Project 2: Classification of Subscribers to a Time Deposit Account

Walter Lai, Rohit Channe, Adam Canton

Part 1: EDA and Simplistic Model –

Our Data set consists of 41,188 observations of individuals contacted by a business in an effort to sell them a time deposit account. The data is composed of 20 variables collected about the marketing campaign, the individuals contacted, economic indicators, and other attributes, as well as the 21st variable, which is the response - subscription decision. The data contains no explicit NA values, however, there are unknown categories in some of the categorical variables’ levels. An analysis was performed on both the original data set and a second data set in which the unknowns were removed.

Our goal with this analysis is to better understand what variables are informative predictors of a customer’s subscription decision to a timed account. As well as trying to predict, with the best accuracy possible, what a customer’s subscription decision will be given a set of variable inputs. Our method to classify these customers, will be to examine the current trends in our data and try to manually pick out what variables contribute to the success most, as well as employ at least 2 selection techniques for further inference on variable contribution.

To find which variables may contribute the most to subscription decision we looked at distributions of continuous variables split by subscription decision, and for categorical variables we looked at percentage-based bar graphs split by both the explanatory and response levels. There were some interesting trends to note:

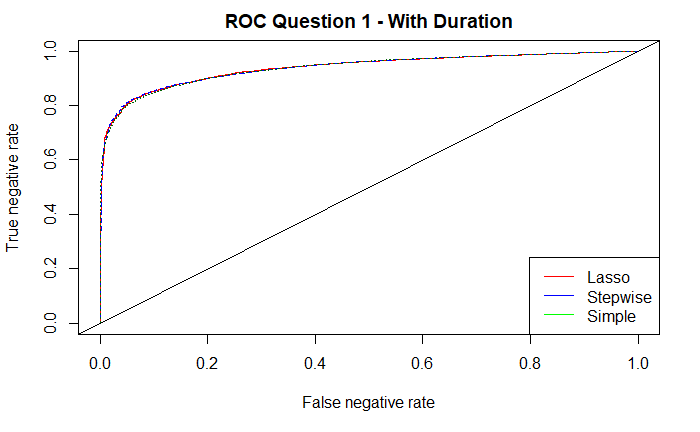
* In general the age distribution was equal among No and Yes
* We see large differences in subscription medians in emp.var.rate, euribor3m, nr.employed, and duration
  + To lesser extent cons.price.idx and previous had differences
* Month sees drastic changes in percent of success
  + the highest percentage success months are also the ones with the least call volume
* Nearly double the success rate when contacted on cellular vs telephone
  + Stranger, this does not appear to be age dependent.
* Those who had previously subscribed are more likely to subscribe in the future.
* Largest single group of education who subscribe is Illiterate.
* Most success amongst those who are retired or are still students.

Some of these continuous variables make sense with respect to the response. We would expect employment variation rate to play a role. A person who is higher in employment variation may be less stable professionally, or in general. One would expect this to have adverse effects on a subscription to a long-term deposit. We also saw an increase in the median number of subscriptions at lower levels of the Euribor3m rate, an interbank interest rate based of averages of interest rates around the Eurozone. Number employed also has a stark difference between medians and spreads by subscription. It seems that subscription decision is inversely related with gains in number employed, though we are at a loss for the mechanism by which this influence occurs.

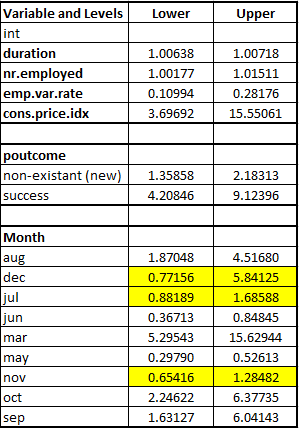
Finally, we have duration. This variable turns out to be a great predictor of success (approximately 73% accuracy alone). However, it may be a bit disingenuous to use it in our predictive analysis. This variable tracks how long the sales calls are, thus shorter times are generally No answers, while longer times are much more likely to be Yes answers. Unless this model was going to be used in benchmarking, duration would be redundant, because if I know the duration of the call, then the call is over, and the truth would be known. I have used duration in our models for comparison purposes.

In addition to the EDA, we looked at some feature selection out of a step model (based on AIC) and a LASSO modelcompared with our custom group-chosen models. Through the selection we learned a few new things about the data/variable relationship.

* Number employed and the Euribor3m rate – we only need one of these.
* Somehow our Lasso model was able to achieve comparable accuracies without either nr.employed or euribor3m, and went heavy on emp.var.rate and cons.price.idx
* Lasso placed more importance on categorical variables than any other model
  + Found this to be counter to how we built our simple model

The models performed quite close together (86-87%), generally only being different by 1-2% in the relevant statistics. Our simple model used the least variables and all variables were significant, though some levels of the categorical variables were not. As can be seen from the ROC curve set to the right, the three models are nearly indistinguishable from each other. The practical significance of their differences may be irrelevant depending on how critical those few percentage points are to our overall business plan.

Our simple model is as follows:

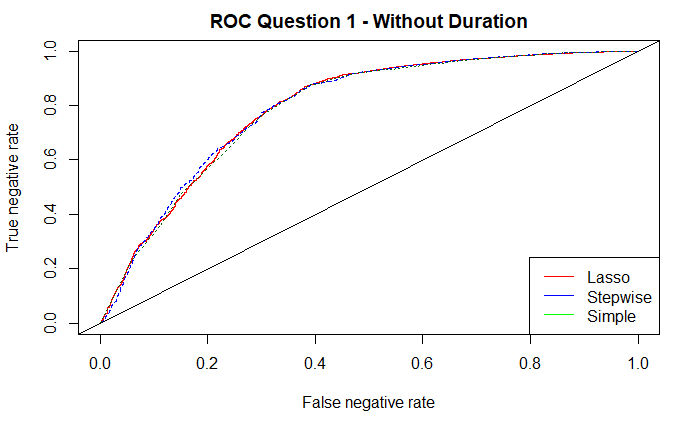


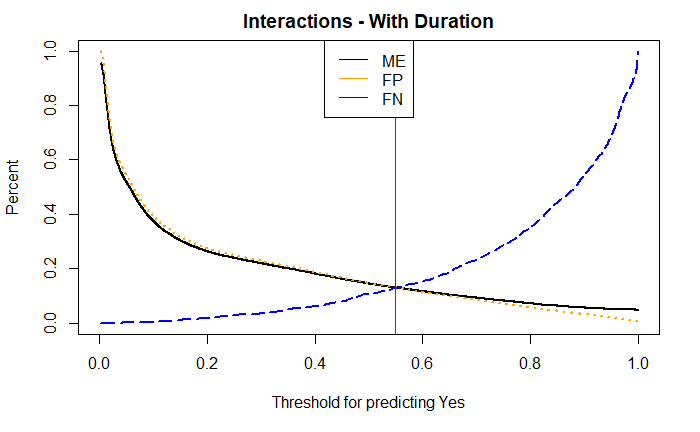
*Note: reference levels are “no” for poutcome and “April” for month.*

*Point estimates can be found in appendix*

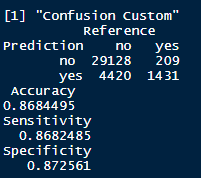
Here we can see the equation that our simple logistic model resulted in. The 95% confidence intervals to the right display the expected average changes in the odds ratio from a unit change to each variable or change in variable level – all else held constant. The highlighted portions of the table show the insignificant levels for an otherwise significant categorical.

