6372: Project 2

Simerpreet Reddy, Rinku Lichti, Megan Ball

# Introduction

The goal of this analysis was to develop an optimal classification model to predict the success rates of a Portuguese bank telemarketing campaign. As part of this, a total of four models were developed for use in comparison for best overall combined accuracy, sensitivity, and specificity:

* Simple, interpretable linear regression model
* Complex linear regression model
* Linear discriminant analysis (LDA) model
* Random Forest model

The subsequent discussion will review the analysis steps, model metrics, and ultimate best model for classification.

# Data Description

We used bank-additional-full.csv as our base dataset. This dataset consisted of 41,188 observations with a total of 21 different variables that consisted of various factors pertaining to the following:

* bank client data: age, job, marital status, education, default (whether or not the client has credit in default), housing (whether or not the client has a housing loan) and loan (whether or not client has a personal loan).
* previous campaign attributes related to the last time the customer was contacted:  contact (cellular or telephone), month, day\_of\_week and duration (of the call in seconds).
* Other campaign attributes: campaign (times contacted in the current campaign), pdays (days since last call), previous (times contacted in last campaign), poutcome (outcome of the previous campaign).
* Social and economic context attributes:  emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed.
* Response variable: y (whether or not a client subscribed to a term deposit)

Out of these the following were categorical variables: job, marital status, education, default, housing, loan, contact, month, day\_of\_week, poutcome and the response variable y. The following were the continuous variables: duration, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m and nr.employed.

Derived variables

We derived the following variables using the provided variables:

Age\_Group from age: We used the variable age to create variable Age\_Grp (age groups) with values "17-31","32-37" ,"38-47", "47-55", ">55". We did this based in IQR for age.

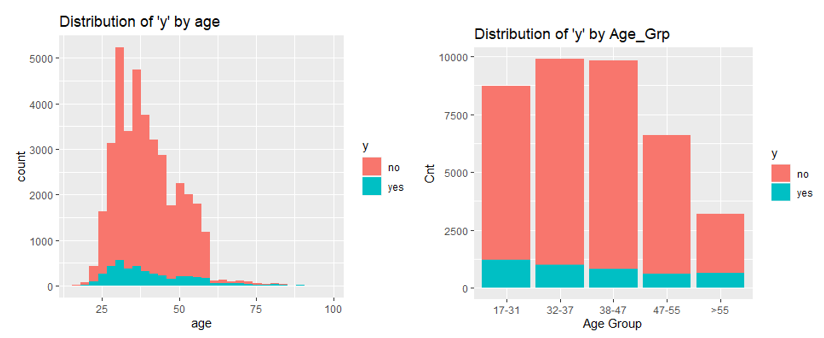


Figure 1. Age distribution and grouping coded by response variable

prevly\_Cntctd from pdays: We created variable prevly\_Cntctd with values yes/no to see if the client has been previously contacted ever. Pdays**=**999 meant that the client has never been contacted before.

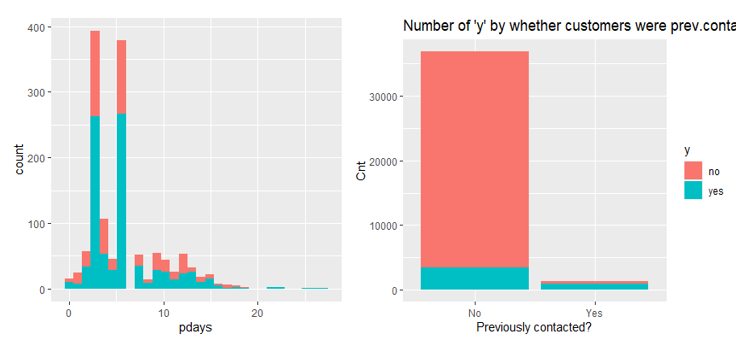


Figure 2. Distribution of pdays (excluding the ‘999’ group) and previously contacted grouping coded by response.

duration\_group from duration:We created variable duration group to determine what range did the last call duration fall in - "0-5min", "5-10min" or "10+ min".

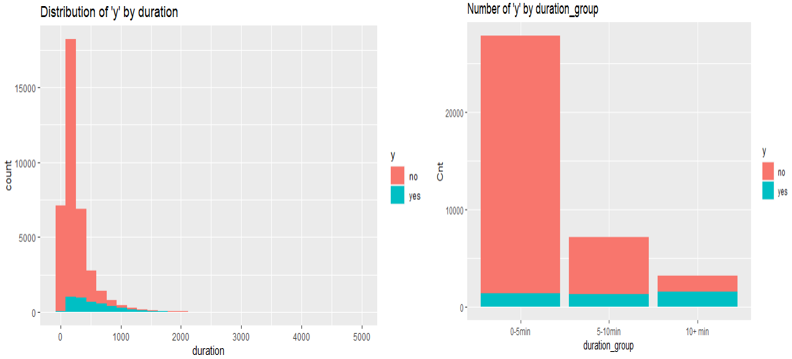


Figure 3. Distribution of duration and duration group coded by response.

