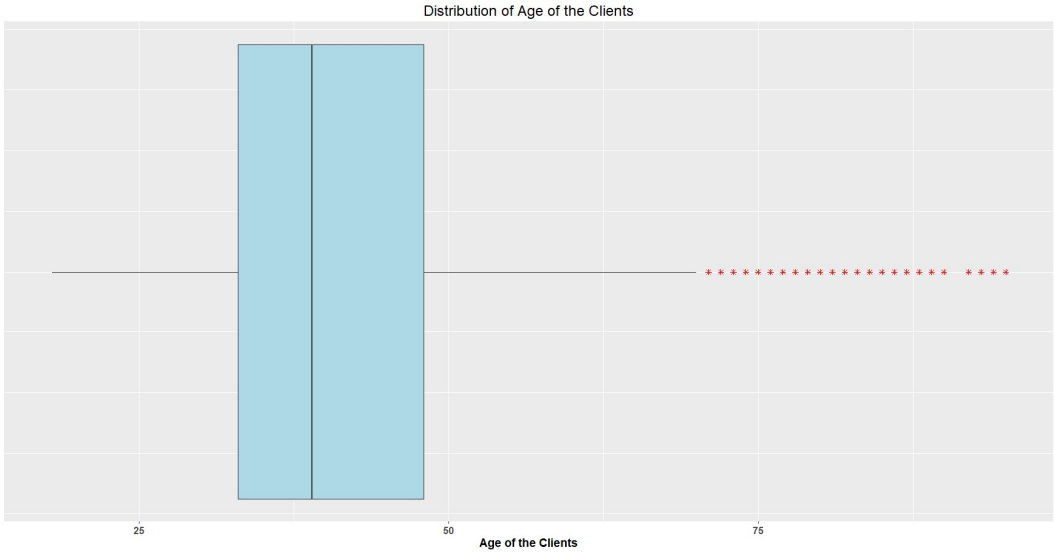
CONCLUSION

To sum up, C5.0 algorithm is quite accurate and a better model to predict the term deposit subscription from the corresponding banking factors that are provided. The variables 'duration' is the most important factor (among 16) to predict the success of the term deposit which is followed by poutcome, contact, housing, campaign and month. The duration of the call almost assures a success to the term deposit subscription and vice-versa. According to the results, the second crucial factor is 'poutcome'. Moreover, the 'poutcome' plays a key role because it is more likely that a client will subscribe for a term deposit given that he/she has already opted for it during the previous campaign. The major limitation of the bank marketing dataset is that most of the data is right skewed. Majority of the variables were categorical in nature and one of them was converted into categorical (age) to increase its predictor importance. There is an abundant scope for research because this data set used conventional medium of marketing i.e. telemarketing. However, in this modern world, there is a potential scope for marketing using digital platforms. As we all know, the power of social media is growing rapidly, advertising on social media can create considerable impact.

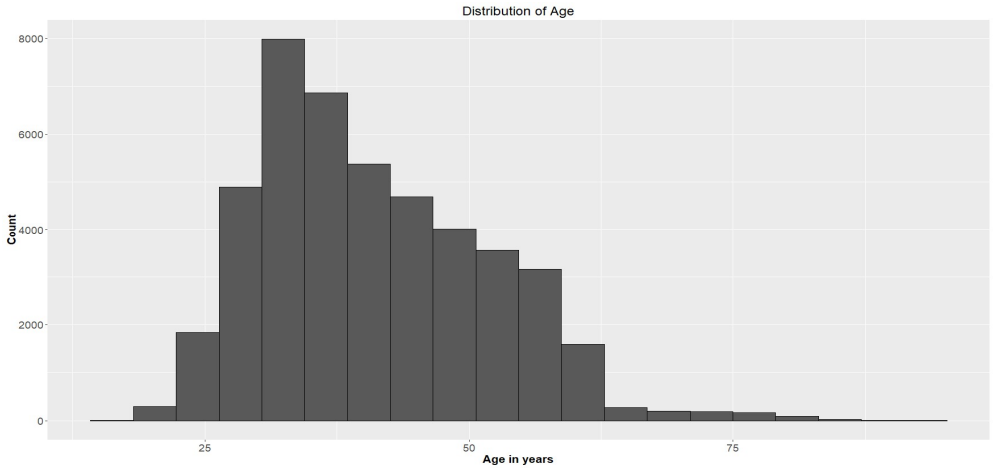
Fig C5.0


term_deposit_subscribed
Node 0
Category % n
no 88.198 11905
yes 11.802 1593
Total 100.000 13498
duration
<= 208.500
> 208.500
Node 1
Category % n
no 96.386 7361
yes 3.614 276
Total 56.579 7637
Node 7
Category % n
no 77.529 4544
yes 22.471 1317
Total 43.421 5861
poutcome
duration
failure
success
unknown
<= 521.500
> 521.500
Node 2
Category % n
no 96.643 806
yes 3.357 28
Total 6.179 834
Node 3
Category % n
no 56.757 105
yes 43.243 80
Total 1.371 185
Node 4
Category % n
no 97.461 6450
yes 2.539 168
Total 49.029 6618
Node 8
Category % n
no 85.508 3735
yes 14.492 633
Total 32.360 4368
Node 22
Category % n
no 54.186 809
yes 45.814 684
Total 11.061 1493
month
contact
apr; mar; oct; sep
aug; dec; feb; jan; jul; jun; may; nov
cellular
telephone
unknown
Node 5
Category % n
no 83.096 408
yes 16.904 83
Total 3.638 491
Node 6
Category % n
no 98.613 6042
yes 1.387 85
Total 45.392 6127
Node 9
Category % n
no 80.569 2293
yes 19.431 553
Total 21.085 2846
Node 20
Category % n
no 76.210 189
yes 23.790 59
Total 1.837 248
Node 21
Category % n
no 98.352 1253
yes 1.648 21
Total 9.438 1274
poutcome
failure; success
unknown
Node 10
Category % n
no 65.714 414
yes 34.286 216
Total 4.667 630
Node 11
Category % n
no 84.792 1879
yes 15.208 337
Total 16.417 2216
housing
yes
no
Node 12
Category % n
no 92.331 927
yes 7.669 77
Total 7.438 1004
Node 17
Category % n
no 78.548 952
yes 21.452 260
Total 8.979 1212
duration
balance
<= 381.500
> 381.500
<= 106.500
> 106.500
Node 13
Category % n
no 94.386 723
yes 5.614 43
Total 5.675 766
Node 16
Category % n
no 85.714 204
yes 14.286 34
Total 1.763 238
Node 18
Category % n
no 92.188 295
yes 7.812 25
Total 2.371 320
Node 19
Category % n
no 73.655 657
yes 26.345 235
Total 6.608 892
age
Retired
Workforce; Youth
Node 14
Category % n
no 70.000 7
yes 30.000 3
Total 0.074 10
Node 15
Category % n
no 94.709 716
yes 5.291 40
Total 5.601 756

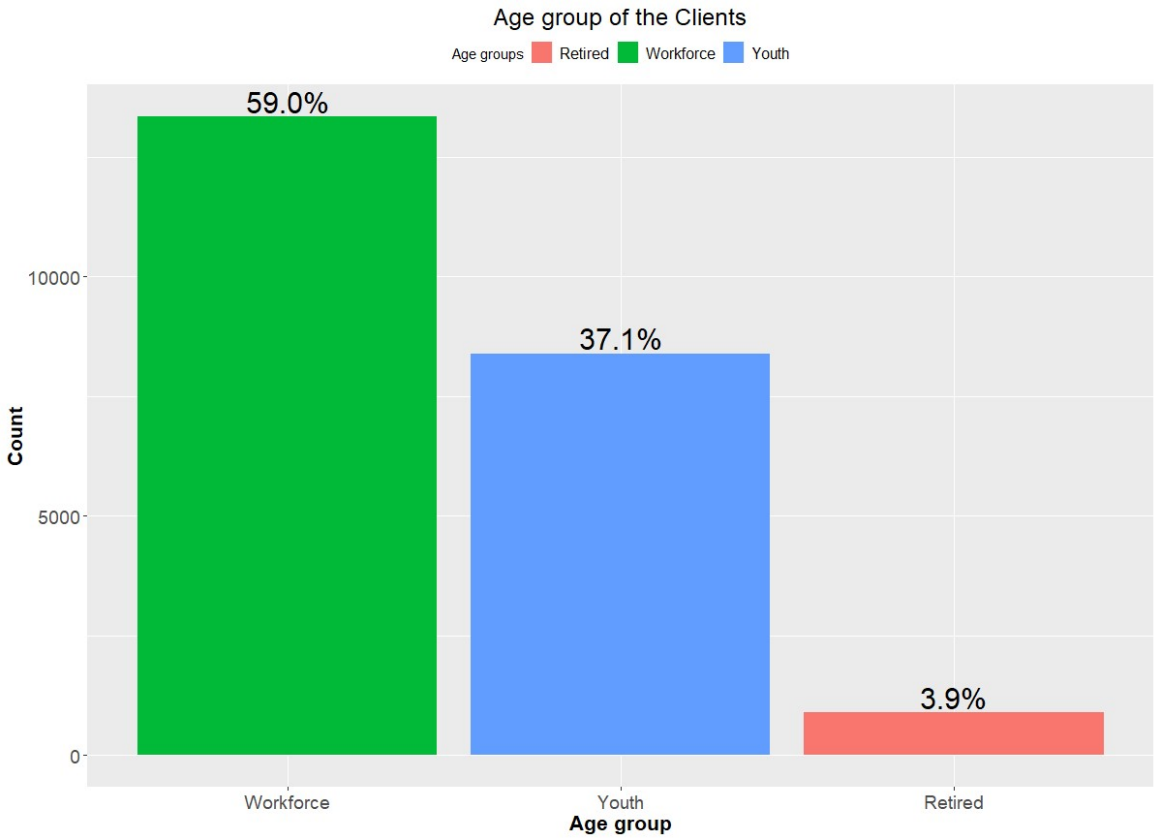
Appendix 5: Boxplot - Distribution of the Age of the Clients



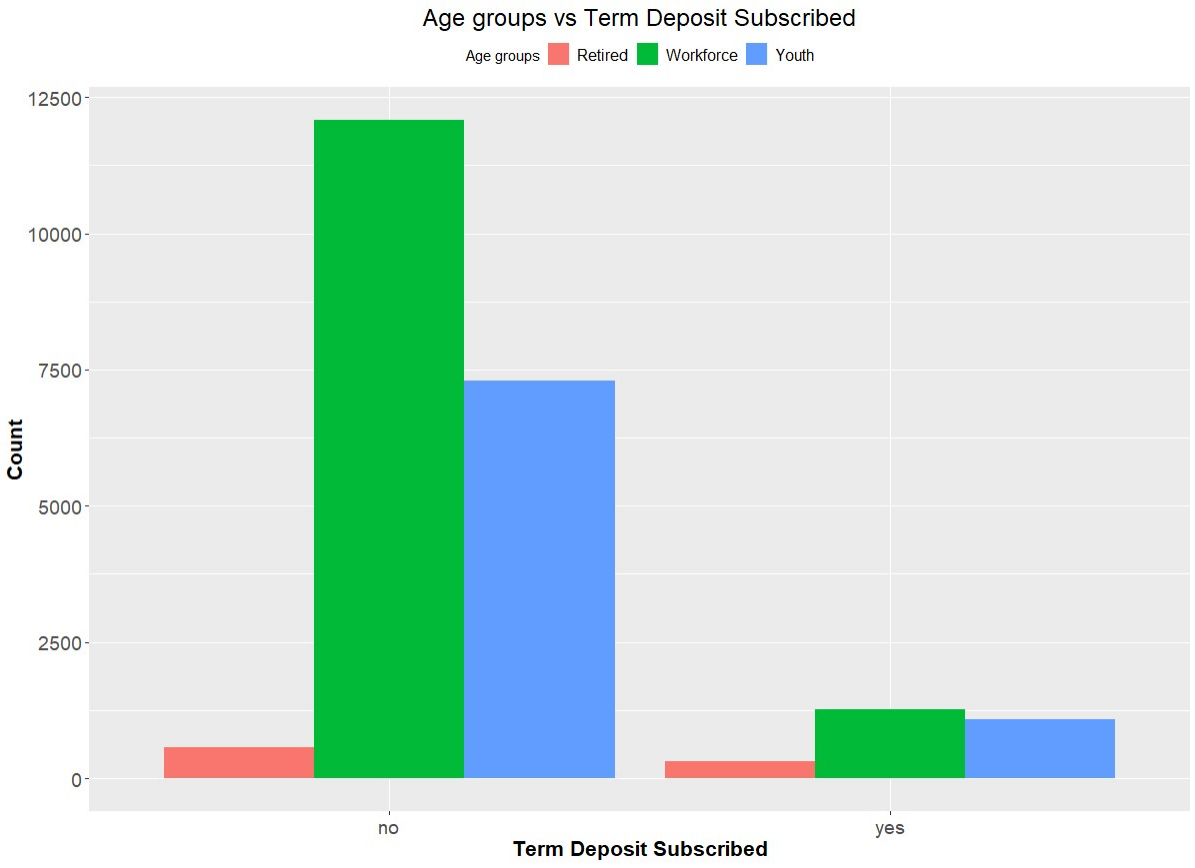
Appendix 6: Histogram - Distribution of the Age of the Clients



