7/8/2019

ACTL3142

Predicting Telemarketing Bank Customers

James To
Z5113921

Executive Summary

This report aims to predict the best customers and economic situation to target when telemarketing bank assets. The report will go into detail how the raw data was prepared for the prediction, and which models were tested for the best result. An overview of the average customer will also be explained, and the adjustments to the optimal predictive model will be explained.

The data had numerous issues to be addressed before a predictive model could be implemented. These issues included data obtained during the global financial crisis, missing values, imbalanced data, and others. It was concluded that missing values had to either be deleted or imputed using the best method possible. Imbalanced data was looked at in two different ways – SMOTE and random up-sampling – and both methods are discussed in the report. The global financial crisis involved a lot of disproportionate data, in the end the 2008 data was not used for predicting.

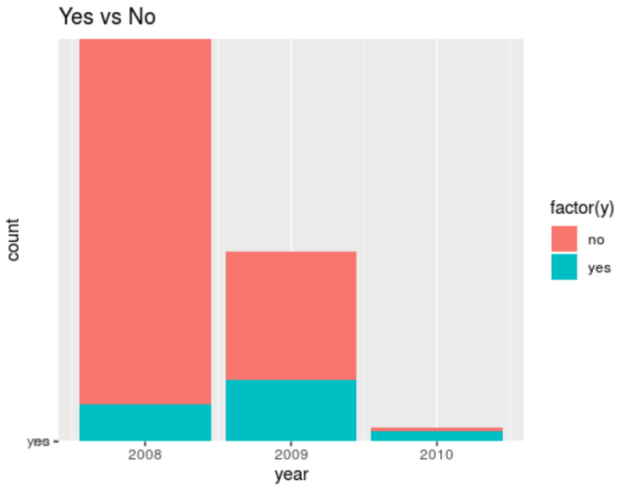
A total of 9 models are discussed in this report but only one was considered to be the best for predicting customers. The Generalized Boosted Model was the final model decided upon due to its high AUC, tuning capabilities, and decent computational speed. Other models discussed were Decision Trees, Logistic Regression and its variations, and Support Vector Machine Linear, Radial, and Polynomial. Furthermore, a few other models were tested, however, the were impractical due to computing issues. The model analysis will also go through the pitfalls of each model.

The average customer is then analysed in this report alongside a review of the optimal economic conditions for a successful sale. A discussion about the benefits of the final model and these parameters are also discussed.

Data Analysis and Preparation

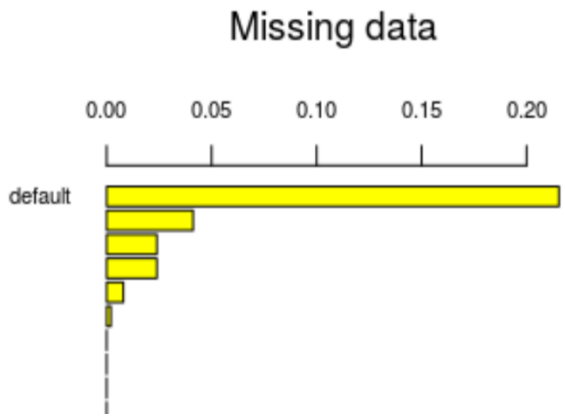
The raw data contains information on 39,870 customers and 20 variables for each customer - seven of which are personal details of the customer. Issues pertaining to the global financial crisis, missing data, imbalance of responses, and others are addressed in this section.

Global Financial Crisis



The global financial crisis turned the financial sector into turmoil when it struck, causing distrust between consumers and banks. In 2009, this trust began to recover and lower risk bank assets such as term deposits were being bought. This explained the difference between customer responses of 2008 and 2009; recession and recovery. A total of 1,339 'yes' responses were recorded in 2008, whilst there were 26,351 'no' responses. Furthermore, since 2008 was considered a recession, the data isn't as reliable as 2009 – where the economic situation was recovering – for predicting 2010 data points. Thus, due to the outlandish socio-economic conditions – which caused a highly disproportionate response data points – 2008 was deleted from the data set.

