TELEMARKETING PREDICTION

Prepared for the Union Bank of Australia

Name: Samuel Ting
Student ID: z5114381
Date: 07/08/2019

EXECUTIVE SUMMARY

This report provides an analysis of the characteristics that contribute to a successful telemarketing call for long-term deposits and a statistical model to predict high likelihood campaign call clients. The models considered and applied within this study include the Logistic Regression, Gradient Boosted Machines (GBM), Support Vector Machines (SVM), Decision Trees and Neural Networks. Recommendations for the final model include:

- Use of Logistic Regression to maximise both sensitivity (true positives) and specificity (true negatives).
- Use of GBM to maximise sensitivity when cost of unsuccessful calls is not within the campaign scope.

The bank also requested a non-technical description of the client and economical characteristics for a successful telemarketing campaign. The top 5 characteristics for a higher probability of making a long-term deposit include:

- Age: Low and high aged clients (excluding middle aged).
- Job: Students and Retired Clients.
- Day of Week: Calls made between Tuesday and Thursday.
- Campaign: Clients who have least been contacted during the respective campaign.
- Euribor: A low 3 month Euribor (short-term interest rate).

Finally, the report also discusses the limitations within the study and further recommendations for improvements.

- Main limitation within this study was the removal of 2008 GFC data, 70% of the training set.
- Memory and hardware limitations reduced algorithm choice.
- Higher quality data will enable Neural Networks which have higher predictive accuracy in detecting true positives and will also allow unsupervised learning algorithms such as k-means clustering to provide more distinguished successful characteristics.

I. DATA PRE-PROCESSING AND EXPLORATION

A. Imputations and Imbalanced Data

The pre-processing stage of the data began with the identification of missing attribute values and the imbalanced distribution of successful (yes) and non-successful (no) telemarketing calls for long-term deposits. Missing attribute values needed to be removed or replaced with another value before fitting statistical learning algorithms. Removing selections of data with missing attribute values remove a large proportion of the data and as such not optimal considering the required flexible models which require more data to train the higher number of parameters. Therefore, multivariate imputations replaced missing values using Logistic Regression for binary predictors, Bayesian Polytomous Regression for factor variables greater than two levels and Proportional Odds model for ordered greater than two levels predictors (Yadav and Roychoudhury, 2009). The Default predictor was removed as it had majority unknown values and would not contribute significant information to the statistical algorithms.

