Statistics Project II

## Introduction

The study relates to a phone call marketing campaign directed by a banking institution to predict whether or not a client will participate in a term deposit. Term deposits are considered to be a more secure investment opportunity, somewhat protected from market fluctuations, as opposed to stocks. Generally, a client will invest a specific sum for a set amount of time (e.g. 5 months) with a predetermined interest rate. The investment is then pulled after the time has passed or prior to the acceptance date, typically with a penalty cost.

The dataset contains the latest contact attempt to the clients, of which, the contact can be multiple times to determine whether or not the client will subscribe to a term deposit (campaign). In total, there are 41,188 observations. For social and economic context attributes, it should be noted that the indicators are assumed to be pulled from the general demographic, and are hence normalizing the data.

Additional information concerning the dataset can be found at [UCI’s learning repository](https://archive.ics.uci.edu/ml/datasets/bank%20marketing).

## Data Description

The full dataset provided by UCI is leveraged in the analysis. Variable descriptions and data types are listed in a comprehensive [table](#_ueyfupkjrepv) in the appendix. UCI’s website describes that more than one contact to the same client is generally required to determine if the client has acquired a bank term deposit, thus our observations are not independent. However, the data does not outline a unique identifier for the individual contacts of any repeated measures analysis and it is assumed that the observations are independent. The variable duration is reported in seconds and scaling is recommended for linear regression or random forest. Majority of the models applied in the analysis mapped below focuses on logistic regression, where scaling is not required due to the logistic transformation.

## Exploratory Data Analysis (EDA)

Initial inspection of the data reveals numerous factors with “unknown” listed as a viable category.

After determining that 26% of the data would be lost in removal of the observations in question, “unknown” is applied as its own level in the analysis.

The variables with high levels of “unknown” include: job, marital status, education, default (whether or not the customer has credit in default - failure to pay), housing loans, and personal loans. The number of days (pdays) passed since last contact also filled nulls with “999”. These are assumed to be the same as “has not been contacted” and treatedas zeros.

Trends discovered in relation to whether or not clients acquired a term deposit includes the [job description](#_ef11jtpaey5f), where students and retirees have a higher than average amount of term deposit acquisitions, with 25 and 31%, respectively, as compared to the 11% average. Also, term deposits may be subject to [seasonality](#_jszdeuhlhqv9) as March and December have higher amounts of acquisitions, ~ 50%.

In predetermining whether columns would be significant, pdays was a column under consideration for being dropped. A significant number of the observations do not have prior contacts, hence they do not have days after the first contacts filed (pdays not equal to 999). Considering that 3.68% of the observations (1,151 records) have had more than one contact, this appears to be an insignificant column, also supported by [the matrix](#_l2zte2fq7wi2) in the appendix.

The field “previous” shows that while ~14% of the clients were contacted prior to the current campaign, 11% were contacted only once prior; leaving 3% that were contacted more than once. Campaigns were determined to be potentially significant due to the randomness of the data and lack of correlation with other factors.

UCI warns in the data description that the duration variable should be removed, considering that the dataset states: for a realistic predictive model, this factor should not be considered. However, throughout the analysis of adding/excluding duration in the models , very little impact was observed. This will be touched on more when discussing the heatmap results.

#### Correlation Matrix

For the integer values,a [Pearson-Correlation](#_7ovoll8l3zui) is applied to see how predictors are related to one another. There appears to be a handful or correlations for the continuous variables. In the correlation matrix, it can be observed that the days since the last call and previous contacts made by this caller are slightly correlated. The four of the continuous attributes grouped as social and economic context attributes are all heavily correlated: employment variation rate, consumer price index, euribor 3 month rate, and number of employees.

#### Continuous Matrix

The continuous variables are used to build out a [matrix](#_l2zte2fq7wi2) for visual correlation analysis between variables. This time added in the predictor to highlight the result of purchasing or not of a deposit. The relationship between the duration of call and the number of calls made in campaign suggests that more calls did not result in a term deposit.

#### PCA

Keeping correlations in mind, a review was conducted of the continuous variables being transformed into principal component analysis variables (PCA), but the data did not indicate any clear delineation of groups. Only 10 continuous variables were used within the Principle Component Analysis. Although the continuous variables have at least four which appeared correlated in the Pearson-Correlation, the [plot comparisons](#_2z56irms0rqv) of principal components did not yield a clear distinction between groups. The lack of differentiation between the yes/no term deposit group conveys that PCAs are not a good fit for this dataset since it cannot clearly separate whether an observation acquired a term deposit or not. An [LDA ran on the PCA](#_e5mjn0iisk25) variables supported that the conclusions were accurate and that LDA was unable to have a high accuracy leveraging the PCA variables.

#### Heat Map

The [heatmap](#_phpz75q2bmfr) also shows a possible grouping around pday, previous, and separately campaign. It also showed that duration appears to be grouped. For this reason, despite seeing the comment about leaving it out for modelling purposes, duration was experimented with, leaving it in and comparing modeling results.

### Objective 1

The primary objective is to determine whether or not a client will subscribe to a term deposit, depending on the factors presented in the dataset. Objective one’s focus is generating a model which will predict subscription success while maintaining interpretability.

### Model Selection

#### Type of Selection

The data set is split into a training and testing set, with the training set downsampled to result in a balanced response. Standard logistic regression is run on the training data set before applying feature selection. Lasso, stepwise, forward, backward, and the original logistic regression are analyzed to determine best fit. [ROC curves](#_8fxfq357w3bb) are generated to compare accuracies of each model, visible in the appendix.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **AIC w/ Duration** | **AIC** | **AUC w/ Duration** | **AUC** |
| **Original** | 2559.1 | 2578.3 | 0.784 | 0.784 |
| **Stepwise** | 1525.3 | 2555.9 | 0.933 | 0.933 |
| **Forward** | 1530.3 | 2568.2 | 0.933 | 0.782 |
| **Backward** | 2555.9 | 2555.9 | 0.782 | 0.782 |
| **Lasso** | 1542.8 | 2568.4 | 0.934 | 0.785 |

Table 1: Displays the logistic regression models with applied feature selections, resulting in AIC, and AUC from ROC curves.