Additional Data Cleaning

We chose not to balance the data as it would lead to losing a lot of the information. Instead, we decided to plot ROC curves on the chosen models to predict the best cut off to be used for prediction.

There were no null values in the data set but ‘unknowns’. We removed the rows where the following variable had ‘unknown’ values – marital (80 total unknowns), job (289), loan (990) and education (1732). The column default had ‘unknowns’ as well but since that consisted of 20.9% of the total default observations, we decided to keep those as they could be important predictors. However, we removed 3 rows of data with default=’yes’ as such few values were adding error the data model due to the small sample size.

We changed the data type of following variables to factors: job, marital, education, housing, loan, contact, month, day\_of\_week, default, poutcome, and y. We adjusted the level of y to (‘No’, ’Yes’) to make sure ‘yes’ was considered as event=1. We kept the continuous variables as numeric.

While splitting data set into train and test, we made sure these data sets received an 80:20 ratio of both ‘Yes’ and ‘No’ of the response variable to mimic the ration in the complete data set.

Please see Figure 10 in the appendix for summary statistics and frequency values for the data set.

# Exploratory Data Analysis

Performing EDA, we saw some interesting relationships among variables. A few of them are shown below.

Chart, bar chart

Description automatically generated

Figure . Distributions of job and previous by proportion of response variable yes/no.

When looking at job, there is a high ratio of no to yes for most job roles, but most significantly for blue-collar, admin, and technician as they have the largest sample sizes. For previous, the proportion of ‘yes’ seems to increase by each contact.

Next, when reviewing month (see Figure 11), there is a much smaller sample size for the months of December, March, October, and September. However, the data for these months appears to be more significant as the ratio for yes is higher compared to the other months.

Looking at the proportions table below, we saw that almost 50% of the ‘yes’ values for the response variable were after 10+ minutes of calls with the prospective clients. Thus, making duration group one of the important predictors.

Table . Proportion of yes by duration group.

|  |  |  |  |
| --- | --- | --- | --- |
| Duration Group | Y | N | duration\_group\_conv |
| 0-5 min | yes | 1375 | 0.0494 |
| 5-10 min | yes | 1323 | 0.184 |
| 10+ min | yes | 1560 | 0.486 |

The plots below show the variation in the response variable y by whether the client has been contacted before. There is a clear increase in ‘yes’ for the group that has had prior contact (prevly\_Cntctd = ‘Yes’). We also see a possible interaction between duration\_group and prevly\_Cntctd as the proportion of those previously contacted increase slightly by the duration of the call.

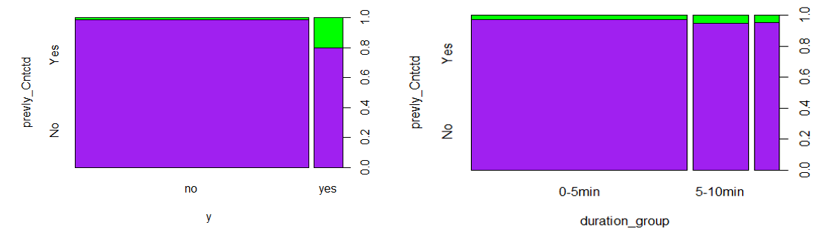


Figure . Mosaic plots for previously contacted vs response variable and duration group vs previously contacted.

Furthermore, for those who subscribed to a term deposit, there appears to be a dependence on their default status and contact method (see Figure 12). There is a difference among the subscription rates for those in the default ‘no’ category and those in the default ‘unknown’ category, indicating that the ‘unknown’ group is useful for some predictive modeling.

Correlation was also checked for all the new and existing variables (see Figure 13in appendix). Although high correlation existed between some variables, such as euribor3m and nr\_employed and pdays and poutcome, we decided to keep these variables in and let VIF and feature selection tools to indicate which to remove.

Finally, PCA was performed but did not yield any interesting information regarding the relationship amongst the continuous variables.

# Objective 1

## Restatement of Problem

Our first goal was to develop a simple and interpretable model to predict and explain the relationship between our variables and the probability of customers opening an account. To do this, we chose logistic regression as our desired model type and performed variable selection and validation using confusion matrices to select the best interpretable model.

Modeling Strategy

After doing EDA and creating a few derived variables, we started with creating a simple model using all the variables and then step by step removing statistically insignificant variables as well as variables with high multicollinearity based on both VIFs and scatter plots. After every column removal, we re-ran the model and repeated the exercise till we found the model with statistically significant as well moderate (generally with values less than 10) VIF variables.

