**Bank marketing prediction**

DS6372: Project 2

By Nick Blair, Dhruba Dey, Bo Yun

## **Introduction**

Term deposit is a risk-free financial instrument and a sort of alternative to saving and money market accounts because it gives a higher interest rate with multiple range of durations such as 1-month, 3-month, 6-month, 1-year etc. Also, it has less restrictive feature or penalty if the term of the loan is violated in case of early withdrawal. Therefore, people can earn some extra money if they plan their spending in accordance with deposit term. Many people, however, do not know the benefits of the term deposit. As a result, the bank proactively markets to customers so that they can apply for the term deposit. However, the bank needs to know the people who would be interested in undertaking this financial instrument so that their marketing effort is better utilized.

Data science can provide the tools and skills to classify people as to who would be more likely to opt for the term deposit. A Portuguese bank undertook such a direct marketing campaign and collected data of people who applied or did not apply for term deposit, and some features related to people to whom they made the call. In this study, the goal is to predict the people who are likely to be opt for the term deposit based on the available profile of data.

### **Data Description**

The data for this project originated from a direct phone marketing campaign conducted by a Portuguese banking institution from May 2008 to November 2010. The data includes 41,188 samples of contacts that the institution called. The response variable (‘y’), is a binary indicator of if the specific contact would subscribe to a term deposit product available from the institution.

The dataset includes 20 input variables including both categorical and continuous variables. These input variables describe the client (age, job, marital status, education, etc.), attributes related to previous attempts at contact with the client (number of days passed since last contact, number of calls to this contact, etc.), and various social and economic context attributes (consumer price index, employment variation rate, etc.).

Through exploratory data analysis (detailed below), we found a selection of the total input variables to be practically significant. Education, default, job, and housing were found to have practical significance as they provide context regarding an individual client’s potential financial situation and thus if they are likely to enroll in new financial programs. Month of previous contact (month), contact type (contact), number of contacts during the campaign (campaign), number of previous contacts (previous), and duration of the contact calls (duration) are practical in order to inform business practices of the call campaign as these variables related to attributes of the marketing campaign calls themselves. Finally, consumer price index (cons.price.idx), employment variability rate (emp.var.rate), and consumer confidence index (cons.conf.idx) carried practical significance in order to provide context for the economic situation which influences consumer habits.

The dataset is complete but features an extremely unbalanced ratio of negative response variable entries compared to the relatively small number of positive response variable entries.

The data was sourced from the following published research paper:  [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

### **Exploratory Data Analysis**

**Overall correlation plot**

#### **Graphical user interface, application Description automatically generated**

Figure 1: Correlation plot

In the correlation plot above, nr.employed and euribor3m are very highly correlated with 94.2 percent. Nr. Employed will be dropped from the model due to its less significance judging from the y vs nr.employed box plot. When looking at the plots on age vs campaign, duration vs campaign, campaign vs cons.conf.idx, and campaign vs euribor3m, it is highly suspected that there are interactions that might exist and further be employed for building prediction models.

**Heatmap**

A Heatmap was created in order to study the relationship among continuous variables. The heatmap contributes validity to the correlation study done above as it shows a similar correlation seen among euribor3m vs. emp.var.rate and nr.employed vs. euribor3m. ***Chart

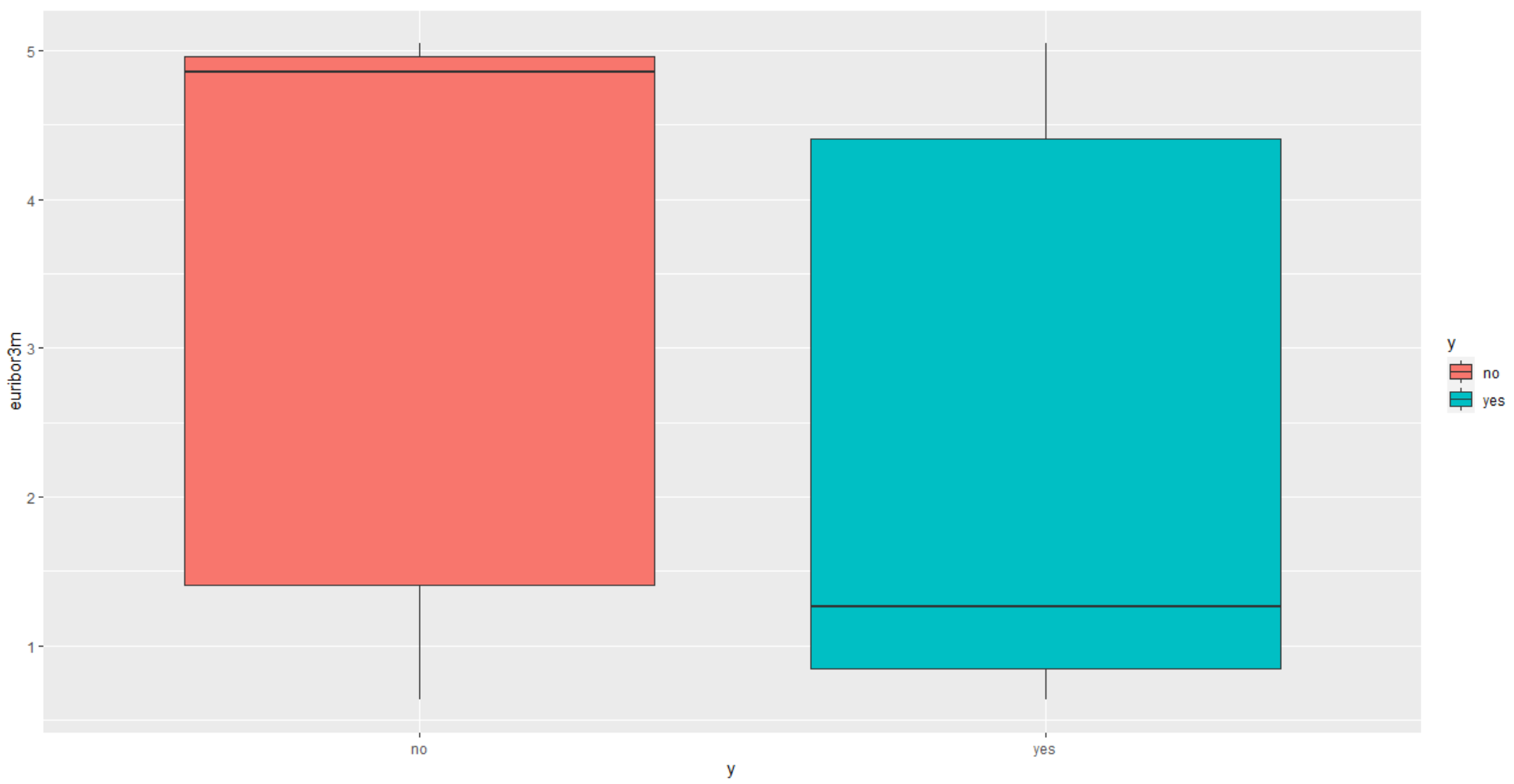
Description automatically generated***

Figure 2: Heatmap

**Boxplot**

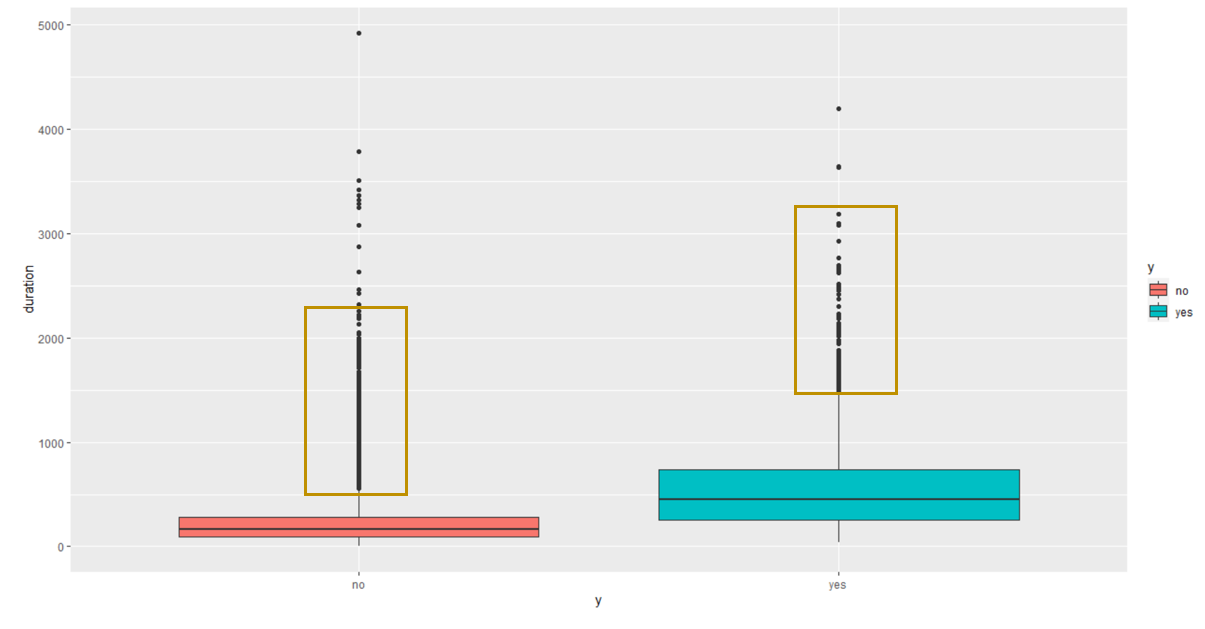
Boxplots were used to see whether a continuous feature can help in explaining the response. In our earlier analysis with the correlation plot and VIF have indicated that nr.employed (numbers of employees) is highly correlated with euribor3m (3-month euribor) and euribor3m is practically very significant because the interest is directly related to it. Therefore, we have dropped nr.employed from our further analysis and conducted the boxplot analysis with the following continuous variables.

*Euribor3m:*



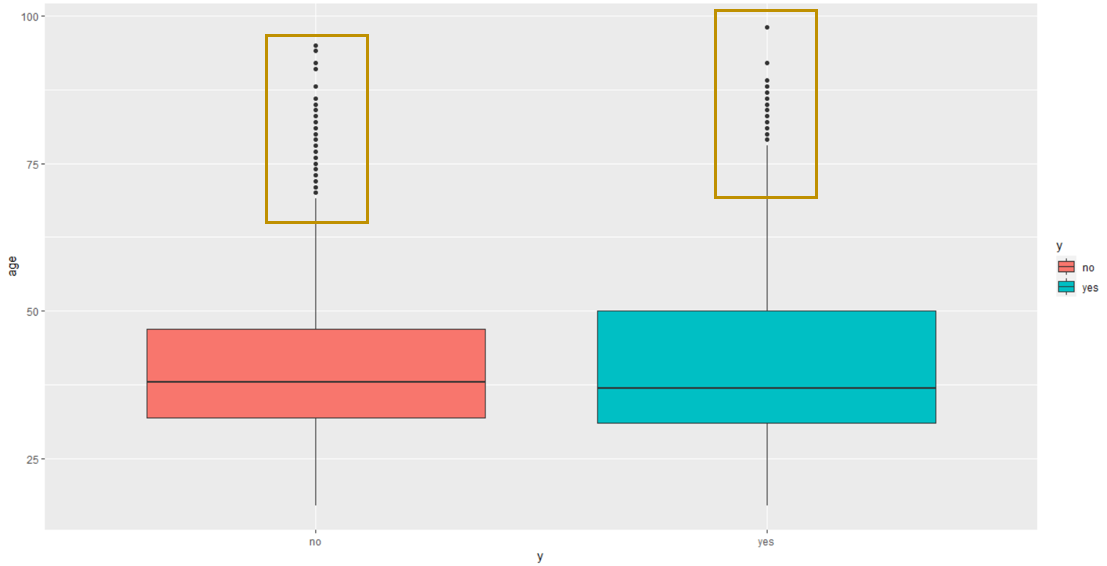
The means of 3-month Euribor for yes and no are far part and 3 quartile range of either yes or no does not cross the mean of one another, meaning this feature can significantly explain the response variable. It has almost no outliers

*Duration:*



Duration manifests that it can explain the response well given the separation of means, but it has some outliers which can adversely impact the model. But at this point, we would like to keep this feature

*Age:*



In case of Age, we do not see a significant difference between the means of yes and no. But we think that Age is a relevant information in regard to whether a person want to commit for a term deposit. It has also some outlier, but not too much

*Cons.price.idx:* (the plot is provided in the appendix)

Consumer price index shows the similar behavior as that of Euribor. The means of y and n are apart from one another and it has almost no outlier. In fact, the consumer price index which influences the inflation, thus the interest rate as a good significance in driving people’s decisions on the term deposit.