#### Assumptions

##### Lack of fit test

In comparing the models from table 1 with duration, the AUC of logistic regression with lasso feature selection applied is highest at 0.94, followed by stepwise and forward at 0.93. The small AUC difference in the models is not sufficient to determine best fit.

Reviewing AIC values with duration, both stepwise and forwards are significantly lower than lasso (1525, 1530, 1542 respectively). Balancing the two, the stepwise model is considered to be the best model of the 5.

When removing duration, as recommended by the data collectors, we notice that Stepwise maintains its high AUC, while the remainder of the models drop below 0.8. The increase in AIC continues to show backward and original with little change. This is due to backwards not containing duration initially and duration having an extremely small coefficient for the standard logistic model (0.007).

##### Influential point analysis (Cook’s D and Leverage)

The [Cook’s D plot](#_ix2tc0uh9o40) reveals a single data point that appears to be further in space than the remainder of the dataset. Despite this, the scale of the plot is ~0.04, supporting the conclusion that the data does not show concerns withoutliers.

This conclusion is supported by the [Leverage plot](#_ix2tc0uh9o40) which reveals no data points within the region of interest.

##### Residual Plots

The outcome meets the requirement of a binary class of yes or no for committing to term deposits. Intercorrelation is removed in the process of feature selection, causing any correlation between variables to result in one of the two being statistically insignificant.

Reviewing the [residual plots](#_ix2tc0uh9o40) in the appendix, the QQ-plot shows normality in the data.

Despite the customers being contacted more than once, independence is not violated as each measurement in the dataset records the final contact with the customer. During EDA, it was also discovered that the majority of the observations were first contact. Linearity is assumed to be present or filtered out through feature selection.

#### Parameter Interpretation

The coefficients and confidence intervals are on the order of odds ratio, allowing for simple interpretation. Although an intercept is generated for logistic regression, it is not typically interpreted.

The variables discovered to be significant in determining if a customer would commit to a term deposit is displayed in table 2, along with an interpretation example for categorical variables, and the true interpretation for a continuous variable.

Note that despite the original model resulting in duration being a significant factor, it is removed from all models, hence why it's not noted below.

|  |  |
| --- | --- |
| Factor | Interpretation |
| education | The odds of committing to a term deposit (opposed to not committing )with an education of 6 years increase by a factor of 2.39 |
| month of contact | The odds of committing to a term deposit (opposed to not committing) with a contact month of aug decrease by a factor of 0.67 |
| previous campaign outcome | The odds of committing to a term deposit (opposed to not committing) with a previous outcome of yes increase by a factor of 1,76 |
| employment variation rate | For a one unit increase in employment variation rate, the odds of committing to a term deposit (opposed to not committing) decrease by a factor of 0.007 |
| consumer price index | For a one unit increase in consumer price index, the odds of committing to a term deposit (opposed to not committing) decrease by a factor of 0.009 |
| number of employees | For a one unit increase in number of employees, the odds of committing to a term deposit (opposed to not committing) decrease by a factor of 0.04 |

Table 2: Maps the factors of importance to classifying term deposit commitment and interpretation of coefficients.

##### Confidence Intervals

The confidence intervals indicate the 95% confidence that the true odds ratio lies within the range specified for lower and upper confidence levels, under the assumption that the model assumptions are met and bias is not prevalent.

##### Performance Metrics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **Average** | **Cutoff** |  |
| Step | 0.8683 | 0.8699 | 0.8513 | 0.8632 | 0.30 |  |
| Step | 0.8740 | 0.8777 | 0.8366 | 0.8628 | 0.40 |  |
| Step | 0.8763 | 0.8810 | 0.8282 | 0.8618 | 0.45 |  |
| Step | 0.8785 | 0.8843 | 0.8189 | 0.8605 | 0.50 |  |
| Step | 0.8809 | 0.8876 | 0.8117 | 0.8601 | 0.55 |  |
| Step | 0.8829 | 0.8908 | 0.8024 | 0.8587 | 0.60 |  |

Table 3: Displays the complex model’s metric through cutoff iterations.

The final cut of 0.40 results in an accuracy of 0.87. The cutoff was selected due to it being the most balanced of the cutoffs tested (ranging from 0.3 to 0.6). Sensitivity and specificity (0.88 and 0.84, respectively) are equally weighted without the loss in either. The overall accuracy, although not the highest, is still considered to be sufficient.

Reviewing the performance metrics in [the appendix](#_ks0w0av0pm1v), notice that backwards and the standard logistic regression have the lowest specificity, with forward and lasso being relatively similar, if not slightly higher. This matches the trend noticed in the ROC curves.

#### Final conclusions from the analyses of Objective 1

The stepwise logistic regression results in high model statistics (AIC: 2555.9, AUC: 0.93), and reasonably high metrics of overall accuracy, sensitivity, and specificity in the range of high 80s. The performance metrics suggest that the model could potentially be improved by reducing bias and increasing complexity since the accuracy is less than 0.9.

The data meets the assumptions for logistic regression, with no concerns of outliers or leverage. Although the resulting coefficient and confidence intervals were interpreted. Object 2 explores additionally complex models which focuses on model performance and prediction rather than interpretation.

It is noteworthy to report that the variables determined to be significant in predicting the term deposit outcome are education, month of contact, previous campaign outcome, employment variation rate, consumer, price index, number of employees, and duration. While duration had a considerable small coefficient (0.007), the removal of the variable resulted in little to no difference in the final outcome.

## Objective 2

The primary focus of object 2 is to create a model with accurate predictability despite the potential loss of interpretability. In order to achieve/test out the theory, additional complexity such as new variable generation, interaction terms, and complex algorithms were introduced.

### Main Analysis Content

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **AIC w/ Duration** | **AIC** | **AUC w/ Duration** | **AUC** |
| **Custom** |  |  |  |  |
| **Stepwise** | 1525.3 | 2555.9 | 0.933 | 0.933 |

### Models

#### Linear Discriminant Analysis

A linear discriminant analysis (LDA) was run on the 10 continuous variables within the data set to determine if just the subset of continuous variables were able to predict a better fit model on the balanced test/training split. LDA on the original variables produced an AUC of 0.917. By leveraging principle component analysis (PCA) variables within an LDA model, no improvement in accuracy is observed, with an AUC score of 0.59. See appendix for ROC curves.