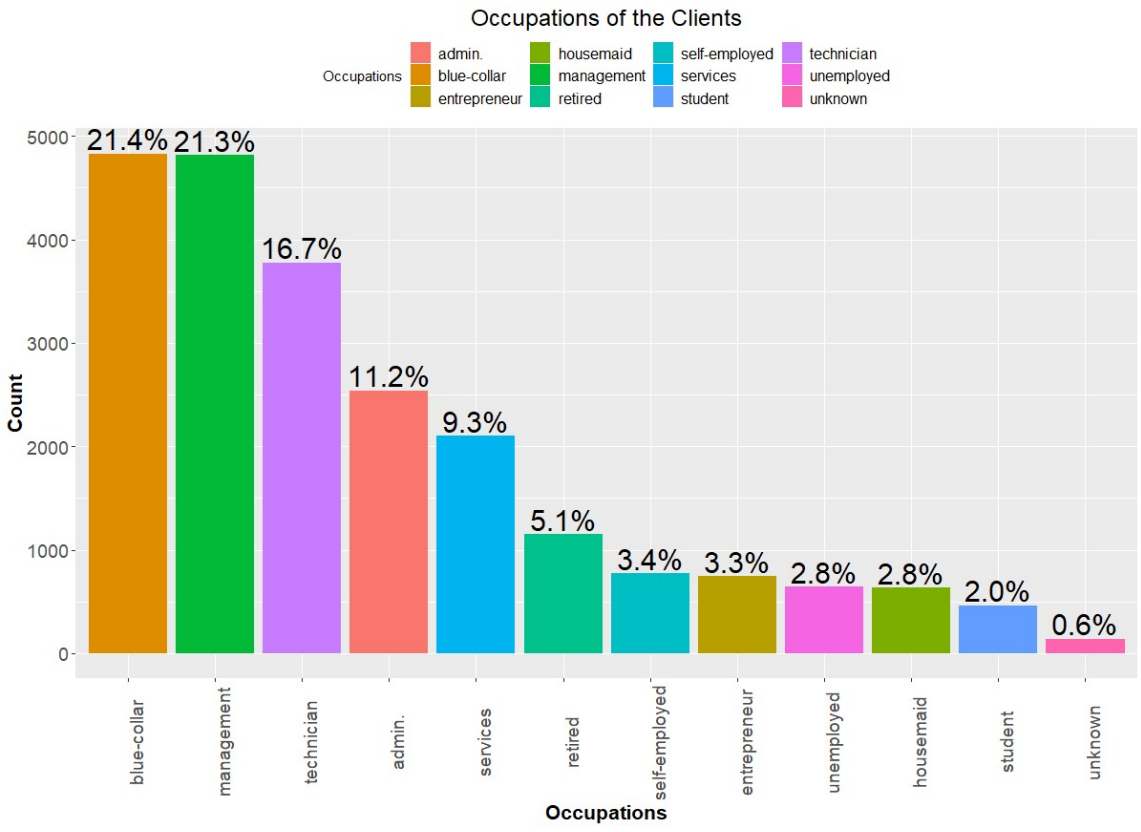
Appendix 7: Bar Graph of the categorized Age groups



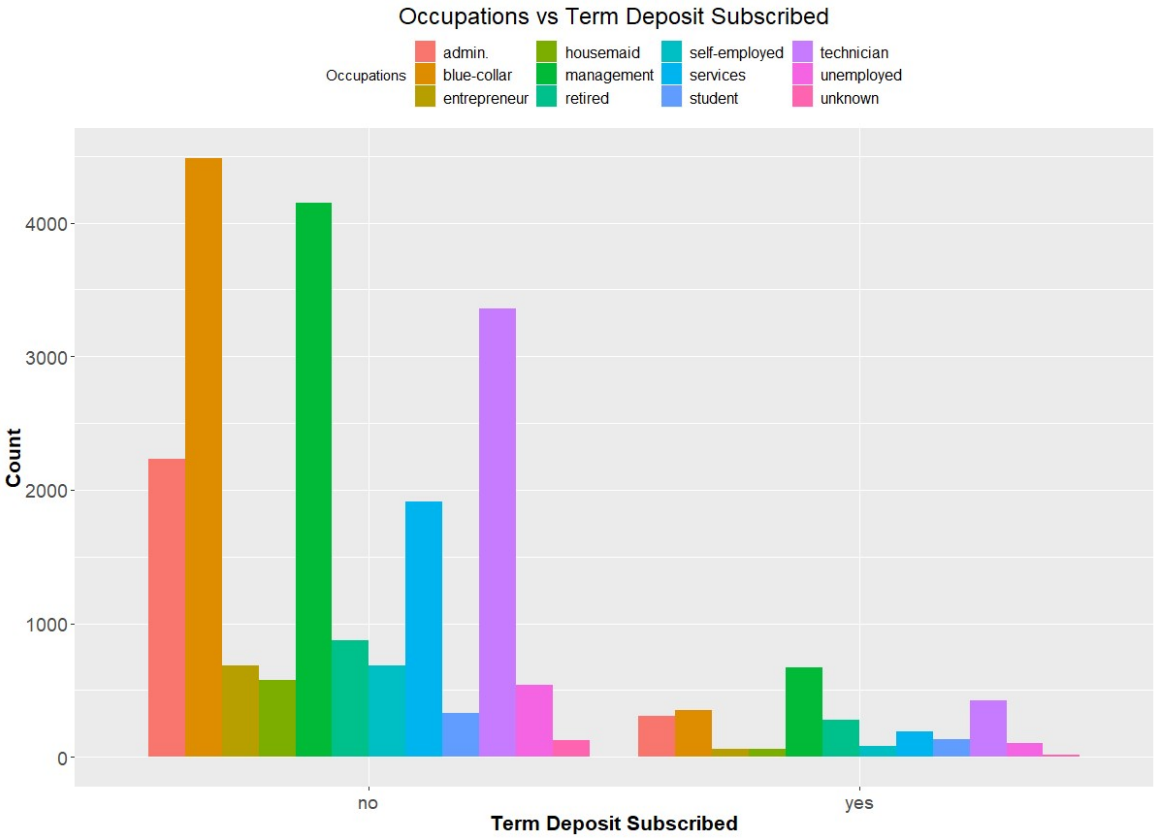
Appendix 8: Comparative Bar Graph of Age Group vs Outcome Variable



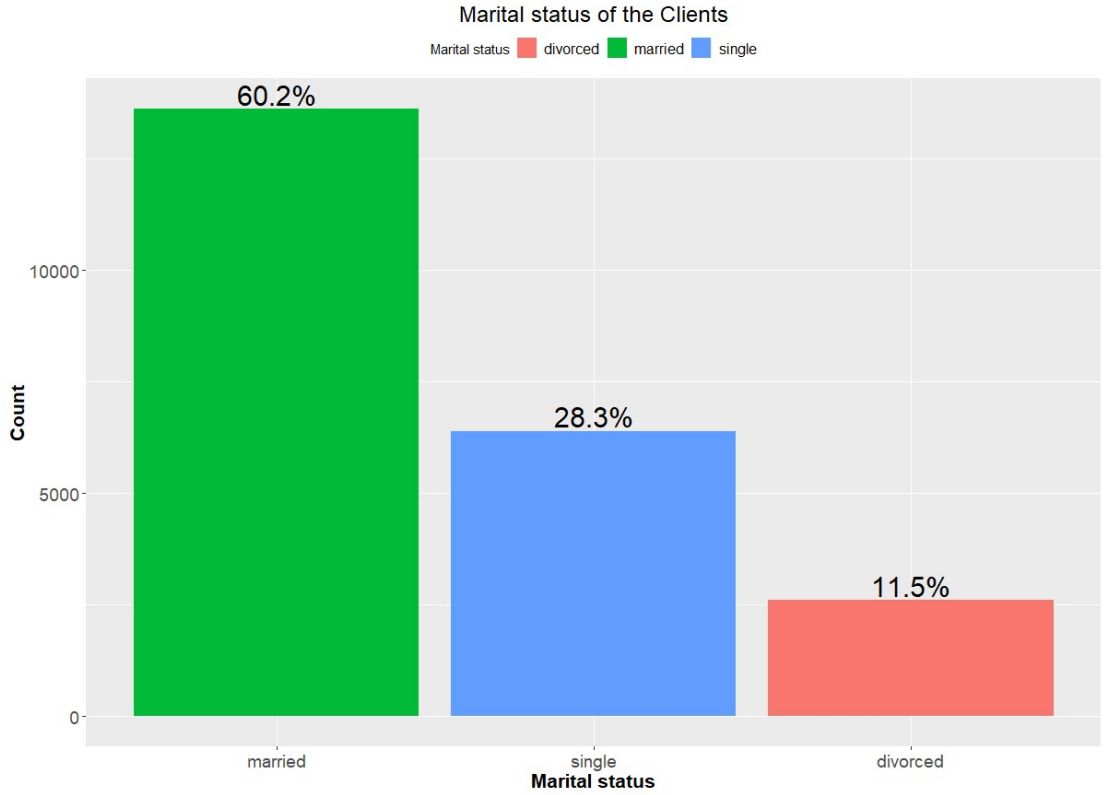
Appendix 9: Frequency plot of various occupations of the clients



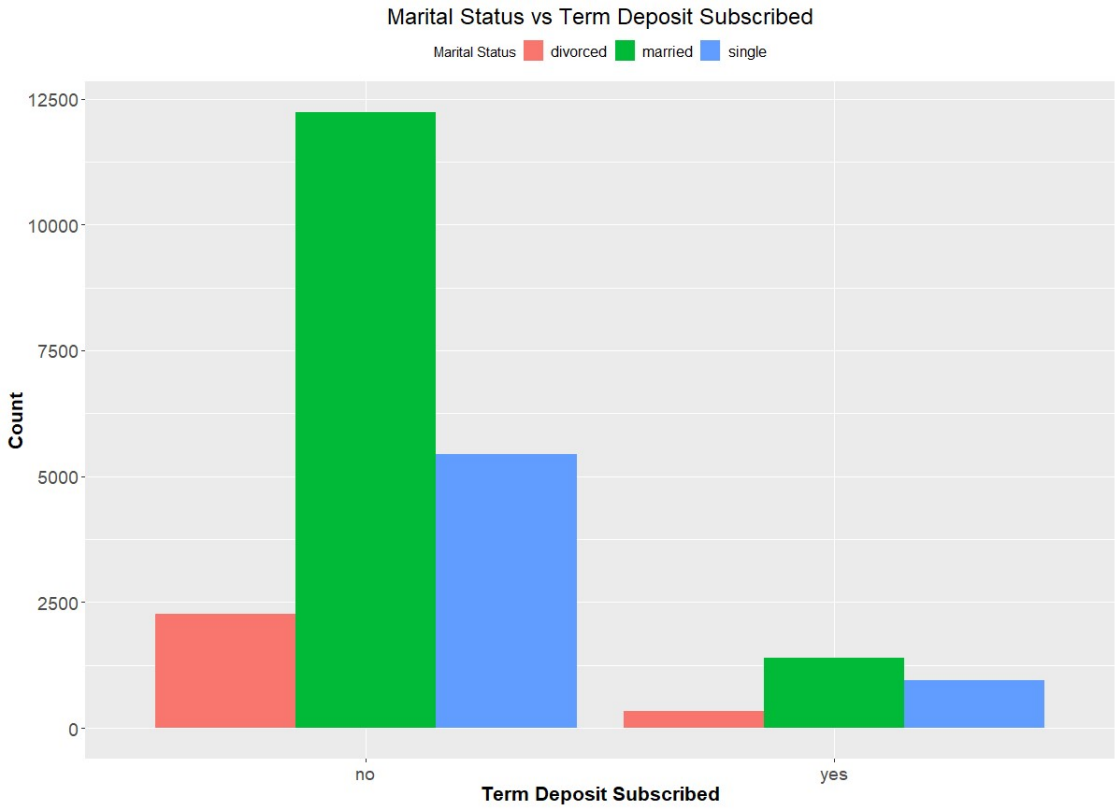
Appendix 10: Comparative Bar Graph of Occupations vs Outcome Variable (Term Deposit Sub)



