

**Project 2:**  
Bank Marketing

**Students**: Matt Farrow, Michael Olheiser, Edward Roske

**Professor**: Dr. Jacob Turner

**Class**: MSDS 6372

**Date**: November 28, 2020

Table of Contents

[Introduction 1](#_Toc56857841)

[Data Description 1](#_Toc56857842)

[Bank Client Data 1](#_Toc56857843)

[Related with the Last Contact of the Current Campaign 1](#_Toc56857844)

[Other Attributes 1](#_Toc56857845)

[Output Variable (Desired Target) 2](#_Toc56857846)

[Exploratory Data Analysis 2](#_Toc56857847)

[Data Cleanup 2](#_Toc56857848)

[Model 1 – Interpretable Model 2](#_Toc56857849)

[Model 2 – Complicated Model 2](#_Toc56857850)

[Model 3 – Discriminant Analysis Model 3](#_Toc56857851)

[Model 4 – Non-parameterized Model 3](#_Toc56857852)

[Overall Conclusion 3](#_Toc56857853)

[Appendix 4](#_Toc56857854)

[SAS Code 4](#_Toc56857855)

[Exploratory Data Analysis 4](#_Toc56857856)

[Model 1 5](#_Toc56857857)

[SAS or R Code – Exploratory Data Analysis 10](#_Toc56857858)

[R Code – Data Cleanup 11](#_Toc56857859)

[SAS Code – Model 1 (Interpretable Model) 12](#_Toc56857860)

[R Code – Model 2 (Complicated model) 13](#_Toc56857861)

[R Code – Model 3 (Discriminant Analysis Model) 14](#_Toc56857862)

[R Code – Model 4 (Non-parametric Model) 15](#_Toc56857863)

# Introduction

The project utilized data from the *UCI Machine Learning Repository* (can be found [here](https://archive.ics.uci.edu/ml/datasets/Bank%20Marketing)). This is data set related to direct marking campaigns of a Portuguese banking institution via phone calls from May 2008 to November 2010. This data provides a wealth of possibly explanatory variables including age, marital status, and job that may correlate to why someone would subscribe to a possible banking offer.

The research team was challenged to develop models to identify which explanatory variables tend to have the biggest impact on whether a subscription was made. The primary purpose of the project was deemed to be explanation not “pre-phone call prediction” so all variables were left in the model to begin with (UCI suggests discarding the duration variable if the primary purpose is prediction).

The team decided to create 5 types of models:

1. Easily interpretable model (in SAS)
2. Complicated model with transformations and interactions (in SAS)
3. Linear/Quadratic Discriminant models (in R)
4. Random Forest model (in R)
5. Model to predict duration then use predicted duration in the final complicated model (in R for prediction then SAS for final model based on #2 above).

# Data Description

The data was provided in multiple files (some of which contained additional explanatory variables). The data set the researchers were given contained 20 explanatory variables related to the response variable, ‘y’ (which we renamed to ‘subscribed’ to be more descriptive). The final data set the team used (after cleaning up data, removing redundancies, and so on) contained 41,188 observations. See “*Data Cleanup*” section below for a deeper discussion of variable observations. Here is a table of variables and their descriptions from the UCI Machine Learning Repository (including additional fields not found in the primary data file):

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type** | **Description** |
| **age** | *numeric* | Age of the customer in years |
| **job** | *factor* | Type of job the customer holds |
| **marital** | *factor* | Customer’s marital status |
| **education** | *factor* | Customer’s education level |
| **default** | *factor* | Whether the customer has credit in default |
| **housing** | *factor* | Whether the customer has a housing loan |
| **loan** | *factor* | Whether the customer has a personal loan |
| **contact** | *factor* | Method of communication |
| **month** | *factor* | Month of most recent contact |
| **day** | *factor* | Day of the week of most recent contact |
| **duration** | *numeric* | Length of most recent contact |
| **campaign** | *numeric* | Number of customer contacts during the campaign |
| **pdays** | *numeric* | Number of days since the customer’s most recent contact for a previous campaign |
| **previous** | *numeric* | Number of times the customer was contacted prior to this campaign |
| **poutcome** | *numeric* | Outcome of the previous campaign |
| **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 (Euro Interbank Offered Rate) 3-month rate (daily indicator) [reference](https://www.euribor-rates.eu/en/) |
| **nr\_employed** | *numeric* | Number of employees (quarterly indicator) |
| **y** | *factor* | Subscription completion |

The data contains a fairly even split of continuous and categorical variables in addition to the dependent response variable ‘y’ (renamed to ‘subscribed’ and hereafter referred to as such). The response variable has two factor levels, *yes* and *no*, that correspond to whether or not a bank customer signed up for a subscription (Figure 6).

# Exploratory Data Analysis

To begin our EDA (Exploratory Data Analysis), the research team began by creating a summary statistic table (Figure 1: Summary Table - All Variables). The table reveals that there are many missing values (NA) in Job, Education, Default, Housing, Loan, Pdays (the only continuous variable with missing data), and Poutcome. The summary statistics also revealed that Default is virtually useless because across the 41,188 observations, only 3 have Default=Yes. The team discarded Default and focused on imputing the remaining missing values (discussed below in “*Data Cleanup*”).

