**Bank Marketing Case Study**

1. **Executive Summary**

In this case study, we were tasked to find a model that best identified whether a customer would subscribe to a long-term deposit offered by a Portuguese bank during a call campaign. For our analysis, we reviewed variations of the logistic model and the linear discriminant analysis (LDA) model. In summary, a simple logistic regression model proved to be the best overall model with high accuracy and sensitivity. The LDA model did not prove to have comparable accuracy because the underlying predictor data did not have a normal distribution. Finally, a review of model coefficients and p-values indicate that to increase the likelihood of having a customer accept the bank product, the bank should consider the month in which the campaign call is conducted and the duration of the customer call.

1. **The Problem**

A Portuguese banking institution has provided direct marketing campaign data. The data includes various categorical and numerical customer data such as customer age, job, marital status, education, in addition to marketing campaign data such as how the customer was contacted for the campaign, the duration of contact, number of contacts, and outcome of the campaign. The purpose of the case study is to create a model that will accurately identify if a customer will subscribe to a term deposit. The problem is significant because customers who subscribe to long-term deposits increase assets available to the bank which are then used to create revenue earning debt instruments. This is because banks can obtain assets from customers at a lower cost of capital (interest paid to customers) than the interest earned on customer loans.

In this study, we will use a logistic and Linear Discriminant Analysis (LDA) model to determine which model can best predict whether a customer will subscribe to a long-term deposit. We will clean data, test assumptions, and refine models to achieve the best results.

1. **Review of Related Literature**

We reviewed various articles to obtain a better understanding of models that would create accurate binary predictions. Per ‘*Data Mining’* by John A. Bunge, Dean H. Judson (2005) in Encyclopedia of Social Measurement, the logistic regression can be used for classification where the response, or class, is a binary random variable. One advantage is that the binary logistic regression can provide a probability of success rather than just a classification. This enables adjusting the cut-off rather than setting arbitrary numbers such as 0.5.

In addition, as mentioned in ‘*Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study*’ by Maja Pohar, Mateja Blas, and Sandra Turk, the LDA model, in comparison to the logistic regression, assumes features are normally distributed. In addition, given the normal distribution requirement is satisfied, the LDA becomes able to produce lower error rates, but may sacrifice sensitivity.

1. **Methodology**

The data provided consisted of 20 predictors and one binary response variable. The predictor variables were grouped into three general categories: customer descriptors, marketing campaign attributes, and social and economic context predictors. At first glance, the predictors consisted of 7 nominal, or categorical variables, 3 binary variables, and 10 continuous variables.

The data set included 4,119 records. The target variable for these records were predominantly “No” responses. To fix the imbalance between the “No” and “Yes” responses in the target variable, we first removed the records with missing values, then randomly sampled 370 observations with “No” responses and appended the observations to the “Yes” responses to create a new balanced data set with 740 total observations from which we trained and tested models. Then, the new data set was split 80/20 in a random manner for training and testing purposes. The resulting train and test sets had 592 and 144 observations, respectively.

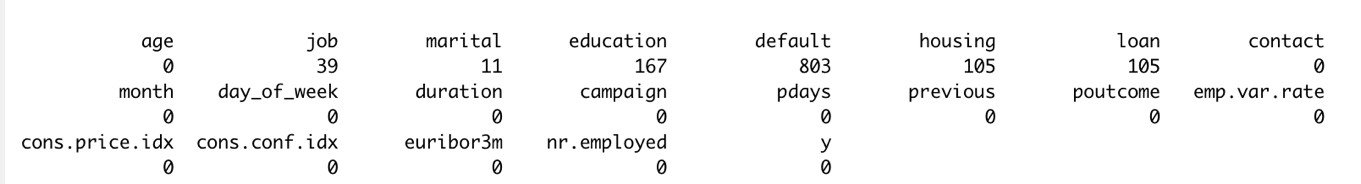
For this study we used both a logistic and a Linear Discriminant Analysis (“LDA”) model. The logistic model although simplistic, is a powerful baseline model on which to test other more complex models. The logistic model assumes that the underlying data has 1) a binary response variable, 2) no multicollinearity between variables, 3) no extreme outliers in continuous predictors, 4) independent observations, and 5) the sample size is sufficiently large. These assumptions were explored and addressed in the Data section of this report. One benefit of the simplicity that comes with using the logistic models is that the relationships are easy to understand. In an effort to refine the results of the logistic model, we compared results using a complex (full) predictor model, a simple model, a simple model with an adjusted cut-off and a simple model with pre-processed predictors. Details of the refinement can be seen in the following sections of this report.

To compare performance of the baseline logistic model, we selected the LDA model. The LDA model assumes a normal distribution and will produce more accurate results when predictors are normally distributed. As seen on section ***V. Data***, numeric predictors were not normally distributed. As a result, we do not expect the LDA model to have more accurate results compared to the logistic regression model.

1. **Data**
   1. **Missing Values**

We began our analysis by searching the data set for missing records. We found that 1,029 of 4,119 records (25% of total) had values designated as ‘unknown’, which the case information indicated as designations for missing values. Of the records with missing values, we found that the *default*, *education*, *housing*, and *loan* predictors had the most missing values in the amount of 803, 167, 105, and 105, respectively. See Figure 1 for results. These missing values were converted to ‘NAs’ and the corresponding records were dropped from the data set. The remaining records with non-missing values totaled 3,090.

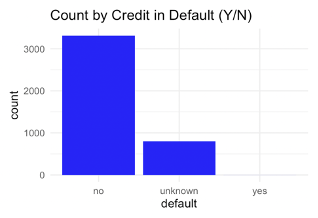
**Figure 1: Sum of Missing Values by Predictors**



* 1. **Predictor Cleaning and Preprocessing**

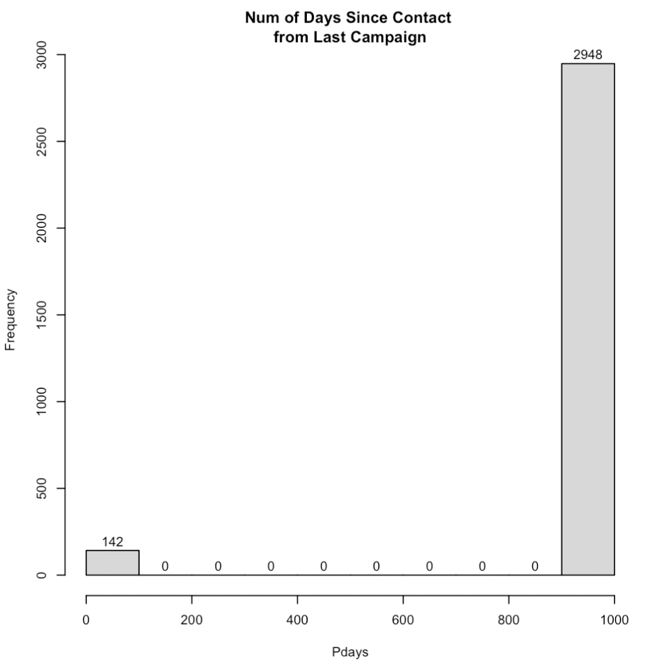
Next, we explored the predictor distributions by plotting histograms of numerical data and count frequencies of categorical data. See ***Appendix A: Plots of Predictor Distributions*** in section IX of this report for all predictor bar graphs. During our review, we found that the *default* predictor was binary, and all non-missing values had the same “No” response, as seen on Figure 2 below. Consequently, this predictor will be eliminated from our model as it does not appear to add any value to the model because all non-missing values are exactly the same.

