Evaluating Bank Marketing Success Based on Client Profile and Economic Conditions

Harry Dong

1 Motivation

Telemarketing is a typical strategy adopted by banks to drive sales. However, there is significant time and human resources expended, which may not be recouped by a strategy based on random calling. Thus, it is best for a bank to identify the ideal client profile and prevailing economic conditions which maximise the success rate of telemarketing calls. Through this, banks can achieve higher returns on resources invested into direct marketing campaigns.

Telemarketing data from a Portuguese banking institution, collected from May 2008 to May 2010, was analysed. This involved 20 variables describing client details, campaign-related data and economic variables at the time of contact, across 41,888 contacted clients (Appendix 1). The goal was to train various statistical learning models on a 'train' dataset. These models were then used to identify the clients in the 'evaluation' dataset with the highest likelihood of subscribing to a term deposit due to a successful telemarketing call.

2 Data Exploration

The numerical and categorical variables were visualised, and interesting findings recorded. The effect of the 2008 Global Financial Crisis on the willingness to subscribe to a term deposit was examined by looking at a time series of the success rate of calls (% of 'yes' responses) from May 2008 to May 2010 (Appendix 2). This was overlayed with a plot of the variation in the Euribor interest rate. From the graph, we see there is a strong inverse relationship between success rates and the Euribor. Thus, it was assumed that the effects of the GFC could be directly captured by examining the economic variables.

The distributions of 'campaign' and 'age' for the 'yes' and 'no' cases had unexpected results (Appendix 2). Increasing the number of contacts during a campaign did not seem to increase the effectiveness of the campaign, evidence by the short tail of the 'campaign' histogram for the 'yes' cases. Additionally, the age distribution for the 'yes' cases had longer tails than the 'no' cases, indicating that elderly retirees and young adults had a higher propensity to buy a term deposit. This seems to be confirmed by abnormally high success rates in the 'student' and 'retiree' job levels.

The 'default' status of clients were also visualised, and clients with 'unknown' default status had half the success rate of clients with 'no' default status (Appendix 2). This is expected, as the 'unknown' level would contain individuals with credit issues that they withheld from the bank, and who are less likely to place their cash in an illiquid term deposit.

From this initial exploratory analysis, we would expect that the ideal target client would be a student or retiree, with no credit in default. The ideal economic conditions would be during a period of low Euribor interest rates.

3 Data Cleaning

3.1 Missing Values

The frequency and pattern of missing data values were visualised (Appendix 3). The missing values all occurred in categorical variables describing personal details of the client. A consequence is that the missing values can be considered "Missing Not at Random" - the occurrence of the missing values relate to the value itself. For example, if a client's 'education' level is 'illiterate', they may be unwilling to disclose this out of embarrassment.

Most available imputation algorithms operate on the assumption that missing values are Missing At Random i.e. independent of their real value. Additionally, imputation can be computationally expensive, especially on this large dataset of 40,000 observations.

Thus, missing entries in the 'default', 'education', 'housing' and 'loan' variables were treated as an additional 'unknown' category level. Rows with missing 'job' and 'marital' variables were deleted as these had a low incidence rate (<2.5%), so there is not a significance amount of valuable data discarded.

3.2 Sparse Factor Levels

Based on the frequency plots of the categorical variables, some categories had quite infrequent occurrences of levels (Appendix 2). When a categorical variable with n levels is encoded, $n-1$ dummy variables must be created. Many sparse levels introduces the risk of overfitting models on too many dummy variables. High dimensionality also bloats the runtime of statistical learning algorithms.

Thus similar category levels were aggregated. For the 'job' variable, very niche 'entrepreneur' and 'housemaid' levels were absorbed into the more general 'self-employed' and 'services' levels respectively. Additionally, for the 'education' variable, the 'illiterate', 'basic.4y', 'basic.6y', and 'basic.9y' variables were combined in a single 'basic' level. The 'student' and 'retired' levels were combined due to similar success rates (Appendix 2).

3.3 Manipulating Predictors

Predictors that are highly correlated to another predictor are non-informative. The correlations of numerical predictors were found (Appendix 4). The 'year', 'nr.employed' and 'emp.var.rate' predictors were dropped due to their high correlation ($|\rho| > 0.9$) with 'euribor3m'. For categorical variables, the Cramer V statistic, a measure of dependency, was calculated. The 'loan' predictor was dropped due to its high Cramer V (>0.7) with 'housing'.

The 'pdays' (previous days since last contact) was incorrectly assigned as 999 to 4080 clients who had already been contacted. Thus this predictor was dropped.

Finally, the categorical variables were encoded as dummy variables in order to be used for model training.

4 Model Fitting

Three main approaches were used to train models with the cleaned training data: logistic regression variants, Linear Discriminant Analysis (LDA), and tree-based methods.

Before fitting any of the models, any near-zero variance predictors were removed. These predictors do not provide useful information for statistical learning methods as they essentially maintain a constant or near-constant value across all observations. They also have a tendency to cause errors or crashes when training models.

The following process was applied for each of these learning methods:

1. Predictor selection and/or parameter tuning.
2. Fitting the model using 5-fold cross validation, resampling the training folds with SMOTE each of the 5 times.
3. Finding the cross-validated estimate of AUC of the ROC curve.

Modelling on too few predictors will lead to underfitting, and excessive bias, while using too many predictors will overfit and introduce excessive variance. To optimise the bias-variance tradeoff, techniques like backward selection and parameter tuning were implemented, which will be discussed in sections 4.2, 4.3 and 4.4.