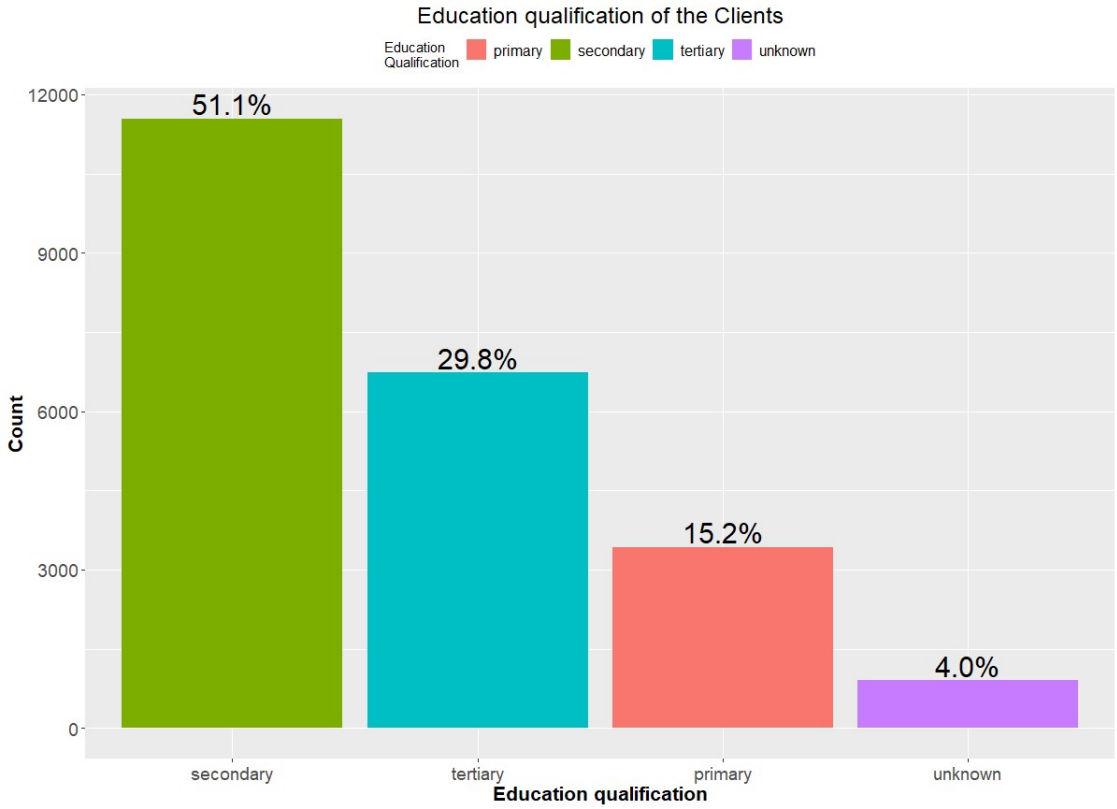
Appendix 11: Frequency plot of Marital status of the clientele



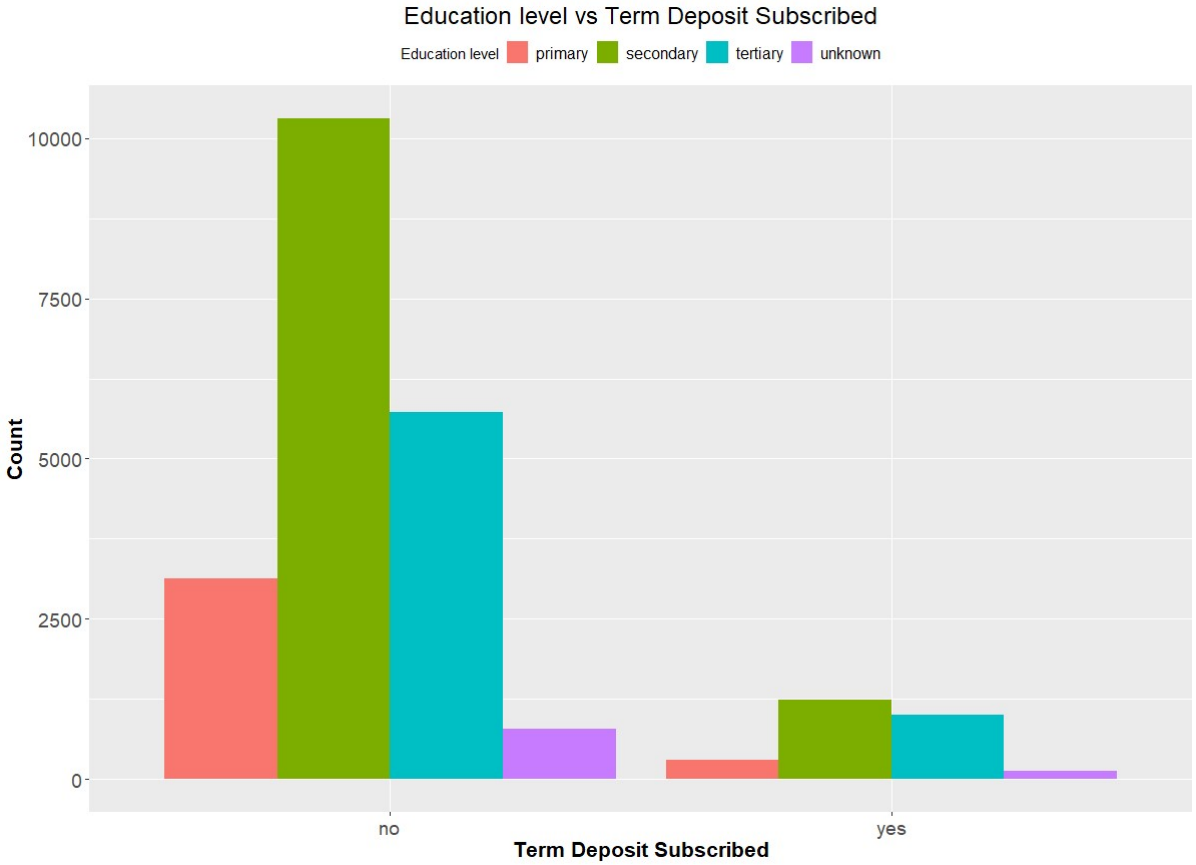
Appendix 12: Comparative Bar Graph of Marital Status vs Outcome Variable (Term Deposit Sub)



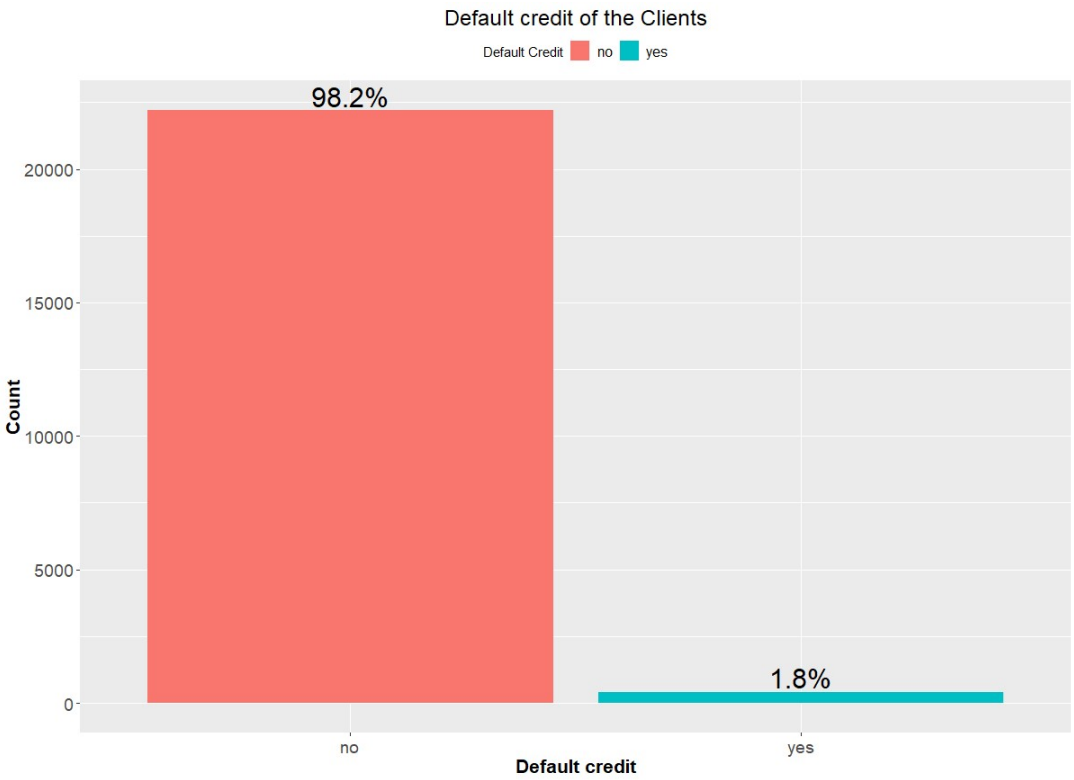
Appendix 13: Frequency plot of the Education Qualification of the Clientele



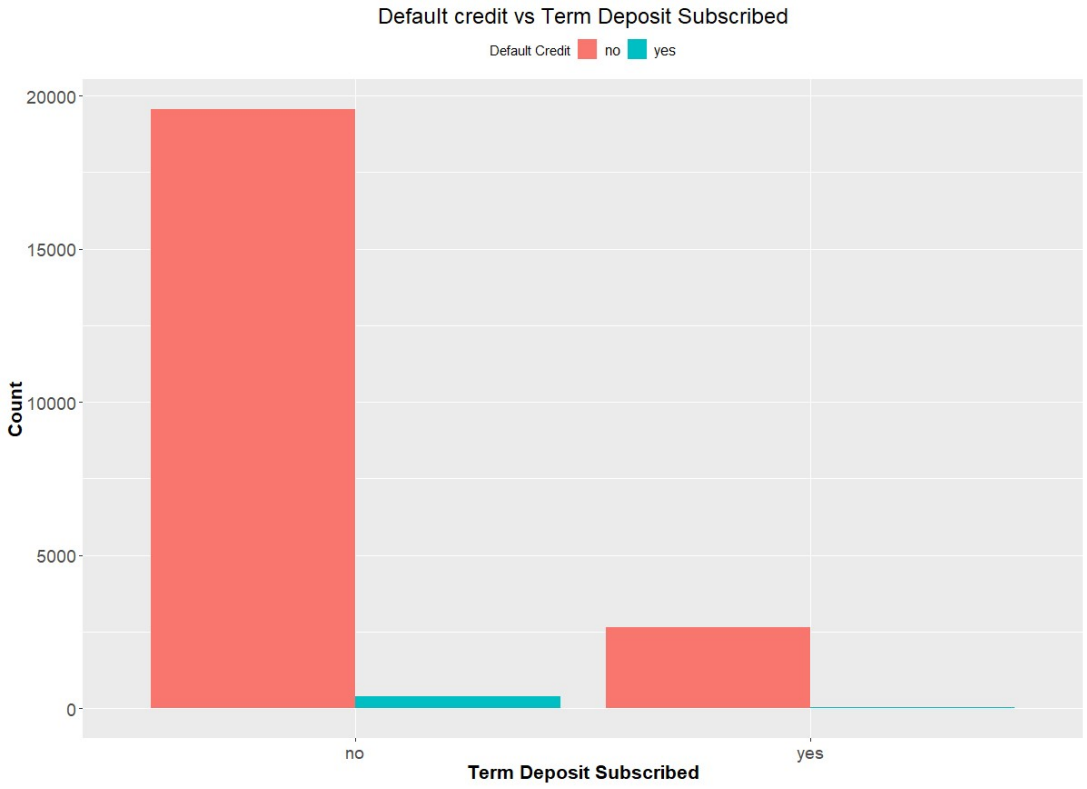
Appendix 14: Comparative Bar Chart of Education Level vs Outcome Variable (Term Deposit Sub)