Some of our odds multipliers when compared to their respective graphs seem a bit off, such as: duration, nr.employed and cons.price.index. Duration here is measured in seconds, so a 5 minute call is duration 300! This is why it seems like such a minor multiplier. Likewise, nr.employed is measured in the thousands. Cons.price.index appears to have a massive effect, but the underlying variable values have a range of approximately 2.5, so has small unit changes.

After having found what we felt was an easily interpretable model with sufficient accuracy toward ours problem, it was time to pull out duration and check how this model would do without information we likely wouldn’t have if we were trying to predict a customer’s subscription decision.

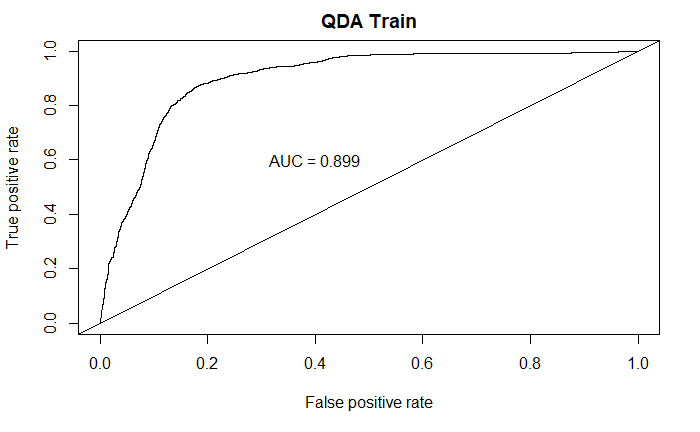
Without duration in the model, the accuracy and classification of both the positive and negative classes dropped by approximately 10-15% (70-75%). This drop in accuracy is nearly independent of what model we were running (simple, step, or lasso). Our simple model was behind by 1-2% with duration, however without it the simple model is about 3-5% better in total accuracy than the step and lasso models. I think this lends some credibility to our simple model’s accuracy in the face of more uncertainty. Though this would depend on whatever penalty is experienced for incorrect predictions.

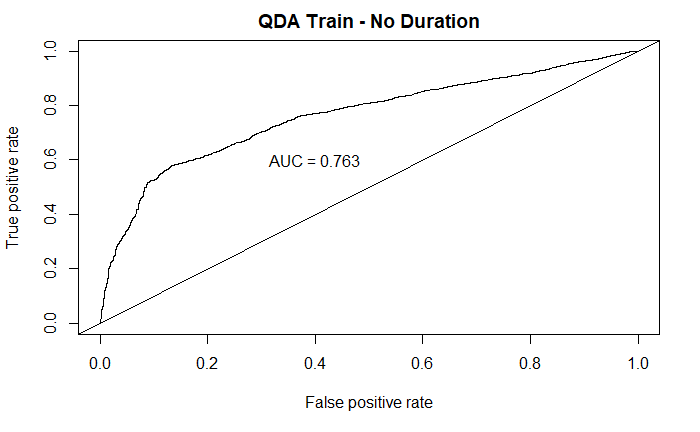
Part 2: Complex and Non-Parametric Models

 In our first attempt at creating a more complex logistic model (referred to as custom) we set out to identify possible interactions between our already included variables. In order to do this, we started with our simplistic model and used the stepAIC function to search for significant interactions between the 6 variables we had included already as well as euribor3m. Similarly, to part one we started with duration in and pulled it later to compare models.

The step process produced 3 candidate interactions: duration\*month, euribor3m\*duration, and duration\*poutcome. We included these interactions and scored the model. The results were about where we were in the simple model, approximately 87% accuracy, a rather marginal difference. We retried the selection with a different subset of variables and got a few new interactions, though none of them improved the model prediction. We felt this marginal difference in prediction was not worth the increase in complexity. Moreover, since all the new interaction terms were dependent on duration being in the model, the model would essentially default back to the simple version when that variable was pulled.

We further tried a KNN model, though our hopes were not high that this model would do better since we could not find any good separations in visual comparisons of our variables. We were quite surprised to find that the KNN model was relatively close in performance to our Logistic regression models. We scaled all 9 continuous variables (including duration) and used them in a KNN classifier where k = 7. The choice of k is dependent on how sensitive and specific you would like the model to be in classifying. Higher k’s were biased toward more negative classification. Performance metrics can be found in the appendix (Section 6).

Encouraged by our KNN model on the scaled original variables we further pushed into looking at a KNN model using Principal components, since some of our continuous variables were highly correlated. This model did not perform well against the previous model runs. We topped out at around 70% specificity and pushing specificity this high had the adverse effect of cratering our accuracy and sensitivity below 50%. However, we recycled these principal components to use in a QDA model. We chose a QDA model after only visual inspection of the PC graphs. This QDA model proved to be the best model we had found. On the training set it scored near 90% which is at least 3% above any other model. And on the test set scored 90.5. However, we need to run these without the duration crutch.

 As was observed when duration was removed from the simple model, removing duration from the QDA principal component’s analysis dropped the accuracy of the models by about 13%. The same effect occurred to the KNN model though not quite as drastically – only a 7% drop.

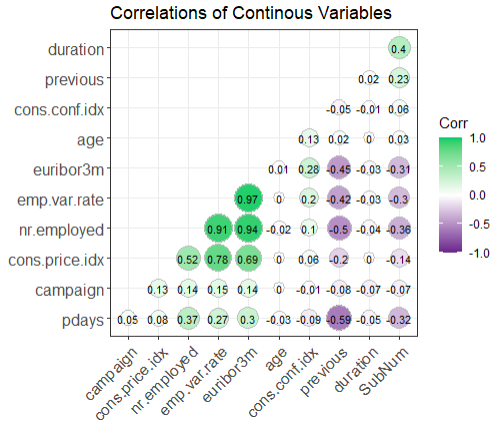
A second round of analysis was completed on the data set, except with all the unknowns removed. This reduced the data set to 30,488 observations a reduction of approximately 25%. This had the largest effect on categorical variables with large percentages of unknowns which were education and default. These two variables were not selected for when step selection was running, they were also not favored by lasso. All in all, removing the unknown categories did little to nothing for our models.