We then ran both stepwise selection and LASSO as competing feature selection methods, and further limiting our variables based on VIF. Using ROC curves to find a good cut off range, we predicted the response variable with multiple cut off values in the range 0.1 and 0.5 to determine optimum value. We then compared the accuracy, sensitivity and specificity of all the different models – simple, step, and lasso – on various cut offs. Since the primary goal is to predict when a customer will subscribe to a term deposit (positive response variable = ‘yes’) we chose a cut-off value based on balancing our overall accuracy and sensitivity over specificity. Please see Table 6in the appendix for the metric comparisons for each of the models developed.

Model

For the simple interpretable model, the following model gave us the best combined accuracy, sensitivity, and specificity at a cut off probability value of 0.15.

## Assumptions

* Independence of data points: There is no evidence of any repeated measures or any other interdependence in the data set.
* Multicollinearity: We removed the variables that had high correlations according to VIF.
* Outliers/Influential observations: We looked at Cook’s D, leverage, deviance and residual plots to examine outliers. We did not see any significant difference in the model performance after removing these. See details in Figure 15 and Figure 16 in the appendix.

Lack of fit

* We performed a global test for the coefficients (beta) being equal to zero. With Likelihood ratio, Score and Wald test all agreeing at p-value <0.001, we reject the null hypothesis and conclude that the logistic regression model is valid.

Table . Logistic regression overall model significance results.



* Cross validation was done in the form of a test/train split which demonstrated good metrics on our test data set.

Table . Metrics for final simple logisitc model at cut-off of 0.15

|  |  |
| --- | --- |
| Metric | Result |
| Accuracy | 86.6% |
| Sensitivity | 83.7% |
| Specificity | 87.2% |

## Model Interpretation

Table . Odds ratios and 95% confidence intervals for logistic regression model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Odds Ratio | 95% CI: Lower Limit | 95% CI: Upper Limit | P-value |
| Intercept | 1.486 x 10-21 | 5.619 x 10-26 | 3.934 x 10-17 | < 0.001 |
| Job: blue-collar | 0.707 | 0.610 | 0.820 | < 0.001 |
| Job: entrepreneur | 0.871 | 0.600 | 1.151 | 0.332 |
| Job: housemaid | 0.818 | 0.589 | 1.136 | 0.230 |
| Job: management | 0.950 | 0.784 | 1.150 | 0.600 |
| Job: retired | 1.085 | 0.843 | 1.397 | 0.528 |
| Job: self-employed | 0.893 | 0.690 | 1.155 | 0.388 |
| Job: services | 0.741 | 0.615 | 0.892 | 0.002 |
| Job: student | 1.192 | 0.917 | 1.550 | 0.190 |
| Job: technician | 0.994 | 0.863 | 1.146 | 0.938 |
| Job: unemployed | 0.884 | 0.655 | 1.195 | 0.423 |
| Default: unknown | 0.722 | 0.620 | 0.841 | < 0.001 |
| Contact: telephone | 0.691 | 0.588 | 0.811 | < 0.001 |
| Month: August | 1.264 | 1.000 | 1.600 | 0.050 |
| Month: December | 1.052 | 0.643 | 1.720 | 0.841 |
| Month: July | 1.396 | 1.123 | 1.736 | 0.003 |
| Month: June | 1.485 | 1.202 | 1.835 | < 0.001 |
| Month: March | 5.741 | 4.332 | 7.610 | < 0.001 |
| Month: May | 0.530 | 0.446 | 0.631 | < 0.001 |
| Month: November | 1.072 | 0.854 | 1.346 | 0.545 |
| Month: October | 1.514 | 1.135 | 2.021 | 0.005 |
| Month: September | 0.992 | 0.725 | 1.359 | 0.961 |
| Duration | 1.002 | 1.002 | 1.003 | < 0.001 |
| Campaign | 0.956 | 0.931 | 0.981 | < 0.001 |
| Previous number of contacts | 0.762 | 0.694 | 0.836 | < 0.001 |
| Cons\_price\_idx | 1.692 | 1.513 | 1.893 | < 0.001 |
| Cons\_conf\_idx | 1.060 | 1.047 | 1.072 | < 0.001 |
| Euribor3m | 0.491 | 0.470 | 0.513 | < 0.001 |
| Age Group (32-37) | 0.829 | 0.727 | 0.946 | 0.005 |
| Age Group (38-47) | 0.743 | 0.646 | 0.855 | < 0.001 |
| Age Group (47 – 55) | 0.859 | 0.734 | 1.004 | 0.057 |
| Age Group (55+) | 0.930 | 0.753 | 1.149 | 0.502 |
| Previously contacted: Yes | 6.190 | 5.048 | 7.590 | < 0.001 |
| Duration Group: 5-10min | 2.983 | 2.600 | 3.424 | < 0.001 |
| Duration Group: 10+ min | 7.200 | 5.563 | 9.319 | < 0.001 |

The variables with the largest odds ratios were categorical variables and included the month of March and each of the duration groups. Holding all other variables constant, the odds of a client saying yes to a term deposit in the month of March is estimated to be 5.7 times the odds of a client saying yes in the month of April. A 95% confidence interval associated with this odds ratio is 4.3 to 7.6. Additionally, the odds of a client saying yes to a term deposit if the duration of the last call is 10 minutes or greater is 7.2 times the odds of a client whose final phone call duration was less than five minutes. A 95% confidence interval associated with this odds ratio is 5.6 to 9.3. We will discuss the relationship between duration and our response variable in more detail in the following sections.