#### Custom Complex Model

When exploring other custom variables to include in the model, the stepwise variables were leveraged in order to potentially improve on the best fitting model of the data. Initially when creating the interactions,practical knowledge of the data and past experience was applied to hypothesize on their relationship. For example, looking to see if campaign and month had any impact on whether or not clients acquired a term deposit. After investigating the interaction variable’s VIF compared to the other variables included in the model, the conclusion was to either toss it, iterate, or keep it in. The following interactions were explored: campaign\*month, campaign\*job, month\*campaign, job\*duration, duration\*day\_of\_week, duration\*month, campaign\*contact, contact\*job, contact\*duration. Additionally, categorical cuts of some of the variables were made to see the impact, such as bucketing the ages into groups. Pdays were also alteredto have the original data and another field that converted the pdays =999 into 0’s, because in effect those observations with ‘999’ had not been contacted.

After including 3 additional interactions to the stepwise logistic model, an accuracy and confusion matrix with a cutoff of 0.1 were run. The custom model proved to not provide any additional predictive qualities than the stepwise model. The interactions made the model more complex, whereas the stepwise doesn’t suffer from the added bias.

Working through the iterations, roughly three dozen variations of the model were run and the selected complex model contains the duration variable. Based on the confusion matrix results of the models run in objective 1 and objective 2, stepwise proved to have the best accuracy with XX score.

#### Decision Trees

An attempt to leverage a decision tree to see if it was able to provide any insightful or accurate predictions resulted in a tree with few splits (not using the gini method) containing both continuous and categorical variables. These default splits of the decision tree did not prove to have a fruitful explanation and is not recommend for the final model analysis (See Appendix: [Decision Tree](#_r7vo1suzo64c)).

### Performance Summary

#### Conclusion/Discussion

The conclusion should reprise the questions and conclusions of objective 2 with recommendations of the final model, what could be done to help analysis and model building in the future, and any insight as to why one method outshined all the rest if that is indeed the case. If they all are similar why did you go with your final model?

Based on the analysis, it wasdetermined that the logistic regression using stepwise variable selection had the most accurate predictions considering it’s low complexity and easy interpretation. Little differences were observed in the models including or excluding the duration variable.

Will probably remove…

Duration in random forest is not scaled and can present as more significant than it really is. Logistic regression scales the variables and duration ends up having a very small coefficient, so when it is included in the model it has very little impact on the overall results.

## 

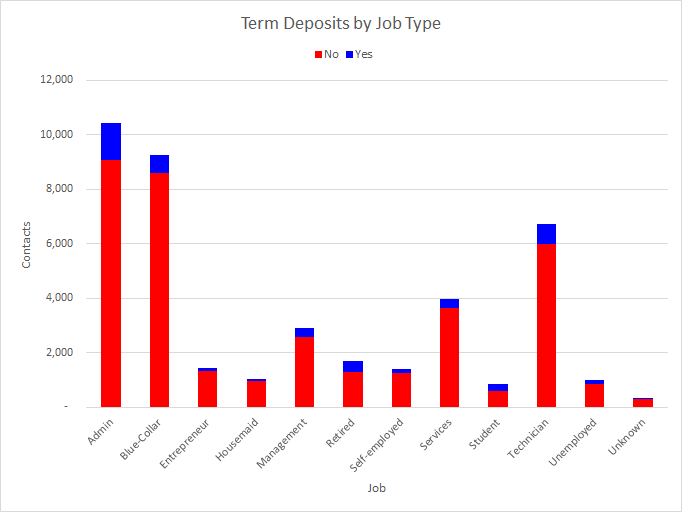
## Appendix

### Graphics

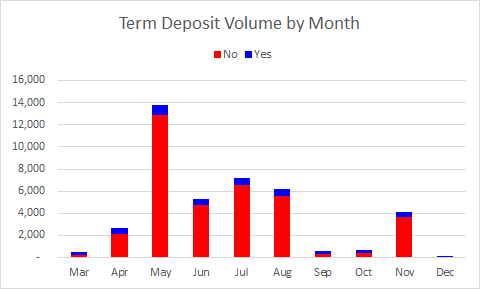
#### Data Description

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Variable Type** | **Variable Description** |
| age | numeric | Age of contact |
| job | Factor/Categorical | Contact’s type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') |
| marital | Factor/Categorical | Contact’s marital status ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed) |
| education | Factor/Categorical | Highest level of Contact’s education ('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown') |
| default | Factor/Categorical | Does Contact has credit in default? ( 'no','yes','unknown') |
| housing | Factor/Categorical | Does Contact has housing loan? ('no','yes','unknown') |
| loan | Factor/Categorical | Does contact has personal loan? ('no','yes','unknown') |
| contact | Factor/Categorical | How the contact was contacted, communication type ('cellular','telephone') |
| month | Factor/Categorical | Last contact month of year ( 'jan', 'feb', 'mar', ..., 'nov', 'dec') |
| day\_of\_week | Factor/Categorical | Last contact day of the week ('mon','tue','wed','thu','fri') |
| duration | Numeric | Last contact duration, in seconds |
| campaign | Numeric | Number of contacts performed during this campaign and for this client (includes last contact) |
| pdays | Numeric | Number of days that passed by after the client was last contacted from a previous campaign ( 999 means client was not previously contacted) |
| previous | Numeric | Number of contacts performed before this campaign and for this client |
| poutcome | Factor/Categorical | Outcome of the previous marketing campaign (: 'failure','nonexistent','success') |
| emp.var.rate | Numeric | Employment variation rate - quarterly indicator |
| cons.price.idx | Numeric | Consumer price index - monthly indicator |
| cons.conf.idx | Numeric | Consumer confidence index - monthly indicator |
| euribor3m | Numeric | Euribor 3 month rate - daily indicator |
| nr.employed | Numeric | Number of employees - quarterly indicator |
| y | Binary | Has the client subscribed a term deposit? (: 'yes','no') |