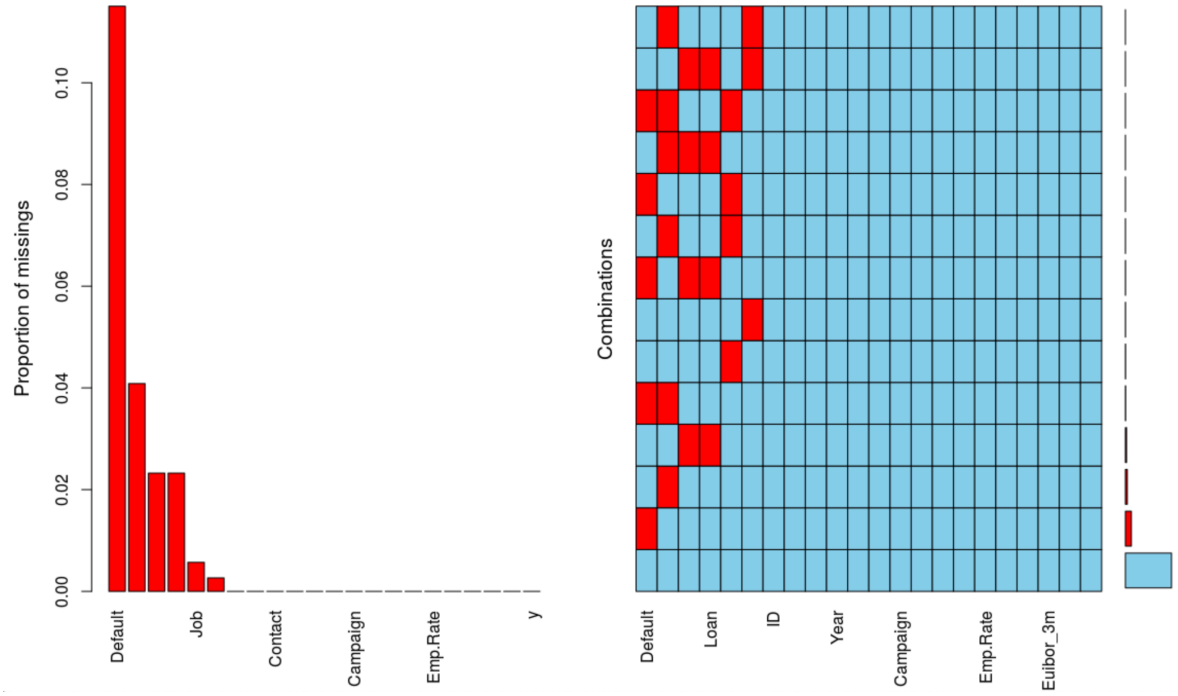

Figure 1: Proportion of missing attributes

Imbalanced data can be detrimental during the training process of statistical learning algorithms as predictions would be biased towards the majority class. The imbalanced data can lead to many false negatives as the algorithm prediction will predict majority to be a false class. These false negatives will result in large costs for the bank particularly within the scope of telemarketing to only 500 clients. Therefore, we balance the data using Synthetic Minority Over-Sampling Technique to balance the proportion of the classes through adding unique synthetic samples to the minority class and removing samples from the majority class (Haibo and Garcia, 2009).

