Decrypting Customer Behaviour: Implementation of Logistic Regression to Predict Subscription Term Deposit.

By Justin Grima (14248599)

I. Abstract

Marketing is important for the success of companies such as banks and strategies like "directed marketing through a focus on targets that assumable will be keener to that specific product/service" (Moro et al., 2011)[1] are adopted in hope of increasing product interest and success, such as subscription term deposits. However, customers' "generally evince a negative attitude toward banks using direct marketing strategies" which can result in "low intention to purchase as an outcome of direct marketing" (Page & Luding, 2003)[2]. Advances in technology and data analysis provide new and exciting opportunities to understand customer behaviour and predict outcomes. The **purpose** of this report is to analyse the "bank-full.csv" dataset, which contains "marketing campaigns of a Portuguese banking institution" (Moro et al., 2011)[1], to **i)** investigate and identify significant variable to **ii)** develop a binomial generalized linear model (GLM) to predict subscription term deposits. The **methodological approaches** use RStudio to conduct data exploration, feature selection, regression analysis, and model refinement, summarizations, and visualisations

The **results** of this study found that the **i)** binomial GLM was effective in predicting subscription term deposit, and ii) a thorough exploratory data analysis (EDA) revealed that the marketing campaign was unsuccessful in generating high subscription term deposits among the bank's clients. Overall, we can **conclude** that by using a subset of variables from the 'bank-full.csv' dataset, a successful predictive model with approximately 90% accuracy can be used to predict customer term deposit subscription (CTDS) and alternative approaches, such as offering better interest rates or additional features can produced more promising results for the campaign. Additionally, the bank can explore offering new products or other existing products to current and potentially new clients.

II. Introduction

Technology enables easy access to various products, services, entertainment, and information through the internet, making companies more competitive as they strive to acquire and retain customers. In the banking industry, competition has evolved from regional to global with "the entry of foreign banks into new markets irrespective of physical presence" while "consumers are now more informed, face lower switching costs, and are showcasing an ever-increasing set of diverse needs. The combination of these various forces has escalated the competitive forces faced by retail banks" (Neilson & Chadha, 2008)[3]. Due to fierce competition and the financial crisis, banks are under pressure to increase customers financial assets or risk losing them to better offers. Offering term deposit subscriptions is one possible solution which provides "a way to invest your money with an authorised deposit-taking institution (ADI) and earn a fixed rate of interest. Your money is locked away for the time that you choose (the term)" (Term Deposits - Moneysmart.gov.au, n.d.)[4] and a minimum deposit amount is required. The objective of the direct marketing campaign was to offer term deposit services to customers in order to boost their financial assets, retain customers, and ensure the success of the company. This report uses the "bank-full.csv" dataset to select a suitable GLM to create, refine, train, and test the model (with high predictive power) to predict term deposit subscriptions. It also provides insights into the effectiveness of the marketing campaign, provide improvements, explore alternative marketing campaign avenues and products.

III. Data

To conduct the study, we will be using the "bank-full.csv" dataset, as stated in the abstract, which has been original sourced from Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012 and can be downloaded from "UC Irvine Machine Learning Repository" (UC Irvine Machine Learning Repository, 2012)[5]. The population from which the sample/ data has been drawn from are customers who are members of a Portuguese bank. The data was collected by Paulo Cortez and Sérgio Moro from a "Portuguese bank that used its own contact-centre to do directed marketing campaign" from "17 campaigns that occurred between May 2008 and November 2010" (Moro et al., 2011)[1], specifically from the concluded contact group. The sample size for this dataset is 45211 instances for which we can assume a simple random sampling process was conducted: "each item in a population has an equal chance of inclusion in the sample" (Statistics, 1998)[6]. Using RStudio functions (see Appendix A for raw code) we conducted an EDA of the dataset and found there to be 45211 observations and 17 variables: 7 'integer' class, continuous numerical variables, 3 'character' class, binary categorical variables, 6 'character' class, multi-level categorical variable and one response variable which is a 'character' class, binary categorical variable (see Appendix B for full variable description). There were also 0 duplicates and no missing values. No interventions, treatments, grouping structures or procedures invoked to reduced sampling bias or sampling error was conducted in this study. The original study describes their process for variable and observation elimination, see Appendix C for full breakdown.

IV. Methods

The following report was conducted in RStudio: Elsbeth Geranium, version 2022.12.2+576. This report uses several methods to create an effective GLM model for predicting CTDS. As mentioned in the 'Data' section an EDA was conducted and found the response variable to be an ungrouped binary categorical variable. Therefore, we use a binomial GLM to predict subscription term deposit for a Portuguese bank. A binomial GLM has three components: random component, link function and systematic component. The random component for an ungrouped binomial GLM is y_i ~ Bin(1, π_i): y_i is our response variable with a binomial ('Bin') distribution, the sample size is 1: one person per outcome, and π_i is the probability of a CTDS. Therefore, our expected value of y_i is π_i: ($E(y)= \pi_i$) (see Appendix D for full breakdown). The link functions: g(.), for binomial GLM are 'logit', 'probit', 'Cauchy', 'cloglog' and 'log', each creating different links between random and systematic components. By modelling all link functions, we compared and evaluated the most effective choice, using the Akaike information criteria (AIC) score as the determinant; AIC chooses the model with the smallest Kullbak – Leibler divergence which is essentially the distances between the probability density function (PDF) of model *M* and the data. (See Appendix E for mathematic breakdown). Overall, the most practical link function for this study was a 'logit' link function with a trade-off between a slightly higher AIC score, but an easier model to explain versus a model with a lower AIC score, but more complex model to interpret and explain ('probit'). Therefore the 'logit' link function with our random component give us $g(\pi_i) = \log(\pi_i /1- \pi_i)$ where g(.) is the link function, π_i is our expected value of y ($E(y) = \pi_i$): the mean for a binomial distribution. Finally, we have the systematic component, $n_i = (\Sigma_j x_{ij}\beta_j)$ which is essential a linear component $\beta_0+ \beta_1x_1+ \beta_2x_2\ldots \beta_nx_n$. Therefore we have a binomial GLM with a logit link function where we model the log of odds using the linear component: $\log(\pi_i /1- \pi_i) = \beta_0+ \beta_1x_1+ \beta_2x_2\ldots \beta_nx_n$. This then associates the magnitude of beta with the log of odds. Our final model mathematical notation, using all 3 components, is $g(\pi_i) = \beta_0+ \beta_1x_1+ \beta_2x_2\ldots \beta_nx_n$. (See Appendix F for rearrangement of model to solve for π_i).