##### EDA: Job Type

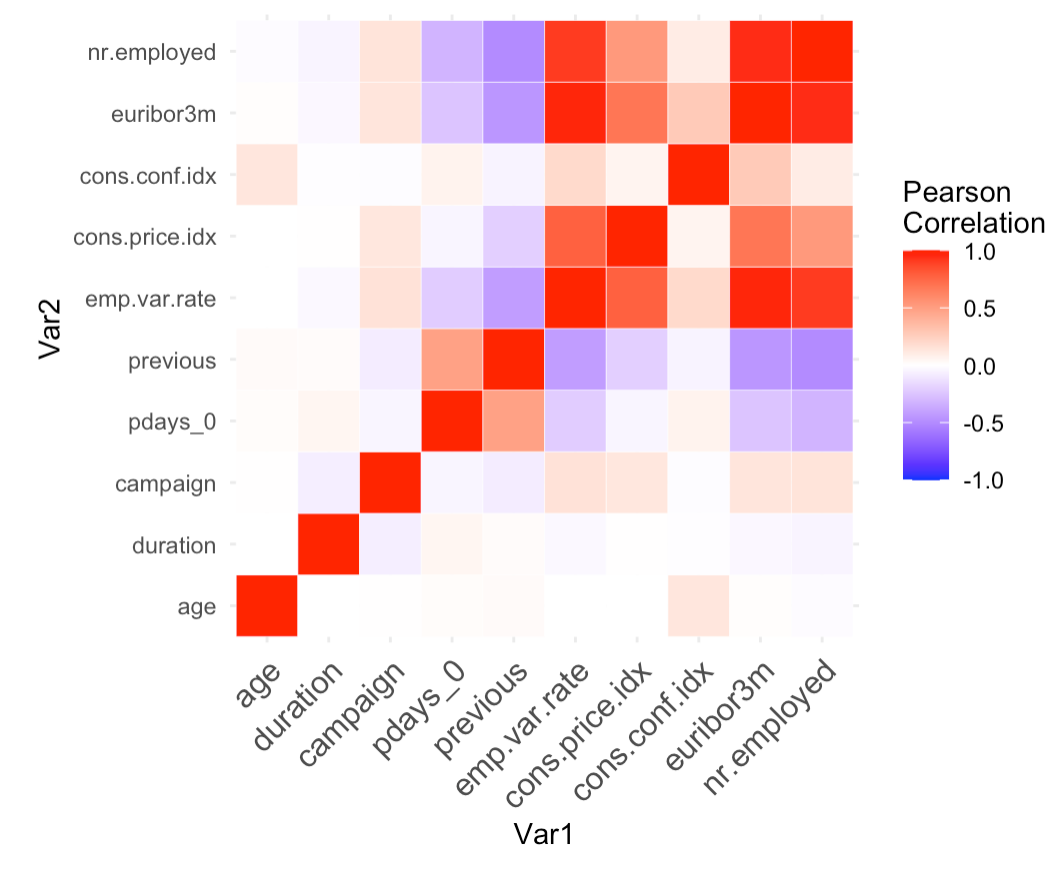


##### EDA: Month

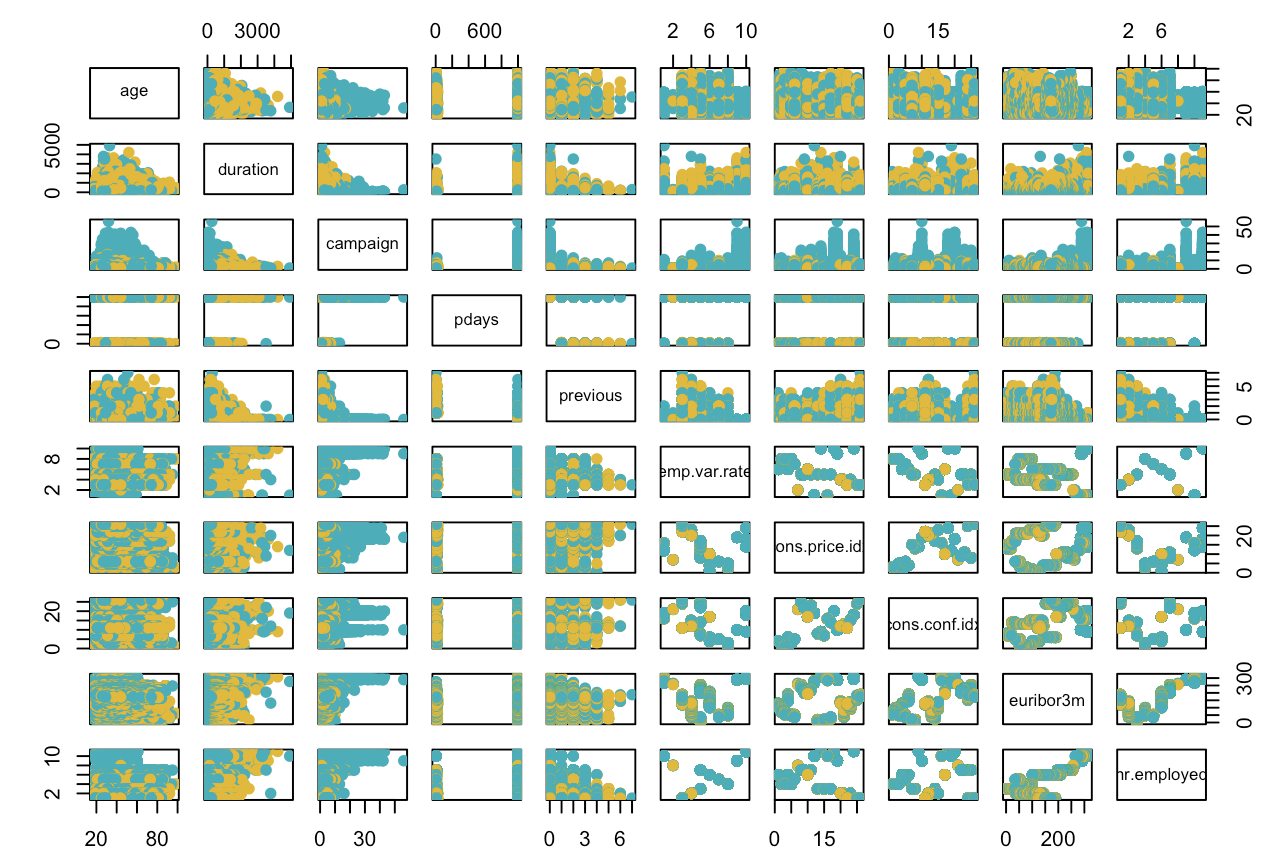


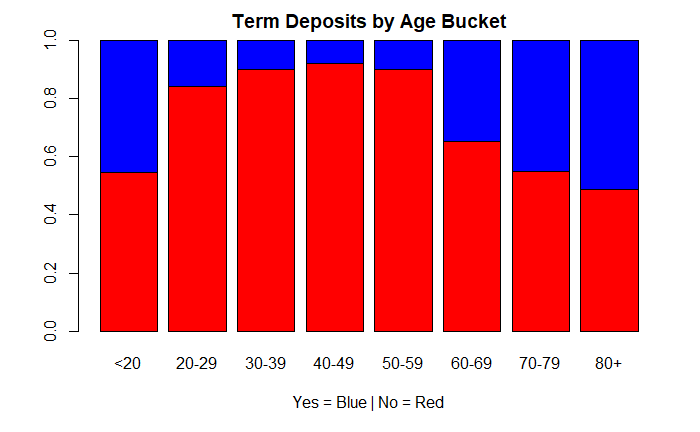
#### EDA

##### Pearson-Correlation Matrix

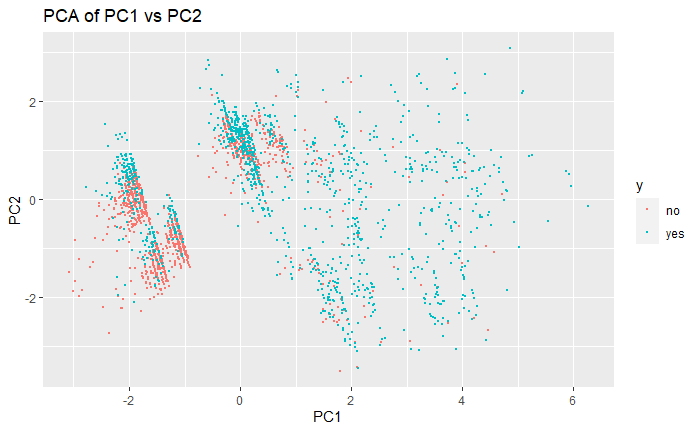