Missing Values

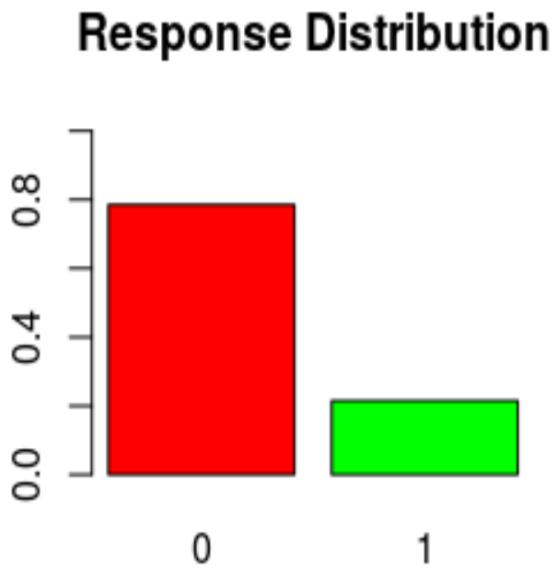


The default category was removed from the data set due to an abundance of missing values, and heavily disproportionate responses. There was a total of 12,509 missing values in the raw data, and 8,570 of those came from the default variable. Furthermore, the default category only had 3 'yes' responses among 31,307 'no' responses, which would have led an even greater disproportion of yes/no if imputed.

Deleting the other data points that had unknowns was not a viable option since it would leave just a fraction of the data set after the removal of 2008 data. Therefore, the data was imputed through the MICE package, which utilized the Predictive Mean Matching (PMM) method to impute the numerical data. The categorical variables in the dataset were made into dummy codes to impute. PMM selects the closest observed values to the missing data and randomly selects from them.

Imbalanced Data

After removing 2008, the dataset had 79% ‘no’ responses, and 21% ‘yes’ responses. This imbalance of data is an issue since a majority of algorithms are based on probability to classify data, and if there is a high imbalance the new observations will most likely be classified in the larger set.



Two methods were used to deal with the imbalance, random up-sampling and SMOTE. Random up-sampling takes the current data and adds the underweighted response points into the dataset till there is equivalent responses. The second method is an oversampling method, which takes the underweighted response variable and creates new datapoints. The main advantage of the random up-sample method is that it takes straight from existing data, however, this may cause overfitting, which contrasts with SMOTE which may make new datapoints that create a lot of noise *(He and Garcia, 2009)*. SMOTE was chosen for the majority of the modelling, however, technical difficulties often occurred, which led to switching to random up sampling method for other models.

Other Issues

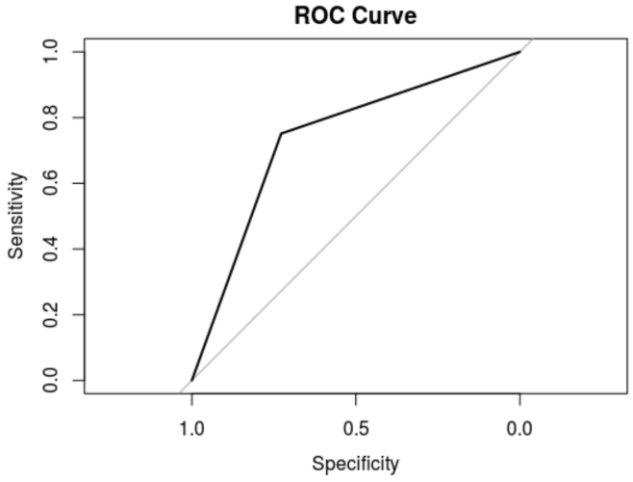
Some other issues that should be noted for modelling data is correlation between variables, and noisy data. The correlation between number of employees and the three-month Euribor rate is 0.9, making it an issue for models that cannot deal with collinearity. The dataset itself is troublesome due to a large number of variables that aren’t as significant as the data variables such as call duration, and directionality of the call (inbound outbound) – these variables were both present for *Moro, Cortez, and Rita (2014)* data analysis. Due to the noisy dataset, a more aggressive data partitioning was implemented of 80/20.