Proportion	Yes	No
Before resampling	11%	89%
After resampling	50%	50%

Table 1: Over-sampling

B. Data Exploration

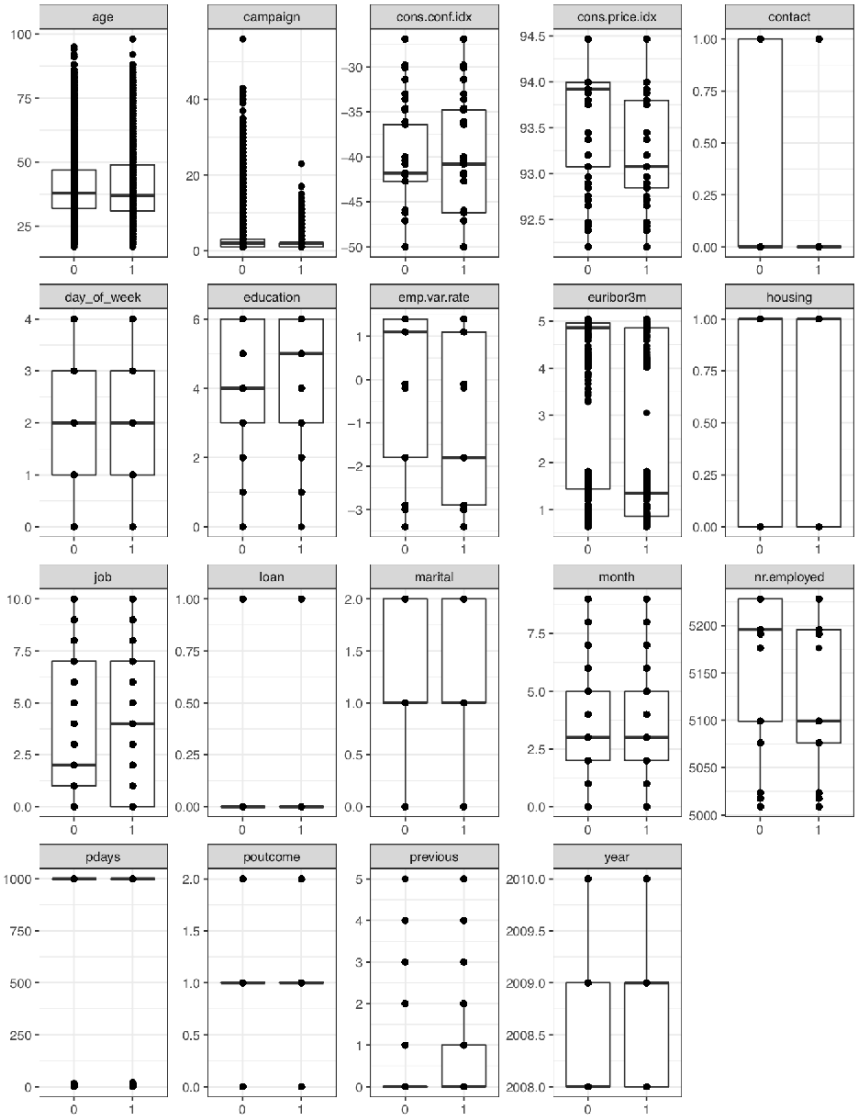

Figure 2: Predictors

After initial feature exploration, the predictors of interest include social and economic indicators euribor 3m, campaign, job, and education. Figure 2 demonstrates that whereas a higher consumer confidence index is an indicator for a higher likelihood of client making long-term deposits, higher consumer price index, employment variation rate and Euribor rate are lower likelihood indicators. The results contradict expectations of lower Euribor to result in decline in savings can be explained by the GFC in 2008 as consumers may feel the need to consider saving for the future in times of economic depression (Moro, Cortez and Rita, 2014).

Furthermore, a higher CPI (inflation rate) also implies lower deposits which is reinforced with the low interest rates as the real return on long-term investments such as the bank's deposits would be low to negative. With the evaluation predictions to be made in 2010 would be significantly biased using the same predictors, the year 2008 was removed from training of algorithms.

Predictor groupings of Other and Client characteristics are of interest such as campaign indicating the number of contacts performed during the respective campaign for the client and the client's education level. Lower campaign contacts for a client show higher likelihood of a successful outcome. A higher education level also indicated a higher likelihood of making long-term deposits with 12% of university clients against 9% of those with a basic four-year education. A standardization transformation was applied due to highly skewed nature of the predictors which ultimately increased hyperparameter tuning and computation speeds of training algorithms.

II. TECHNICAL MODELLING

The analysis of Client, Socio and Economical characteristics was performed through feature selection using Principal Components Analysis and Random Forest to perform implicit variable selection, identifying which parameters would have high influence on the fitted statistical algorithms. Algorithms considered include Logistic Regression (LR), regularized LR using Lasso and Ridge, Gradient Boosted Machines, Decision Trees and Neural Networks (refer to section III technical appendix). Removing 2008 data resulted in a 30% remaining dataset, which meant that flexible models such as neural networks would perform poorly from insufficient training data. In describing the successful-deposit characteristics, interpretable models were better suited.

A. Fitted Model Results and Summary

The final model is the Logistic Regression, which was followed in performance by an ensemble classifier, GBM. As seen in Table 2, the Logistic Regression displays a slightly lower ROC & AUC than the GBM. This can be explained by the GBM being a comparatively more flexible model using an ensemble of decision trees which consequently sacrifices its interpretability. The Logistic Regression, on the other hand, is less flexible but exceeds in interpretability and also reduces variance of estimates. With greater interpretability, the statistical model reveals greater insight into the characteristics contributing to a successful telemarketing call. Furthermore, the fitted Logistic Regression was unregularized to increase predictive capacity as a 10-fold cross-validation using Lasso and Ridge revealed the optimal penalty parameter as 0.006 implying an extremely low penalty should be applied.

	Logistic Regression	**Gradient Boosted Machines**
AUC	72.9%	74.0%
Sensitivity (TP Yes)	70.3%	73.9%

Specificity (TN No)	75.5%	73.0%

Table 2: Final fitted models

The Logistic Regression has significantly higher specificity which implies a higher probability of correctly predicting the true negative case of an unsuccessful call. The Logistic Regression has lower sensitivity which implies it has lower ability to predict a successful call. After the GFC many banks are pressured to increase their long-term deposits so increasing costs from making unsuccessful calls may be insignificant (Moro et al., 2014). As the bank requires only 500 likely clients to reduce the costs of unsuccessful contacting, a higher specificity is desired. Accordingly, the GBM model with a superior sensitivity should be undertaken proportional to the cost limitations of the bank. Table 3 provides advantages and disadvantages of other models fitted within this study.