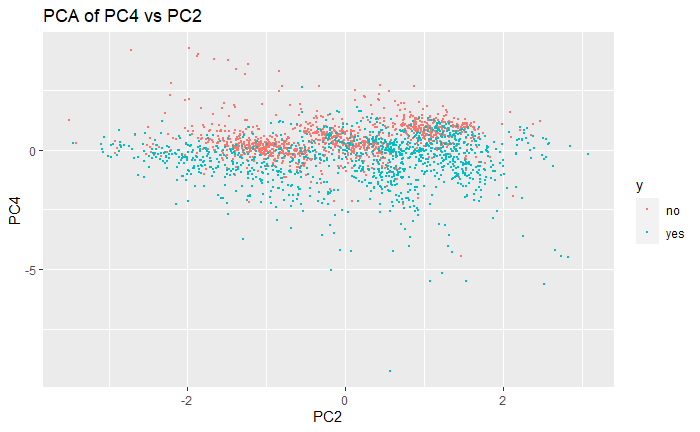
##### Continuous Matrix



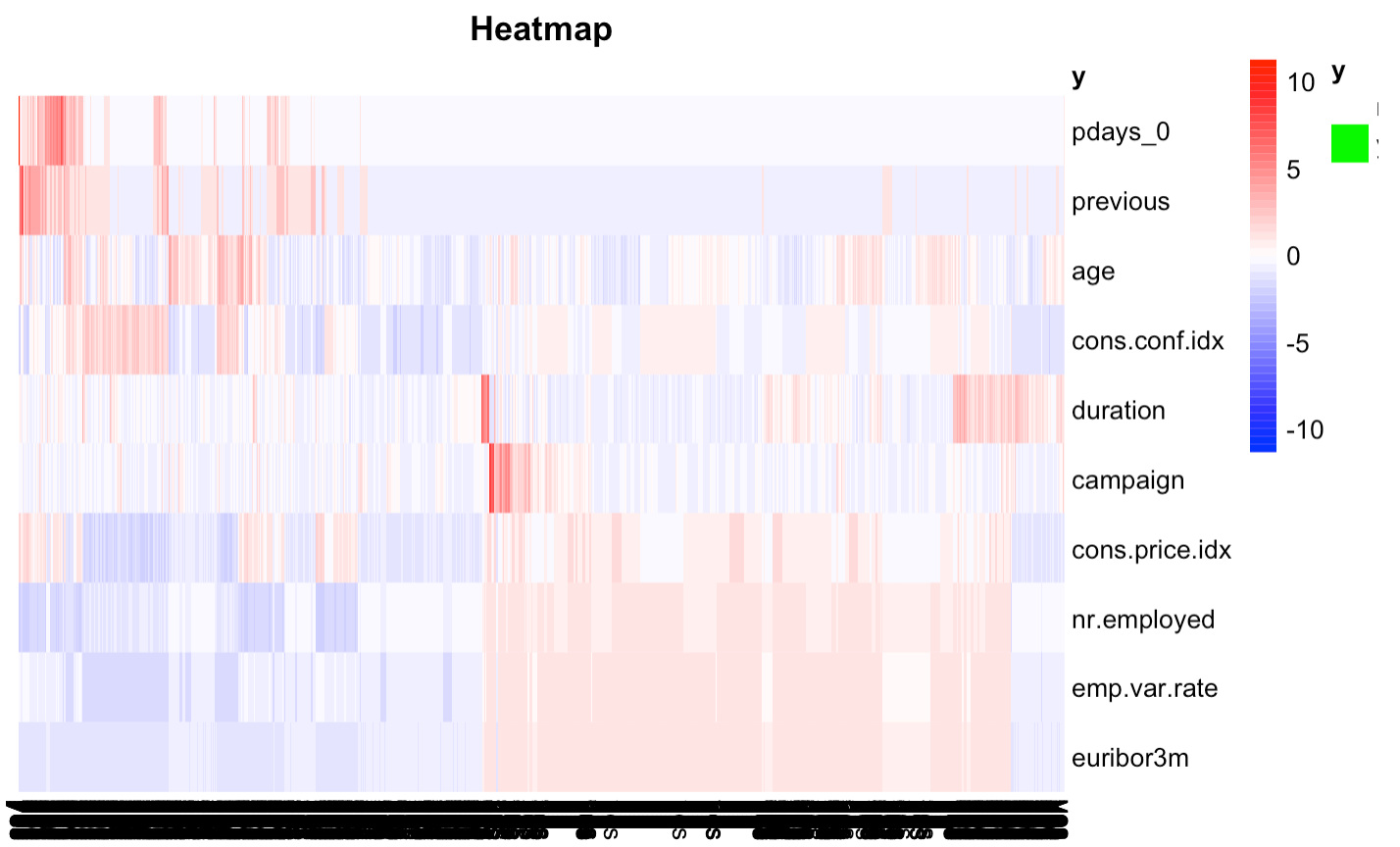


##### PCA



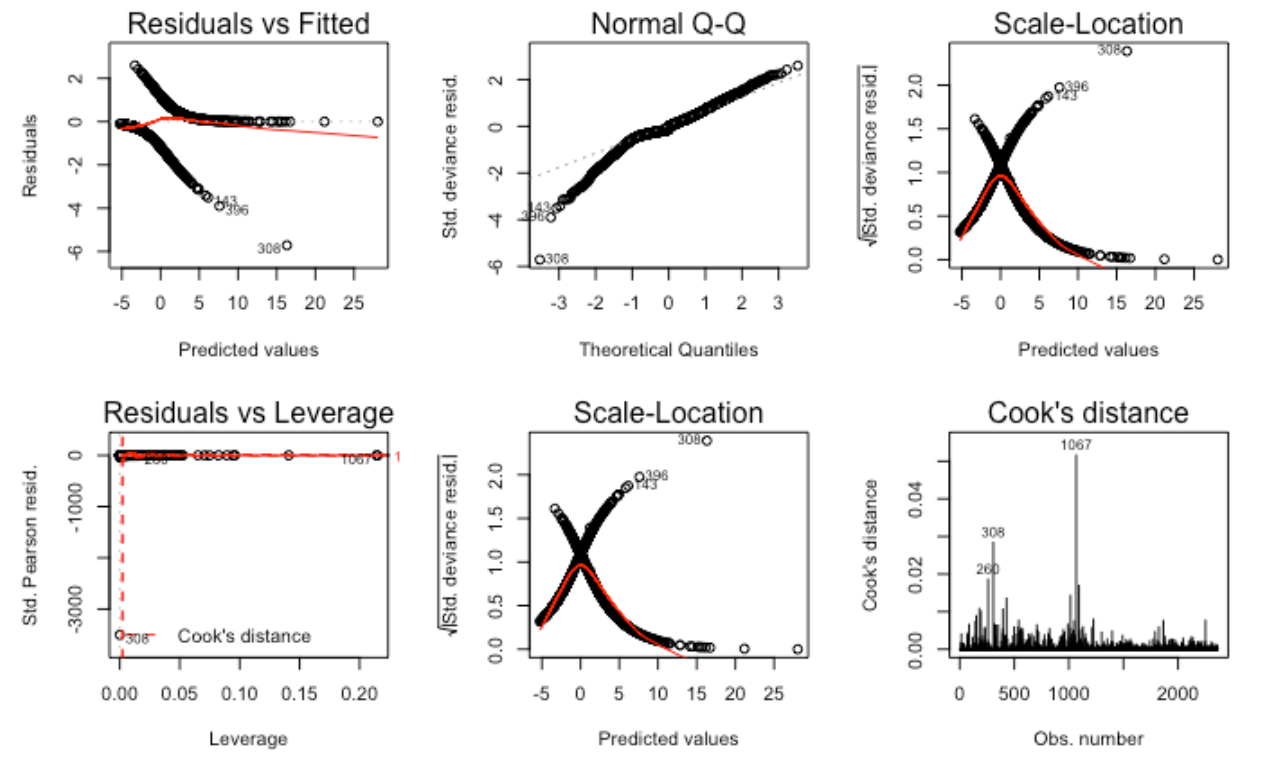