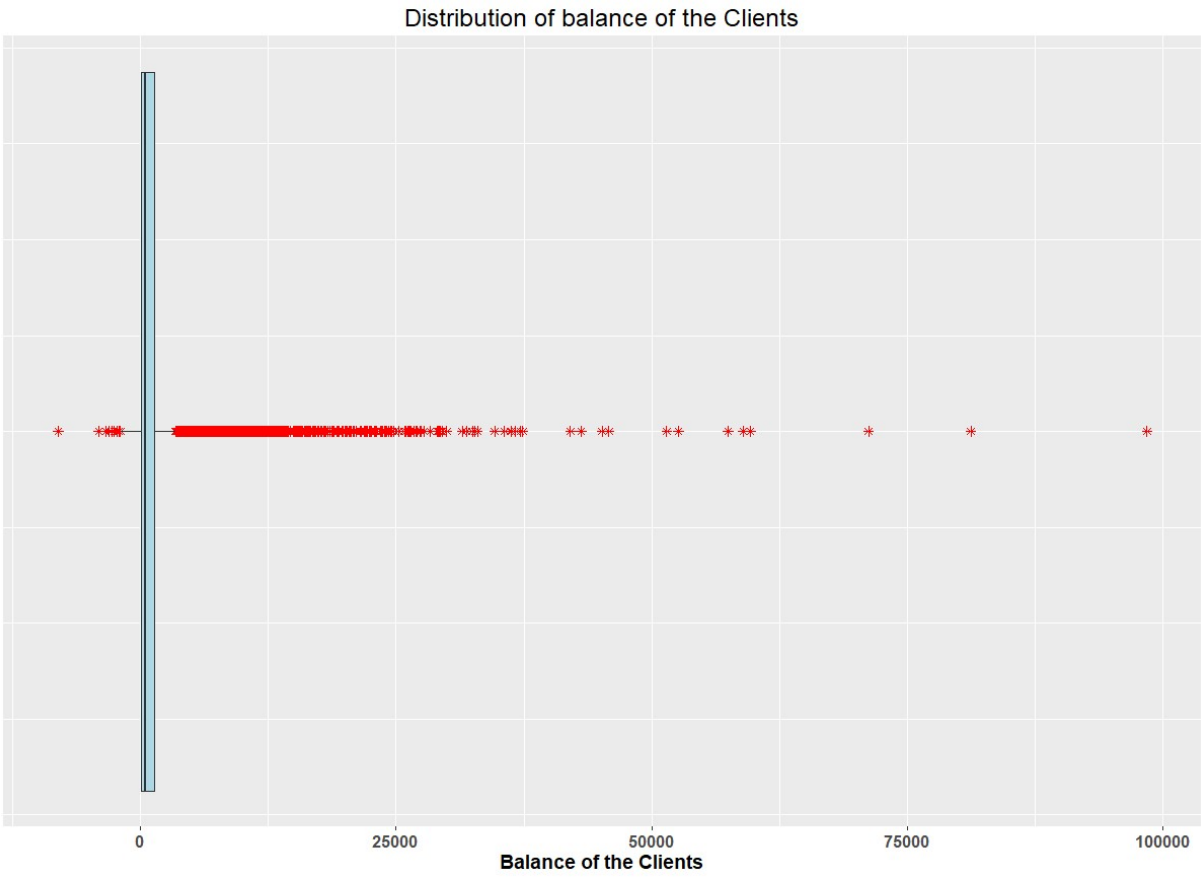
Appendix 15: Frequency plot of the Default Credit of the Clientele



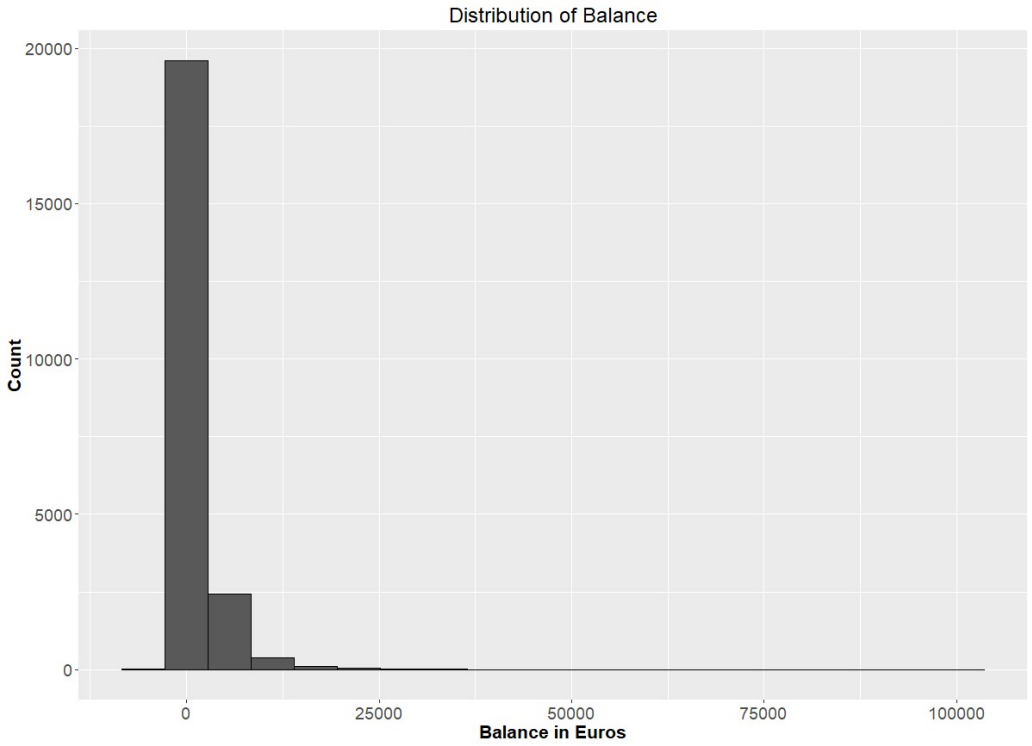
Appendix 16: Comparative Bar Graph of Default credit vs Outcome Variable (Term Deposit Sub)



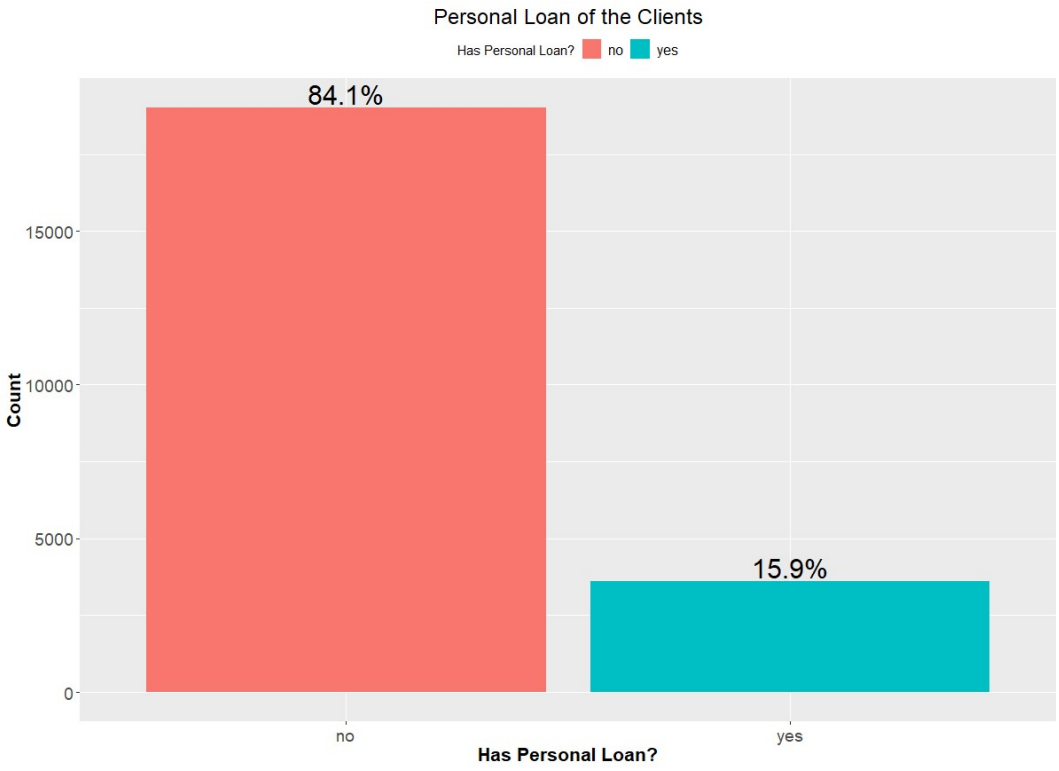
Appendix 17: Boxplot - Distribution of balance of the Clientele



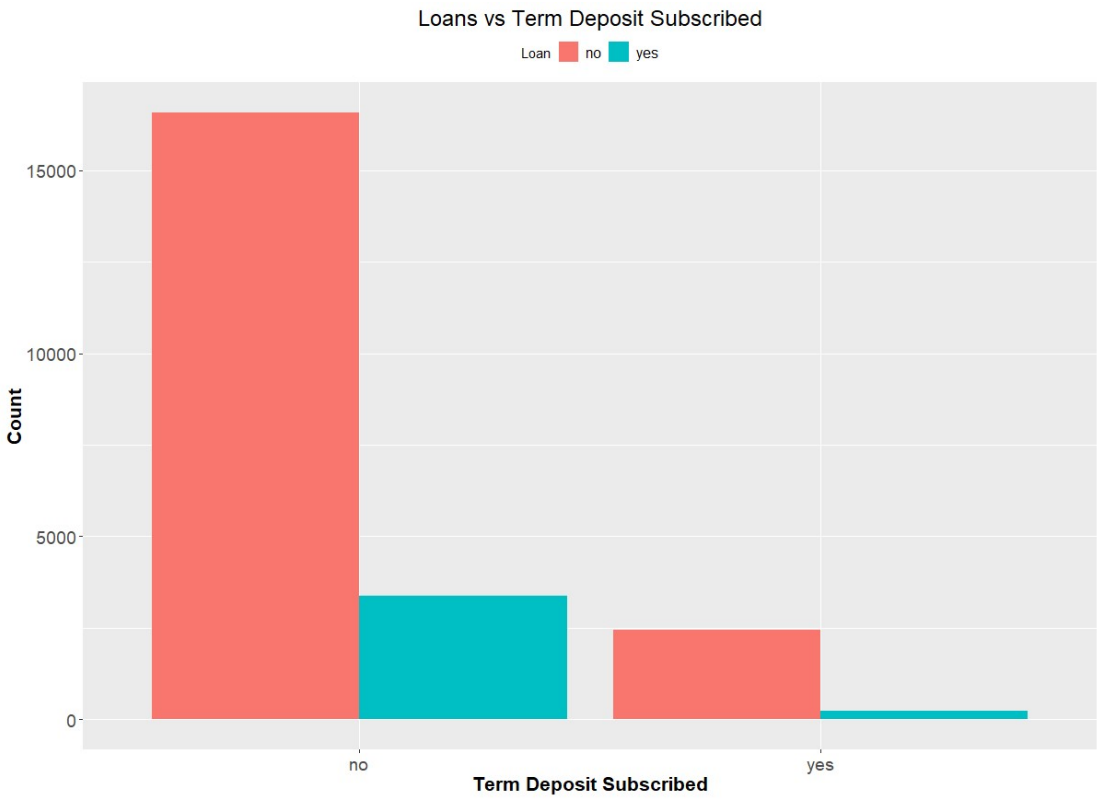
Appendix 18: Histogram - Distribution of balance of the Clientele



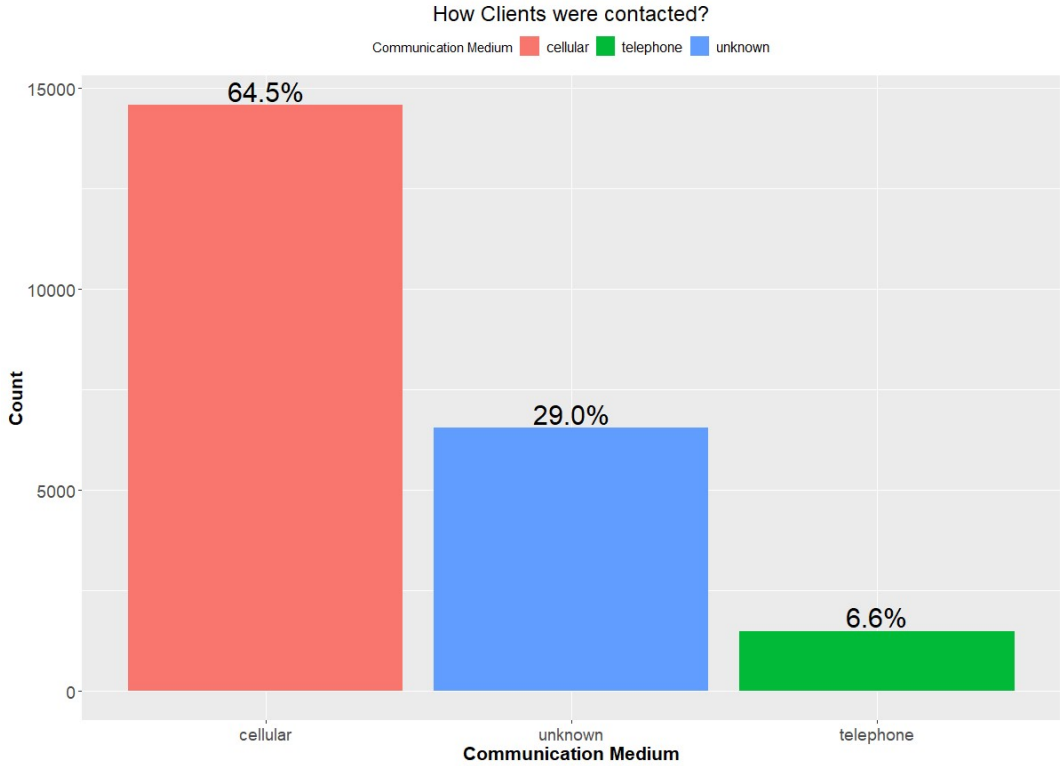
Appendix 19: Frequency plot of the Default Credit of the Clientele