4.1 Imbalanced Data Issue

A major issue was that only around 10% of clients subscribed to a term deposit, thus 'yes' was a small minority of the recorded response variables. Imbalanced datasets distort the performance of statistical learning methods, so we must ensure there are an even number of 'yes' and 'no' responses. Oversampling tends to encourage overfitting, as we are essentially giving more weight to it by duplicating it. Thus, Synthetic Minority Oversampling Technique (SMOTE) was applied to any data before it was used to train a model. This creates synthetic 'yes' values by aggregating nearby data points.

The area under the ROC curve (AUC-ROC) was chosen as the performance metric because it gives an picture of the performance of the model across different probability thresholds.

4.2 Cross Validation

To calculate the AUC-ROC, a 5-fold cross validation approach was used. The 'train' data was split into 5 randomly sampled folds $\{F_1, F_2, F_3, F_4, F_5\}$, and for $i = 1, .., 5$ a model was trained on $F_{j \neq i}$ and evaluated using F_i to get a AUC-ROC. This approach is better than simply fitting a single model to a portion of the data and assessing it on the remainder of the data. By taking an average of the AUC-ROC across 5 models, we use more of the data in fitting each model, resulting in a less biased estimate of AUC-ROC.

4.3 Logistic Regression Variants

A preliminary model fitted was a logistic regression, as it is a simple and interpretable model. In logistic regression, the log-odds of the 'yes' case are estimated by fitting a linear model to the predictors. A backward selection process was performed - a model with all predictors was fitted, and the variable with the lowest individual importance (measured by the t-statistic) was discarded, and the model refitted. This process was repeated until models with $n = 1, ..., 30$ predictors were compared to find the highest AUC-ROC. The optimal number of predictors was found to be 18, with the most important being the economic variables ('cons.price.idx', 'cons.conf.idx','euribor3m') (Appendix 5).

To further reduce overfitting, a ridge regression and lasso regression were trained. Ridge and lasso are extensions of logistic regression, which prevent overfitting by shrinking excessively large regression coefficients. The higher the regularisation parameter λ, the more heavily large parameters are penalised. The regularisation parameter λ was tuned by a grid search - the AUC-ROC was found for models built with $\lambda = \{0, 0.005, ..., 0.1\}$ (Appendix 5). The optimal regularisation parameter was found to be $\lambda = 0$. Increases in λ decreased the performance of both the ridge and lasso models. Evidently, the backward selection process prevented overfitting sufficiently, so that regularisation was not necessary, and the logistic regression proved to outperform the ridge and lasso models.

4.4 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was chosen as a alternative model as it is computationally cheap to fit. In LDA, the probability of the 'yes' case is estimated by assuming that $X|Y = yes$ is normally distributed. Again, to optimise the AUC-ROC, backward selection was performed (Appendix 6). The optimal number of predictors was found to be 25, and again the economic variables ('cons.price.idx', 'cons.conf.idx','euribor3m') ranked high in terms of variable importance.

4.5 Tree-based Models

A tree based approach to modelling was selected due to its simplicity and its similarity to human decision making. In a Classification and Regression Tree (CART) model, at each node of the tree, features are split according to a rule based on inputted predictors. The complexity parameter cp was tuned using a grid search, and CART models were built with $cp = \{0, 0.00075, ..., 0.015\}$ (Appendix 7). The optimal value of cp chosen was 0.00079.

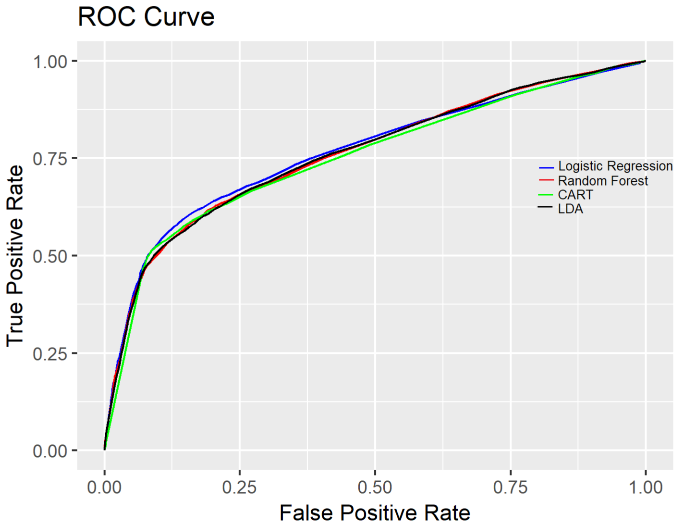
However, single decision trees tend to have high bias. In random forest models, multiple decision trees are aggregated, leaving a lower level of bias.The parameter mpart was tuned, with random forest models built for $mpart = \{1, ..., 10\}$ (Appendix 7). The optimal value of $mpart$ was found to be 3.

4.6 Results Summary

The following table summarises the performance of the final models for each learning method:

Method	AUC-ROC
Logistic Regression	0.7615
LDA	0.7618
CART	0.7353
Random Forest	0.7667

The ROC curves of the these models are shown below:



5 Final Outcome

The random forest model was selected as it had the highest ROC-AUC metric. Using this model, the probabilities for the 'evaluation' set were predicted. The clients with the 500 highest probabilities were extracted (Appendix 8).