	Advantage	**Disadvantage**
Decision Trees	- Interpretability of successful characteristics - Low computation intensity	- Biased towards high variance predictors - Overfit
Random Forest	- Accurate - Less variance and overfitting than a single decision tree	- Computationally expensive learning - Memory intensive
SVM	- Regularization capabilities - Overfitting robustness	- Difficult to interpret - Requires feature scaling - Computationally expensive for hyperparameter tuning - Memory intensive
Neural Networks	- Accurate with large amounts of labelled data	- Difficult to interpret - Requires feature scaling - Requires large amount of labelled data

Table 3: Fitted model's advantages and disadvantages

B. Feature Selection and Successful Telemarketing Characteristics

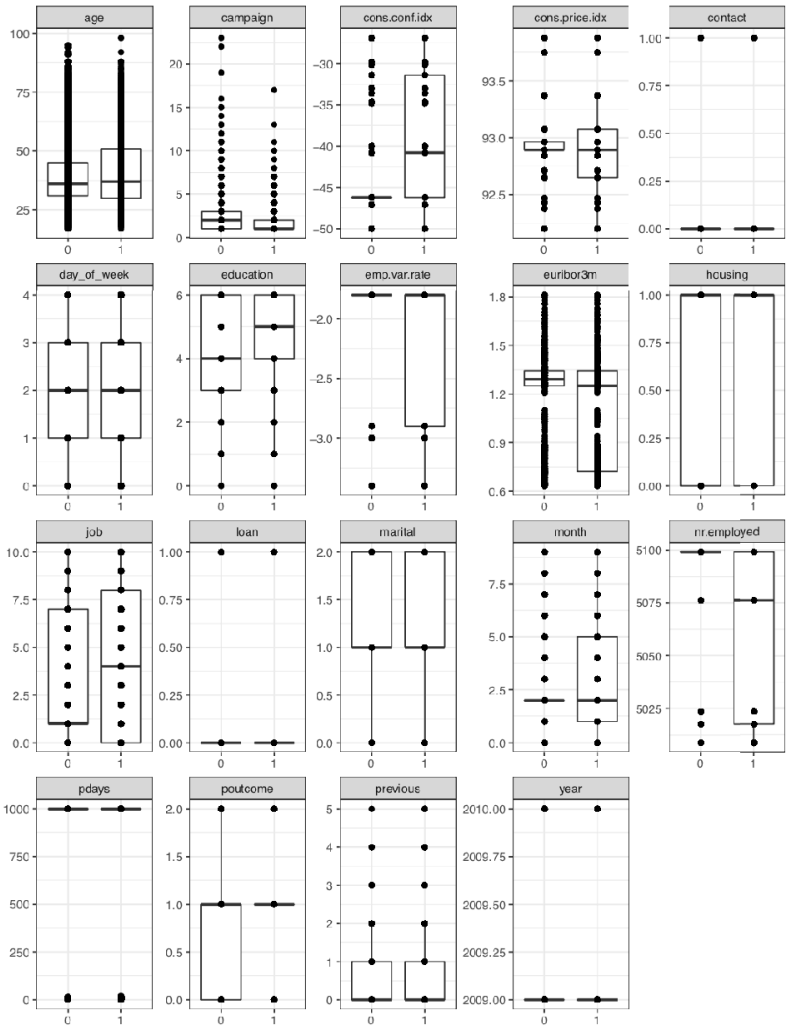

Figure 3: Predictors and Response excluding 2008

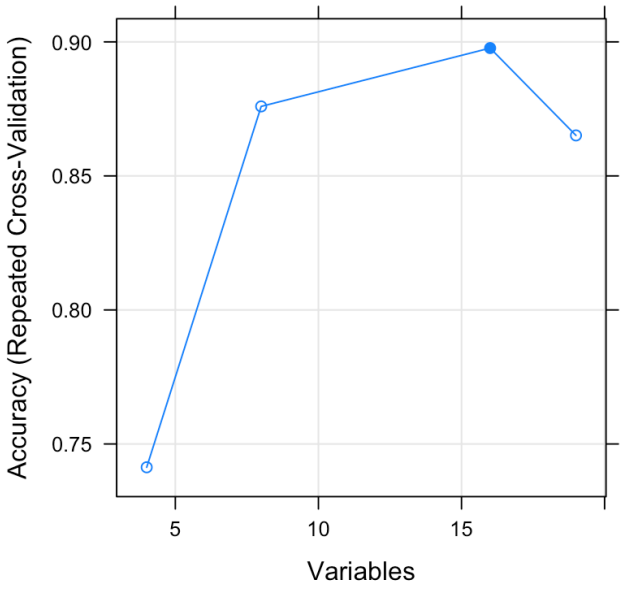

Figure 4: Recursive Feature Elimination

Feature selection and analysis was performed through implicit variable selection on 19 variables excluding ID and Default using random forest algorithm and a principal component analysis. The recursive feature elimination feature selection technique which uses iterative random forests displays the optimal variable count at 16 using 10 cross-validations. The variables excluded within the feature selection were emp.var.rate, year and nr.employed. The exclusion of only 3 features is supported by principal component analysis as 95% of variance was captured in the 32 of 47 principal components where categorical factors were created as dummy variables. Employment variability rate also demonstrated collinear relationships with CPI, CCI and Euribor showing 0.64, -0.88 and 0.67 correlations respectively. When removing the effects of economic effects of the GFC in 2008, an interpretation is that year itself does not contribute to higher call success.

The top 10 variables from recursive feature elimination is shown in Table 4:

Age	Education
Job	Housing
Day of Week	Marital
Campaign	Loan
Euribor	Contact

Table 4: Top ten implicit variable selection features

	Statistic	
Age Band (yes %)	**High** - A67+: 50%, A47-57: 16%	**Low** - A37-57: 7%
Job (yes %)	**High** - Student: 28%, Retired: 22%	**Low** – Services: 7%, Blue-collar: 7%

Day of Week (yes %)	**High** - Thu: 22%, Tue: 21%	**Low** – Mon: 18%, Fri 19%
Campaign Median	**Yes**: 2.08	**No**: 2.65
Euribor Mean (excluding 2008)	**Yes**: 1.07	**No**: 1.24
Education (yes %)	**High** – Illiterate: 22%, University Degree: 12%	**Low** – Basic 9Y: 7%, Basic 6Y: 8%