*Cons.conf.idx:* (the plot is provided in the appendix)

Consumer confidence index does not exhibit the convincing visual as the means are not separated enough. We think if the consumers are not confident enough, it would not commit to lock their funds, rather they would resort to liquid cash in their saving or money market accounts. That is the reason we would like to keep this feature at this point.

*Previous:* (the plot is provided in the appendix)

Numbers of previous calls made do not seem to be an important feature in explaining the response variable as the means are close enough, but interestingly their ranges are significantly different. We will let the feature selection or feature extraction decide whether this feature will go the final selection.

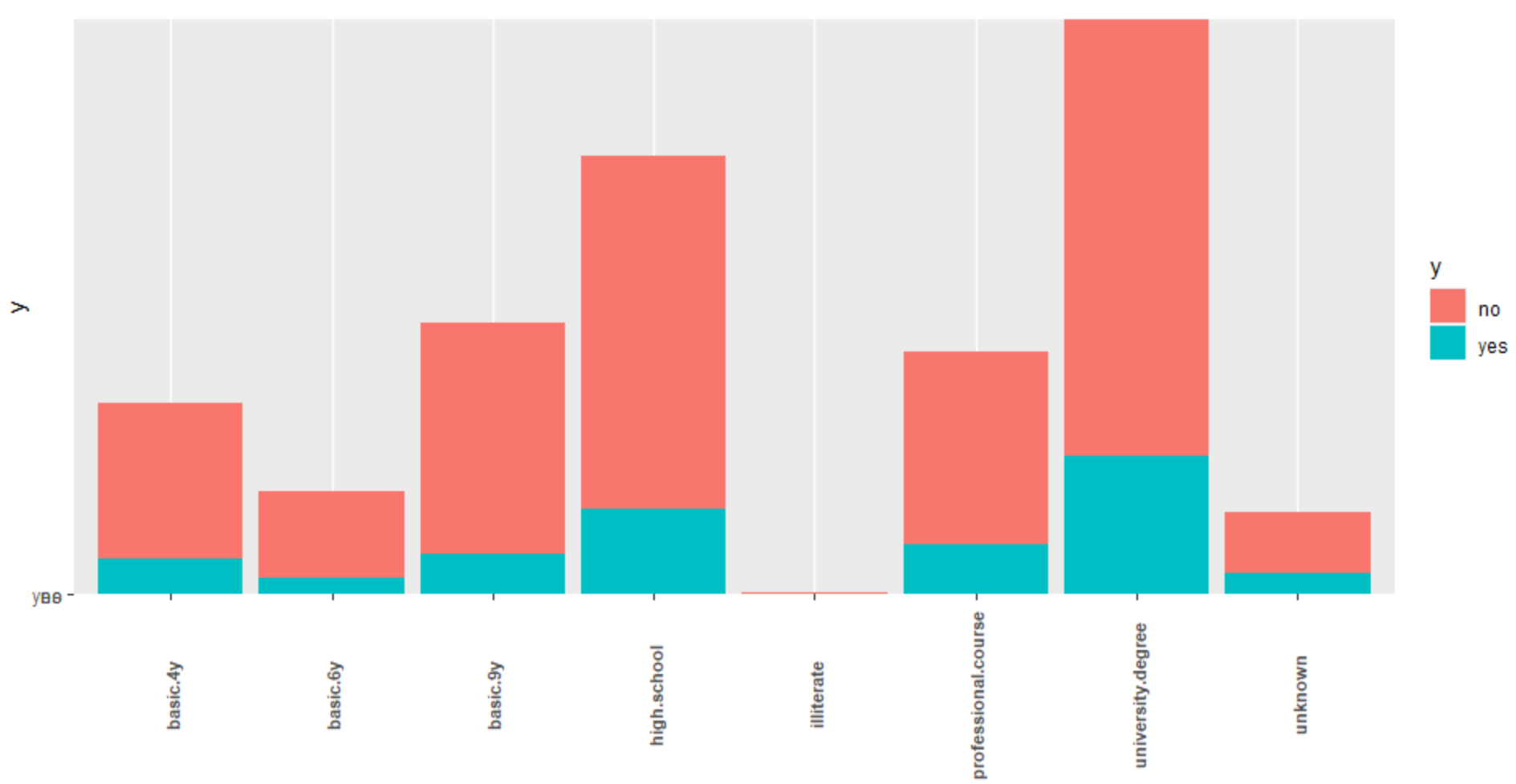
*Campaign:* (the plot is provided in the appendix)

“Numbers of contact performed during the campaign” presumably is an important feature. The 3 quartile ranges for yes and no are pretty small, and their means are apart some extent.

**Bar plot**

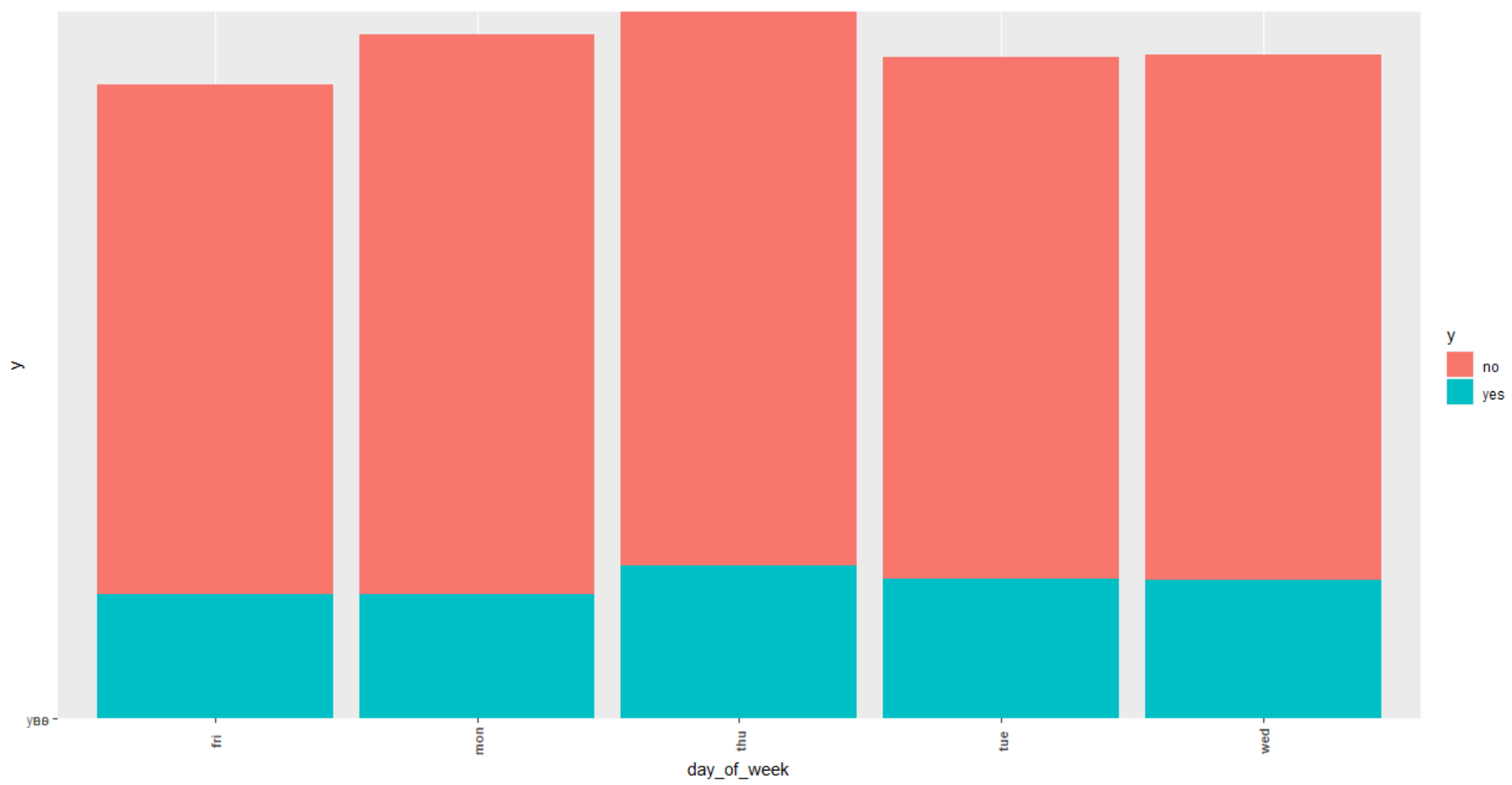
Bar plots were used to show whether the categorical variables can meaningfully contribute to explain the response. We have identified some anomalies in the date of the feature P-days. Also, we do not think it is an influencing factor to drive the decision for people. Therefore, we have decided to drop this feature at the outset and analyzed the following features.

*Education:*



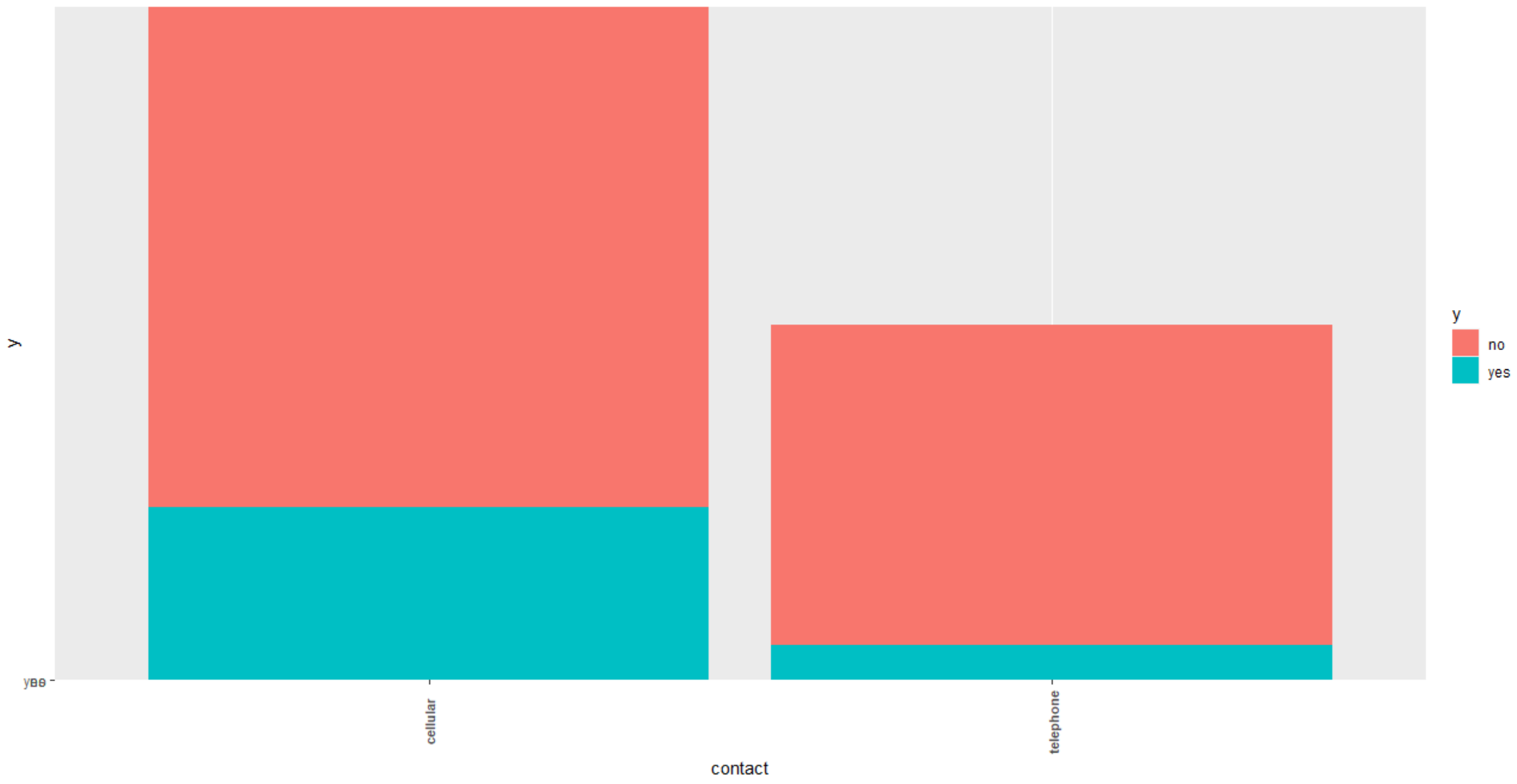
The proportion of yes and no varies from one level of education to another, meaning the success rate would vary depending on educational level of the people to whom the calls are placed.