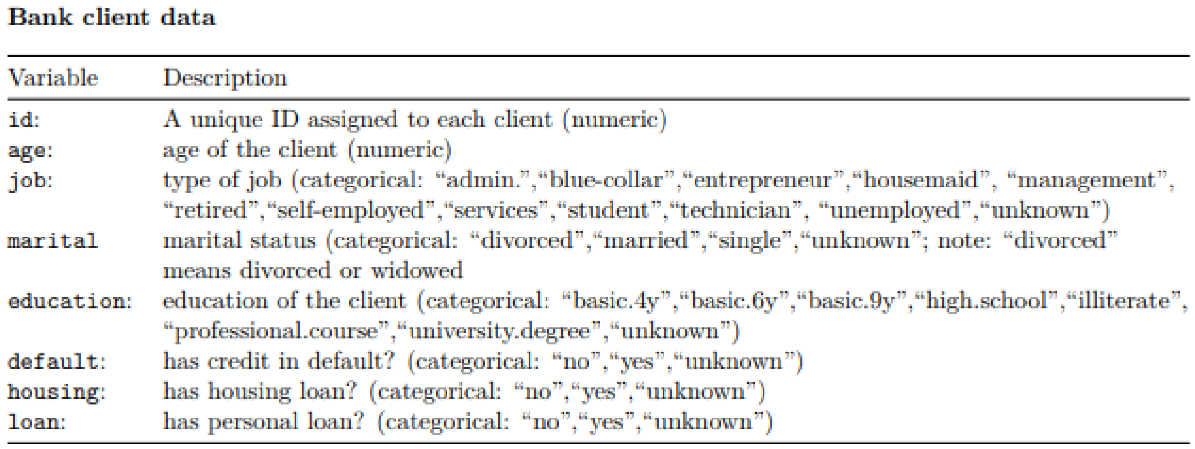
A limitation of the random forest model is that it is a 'black box' model with low interpretability, thus the fitted logistic regression model was also used to assist with the identification of a target client profile, and ideal economic conditions.

In line with our expectations, economic variables (cons.price.idx, cons.conf.idx, euribor3m) were highly important predictors of telemarketing success, having logistic regression coefficients of -1.1969, -0.04747, and 0.7869 respectively. Thus to maximise the probability of a successful call, the campaign should be conducted during a time of high Euribor interest rates, and low consumer price index.

In terms of personal and campaign details, 'age', 'contact' and 'campaign' had the most predictive power. Contrary to our expectations, 'campaign' did have a marginal effect on the probability of success, with a logistic regression coefficient of 0.0713. Based on the distributions of ages for the best clients (most likely to subscribe) and worst clients (least likely to subscribe), the age distribution is skewed towards slightly younger ages for the best clients. In terms of contact method, the majority of the best clients were contacted by mobile, while a large portion of the worst clients were contacted by telephone. This points to cellular communication as the ideal method of contact.

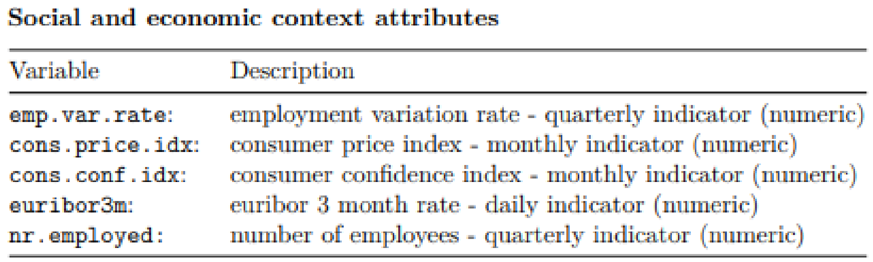
Appendices

Appendix 1: Variables



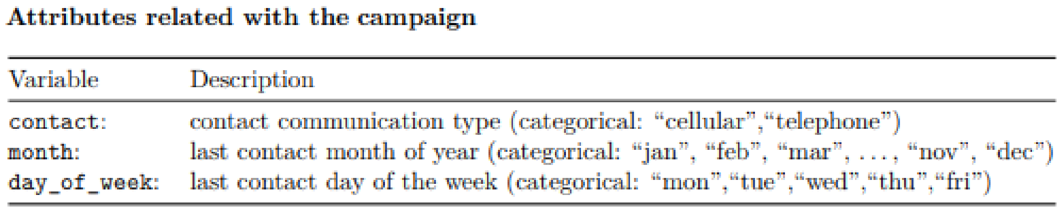
Bank client data

Variable	Description
id:	A unique ID assigned to each client (numeric)
age:	age of the client (numeric)
job:	type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid", "management", "retired","self-employed","services","student","technician", "unemployed","unknown")
marital	marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed
education:	education of the client (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate", "professional.course","university.degree","unknown")
default:	has credit in default? (categorical: "no","yes","unknown")
housing:	has housing loan? (categorical: "no","yes","unknown")
loan:	has personal loan? (categorical: "no","yes","unknown")