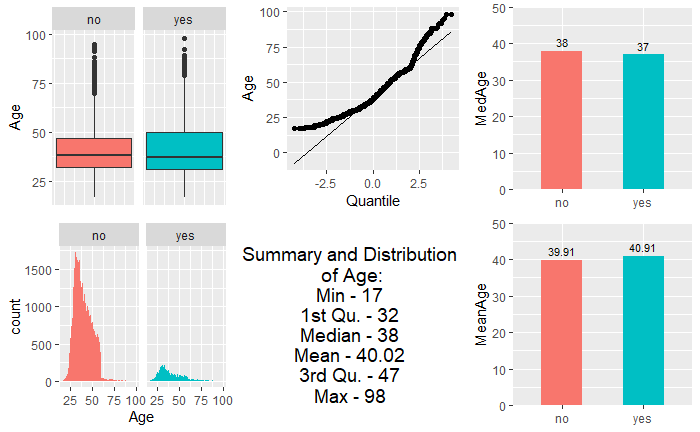
At the end of our analysis it appears as though the original simple logistic model was the optimal prediction tool both in terms of overall accuracy and low complexity. It was very difficult to find anything that would up our prediction accuracy that didn’t rely on duration. In the original data set we have an approximate 90/10 split of no to yes. We could be right trivially 90% of the time by just predicting no every time. If we allow duration – which seems to be bypassing the spirit of the analysis – then we can almost beat out our trivial accuracy. However, depending on the penalty of false positives/negatives we do seem to be able to push identification of those that will subscribe much higher than 10%, which is likely more valuable. It should be noted that a larger percent of subscriptions were successes in the months with the least sales calls. This may provide avenue for future study on the company (instead of customer) side of things. It may be that in the low sales call volume months that only the most likely to subscribe to the accounts are called, or it could be that the low volume of calls allows for longer duration/call and less burnout in the sales force. Further, since duration was such a good predictor of sales success, it might be worth it to try and find the variables that influence call duration.

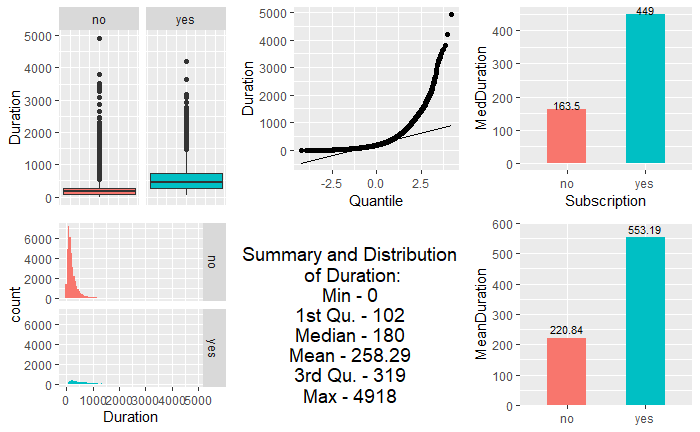
Appendix: Note Red is Subscription NO, Blue is subscription YES

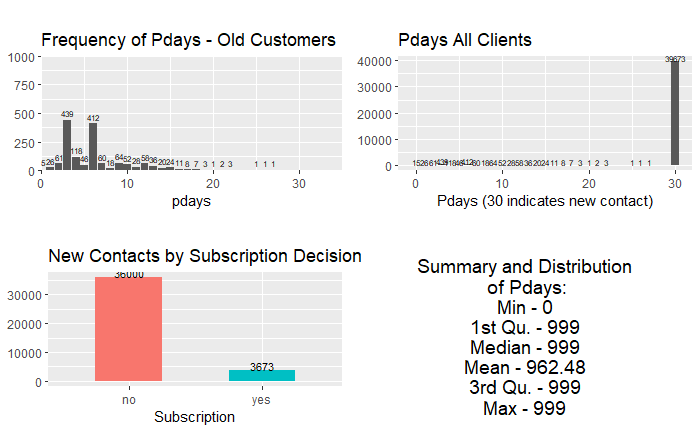
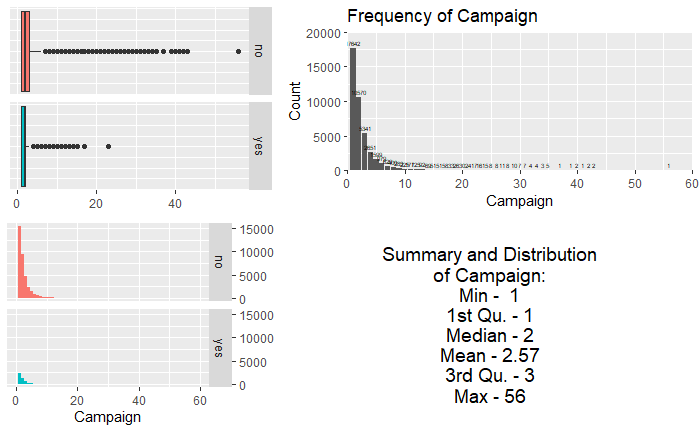
Sampling: We forced out training set to be 50/50 yes/no by sampling a random 3000 yes’s and 3000 no’s. This made our test set approximately 5% yes and 95% no