*Day\_of\_week:*



Days of the week does not exhibit the variation of yes and no. Therefore, we think it would not add much value in explaining the response variable. Also, it is not practically significant in our assessment.

*Contact:*



It seems the callers have contacted people disproportionately via their cellular phone compared to their regular home or office phone. But the proportion of success is almost same regardless of method used to reach. So, we will let the feature selection decides the overall importance.

*Job:* (the plot is provided in the appendix)

Job type has impact on people’s decision on term deposit as it is manifested in the differentiated proportion of yes and no in each type of job.

*Month:* (the plot is provided in the appendix)

Interestingly, the month in which the calls were made had a meaningful impact on people’s decision on term deposit as yes and no were not evenly proportioned across different months.

*Housing:* (the plot is provided in the appendix)

People’s housing (whether they own a house or not) has some impact on the response variable. The proportion of y and n varies to some extent.

*Default:* (the plot is provided in the appendix)

People who defaulted their loans may be in a more delicate financial state or may not have enough deposable income for savings. Variation across this variable is observed.

*Marital:* (the plot is provided in the appendix)

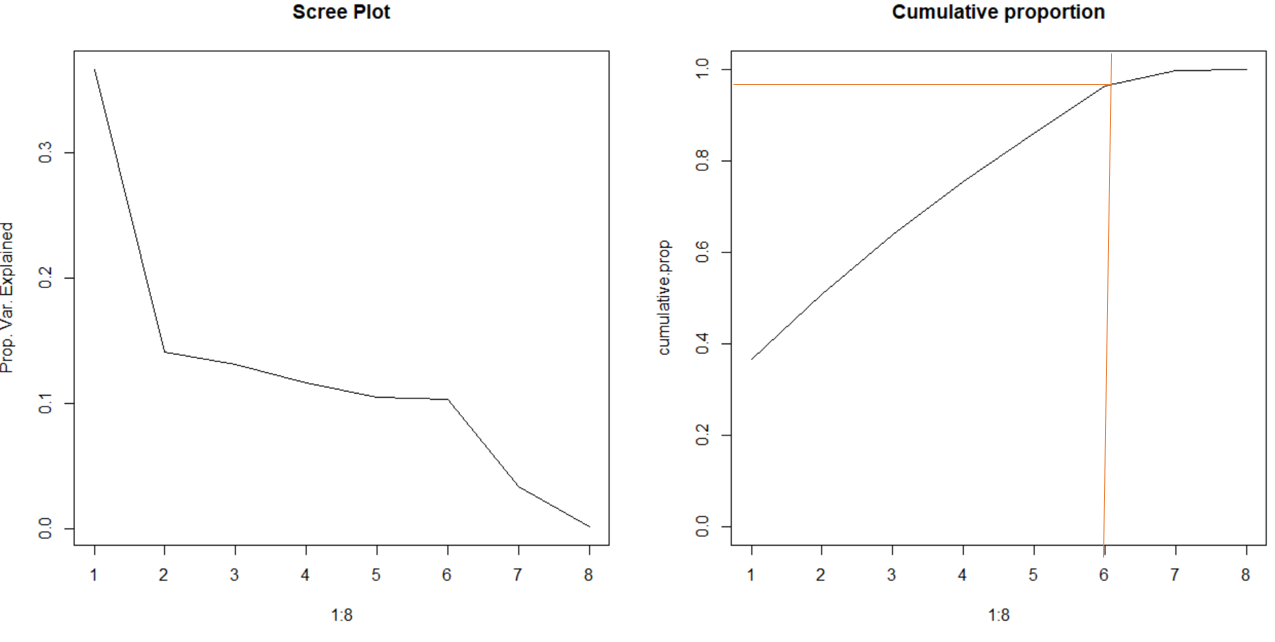
We thought that the married people will be inclined towards more savings instruments, but it seems they are using other financial instruments rather than term deposits (or at the very least are saving at similar rates as other people). The proportion of y and n across married, unmarried and divorced are almost same.

**Feature Selection**

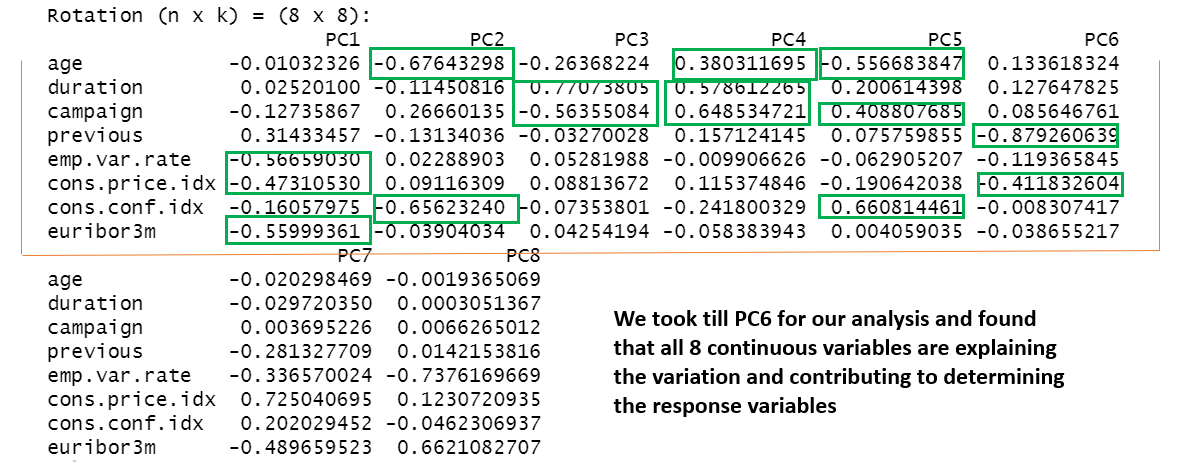
We have taken multiple approaches in feature selection to diversify our methodology. We used a combined approach of feature extraction (PCA) and feature selection (Stepwise, Backward Selection, LASSO) in combination with decisions based on practical significance. In our entire course of selection, we have remained cautious about the bias and variance trade off.

We have down-sampled the data for the feature selection analysis so that we do not need to worry about the unbalanced response rate and its consequence in our feature selection process. First, we attempted to find the important continuous variables using PCA. Then, we used the take-one-feature out from the logistic regression for categorical variables to see how a particular categorical variable added value in terms of deviance and AIC.

We have run PCA with a set of continuous variables and found the following results:



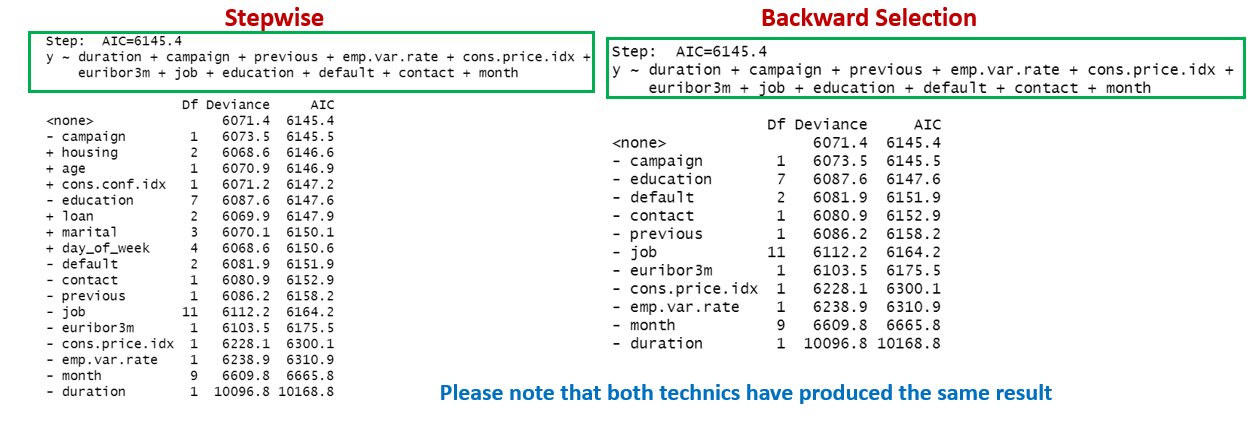
The cumulative plot that with PC6, we can achieve 95% variation. So, we will do our analysis based on PC1 through PC6



For categorical variables, we have run the logistic regression with all categorical features and taken one feature in each subsequent run to see whether the model is performing better or worse in terms of AIC and deviance. The results are shown below:

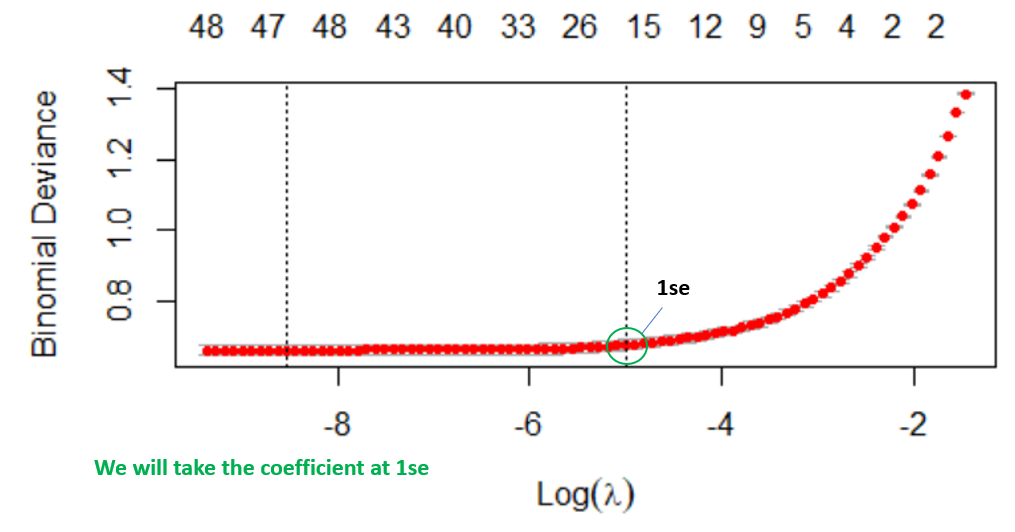


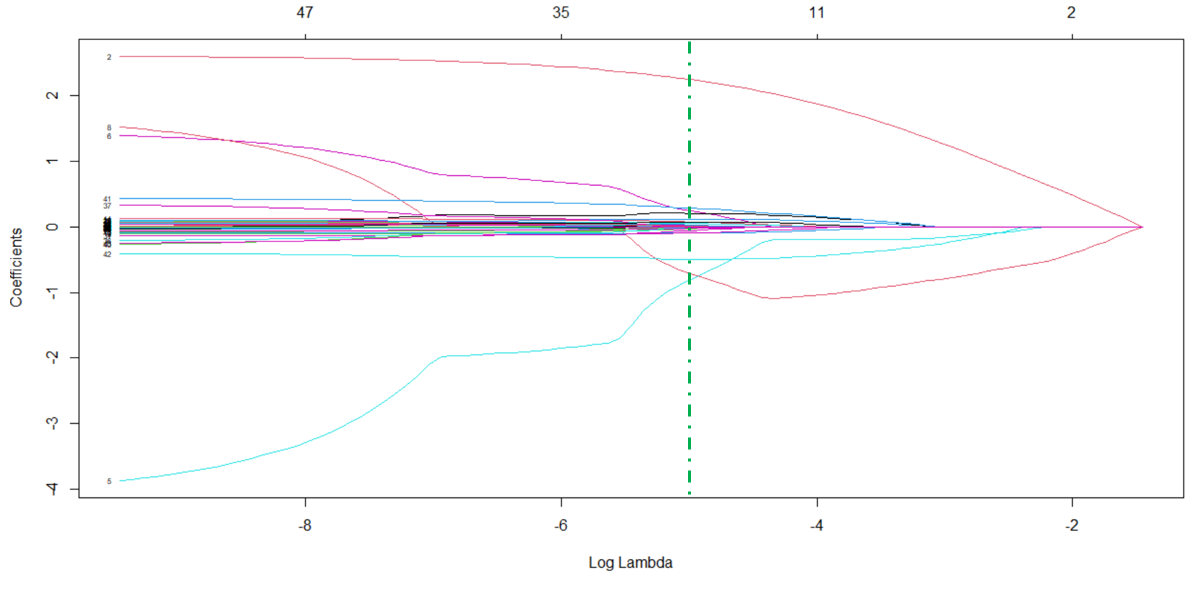
After conducting EDA, analysis of practical significance, PCA and take-one-out logistic regression analysis on categorical variables, we have a good picture about the features and the data; however, we would like to verify further using feature selection techniques. So, we have run stepwise and backward selection:

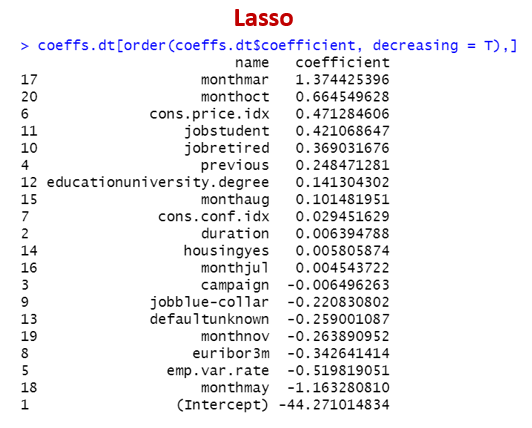


Note that both techniques have produced the same result

Next, we have run the LASSO and got the following results:





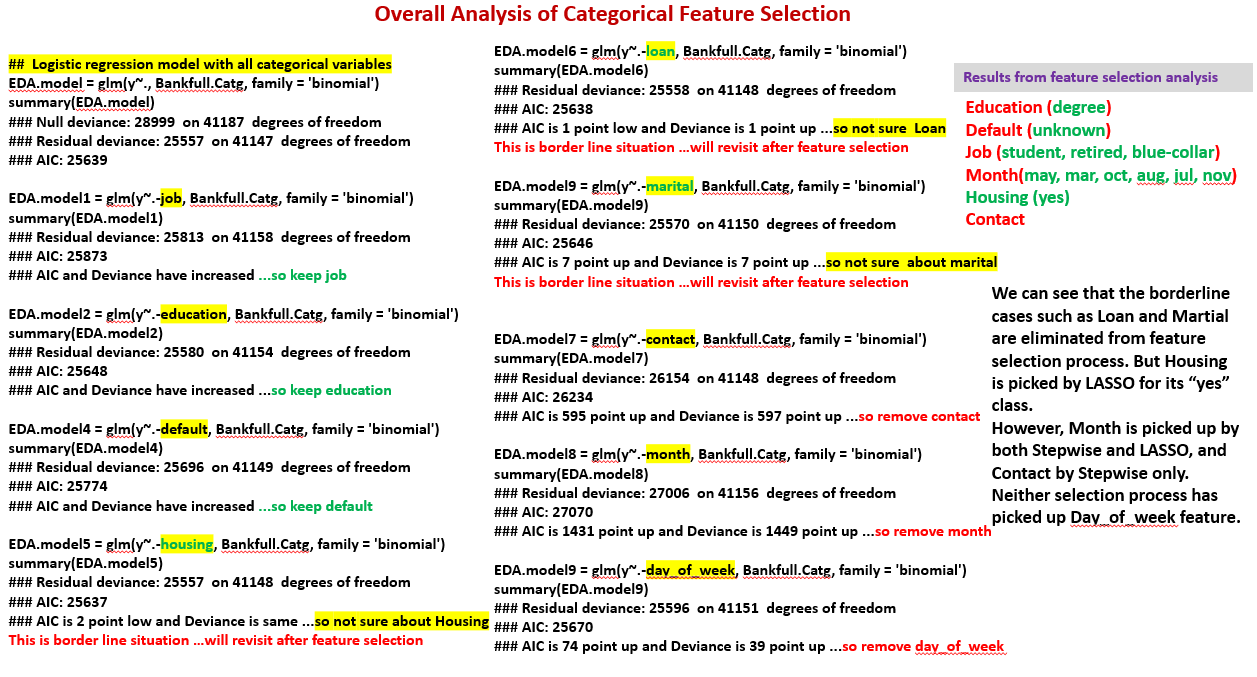


Now, we would like to compare the results of Stepwise and LASSO outputs (ignoring Backward Selection because it has generated the same result as that of Stepwise) in terms of **categorical features.** And we tie back that analysis with our earlier findings from take-one-out feature selection from logistic regression.



At this point, we can see that Education, Default, Job, Month, Housing and Contact are selected by the feature selection process. Though LASSO has not picked up all classes with the selected categorical features, we will not drop other not-significant classes from these features given that the loss of observations may cause errors in our prediction.

Now, we will try to validate these results with our earlier take-one-out approach for categorical variables.



Now, we are looking back to EDA and practical significance analysis to see whether it makes sense to have contact and month in the model. Looking at the bar-plots for month and contact, we have realized that the month has a significant differentiating proportion of yes and no across different months, and contact has shown the similar display but to a milder degree

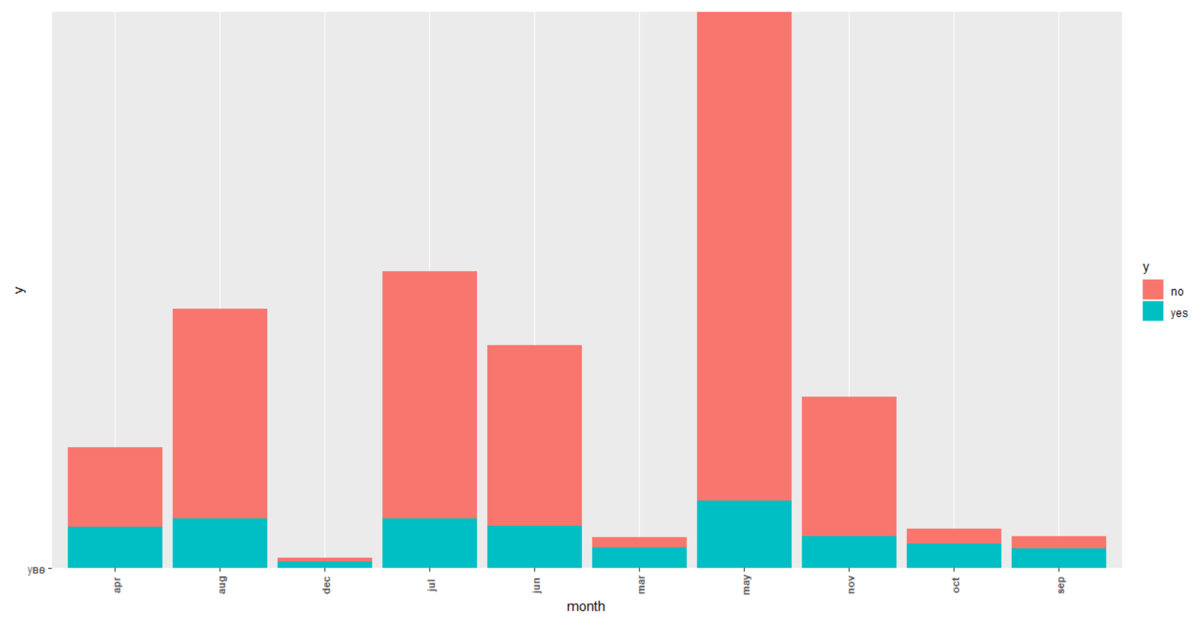
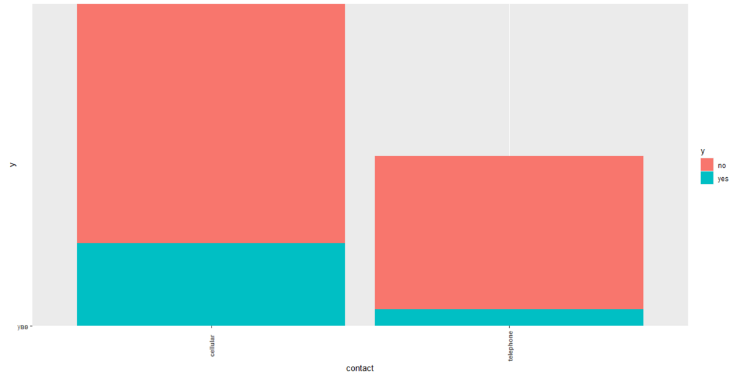


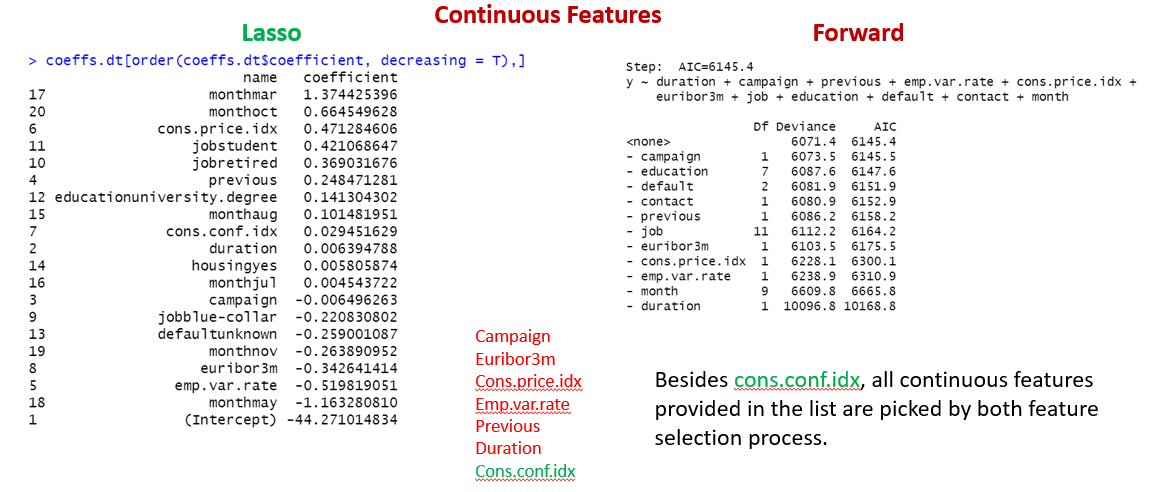
Figure 3: Bar plot of month

Figure 4: Bar plot of contact

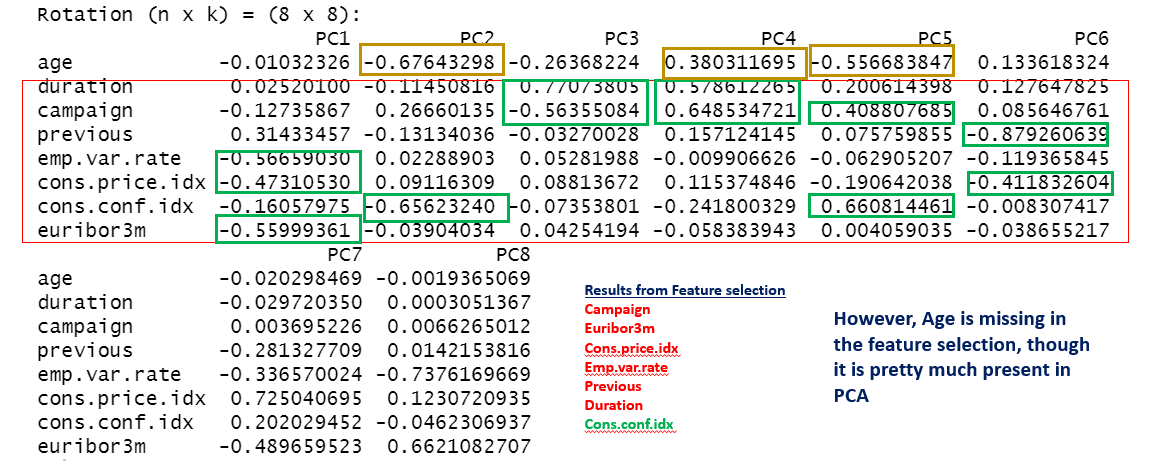
Finally, we have the following categorical features for our model building.