**Figure 2: Count of Responses for ‘Has credit in default?’**



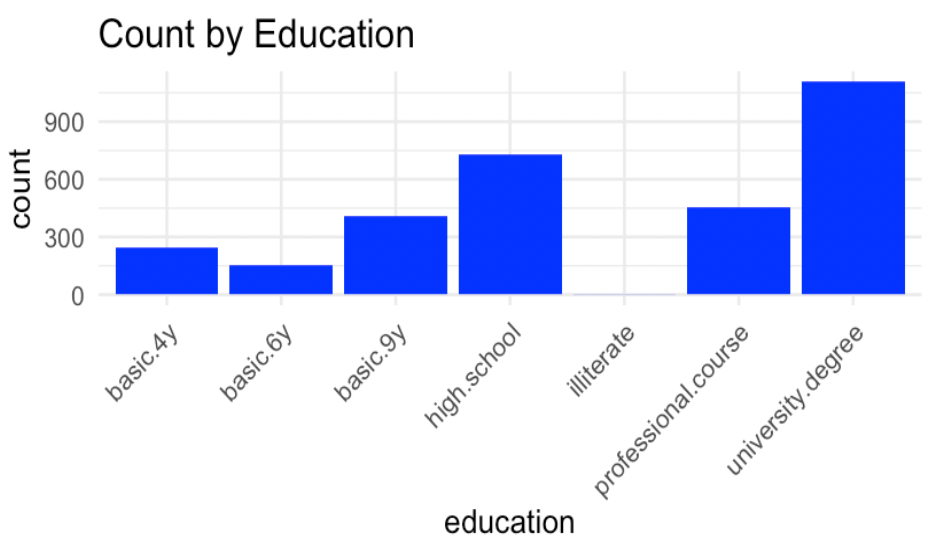
We also noted that the numerical *pdays* predictor (which indicates the number of days that passed after the client was last contacted from a previous campaign) had values that can be summarized into two categories. As seen on Figure 3, customers were either contacted in a previous campaign (as noted by values other than 999) or not contacted in a previous campaign (categorized as 999). We reformatted this predictor as a binary predictor where 0 indicated that the customer was not contacted in a previous campaign and 1 indicated that the customer was contacted in a previous campaign.

**Figure 3: Histogram for ‘Number of days that passed by after the client was last contacted from a previous campaign’**



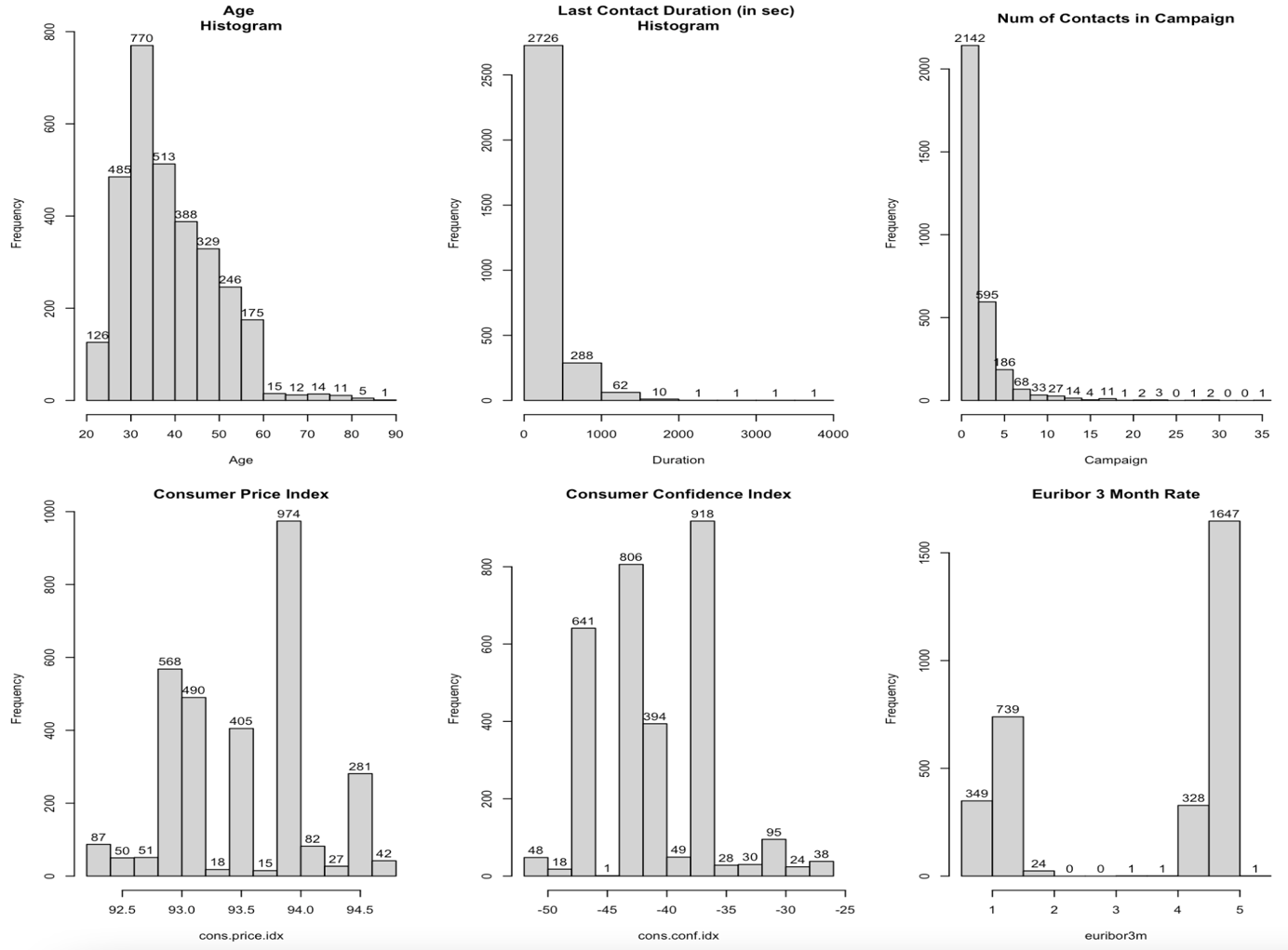
Moreover, the *education* predictor also had an opportunity for improvement. As seen in Figure 4 below, responses for the level of education obtained included seven different response categories of which four can be summarized into one single category. For this reason, we simplified the values in this predictor by grouping illiterate and all basic education beneath high school to a *‘< high school’* category. This reduced the number of responses from seven to four.