##### Heat Map

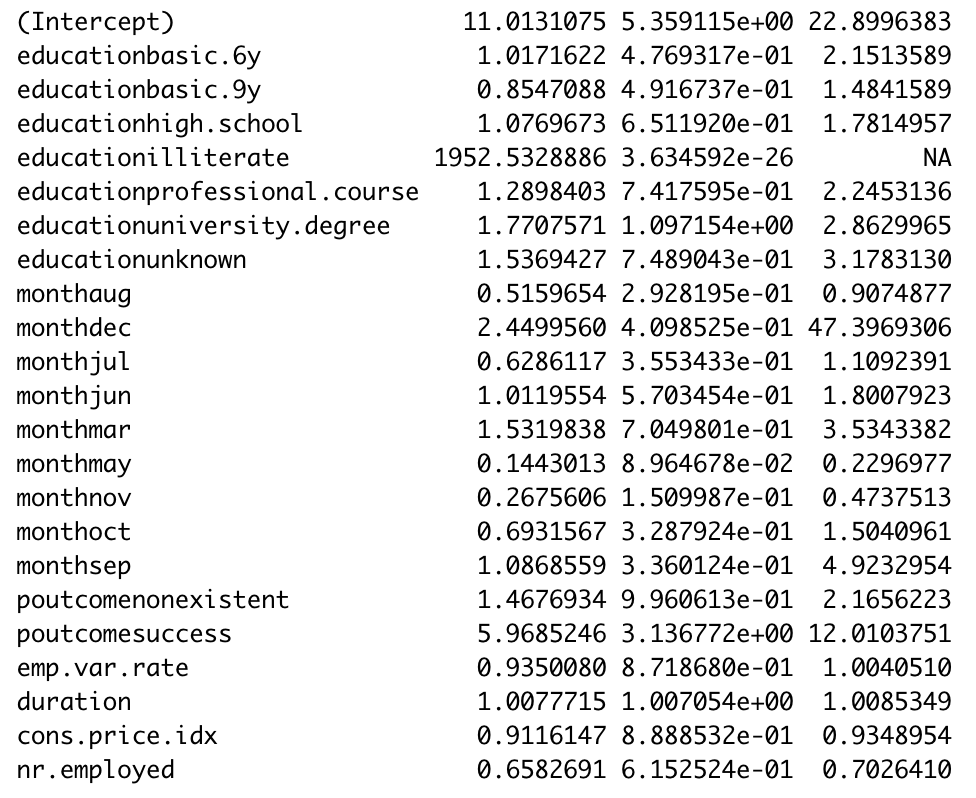


#### Model 1

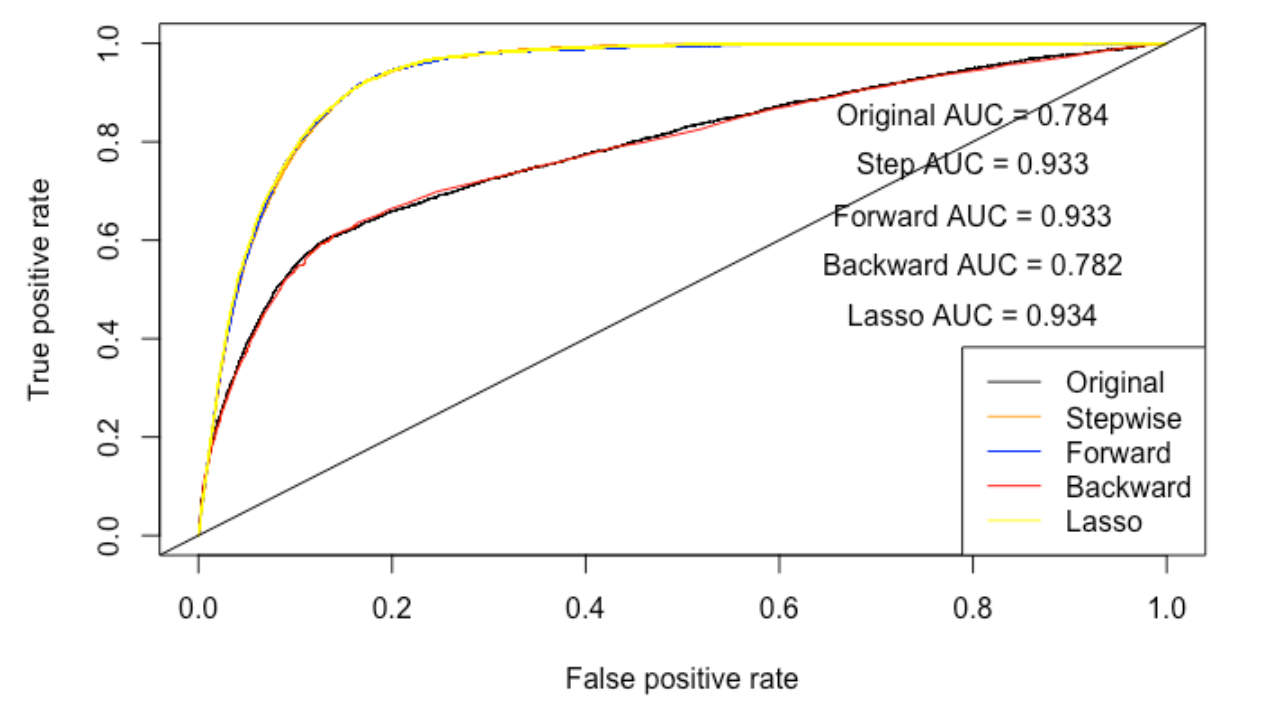
##### Residuals



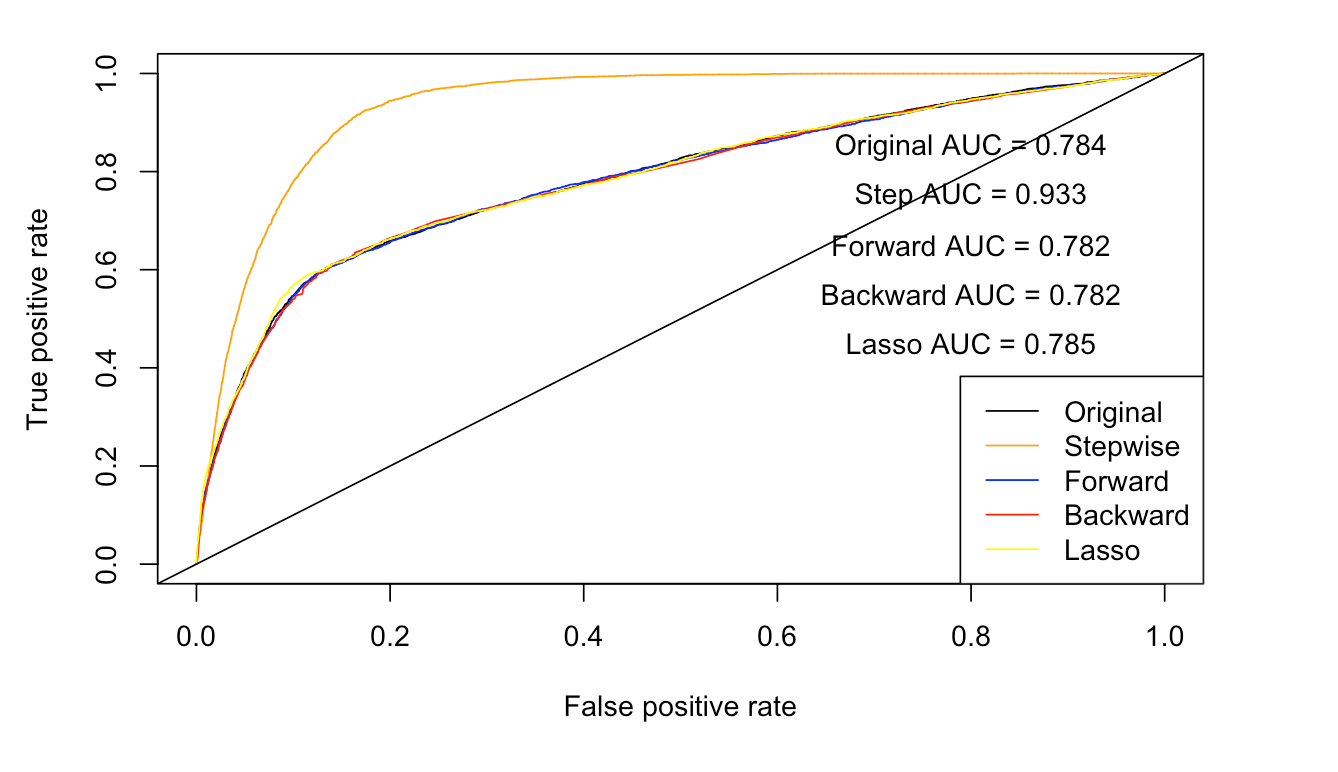
##### Odds Ratio/Confidence Intervals



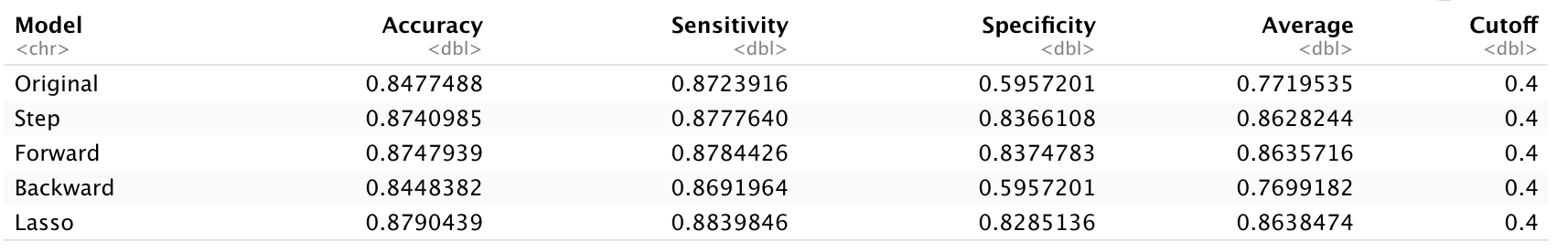