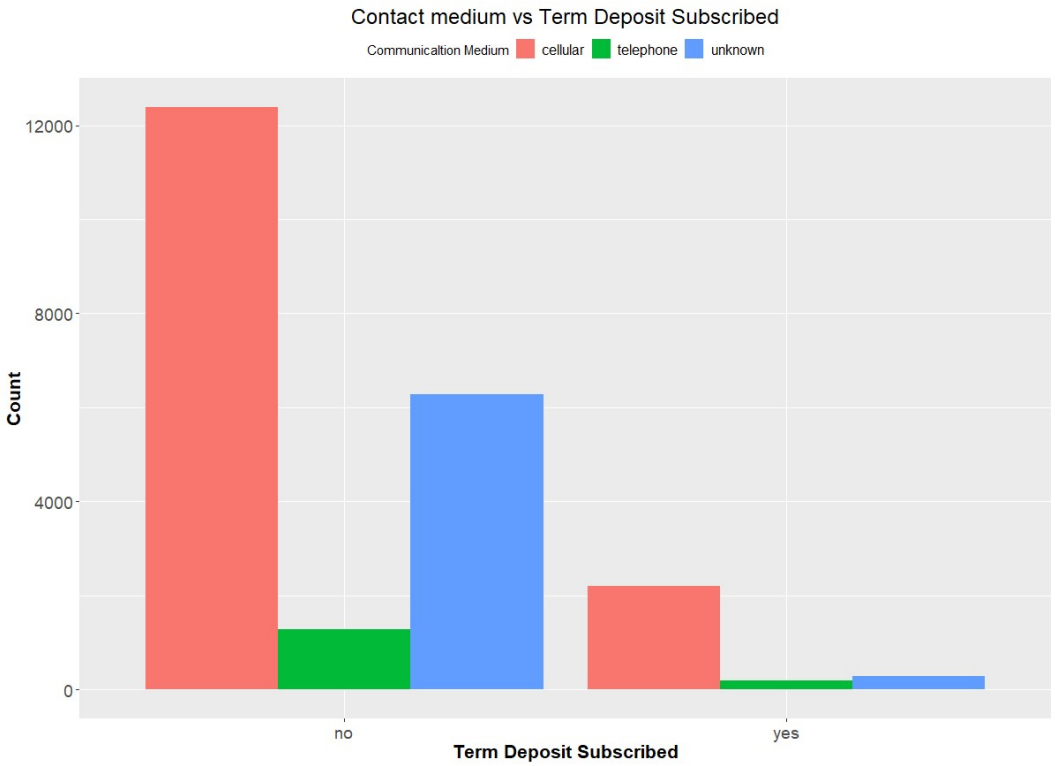
Appendix 20: Comparative Bar Graph of Loans vs Outcome Variable (Term Deposit Sub)



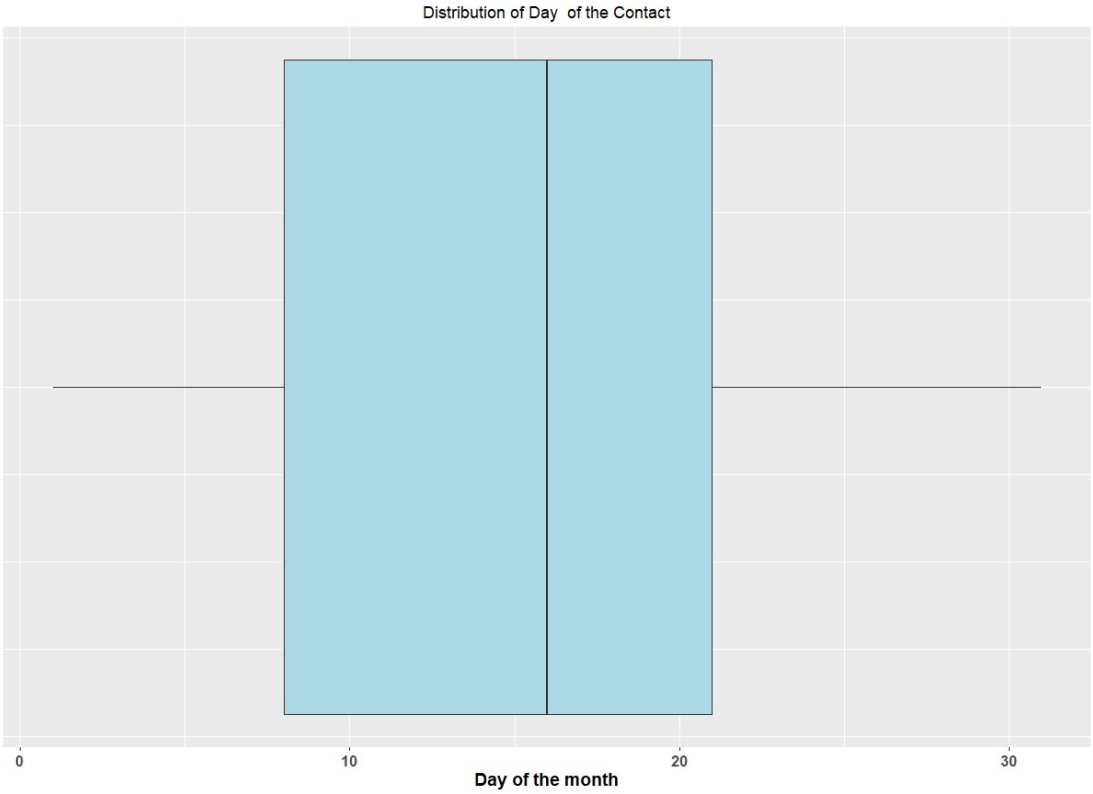
Appendix 21: Frequency Plot of Contact



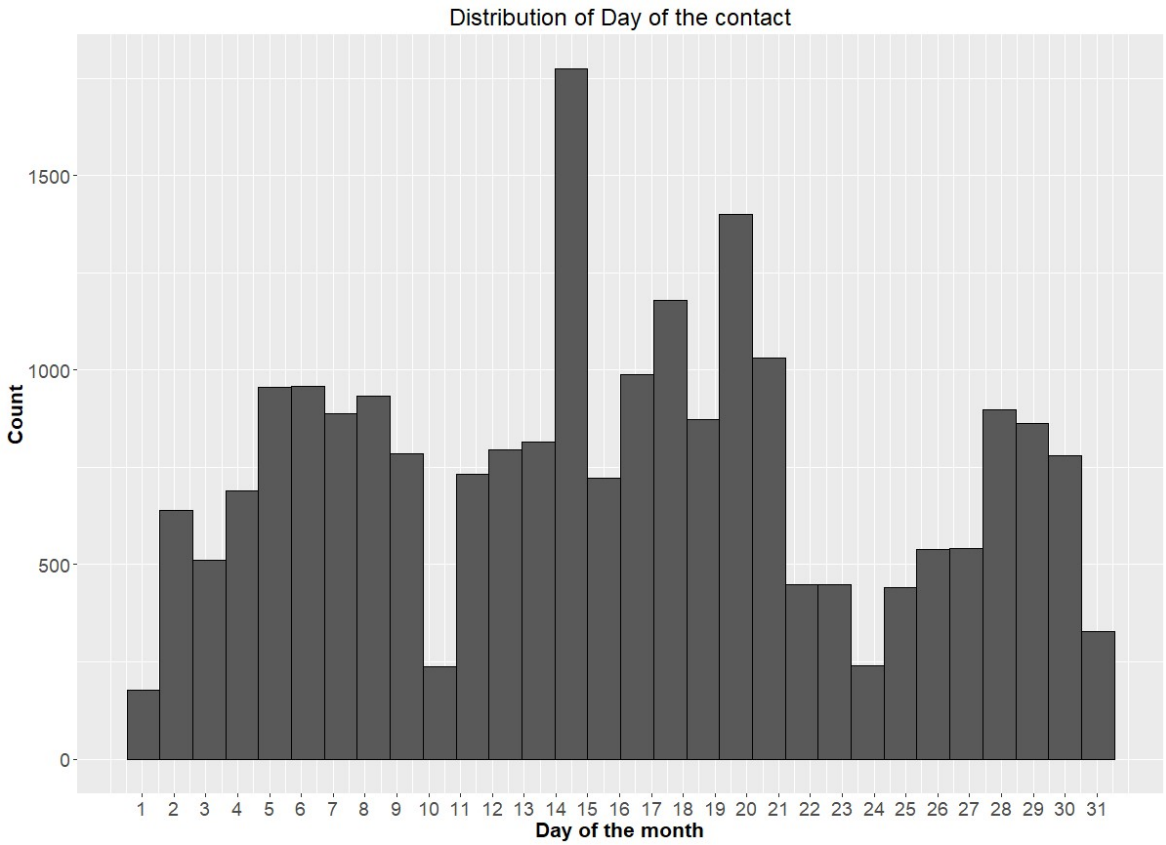
Appendix 22: Comparative Bar Graph of medium of contact vs Term Deposit Sub



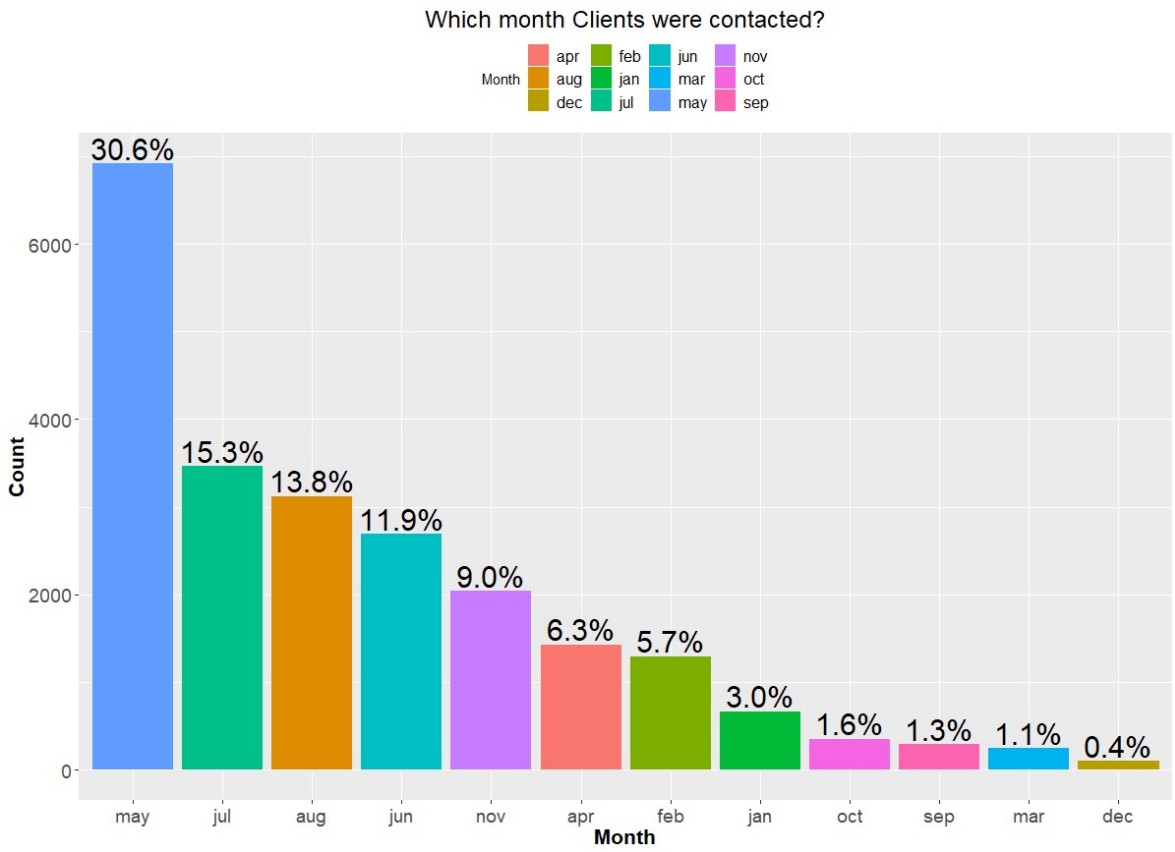
Appendix 23: Boxplot – Distribution of the day of contact



