MGT 6203 Group Project

Classification of Portuguese Bank Telemarketing Campaign Outcome on Long-Term Deposits

Final Report

Team 39:

Mangesh Madhavrao Pohanerkar

Mayur Balchandra Uttarwar

Shameema Shahul Hameed

Paul Kim

Lawrence Eun

Contents

Final Report 1
Introduction 3
Data Exploration 4
APPROACH/METHODOLOGY 8
CONCLUSION 19
APPEXDIX 20

Introduction

This document covers the progress of the project detailed below.

Problem:

Marketing selling campaigns such as telemarketing have been implemented by various organizations to attract more customers to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. To maximize the profits earned by the banks, it is ideal for the institutions to invest fewer resources in the campaign and sell the maximum number of term deposits by targeting the customers who are most likely to subscribe.

Background:

Marketing selling campaigns constitute a typical strategy to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. Centralizing customer remote interactions in a contact center eases the operational management of campaigns. Such centers allow communication with customers through various channels, telephone (fixed-line or mobile) being one of the most widely used.

Objectives:

The main objective of the project titled "Classification of Portuguese Bank Telemarketing Campaign Outcome on Long-Term Deposits" is to determine how to classify the prospective customers to those who will subscribe and those who will not subscribe for the long-term deposits. Our project focuses on analyzing the outcome of the telemarketing phone calls made by a Portuguese bank as part of their marketing campaign to sell long-term deposits to the clients. The result of the campaign is a binary unsuccessful or successful contact made with the clients, where subscription of the client in the bank's long-term deposit is considered as a successful outcome. If the contact made with the client does not result in the subscription of the client for a long-term deposit, then the outcome is unsuccessful.

The data was collected from a Portuguese retail bank, from May 2008 to June 2013, in a total of 52,944 phone contacts will be used in the proposed project.

We also leveraged supplemental data "Bank_personal_loan_Modelling" to analyze and determine the attributes having significant relationship with personal loan status of the

customers. These attributes could be used by the bank to decide whether to offer a personal loan to the customers or not.

Business Justification:

Telemarketing has been implemented by various organizations in the following ways, to enhance business outcomes:

1. Term deposit accounts provide banks with the cash flow they need to lend money to other customers.
2. The bank makes a profit by lending the funds held in term deposit accounts for a higher interest rate than the rate it pays on the term deposits.
3. The bank may also invest the money from the term deposits in other securities that pay a higher return than it is paying the customers.

It is ideal to target customers who are most likely to apply for term deposits for financial institutions to invest less resources in campaigns, sell the maximum number of term deposits and maximize bank profits. The classification task we want to perform in this project will help financial institutions identify ideal customers for making a profit.

Data Exploration

The project uses the following datasets:

1. bank-additional.zip from https://archive.ics.uci.edu/ml/machine-learning-databases/00222/

   The dataset is referred to as Portuguese bank dataset in the report.

2. Bank_Personal_Load_Modelling.csv from https://www.kaggle.com/code/somnathpathak/bank-personal-loan-modelling-supervised-learning/notebook

   The second dataset is a Bank's Personal Loan dataset

Dataset 1 - Portuguese Bank Telemarketing Campaign Dataset

Dataset size: There are 41,188 data points and 21 attributes in the Portuguese bank dataset.

Dataset Cleaning: The dataset was observed to be clean, hence no extra cleaning was performed on the dataset. In the source dataset, the categorical attributes with missing data were assigned the value of "unknown" and we decided to keep the attributes and the results unaltered for the planned analyses.

A snapshot of the attributes is shown here. More information about the attributes is provided in the appendix

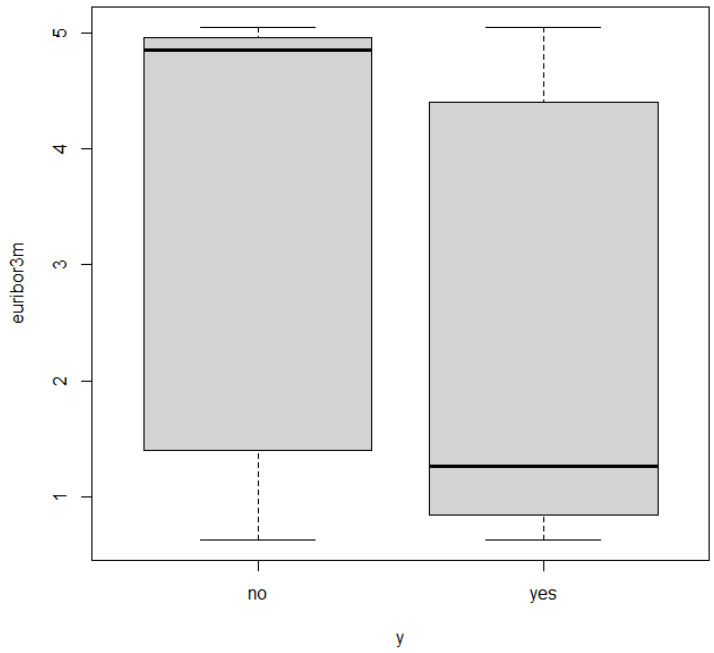
age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no

Data Exploration: The datasets were visually explored using different plots. Some of the most prominent graphs are included in this section whereas the others are included in the Appendix.

The boxplots were created to observe the distribution of the numerical attributes for the customers who enrolled into the long-term deposits versus the customers who did not.

1. The following attributes visually appeared to have somewhat different distributions between the customers who enrolled in the long-term deposits versus those who did not: Employment variation rate (emp.var.rate), Consumer price index – monthly indicator (cons.price.index), Consumer confidence index – monthly indicator (cons.conf.index), Euribor 3-month rate – daily indicator (euribor3m), Number of employees – quarterly indicator (Nr.employed), Number of contacts performed before this campaign for the given customer (previous)

   The box plot below shows that the distribution of Euribor 3-month rate is significantly different between the customers who enrolled and those who did not.

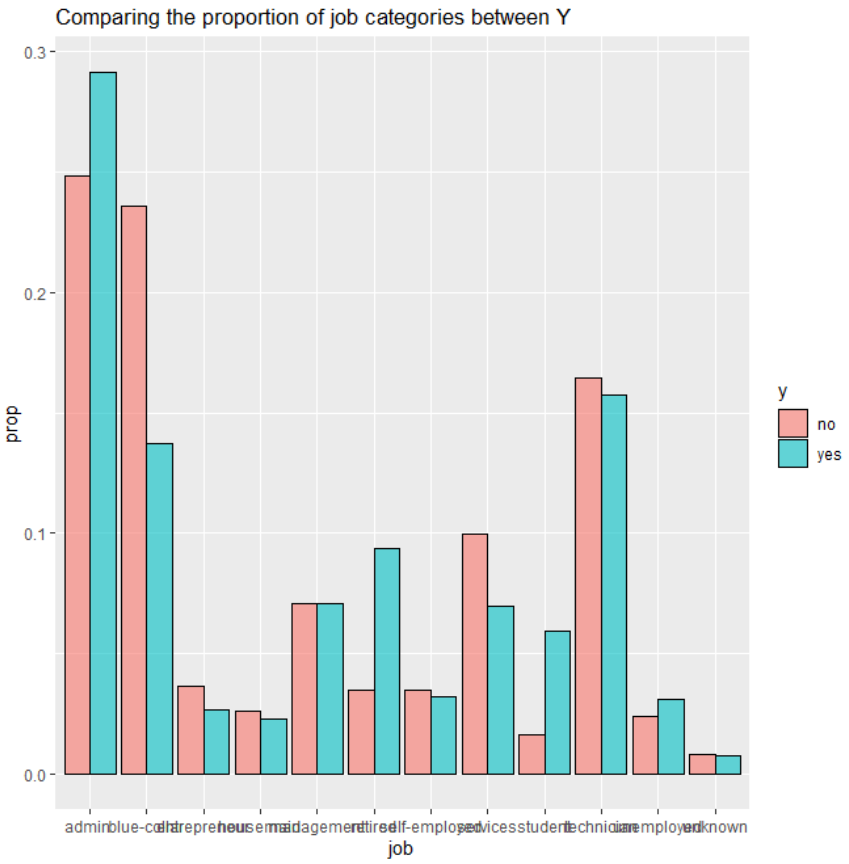


2. The attributes like, Age (age), Number of contacts performed during the campaign with the given customer (campaign), and Number of days that passed by after the client was last contacted from a previous campaign (pdays) appeared to have similar distribution for the customers who enrolled and who did not in the long-term deposits.

Histograms were created to observe the differences in the proportions of customers for the categorical attributes between the customers who enrolled versus those who did not.

1. A higher proportion of the customers who were retired (job=" retired") or the customers with administration jobs (job=" admin"), or the customers who were students (job==" student) appeared to have enrolled in the long-term deposits than other types of jobs. A higher proportion of the customers with blue-collar jobs (job=="

blue-collar") and any services (job=" services") appeared to not enroll in the long-term deposits.



2. The customers who were single (marital=" single") were observed to enroll in the long-term deposits more than the ones who were married (marital=" married") or divorced (marital=" divorced").
3. A higher proportion of customers with university degrees were observed to have enrolled in long-term deposits than other types of education.
4. Most of the customers were contacted by cellular phones (contact=" cellular"), however, the customers were also contacted by telephone (contact=" telephone"). A higher proportion of customers who were contacted by cellular phones were observed to enroll in long-term deposits.
5. The customers who previously enrolled in other services offered by the bank (poutcome="success") also appeared to enroll in the given long-term deposits. However, most of the customers were nonexistent (poutcome="nonexistent") in the bank database.
6. A higher proportion of customers who were contacted on Tuesday, Wednesday and Thursday enrolled in long-term deposits with the bank than the ones who were contacted on the first two working days of the week.
7. The proportion of customers with housing loans (housing) and Personal loan (loans) was similar for both types of campaign outcome (i.e., those who enrolled and those who did not). The largest proportion of customers who did not enroll in the long-term deposits

were contacted in May (month=" May"), but it was also observed that most of the customers were contacted in May.