Overall, the Subscribed variable is No=88.7% and Yes=11.3% (Figure 3: Frequency - Subscribed (Class)). The team looked to the frequencies (PROC FREQ in SAS) where a categorical response revealed percentages substantially different than 88.9% (no) and 11.3% (yes). For instance, for the categorical variable Job, at one extreme, 31% of Job=Student subscribed (substantially more than 11.3%) and at the opposite extreme, 7% of Job=Blue-Collar subscribed (substantially less than 11.3%). Both of these jobs are far enough away from the expected values that we would include Job for further evaluation whether or not the Chi-Square test said we should. That said, the team did a Chi-Square test (also from PROC FREQ in SAS) to see which variables correlate to Subscribed (p-value of < 0.05).

Without repeating the entire PROC FREQ tables, the variables that had strong significance levels or seemed to have explanatory values that led to different no/yes percentages were:

* **Categorical**: Job, Education, Contact, Month, Marital, Day\_of\_week, Housing
* **Continuous**: Duration, Campaign, Emp\_var\_rate, Euribor3m, Nr\_employed, Previous

The variables that we rejected (p-value > 0.05 and/or data seemed to be junk/incomplete) were:

* **Categorical**: Default, Loan
* **Continuous**: Age, Pdays

While there was only weak evidence that Cons\_price\_idx and Cons\_conf\_idx were significant, for now, we left those variables in. If they turn out to be insignificant, they will hopefully be discarded by the feature selection.

In terms of multicollinearity, Emp\_var\_rate, Euribor3m, Nr\_employed, Cons\_price\_idx, and Cons\_conf\_idx are all really just measures of what is happening in the outside world at that moment in time. In other words, they coordinate heavily with “what month is it?” The main purpose of this research is explanation, so we will not be creating a Principle Component with all of these combined, but it is expected that the feature selection in model 1 (the interpretable model) which will then be used as the starting point for model 2 (the complex model) will reject the majority of these continuous variables.

Though Model 1 will be using only the untransformed values (since the goal for that model will be ease of interpretability), the team did add in logs and squares of all possible explanatory variables for Model 2. The team calculated the linear correlation of all explanatory variables (original, log transformed, and squared) to produce the correlation calculations (Figure 4: Correlation – Continuous Variables).

The six most *negatively* statistically correlated variables (meaning, variables that as they increase, the response variable is most significantly decreased) are forms of the Euribor3m and Nr\_employed variables. This makes sense for the first because *higher* interest rates (Euribor3m is an interest rate metric) would logically make one less likely to want to deposit money in a Portuguese bank based on a marketing call to solicit deposit subscriptions. It’s less clear why an increase in the number of employed people would make one less likely to subscribe.

Of the top five most *positively* statistically correlated variables, two are related to the duration of the call and three are related to the consumer price index (inflation, basically). Duration clearly makes sense: the longer the person stays on the phone with the telemarketer, the more likely they are to subscribe. Consumer price index is less clear: if inflation is rising, one would tend to want to save less and spend more. This might possibly be an issue in the data set: the CPI is coming out substantially negative and “negative inflation” is rare and in these large of numbers, extremely rare. We will leave in CPI for now and just point out that the data may be incorrect.

**Note**: though these variables are correlated statistically, we cannot say that an increase in any of these variables causes an increase in subscriptions (or vice versa) because this is an observational study. The team will not be calling this out during the rest of the paper, but it’s important to remember.

The team also did some additional data profiling (beginning with Figure 5: Data Profiling Report - Continuous Variables) and it further supports the EDA findings. Based on this exploratory analysis, it is likely that these key variables listed above will end up being present across the majority of the models developed by the team.

# Data Cleanup

The research team used R to cleanup the data that was then used by the five separate models.

## Missing Values

Our first step in cleaning up the data involved replacing placeholder values with *NA;* for most variables this placeholder value was *“unknown”*, but for *pdays* it was *“999”* and for *poutcome* it was *“nonexistent.”* Once all missing values had been identified, we replaced the missing values in each of the categorical variables with their corresponding mode (Table 1). We also made the decision to drop *default* since there were only three *“yes”* values and replace missing values of *pday* with the continuous variable’s median, *“6.”*

## Other Cleanup

Our final cleanup step was to convert all character variables into factors.

## Test & Training Data Sets

In order to build our models off of the same data sets, we built our test and training data sets at the end of our data cleanup process. Two different 75/25 test and training sets were built:

1. The first test/train set was built using all of the available data.
2. The second test/train set was build using a down-sampled set that contained an equal (4,640) number of *“yes”* and *“no”* responses from the *subscribed* variable.

# Model 1 – Interpretable Model

For our interpretable model we began with 19 statistically relevant variables: *job, marital, education, housing, loan, contact, month, day\_of\_week, poutcome, age, duration, campaign, pdays, previous, emp\_var\_rate, con\_price\_idx, cons\_conf\_idx, euribor3m, and nr\_employed*. We knew that many of these would be discarded based on the EDA above, but we left them in for the moment as a sanity check.

We ran a logistic regression (discarding no variables) and based on the Type 3 Analysis of Effects (Figure 10), several variables had p-values > 0.05: *marital* (0.5086), *education* (0.0591), *housing* (0.8357), *loan* (0.4502), *age* (0.5379), *pdays* (0.6220), and *nr\_employed* (0.1329). Note that many of the more egregiously unimportant variables (age, housing, pdays) were previously identified in the EDA as unimportant. We did leave education in the model since it was just outside the alpha of 0.05 and a change in the other variables might make it statistically significant.

## Feature Selection: Stepwise

We then took the remaining variables and tried forward, backward, and stepwise feature selection. Stepwise gave us the highest model performance (versus forward and backward). The final variables left in the model (Figure 15: Logistic Regression (Final Interpretable Model) - Type 3 Analysis of Effects) were *job, education, contact, month, day\_of\_week, poutcome, duration, campaign, previous, emp\_var\_rate, cons\_price\_idx,* and *eruibor3m*.