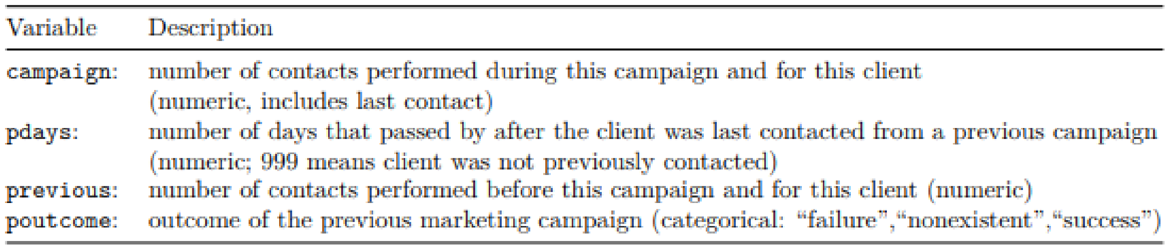
Social and economic context attributes

Variable	Description
emp.var.rate:	employment variation rate - quarterly indicator (numeric)
cons.price.idx:	consumer price index - monthly indicator (numeric)
cons.conf.idx:	consumer confidence index - monthly indicator (numeric)
euribor3m:	euribor 3 month rate - daily indicator (numeric)
nr.employed:	number of employees - quarterly indicator (numeric)

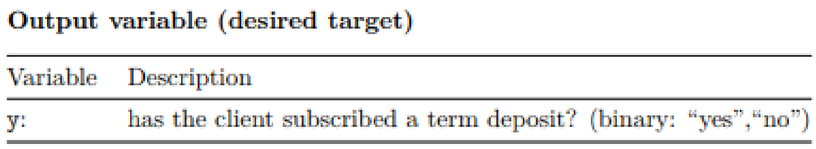


Attributes related with the campaign

Variable	Description
contact:	contact communication type (categorical: "cellular","telephone")
month:	last contact month of year (categorical: "jan", "feb", "mar", . . . , "nov", "dec")
day_of_week:	last contact day of the week (categorical: "mon","tue","wed","thu","fri")



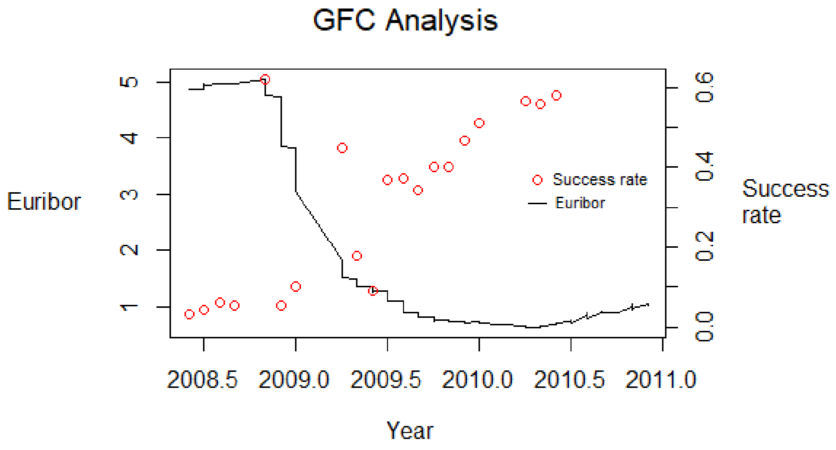
Variable	Description
campaign:	number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays:	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
previous:	number of contacts performed before this campaign and for this client (numeric)
poutcome:	outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

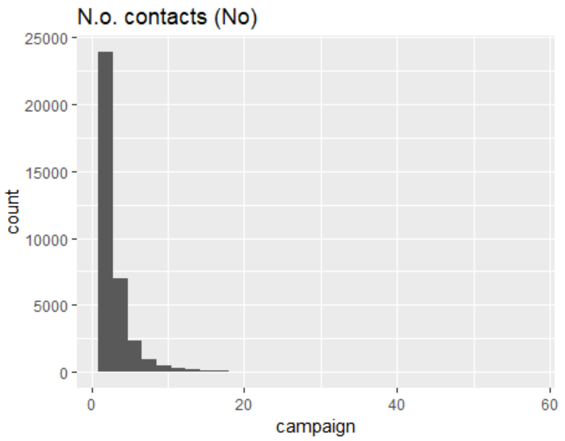


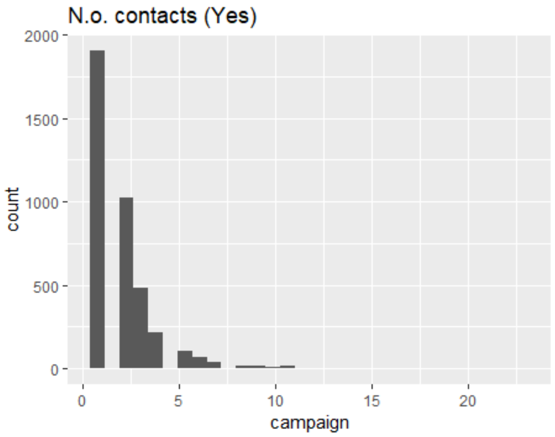
Output variable (desired target)

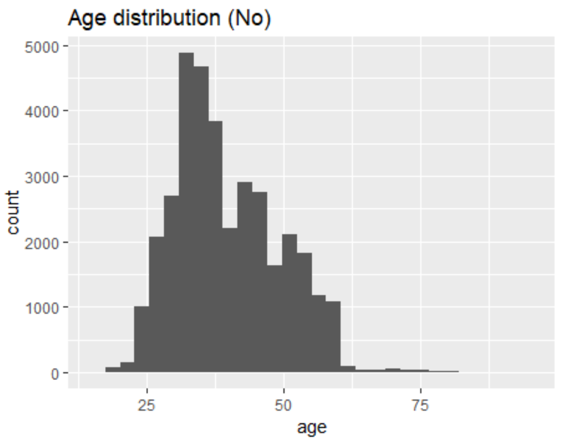
Variable	Description
y:	has the client subscribed a term deposit? (binary: "yes","no")