Models

It's noteworthy to mention that when comparing models, the most important statistic for this case scenario is AUC and false positive rate. The AUC is for binary classification in cases of skewed sample distribution (like this data set) since it doesn't overfit to the data (unlike accuracy). False positive rate also needs to be low, because this is a marketing campaign aiming to target customers with desired attributes. A higher false positive rate would result in selecting unwanted customers for this marketing campaign resulting in a loss in costs committed to selling to that customer.

Generalized Boosted Model

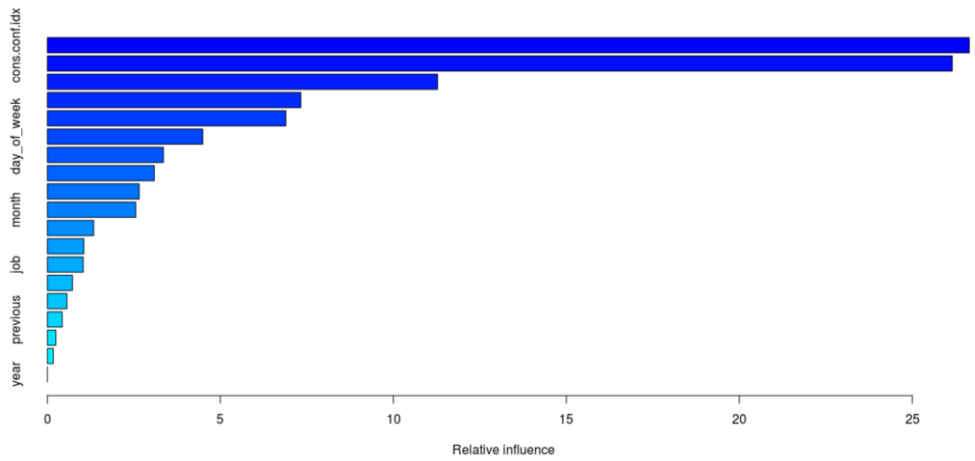
The Generalized Boosted Model (GBM) was chosen to be the best model for predicting the data. GBM is a tree-based model that adjusts weights on variables based on their important. GBM is particularly good with categorical data as it creates sub-models through decision trees.



The predictive model used for the training data set was set at cross-validation of 5 folds. The AUC returned was 0.7395, the highest of all models with cross-validation. An accuracy of 73.27% was achieved with this model, and a false positive rate of 24.86% - the lowest of tested models.

	Predicted No	Predicted Yes
Actual No	1381	129
Actual Yes	517	390

GBM adds weights to its variables in accordance to their importance. This is a double-edged sword in modelling as it makes the best model using its systemic algorithm but may tend to overfit to the data. The GBM model found the consumer confidence index to be the most influential variables, followed by the 3-month Euribor, number of days since the last call, number employed, and age. GBM also noted that the year of the call wasn't influential and removed it from the prediction.



Decision Trees

Decision trees are good models for predicting both qualitative and quantitative variables due to its flexibility, and nonlinear parameters don't affect its performance. However, decision trees often overfit to the data, capturing the noise.

Decision tree turned out having an AUC of 0.6237, despite an accuracy of 74.37%. This is unsurprising due to the overfitting disadvantage of using decision trees. The false positive rate of 58.87% depicts clearly that Decision Trees are not appropriate for this situation.

	Predicted No	Predicted Yes
Actual No	1596	312
Actual Yes	313	218

Logistic Regression

Logistic Regression is a good binary classification model that calculates the probability of a response belonging to a class. This isn't a computationally intensive model to use and is simple o implement. However, logistic regression creates a linear boundary, which is particularly detrimental for classifying noisy data.