Table 5: Successful and unsuccessful call characteristics

The characteristics of a successful telemarketing call for long-term deposits is primarily separated into characteristics of the client and of the current economic climate. Age is a primary influence on the outcome of a deposit call with 50% of those aged above 67 making long-term deposits after a campaign. Table 5 also displays a parabolic relationship with Job and Education to success; that the highest successes are within young students, old retirees, illiterates and those with university degrees. Age acts as a likely confounding variable for these other factors. There is also higher likelihood of successful calls on days other than Monday and Friday as people may not appreciate being contacted during those workdays. Campaign median is higher for unsuccessful calls and suggests the bank contact those who have been contacted the least during the respective campaign. Finally, the economic indicator of the short-term Euribor mean is lower for successful calls indicating low short-term interest rates imply an upwards yield curve with higher long-term interest rates and as a result, more incentive for long-term deposits (between 2008 and 2010).

C. Limitations and Recommendations

Removal of 2008 data due to the GFC drastically decreased the sample dataset's size by 70%. As a result, a lower training dataset implied the limited use of highly flexible methods which require more data to train the many parameters to predict a higher probability of successful clients. A lower sized training set limited flexible learning algorithms such as Neural Networks within this study as the model was applied to a Portuguese bank with 50,000 training points and provided a higher 0.79 AUC (Moro et al., 2014).

The results of this study for the bank present the applications of the Logistic Regression model within a cost-constrained telemarketing campaign against the more flexible, ensemble GBM classifier for a higher predictive accuracy of true positives. The Logistic Regression and Decision Tree's enabled a profiling of successful telemarketing characteristics for long-term loans due to their interpretability. Further recommendations consider the limitations of the 2008 data and suggest increasing the amount of quality dataset to enable an accurate Neural Network, which had the highest sensitivity but lowest specificity, to be trained and also unsupervised methods such as k-means clustering algorithm to group the successful characteristics (Jayabalan, 2017).