* Education
* Default
* Job
* Month
* Housing
* Contact

Now, we would like to compare the results of Stepwise and LASSO outputs (ignoring Backward Selection because it has generated the same result as that of Stepwise) in terms of **continuous features.** And we tie back that analysis with our earlier findings from PCA.



Now we go back and compare this result with PCA.



In addition, we think Age is practically significant. But there is lack of proof from EDA (boxplot) as shown above. So, we want to see whether adding this feature will harm our model. We ran logistic regression with the categorical features selected earlier and continuous features selected by feature selection in order to capture AIC and deviance. Then, we compare AIC and deviance with the same model plus age.



There is no significant difference in error in having age or not. But strong indicators from PCA and our notion of practical significant have prompted us to include Age in the mix of continuous features. However, we must pay attention to variance because adding unnecessary features may result in overfitting the model. We will deal with the bias-variance issue at the time of model cross validation.

Here are the final set of features for our model selection.

* Campaign
* Euribor3m
* Cons.price.idx
* Emp.var.rate
* Previous
* Duration
* Cons.conf.idx
* age
* Education
* Default
* Job
* Month
* Housing
* Contact

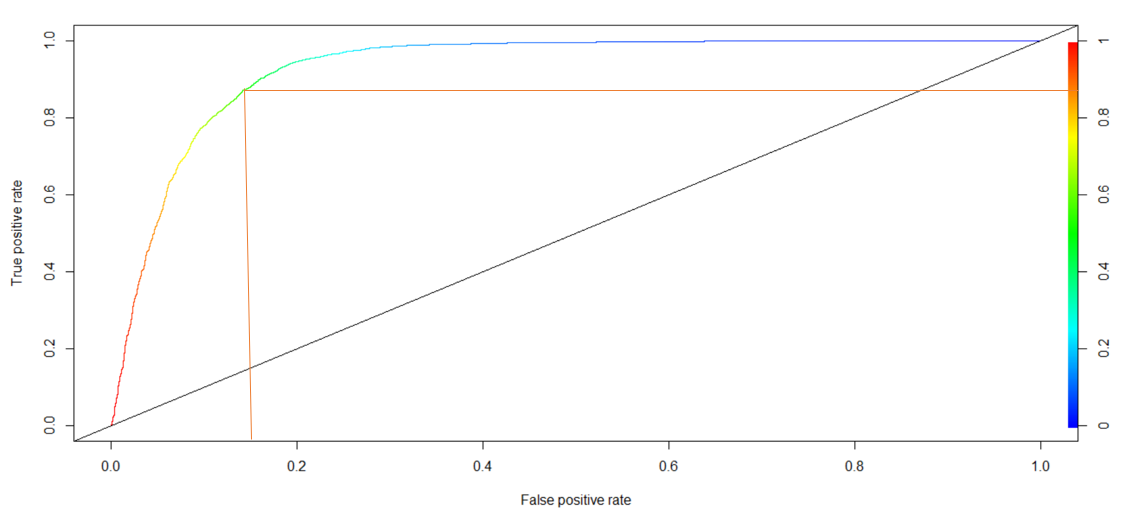
**Logistic Regression:**

We have the logistic regression using two approaches.

1. Created a large, balanced sample with equal numbers of yes and no; train a model with the balanced sample and then use the model to validate the entire population of the data.
2. Adopted the traditional approach of 70-30 split for training and test.

Approach 1 (Train – Balanced and Test – Entire Population)

We have built the logistic regression model based on balanced sample and achieved 86.58% model accuracy on balance sample using cutoff at 0.5 in the first cut. Then, we optimize the cutoff value through ROC curve as shown below.



As expected, the cutoff of 0.5 is the most optimized for the balanced sample. Using a cutoff value of 0.5, we have cross validated the entire population of data and achieved 86.15% accuracy. This model looks good.

Approach 2 (Split Sample 70-30 for Train and Test, Cross-validate on Test)

We have split the data 70-30 to create Training and Test data and used Training data for model building. Off course, Training data is unbalanced, and we have planned to optimize it using the ROC curve.

We have built the model and achieved 90.58% accuracy at cutoff 0.5 on the training model, which is good. But we know that the model can be further optimized as the training data set is not balanced. We have generated a ROC curve and found the optimum cutoff of 0.4. Then, after updating the model with a cutoff of 0.4, the accuracy is improved to 90.73%.

After that, we have cross validated the model with the Test data and found the accuracy to be 91.03%, which exceeds the accuracy of the model built on the balanced data.

Therefore, we have selected the regression model generated via Approach 2.

### **Assumptions:**

1. Residual plots as follows shows that residuals are normally distributed.

Graphical user interface, diagram

Description automatically generated

1. Cook’s D - There is no significant outlier or leverage point in this model.

Chart

Description automatically generated

1. Lack of fit test - Using deviance test, the interpretable model will be tested to see if it fits the data and if it indicates that we haven’t missed big sources of data variation in our model.

1-pchisq(residual deviance,df) = 1-pchisq(6067.3,9240) = 1. Since the result comes out to 1, it means it is a very good fit.

We have conducted thorough analysis of each feature from multiple angles and included those which makes sense from statistical and practical standpoints. In addition, we have investigated the study carefully and thought about the design of the direct marketing. Most of the features are relevant barring a couple, mostly because of sparse data. Moreover, we have not seen any significant evidence against the features that we have selected. ROC and Cross-validation are used as the measures of fit.

**Interpretation model:**

The following effects plot was populated and each variable was analyzed to see how the probability of the response variable (probability of getting the client a term deposit) differs with its own different levels.

Diagram

Description automatically generated

**General behavior of variable probabilities:**

1. The higher the variable, the higher the probability of the client subscribing to a term deposit.
   * Cons.price.indx (Showed the most strong and clear distinction among its different levels)
   * Duration (Showed clear distinction among its different levels)
   * Euribor3m (Showed clear distinction among its different levels)
   * Previous
   * Duration
   * Cons.conf.idx
   * Age
2. The lower the variable, the higher the probability of the client subscribing a term deposit.
   * Campaign
   * Emp.var.rate
3. Other variables
   * Job
   * Month
   * Housing
   * Default
   * Contact (Showed a clear distinction. Making contacts through cellphone(Prob=0.576) vs. telephone(Prob=0.448)

**Odds ratio interpretation on key variables**

* Cons.price.idx
  + The odds of a person subscribing for a deposit term is 8.78 times higher with 1 point consumer price index lower.

* Emp.var.rate
  + The odds of a person subscribing for a deposit term is 0.132 times lower with 1 percent lower employment variation rate.

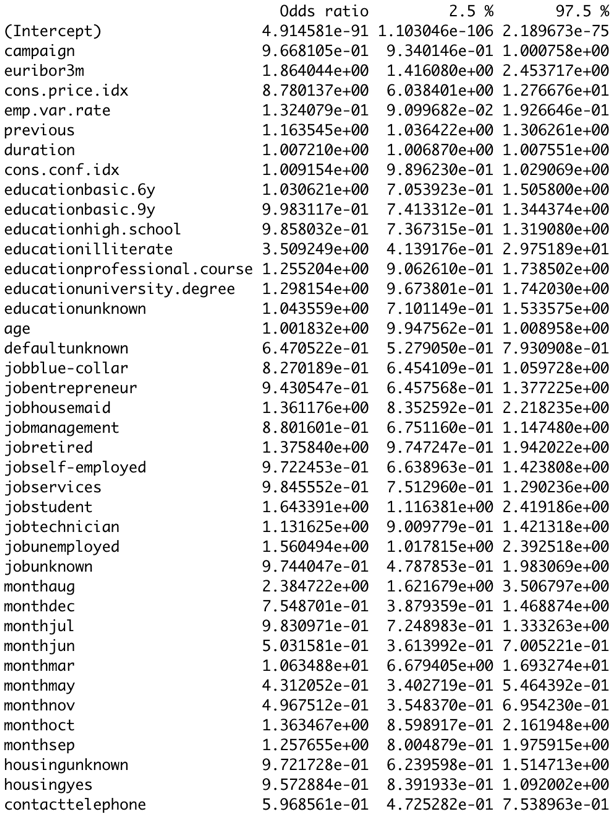
* Euribor3m
  + The odds of a person subscribing for a deposit term is 1.84 times higher with 1 percent lower euribor interest rate.

* Campaign
  + The odds of a person subscribing for a deposit term is 0.9668 times lower than a person with 1 contact less made.

* Contact
  + The odds ratio for subscribing a deposit term for telephone to cellphone is 0.597 after accounting for all the other variables fixed.

**Confidence intervals**

Please refer to the 95% confidence interval chart in the appendix.



Age : A 95% confidence interval for the odds of a person subscribing a deposit term is (0.99475, 1.00895)  higher than a person 1 year younger holding all other variables fixed.

Duration : A 95% confidence interval for the odds of a person subscribing a deposit term is (1.0068, 1.0075)  higher than a person with one second less contacted while 1 year holding all other variables fixed.

Prediction Models

##### **First prediction model**

Final variables were chosen from EDA and feature selection(Age and euribor3m were removed via repeated cycle of trial and error checking significance of individual variables after each model is established).

*Variable removed through trial and error : Age, euribor3m*

*CV Misclassification rate : 12.8%*

*AUC : 0.937*

Table

Description automatically generated

Figure 4: Coefficients

Chart

Description automatically generated

Figure 5: ROC Curve

##### **Second prediction model**

*Interaction terms used: duration:campaign, cons.conf.idx:campaign*

*Variable removed through trial and error : Campaign, Age, housing, Euribor3m*

*CV Misclassification rate : 13.1%*

*AUC : 0.937*

A picture containing text, newspaper, receipt

Description automatically generated

Figure 6: Coefficients

Chart

Description automatically generated

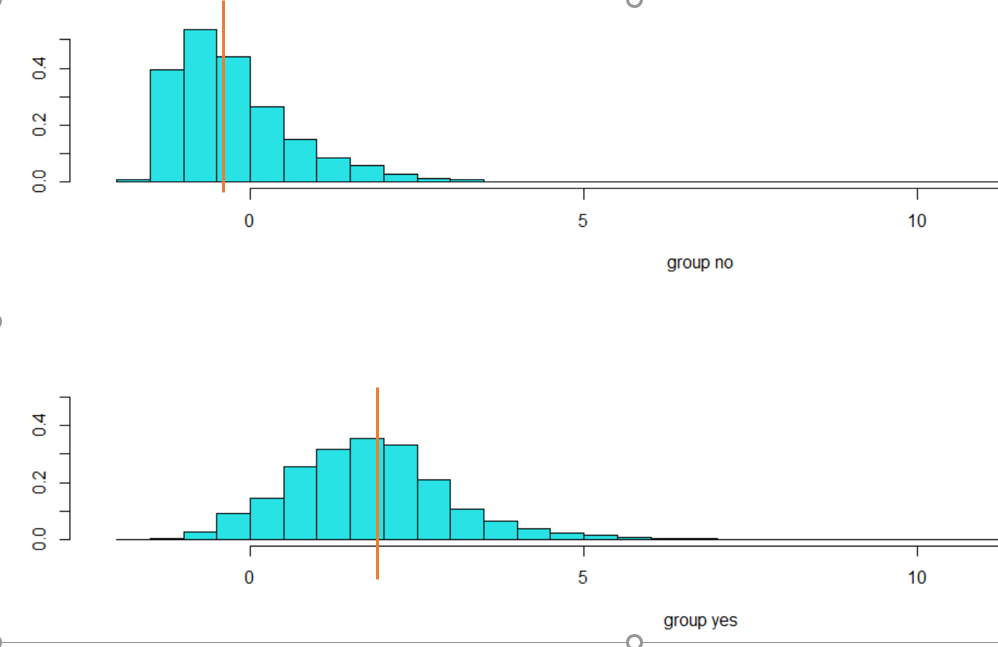
Figure 7: ROC Curve

**Discriminate Analysis**

In the case of LDA, we have not used balanced data as the posterior probability plays a role in classification. The data set is split to 70-30 for Train and Test. We have built the model on Train and then cross validated against Test.