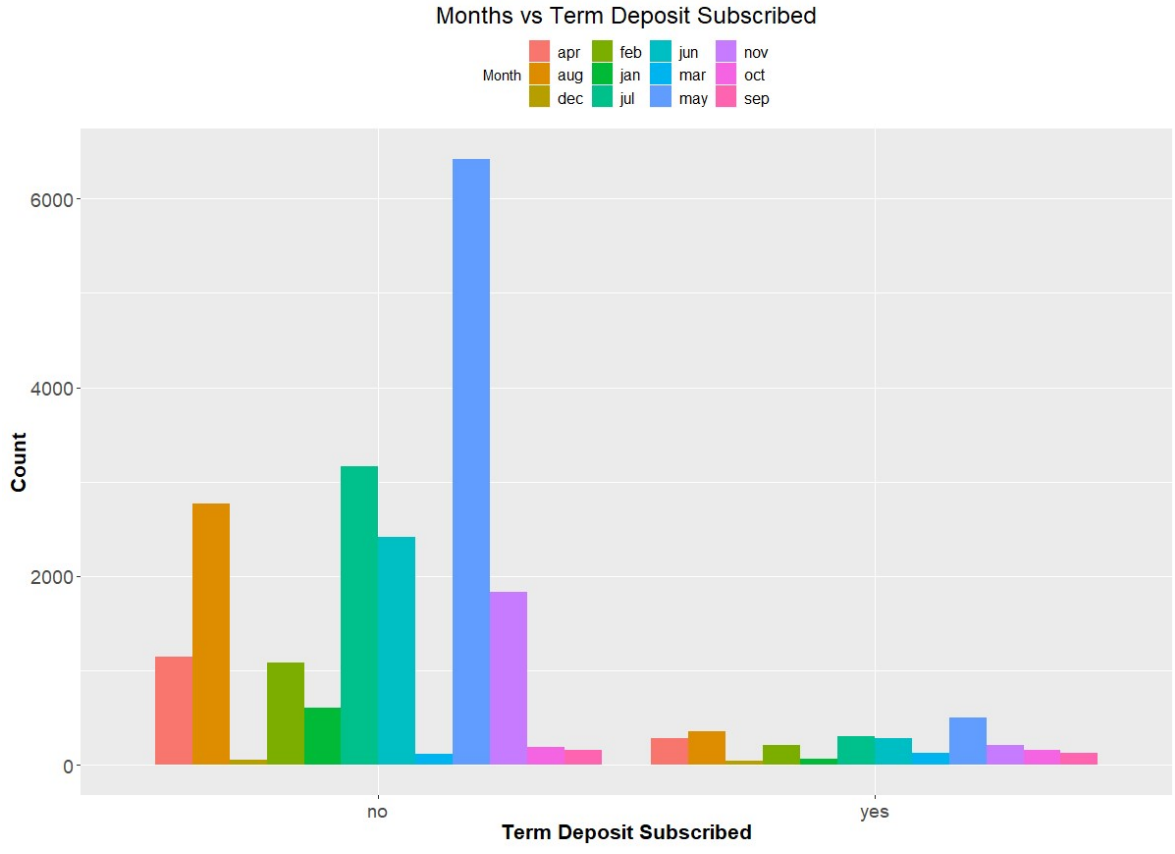
Appendix 24: Histogram – Distribution of the day of contact



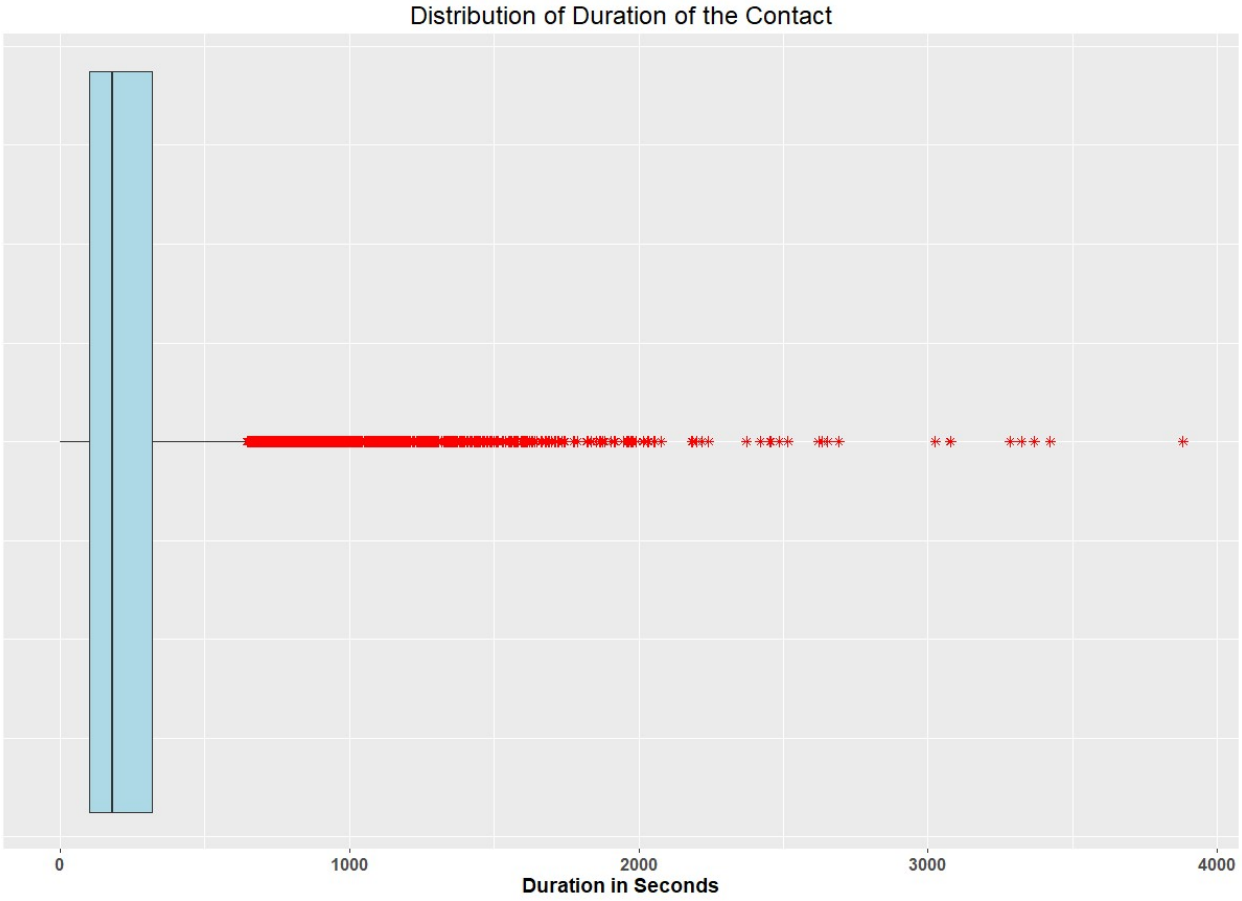
Appendix 25: Frequency Plot of Month of Contact



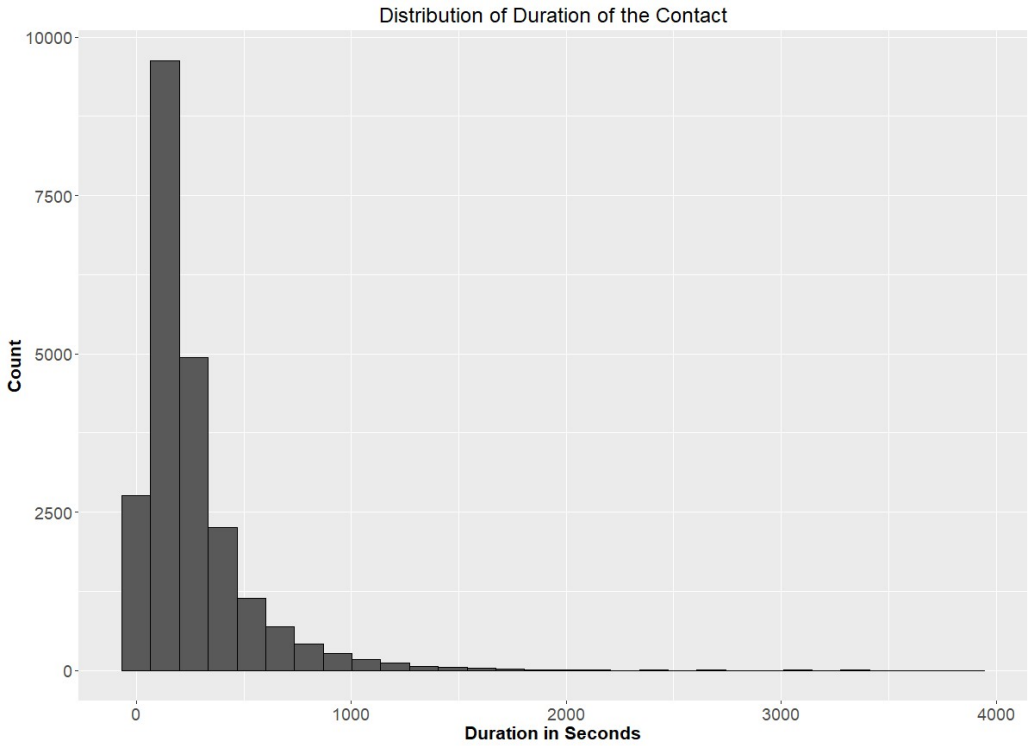
Appendix 26: Comparative Bar Graph of month vs Term Deposit Sub



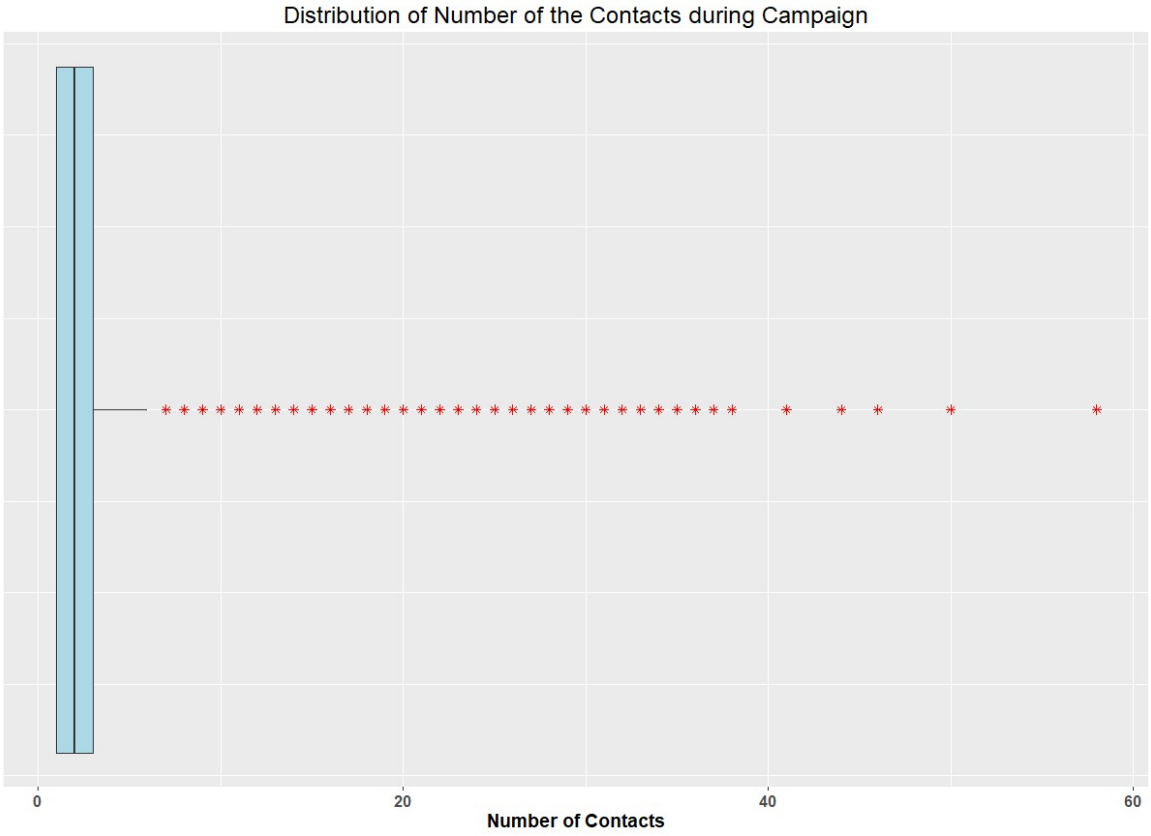
Appendix 27: Boxplot – Distribution of the Duration of contact