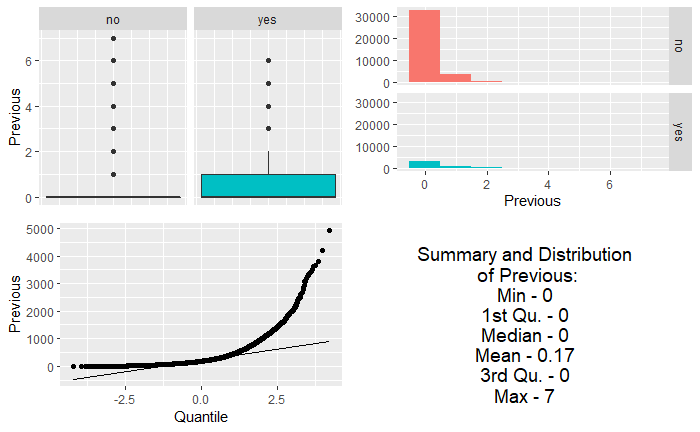
**Section 1 - continuous graphics**

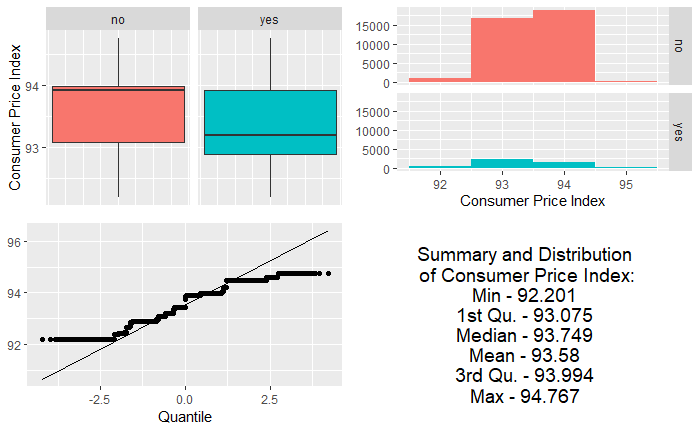


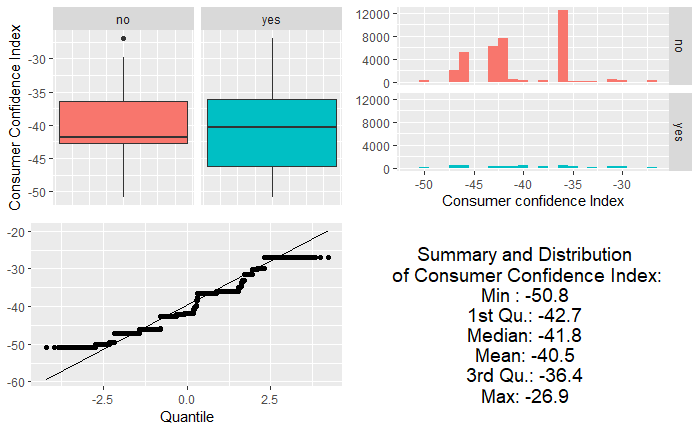


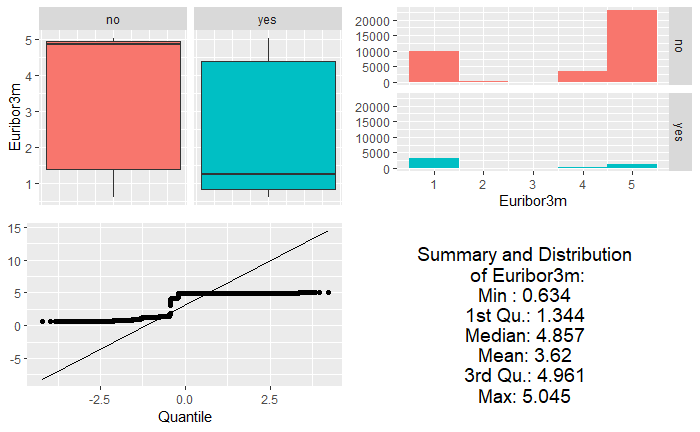


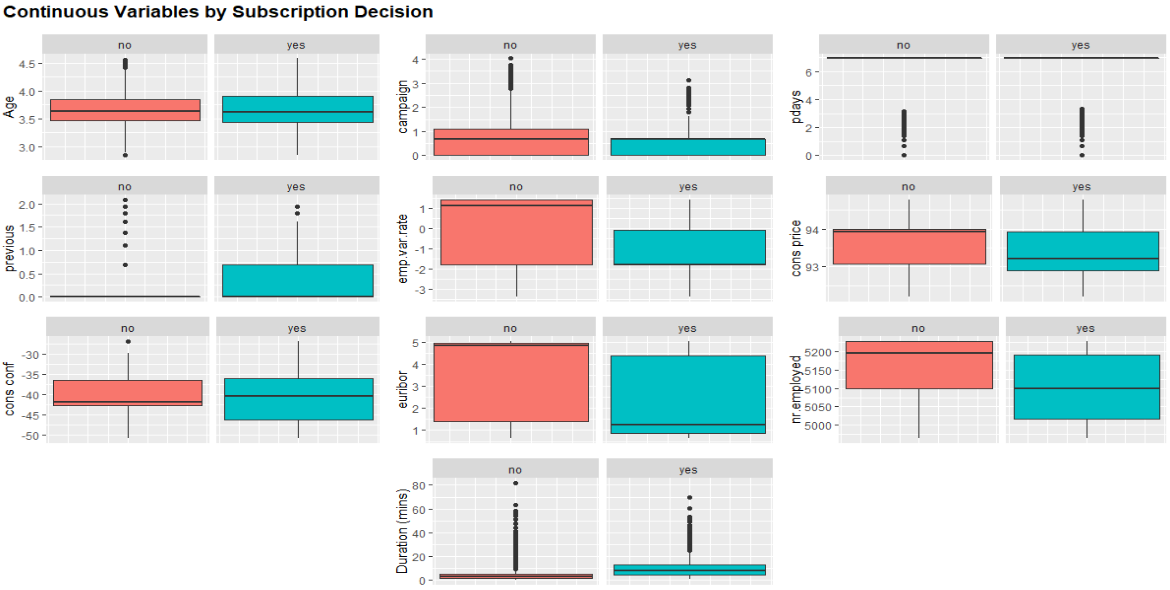








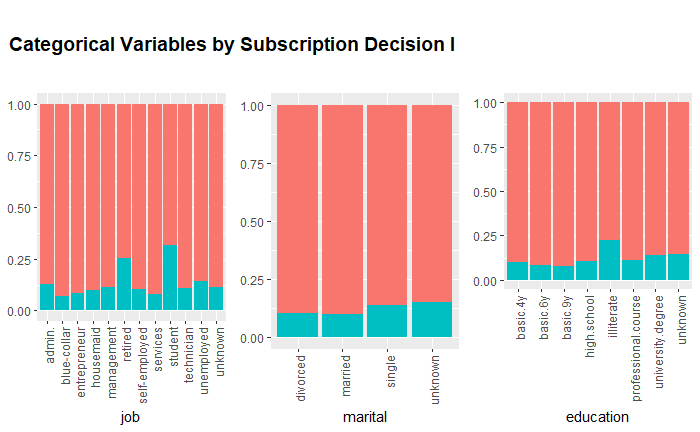


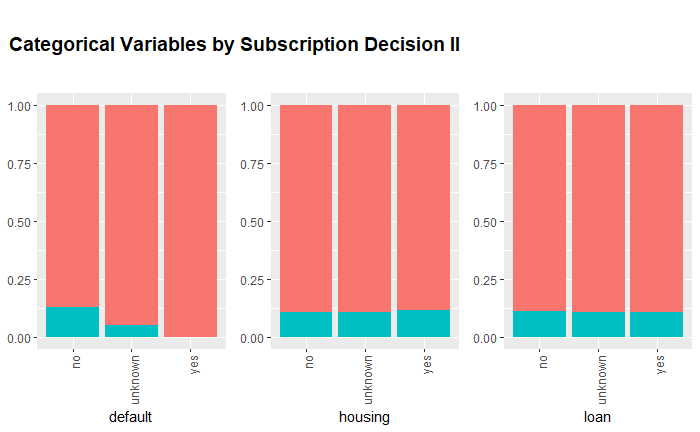


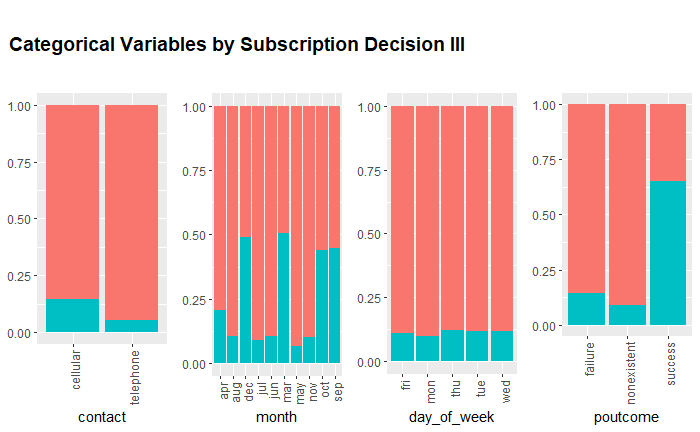
**Section 2 – categorical graphics**

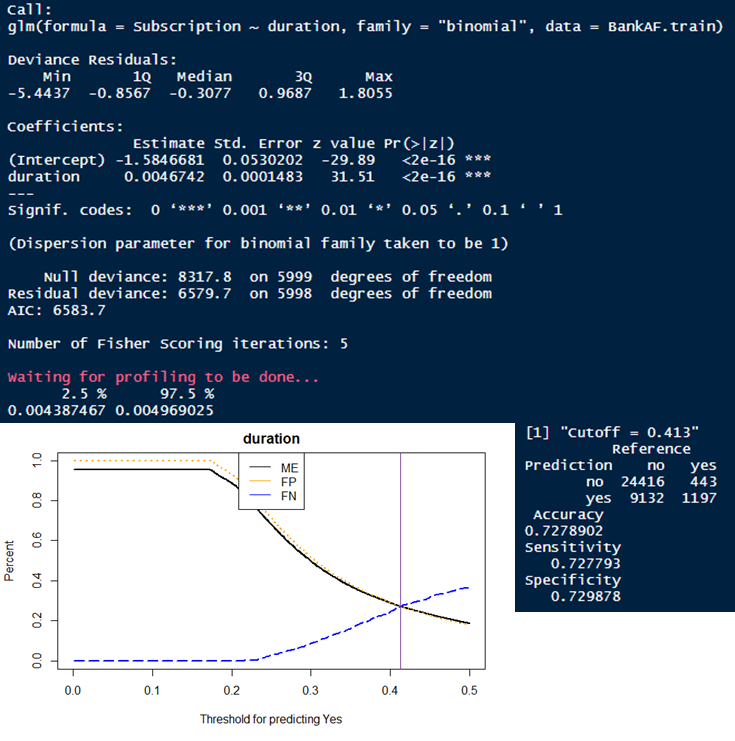
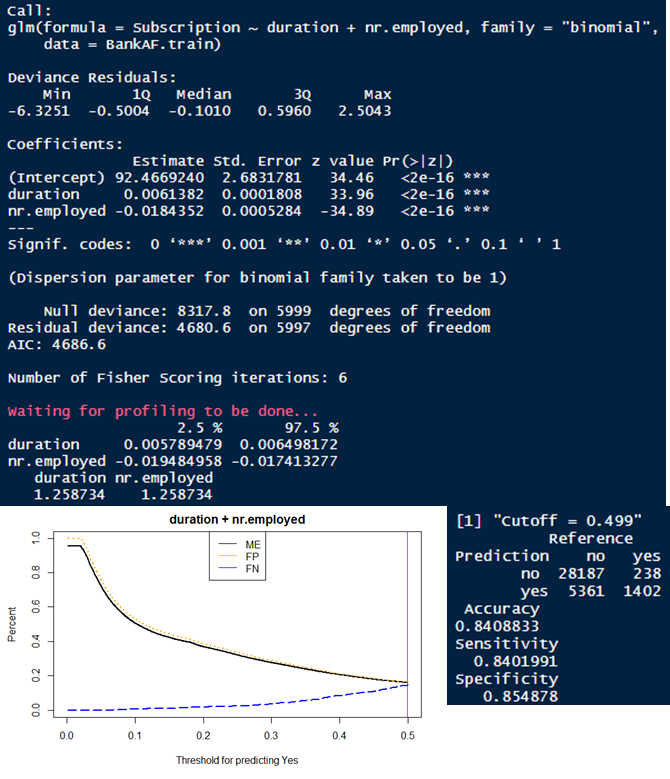
Note – 3 custom categoricals were made

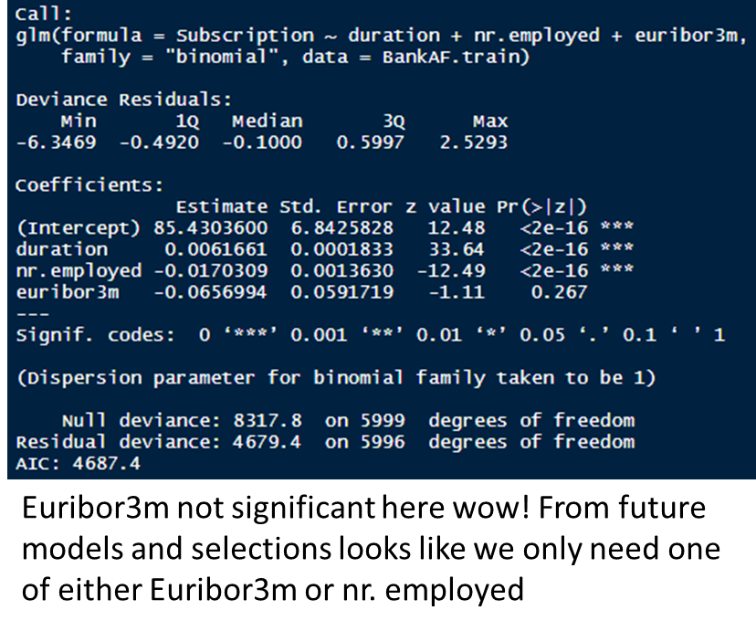