Dataset #2: Bank Personal Loan

Dataset size: The second dataset includes customer information, and if a customer has taken a loan or not. It has 5000 rows and 13 features and 1 target variable (loan). A snapshot of the dataset is provided below.

Age	Experience	Income	ZIP.Code	Family	CCAvg	Education	Mortgage	Personal.Loan	Securities.Account	CD.Account	Online	CreditCard
56	30	73	94035	2	1.1	1	0	0	0	0	0	0
60	35	153	95136	3	2	3	0	1	0	0	0	1
33	3	59	91040	2	1.75	3	0	0	0	0	1	0
31	6	62	95630	1	1	1	0	0	1	0	1	0
28	4	101	95136	3	2.5	1	270	0	0	0	0	0

Dataset Cleaning: No explicit cleaning was required.

Data Exploration: The distributions of the categorical variables as seen in the histograms tend to suggest that it is a fair sampling. Such as income, where we have much fewer higher incomes than medium to lower income, which is expected of a random sampling of the population.

The following plot shows a heatmap of the correlation between variables. The heatmap shows that there are not necessarily any features very strongly correlated except the obvious (such as Age and Experience). Some interesting things to note are Income - Personal Loan, and Income and CCAvg (average spending on Credit Card per month). Both these pairings have some type of correlation. This is worth noting as Personal Loan is our target variable.

The plot showing Principal Components which can group certain features into a dimension where it captures as much variation between the features and the target variable (Personal Loan) was created. The graph suggested that Principal Component #1 and #2 could explain up to 30% of the variability combined.

Examining Principal Component #1 and 2, we saw that it was composed of obvious features such as Income, Experience, Mortgage, Education, CC_Avg, and Age in predicting Personal Loan. By using this, it could be possible to reduce the features to only the ones that give the most variability or signal.

APPROACH/METHODOLOGY

The Portuguese bank telemarking dataset is randomly divided into Training, Validation and Testing datasets in the ratio **60%: 20%: 20%**. The distribution of the customers who enrolled and who didn't in all three datasets is given below:

Dataset	Customers who Enrolled	Customers who Didn't Enroll
Training	11.27 %	88.73 %
Validation	11.17 %	88.83 %
Testing	11.34 %	88.66%

The distribution of customers was equitable in all three datasets, assuring that the models that were built using the training dataset would be good enough for prediction in the testing dataset.

The Bank_personal_loan_Modelling dataset is also randomized into Training, Validation, and Testing datasets in ratio 60%:20%:20%. The distribution of the customers who accepted the loan (Buyer) and who didn't (Rejector) in all three datasets is given below:

Dataset	Personal Loan Buyer	Personal Loan Rejector
Training	10.03 %	89.97 %
Validation	09.00 %	91.00 %
Testing	08.90 %	91.10%

The following 5 classification models are selected to classify the prospective customers who subscribe to long-term deposits and those who do not subscribe to long-term deposits. The details of the approach for each model are given below.

1. Logistic Regression Model

What is the model about: One of the models we used for classification is Logistic Regression Model – which is a supervised learning algorithm to predict the response variable. We also used the stepwise regression model which is an iterative regression model of identifying the significant independent variables that have a substantial impact on the dependent variable. Binary logistic regression is used in this project as the response variable has only two values Yes/No for the first dataset and 0/1 for the second dataset.

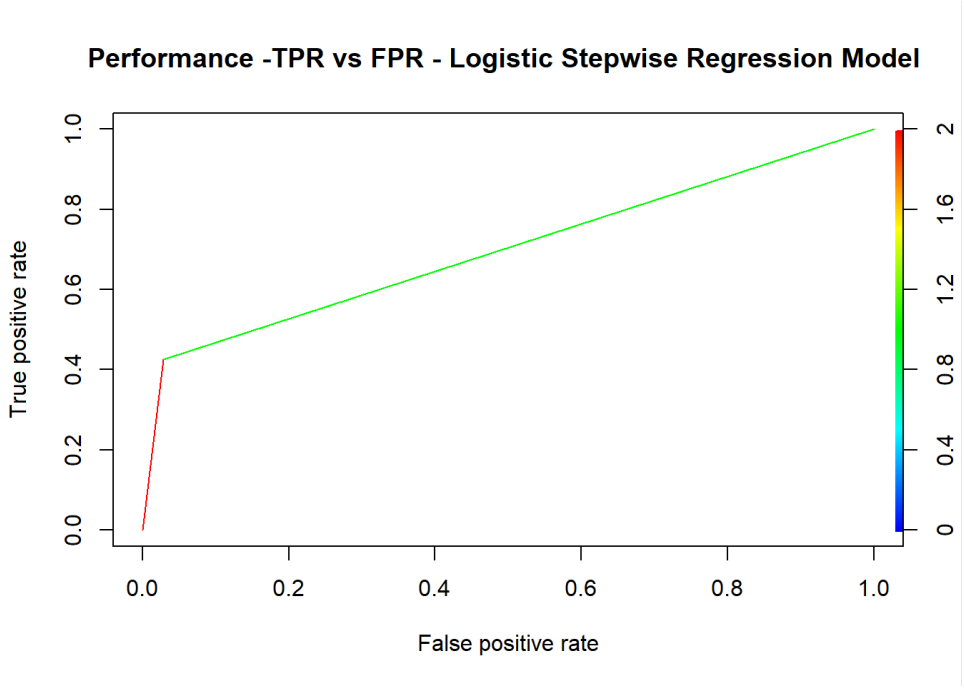
How the model was used: **Dataset 1:** The first logistic regression model is built using the training dataset with all 20 independent variables and y as the dependent variable. The resulting model indicated that two or more of the independent variables had perfect multi-collinearity and thus, not all variables could be used in the model for prediction. Also, the predictions based on a model from a rank-deficient fit could be misleading. The relevant warning obtained from the model fit could have been caused due to 2 reasons; multicollinearity or the number of variables being greater than the number of observations. In our case, the datapoints exceeded the number of variables by many folds and hence the warning indicated that there was a multicollinearity among 20 independent variables. Further investigation using the "alias" method indicated that there was a strong correlation between the dummy variables "housing unknown" and "loan unknown" (where "unknown" being one of the categories of the results for the variables "housing" and "loan" respectively). So, a second logistic regression model was built by removing the variable "loan". The second model was then used to predict the customers who will enroll and who will not enroll in the long-term deposits using the testing data.

The warning message observed in the first model disappeared when the second model was built. A cutoff value of 0.5 was used to classify the customers i.e., y=1 or y=0. The area under the curve (AUC) was observed to be 0.6963.

The sensitivity, specificity, and overall accuracy were observed to be 42.18%, 97.07%, and 90.85% respectively.

To produce a simpler and more efficient model, we built a third logistic regression model using stepwise variables selection method. This model had retained the following variables: duration, month, poutcome, emp.var.rate, cons.price.idx, contact, euribor3m, job, default, pdays, campaign, and day_of_week. The rest of the variables were discarded by the stepwise method.

This model was finally used to make the predictions in the testing data. The ROC curve is plotted below. It was observed that the AUC of the ROC curve was 0.6988.



The confusion matrix is given below:

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	7095	208	7303
	Yes	536	398	934
	Total	**7631**	**598**	**8237**

The sensitivity, specificity, and overall accuracy were observed to be 42.61%, 97.15%, and 90.97% respectively.

There is only a slight improvement in the model's performance selected using the stepwise selection method compared to the second model which included all but one predictor (loan). The sensitivity of both the models was slightly greater than 42%, but the specificity was observed to be greater than 97%. The overall accuracy of both the models was close to 91%.

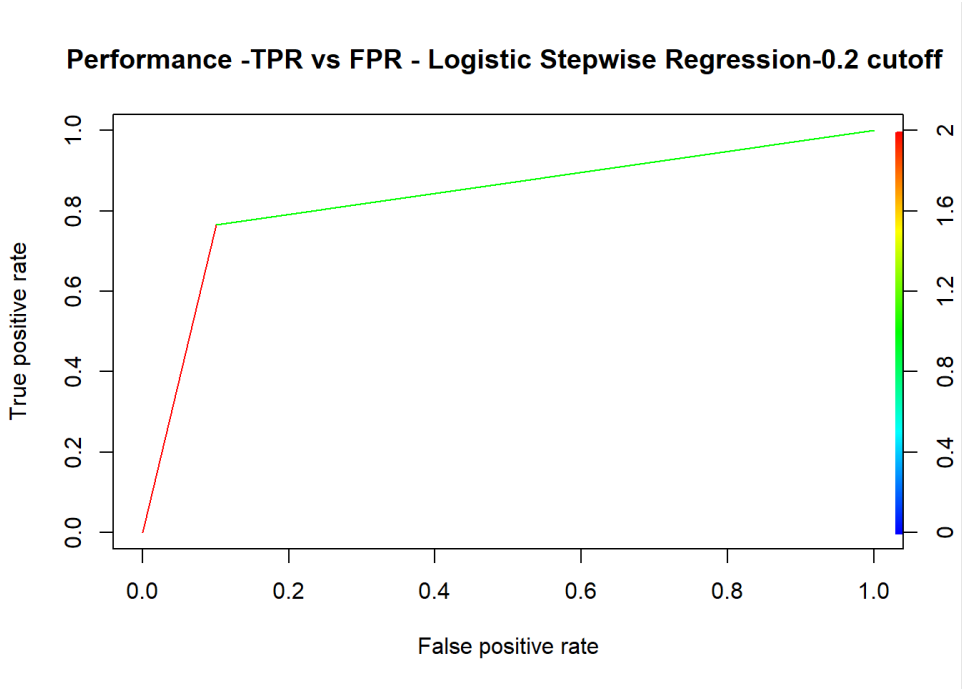
Misclassification cost with different cutoffs with Dataset 1

The stepwise model was also used to make predictions on validation data with multiple cutoff values of 0.5, 0.9, 0.3, 0.4, and 0.2. The total cost of misclassification was then calculated as the sum of false positives and 10-fold false negatives for each cutoff. This formula is used based on the assumption that the banks will lose about 10 times more money by not correctly identifying the customers that will subscribe to the long-term deposits than the ones that are not correctly identified for not subscribing. A cutoff of 0.2 is observed to have the lowest misclassification

cost. So, we decided to use the test data to predict the model with a cut-off of 0.2 to classify the customers.