To complete the model, we need to choose the appropriate variables to include. This was done by running the binomial GLM in R with the response variable and all other appropriate variables as the predictors. Using the summary() function, we analyse each predictor variable; estimate (β) values, standard error, z-value and p-value to find variables that have a significant influence on the response variable; determined by their p-value being less than 0.05. Variables with a p-value greater than 0.05 infers that their estimate values are not significantly different from zero, do not influence the response variable and were removed from the model. Further details regarding variable selection are mentioned in Results and Discussion (see appendix G for raw code and initial binomial GLM summary output). To verify our newly selected model we can use the likelihood test ratio; comparing the complex and simple (nested) model based on their difference in deviance ($\sim X^2_{p1-p0}$), and the backward elimination process (stepwise method in R, see Appendix H for raw code and explanation). Based on the results using both methods (see results and discussion for full analysis), we are reassured in the model chosen.

Next, the multicollinearity assumption is checked using variance inflation factor (VIF): the ratio of variance for estimated parameters in a model, that includes several other terms, by the variance of a model made using only one term ($VIF_j=1/(1-R^2_j)$). With a maximum variance value of 1.19 (see Appendix I for raw code and output) we can confirm no multicollinearity. Using maximum likelihood estimation (MLE); $L(\pi,n:y) = \Pi_i(n_i/n_iy_i)\ \pi^{niyi}(1- \pi_i)^{(ni - niyi)}$, our goal is to solve β for the predictor variables, therefore, using the canonical link (logit) function, we calculate the score using the equation $\partial\ell/\partial\beta_j = \Sigma_jn_i(y_i- \pi_i)x_{ij}$, which is unfortunately not linear, so we need a iterative method to find β_j; Newton Raphson iteration. In our model, when the β values for the predictor variables is greater than zero ($\beta_j>0$), it indicates a positive association between those predictors and the response variable and vice versa for β values less than zero ($\beta_j < 0$) (see Appendix J for final model mathematical notation and parameter estimate summary). Observing the standard error in a model summary, if very large values are present, and the maximum log-likelihood value is extremely small we can assume complete or quasi-separation. In this study no complete or quasi-separation was found in the model.

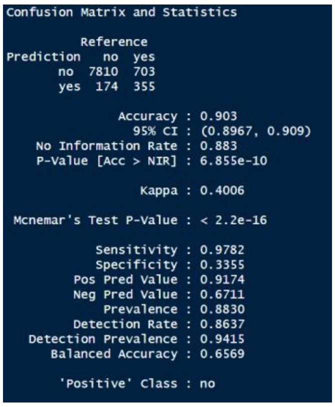
V. Results and Discussion

In the initial EDA we observed the response variable as binomial categorical, with 39922 customers (88.30%) not subscribing to a term deposit and 5289 customers (11.79%) who did (see Appendix K for raw code and output). Based on the response variable characteristics, a binomial GLM (logistic regression) is used. Following this, a 'logit' link function was chosen as a result of conducting logistic regression with 'sub_term_dep' as the response and all other variables as predictors (see Appendix L for raw code). Our initial results showed using the 'probit' link function was the most effective; lowest AIC score (21610) followed closely by 'logit' link function (21648), then 'cauchit' (23180) and finally 'cloglog'(23634). For the sole purpose of performance ability, we would choose the 'probit' link function, but due to its model complexity, it is hard to describe to other people. Alternatively, the 'logit' link function is simpler, and we can describe the model to other people; if we increased an x value of a predictor variable, we could explain the quantification of the effect of its beta value to the response variable along the implementation of the odds and log odds in the discussion. With 'probit' link function it becomes difficult to interpret the meaning of the beta value and therefor describe. Therefore, the slight trade-off between AIC score to give a model that we can describe is preferred. Binomial GLM with the 'logit' link function is the chosen model.

Next, we proceeded to feature selection and choosing appropriate predictor variables for our model. This process involved running the binomial GLM model with 'sub_term_dep' as the response and all other variables as predictors and using the summary() function in R to evaluate each of the predictors and their influence on the response variable, specifically observing the p-value. (See appendix G for raw code and Summary). In the end its was found that 'age', 'pdays' and 'previous' p-values were greater than 0.05, which infers that their β values are not significantly different from zero and therefore do not have an influence on the response variable and were removed. In the final model's summary (see appendix M for raw code and output) with all the significant variables, we have an AIC value of 21643 and no large standard errors, therefore concluding no complete or quasi - separation. Additionally, a model with the 'previous' variable included was considered, with an AIC score of 21643 but, with only a difference of 1, the simpler model was chosen. In the final model summary (Appendix M) it shows that for each categorical variable, one of the levels, chosen by R, is set as the base level ($x = 0$, therefore $x = 0$ is β_0) and the remaining levels are broken down to identify their own influence on the response variable. Some of these levels do not have a significant influence on the response ($p\text{-value} > 0.05$) but are kept in the model because they are included in the predictor variable that is overall an important predictor in the model.

In our final model (see appendix J for mathematical notation of model), when using a binomial GLM with a 'logit' link function, we are taking the log of the odds using our linear component, $\log(\pi_i/1- \pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .. \beta_p x_p$. Therefore, the magnitude of β is associated with the log of odds. If x_p is a quantitative variable; a numerical measure, and is increased by one unit, the $\log(\pi_i/1- \pi_i)$ increases by β_p. Therefore, β_p is how much the log of odds will increase if x_p increases by one unit. A parsimonious model with a good balance between variance and bias is also desired, therefore we can use likelihood test ratio to compare models (complex versus simple (nested) model) to see which one is preferred. The results showed deviance value of -2.677 and a Pr (>Chi) of 0.5959, resulting in a preference for the simpler model (see Appendix N for raw data and output). Interactions were also conducted to investigate to see if a more effective model was produced; lower ACI scores, which was the case. After conducting the likelihood test ratio, the more complex model was preferred over the simple interactive models. Therefore, the simple model without interaction was chosen. Backward elimination was also used (see appendix H for raw code) to verify our model choice. As expected, the model produced was the model mentioned previously which included the 'previous' variable because of its low AIC score, but with an AIC difference of one, the simpler model is preferred and is continued to be used.