The continuous variables with the largest odds ratios included duration, consumer price index, and consumer confidence index. For each additional minute spent on the phone, it is estimated that the odds of a client opening an account increase by 16%, holding all other variables constant[[1]](#footnote-1). For each one unit increase in consumer price index, it is estimated that the odds of a client opening an account increase by a multiplicative factor of 1.7, holding all other variables constant. A 95% confidence interval associated with this odds ratio is estimated to be between 1.5 and 1.9.

# Objective 2

In an attempt to further improve the accuracy and sensitivity of our classification model, we also built a more complex logistic regression model including interaction, made classifications using linear discriminant analysis, and also developed a random forest model.

#### Complex Model

A complex logistic regression model was built off of the variables deemed most significant from our simple regression model. Based on EDA, and as mentioned above, the factors that appeared to have some sort of interaction included default with duration, contact with duration, default with month, and month with euribor3m.

#### Linear Discriminant Analysis

In addition to a complex logistic model, both a linear and quadratic discriminant analysis were run on the numeric variables for comparison. Since both LDA and QDA require numeric values only, this limited our data set to continuous variables only. Two subsets were tested – one with all of the continuous variables and one with only the continuous variables used in our best logistic model from objective 1. The overall performance of the LDA model was comparable to all other models developed but did not perform as well as the other models in terms of specificity. This is likely due to the fact that our categorical variables added additional explanation for variation within the model.

#### Random Forest

To tune Random Forest, we chose mtry and the split function. Mtry was an obvious choice as the default Random Forest hyperparameter, which tells Random Forest how many variables to consider when performing each split. For the split function, we chose gini, extratrees, and Hellinger, all of which are suitable for classification, but of which we were not sure which one would perform best given our data. Optimizing for AUC, the winning parameters are an mtry of 7 predictors considered at each split, and the Hellinger split rule (see Figure 17).

The most influential variables in the model are both duration and nr\_employed. From Figure 6, we can see the more important variables are continuous for the random forest model. This is contrary to our linear regression models which tended to find several categorical variables significant in addition to our continuous variables. This is likely because continuous variables allow for more split possibilities within the trees versus the categorical variables.

Duration of last call may be a strong predictor as those who do subscribe to a term deposit will typically have a longer final phone call. If so, then this model may be not so useful for predicting those who are likely to convert in the future. Predicting on duration feels like predicting rain based mostly on feeling raindrops starting to fall on my head. Maybe duration should be replaced with the duration of the second to last call? Or perhaps average duration of all calls? It may be worth following up with the company to refine the available variables.