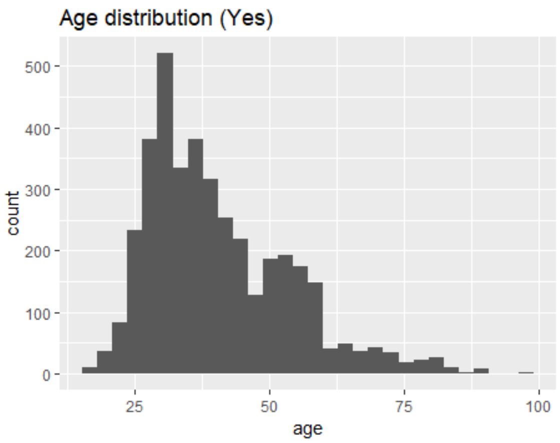
Appendix 2: Data Exploration

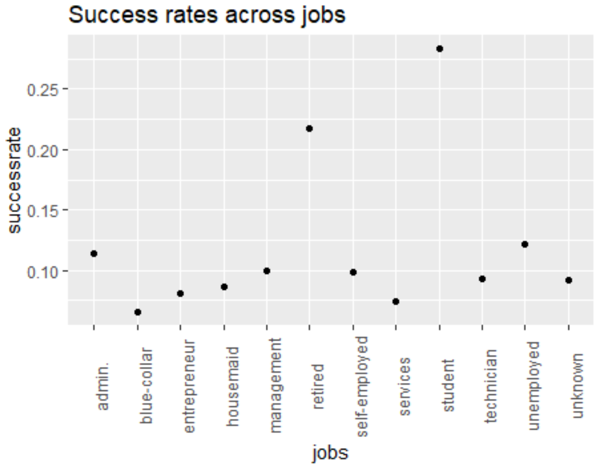


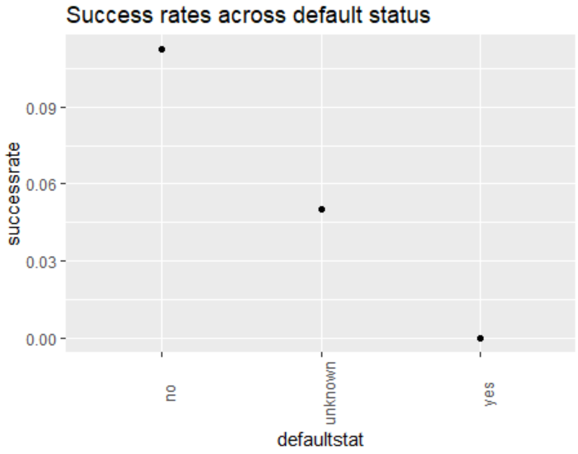


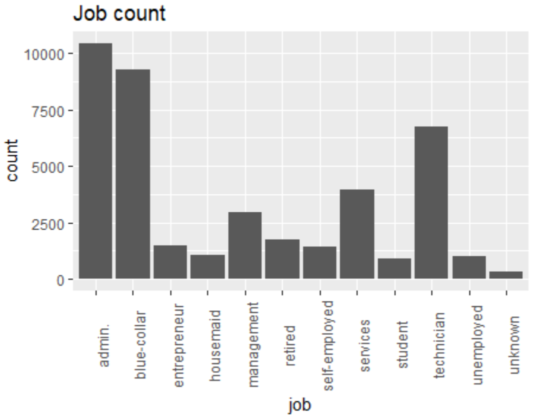


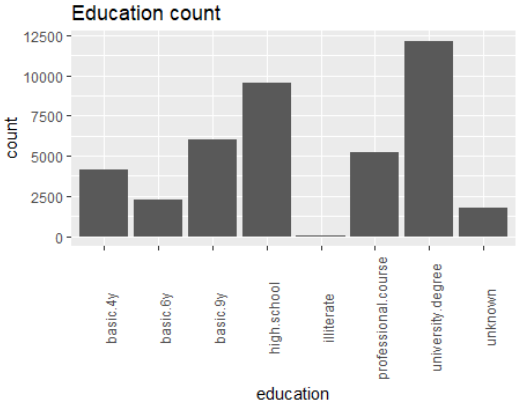




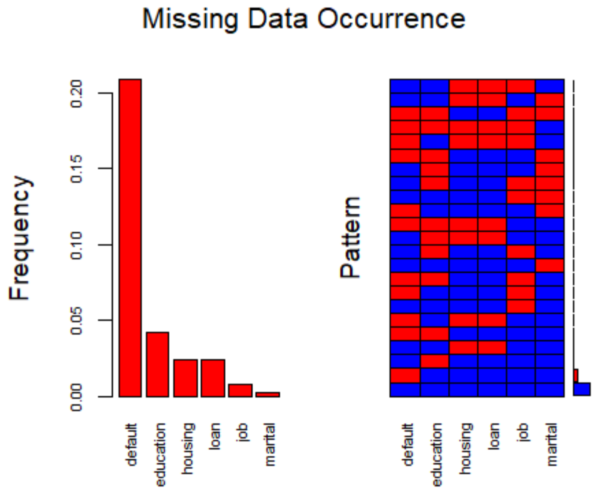




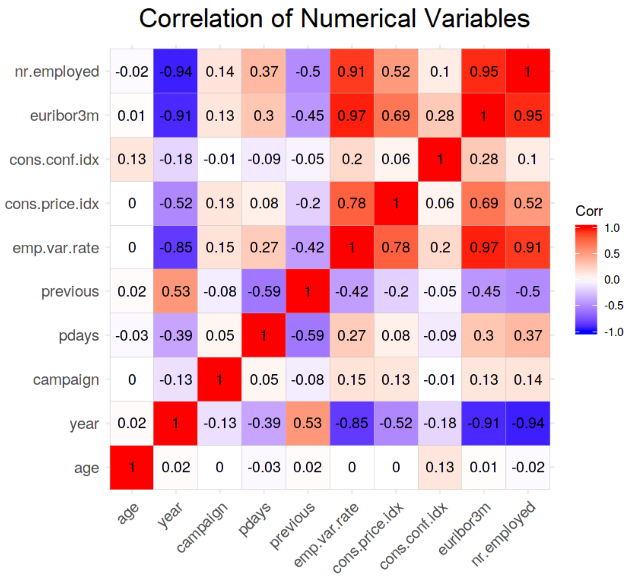




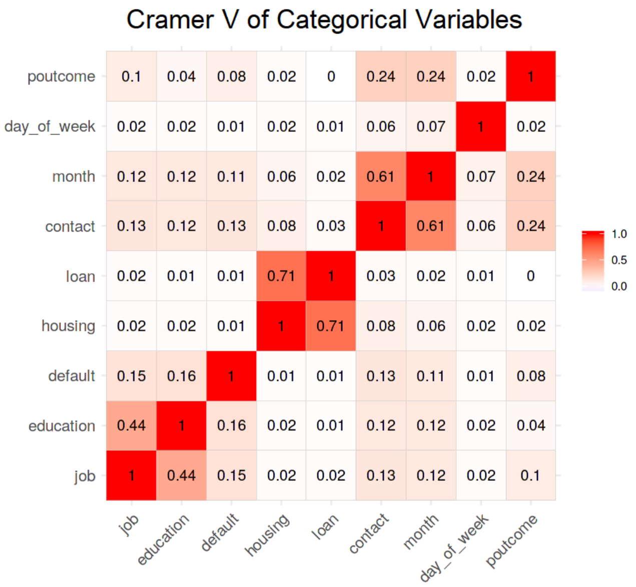
Appendix 3: Missing Data



Appendix 4: Predictor Correlations

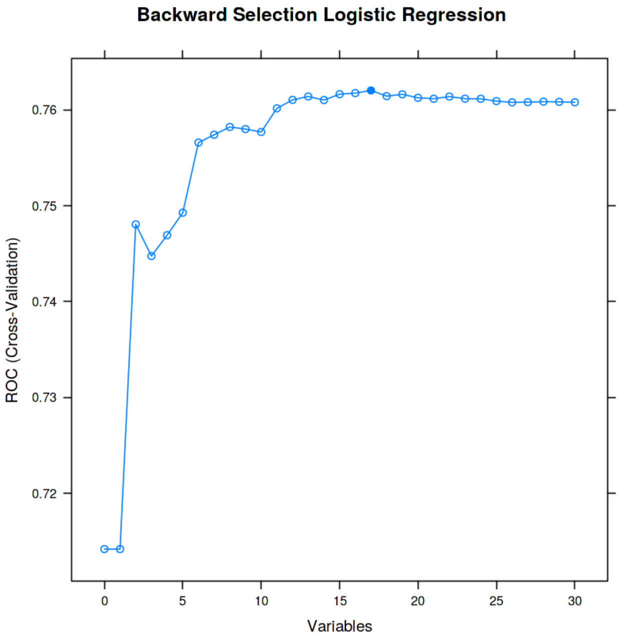

Correlation of Numerical Variables

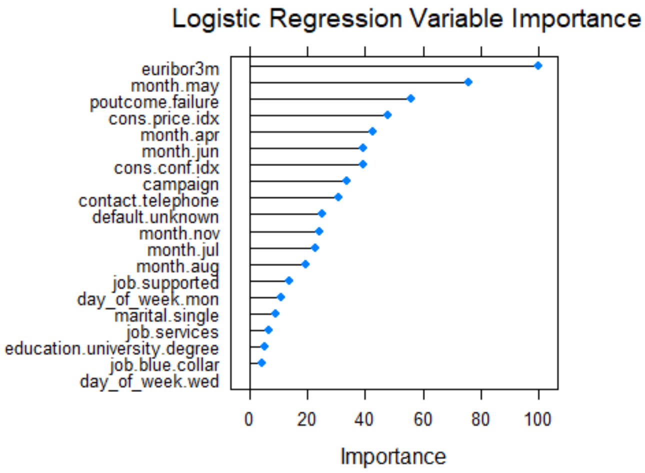
	age	year	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
nr.employed	-0.02	-0.94	0.14	0.37	-0.5	0.91	0.52	0.1	0.95	1
euribor3m	0.01	-0.91	0.13	0.3	-0.45	0.97	0.69	0.28	1	0.95
cons.conf.idx	0.13	-0.18	-0.01	-0.09	-0.05	0.2	0.06	1	0.28	0.1
cons.price.idx	0	-0.52	0.13	0.08	-0.2	0.78	1	0.06	0.69	0.52
emp.var.rate	0	-0.85	0.15	0.27	-0.42	1	0.78	0.2	0.97	0.91
previous	0.02	0.53	-0.08	-0.59	1	-0.42	-0.2	-0.05	-0.45	-0.5
pdays	-0.03	-0.39	0.05	1	-0.59	0.27	0.08	-0.09	0.3	0.37
campaign	0	-0.13	1	0.05	-0.08	0.15	0.13	-0.01	0.13	0.14
year	0.02	1	-0.13	-0.39	0.53	-0.85	-0.52	-0.18	-0.91	-0.94
age	1	0.02	0	-0.03	0.02	0	0	0.13	0.01	-0.02