Next, assumptions for binary GLM were investigated. The first assumption is that the response variable is binomial, which we verified visually (see appendix K) in our EDA. Next, we check for multicollinearity, using the variance inflation factor (vif()) in R (see Appendix I for raw code and output). The results showed no multicollinearity between the predictor variables: the results ranged from 1.029 and 1.19 which is a low value that correlates to no multicollinearity. Checking for overdispersion, necessary for most GLM, but for logistic regression with ungrouped data "there is no overdispersion" because "overdispersion is not possible if n_i=1. If y_i only takes values 0 and 1, then it must be distributed as Bernoulli(π), and its variance must be π_i (1− π_i)" (7.3 - Overdispersion, n.d.)[7] and is therefore not checked. Also, in the GLM framework, leverage is also investigated, but is not important for a binomial regression because y only takes the value of 0 and 1, so therefore the impact of x on y is not great. The final part of the study was to investigate the predictive power of our model using a classification table and ROC Curve. The data is first split into testing data (80% of the dataset used) and training data (20% of the dataset used), then, we use the training data to train our model and use the testing data to test our models' predictive capabilities.



```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  7810  703
       yes  174  355

               Accuracy : 0.903
                 95% CI : (0.8967, 0.909)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 6.855e-10

                  Kappa : 0.4006

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9782
            Specificity : 0.3355
         Pos Pred Value : 0.9174
         Neg Pred Value : 0.6711
             Prevalence : 0.8830
         Detection Rate : 0.8637
   Detection Prevalence : 0.9415
      Balanced Accuracy : 0.6569

       'Positive' Class : no
```

Figure 1: Classification table and predictive accuracy.

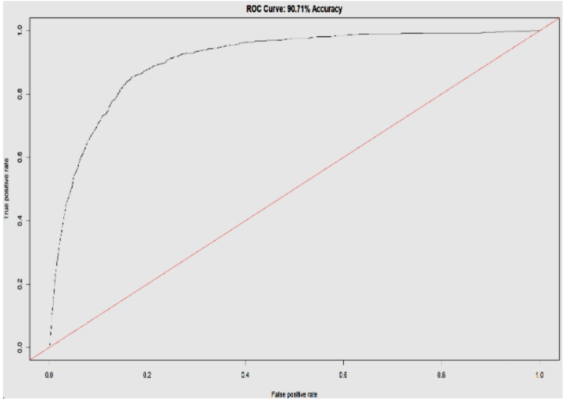


Figure 2: ROC Curve model performance.

Figure 1 displays the results of testing our mode to predict CTDS. The results provided are for when event = 0 (no CTDS), defaulted by R. The classification table results in a sensitivity value (True Negative in this case), of 97.82%; 7810 true negatives and 174 false positive and a specificity (True Positive in this case) of 33.55%; 355 true positives and 703 false negatives. This gives an overall accuracy of 90.30%, However, this doesn't consider the overall cost of those incorrect predictions. Using simple accuracy metrics, problems can give an inflated sense of confidence in model predictions that can be detrimental to some objectives. The ROC curve measures the ability of a logistic regression model to identify between positive and negative classes and evaluates the performance of the model across all possible thresholds. Overall, the ROC curve is a more comprehensive measure of the model's overall performance. In figure 2, the red diagonal line represents 50% accuracy and the closer to value of 1, the better the classification performance. Our ROC curve is relatively close to 1, specifically, .9071 (90.71 %) which equivalates to a high predictive performance (see Appendix O for raw code for testing predictive power). As a final remark, based on the EDA of the dataset, we can conclude very little success in the marketing campaign as only 11.79% of customers subscribed. From these results, there needs to either be a restructuring of the product to make it more desired such as increased interest rates or lower minimum deposit amount, try an alternative marketing campaign such as video or influencer marketing, or transition to other existing products such as personal or bank loans to which the outcomes are more successful (see Appendix P for raw code and visualisation).

VI. Conclusion

Technology provides easy access to products, services, entertainment, and information through the internet. This abundance of opportunities has made companies, including banks and global industries, more competitive in acquiring and retaining customers. Direct marketing is one strategy used to retain customer loyalty and satisfaction, while also attracting new customers and achieving overall success. This report investigated the "bank-full.csv" dataset which comprised of data from 17 direct marketing campaigns deployed by a Portuguese bank with the goal of having clients subscribe to term deposits. In this report an EDA was conducted to determine an effective GLM that could be used to predict CTDS, based on the characteristics of the response variable. The response variable was a binary categorical variable, therefore a binomial GLM was selected. Based on multiple binomial GLM test using different link functions, refinements of the model's predictor variable based on their influences on the response variable, and training and testing the model, we were able to create an effective model to predict CTDS. The final binomial GLM comprised of a logit link function with 12 predictor variable and had a predictive power of 90.30% (classification table). The ROC Curve produced a predictive power of 90.71%. After performing EDA of the dataset, it was discovered that the marketing campaign aimed at increasing term deposit subscriptions was largely ineffective, achieving only an 11.79% success rate. The lack of desired features like interest rates or minimum deposit amounts, negative attitudes toward direct marketing, or both, may have contributed to this result. To improve success rates, options include restructuring the product, exploring alternative marketing campaigns like video or influencer marketing, or promoting other existing products like personal or bank loans. Using only this data, our findings are limited to a specific demographic. However, this report can serve as a starting point for further investigation into the association between direct marketing success and banks' ability to acquire customer term deposit subscriptions. Improvements to the model could be made by incorporating more observations and significant variables to benefit the wider banking industry.

VII. References

1. Moro, S., Laureano, R., & Cortez, P. (n.d.). *Using data mining for bank direct marketing: an application of the CRISP-DM methodology*. Retrieved from Core.ac.uk. https://core.ac.uk/display/55616194?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1 on February 26th, 2023.

2. Page, C., & Luding, Y. (2003). *Bank managers' direct marketing dilemmas – customers' attitudes and purchase intention*. International Journal of Bank Marketing, 21(3), 147–163. Retrieved from https://doi.org/10.1108/02652320310469520 on February 26th, 2023.

3. Neilson, L. C., & Chadha, M. (2008). *International marketing strategy in the retail banking industry: The case of ICICI Bank in Canada*. Journal of Financial Services Marketing, 13(3), 204–220. https://doi.org/10.1057/fsm.2008.21 on February 27th, 2023.