With this model, the *duration* variable had a extremely high Wald Chi-Square score (greater than all the other variables combined) of 4010.6589 with a p-value < 0.0001 (Figure 10: Logistic Regression (Final Interpretable Model) - Type 3 Analysis of Effects).

## Parameter Interpretation

Figure 1 - Estimates of Coefficients

Our goal is to produce a formula like the following:

Log(odds) =

A logistic regression of the explanatory variables by the response variable (*subscribed*) produced the coefficients (the Betas in the above formula) shown to the right in the column labeled “Estimate”. The intercept is admittedly extrapolation but it shows that if all other explanatory variables were zero (both the negatively and positively correlated continuous variables) or the reference (unemployed for job, Sep for month, etc.), the odds of someone subscribing are e-182.9 = 3.7\*10-80 = a really, really, really small number (but still greater than zero… barely).

Many of the most predictive variables are ones we highlighted in our EDA section including *duration, month*, and *euribor3m* are present in this logistic regression formula for the log(odds).

It is easier to interpret this parameter table if we convert it to odds ratios (found in Figure 19: Logistic Regression (Interpretable) – Odds Ratio Estimates). For odds ratios, the point estimate tells us “for each 1 unit increase in the effect, it is predicted that the odds will increase by a multiplicative factor of the point estimate holding the other variables constant.”

For example, let’s take *Duration*. For each 1 second increase in *Duration*, it is predicted that the odds will increase by a multiplicative factor of 1.005 holding the other variables constant (note that this 95% confidence interval is from 1.005 to 1.005 because of rounding). Confidence intervals for all of the odds ratios can also be found in Figure 19. While 1.005 multiplicative factor may not seem that great, that’s for only a 1 second increase. If we want to determine the odds ratio for one minute (60 seconds), that would be e60\*0.00471 = 1.327 multiplicative factor increase for each additional one minute the conversation lasts.

## Conclusion

The receiver operating characteristic (ROC) curve for this model produced an area under the curve (AUC) of 0.9358 (Figure 11: Logistic Regression (Interpretable) - ROC Curve). Out of interest, we also compared the ROC curve to one containing only the *duration* variable and were able to calculate an AUC value of 0.8184 (Figure 12: Logistic Regression (Interpretable) - ROC Curve (just duration)) showing the dramatic significance of the *duration* variable.

Overall, the logistic regression model generated in this case is easy to interpret, easy to explain (telemarketers: call people on their cell phone generally during the cooler months and keep them on the phone for a while), and has a fairly high accuracy rate. Is it possible to do better? Let’s try complicating the model a bit.

# Model 2 – Complicated Model

We began our complicated logistic regression model with our previously identified statistically relevant variables and added transformations. For *duration, campaign, previous, emp\_var\_rate, cons\_price\_idx, and euribor3m*, if the value was greater than 0, we created a log-transformed version[[1]](#footnote-1) of that variable; we also created quadratic versions[[2]](#footnote-2) of those same variables.

The final model, based on stepwise selection, returned an AUC of 0.9471 (Figure 13) and was defined by the formula:

\*\*\*Edward, CONTINUE HERE.\*\*\*

# Model 3 – Discriminant Analysis Model

The Linear Discriminant Analysis (LDA) model began by selecting the numeric variables along with the *subscribed* variable from the previously down-sampled data, creating training and testing splits, and preprocessing the data using the *“center”* and *“scale”* methods[[3]](#footnote-3). The results of the LDA model showed an accuracy of 82.63% (95% confidence interval: 81.03%, 84.15%) (Figure 5), as well as a sensitivity and specificity of 0.8431 and 0.8095, respectively. The accompanying ROC curve showed an AUC value of 0.9126 (Figure 6).

Interestingly, running the QDA model showed an accuracy of 78.32% (95% confidence interval: 76.59%, 79.98%) (Figure 7), as well as a sensitivity and specificity of 0.9043 and 0.6621, respectively. The accompanying ROC curve showed an AUC value of 0.8997 (Figure 8).

# Model 4 – Non-parameterized Model

\*\*\*Mike, INSERT TEXT HERE. Use Styles for Headings (next level down is Heading 2).\*\*\*

# Model 5 – Predicting Duration Model

\*\*\*Mike, INSERT TEXT HERE. Use Styles for Headings (next level down is Heading 2).\*\*\*

# Overall Conclusion

## AUC Analysis

We used the AUC of the ROC curves as our performance metric for evaluating the various model’s performance. The complex logistic regression model, with an AUC of 0.9471, was the best performing model. However, for interpretability, the interpretable logistic regression model offers marginally less performance with a significantly simpler interpretation.

|  |  |
| --- | --- |
| **Model** | **ROC AUC** |
| Interpretable Logistic Regression | 0.9358 |
| **Complex Logistic Regression** | **0.9471** |
| Linear Discriminant Analysis | 0.9126 |
| Quadratic Discriminant Analysis | 0.8997 |
| K-Nearest Neighbor |  |
| Random Forest  Predicting Duration |  |

## Final Conclusions

The decision to leave *duration* in our models is one that may prove controversial; it is not a variable that can be known beforehand, nor can it be used to predict which clients are most likely prospects. That being said, *duration* was far and away the single greatest predictor of success in our analysis, and that knowledge may prove more crucial to long-term success than client prediction. It is our recommendation that the bank undertake a study of the calls themselves to try and identify whether there are callers, techniques, campaigns, and compounding factors that lend themselves to keeping a prospective client on a call longer.