The AUC returned for logistic regression with 5 fold cross-validation was 0.7098, and an accuracy of 72.28%. The false positive rate was 31.32%, which is not as good as GBM but much better than Decision Trees.

	Predicted No	Predicted Yes
Actual No	1399	166
Actual Yes	510	364

This model was further improved by removing variables with a high p-value – deeming them insignificant. The variables found with high p-values was jobs, housing, and loans.

This new logistic regression returned an AUC of 0.7731 without cross-validation, but only 0.7131 with 5 fold cross-validation. The discrepancy between these two numbers depicts the instability of the model between different groups, thus making it an unreliable model.

Logistic Regression with Lasso

Logistic Regression with Lasso (LAS) is a variation of logistic regression, which introduces the tuning parameter lambda to reject insignificant variables.

	Predicted No	Predicted Yes
Actual No	1386	156
Actual Yes	512	363

LAS returned an AUC of 0.7148, which is higher than ordinary logistic regression, and an accuracy of 72.36%. Furthermore, the false positive rate became lower with 30.30%.

Support Vector Machines – Linear/Radial/Polynomial

Support Vector Machines (SVM) takes the extremes of a boundary and makes a hyperplane to segregate two classes. Three ways to draw these hyperplanes are known as kernels linear, radial, and polynomial, and each kernel becomes more difficult to compute.

	Linear	Radial	Polynomial
AUC	0.691	0.717	0.698

The Radial SVM had the best AUC among the three, and this could be attributed to Radial handling noisy data a lot better than the other two. In addition, the false positive rate 29.96% making it fairly suitable for predicting telemarketing, but not good enough to beat GBM.

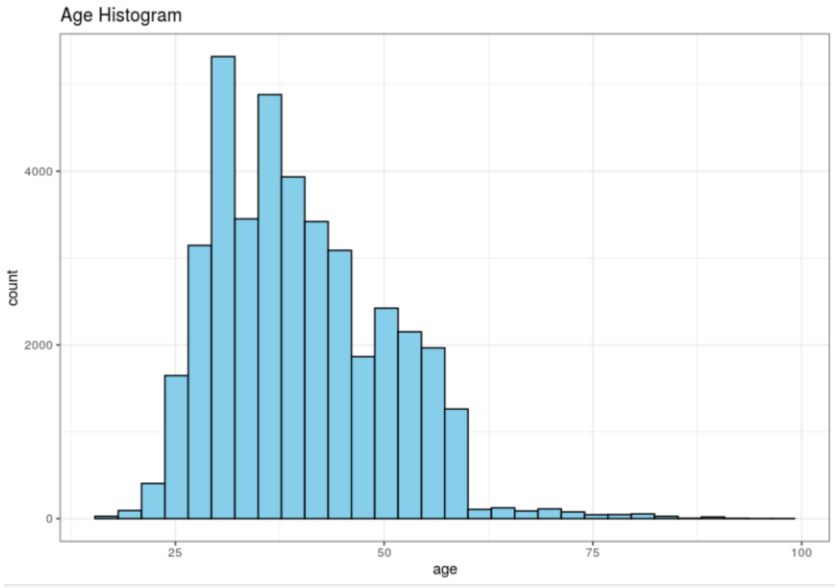
	Predicted No	Predicted Yes
Actual No	1380	157
Actual Yes	518	367

Analysis of Selected Customers

Using the Generalized Boosted Model, 500 clients were selected from the Evaluation dataset. By analyzing the customers, it was found that people with the following attributes were best target market for the selling bank assets:

Target Demographic	
Age	42
Job	Administration
Marital Status	Married
Education	University Degree
Housing	Yes
Outstanding Loan	No

Seemingly, the target demographic for improving sales is middle aged married individuals who have had university education. Noting that GBM placed heavy influence on the bank should ensure it's calling the right age demographic for success. Age The standard deviation of age for the predicted clients was 16, thus, campaigning to ages 26-58 is an appropriate choice of customers.