**Figure 4: Count of Response for ‘Education’ Predictor**

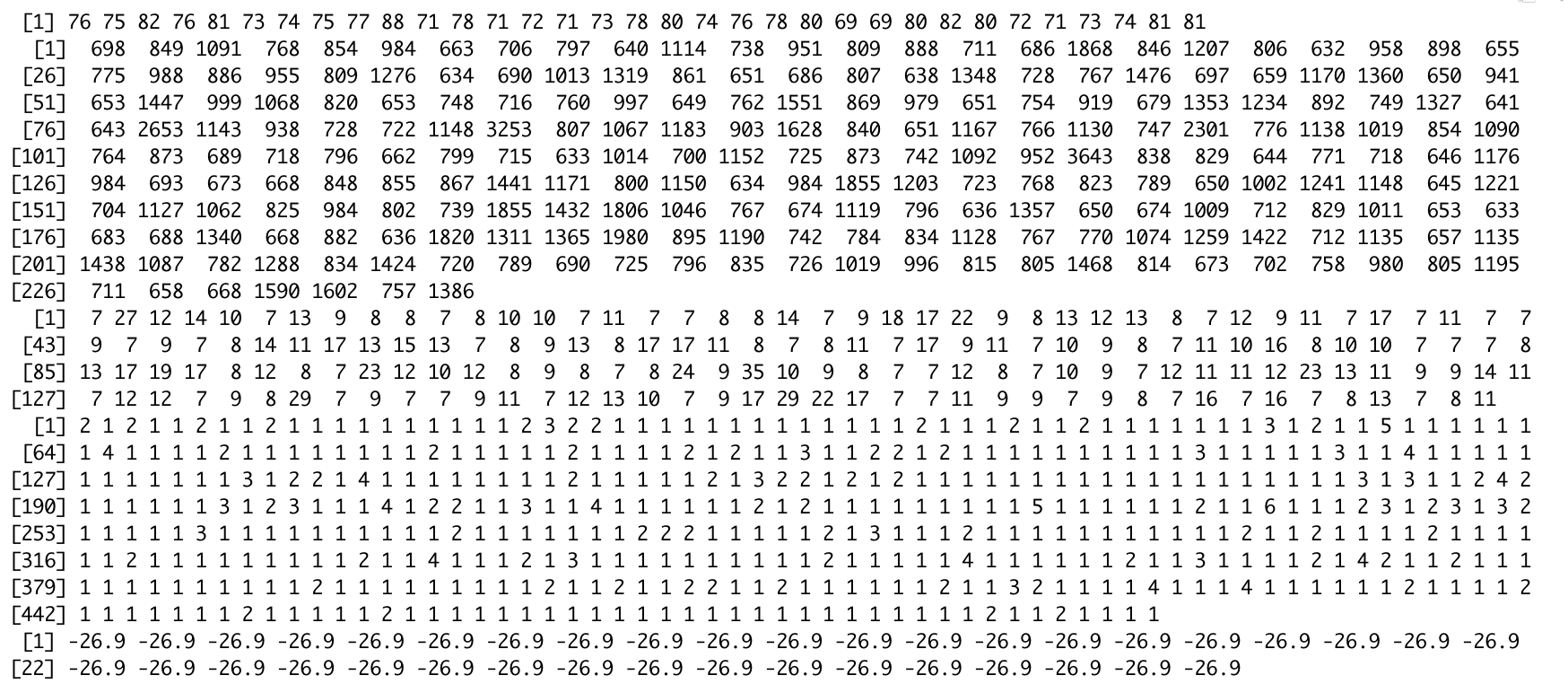


Furthermore, we noted that the *age*, *duration*, *campaign*, and *previous*, predictors had a right skewed distribution while other predictors appear to need transformations. See Figure 5 for an example of predictor histograms observed during our review. For the complete list of histograms observed, see the ***Appendix A*** in section ***IX. Appendix*** of this report. These distributions violate the key assumption of normal distribution. In addition, various predictors appear to have numerous outliers. See outliers identified using the *boxplot.stats* function in Figure 6. To correct, we used a preprocess function to transform (Box-Cox transformation), center, and scale the numeric data.

**Figure 5: Sample Plots for Predictor Distributions**



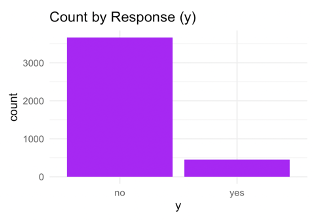
**Figure 6: Results of Predictor Outlier Search**

****

* 1. **Imbalanced Target Variable**

In addition to graphing the predictors, we also graphed the response variable to see the distribution between individuals who accepted and declined the offer. As seen on Figure 7, the response variable appears to be binary (in “Yes”/ “No” format), but does not appear to be balanced. We found that 3,668 customers sampled (89% of total records) declined the offer. To balance the target variable, we randomly sampled 370 observations from the customers who declined the offer (“No” responses) to match the number of offer acceptances. This produced a total data set of 740 observations.

**Figure 7: Count of Target Variable Responses**



* 1. **Identify Multicollinearity in Predictor Variables**

To understand the relationship between numeric predictors, we used a pairs scatter plot. The complete scatter plot can be seen on ***Appendix B: Scatter Plot of Predictor Pairs*** of section ***IX. Appendix*** in this report. From the scatter plot, we found weak correlation between a couple of predictors, and thus concluded that the requirement of no multicollinearity had be satisfied. No adjustments were necessary in this area.

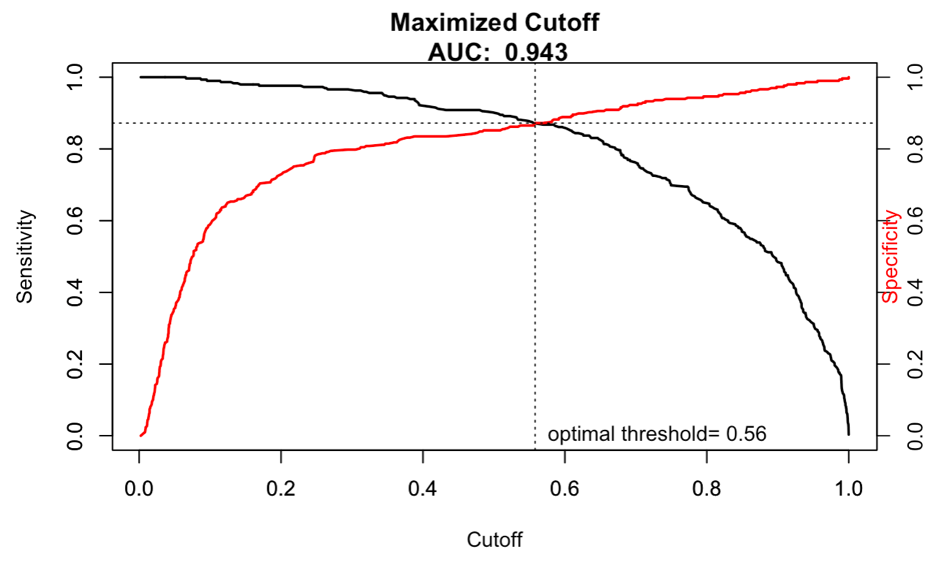
1. **Findings (Results)**
   1. **Model Accuracy, Sensitivity, and Specificity**

In total, we modified the logistic regression model four different times and compared to the LDA model results. First, began with the logistic regression model with all predictors, then simplified the model by reducing the number of predictors, followed by adjusting the cut-off and applying the pre-processing transformations. Finally, we compared the results with the LDA model. The model accuracy, sensitivity, and specificity results are seen on Table 1 below, with green highlights indicating the highest scores.

**Table 1: Model Accuracy, Sensitivity and Specificity Results**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** |
| Logistic Regression - Full Model | 87.84 | 86.30 | 89.33 |
| Logistic Regression - Simple Model | 87.16 | 86.30 | 88.00 |
| Logistic Regression – Simple w/ Cut-off Adjustment | 85.81 | 86.30 | 85.33 |
| Logistic Regression – Simple w/ Pre-Processed | 87.16 | 83.56 | 90.67 |
| Linear Discriminant Analysis (LDA) | 85.81 | 89.04 | 82.67 |

The logistic regression model that incorporated all the predictors (Full Model) produced the highest accuracy across all models tested. However, as we saw in the data section of this report, not all predictors appeared to add value. In addition, a highly complex model with 20 predictors is susceptible to overfitting. As a result, we used the step function to create a smaller, simple model. The resulting simple logistic regression model still had high accuracy, and a sensitivity rate that held constant, but a specificity that was slightly reduced from the full model. Then, to further refine the simple logistic model, we refined the cut-off by finding the optimal point between sensitivity and specificity as seen on the graph below.

******

By adjusting the cut-off to 0.56, we actually lost accuracy and specificity. To further refine the logistic regression model, we centered, scaled, and transformed the predictors through a pre-process function. Additional details behind the pre-processing of predictors can be found on the section ***V. Data*** of this report. As expected, the resulting model increased accuracy and specificity, but not sensitivity. As sensitivity is the measure of accurately predicting true positives, for this case, sensitivity is very important, as the purpose of the case is to correctly identify customers who will purchase the bank product.