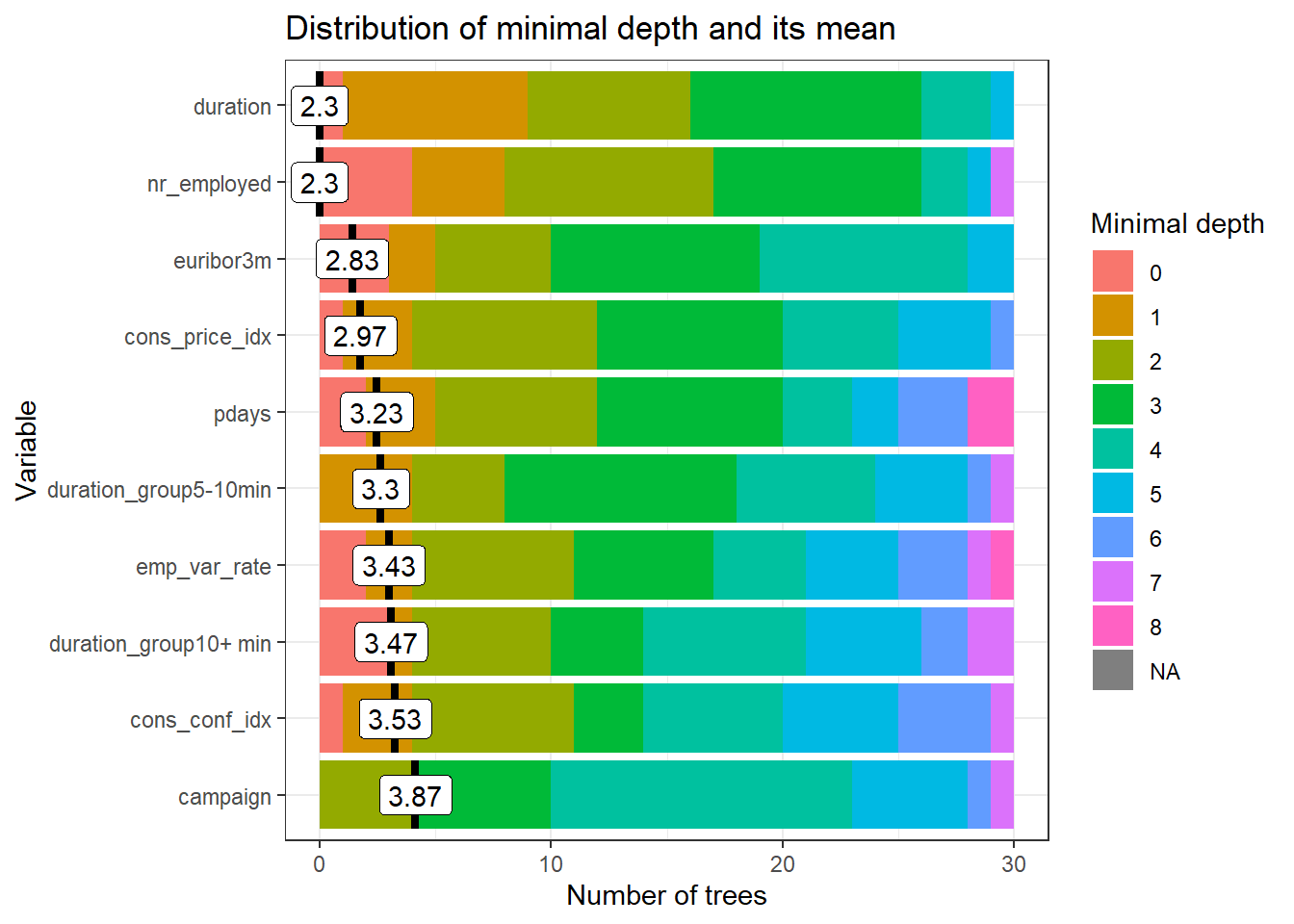


Figure 6. Minimal depth chart for random forest model

Another way to interpret the model is to look at the influence of interactions (see Figure 18). Duration and euribor3m most frequently interacted with other variables, like age, cons\_price\_idx, and campaign, and in fact the minimum depth interaction, and therefore most important one, is cons\_conf\_idx:duration.

## Metrics & Model Comparison

To determine optimal balance between accuracy, sensitivity, and specificity since we had an unbalanced data set, we used different cut-offs for each model. For our simple, complex, and LDA models, the cut-off used was 0.15. For the random forest model, the optimal cut-off used was 0.108. Based on optimal sensitivity with good accuracy and specificity, as well as largest AUC, our random forest model performed the best.

Table . Model comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Sensitivity | Specificity |
| Stepwise | 0.87 | 0.83 | 0.87 |
| Complex | 0.87 | 0.86 | 0.87 |
| LDA | 0.89 | 0.91 | 0.71 |
| Random Forest | 0.84 | 0.94 | 0.83 |

Chart

Description automatically generated

Figure . ROC curves for the top four models.

# Conclusion & Final Recommendations

In conclusion, random forest performed the best likely due to a combination of the benefits of using a Hellinger distance measurement for the splits (as it is optimized for unbalanced data) and the bagging done in this ensemble method. Further investigation is needed on the specifics of Hellinger distance with respect to building trees and using categorical variables.

Further improvement could be made by revisiting the problem with duration. It appears to have the most influence yet logically seems like any model based on it would only be useful to predict conversion after people have already converted. We would encourage rethinking about which data to include in the training set based on that data’s availability about people at the time accurate predictions would be most valuable to the business. A simple experiment of running random forest without duration variables can be found in the appendix.

# Appendix

**Exploring an alternative model without duration…**

We didn’t have time to fully explore this, but as an experiment, a Random Forest model was built that excluded duration and duration\_group entirely, and here are the variables with min depth:

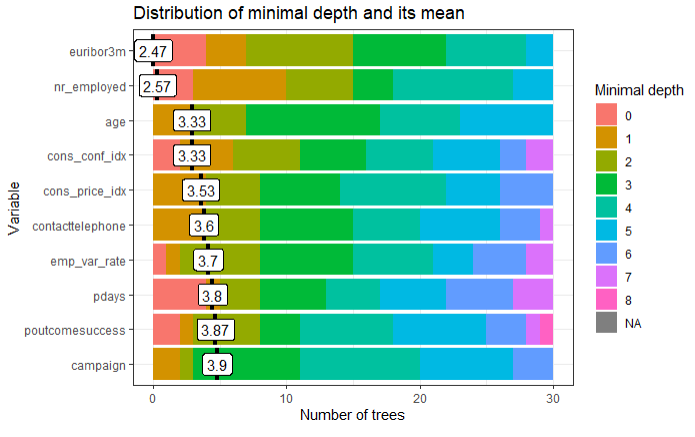


Figure . Most influential observations from random forest model, excluding duration.