4. *Term deposits* – Moneysmart.gov.au (n.d.). Moneysmart.gov.au. Retrieved from https://moneysmart.gov.au/saving/term-deposits on February 27th, 2023.

5. *UC Irvine Machine Learning Repository*. (2012). Uci.edu. Retrieved from https://archive-beta.ics.uci.edu/dataset/222/bank+marketing on February 27th, 2023.

6. Statistics, c=AU; o=Commonwealth of A. ou=Australian B. of. (1998, July 31). *Chapter - Sampling Methods - Random Sampling*. Www.abs.gov.au. Retrieved from https://www.abs.gov.au/Ausstats/abs@.nsf/2f762f95845417aeca25706c00834efa/A493A524D0C5D1A0CA2571FE007D69E2?opendocument on February 27th 2023.

7. 7.3 - Overdispersion. (n.d.). Online.stat.psu.edu. Retrieved from https://online.stat.psu.edu/stat504/book/export/html/779#:~:text=Note%2C%20there%20is%20no%20overdispersion%20for%20ungrouped%20data on February 28, 2023.

Appendix

Appendix A: raw code for exploratory dataset analysis of dataset in RStudio.

```
View(bank) #contact has the same value throughout the dataset and is therefore not needed.

str(bank) #10-character type; 6 categorical, 4 binary variables and 7 integers. Response variable 'y' is binary. Ungrouped data where response is 1 or 0 for each observation.

summary(bank)

dim(bank) #45211 observations and 17 attributes (including response variable)

#Missing values
count_miss_values = function(x) sum(is.na(x))
apply(bank, MARGIN = 2, FUN = count_miss_values) #0 missing values

#Duplicate values
sum(duplicated(bank)) #0 duplicate values.
```

Appendix B: Variable description.

Variable Name	Variable Type	Variable Levels	Description
age	Numerical	NA	Age of contacted clients

job	Categorical	admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services	Type of job
marital	Categorical	married, divorced, single	Marital Status. Note: divorced mean divorced or widowed
education	Categorical	unknown, secondary, primary, tertiary	Education Level
default	Binary	yes, no	Has a credit default?
balance	Numerical	NA	Average yearly salary.
housing	Binary	yes, no	Has a housing loan?
loan	Binary	yes, no	Has a personal loan?
contact	Categorical	unknown, telephone, cellular	Contact communication type.
day	Numerical	NA	Last contact day of the month.
month	Categorical	jan, feb, mar, …. nov, dec	Last contact month of the year.
duration	Numerical	NA	Last contact duration, in seconds.
campaign	Numerical	NA	Number of contacts performed during this campaign and for this client.
pdays	Numerical	NA	Number of days that passed by after the client was last contacted from a previous campaign (-1 means the client was not previously contacted)
previous	Numerical	NA	Number of contacts preformed before this campaign and for this client.
poutcome	Categorical	unknown, other, failure, success	Outcome of the previous marketing campaign
y	Binary	yes, no	Has the client subscribed a term deposit?

Appendix C: Original study reasoning for variable and observation elimination.

In the original study “USING DATA MINING FOR BANK DIRECT MARKETING: AN APPLICATION OF THE CRISP-DM METHODOLOGY” (Moro et al., 2011)1 there were 79354 contacts (observation) that was reduced to 45211after removing non-conclusive instances and examples with values missing. Originally there were 59 variables, which reduced to 30 (response variable included) after conducting variable influence on the response and removing non-influential variables. No addition reasoning was provided for further reducing the variable count to 17, which is the amount in the dataset used.

Appendix D: Mathematical breakdown of logit function with binomial GLM.

The canonical link function is $\mathrm{logit}(\theta) = \log(\theta/1- \theta)$, applying the logit function to our binomial GLM we have, $E(y)= \pi_i$ with logit link function becomes logit $(E(y) = \pi_i)$, which is equal to $\mathrm{logit}(\pi_i)$ which is the same as $\mathrm{logit}(\theta)$. Therefore, $\mathrm{logit}(\pi_i) = \log(\pi_i /1- \pi_i)$ which is equivalent to $g(\pi_i) = \log(\pi_i /1- \pi_i)$.

Appendix E: AIC mathematical breakdown.

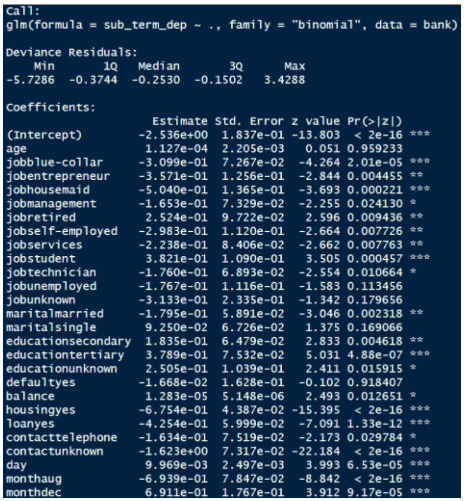
$AIC = -2\ell(\hat{\beta}_M) + 2_p$, where $_p$ is the number of parameters in model *M*. Essential this equation relies on the log likelihood $(-2\ell(\hat{\beta}_M))$ with a penalized term 2_p: the more complex the model (more predicts) the larger the penalty. With likelihood we want the highest values but because we have the negative value at the beginning, we are then looking for the lowest AIC value. There is also Bayesian Information Criteria (BIC): $-2\ell(\hat{\beta}_M) + {}_p \log {}_n$, which is the same concept as AIC but has a heavier penalty on model complexity.