The last model that was reviewed was the LDA model. As expected, the LDA model had the lowest accuracy. This is because the LDA model works best when the data has a normal distribution. Our data, as seen in the Data section of this report, did not have normal distributions.

Finally, as the simple model produced high accuracy and a consistent sensitivity, we chose to keep this model as the overall best.

**b. Key Predictors**

During our review, we noted a consistency in significant predictors across the logistic model summary results. As seen in ***Appendix D: Model Summary Results***, the following predictors were significant with p-values under 0.05:

1. *Month-MARCH* (the month in which the customer was contacted), p-value of 0.00174
2. *duration* (the duration of the customer call), p-value of 2e-16
3. *contact-TELEPHONE* (whether the customer was contacted via cellphone or telephone), p-value of 0.00114
4. *emp.var.rate* (quarterly employment variation rate ), p-value of 7.2e-11
5. *cons.price.idx* (consumer price index or CPI), p-value of 9.11e-06
6. *cons.conf.idx* (consumer confidence index), p-value of 0.02121 and
7. *pdays-CONTACTED* (whether the customer was contacted in a previous campaign), p-value of 0.047

Specifically, we used the results of the Simple Model (Attempt #2 under ***Appendix D***) to observe predictor relationships with the target variable. We found that a call in the month of March positively changed the odds of the customer accepting the bank product by 3.797. Also, a unit increase in the phone call duration (*duration* variable) increased the odds of a customer accepting the bank product by .0071. Other predictors that increase the odds of a customer accepting the bank product include the *consumer price index* (1.6 increase in odds per unit change) and whether the customer was recently *contacted* during a previous campaign (produces a change in odds of 0.089).

Predictors that negatively affect the odds of a customer accepting the bank product include *contacting the customer on a telephone* (decreases the odds by 0.00149) and an increase in the *employment variation rate* (decreases the odds by 0.0986). Intuitively, variations in employment such as increase in unemployment and decreases in consumer confidence, all indicators of a struggling economy, would affect a customer’s willingness to purchase a long-term deposit, which limits liquidity.

1. **Conclusions and Recommendations**

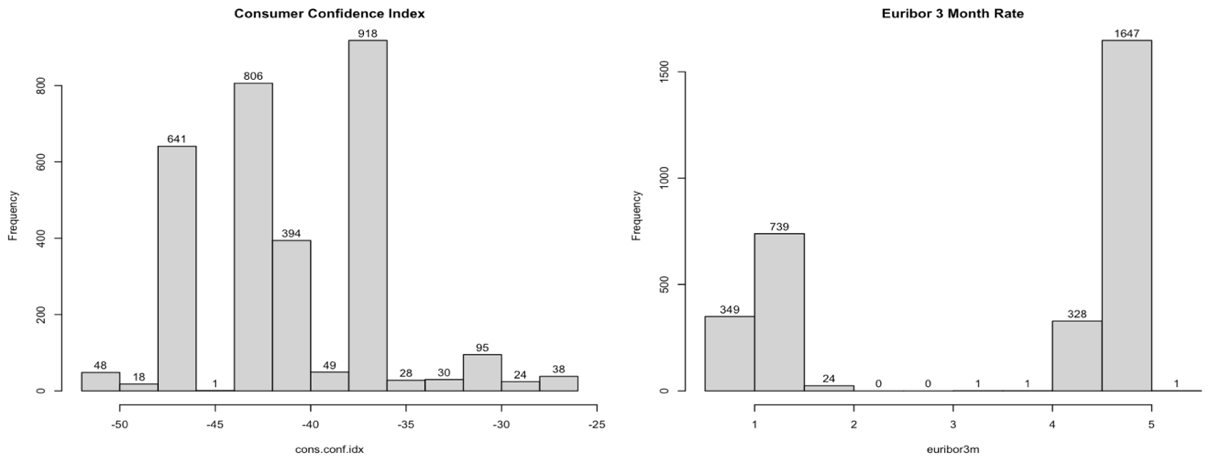
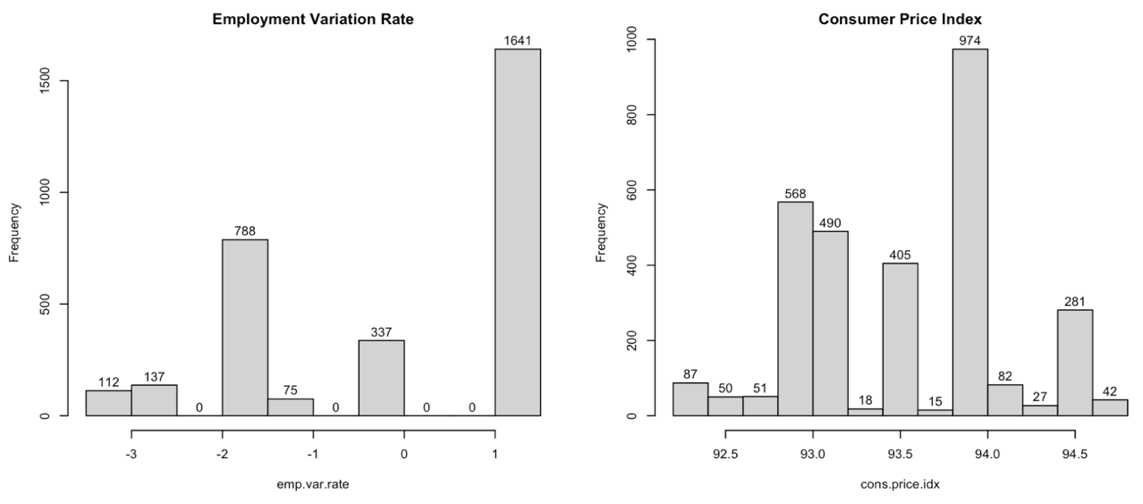
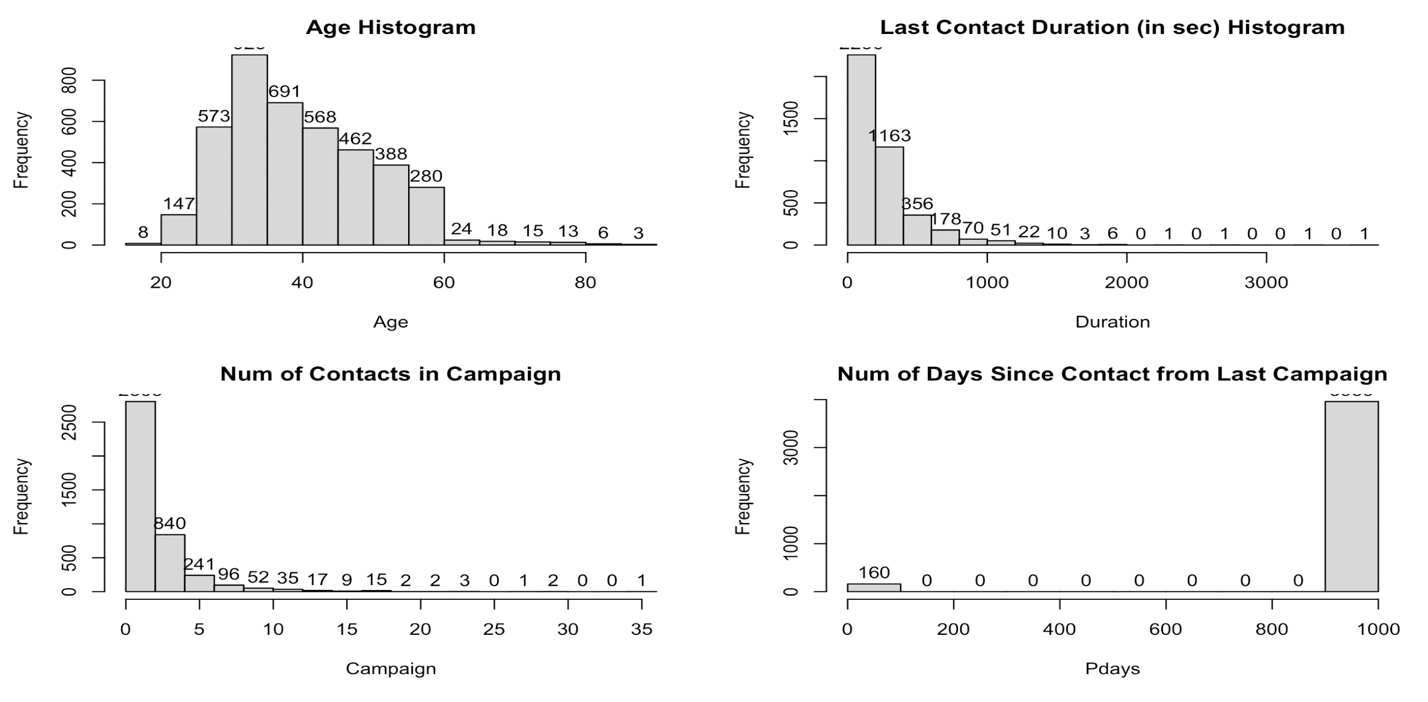
In the beginning, we noted that the case problem was to find a model that can accurately predict which customers will subscribe to the long-term deposit. After our analysis, we found that the best overall model was the simple logistic regression model with the following seven predictors: whether the customer was previously contacted, the month of contact, the duration of call, consumer price index, consumer confidence index, Euribor, and number of days since last contact. Although the simple logistic regression model had the second highest accuracy rate of 87.16%, it also had very good sensitivity and specificity rates of 86.30% and 88.00%, respectively.