For the purposes of this analysis, we recommend the interpretable logistic regression model due to its high performance as well as its parameter interpretation.

**OVERALL PAPER (not including title, table of contents, figures, and appendix) should be approximately 7 pages.**

# Appendix

## Exploratory Data Analysis

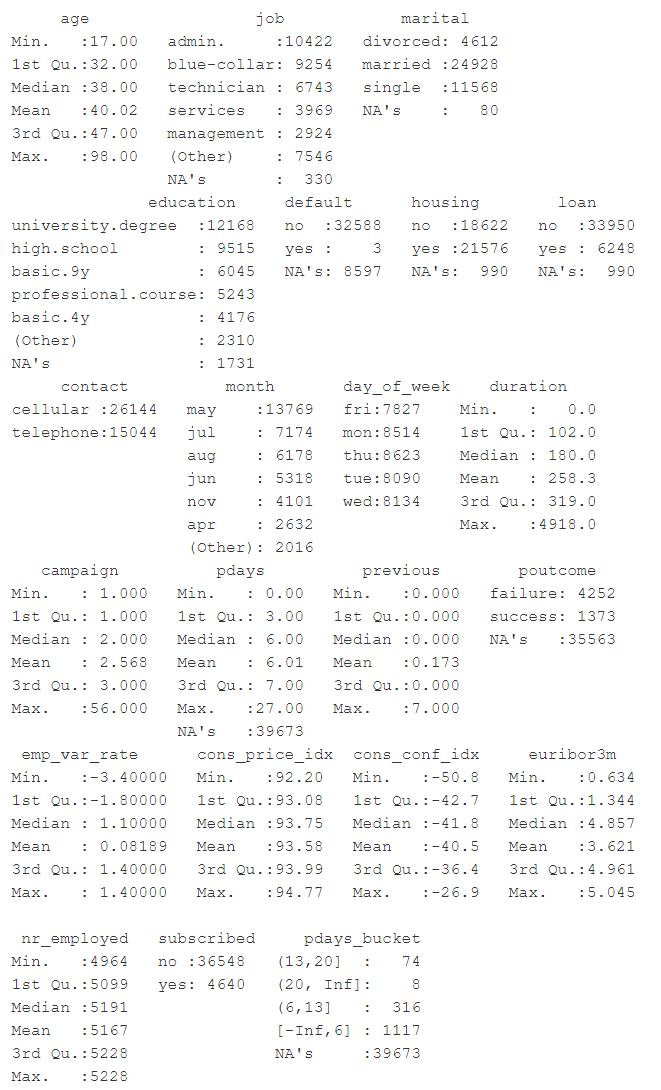


Figure 2: Summary Table - All Variables

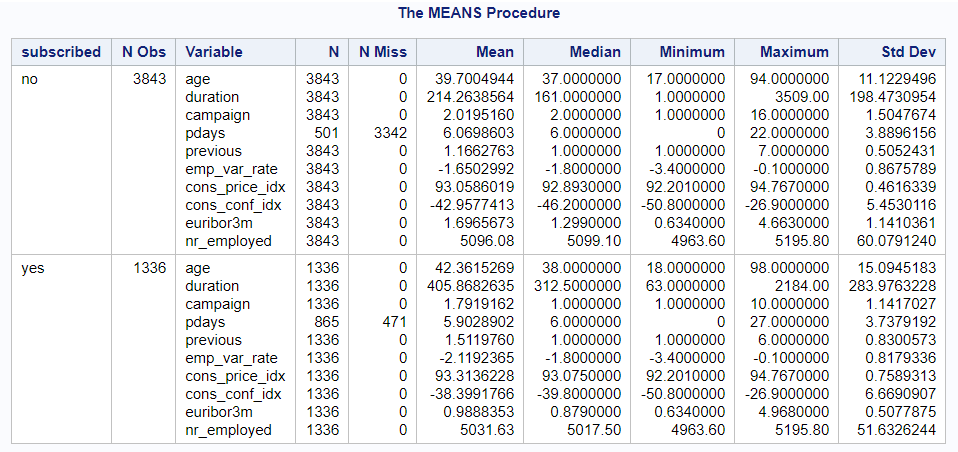


Figure 3: Summary Table - Continuous Variables

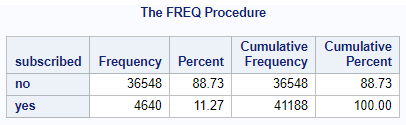


Figure 4: Frequency - Subscribed (Class)

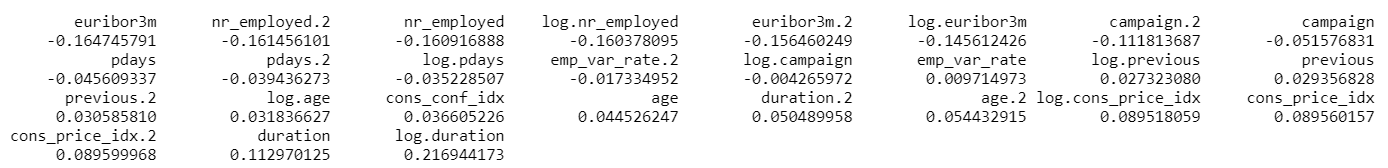


Figure 5: Correlation – Continuous Variables

## Data Profiling Report from R

*Diagram

Description automatically generated*

Figure 6: Data Profiling Report - Continuous Variables

Graphical user interface, diagram

Description automatically generated

Figure 7: Data Profiling Report - Categorical Variables

Chart, scatter chart

Description automatically generated

Figure 8: Data Profiling Report - Covariance Matrix

A picture containing graphical user interface

Description automatically generated

Figure 9: Data Profiling Report - Principal Component Analysis

Chart, line chart

Description automatically generated

Figure 10: Data Profiling Report - QQ Plot

Chart

Description automatically generated

Figure 11: Data Profiling Report - Response Variable Distribution

Table 1: Mode of categorical variables with missing values

|  |  |
| --- | --- |
| Variable | Mode Value |
| job | admin |
| marital | married |
| housing | yes |
| loan | no |
| contact | cellular |
| education | university degree |
| month | may |
| day\_of\_week | thu |
| poutcome | failure |