III. TECHNICAL APPENDIX

	AUC	Accuracy	Specificity (TN No)	Sensitivity (TP Yes)
Logistic Regression Lasso	72.1%	72.9%	73.6%	70.52%
Logistic Regression Ridge	71.68%	72.8%	73.6%	69.8%
SVM: Linear	70.6%	74.3%	77.0%	64.0%
SVM: Radial	71.0%	78.3%	99.5%	0.38%
SVM: Polynomial	71.2%	72.5%	76.0%	54.0%
Decision Trees	73.0%	73.0%	73.0%	72.0%
Neural Networks	68.4%	59.2%	52.2%	84.6%

Table 1: Model Results

Each model fitted using a 10-fold cross-validation grid search for hyperparameter tuning to improve results. Data was also standardized to increase computation speed for models such as SVM and Neural Networks.

A. Logistic Regression Regularised

Logistic Regression is a statistical model that uses a logistic function to model a binary dependent variable. Lasso and Ridge are regularization techniques to penalize the logistic regression's complexity, avoiding overfitting. The optimal penalty parameter was approximately 0.006 for both Lasso and Ridge indicating almost no complexity penalty. The low-cost parameter was used to increased predictive accuracy.

RESPONSE	NO	YES
NO	1396	153
YES	502	366

Table 2: Confusion Matrix Logistic Regression Lasso

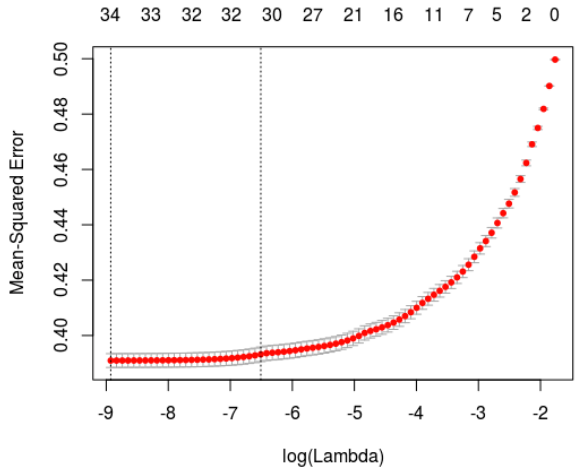

Figure 1: Lasso Lambda Tuning

RESPONSE	NO	YES
NO	1396	153
YES	502	366

Table 3: Confusion Matrix Logistic Regression Ridge

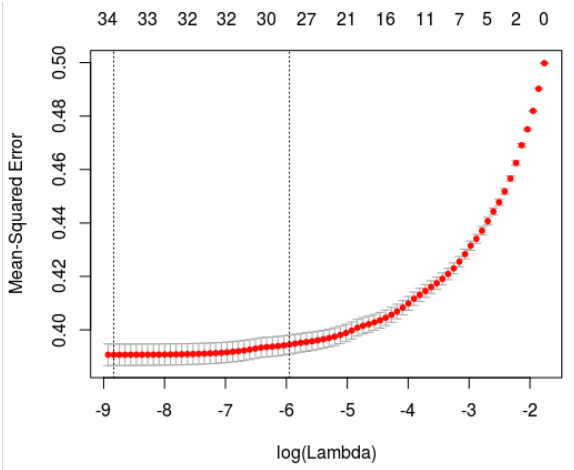

Figure 2: Ridge Lambda Tuning

B. Support Vector Machines (Linear, Radial and Polynomial)

Support vector machines is a supervised learning technique that use points in space, mapped so that the points are divided by a maximum clear gap. The optimal tuning parameter was a cost penalty of 0.01 to maximise their predictive accuracy.

RESPONSE	NO	YES
NO	1465	187
YES	433	332

Table 4: Confusion Matrix SVM Linear

RESPONSE	NO	YES
NO	1890	517
YES	8	2

Table 5: Confusion Matrix SVM Radial

RESPONSE	NO	YES
NO	1455	239
YES	443	280

Table 6: Confusion Matrix SVM Polynomial

C. Gradient Boosted Machines

Gradient Boosted Machines is an ensemble method of refitting decision trees and increases the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The parameter tuning grid search optimised 150 trees with a depth of 3 and a shrinkage of 0.1 providing the model high predictive accuracy.