If the purpose of the model is to predict which people to contact in the first place based on who is more likely to convert, then this model may be more effective since it seems to be driven mainly by predictors known in advance of any contact.

As for interactions:

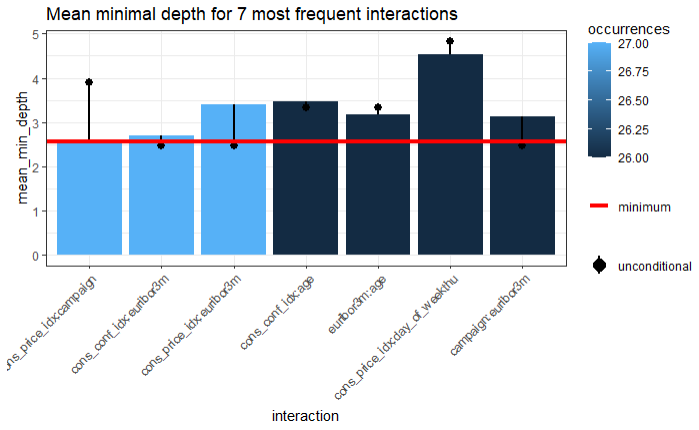


Figure . Top 7 interactions for the random forest model, excluding duration.

Consumer confidence index is now the most influential interaction, this time with campaign.

More work would need to be done to clear away other variables that would be unknown for people who have not yet been contacted, or perhaps were only contacted in the distant past, or for other campaigns. Then we would need to redo this analysis, but unfortunately we are out of time. Still, I wanted to call out this finding and possible next steps.

## Tables & Figures

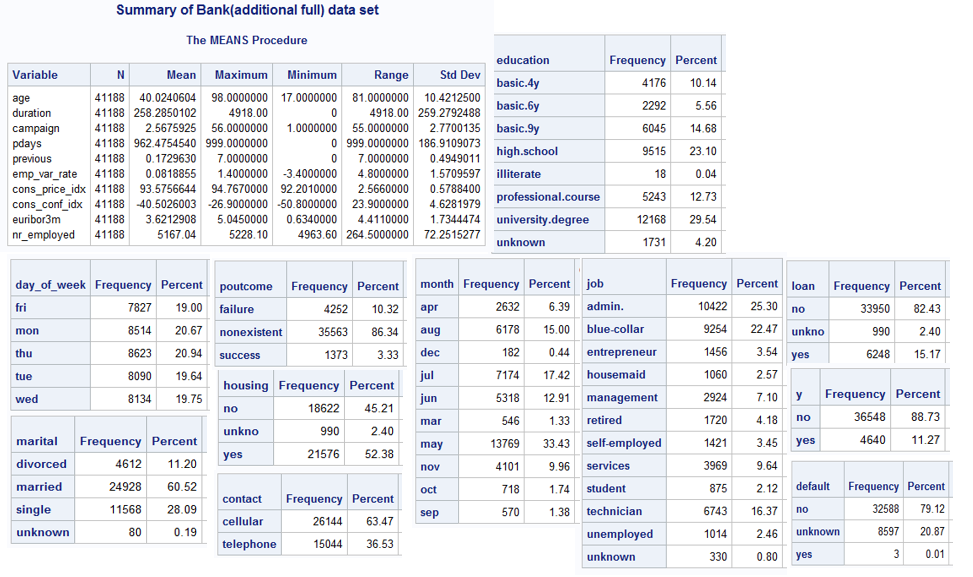


Figure . Summary statistics and frequencies for data set.

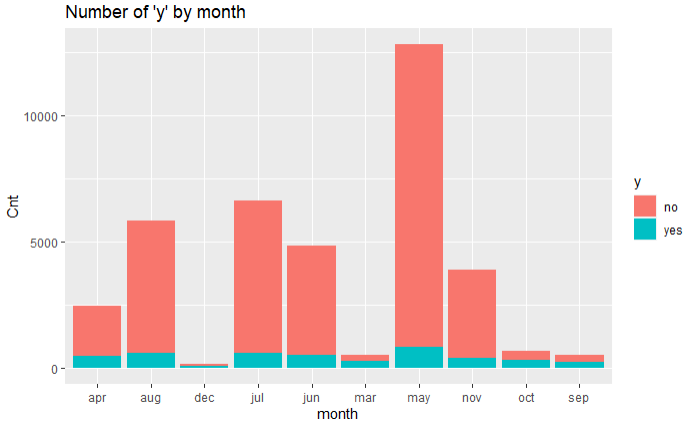


Figure 11. Distribution of months by response variable.