Cutoff	False positives	False negatives	Total Cost
0.5	177	540	**5577**
0.9	22	851	**8532**
0.4	264	464	**4904**
0.3	403	366	**4063**
0.2	633	244	**3073**

The ROC curve is plotted below. It was observed that the AUC of the ROC curve was 0.833.



The confusion matrix is given below:

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	6570	733	7303
	Yes	219	715	934
	Total	**6789**	**1448**	**8237**

The sensitivity, specificity, and overall accuracy were observed to be 76.56%, 89.96%, and 88.44 % respectively.

Dataset 2:

The logistic regression model was also built using the training data of second bank dataset. First, the model was built with no predictors and "Personal.Loan" as the dependent variable. Then, the model was fitted using all the 13 predictor variables. The significant variables in the

model are as follows: Income, Family, CCAvg, Education, Securities.Account, CD.Account, Online and CreditCard. This model was then used to predict the "Personal.Loan" using the test data with a sensitivity of 61% and specificity of 99%. We also built a stepwise regression model for the second dataset using the training dataset and predicted with the testing set. This stepwise model retained the following variables, and the rest are eliminated in the last step: Income, Education, CD.Account, Family, CreditCard, CCAvg, Online, Securities.Account. It was observed that the AUC of the ROC curve was 0.796.

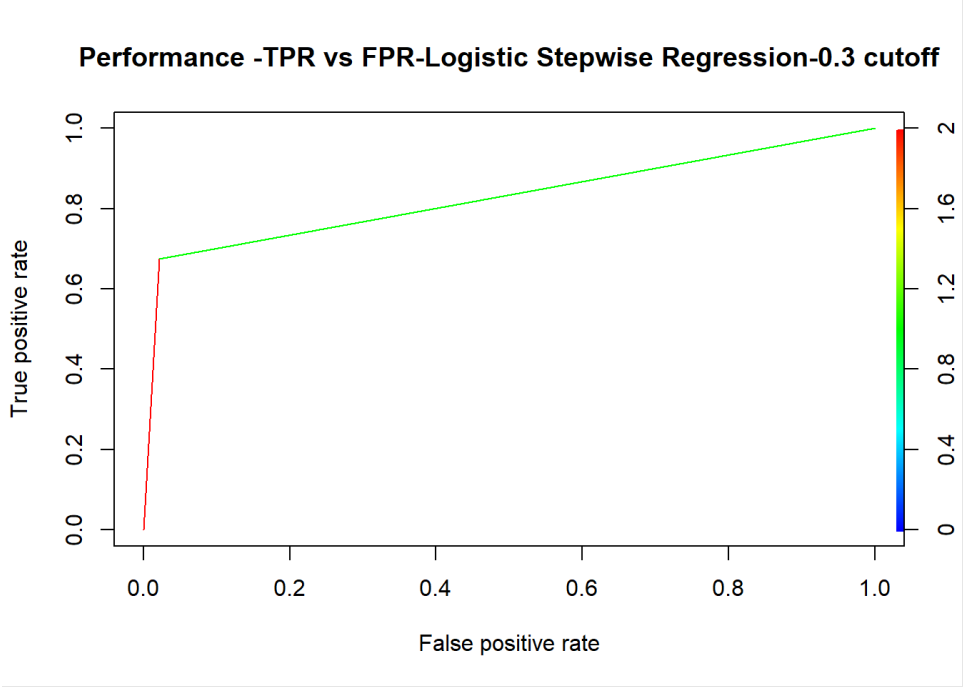
The sensitivity, specificity, and overall accuracy are observed to be 61%, 98.57%, and 95.2% respectively.

Misclassification cost with different cutoffs with Dataset 2

The stepwise model was also used to make predictions on validation data with multiple cutoff values of 0.5, 0.9, 0.3, and 0.4. The total cost of misclassification was then calculated as the sum of false positives and 10-fold false negatives for each cut-off (using the same assumption for dataset 1). The misclassification cost associated with False Positives and False Negatives is minimum for the predictions with a cutoff value of 0.3. Therefore, we used this cutoff to classify customers (Personal.Loan = 0 and Personal.Loan = 1) and predict customers from test data. ROC curve is plotted below, with an AUC of 0.83. The misclassification cost with 0.3 cutoffs on test data is 310(with 20 FP and 29 FN).

Misclassification cost on validation dataset:

Cutoff value	False positives	False negatives	Total Cost
0.5	15	30	**315**
0.9	1	56	**561**
0.4	27	26	**287**
0.3	44	21	**254**



The confusion matrix is given below:

		Predicted Outcome		
		0	**1**	**Total**
True Outcome	**0**	891	20	911
	1	29	60	89
	Total	**920**	**80**	**1000**

The sensitivity, specificity, and overall accuracy are observed to be 67.42%, 97.80%, and 95.10% respectively.The prediction has a low false positive rate of 2%, a low false negative rate of 33%, and a high prediction accuracy of 95%.

2. K-nearest neighbors (knn)

What is the model about: The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label in the case of classification.

How the model was used: The training dataset was used to build the k-nearest neighbors (knn) model. The knn is a supervised learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points. To build the model, the value of k was varied from 3 to 401 and the predictions were made in the validation dataset.

Data Massaging, esp. for Dataset 1: Knn model required numeric data to be segregated from rest of the non-numeric attributes. The outcome variable 'y' was converted from "yes" to numeric

value 1 and rest to a numeric value of 0. Post that, all the columns with numeric values were extracted and further processed for knn.

The Key Parameters: The sensitivity of the models observed in the validation dataset are presented in the table below. The highest sensitivity (47.28%) was observed when k=15

Number of Neighbors (k)	Sensitivity on Validation Data (%)	Number of Neighbors (k)	Sensitivity on Validation Data (%)
3	45.97%	151	46.84%
4	45%	157	46.4%
5	46.5%	158	46.52%
10	47.1%	201	44.02%
15	47.28%	251	41.73%
51	46.30%	301	40.4%
101	44.23%	401	38.9%

When we executed the k=15 on the test data, the sensitivity improved to 50.4%.

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	7010	293	7303
	Yes	463	471	934
	Total	**7473**	**764**	**8237**

The sensitivity, specificity, and overall accuracy are observed to be 50.4%, 95.98%, and 90.8% respectively.

Dataset 2 for knn:

The training dataset was used to build the k-nearest neighbors (knn) model. The knn is a supervised learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points. To build the model, the value of k was varied from 1 to 301 and the predictions were made in the validation dataset. The sensitivity of the models observed in the validation dataset are presented in the table below. The highest sensitivity (15.38%) was observed when k=2

Number of Neighbors (k)	Sensitivity on Validation Data (%)
1	13.3%
1.9	13.3%
2	11.1%
2.5	11.1%
3	6.66%
4	6.66%
5	2.22%

10 to 100	0%

When we executed the k=2 on the test data, the sensitivity improved to 19.01%.

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	848	63	911
	Yes	72	17	89
	Total	**920**	**80**	**1000**

The sensitivity, specificity, and overall accuracy are observed to be 19.01%, 93.08%, and 86.5% respectively.

3. Random forest (RF) model

What is the model about: Random forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random forests are frequently used as black box models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

How the model was used: The training dataset was used to build the random forest with number of trees varying from 500 to 2000 with an increment of 500 trees. The number of attributes that could be randomly selected to build the trees varied from 4 to 6. The validation dataset was used to make predictions on all the 12 random forests that were built using the training dataset. It was observed that the smallest misclassification error rate of 8.558% was observed with a random forest having 1000 trees that were built by randomly selecting 6 attributes. It can be noted that each tree used different attributes. Since the random forest model has inbuilt validation procedure based on out-of-bag samples, the validation dataset in this exercise was simply used to select the random forest that results with the lowest misclassification error rate. The misclassification error rates observed with various random forests are given in the table below.

# of Trees	# of attributes	Error Rate (%)	# of Trees	# of attributes	Error Rate (%)
500	4	8.667	1500	4	8.679
500	5	8.619	1500	5	8.643
500	6	8.570	1500	6	8.655
1000	4	8.679	2000	4	8.619
1000	5	8.667	2000	5	8.631
1000	6	8.558	2000	6	8.619

The random forest with the smallest misclassification error rate (1000 trees, 6 attributes) is highlighted in the table above. This model was selected as the final model and was used to make predictions in the testing dataset. The confusion matrix is presented below.