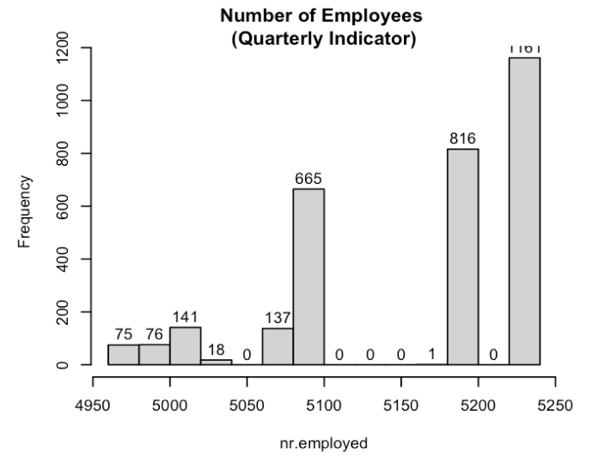
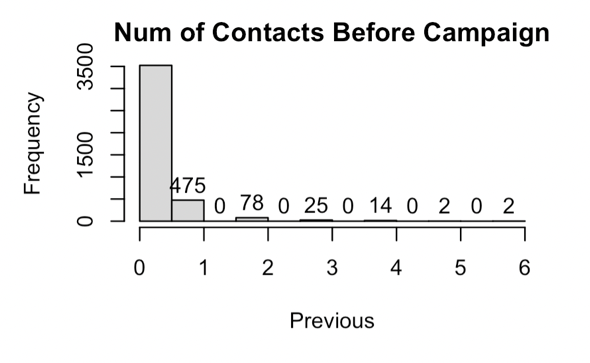
In addition to finding a highly accurate model, we also found key predictors that significantly influenced whether a customer subscribed to the bank product. Such indicators include the month of contact, duration of the campaign call, whether a customer was contacted in a previous campaign, in addition to economic factors such as CPI and the employment variation rate. For increased success in future campaigns, we recommend that the bank contact customers during the month of March and increase the duration of the telephone calls. Furthermore, it is to the bank’s advantage to monitor economic factors such as the employment rate and consumer confidence, as they may also impact the customer’s likelihood of subscribing to the long-term deposit.

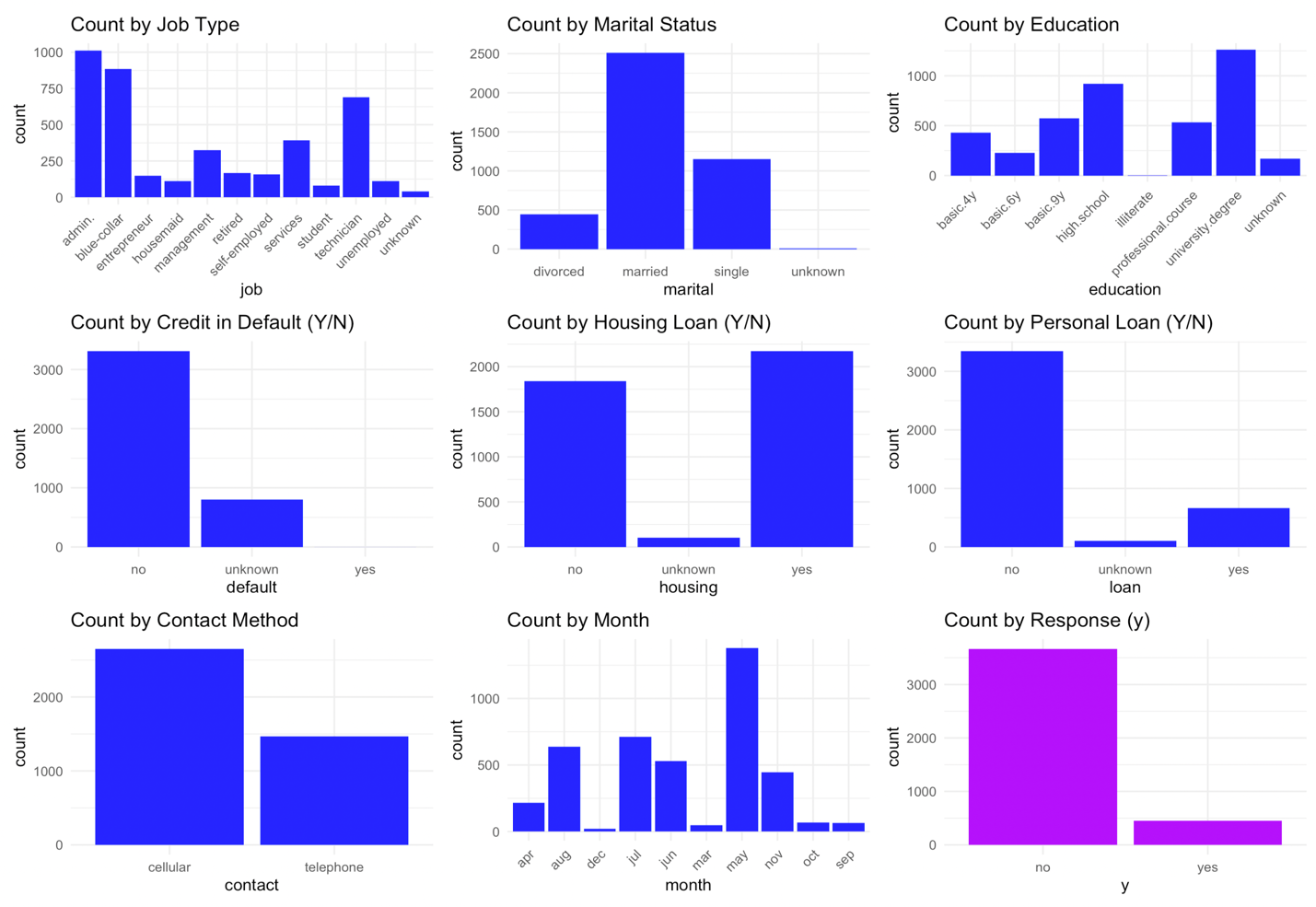
Finally, in the future, a more comprehensive analysis of the case would include the evaluation of non-parametric models such as Support Vector Machine (SVM). Although the model may be more complex and harder to interpret, it may result in more accurate results with larger sample sizes as the Portuguese bank continues to campaign and collect more customer data.