A picture containing graphical user interface, application, table

Description automatically generated

Figure 12: skimr data summary

Table

Description automatically generated

Figure 13: skimr factor variable report

Table

Description automatically generated

Figure 14: skimr numeric variable report

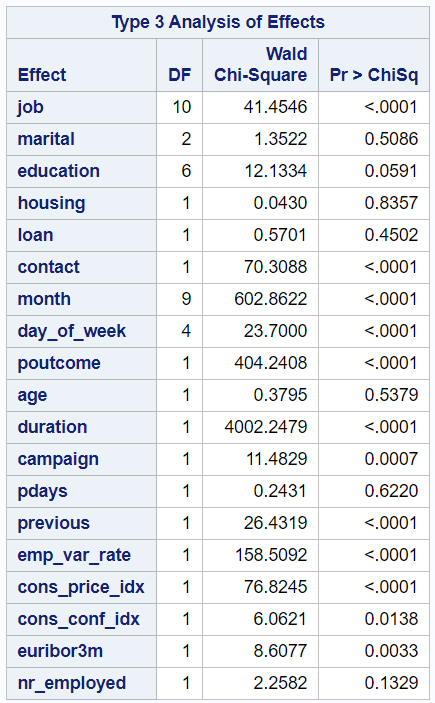


Figure 15: Logistic Regression (All Variables from EDA) - Type 3 Analysis of Effects

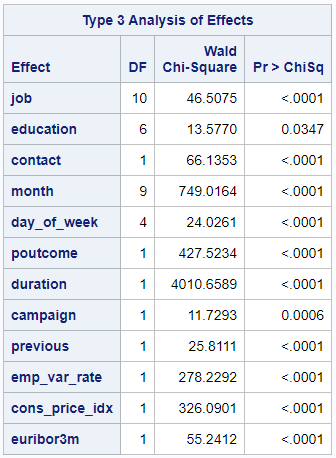


Figure 16: Logistic Regression (Final Interpretable Model) - Type 3 Analysis of Effects

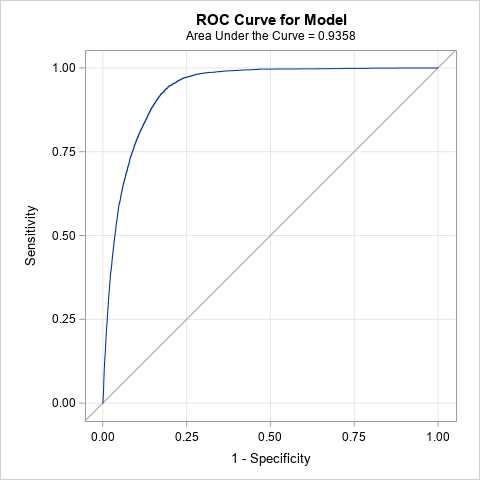


Figure 17: Logistic Regression (Interpretable) - ROC Curve

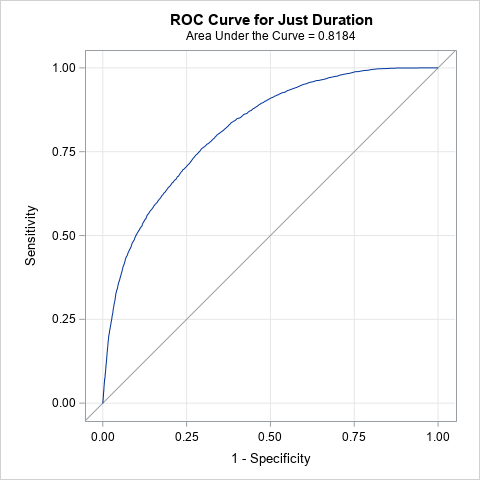


Figure 18: Logistic Regression (Interpretable) - ROC Curve (just duration)