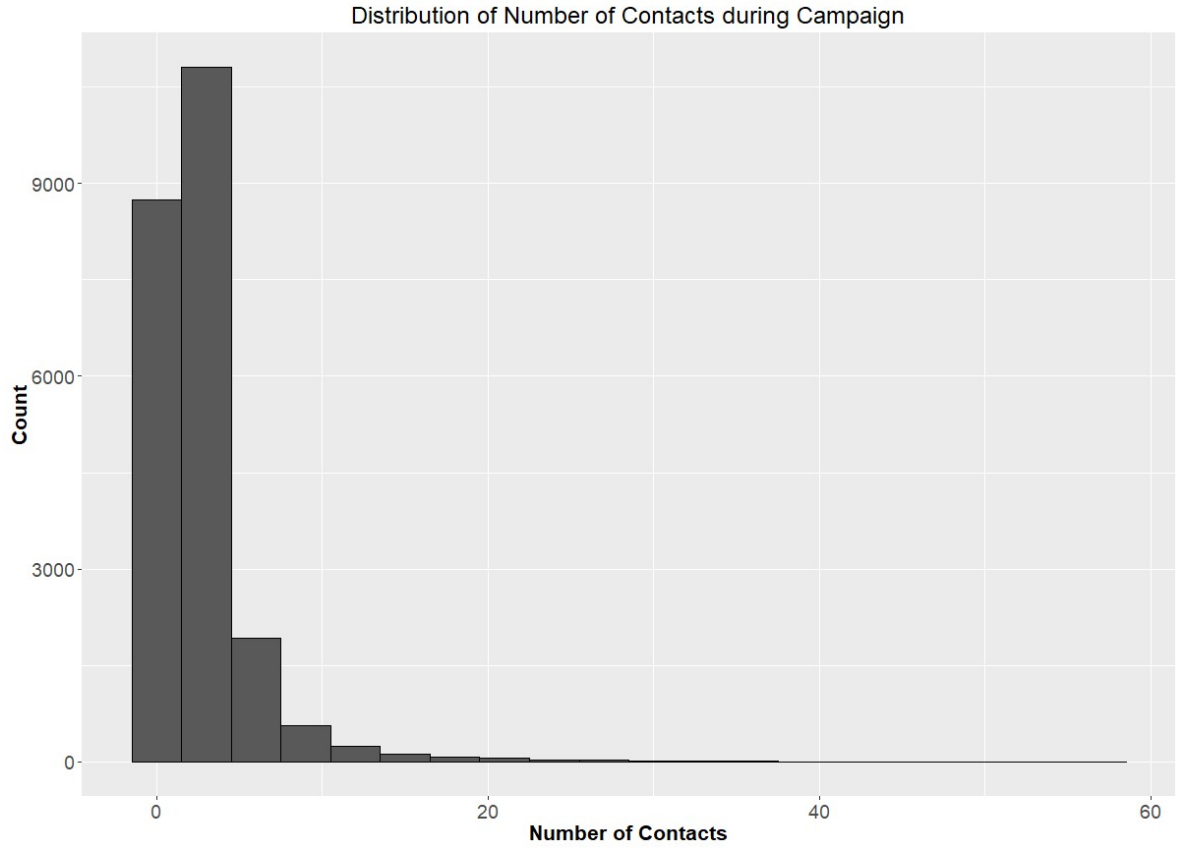
Appendix 28: Histogram– Distribution of the Duration of contact



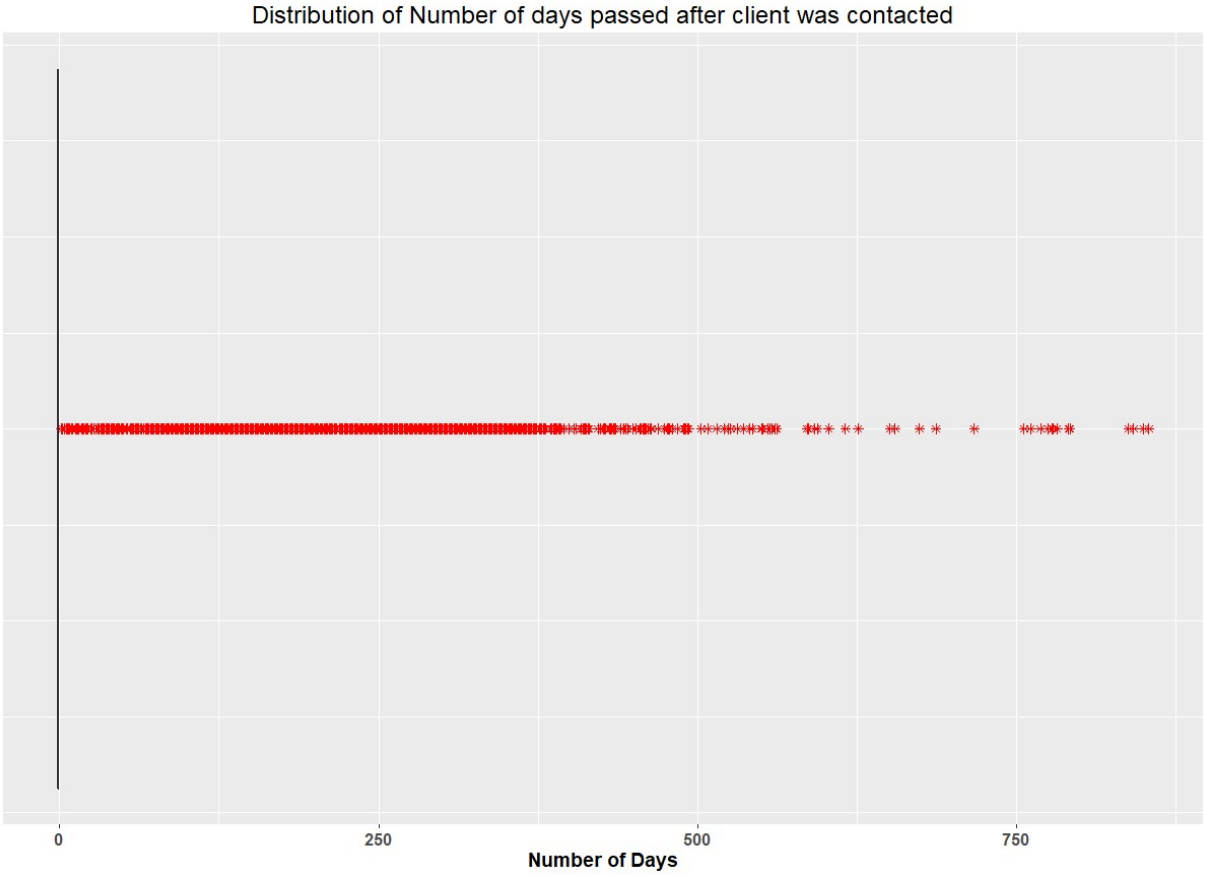
Appendix 29: Boxplot – Distribution of number of contacts during campaign



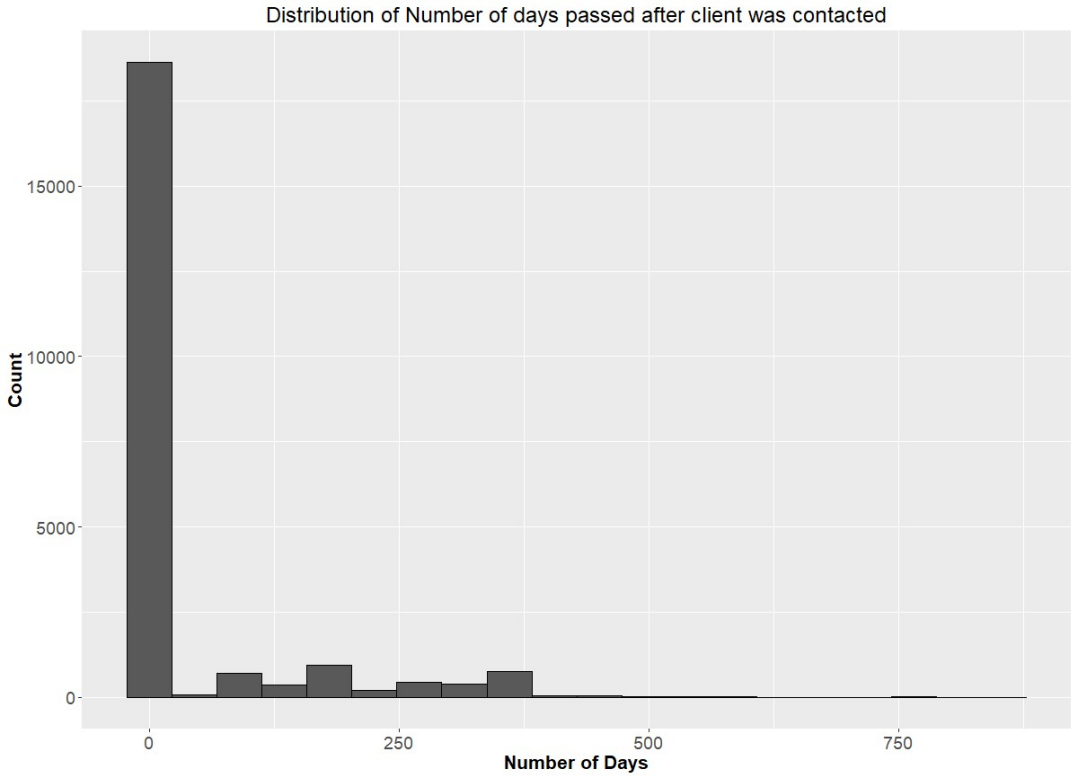
Appendix 30: Histogram – Distribution of number of contacts during campaign



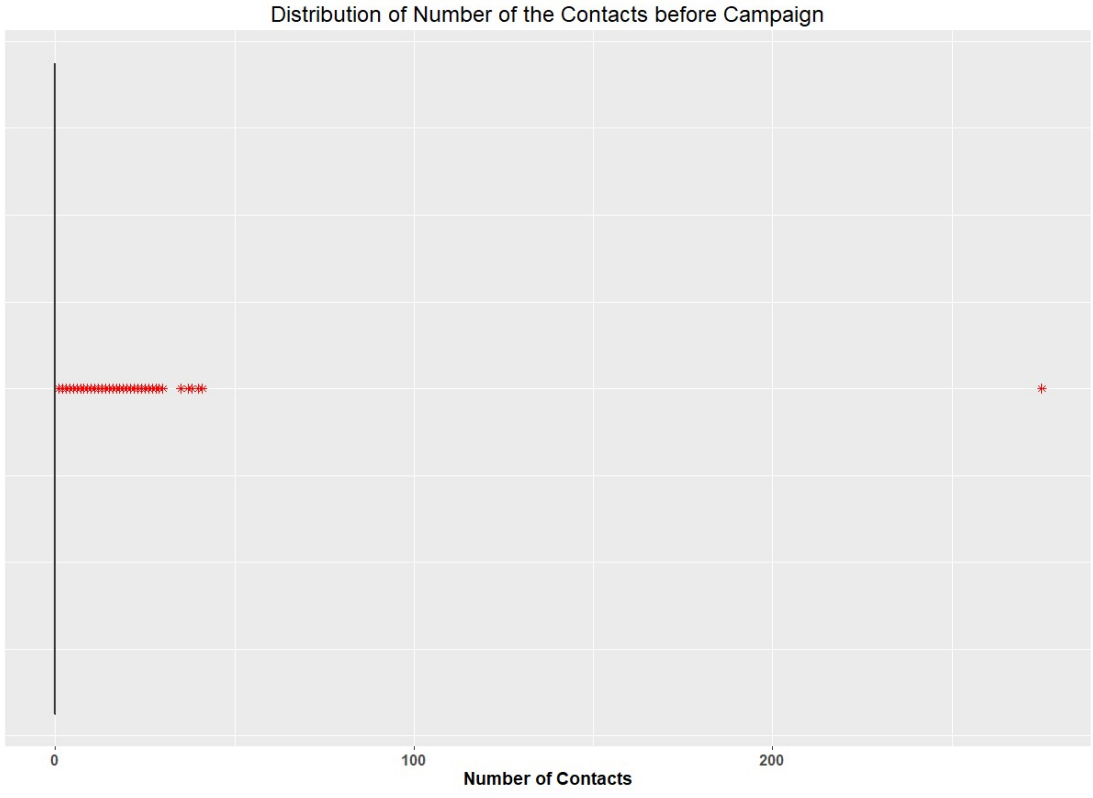
Appendix 31: Boxplot – Distribution of number of days passed after last contact