RESPONSE	NO	YES
NO	1386	135
YES	512	384

Table 7: Confusion Matrix GBM

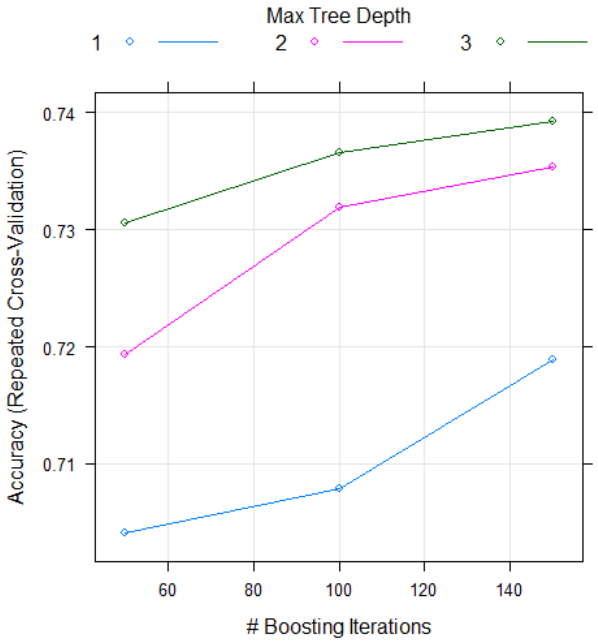

Figure 3: GBM Boosting Iterations

D. Decision Trees

Decision trees is an algorithm that only contains conditional control statements. The decision tree's classification tree was analysed in exploration of successful characteristics due to its extreme ease of interpretability. The optimal cost parameter was 0.00157 which enabled the decision tree to become more complex by increasing the size of the tree. The model was not used as the final fit due to an inferior AUC.

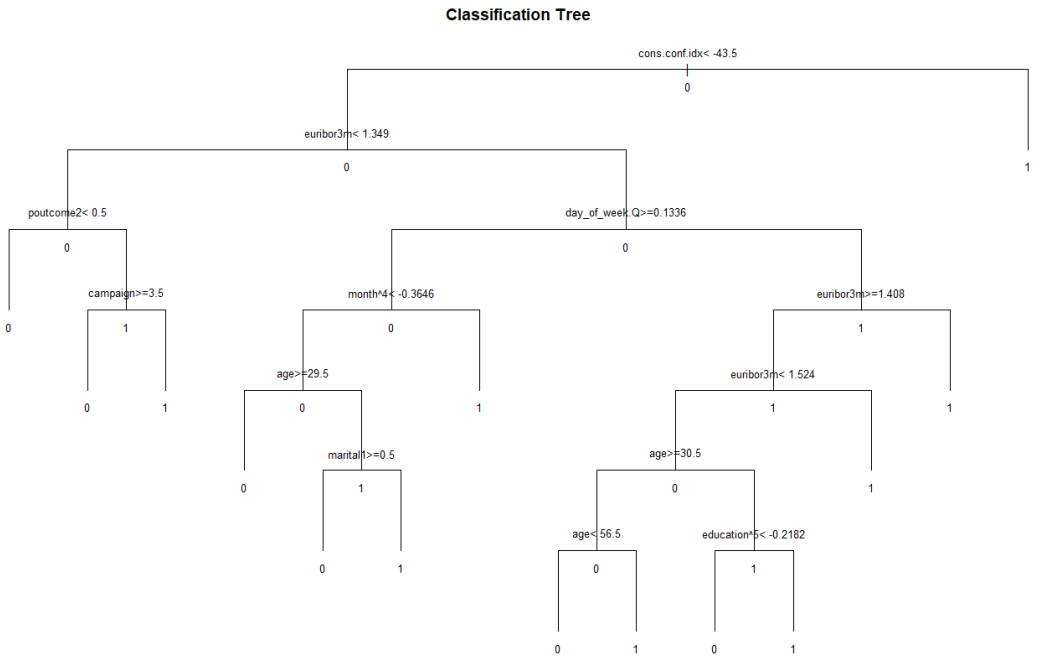

Figure 4: Classification Tree

RESPONSE	NO	YES

NO	1391	143
YES	507	376

Table 8: Confusion Matrix Logistic Regression Lasso

E. Neural Networks

Neural Networks are an adaptive system that change its structure based on information that flows through the network and used for non-linear statistical data modelling. The NN had the lowest overall accuracy rate but highest specificity which implies it has a higher predictive ability to distinguish True Positives (Yes). The bank may want to pursue using this model if making unsuccessful call's is not a significant cost.