The decision rule is based on the Linear Score Function, a function of the population means for each of our sub- populations, as well as the pooled variance-covariance matrix. The plot below shows how the mean of two populations are separated by the discriminate analysis.



We have cross validated against the Test data and found the accuracy 97.32%, which is pretty good. Since the linear model have given us such a good result and it is a linear separation, we feel confident with LDA. However, Levin’s test show that there is a bit violation of homogeneity. Therefore, we have run Quadratic Discriminate Analysis (QDA) and found the accuracy stands at 96.37%, which is less than 97.32% achieved by LDA. Moreover, with LDA, we have a lesser chance of overfitting. Therefore, we would opt for LDA when conducting Discriminate Analysis.

The QDA summary statistics are provided in the appendix.

**Non-Parametric Model:**

Given that we utilize both categorical and continuous data in our models, KNN was not suitable as a non-parametric approach. This left the decision between decision trees or a Random Forest algorithm.

The Random Forest algorithm utilizes a group of randomly created decision trees with each node in the tree working on a subset of features to calculate a decision. The random forest algorithm then combines the decisions of the individual trees for a final output.

The strengths of Random Forest are that it can determine the importance of each variable in the dataset, it predicts continuous or categorical responses, and it handles missing data well. The weaknesses of Random Forest is that while it is conceptually easy to understand, the inner workings of the model are not. Further, it does not predict outside the range of the data, and it can overfit with noisy data.

Ultimately, the Random Forest algorithm was chosen over a standard decision tree approach because random forest is suitable for a large dataset and trades some amount of interpretability for a stronger predictive model.

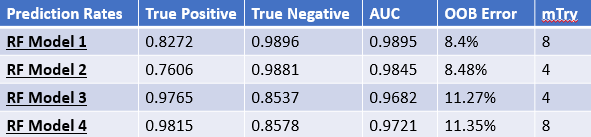
Several Random Forest models were created to calculate predictions.  With Random Forest, the training data is split into multiple regions or trees to create a mean prediction.  The predictions are then aggregated into a single prediction. The overview of our 4 random forest models are:

1. Model built on full dataset (no down-sampling) with all variables included in the model
2. Model built on full dataset (no down-sampling) with a selection of variables determined by backward feature selection included in the model
3. Model built on a down-sampled dataset to balance responses with variables selected from a combination of methods during the EDA process.
4. Model built on down-sampled dataset to balance responses with all variables included in the mode.

*Model Tuning Parameters*

After running the tuneRF function in R and observing the Out of Bag errors, values of 4 or 8 were chosen for the models (see table below) to avoid overfitting.  The OOB Error ranged from 8.4% (Model 1) to 11.35% (Model 4). The number of trees variable was set to the default value.

Table 1: Comparison of various metrics of the 4 Random Forest models. Model 1 (full data, all variables), Model 2 (full data, backward feature selection), Model 3 (down-sampled data, EDA variable selection), Model 4 (down-sampled data, all variables).



**Comparing Competing Models:**

|  |  |  |
| --- | --- | --- |
| **Predictive Models** | **Classification Error Rate** | **AUC** |
| *Logistic Regression (model 1)* | *8.97%* | *0.937* |
| Logistic Regression (model 2) | 13.1% | 0.937 |
| LDA | 3.63% | 0.976 |
| Random Forest (model 3) | 12.83% | 0.968 |

**Conclusion:**

Based on the classification error rate and AUC, LDA has had the lowest classification error rate and highest AUC. On the other hand, the second Logistic Regression model had the highest error rate and tied for lowest AUC. Generally, parametric models are more reliable than non-parametric models if all assumptions are met. This holds true in our case as well.

In addition to reliability, however, we are also seeking to maximize interpretability to improve ease of use by the client. With this in mind, the first Logistic Regression model is the best compromise between effectiveness (low error rate, high AUC) and simplicity for the sake of interpretation. Though it is a classification model, it optimizes the distance between the mean

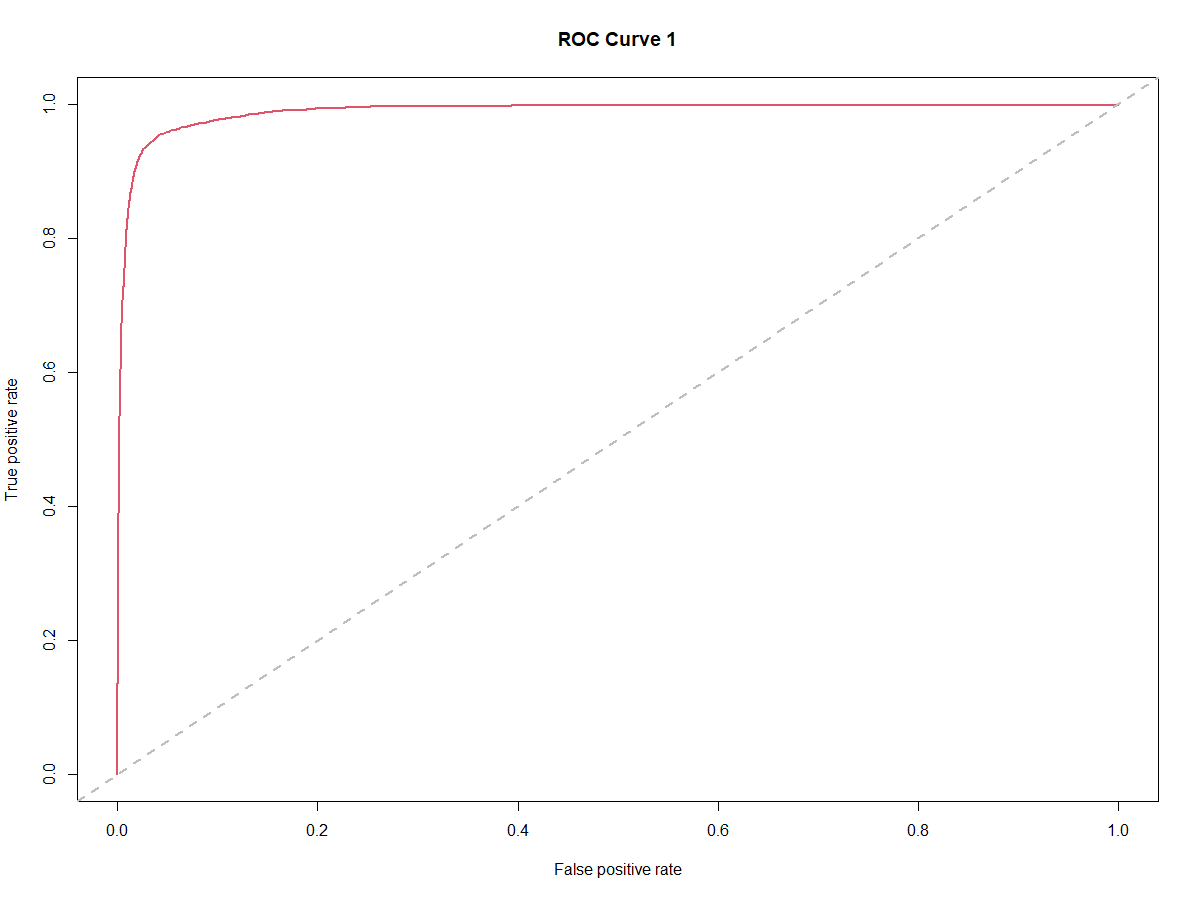
and the variance. When comparing logistic regression and LDA, we lean towards logistic regression given the higher predictive power.

Our conclusion is that the first logistic regression model is that best option for this scenario.