Appendix F: Binomial GLM equation rearrangement to solve for π_i

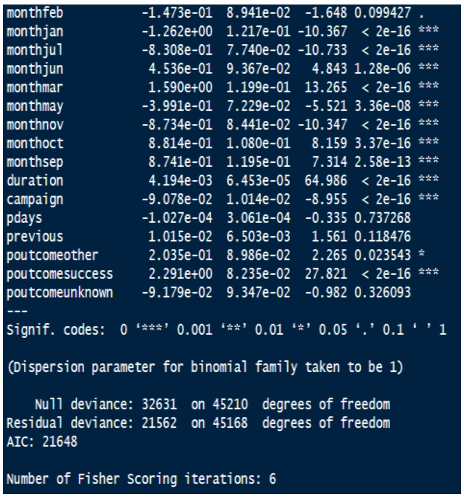
Here we have our random component: $y_i \sim \mathrm{Bin}(1, \pi_i)$, which using a logit link function with our systematic component, we have $\mathrm{logit}(\pi_i) = \log(\pi_i /1- \pi_i) = n_i$, where $n_i = (\Sigma_j x_{ij}\beta_j)$ is our linear model: $\beta_0+ \beta_1x_1+ \beta_2x_2\ldots \beta_nx_n$. If we rearrange this equation we can see that π_i, our expected value of 'y' which is the probability of the desired outcome which is the expect value of the linear model divided by one plus the expected value of our linear model: $\pi_i = \exp(\Sigma_j x_{ij}\beta_j) / \exp(1+ \Sigma_j x_{ij}\beta_j)$

Appendix G: Raw code and Summary of binomial GLM with all predictor variables.

#Logit-link function.
m1 = glm(sub_term_dep ~ ., data = bank, family = 'binomial')
summary(m1) *#AIC: 21648****



```
Call:
glm(formula = sub_term_dep ~ ., family = "binomial", data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7286  -0.3744  -0.2530  -0.1502   3.4288

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.536e+00  1.837e-01 -13.803  < 2e-16 ***
age                 1.127e-04  2.205e-03   0.051 0.959233
jobblue-collar     -3.099e-01  7.267e-02  -4.264 2.01e-05 ***
jobentrepreneur    -3.571e-01  1.256e-01  -2.844 0.004455 **
jobhousemaid       -5.040e-01  1.365e-01  -3.693 0.000221 ***
jobmanagement      -1.653e-01  7.329e-02  -2.255 0.024130 *
jobretired          2.524e-01  9.722e-02   2.596 0.009436 **
jobself-employed   -2.983e-01  1.120e-01  -2.664 0.007726 **
jobservices        -2.238e-01  8.406e-02  -2.662 0.007763 **
jobstudent          3.821e-01  1.090e-01   3.505 0.000457 ***
jobtechnician      -1.760e-01  6.893e-02  -2.554 0.010664 *
jobunemployed      -1.767e-01  1.116e-01  -1.583 0.113456
jobunknown         -3.133e-01  2.335e-01  -1.342 0.179656
maritalmarried     -1.795e-01  5.891e-02  -3.046 0.002318 **
maritalsingle       9.250e-02  6.726e-02   1.375 0.169066
educationsecondary  1.835e-01  6.479e-02   2.833 0.004618 **
educationtertiary   3.789e-01  7.532e-02   5.031 4.88e-07 ***
educationunknown    2.505e-01  1.039e-01   2.411 0.015915 *
defaultyes         -1.668e-02  1.628e-01  -0.102 0.918407
balance             1.283e-05  5.148e-06   2.493 0.012651 *
housingyes         -6.754e-01  4.387e-02 -15.395  < 2e-16 ***
loanyes            -4.254e-01  5.999e-02  -7.091 1.33e-12 ***
contacttelephone   -1.634e-01  7.519e-02  -2.173 0.029784 *
contactunknown     -1.623e+00  7.317e-02 -22.184  < 2e-16 ***
day                 9.969e-03  2.497e-03   3.993 6.53e-05 ***
monthaug           -6.939e-01  7.847e-02  -8.842  < 2e-16 ***
monthdec            6.911e-01  1.767e-01   3.912 9.17e-05 ***
```



```
monthfeb           -1.473e-01  8.941e-02  -1.648 0.099427 .
monthjan           -1.262e+00  1.217e-01 -10.367  < 2e-16 ***
monthjul           -8.308e-01  7.740e-02 -10.733  < 2e-16 ***
monthjun            4.536e-01  9.367e-02   4.843 1.28e-06 ***
monthmar            1.590e+00  1.199e-01  13.265  < 2e-16 ***
monthmay           -3.991e-01  7.229e-02  -5.521 3.36e-08 ***
monthnov           -8.734e-01  8.441e-02 -10.347  < 2e-16 ***
monthoct            8.814e-01  1.080e-01   8.159 3.37e-16 ***
monthsep            8.741e-01  1.195e-01   7.314 2.58e-13 ***
duration            4.194e-03  6.453e-05  64.986  < 2e-16 ***
campaign           -9.078e-02  1.014e-02  -8.955  < 2e-16 ***
pdays              -1.027e-04  3.061e-04  -0.335 0.737268
previous            1.015e-02  6.503e-03   1.561 0.118476
poutcomeother       2.035e-01  8.986e-02   2.265 0.023543 *
poutcomesuccess     2.291e+00  8.235e-02  27.821  < 2e-16 ***
poutcomeunknown    -9.179e-02  9.347e-02  -0.982 0.326093
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 21562  on 45168  degrees of freedom
AIC: 21648

Number of Fisher Scoring iterations: 6
```

Appendix H: Raw code for likelihood ratio test and Backward elimination (stepwise method) in R and with discussion.

(m1m1c_anova = anova(m1,m1c, test = "Chisq"))

#The likelihood test ratio shows that there is no significant difference between the two models, deviance: -2.7767. Due the models' residuals deviance value being the same (21,562) and the Deviance value of -2.7767 and p-value greater than 0.05 (0.5959), therefore the simpler model is preferred over the complex model. Therefore, we will continue with the simpler model.

#Backward elimination method for model selection.
M6 = step(m1, direction = 'backward')
summary(m6) *#sub_term_dep ~ job + marital + education + balance + housing + loan + contact + day + month + duration + campaign + previous + poutcome*
#AIC: 21642

#Forward selection and backward elimination method for model selection.
M7 = step(m1, direction = 'both')
summary(m7) *#sub_term_dep ~ job + marital + education + balance + housing + loan + contact + day + month + duration + campaign + previous + poutcome*
#AIC: 21642

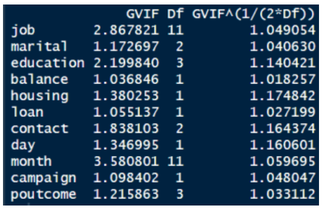
#observing the standard errors. There are no large values, therefore we can conclude that there is no complete or quasi separation.