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	7037	266	7303
	Yes	428	506	934
	Total	**7465**	**772**	**8237**

The sensitivity, specificity, and overall accuracy are observed to be 54.18%, 96.36%, and 91.57% respectively. The overall accuracy achieved by the model is greater than 90%. The specificity appeared to be quite high (> 96%), but the sensitivity of the model was observed to be a bit low. It appears that there are plenty of existing attributes that have similar results for both types of customers (i.e., those who enrolled and those who didn't in the long-term deposits) and hence the model didn't achieve high sensitivity.

The attributes of importance are plotted in the graph. The attributes are ordered in terms of their importance from top to bottom with the attribute at the top to be most important to predict the outcome. The duration of last contact appeared to have the most significant impact on the campaign outcome followed by Euribor 3-month rate. The graph is provided in the appendix.

The RF model was also fitted in the second dataset. Training dataset was used to build the random forest with number of trees varying from 500 to 2000 with an increment of 500 trees. The number of attributes that could be randomly selected to build the trees varied from 2 to 4. The validation dataset was used to make predictions on all the 12 random forests that were built using the training dataset. It was observed that the smallest misclassification error rate of 0.9% was observed with a random forest having 1000 trees that were built by randomly selecting 4 attributes.

As done previously, the validation dataset in this exercise was simply used to select the random forest that results with the lowest misclassification error rate. The misclassification error rates observed with various random forests are given in the table below.

# of Trees	# of attribute	Error Rate (%)	# of Trees	# of attribute	Error Rate (%)
500	2	1.5	1500	2	1.5
500	3	1.1	1500	3	1.0
500	4	1.1	1500	4	1.2
1000	2	1.5	2000	2	1.6
1000	3	1.1	2000	3	1.1
1000	4	0.9	2000	4	1.1

The random forest with the smallest misclassification error rate (1000 trees, 4 attributes) is highlighted in the table above. This model was selected as the final model and was used to make predictions in the testing dataset. The confusion matrix is presented below.

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	910	1	911

	Yes	16	73	89
	Total	**926**	**74**	**1000**

The sensitivity, specificity, and overall accuracy are observed to be 82.02%, 99.89%, and 98.3% respectively. The overall accuracy achieved by the model is greater than 90%. The specificity appeared to be quite high (> 96%), but the sensitivity of the model was observed to be a bit low. It appears that there are plenty of existing attributes that have similar results for both types of customers (i.e., those who enrolled and those who didn't in the long-term deposits) and hence the model didn't achieve high sensitivity.

The attributes of importance are plotted in the graph which is provided in the appendix. The attributes are ordered in terms of their importance from top to bottom with the attribute at the top to be most important to predict the outcome. The income appeared to have the most significant impact on the personal loan approval followed by Education.

4. Support Vector Machine (SVM) model

What is the model about: The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

How the model was used: Using only a training and test dataset, a basic SVM model was built. Currently, the data underwent one hot encoding for the categorical variables.

		Predicted Outcome		
Model #1		**No**	**Yes**	**Total**
True Outcome	**No**	7132	171	7303
	Yes	631	303	934
	Total	**7763**	**474**	**8237**

The resultant SVM model was used to make the predictions in the testing dataset. The sensitivity **32.44 %**, specificity 97.66 %, and overall accuracy was 90.26 %.

It's seen that specificity is rather high as it represents the model's ability to correctly reject clients who will not subscribe to a long-term deposit; this may also be because the class is imbalanced and there is a much larger number of people who don't subscribe compared to those who do. If we focus on the model's ability to identify the clients who will subscribe (sensitivity), we find that the metric is much lower at 32.44%. Please refer to the appendix for more information.

Furthermore, an additional model was built to predict whether someone would take out a personal loan at a bank or not. The following results yielded from the secondary SVM model that predicted on this second dataset.

		Predicted Outcome		
Model #2		**No**	**Yes**	**Total**
True Outcome	**No**	902	9	911

	Yes	39	50	89
	Total	**941**	**59**	**1000**

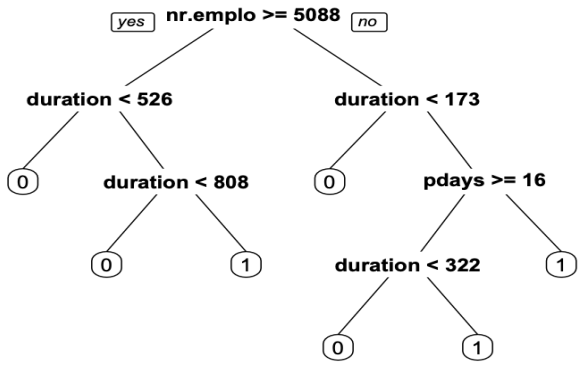
The resultant SVM model was used to make the predictions in the testing dataset. The sensitivity 56.20 %, specificity 99.02 %, and overall accuracy was 95.20 %

The results show that this model performs better in predicting whether someone would take out a loan than the first model/dataset where it predicts whether someone would subscribe to a long-term deposit. Though the second dataset is much smaller, the reason for the increase in accuracy in this second model could be due to the features, quality or importance. As seen in the explanatory data analysis of the two datasets, it can be seen that the dataset has a smaller number and likely stronger features that assist in the model predictions.

5. Classification and Regression Tree (CART) model

What is the model about: A Classification and Regression Tree model explains outputs of response variables based on values of other variables in the data by creating a decision tree.

How the model was used: When running the Classification and Regression Tree model on the likelihood of people taking out a long-term loan using the training dataset, below is the plotted output.



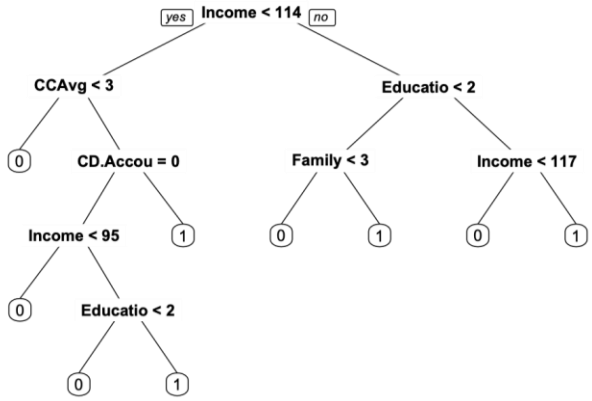
The model, using 5 and 1 for its minimum number of observations , found that the number of employees, duration of the call, and pdays (number of days since contact) were significant in the output. Below is the Confusion matrix that was derived using the testing dataset.

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	7050	253	7303
	Yes	465	469	934
	Total	**7515**	**722**	**8237**

Using this confusion matrix, the sensitivity 50.21 %, specificity 96.54 %, and the accuracy **91.28 %** of the model are below.

The CART model follows similar trends to the other models used to predict the banking dataset (high specificity with a lower sensitivity). Because the nature of the model includes most people not subscribing to a long-term deposit, this is a plausible and logical conclusion to make. Accuracy may be high, but the lowest cost is associated with a higher sensitivity rather than specificity.

Another CART model was created using the training data from the second dataset. Below is the plotted output when regressing on the likelihood of a person taking out a Personal Loan.



The model, using a minimum number of observations of 1, found the variables Income, CCAvg, CD.Account, Family, and Education statistically significant. Below is the confusion matrix derived when using the testing dataset.

		Predicted Outcome		
		No	**Yes**	**Total**
True Outcome	**No**	908	3	911
	Yes	11	78	89
	Total	**919**	**81**	**1000**

Using this second confusion matrix, the sensitivity 87.64 %, specificity, and the accuracy 98.6% of the model were calculated below.

This CART model yielded a higher sensitivity, specificity, and accuracy than the previous model. However, there may be other reasons that have led to these results such as the nature of this data having a lower percentage of people taking out personal loans than those taking out a long-term deposit in the other dataset. The second dataset also had less observations than the first dataset.

CONCLUSION

When the models were fitted in the first dataset (Bank telemarketing dataset), sensitivity was observed to be considerably smaller than the specificity and accuracy. For most of the models, the accuracy was observed to be greater than 90%, and specificity varied from 89.96% to

97.66% while sensitivity varied from 32.44% to 76.56%. The variable Duration was observed to be significant in the logistic, random forest, and CART models.

We observed that the dataset was imbalanced in terms of the customers who enrolled (about 11%) and who did not (about 89%). The measure of accuracy was not ideal in such an imbalanced dataset and hence we decided to put more weight on other measures like sensitivity and specificity to estimate the performance of the models. All the 5 models exhibited smaller sensitivity, which indicated that the given attributes available in the dataset were not sufficient for the models to identify the true customers who would enroll in the bank-offered long-term deposits and more information regarding those customers would be desired.

In the second dataset, sensitivity was still observed to be smaller than specificity and accuracy, but the difference was not as extreme. The accuracy was still generally greater than 90%. sensitivity varied from 19.01% to 87.64%, and specificity varied from 93.08% to 99.89%. The variables Education and Income were commonly observed to be significant in the logistic regression, random forest, and CART models. These variables in general are great indicators of the suitability of the customers for personal loans and the models corroborated that assumption.

In general, we observed that all 5 models we implemented in this project performed quite well in classifying the ideal customers for the banks. The lower sensitivity observed in all the models indicated that more data are required for the models for better performance.