##### Variable Selection ROC Curves with Duration



##### Variable Selection ROC Curves without Duration

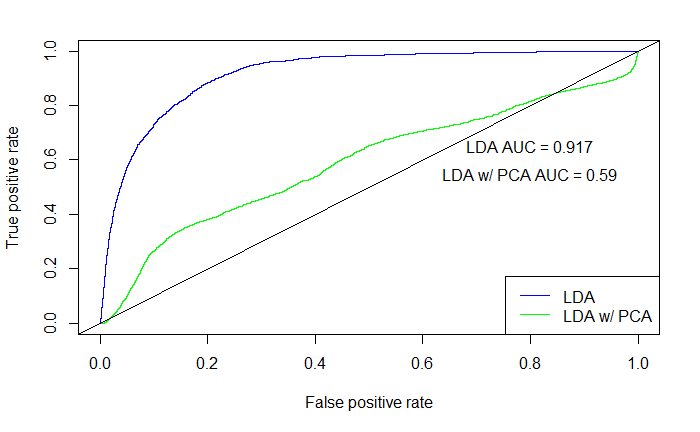


##### Performance Metrics

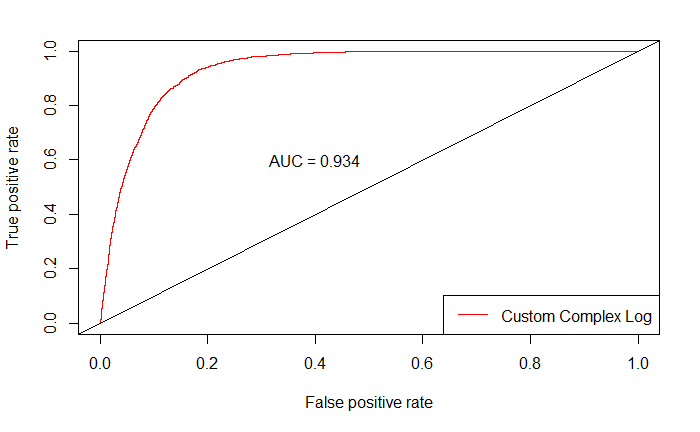