The process for using the backward elimination involves a model that has a response variable and all other variables as predictors. When the model runs through its first iterations its will eliminate the most insignificant (largest p-value) variable and iterate through again without that variable included. One by one non influential variables are removed through each iteration until only significant influential variables are left.

Appendix I: Raw code for variance inflation factor and output

#Variance inflation factor (VIF) model m1c.
var_infl= vif(m1c) *#VIF focuses in turn on each predictor in the model, combining the main effect for that predictor with the main effects of the predictors with which the focal predictor interacts and the interactions; e.g., in the model with formula y ~ a*b + b*c*

var_infl *#A general guideline is that a VIF larger than 5 or 10 is large, indicating that the model has problems estimating the coefficient. The VIF values are between 1.029 and 1.19 which is a low value that correlates to no multicollinearity, which satisfied the multicollinearity assumptions of.*



```
              GVIF Df GVIF^(1/(2*Df))
job       2.867821 11        1.049054
marital   1.172697  2        1.040630
education 2.199840  3        1.140421
balance   1.036846  1        1.018257
housing   1.380253  1        1.174842
loan      1.055137  1        1.027199
contact   1.838103  2        1.164374
day       1.346995  1        1.160601
month     3.580801 11        1.059695
campaign  1.098402  1        1.048047
poutcome  1.215863  3        1.033112
```

The ‘GVIF’ on the right is the initial variance inflation factor score, the equation on the left calculates the variance inflation score using the number of levels in the variable. We refer to the right column for our final VIF score.

Appendix J: Final model mathematical notation and parameter estimate summary.

$$\text{logit}(\pi_i) = -2.527 \left\{\begin{array}{ll} +0 & (\text{job admin}) \\ -3.116\text{e-}01 & (\text{job blue-collar}) \\ -3.586\text{e-}01 & (\text{job entrepreneur}) \\ -5.043\text{e-}01 & (\text{job house maid}) \\ -1.646\text{e-}01 & (\text{job management}) \\ 2.541\text{e-}01 & (\text{job retired}) \\ -2.983\text{e-}01 & (\text{job self-employed}) \\ -2.252\text{e-}01 & (\text{job services}) \\ 3.822\text{e-}01 & (\text{job student}) \\ -1.760\text{e-}01 & (\text{job technician}) \\ +0 & (\text{job unemployed}) \\ +0 & (\text{job unknown}) \end{array}\right. \left\{\begin{array}{l} +0(\text{marital divorced}) \\ -1.785\text{e-}01(\text{marital married}) \\ +0\ (\text{martial single} \end{array}\right. \left\{\begin{array}{l} +0\ (\text{education primary}) \\ +1.822\text{e-}01\ (\text{education secondary}) \\ +3.783\text{e-}01\ (\text{education tertiary}) \\ 2.493\text{e-}01(\text{education unknown}) \end{array}\right. +1.289\text{e-}05\ (\text{balance}) \left\{\begin{array}{l} +0\ (\text{housing no}) \\ -6.755\text{e-}01\ (\text{housing yes}) \end{array}\right. \left\{\begin{array}{l} +0\ (\text{loan no}) \\ -4.254\text{e-}01\ (\text{loan yes}) \end{array}\right. \left\{\begin{array}{l} +0\ (\text{contact cellular})\ -1.614\text{e-}01 \\ (\text{contact telephone}) \\ -1.622\text{e+}00\ (\text{contact unknown}) \end{array}\right. +9.898\text{e-}03\ (\text{day}) \left\{\begin{array}{l} +0\ (\text{month apr}) \\ -6.914\text{e-}01\ (\text{month aug}) \\ +6.943\text{e-}01\ (\text{month dec}) \\ +0\ (\text{month feb}) \\ -1.258\text{e+}00\ (\text{month jan}) \\ -8.288\text{e-}01\ (\text{monthjul}) \\ +4.558\text{e-}01\ (\text{month jun}) \\ +1.592\text{e+}00\ (\text{month mar}) \\ -3.992\text{e-}01(\text{month may}) \\ -8.693\text{e-}01(\text{month nov}) \\ +8.858\text{e-}01\ (\text{month oct}) \\ +8.763\text{e-}01(\text{month sep}) \end{array}\right. +4.194\text{e-}03\ (\text{duration})\ -9.037\text{e-}02\ (\text{campaign}) \left\{\begin{array}{l} +0\ (\text{poutcome failure}) \\ +2.163\text{e-}01\ (\text{poutcome other}) \\ +2.300\text{e+}00\ (\text{poutcome success}) \\ +0\ (\text{poutcome unknown}) \end{array}\right.$$

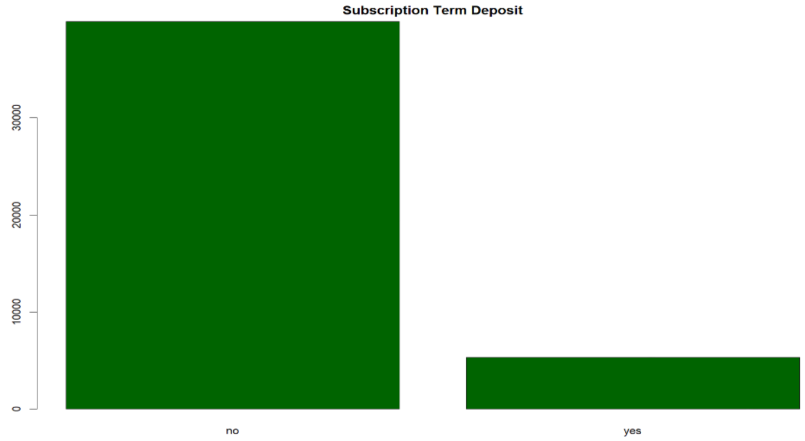
Predictor Variable (x_n)	Event # (1 = Subscription, 0 = no subscription)	Estimate Value (β_n)	Predictors Impact
Intercept	0	-2.527e+00	NA
Job: admin	0	0	No Impact
Job: blue collar	0	-3.116e-01	Negative Impact
Job: house maid	0	-5.043e-01	Negative Impact
Job: management	0	-1.646e-01	Negative Impact
Job: retired	0	2.541e-01	Positive Impact
Job: self employed	0	-2.983e-01	Negative Impact
Job: services	0	-2.252e-01	Negative Impact
Job: student	0	3.822e-01	Positive Impact
Job: technician	0	-1.760e-01	Negative Impact
Job: employed	0	0	No Impact
Job: unknown	0	0	No Impact
Marital: divorced	0	0	No Impact
Marital: married	0	-1.785e-01	Negative Impact
Marital: single	0	0	No Impact
Education: primary	0	0	No Impact
Education: secondary	0	1.822e-01	Positive Impact
Education: tertiary	0	3.783e-01	Positive Impact
Education: unknown	0	2.493e-01	Positive Impact
Balance	0	1.289e-05	Positive Impact