1. **Work References**
2. ‘Data Mining’ by John A. Bunge, Dean H. Judson (2005) in Encyclopedia of Social Measurement
3. ‘Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study’ by Maja Pohar, Mateja Blas, and Sandra Turk
4. **Appendix**

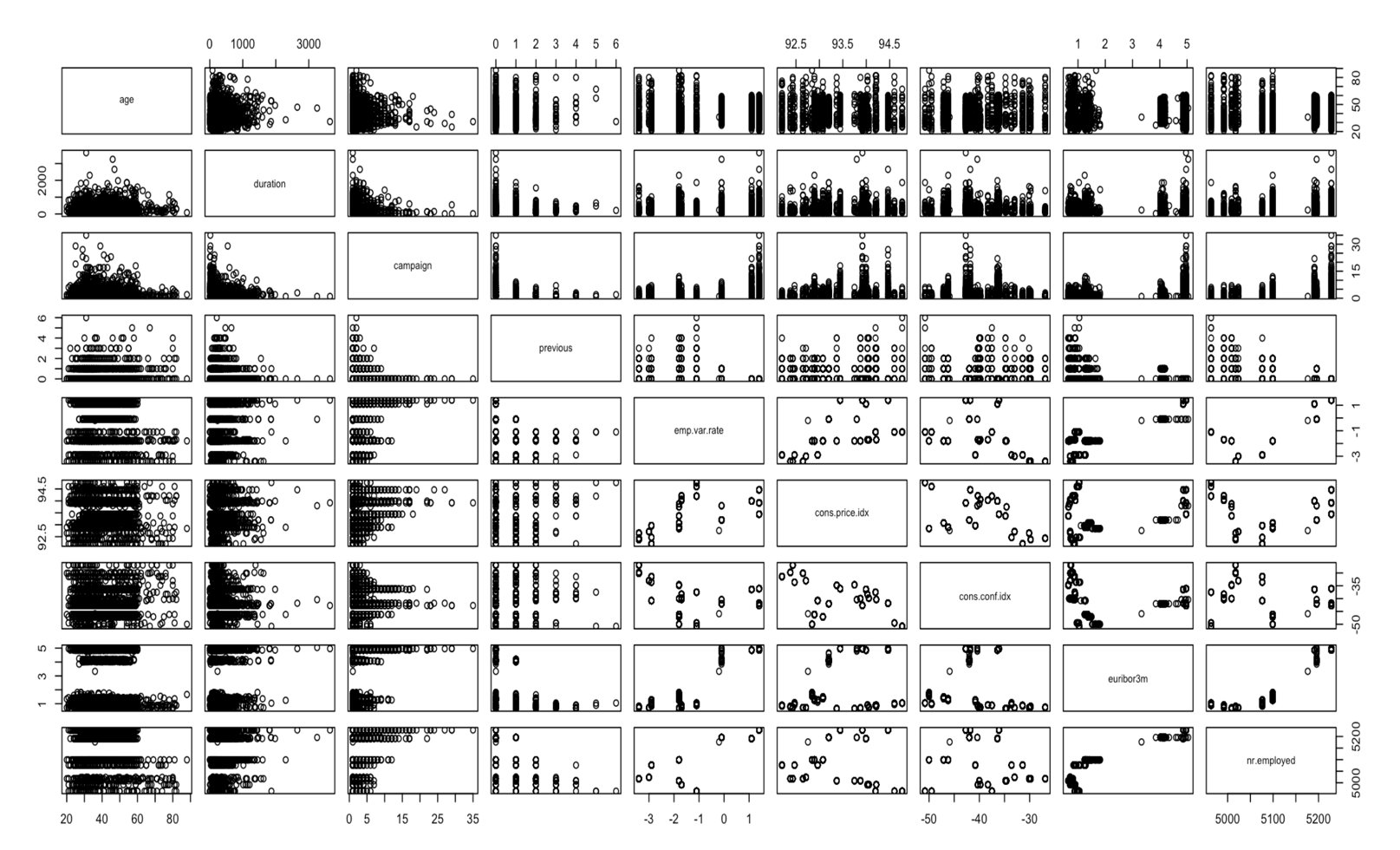
***Appendix A: Plots of Predictor Distributions***







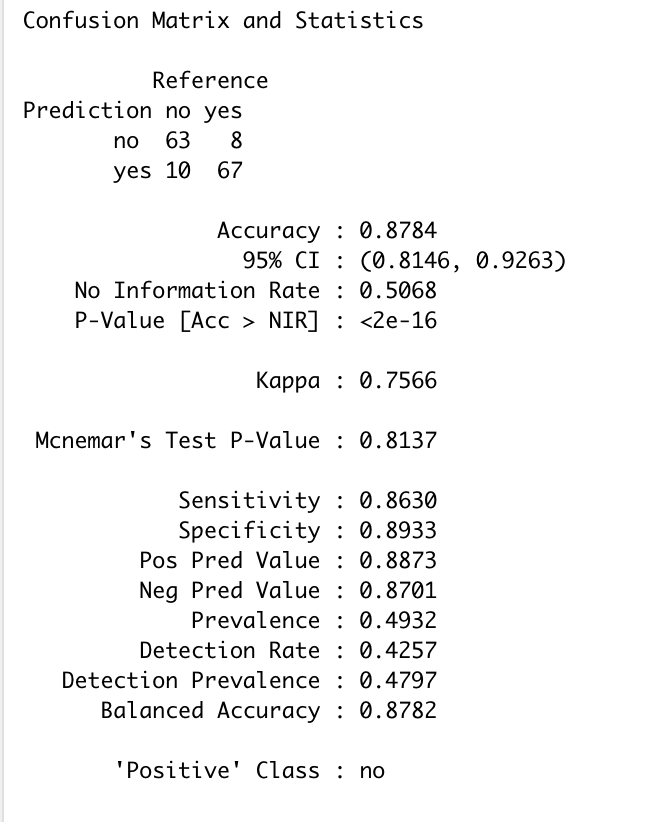
***Appendix B: Scatter Plot of Predictor Pairs***