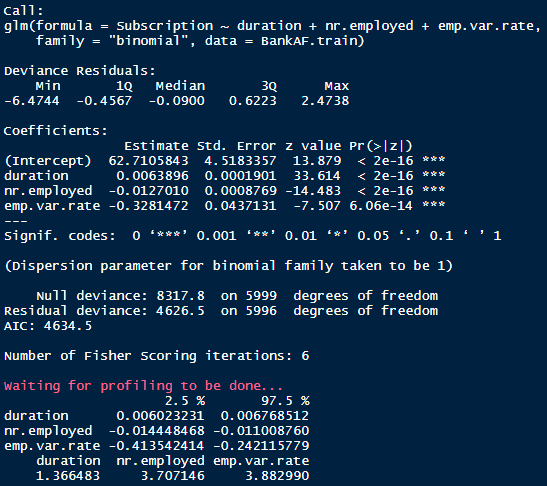
* Poutcome was morphed into new contacts and old contacts where old was previous success + previous failures, and new were the non-existent.
* Job was morphed to Workforce where those who had jobs were grouped and those who did not (student, retired, unemployed) were grouped together.
* Default was re-group to have no and unknown grouped while leaving yes alone. However, there are only 3 yes defaults. This variable did not have much information.
* None of these variables were selected or added any value.