Housing: no	0	0	No Impact
Housing: yes	0	-6.755e-01	Negative Impact
Loan: no	0	0	No Impact
Loan: yes	0	-4.254e-01	Negative Impact
Contact: cellular	0	0	No Impact
Contact: telephone	0	-1.614e-01	Negative Impact
Contact: unknown	0	-1.622e+00	Negative Impact
Day	0	9.898e-03	Positive Impact
Month: apr	0	0	No Impact
Month: aug	0	-6.914e-01	Negative Impact
Month: dec	0	6.943e-01	Positive Impact
Month: feb	0	0	No Impact
Month: jan	0	-1.258e+00	Negative Impact
Month: jul	0	-8.288e-01	Negative Impact
Month: jun	0	4.558e-01	Positive Impact
Month: mar	0	1.592e+00	Positive Impact
Month: may	0	-3.992e-01	Negative Impact
Month: nov	0	-8.693e-01	Negative Impact
Month: oct	0	8.858e-01	Positive Impact
Month: sep	0	8.763e-01	Positive Impact
Duration	0	4.194e-03	Positive Impact
Campaign	0	-9.037e-02	Negative Impact
Poutcome: failure	0	0	No Impact
Poutcome: other	0	2.163e-01	Positive Impact
Poutcome: success	0	2.300e+00	Positive Impact
Poutcome: unkown	0	0	No Impact

Appendix K: raw code and visualisation for response variable ‘sub_term_dep’

```
#Change response variable name for clarification
names(bank)[names(bank) == 'y'] <- 'sub_term_dep'

#convert categorical and binary variables to factors
bank$sub_term_dep = as.factor(bank$sub_term_dep)
```

#Visualisation
plot(bank$sub_term_dep, col = 'darkgreen', main = 'Subscription Term Deposit')



Appendix L: raw code for link function model comparison

#Logit-link function.
m1 = glm(sub_term_dep ~ ., data = bank, family = 'binomial')
summary(m1) *#AIC: 21648****

#Probit-link function
m2 = glm(sub_term_dep ~ ., data = bank, family = binomial('probit'))
summary(m2) *#AIC: 21610*****

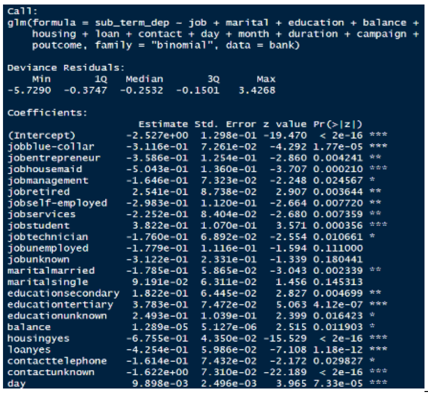
#Cauchit-link function
m3 = glm(sub_term_dep ~ ., data = bank, family = binomial('cauchit'))
summary(m3) *#AIC: 23180***

#cloglog-link function
m4 = glm(sub_term_dep ~ ., data = bank, family = binomial('cloglog'))
summary(m4) *#AIC: 23634**

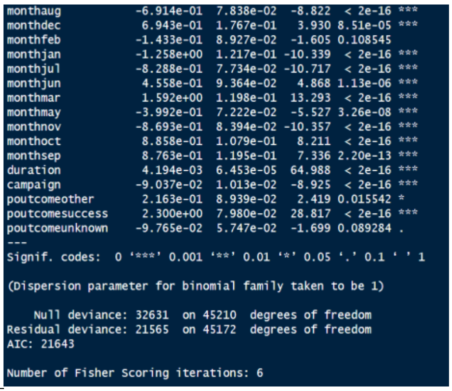
Appendix M: Raw code and output for the final chosen logistic regression mode (binomial GLM).

m1c = glm(sub_term_dep ~ job + marital + education + balance + housing + loan + contact + day + month + duration + campaign + poutcome, data = bank, family = 'binomial')

summary(m1c)



```
Call:
glm(formula = sub_term_dep ~ job + marital + education + balance +
    housing + loan + contact + day + month + duration + campaign +
    poutcome, family = "binomial", data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7290  -0.3747  -0.2532  -0.1501   3.4268

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.527e+00  1.298e-01 -19.470  < 2e-16 ***
jobblue-collar    -3.116e-01  7.261e-02  -4.292 1.77e-05 ***
jobentrepreneur   -3.586e-01  1.254e-01  -2.860 0.004241 **
jobhousemaid      -5.043e-01  1.360e-01  -3.707 0.000210 ***
jobmanagement     -1.646e-01  7.323e-02  -2.248 0.024567 *
jobretired         2.541e-01  8.738e-02   2.907 0.003644 **
jobself-employed  -2.983e-01  1.120e-01  -2.664 0.007720 **
jobservices       -2.252e-01  8.404e-02  -2.680 0.007359 **
jobstudent         3.822e-01  1.070e-01   3.571 0.000356 ***
jobtechnician     -1.760e-01  6.892e-02  -2.554 0.010661 *
jobunemployed     -1.779e-01  1.116e-01  -1.594 0.111000
jobunknown        -3.122e-01  2.331e-01  -1.339 0.180441
maritalmarried    -1.785e-01  5.865e-02  -3.043 0.002339 **
maritalsingle      9.191e-02  6.311e-02   1.456 0.145313
educationsecondary 1.822e-01  6.445e-02   2.827 0.004699 **
educationtertiary  3.783e-01  7.472e-02   5.063 4.12e-07 ***
educationunknown   2.493e-01  1.039e-01   2.399 0.016423 *
balance            1.289e-05  5.127e-06   2.515 0.011903 *
housingyes        -6.755e-01  4.350e-02 -15.529  < 2e-16 ***
loanyes           -4.254e-01  5.986e-02  -7.108 1.18e-12 ***
contacttelephone  -1.614e-01  7.432e-02  -2.172 0.029827 *
contactunknown    -1.622e+00  7.310e-02 -22.189  < 2e-16 ***
day                9.898e-03  2.496e-03   3.965 7.33e-05 ***
```