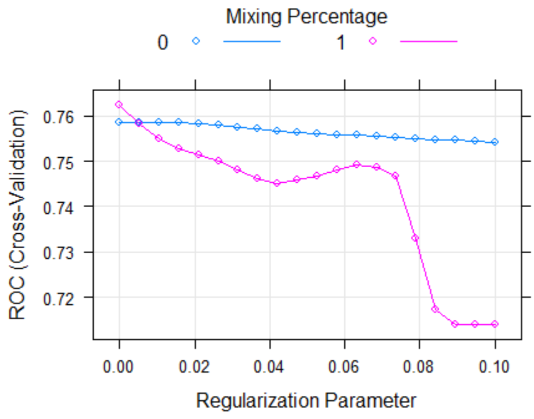

Cramer V of Categorical Variables

	job	education	default	housing	loan	contact	month	day_of_week	poutcome
poutcome	0.1	0.04	0.08	0.02	0	0.24	0.24	0.02	1
day_of_week	0.02	0.02	0.01	0.02	0.01	0.06	0.07	1	0.02
month	0.12	0.12	0.11	0.06	0.02	0.61	1	0.07	0.24
contact	0.13	0.12	0.13	0.08	0.03	1	0.61	0.06	0.24
loan	0.02	0.01	0.01	0.71	1	0.03	0.02	0.01	0
housing	0.02	0.02	0.01	1	0.71	0.08	0.06	0.02	0.02
default	0.15	0.16	1	0.01	0.01	0.13	0.11	0.01	0.08
education	0.44	1	0.16	0.02	0.01	0.12	0.12	0.02	0.04
job	1	0.44	0.15	0.02	0.02	0.13	0.12	0.02	0.1