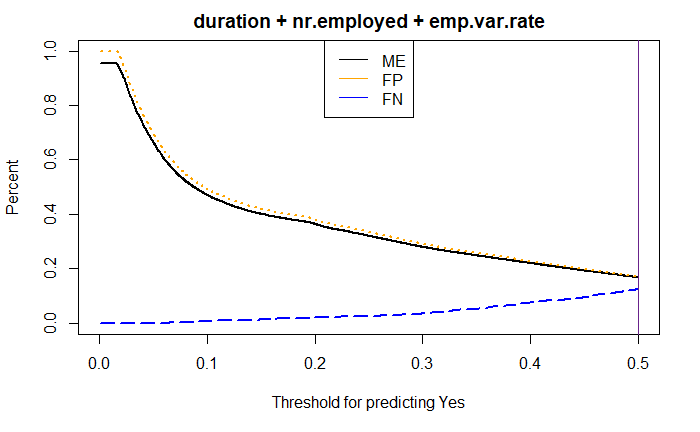
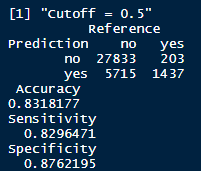


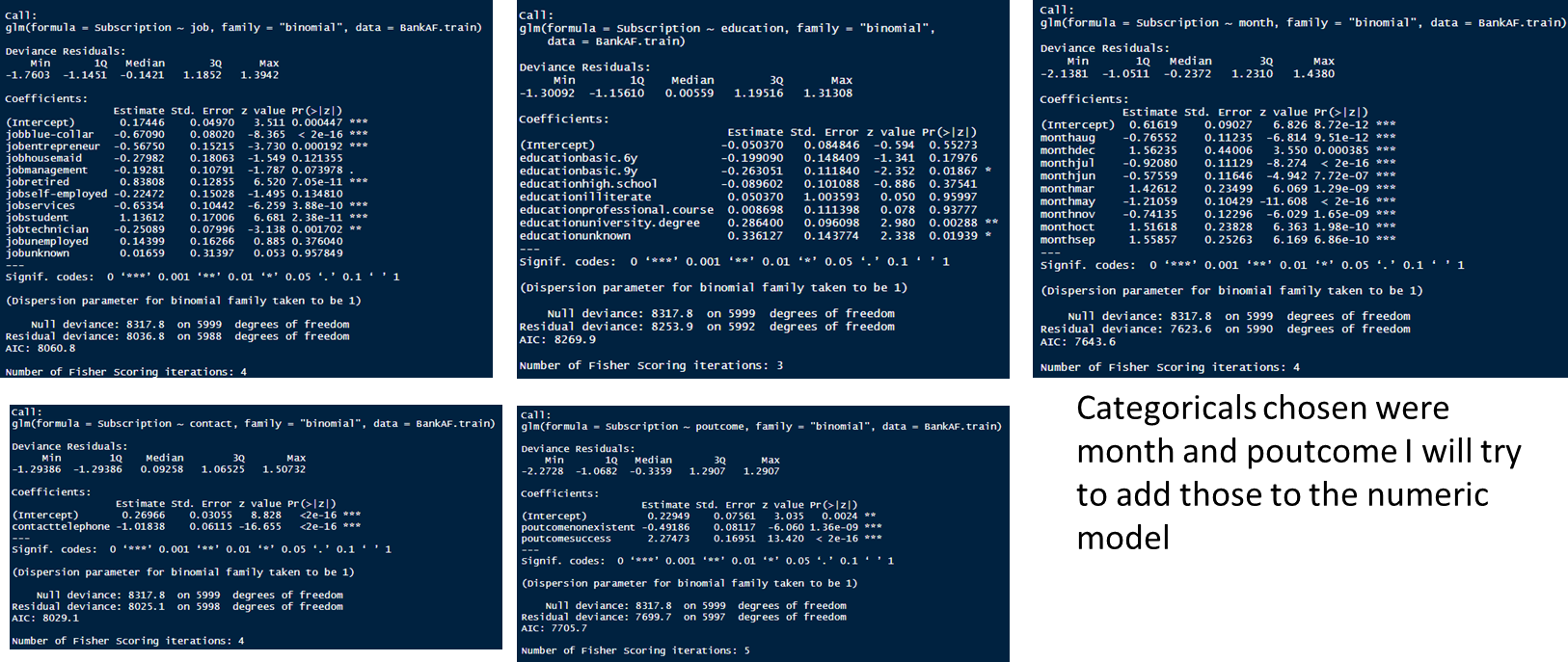


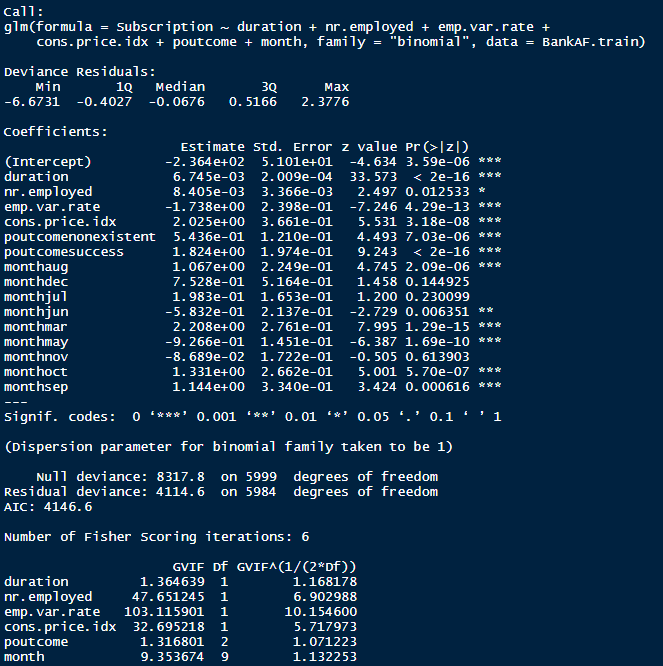
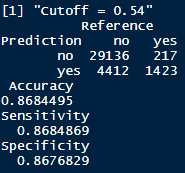