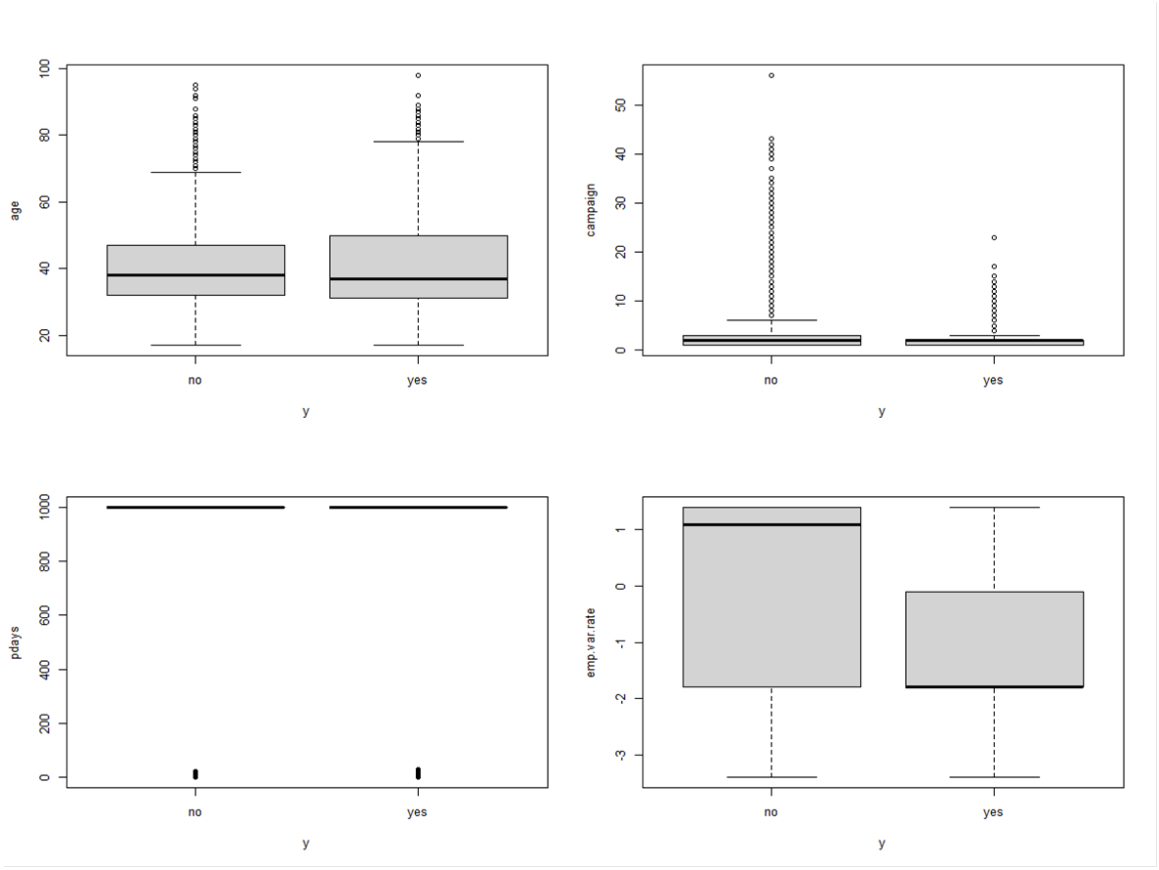
APPEXDIX

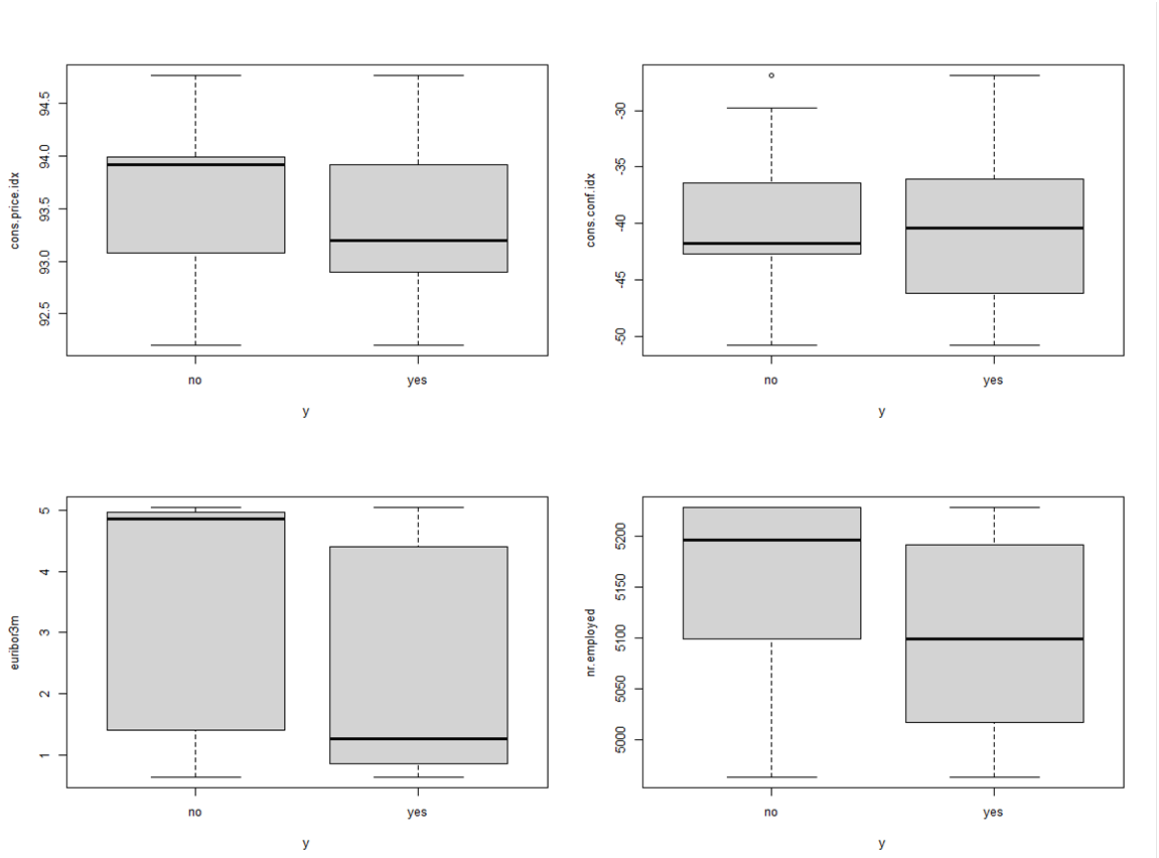
Attributes of Dataset 1:

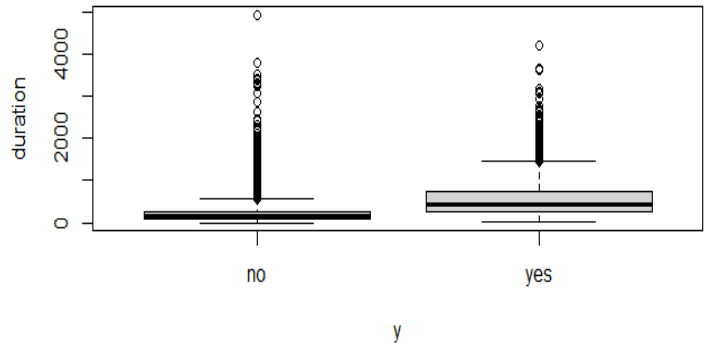
1	Age	Age (Numeric)
2	Job	Type of job (Categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3	Marital	Marital status (Categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4	Education	Education level (Categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5	Default	Has credit in default? (Categorical: 'no', 'yes', 'unknown')
6	Housing	Has housing loan? (Categorical: 'no', 'yes', 'unknown')

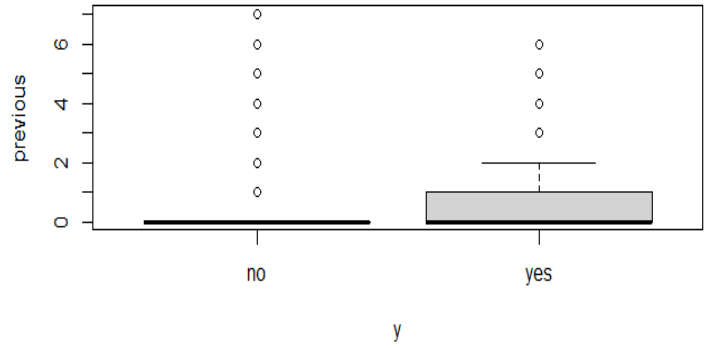
7	Loan	Has personal loan? (Categorical: 'no', 'yes', 'unknown')
8	Contact	Contact communication type (Categorical: 'cellular', 'telephone')
9	Month	Last contact month of year (Categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10	Day_of_week	Last contact day of the week (Categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
11	Duration	Last contact duration, in seconds (Numeric)
12	Campaign	Number of contacts performed during this campaign and for this client (Numeric, includes last contact)
13	Pdays	Number of days that passed by after the client was last contacted from a previous campaign (Numeric; 999 means client was not previously contacted)
14	Previous	Number of contacts performed before this campaign and for this client (Numeric)
15	Poutcome	Outcome of the previous marketing campaign (Categorical: 'failure', 'nonexistent', 'success')
16	Emp.var.rate	Employment variation rate - quarterly indicator (Numeric)
17	Cons.price.idx	Consumer price index - monthly indicator (Numeric)
18	Cons.conf.idx	Consumer confidence index - monthly indicator (Numeric)
19	Euribor3m	Euribor 3-month rate - daily indicator (Numeric)
20	Nr.employed	Number of employees - quarterly indicator (Numeric)
21	Y	Has the client subscribed to a long-term deposit? (Binary: 'yes', 'no')