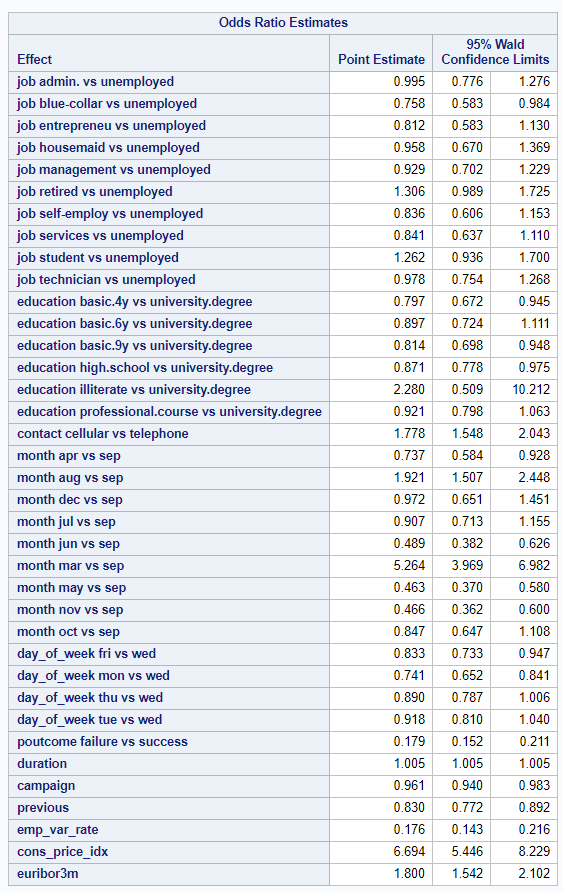


Figure 19: Logistic Regression (Interpretable) – Odds Ratio Estimates

Chart, line chart

Description automatically generated

Figure 19: Logistic Regression (Complicated) - ROC Curve

Chart, line chart

Description automatically generated

Figure 20: Logistic Regression (Complicated) - ROC Comparison

Graphical user interface, text

Description automatically generated

Figure 21: LDA model summary

Graphical user interface, text, application

Description automatically generated

Figure 22: LDA confusion matrix

## Chart, scatter chart Description automatically generated

Figure 23: LDA ROC curve and AUC

Graphical user interface, text, application

Description automatically generated

Figure 24: QDA confusion matrix

Chart, scatter chart

Description automatically generated

Figure 25: QDA ROC curve and AUC

## SAS Code

### Exploratory Data Analysis

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* STUDENT: Edward Roske \*

\* DATE: November 8, 2020 \*

\* CLASS: MSDS6372 - Applied Statistics \*

\* PROJECT: 2 - Logistic Regression \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Header Information;

LIBNAME STATS "/home/u43010517/sasuser.v94/STATS/DS6372";

**RUN**;

TITLE "Logistic Regression";

\* Explore data;

**PROC** **FREQ** DATA=STATS.BANK;

\* default\*subscribed;

TABLES

job\*subscribed

marital\*subscribed

education\*subscribed

housing\*subscribed

loan\*subscribed

contact\*subscribed

month\*subscribed

day\_of\_week\*subscribed

poutcome\*subscribed

/ chisq relrisk;

**RUN**;**QUIT**;

**PROC** **MEANS** DATA=STATS.BANK N NMISS MEAN MEDIAN MIN MAX STD;

\* default;

CLASS subscribed

job

marital

education

housing

loan

contact

month

day\_of\_week

poutcome

;

\* default\*subscribed;

TYPES subscribed

job\*subscribed

marital\*subscribed

education\*subscribed

housing\*subscribed

loan\*subscribed

contact\*subscribed

month\*subscribed

day\_of\_week\*subscribed

poutcome\*subscribed

;

VAR age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed ;

**RUN**;

**PROC** **MEANS** DATA=STATS.BANK N NMISS MEAN MEDIAN MIN MAX STD;

CLASS subscribed

month

;

TYPES subscribed

month\*subscribed

;

VAR Duration Campaign Emp\_var\_rate Euribor3m Nr\_employed Previous ;

**RUN**;

/\*

PROC MEANS DATA=STATS.BANK N NMISS MEAN MEDIAN MIN MAX STD;

CLASS subscribed

month

;

TYPES subscribed

month\*subscribed

;

VAR nr\_employed ;

RUN;

\*/

### Exploratory Data Analysis

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* STUDENT: Edward Roske \*

\* DATE: November 8, 2020 \*

\* CLASS: MSDS6372 - Applied Statistics \*

\* PROJECT: 2 - Logistic Regression \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Header Information;

LIBNAME STATS "/home/u43010517/sasuser.v94/STATS/DS6372";

**RUN**;

TITLE "Logistic Regression";

\* Explore data;

**PROC** **FREQ** DATA=STATS.BANK;

\* default\*subscribed;

TABLES

job\*subscribed

marital\*subscribed

education\*subscribed

housing\*subscribed

loan\*subscribed

contact\*subscribed

month\*subscribed

day\_of\_week\*subscribed

poutcome\*subscribed

/ chisq relrisk;

**RUN**;**QUIT**;

**PROC** **MEANS** DATA=STATS.BANK N NMISS MEAN MEDIAN MIN MAX STD;

\* default;

CLASS subscribed

job

marital

education

housing

loan

contact

month

day\_of\_week

poutcome

;

\* default\*subscribed;

TYPES subscribed

job\*subscribed

marital\*subscribed

education\*subscribed

housing\*subscribed

loan\*subscribed

contact\*subscribed

month\*subscribed

day\_of\_week\*subscribed

poutcome\*subscribed

;

VAR age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed ;

**RUN**;

**PROC** **MEANS** DATA=STATS.BANK N NMISS MEAN MEDIAN MIN MAX STD;

CLASS subscribed

month

;

TYPES subscribed

month\*subscribed

;

VAR Duration Campaign Emp\_var\_rate Euribor3m Nr\_employed Previous ;

**RUN**;

/\*

PROC MEANS DATA=STATS.BANK N NMISS MEAN MEDIAN MIN MAX STD;

CLASS subscribed

month

;

TYPES subscribed

month\*subscribed

;

VAR nr\_employed ;

RUN;

\*/

### Model 1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* STUDENT: Edward Roske \*

\* DATE: November 8, 2020 \*

\* CLASS: MSDS6372 - Applied Statistics \*

\* PROJECT: 2 - Logistic Regression \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Header Information;