RESPONSE	NO	YES
NO	1075	94
YES	823	425

Table 9: Confusion Matrix Neural Network

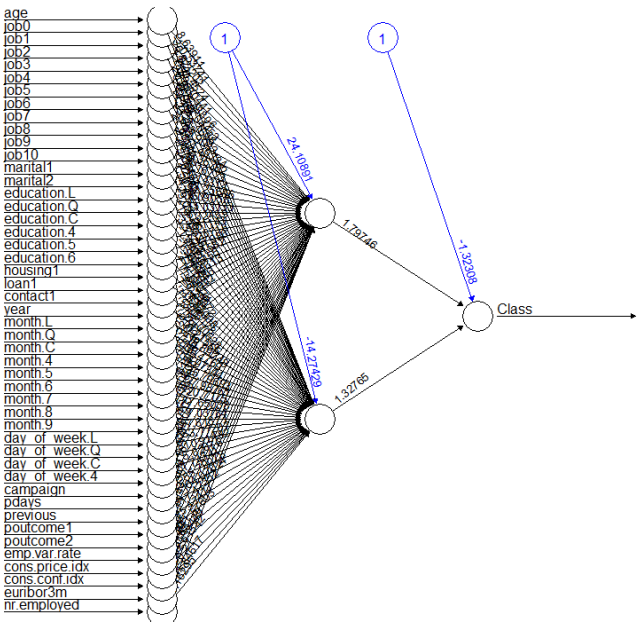

Figure 5: Neural Network

F. Principal Component Analysis

Principal Component Analysis is a dimensionality-reduction algorithm to capture the highest variance in a dataset. The graph below displays that it takes approximately 32 of 47 principal components to capture 95% of the data's variance.

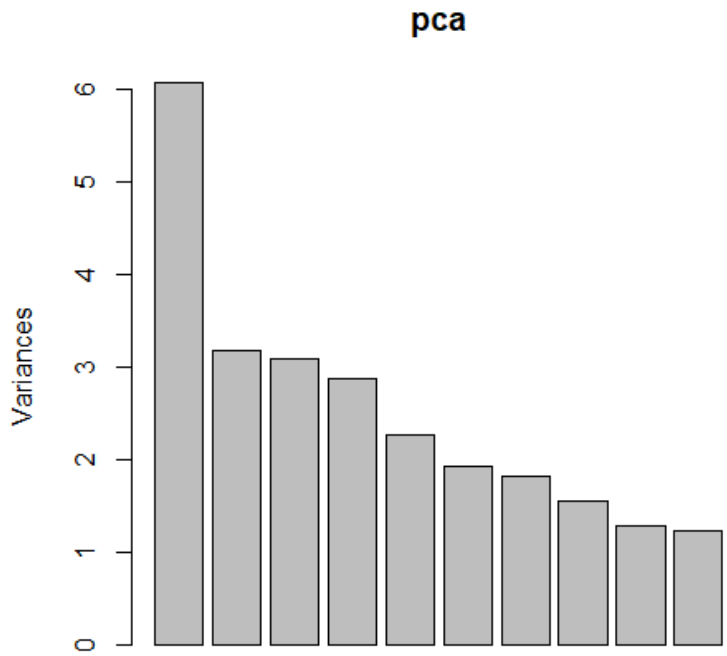

Figure 6: PCA Variance

IV. REFERENCES

Haibo He and Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp.1263-1284.

Jayabalan, Manoj. (2017). Predicting Customer Response to Bank Direct Telemarketing Campaign. *IEEE*,

Moro, S., Cortez, P. and Rita, P. (2019). *A data-driven approach to predict the success of bank telemarketing*.

Yadav, M. and Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, 160, pp.104-118.