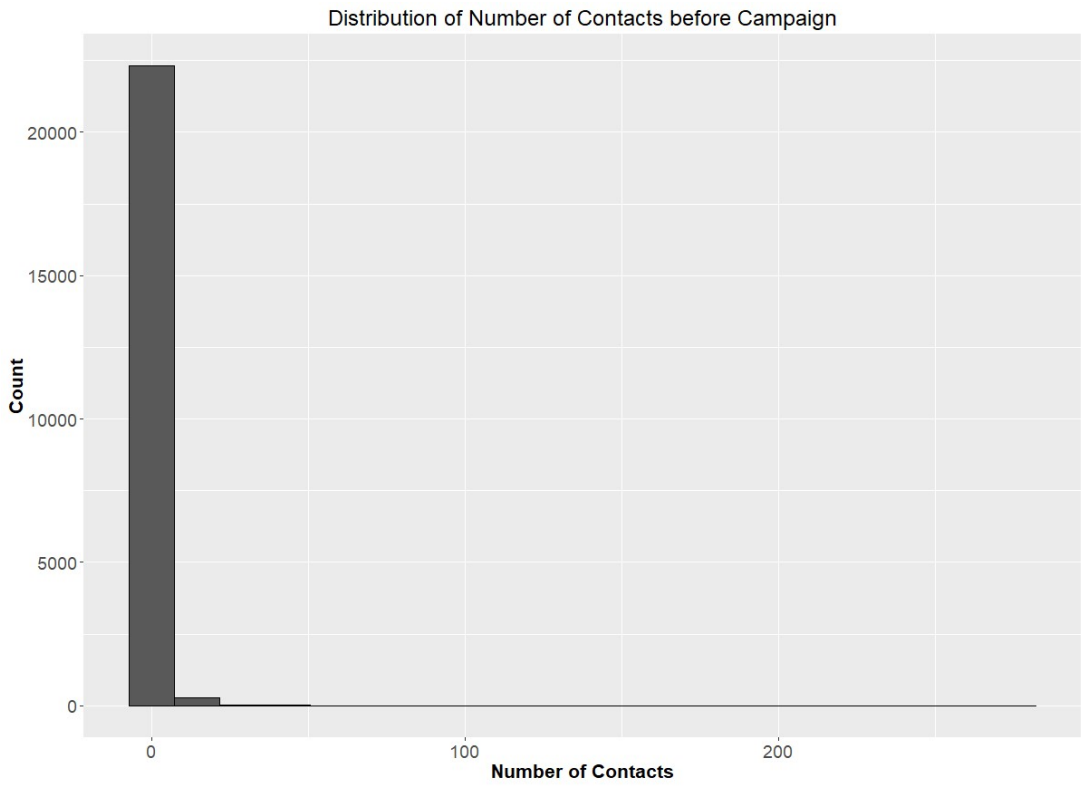
Appendix 32: Histogram – Distribution of number of days passed after last contact



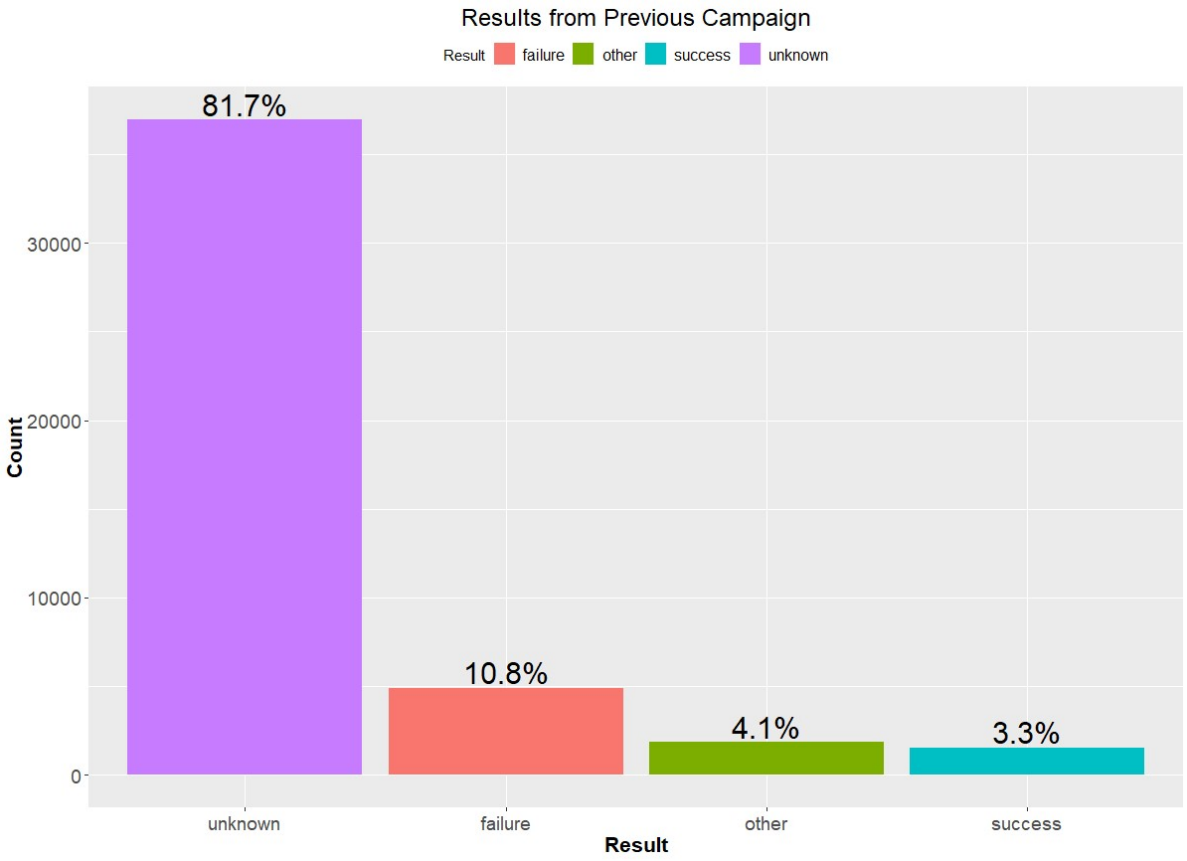
Appendix 33: Boxplot – Distribution of number of contacts before campaign



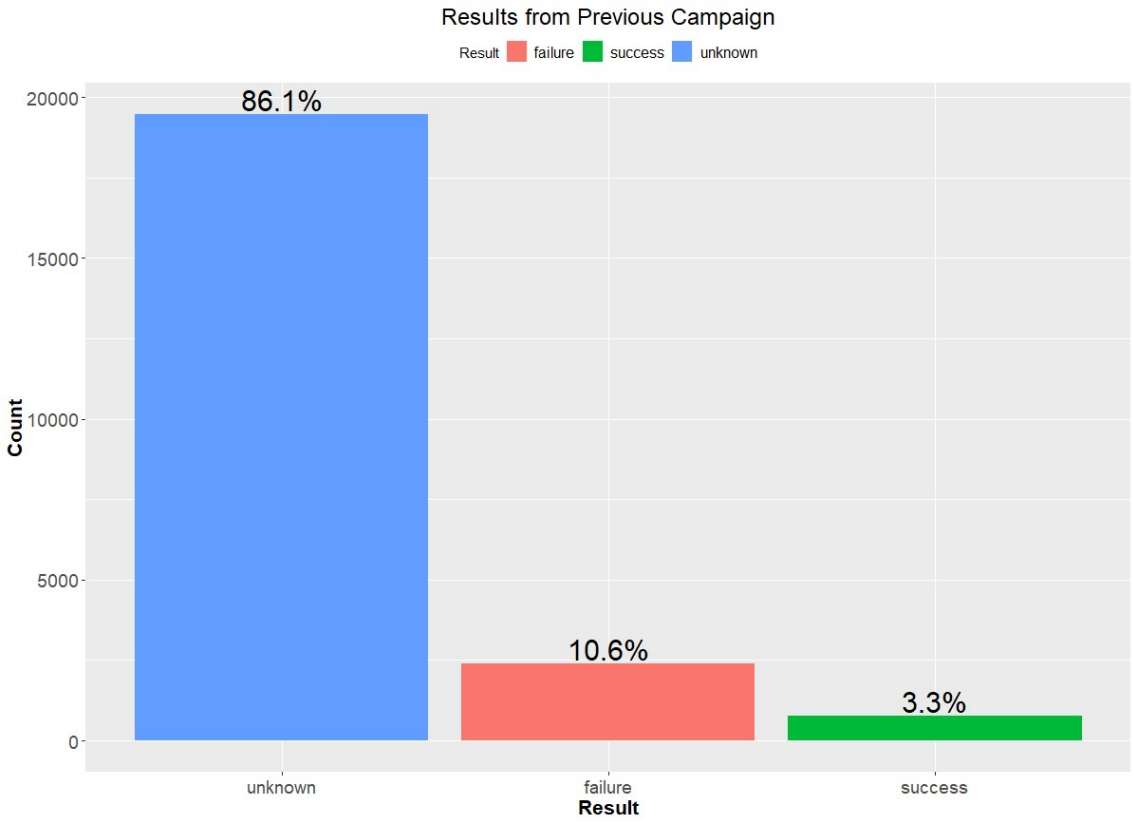
Appendix 34: Histogram – Distribution of number of contacts before campaign



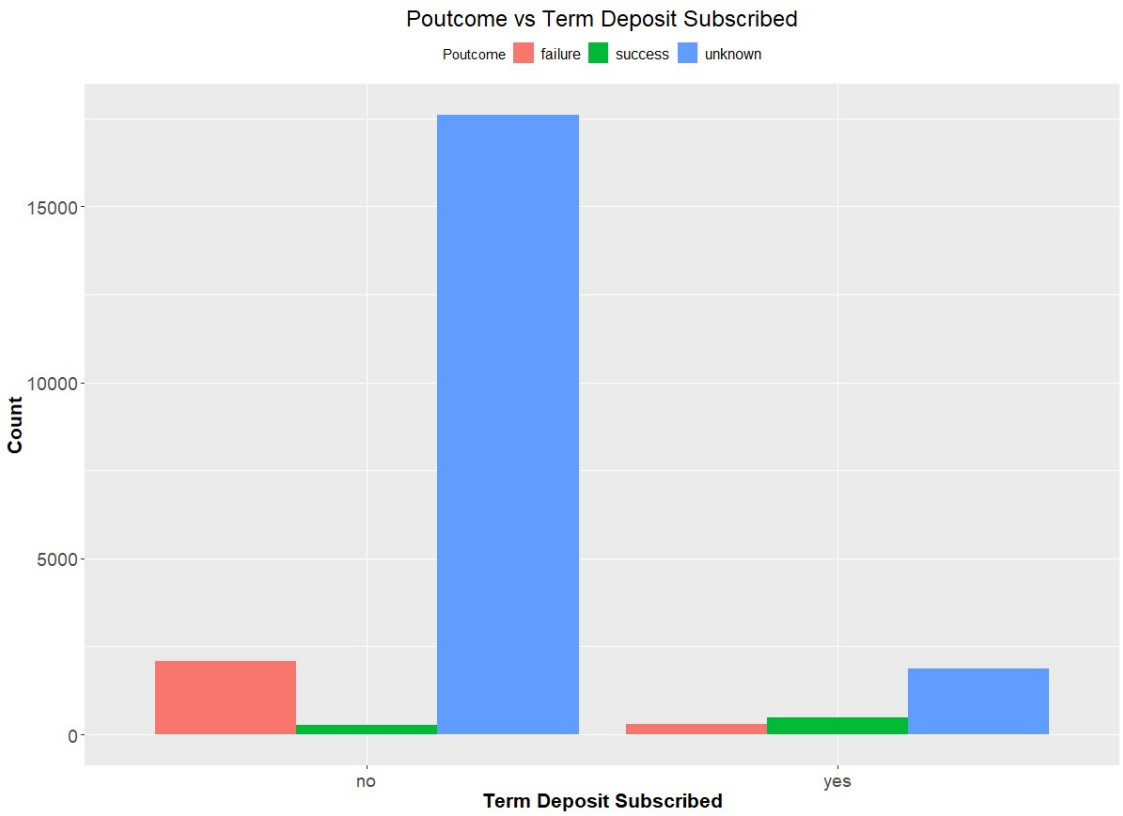
Appendix 34: Frequency plot of results from previous campaign (Unknown + Other)



Appendix 35: Frequency plot of results from previous campaign



Appendix 36: Comparative Bar Graph of poutcome vs Term Deposit Sub


Poutcome vs Term Deposit Subscribed
Poutcome
failure
success
unknown
15000
10000
5000
0
Count
no
yes
Term Deposit Subscribed