#### Model 2

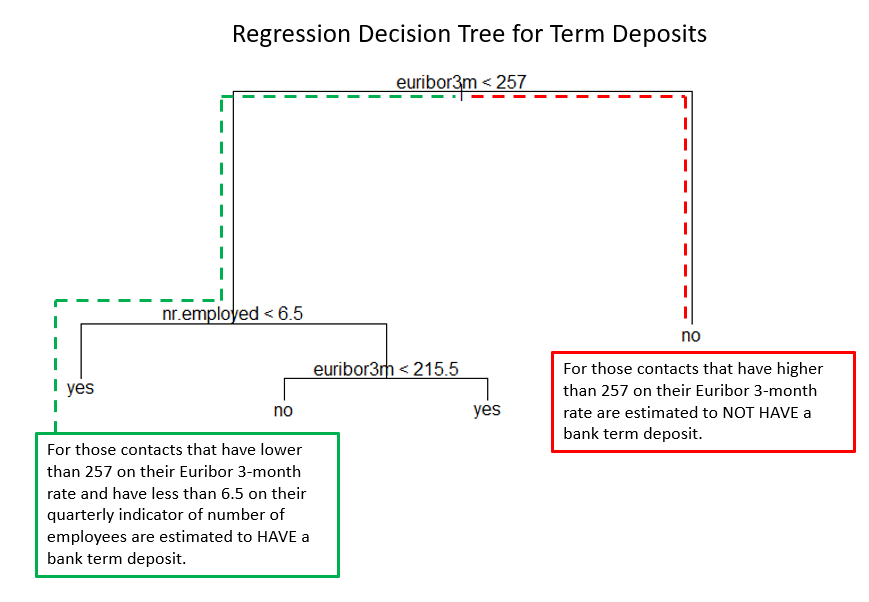
##### ROC Curve of LDA/PCA Models



##### Custom Complex Model



##### Decision Tree



##### Random Forest

### Code

Well commented SAS/R Code **Required**