LIBNAME STATS "/home/u43010517/sasuser.v94/STATS/DS6372";

**RUN**;

TITLE "Logistic Regression";

\* Logistic Regression;

\* - All Continuous Variables;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "All Continuous Variables";

MODEL subscribed(event='ye') =

age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed

/ SCALE=none aggregate lackfit; \* Can add "influence" but it does too many regression diagnostics;

**RUN**;

\* Logistic Regression;

\* - All Categorical Variables;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "All Categorical Variables";

CLASS subscribed

job

marital

education

housing

loan

contact

month

day\_of\_week

poutcome

;

MODEL subscribed(event='ye') =

job marital education housing loan contact month day\_of\_week poutcome

/ SCALE=none aggregate lackfit; \* Can add "influence" but it does too many regression diagnostics;

**RUN**;

\* Logistic Regression;

\* - All Variables;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "All Continuous & Categorical Variables";

CLASS subscribed

job

marital

education

housing

loan

contact

month

day\_of\_week

poutcome

;

MODEL subscribed(event='ye') =

job marital education housing loan contact month day\_of\_week poutcome

age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed

/ SCALE=none aggregate lackfit; \* Can add "influence" but it does too many regression diagnostics;

**RUN**;

\* Logistic Regression;

\* - Statistically Relevant Variables (based on alpha<0.5 p-values from step above);

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Statistically Relevant Variables";

CLASS subscribed

job

contact

month

day\_of\_week

poutcome

;

MODEL subscribed(event='ye') =

job contact month day\_of\_week poutcome

duration campaign previous emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m

/ SCALE=none aggregate lackfit; \* Can add "influence" but it does too many regression diagnostics;

**RUN**;

\* Logistic Regression with Effect Plots;

\* - Statistically Relevant Variables;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Statistically Relevant Variables";

CLASS subscribed job marital contact month day\_of\_week poutcome ;

MODEL subscribed(event='ye') =

job marital contact month day\_of\_week poutcome

duration previous emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m

/ SCALE=none;

EFFECTPLOT slicefit(sliceby=job plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=marital plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=contact plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=month plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=day\_of\_week plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=poutcome plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=previous plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=cons\_price\_idx plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=cons\_conf\_idx plotby=duration) / noobs;

EFFECTPLOT slicefit(sliceby=euribor3m plotby=duration) / noobs;

\* And a few interesting plots;

EFFECTPLOT slicefit(sliceby=month plotby=cons\_price\_idx) / noobs;

EFFECTPLOT slicefit(sliceby=duration plotby=emp\_var\_rate) / noobs;

**RUN**;

\* Logistic Regression;

\* - All Variables

\* - Feature Selection: Forward;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Feature Selection: Forward";

CLASS subscribed

job marital education housing loan contact month day\_of\_week poutcome;

MODEL subscribed(event='ye') =

job marital education housing loan contact month day\_of\_week poutcome

age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed

/ SELECTION=FORWARD start=**3**

SCALE=none aggregate lackfit;

**RUN**;

\* Logistic Regression;

\* - All Variables

\* - Feature Selection: Backward;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Feature Selection: Backward";

CLASS subscribed

job marital education housing loan contact month day\_of\_week poutcome;

MODEL subscribed(event='ye') =

job marital education housing loan contact month day\_of\_week poutcome

age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed

/ SELECTION=BACKWARD start=**3**

SCALE=none aggregate lackfit;

**RUN**;

\* Logistic Regression;

\* - All Variables

\* - Feature Selection: Stepwise;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Feature Selection: Stepwise";

CLASS subscribed

job marital education housing loan contact month day\_of\_week poutcome;

MODEL subscribed(event='ye') =

job marital education housing loan contact month day\_of\_week poutcome

age duration campaign pdays previous emp\_var\_rate cons\_price\_idx

cons\_conf\_idx euribor3m nr\_employed

/ SELECTION=STEPWISE start=**3**

SCALE=none aggregate lackfit;

**RUN**;

\* Logistic Regression;

\* - Statistically Relevant Variables (based on results of STEPWISE);

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Statistically Relevant Variables";

CLASS subscribed

job education contact month day\_of\_week poutcome ;

MODEL subscribed(event='ye') =

job education contact month day\_of\_week poutcome

duration campaign previous emp\_var\_rate cons\_price\_idx euribor3m

/ SCALE=none aggregate lackfit;

**RUN**;

\* Logistic Regression;

\* - Statistically Relevant Variables (based on results of STEPWISE)

\* - ROC Curves;

**PROC** **LOGISTIC** data=STATS.bank ;

TITLE2 "Statistically Relevant Variables";

CLASS subscribed

job education contact month day\_of\_week poutcome ;

MODEL subscribed(event='ye') =

job education contact month day\_of\_week poutcome

duration campaign previous emp\_var\_rate cons\_price\_idx euribor3m

/ SCALE=none aggregate lackfit;

ROC 'Just Duration'

duration

;

ROC 'Just Education'

education

;

ROC 'Random Chance'

;

ROCCONTRAST / estimate e;

**RUN**;

\* Logistic Regression;

\* - Statistically Relevant Variables (based on results of STEPWISE)

\* - ROC Curve;

**PROC** **LOGISTIC** data=STATS.bank plots(only)=roc;

TITLE2 "Model 1: ROC Curve";

CLASS subscribed

job education contact month day\_of\_week poutcome ;

LogisticModel: MODEL subscribed(event='ye') =

job education contact month day\_of\_week poutcome

duration campaign previous emp\_var\_rate cons\_price\_idx euribor3m;