***Appendix C: Confusion Matrix Results***

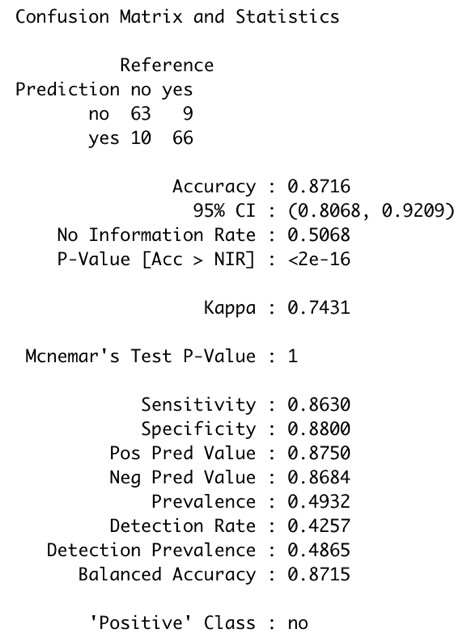
***Attempt #1: Logistic Regression –***

***Full Model***

******

***Attempt #2: Logistic Regression –***

***Simple Model***

******

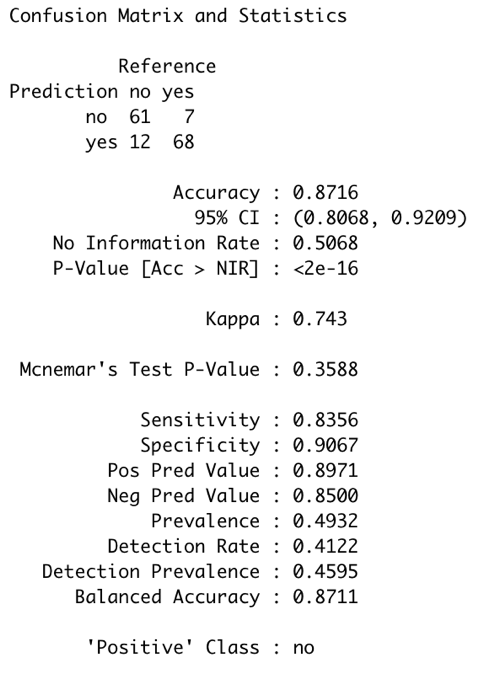
***Attempt #3: Logistic Regression –***

***Simple with Cut-off Adjustment***

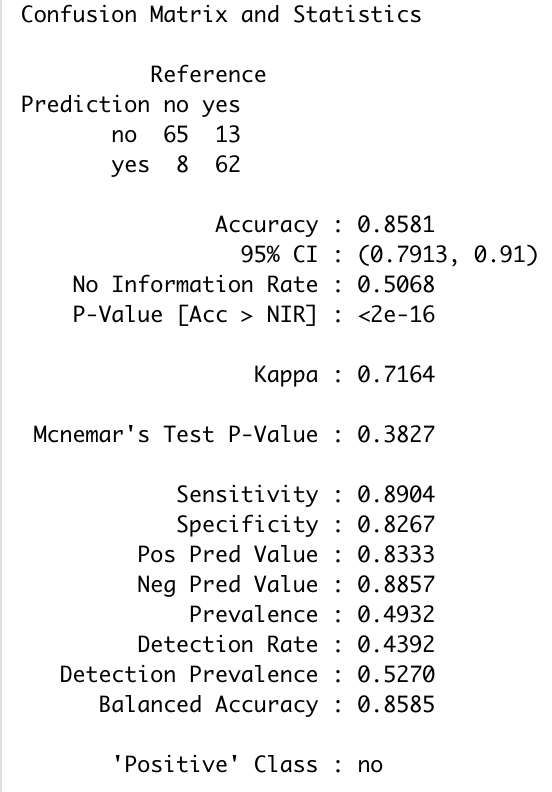
******

***Attempt #4: Logistic Regression –***

***Simple with Cut-off Adjustment and Pre-Processed Predictors***

******

***Attempt #5: LDA Model***

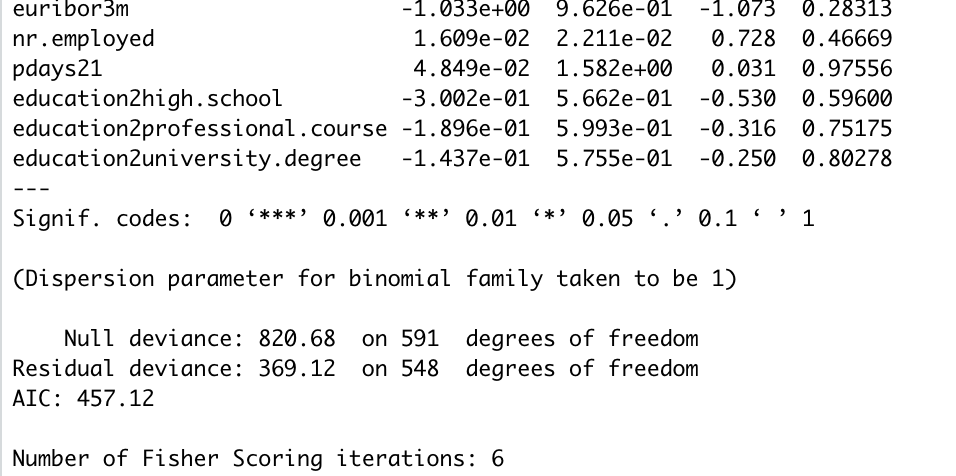