**Section 3 – Manual Iteration Selection – Simple Model**

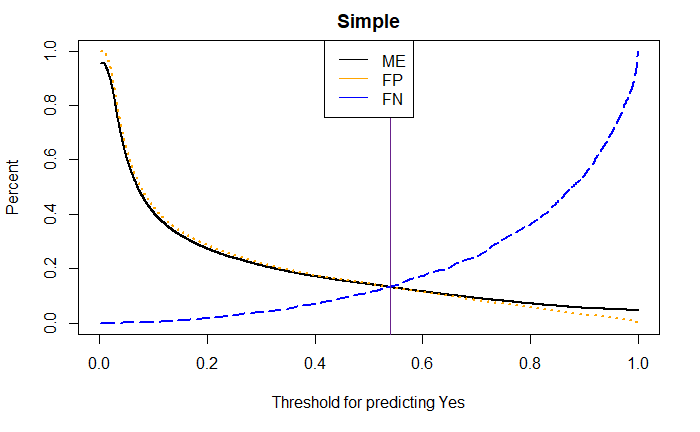


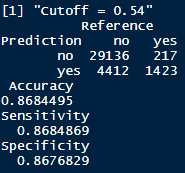


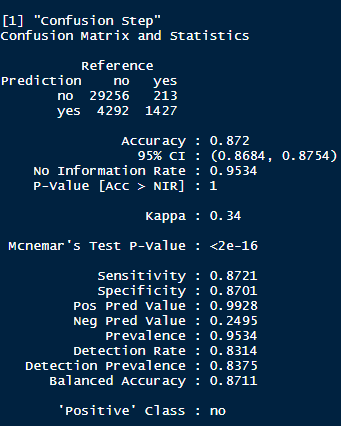
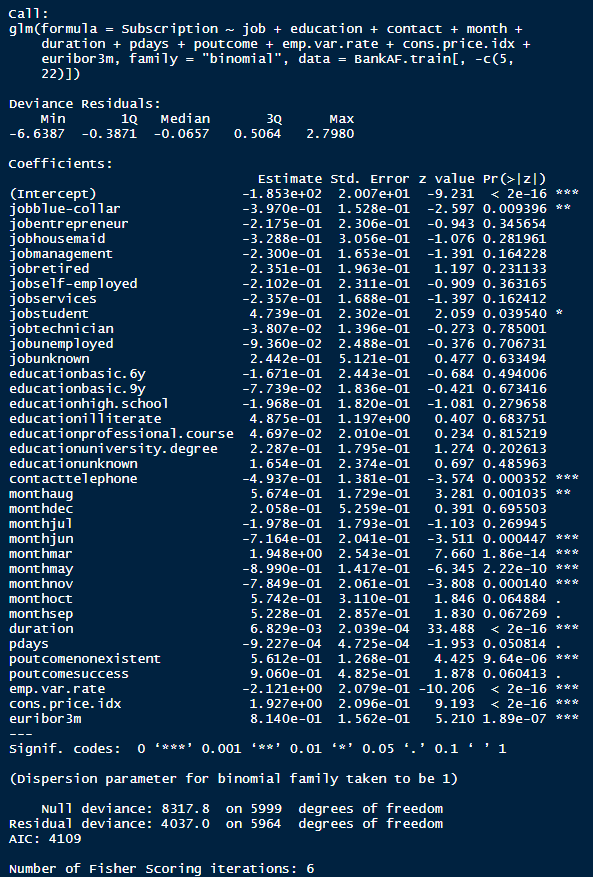
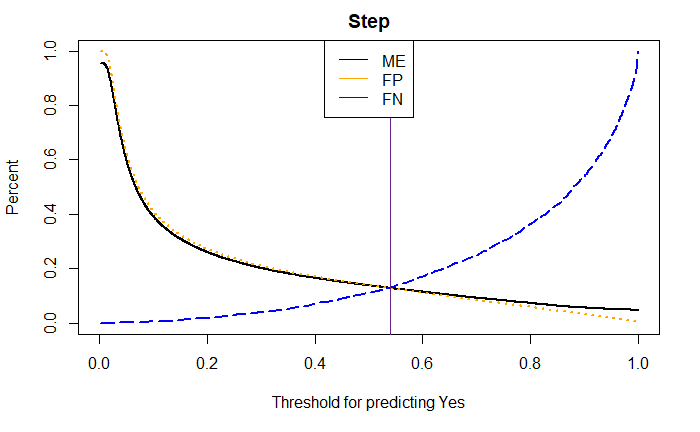




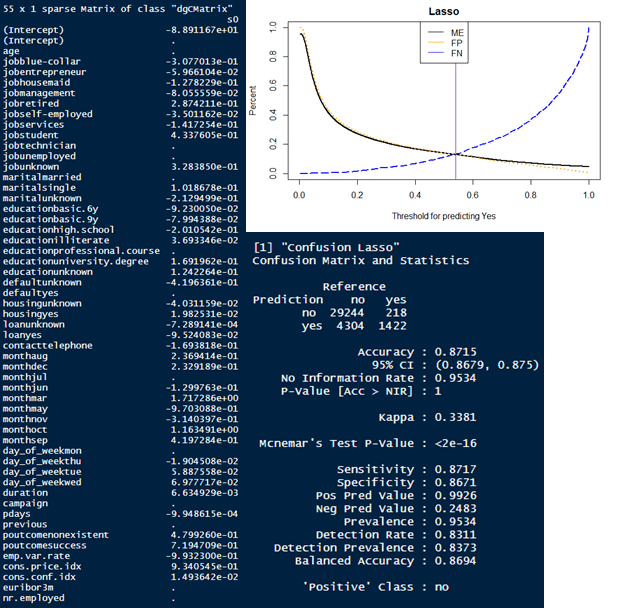
*Final Simple Log Model*



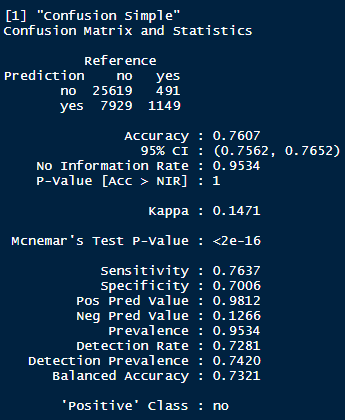
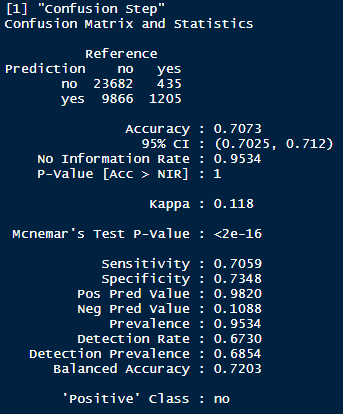
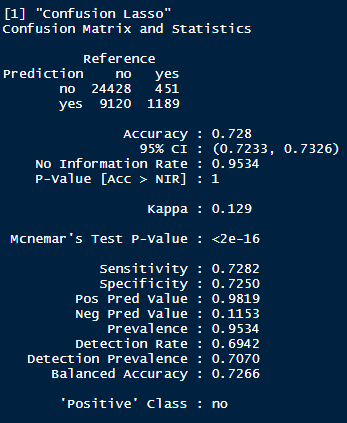
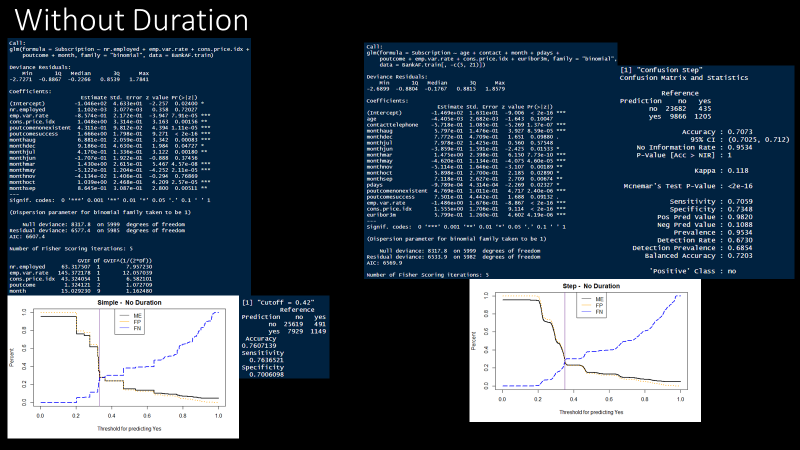
**Section 4 - Step and Lasso Model Selection**