The ideal economical situation is also as follows:

Ideal Economic Situation	
Employment Variation Rate	-1.3666%
Consumer Confidence Index	94.11475
3-Month Euribor	0.9069%

By inspection, when economic prosperity is high, the likelihood of success increases dramatically. GBM weighed the 3 month Euribor to be the most significant factor, and logically this makes sense as the

higher the Euribor, the higher the bank assets interest rates can be. Though the bank is limited in its ability to increase the saving rate, if they can maintain higher interest rates, then they will be able to send more bank deposits.

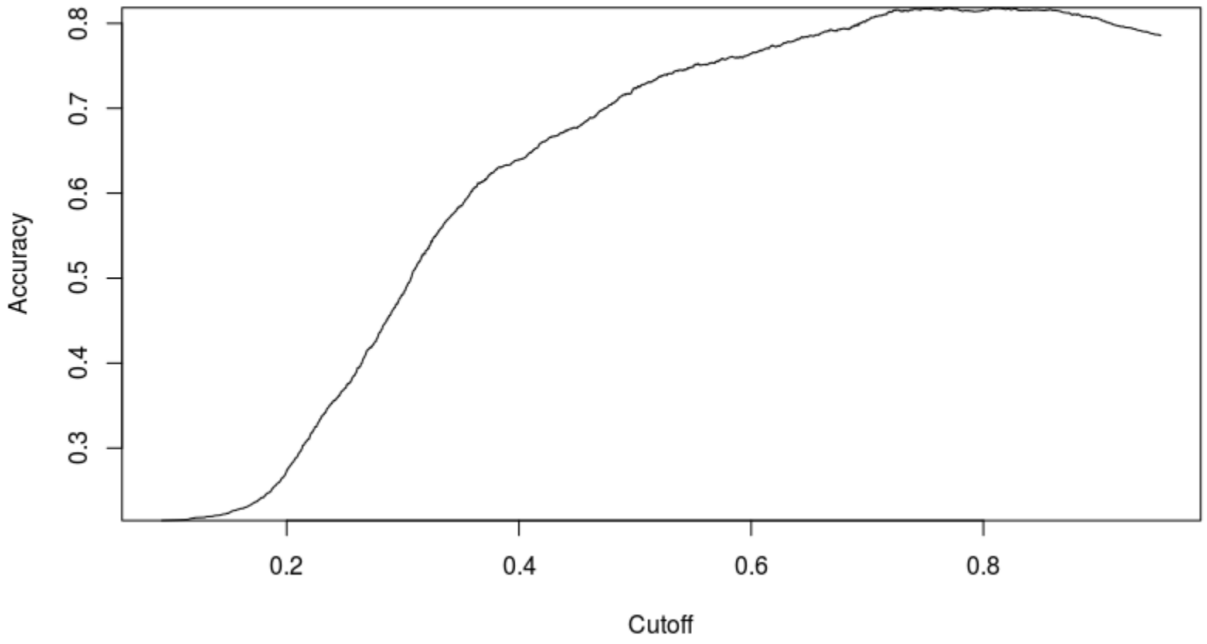
Variables that can improve the banks ideal marketing strategy:

Ideal Marketing Situation	
# of Days Last Called	622
# of Days Last Called Excluding New Customers	7.5
# of Employees	4966
Day of Week	Thursday

The GBM also put heavy importance on the number of employees, and the number of days last contacted. To ensure a successful telemarketing campaign, the bank need to maintain its number of employees, keeping them around 5,000. Furthermore, calling once a week for existing customers will also help improve sales, as GBM predicted that those existing customers who were contacted more often were likely to be positive responses.

In conclusion, the Generalized Boosted Model was implemented for the purposes of classification prediction. Based upon its inbuilt algorithm, a recommendation of strategy for the banking marketing campaign was placed surrounding the customer, economic situation, and bank's situation.

Appendix



Accuracy of LASSO