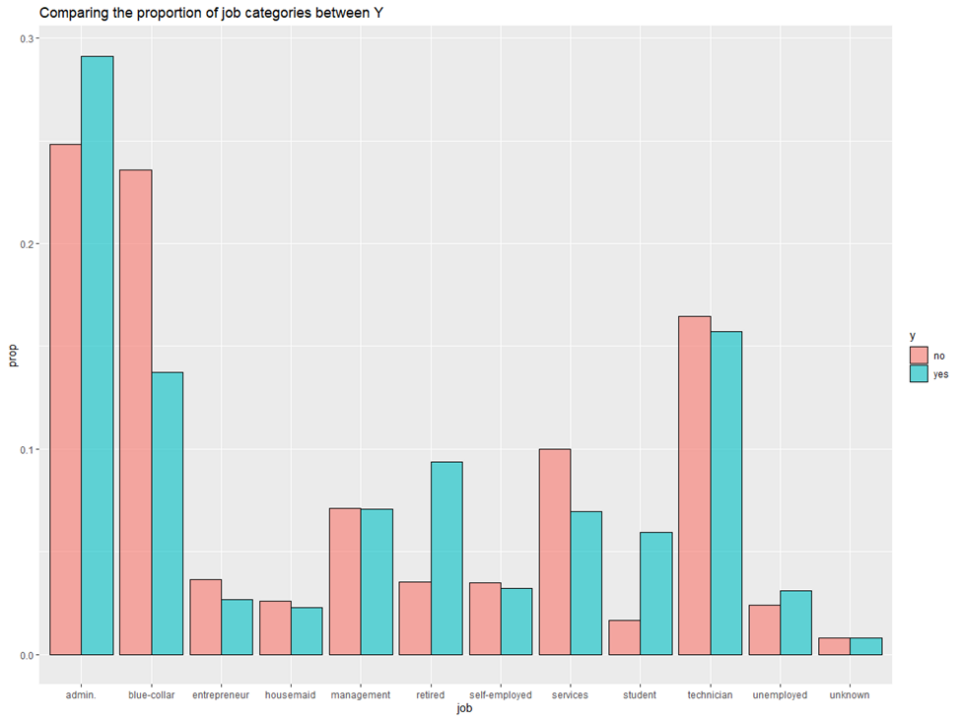
Data exploratory plots for dataset 1:

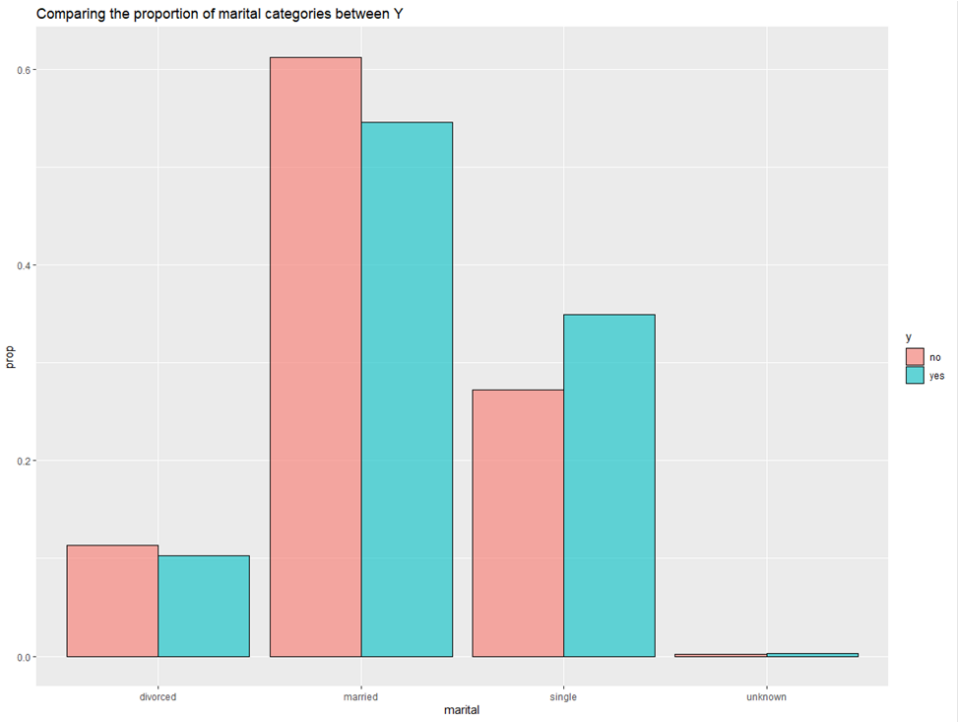

age
campaign
pdays
emp.var.rate
no
yes
y

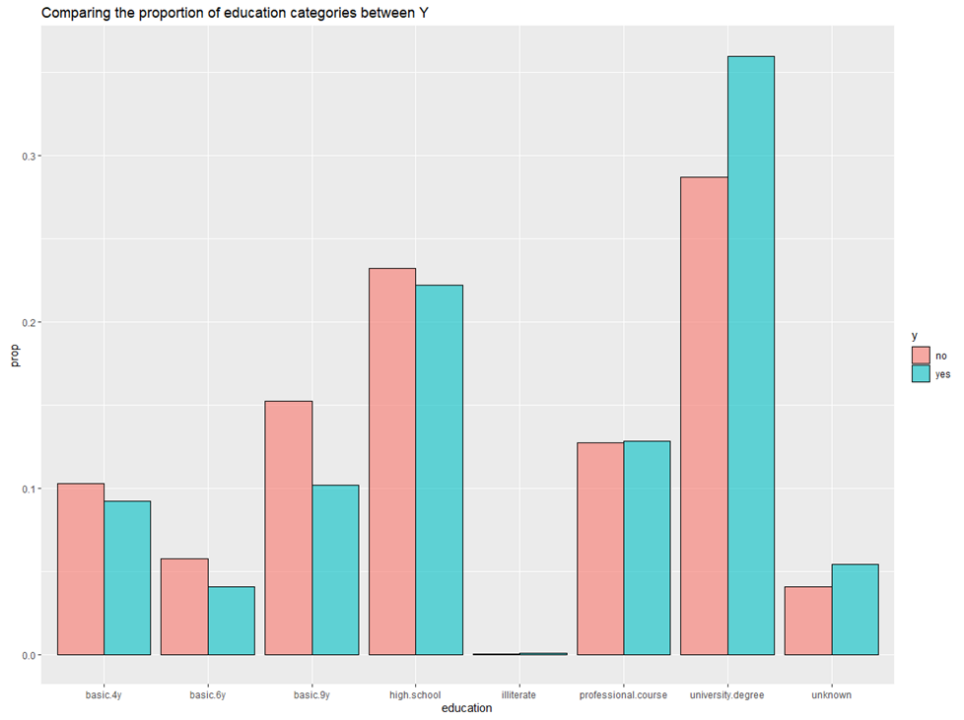

cons.price.idx
92.5
93.0
93.5
94.0
94.5
no
yes
y
cons.conf.idx
-50
-45
-40
-35
-30
no
yes
y
euribor3m
1
2
3
4
5
no
yes
y
nr.employed
5000
5050
5100
5150
5200
no
yes
y