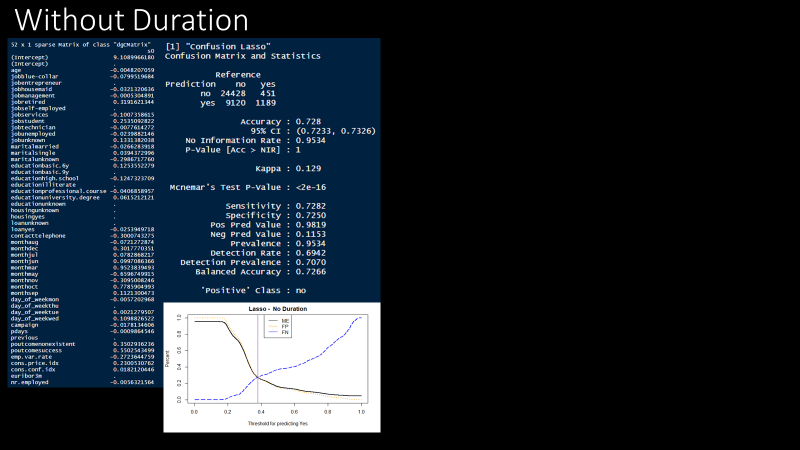


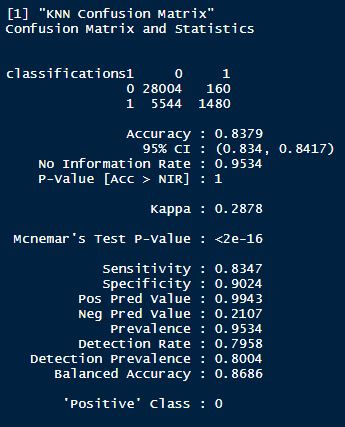
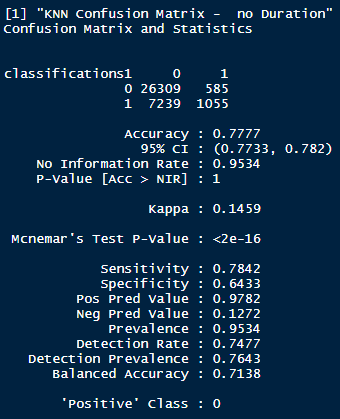
*Step Model*

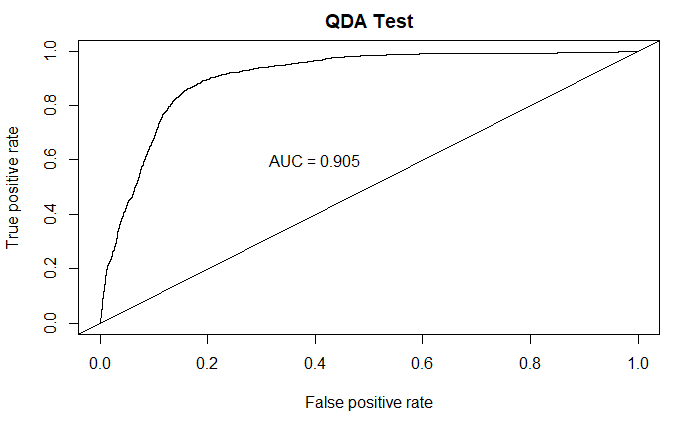


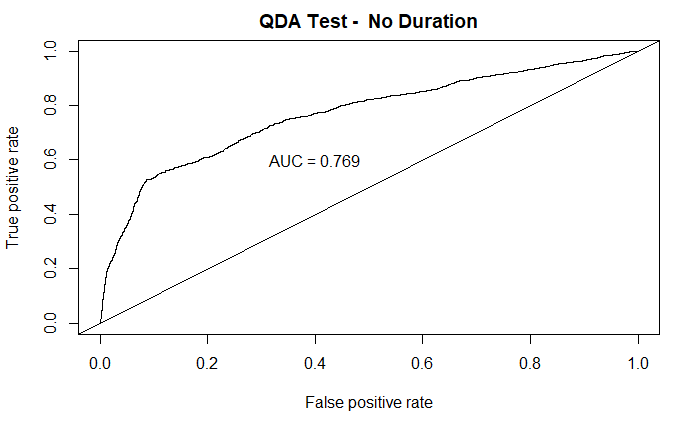
*Lasso Model*

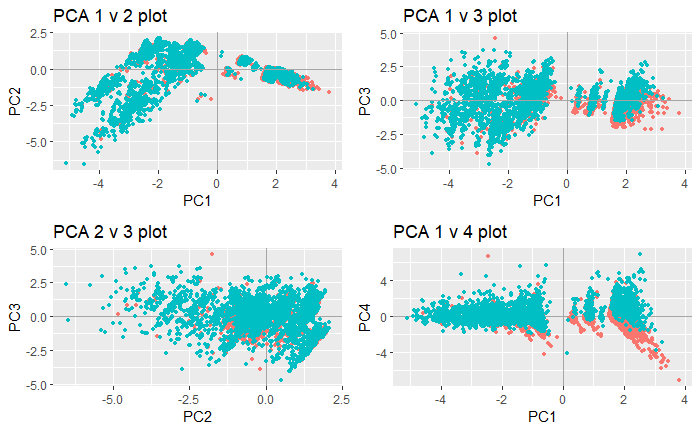
**Section 5 – Comparisons without duration**



**Section 6 – KNN Performance metrics**

**Section 7 – QDA Performance**



**Section 8 – PCA graphics**