***Appendix D: Model Summary Results***

***Attempt #1: Logistic Regression –***

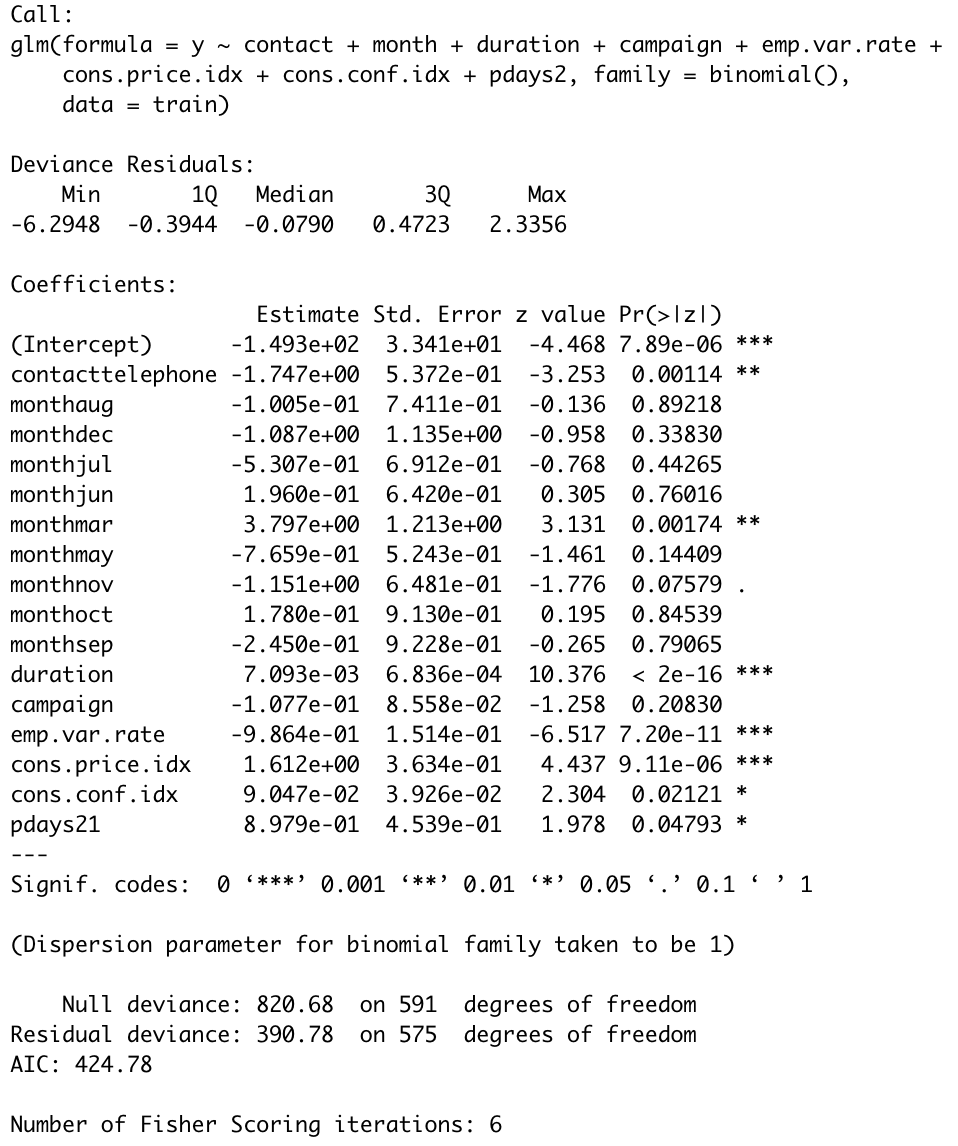
***Full Model***

******

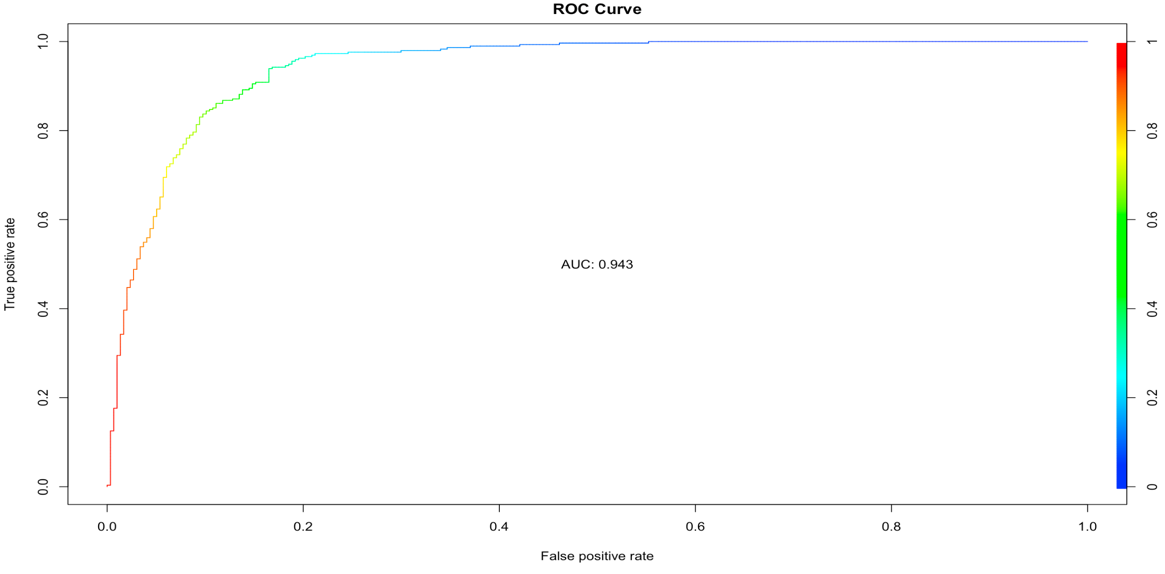
******

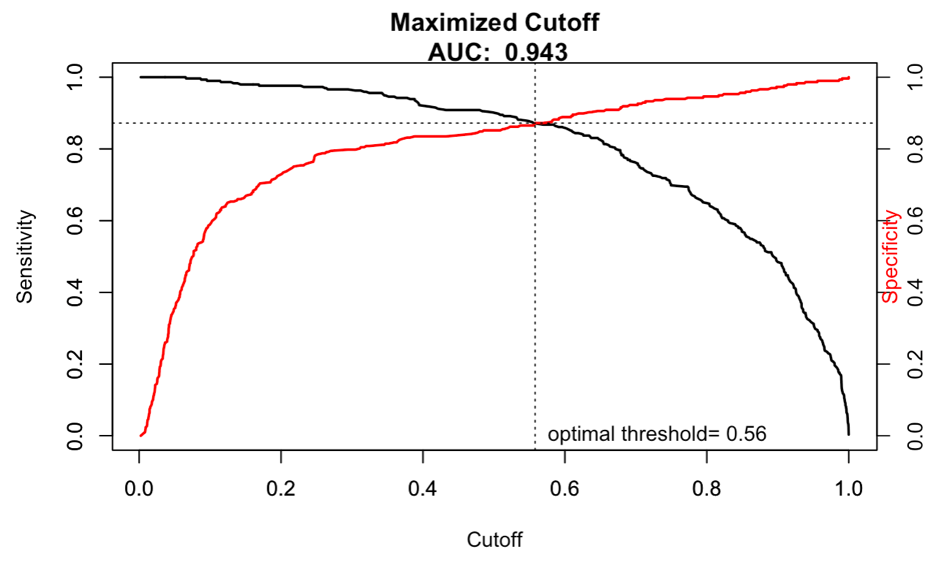
***Attempt #2: Logistic Regression –***

***Simple Model***

******

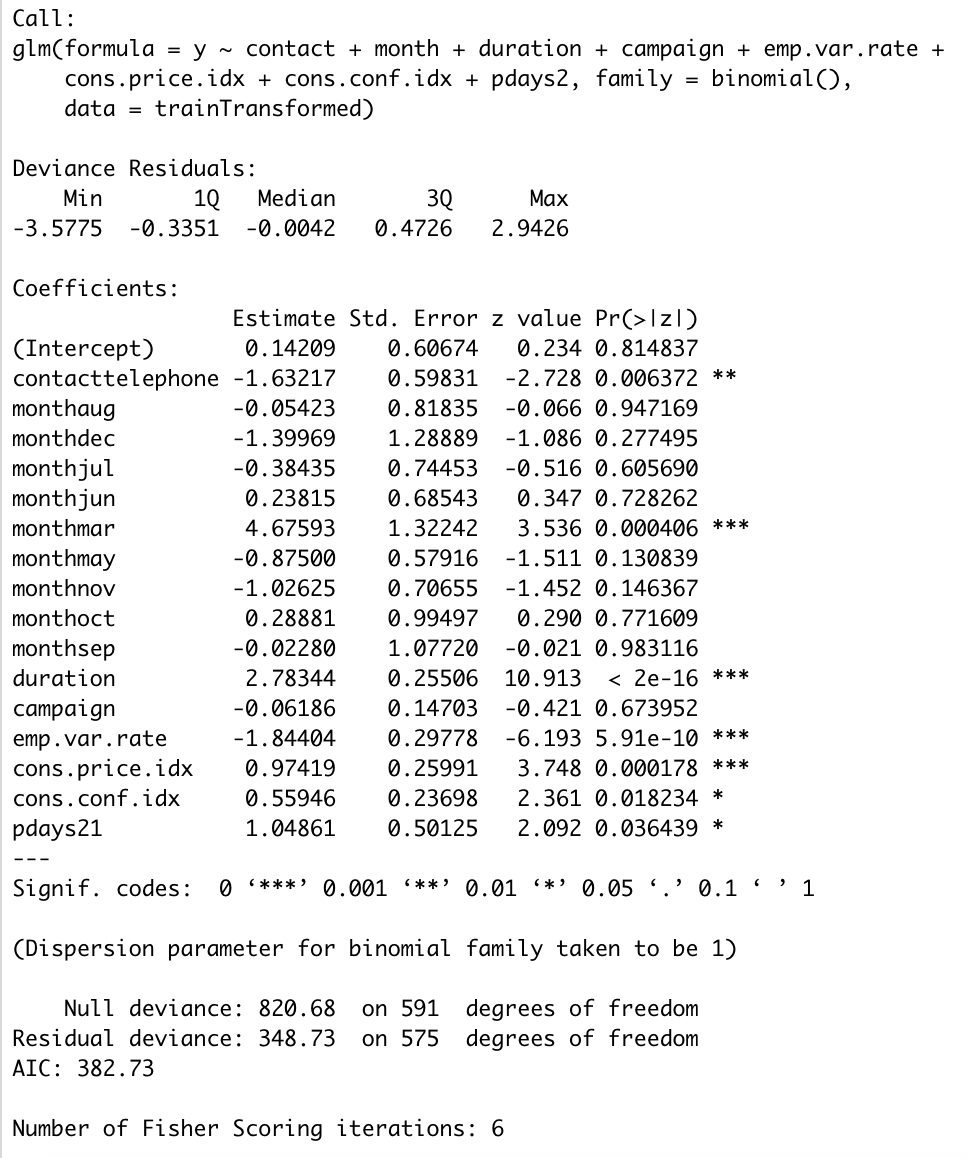
***Adjustments to Cut-off***

******

******

***Attempt #3: Logistic Regression –***

***Simple with Cut-off Adjustment and Pre-Processed Predictors***

******