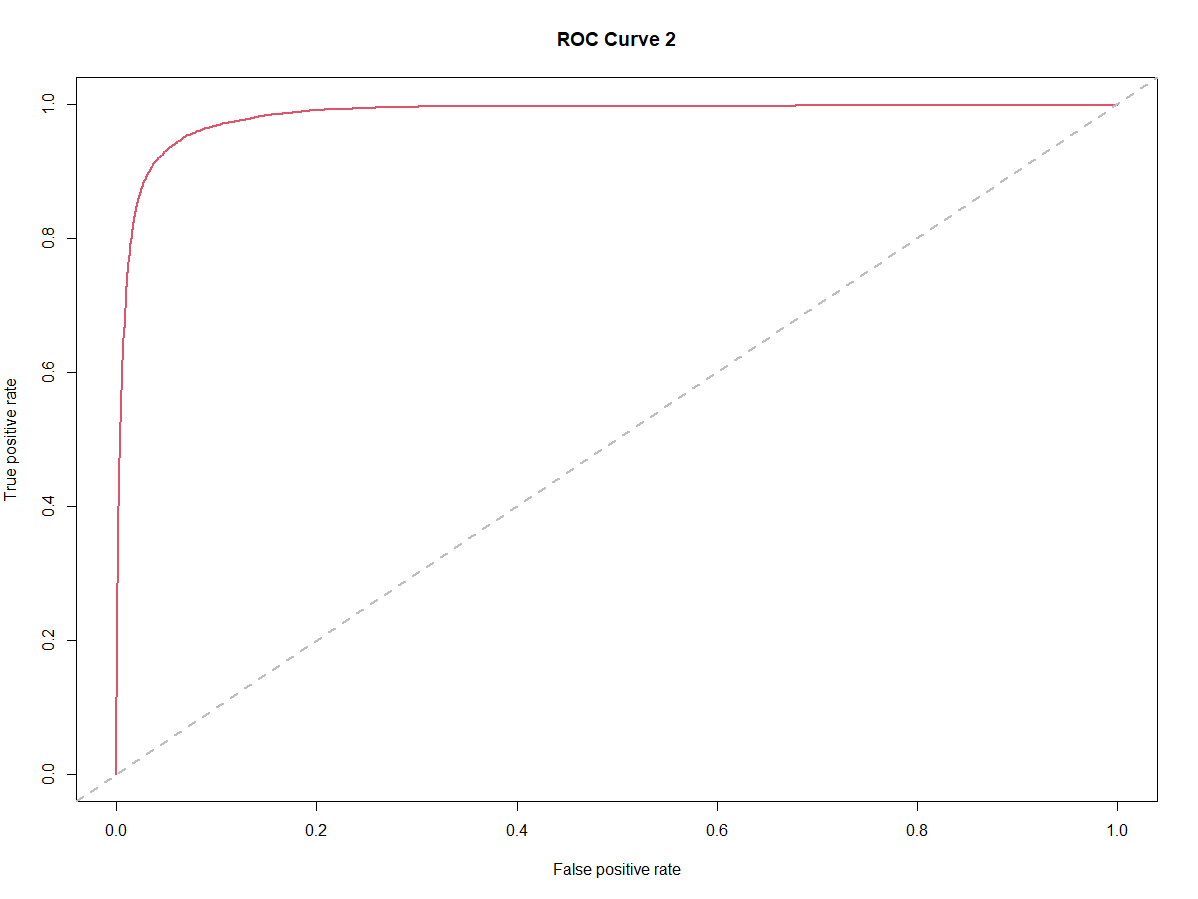
## Appendix

ROC Curves for Random Forest Models 1-4

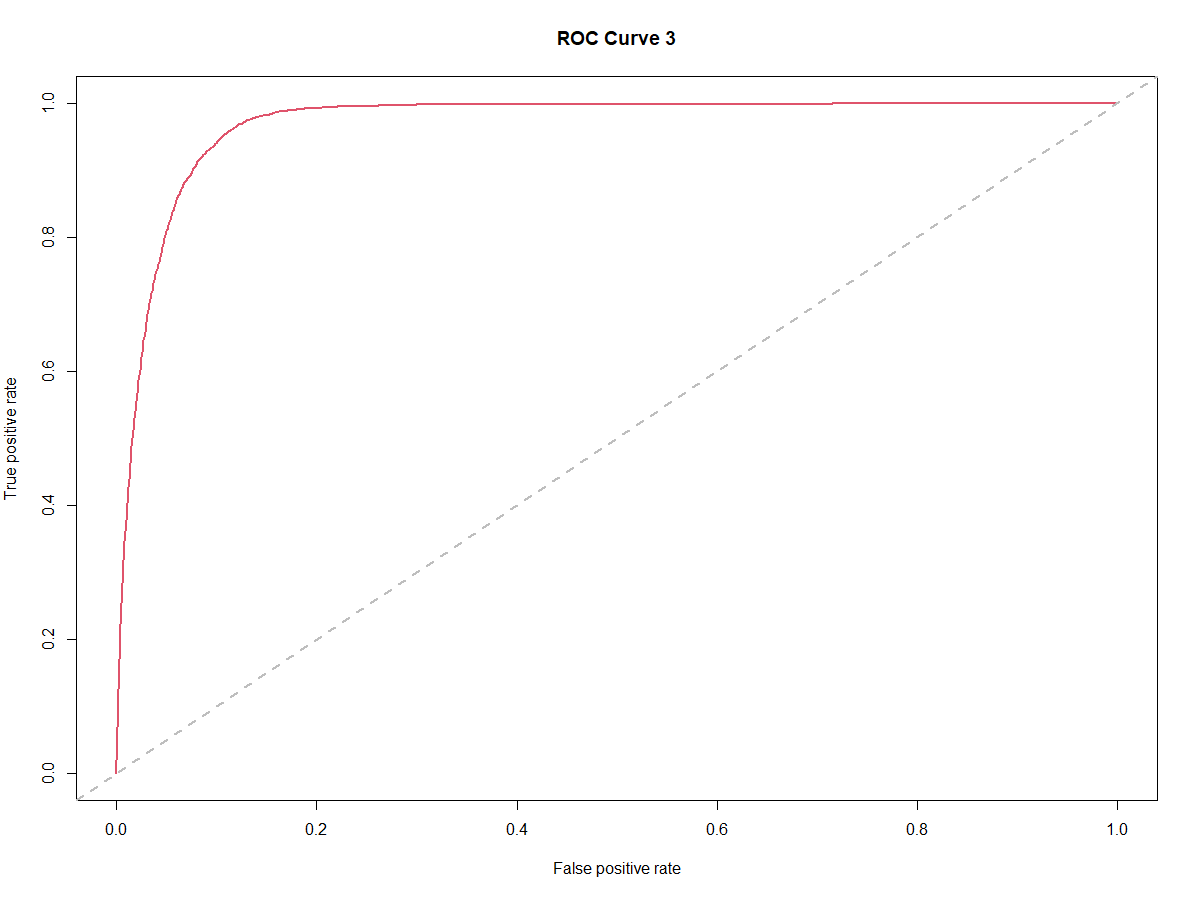
Curve 1:



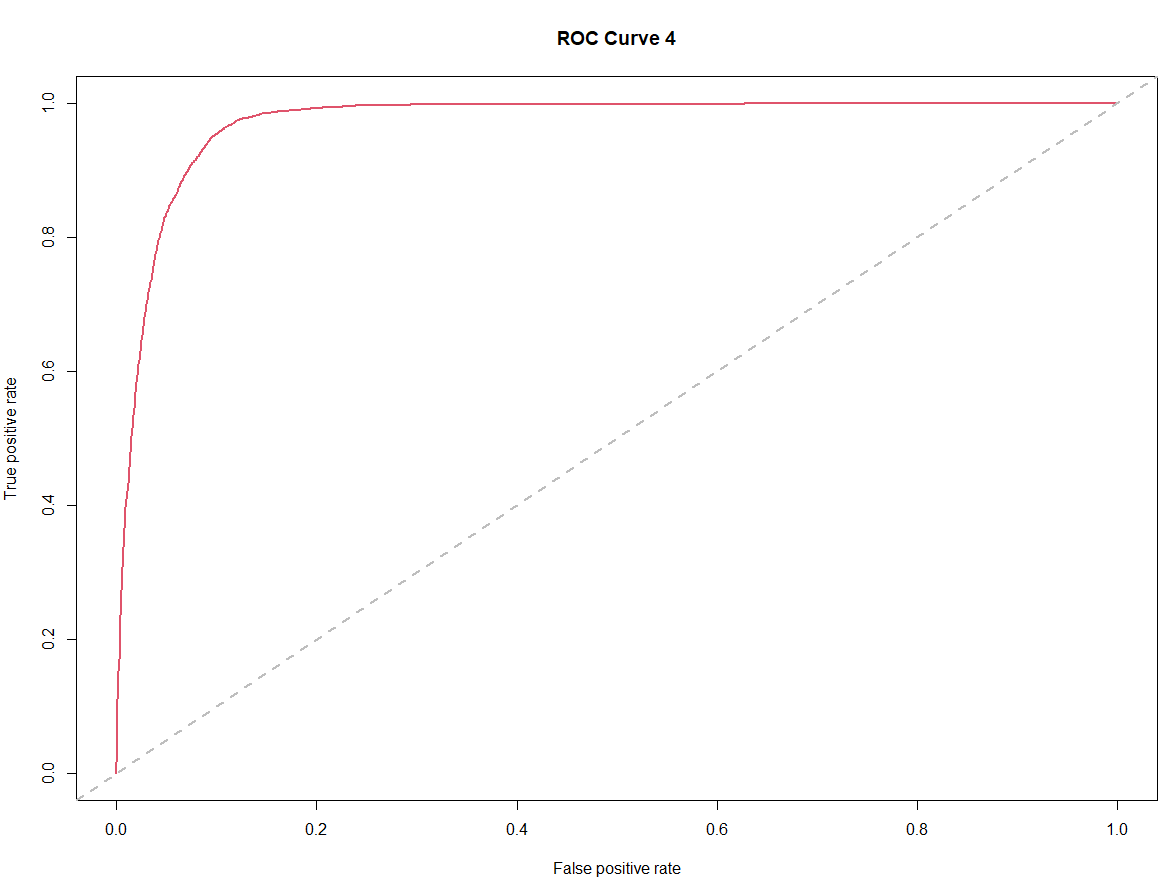
Curve 2:



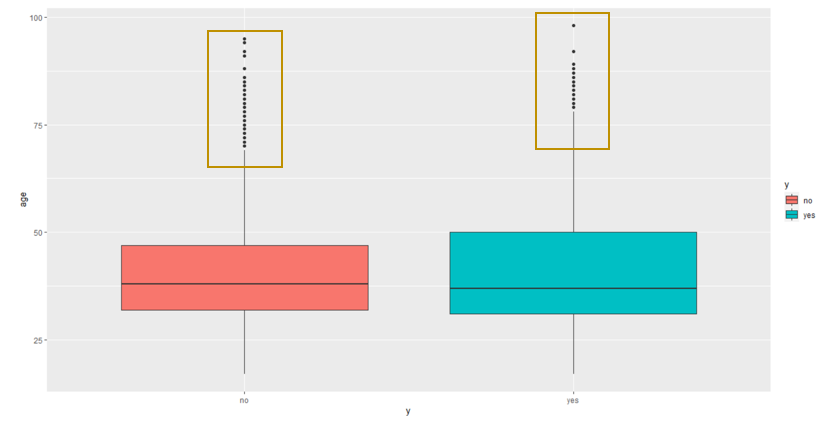
Curve 3:

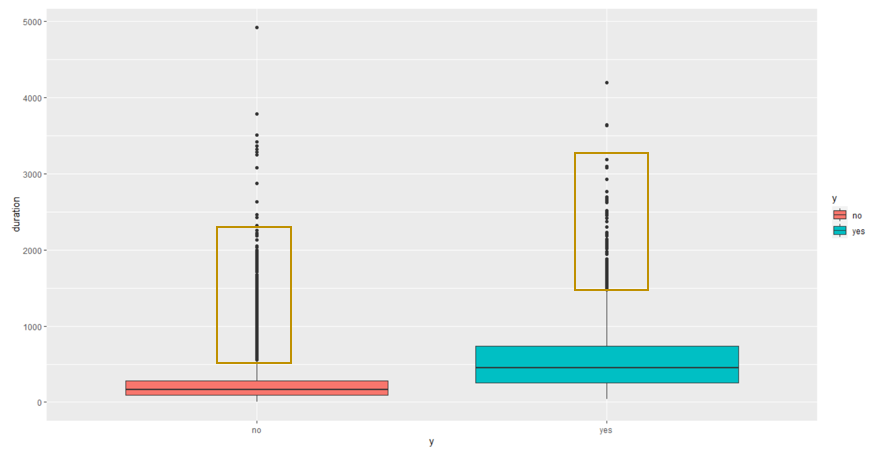


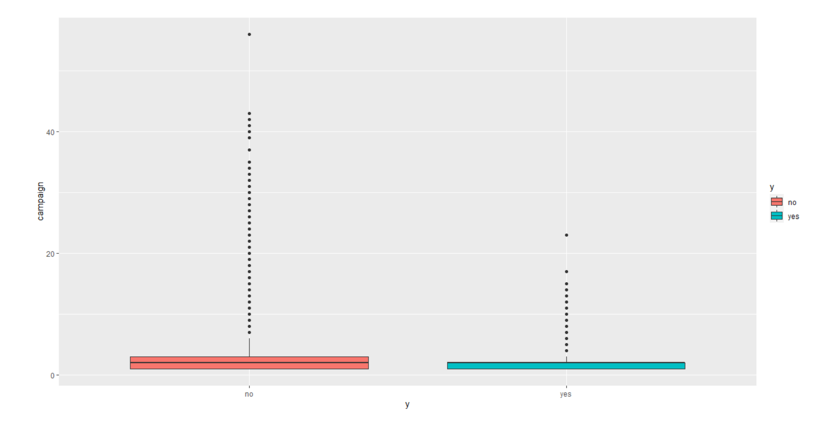
Curve 4:

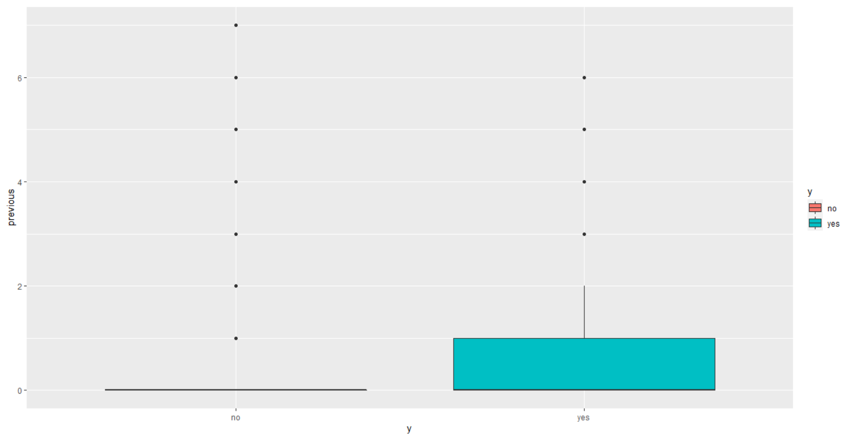


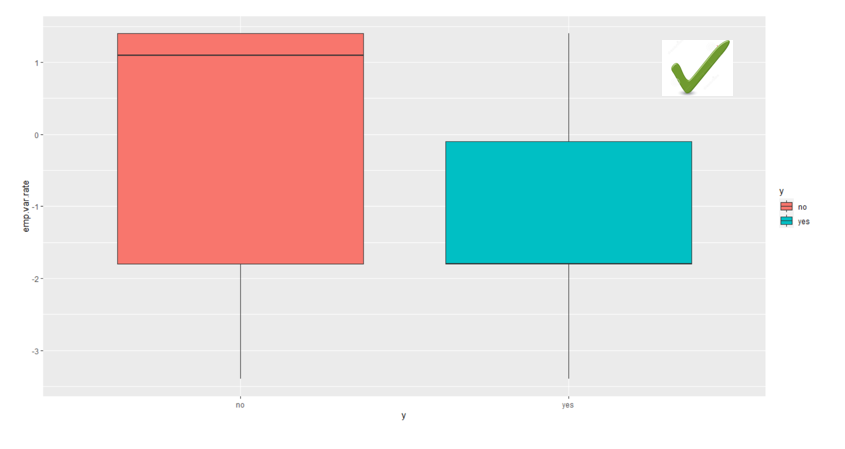
**Boxplot:**

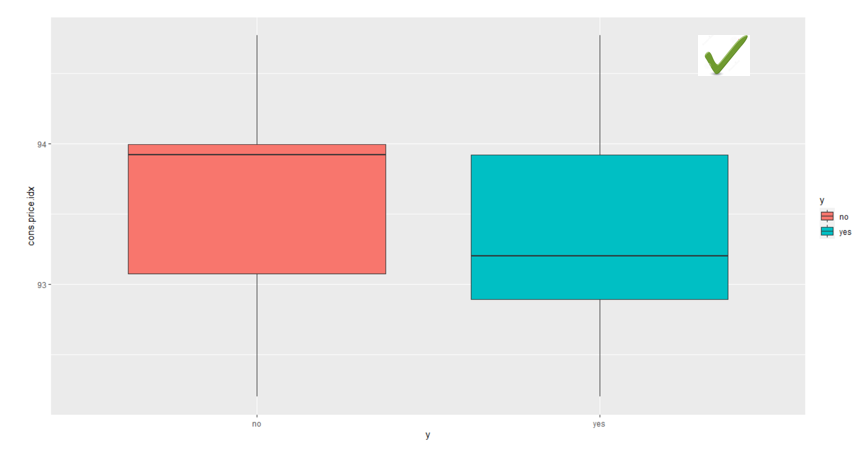


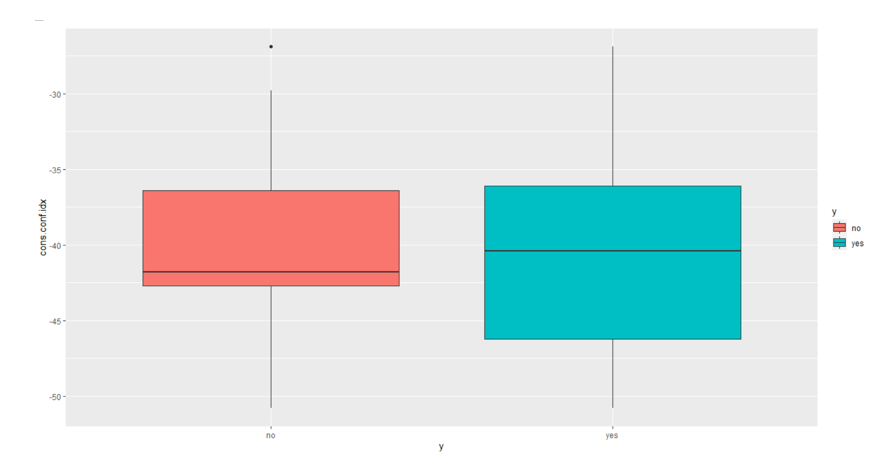


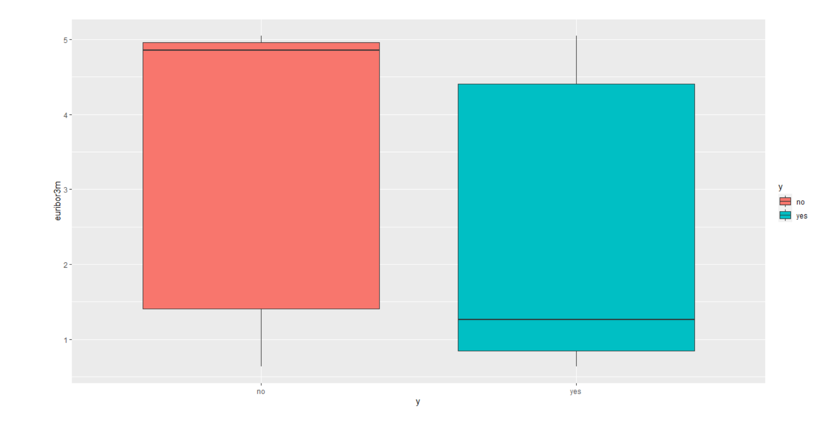
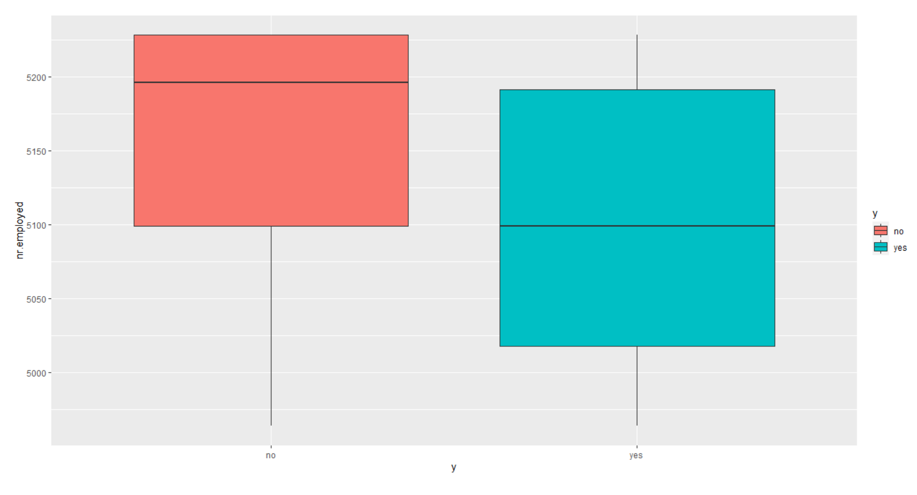




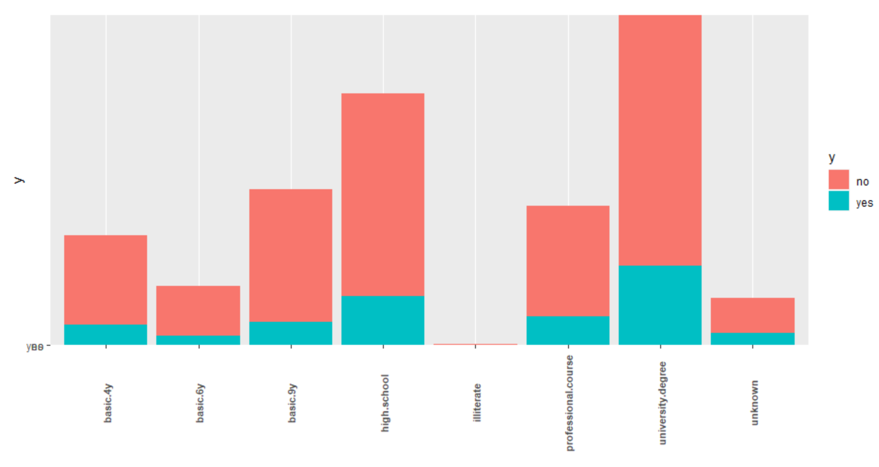


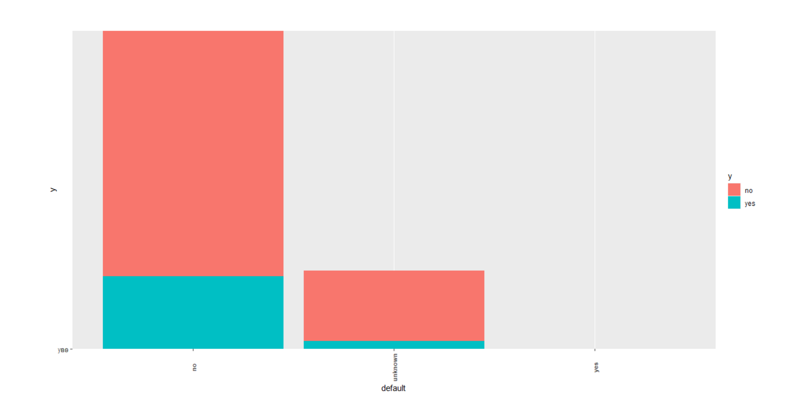


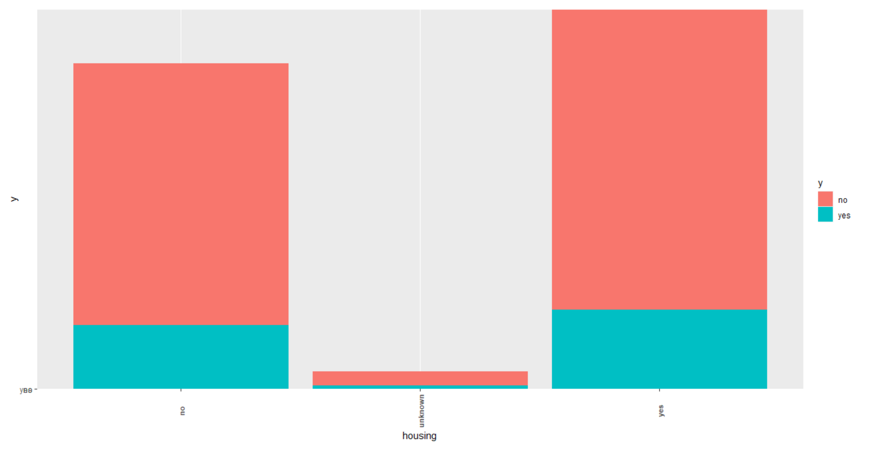


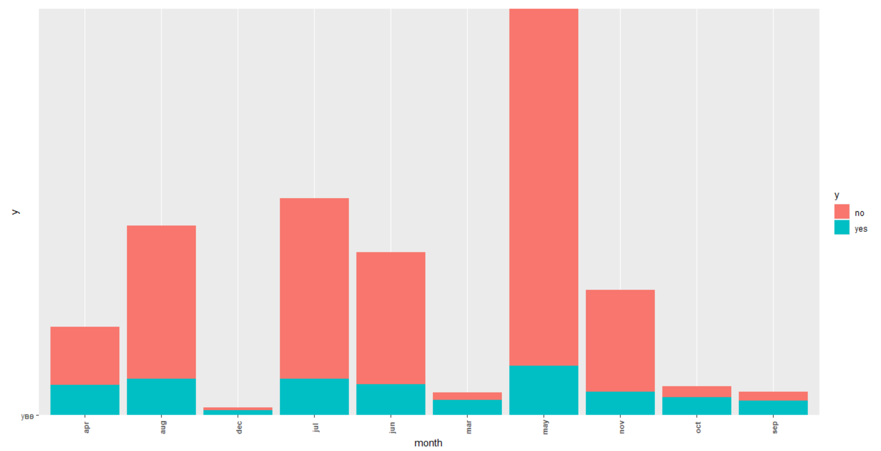


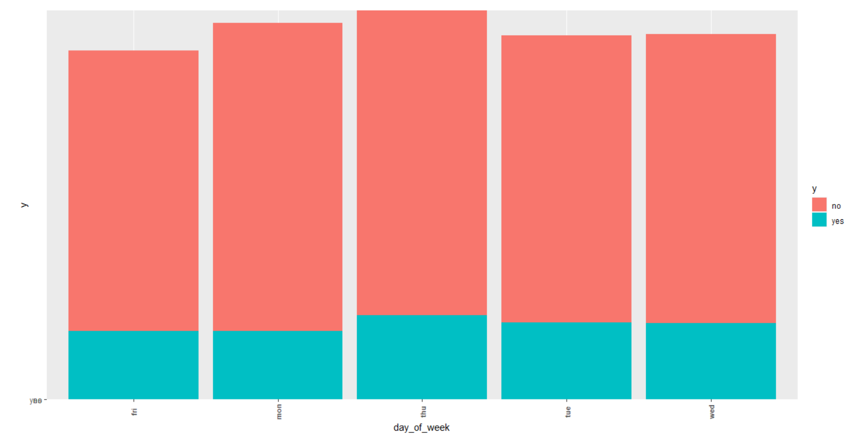
**Bar-plot:**











**QDA Model Summary:**