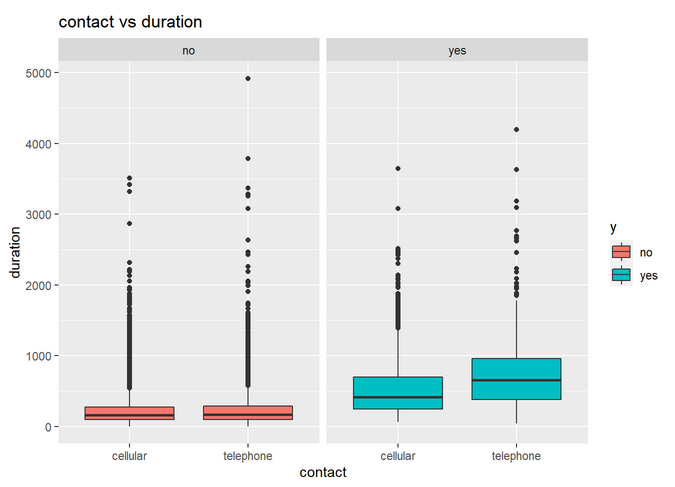
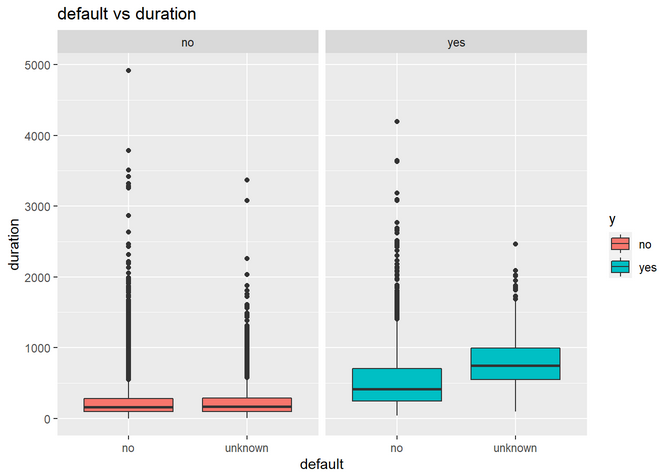
 

Figure . Interaction for contact/duration and default/duration for response variable

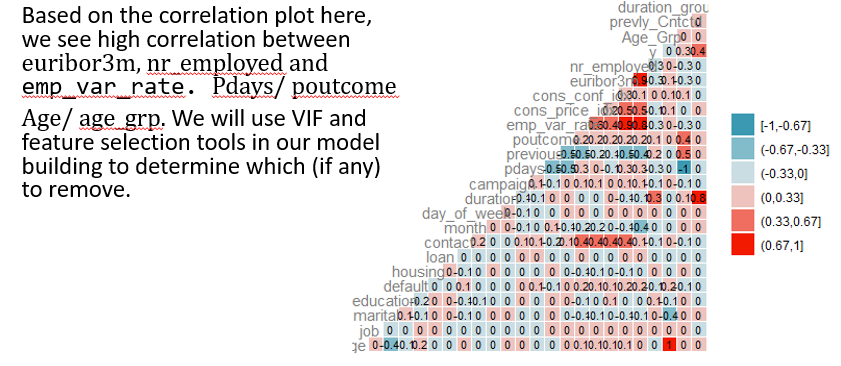


Figure . Correlation matrix for data set.

Graphical user interface

Description automatically generated

Figure . Residuals, Cook’s D, standard deviance and leverage before removing outliers.

Graphical user interface

Description automatically generated

Figure 15. Residuals, Cook’s D, standard deviance and leverage after removing outliers.

Table . Metrics for best simple, interpretable models ran for objective 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Selection | Accuracy | Sensitivity | Specificity |
| Simple | 0.867 | 0.829 | 0.872 |
| Stepwise | 0.867 | 0.827 | 0.872 |
| LASSO | 0.866 | 0.851 | 0.867 |