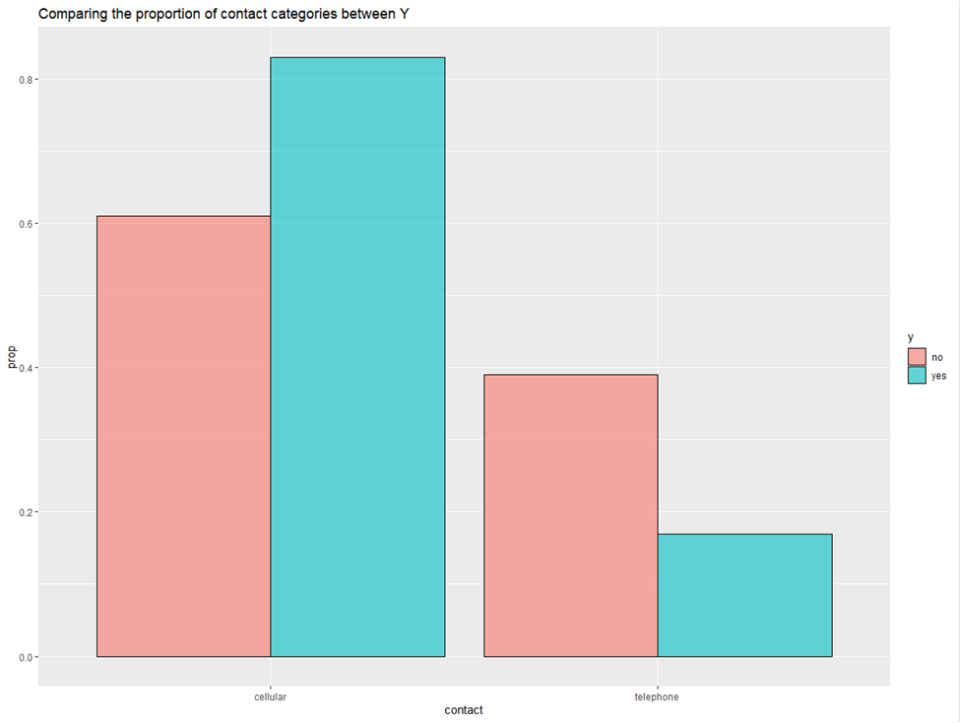

duration
0
2000
4000
no
yes
y

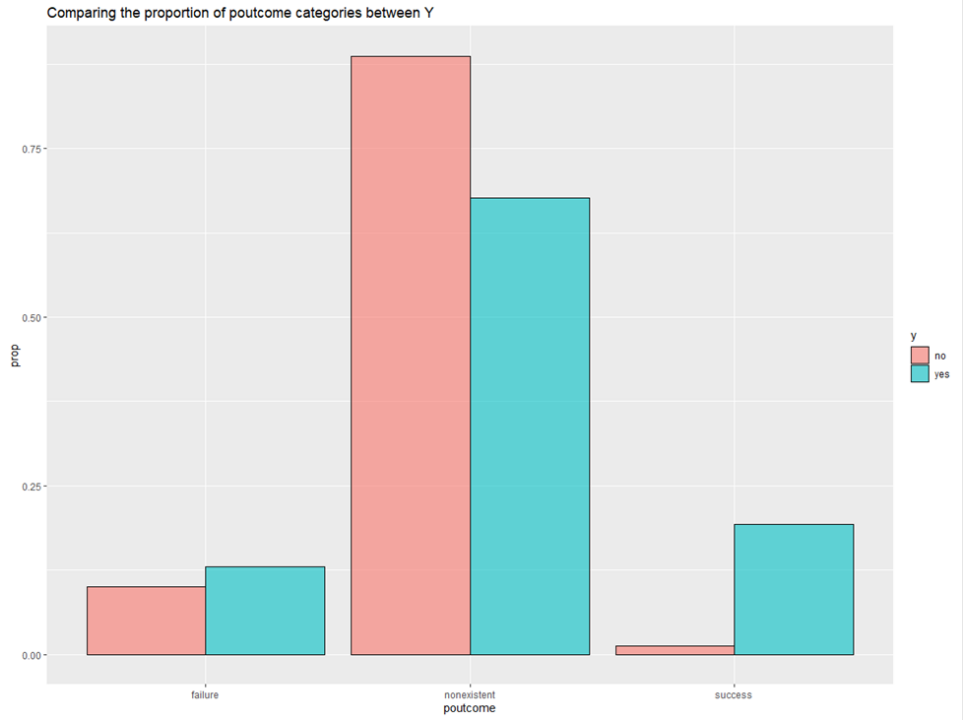

previous
0
2
4
6
no
yes
y


Comparing the proportion of job categories between Y
prop
0.3
0.2
0.1
0.0
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
job
y
no
yes

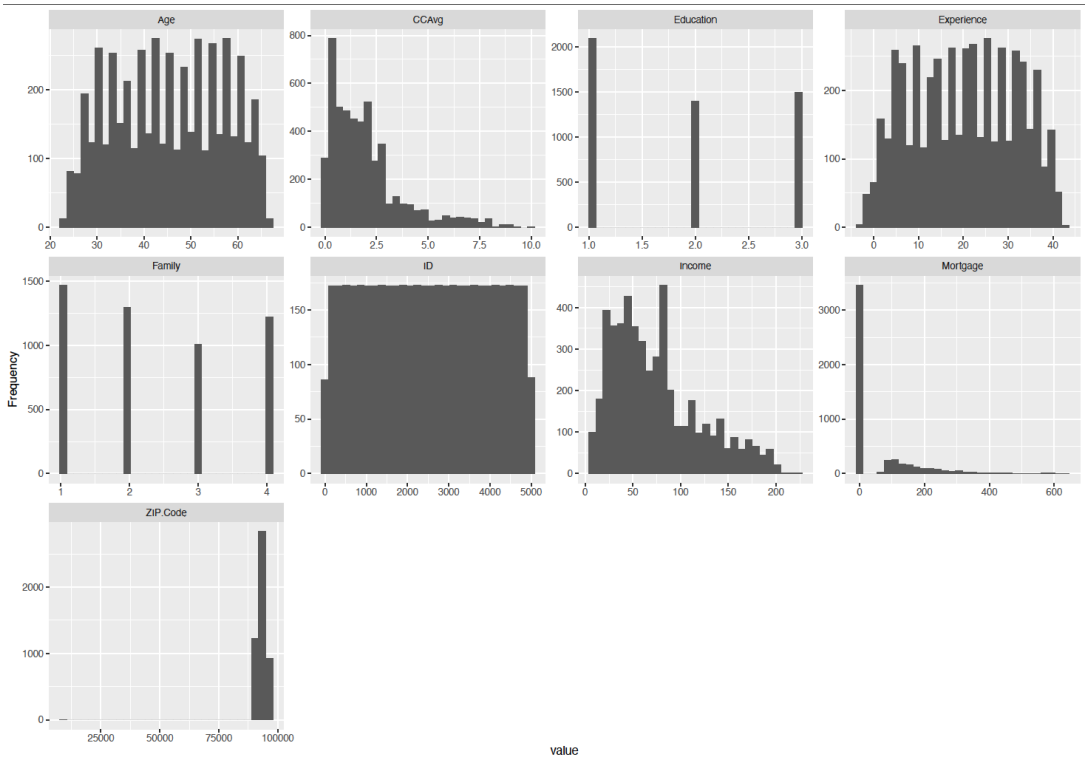

Comparing the proportion of marital categories between Y
prop
0.6
0.4
0.2
0.0
divorced
married
single
unknown
marital
y
no
yes


Comparing the proportion of education categories between Y
prop
0.0
0.1
0.2
0.3
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education
y
no
yes

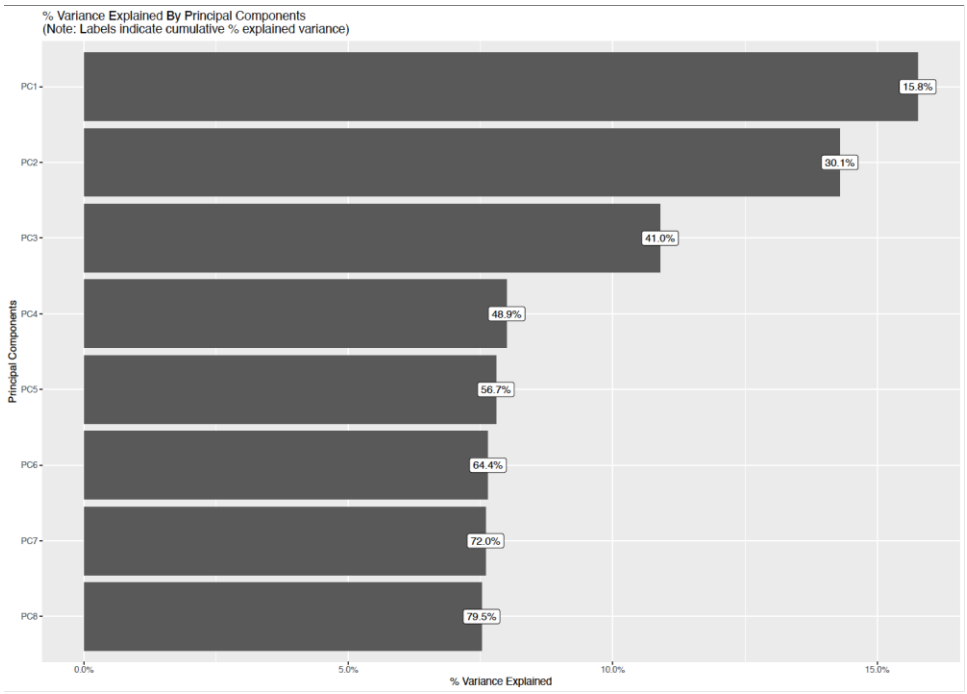

Comparing the proportion of contact categories between Y
prop
0.0
0.2
0.4
0.6
0.8
cellular
telephone
contact
y
no
yes



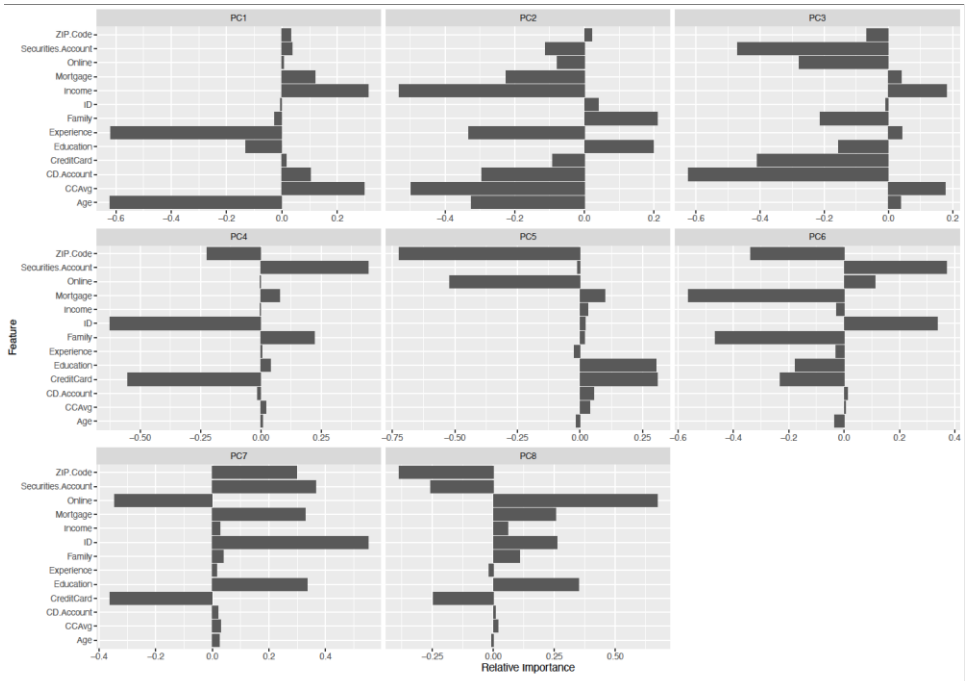
Histograms for Dataset 2:



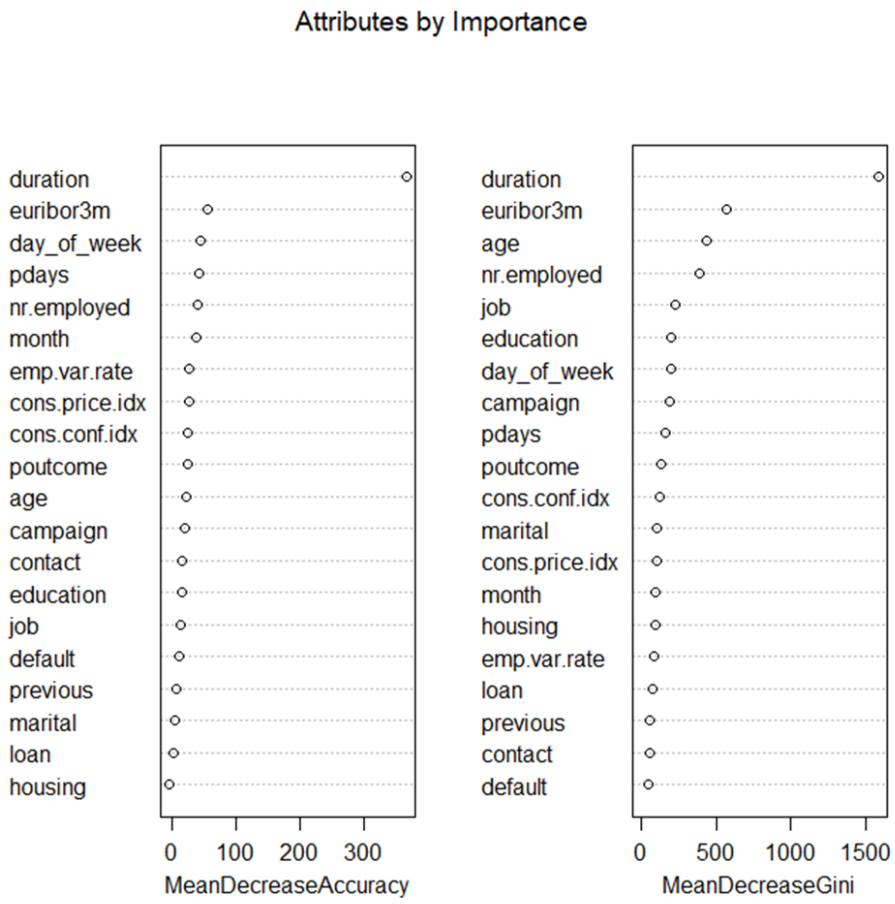
The plot for Principal Components for Dataset 2.



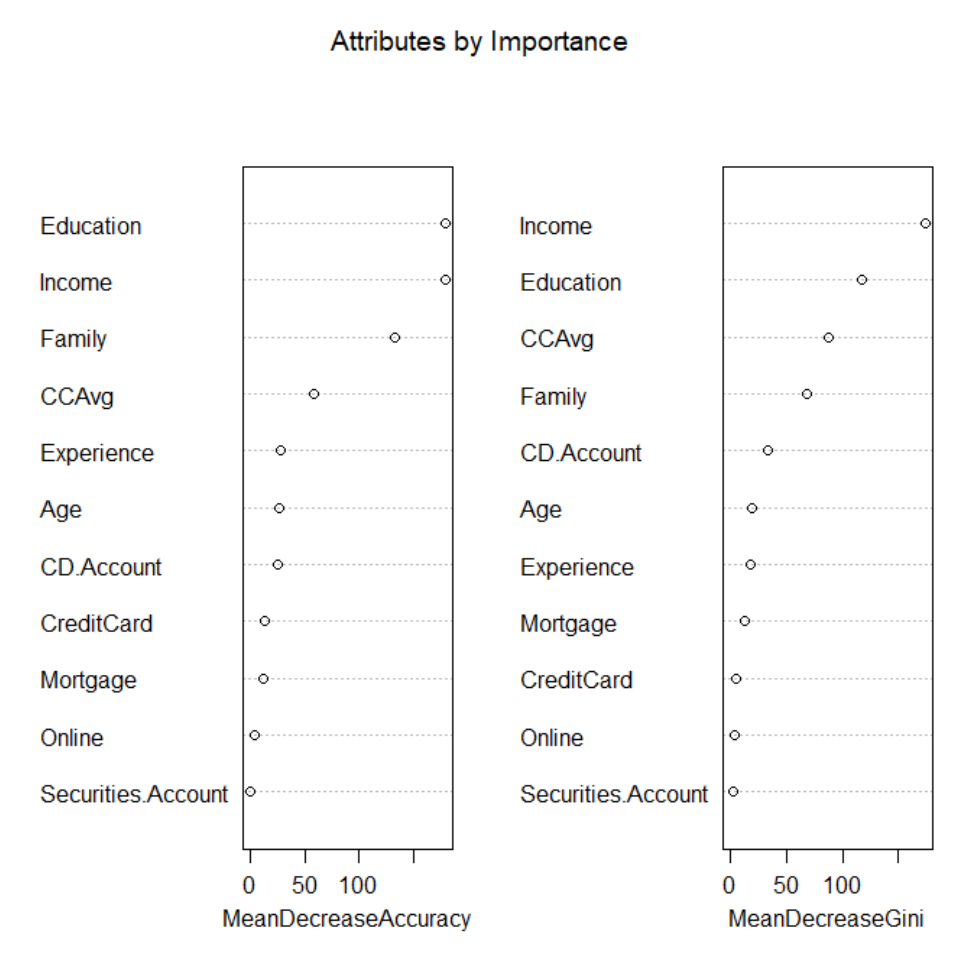
Further plots detailing the first 2 principal components.



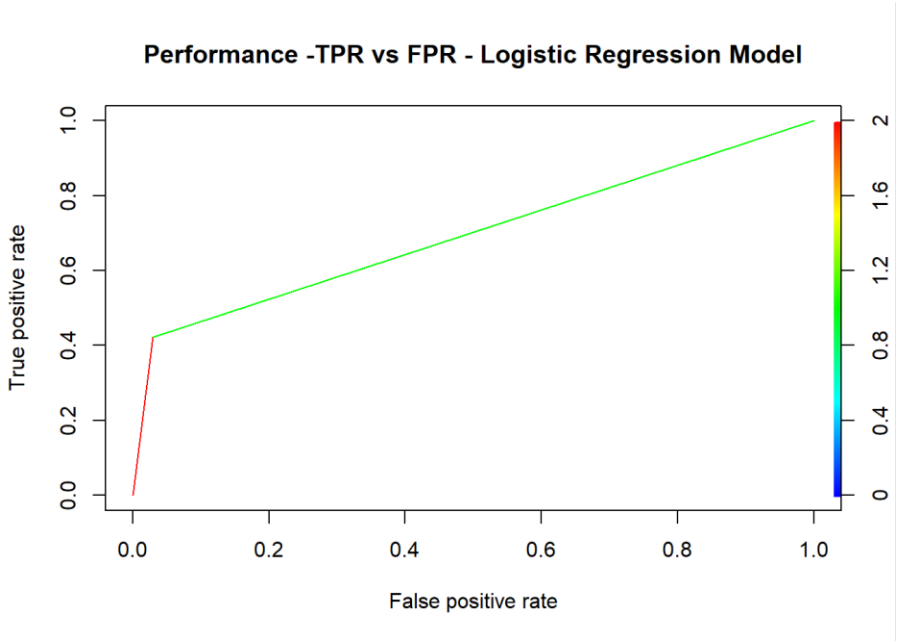
Dataset 1: Attributes of importance as identified by the RF model.

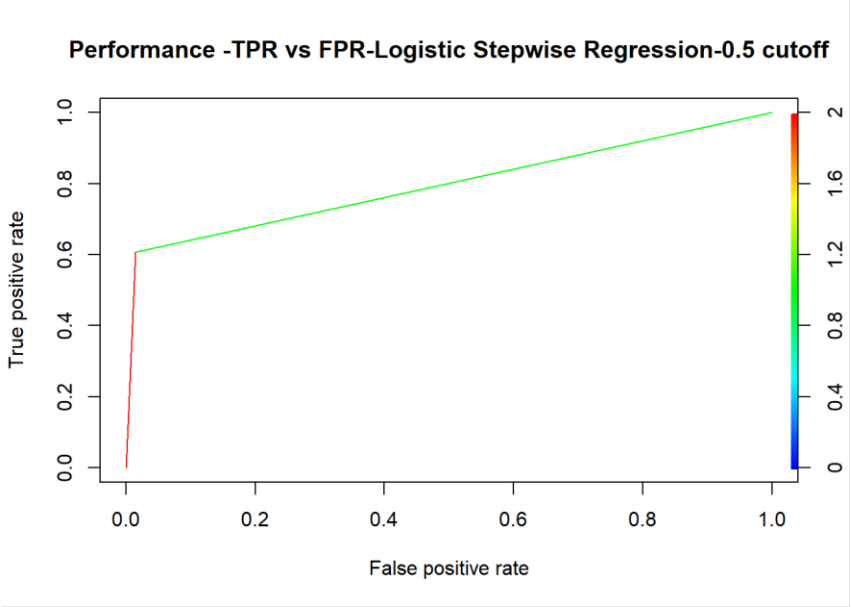


Dataset 2: attributes of importance by RF model



Dataset 1: Logistic Regression model – ROC.





SVM Model:

Based on the business need there is the option of improving sensitivity in exchange for specificity based on the chosen threshold for the classification. This would depend on the primary need or goal of the business whether to find all clients out there that will subscribe to a long-erm deposit (recall) or to prioritize the (precision) of the model, which denotes the number of correctly classified clients who will subscribe.