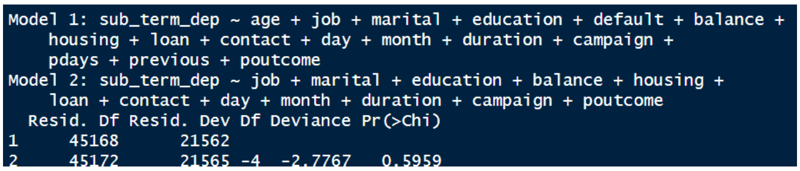
```
monthaug          -6.914e-01  7.838e-02  -8.822  < 2e-16 ***
monthdec           6.943e-01  1.767e-01   3.930 8.51e-05 ***
monthfeb          -1.433e-01  8.927e-02  -1.605 0.108545
monthjan          -1.258e+00  1.217e-01 -10.339  < 2e-16 ***
monthjul          -8.288e-01  7.734e-02 -10.717  < 2e-16 ***
monthjun           4.558e-01  9.364e-02   4.868 1.13e-06 ***
monthmar           1.592e+00  1.198e-01  13.293  < 2e-16 ***
monthmay          -3.992e-01  7.222e-02  -5.527 3.26e-08 ***
monthnov          -8.693e-01  8.394e-02 -10.357  < 2e-16 ***
monthoct           8.858e-01  1.079e-01   8.211  < 2e-16 ***
monthsep           8.763e-01  1.195e-01   7.336 2.20e-13 ***
duration           4.194e-03  6.453e-05  64.988  < 2e-16 ***
campaign          -9.037e-02  1.013e-02  -8.925  < 2e-16 ***
poutcomeother      2.163e-01  8.939e-02   2.419 0.015542 *
poutcomesuccess    2.300e+00  7.980e-02  28.817  < 2e-16 ***
poutcomeunknown   -9.765e-02  5.747e-02  -1.699 0.089284 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 21565  on 45172  degrees of freedom
AIC: 21643

Number of Fisher Scoring iterations: 6
```

Appendix N: raw code for likelihood test ration and output.

#likelihood test ratio for complex model vs chosen model.
(m1m1c_anova = anova(m1,m1c, test = "Chisq"))

#The likelihood test ratio shows that there is no significant difference between the two models, deviance: -2.7767. Due the models' residuals deviance value being the same (21,562) and the Deviance value of -

2.7767 and p-value greater than 0.05 (0.5959), therefore the simpler model is preferred over the complex model. Therefore, we will continue with the simpler model.



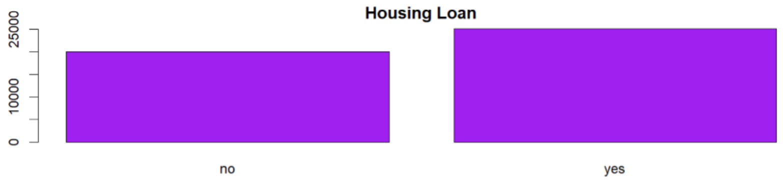
```
Model 1: sub_term_dep ~ age + job + marital + education + default + balance +
    housing + loan + contact + day + month + duration + campaign +
    pdays + previous + poutcome
Model 2: sub_term_dep ~ job + marital + education + balance + housing +
    loan + contact + day + month + duration + campaign + poutcome
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     45168      21562
2     45172      21565 -4  -2.7767   0.5959
```

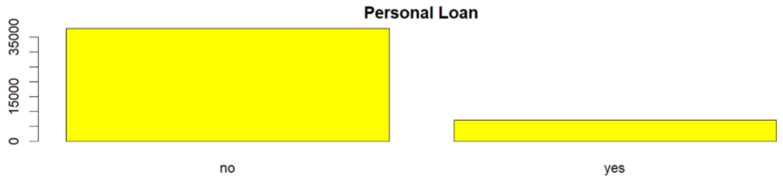
Appendix O: raw code for testing predictive power.

```
# We want to predict probability of subscription, therefore 'no' is the reference level
#Create copy of subset data
bank1 = bank

#Split data
set.seed(123)
bank_sample = sample.split(bank1$sub_term_dep, SplitRatio = 0.80) #sample split.
bank_train = subset(bank1, bank_sample == T) #Training data set.
bank_test = subset(bank1, bank_sample == F) #Test data set.
dim(bank_train)[1] #36169 Observations.
dim(bank_test)[1] #9042 Observations.

#Logistic regression
m8 = glm(sub_term_dep ~ job + marital + education + balance + housing + loan + contact + day+ month + duration + campaign + poutcome, data = bank_train, family = 'binomial')
summary(m8) #AIC: 17345

#Predict probability
(bank_probs = predict(m8, newdata = bank_test, type = 'response'))

# Classification accuracy - test data
bank_pred = as.factor(ifelse(bank_probs > 0.5, 'yes','no'))
(bank_CM = table(bank_pred, bank_test$sub_term_dep))
(bank_acc = sum(diag(bank_CM))/ sum(bank_CM)) #90.30% Accuracy

#More Info
confusionMatrix(bank_pred, bank_test$sub_term_dep)

# AUC — the area under the ROC curve accuracy; measure of accuracy where it graphs the true positive rate and the false positive rate.
bank_pred_probs = predict(m8, newdata = bank_test[,-17], type = 'response')
bank_pr = prediction(bank_pred_probs, bank_test$sub_term_dep)
bank_pfr = performance(bank_pr, "tpr", "fpr")
par(mfrow=c(1,1), mar=c(4,4,2,1))
plot(bank_pfr, main="ROC Curve: 90.71% Accuracy")
abline(a=0, b=1, col='red')
(bank_auc_lgr = performance(bank_pr, "auc")@y.values)
```

Appendix P: raw code and visualisation of ‘house’ and ‘loan’.

```
#Convert to factor.
bank$housing = as.factor(bank$housing)
bank$loan = as.factor(bank$loan)

#Visualisation
plot(bank$housing, col = 'purple', main = 'Housing Loan')
plot(bank$loan, col = 'yellow', main = 'Personal Loan')
```


Housing Loan
25000
10000
0
no
yes


Personal Loan
35000
15000
0
no
yes