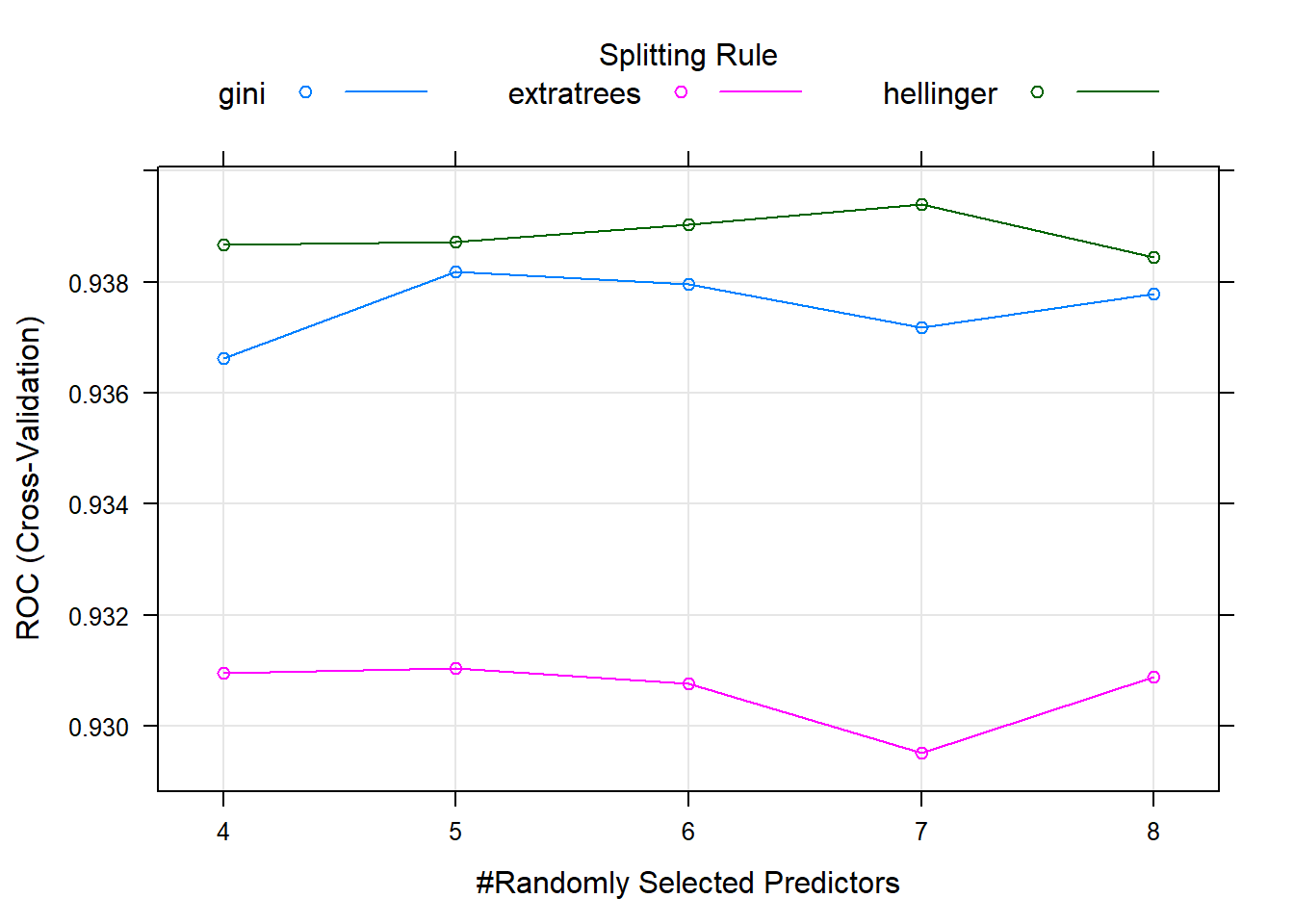


Figure . Random forest performance for each split rule and number of predictors.

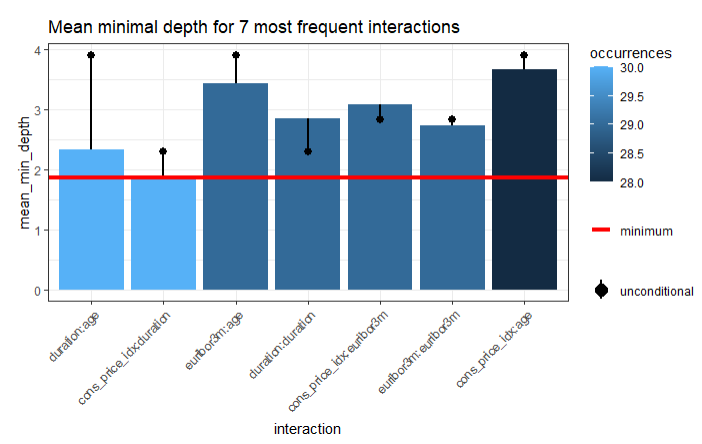


Figure . Top 7 interactions from the random forest model including duration.

Link to Rmarkdown knit file (full notes/comments with graphics): <https://github.com/megball/6372-project-2/blob/main/Rmarkdown/Project2.html>

1. The coefficient of duration is 0.002534 with an associated odds ratio of exp(0.002534) = 1.002, and the units are in seconds, hence a one minute change is equal to an odds ratio of exp(60\*0.0025) = 1.164 [↑](#footnote-ref-1)