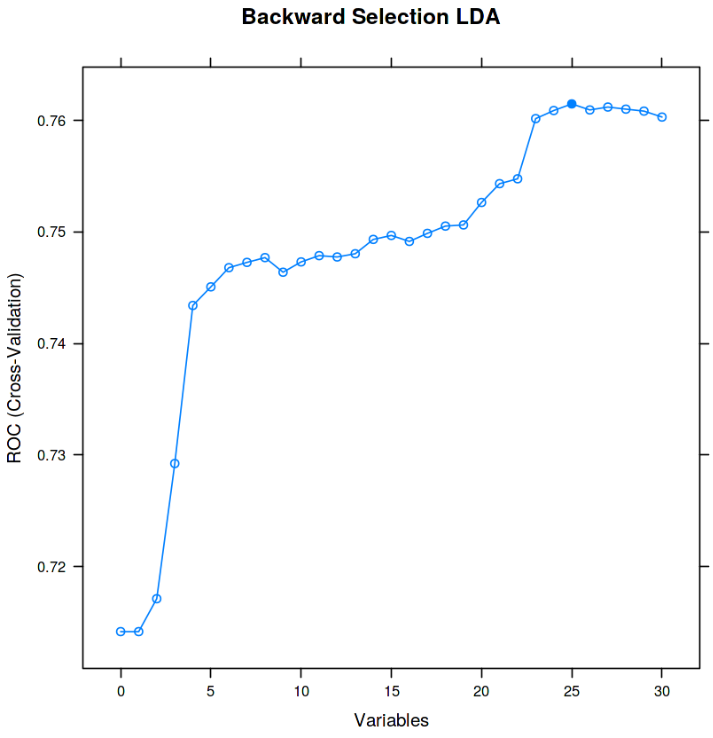
Appendix 5: Logistic Regression

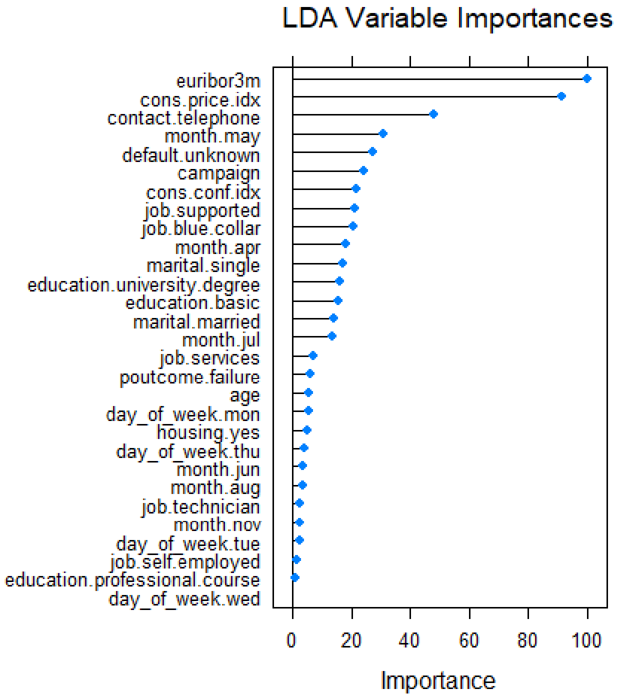

Backward Selection Logistic Regression


Logistic Regression Variable Importance

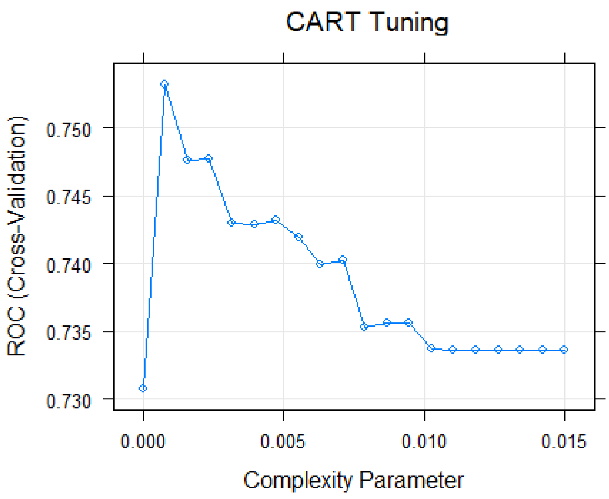


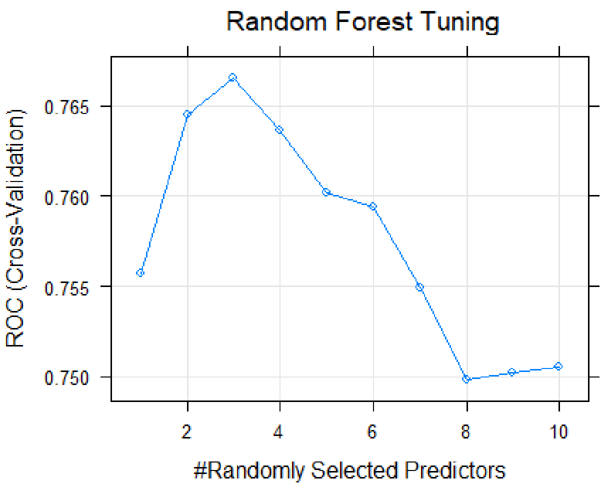
Appendix 6: LDA



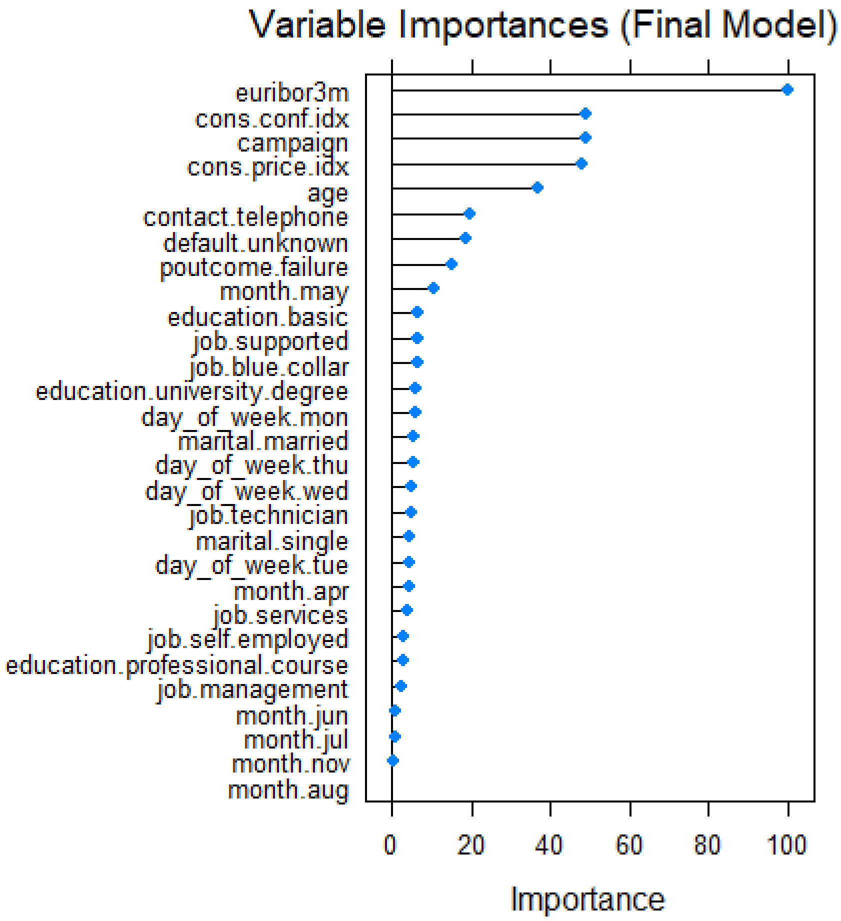


Appendix 7: Tree-based Methods





Appendix 8: Best Client Profile


Variable Importances (Final Model)
euribor3m
cons.conf.idx
campaign
cons.price.idx
age
contact.telephone
default.unknown
poutcome.failure
month.may
education.basic
job.supported
job.blue.collar
education.university.degree
day_of_week.mon
marital.married
day_of_week.thu
day_of_week.wed
job.technician
marital.single
day_of_week.tue
month.apr
job.services
job.self.employed
education.professional.course
job.management
month.jun
month.jul
month.nov
month.aug
Importance