\*/ SCALE=none aggregate lackfit;

OUTPUT out=LogiOut predicted=LogiPred; /\* output predicted value, to be used later if we want to see the predictions \*/

**RUN**;

### Model 2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* STUDENT: Edward Roske \*

\* DATE: November 23, 2020 \*

\* CLASS: MSDS6372 - Applied Statistics \*

\* PROJECT: 2 - Logistic Regression \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Header Information;

LIBNAME STATS "/home/u43010517/sasuser.v94/STATS/DS6372";

**RUN**;

TITLE "Model 2 - More Complicated Model";

\* Add transformations;

\* - Begin with previously relevant variables;

**DATA** STATS.BANKnew

(DROP=age marital housing loan pdays cons\_conf\_idx nr\_employed);

SET STATS.BANK;

\* Logarithms;

if duration gt **0** then durationLog=log(duration);

if campaign gt **0** then campaignLog=log(campaign);

if previous gt **0** then previousLog=log(previous);

if emp\_var\_rate gt **0** then emp\_var\_rateLog=log(emp\_var\_rate);

if cons\_price\_idx gt **0** then cons\_price\_idxLog=log(cons\_price\_idx);

if euribor3m gt **0** then euribor3mLog=log(euribor3m);

\* Quadratics;

duration2=duration\*\***2**;

campaign2=campaign\*\***2**;

previous2=previous\*\***2**;

emp\_var\_rate2=emp\_var\_rate\*\***2**;

cons\_price\_idx2=cons\_price\_idx\*\***2**;

euribor3m2=euribor3m\*\***2**;

\*\*\* Will add interactions later in model code;

**RUN**;

\* Logistic Regression;

\* - All Variables including new transformations;

\* - Had to leave out previousLog and emp\_var\_rateLog because of too many NAs;

**PROC** **LOGISTIC** data=STATS.BANKnew ;

TITLE2 "All Continuous & Categorical Variables";

CLASS subscribed

job education contact month day\_of\_week poutcome;

MODEL subscribed(event='ye') =

job education contact month day\_of\_week poutcome

duration campaign previous emp\_var\_rate cons\_price\_idx euribor3m

duration2 campaign2 previous2 emp\_var\_rate2 cons\_price\_idx2 euribor3m2

durationLog campaignLog cons\_price\_idxLog euribor3mLog

/ SCALE=none aggregate lackfit; \* Can add "influence" but it does too many regression diagnostics;

**RUN**;

\* All variables are still statistically impactful to the model;

\* - Campaign is less relevant but still alpha < 0.05;

\* Logistic Regression;

\* - Add interactions;

\* - All Variables including new transformations;

\* - Had to leave out previousLog and emp\_var\_rateLog because of too many NAs;

**PROC** **LOGISTIC** data=STATS.BANKnew plots(only)=roc;

TITLE2 "All Continuous & Categorical Variables with Interactions";

CLASS subscribed

job education contact month day\_of\_week poutcome

;

MODEL subscribed(event='ye') =

job education contact month day\_of\_week month day\_of\_week poutcome

job\*education job\*education\*month month\*day\_of\_week month\*poutcome

campaign | previous | emp\_var\_rate | cons\_price\_idx | euribor3m

duration campaign previous emp\_var\_rate cons\_price\_idx euribor3m

duration2 campaign2 previous2 emp\_var\_rate2 cons\_price\_idx2 euribor3m2

durationLog campaignLog cons\_price\_idxLog euribor3mLog

/ SCALE=none aggregate lackfit; \* Can add "influence" but it does too many regression diagnostics;

ROC 'Random Chance'

;

ROCCONTRAST / estimate e;

**RUN**;

\* Logistic Regression;

\* - All Variables p-value > 0.05 (from last step)

\* - Feature Selection: Stepwise;

**PROC** **LOGISTIC** data=STATS.BANKnew ;

TITLE2 "Variables: Stepwise Selection";

CLASS subscribed

job education contact month day\_of\_week poutcome

;

StepwiseModel: MODEL subscribed(event='ye') =

education contact month day\_of\_week month day\_of\_week poutcome

job\*education job\*education\*month month\*day\_of\_week month\*poutcome

previous\*cons\_price\_idx emp\_var\_rate\*euribor3m cons\_price\_idx\*euribor3m

emp\_var\_rate\*cons\_price\_idx\*euribor3m

duration previous cons\_price\_idx euribor3m

duration2 campaign2 previous2 emp\_var\_rate2 cons\_price\_idx2 euribor3m2

durationLog cons\_price\_idxLog euribor3mLog

/ SELECTION=STEPWISE start=**3**

SCALE=none aggregate lackfit;

ROC 'All variables'

education contact month day\_of\_week month day\_of\_week poutcome

job\*education job\*education\*month month\*day\_of\_week month\*poutcome

previous\*cons\_price\_idx emp\_var\_rate\*euribor3m cons\_price\_idx\*euribor3m

emp\_var\_rate\*cons\_price\_idx\*euribor3m

duration previous cons\_price\_idx euribor3m

duration2 campaign2 previous2 emp\_var\_rate2 cons\_price\_idx2 euribor3m2

durationLog cons\_price\_idxLog euribor3mLog

;

ROCCONTRAST / estimate e;

**RUN**;

\* ROC for stepwise variables: 0.9349

\* ROC for stepwise if I leave in job\*education\*month: 0.9401

\* ROC for all statistically relevant variables: 0.9471

\* Logistic Regression;

\* - Statistically Relevant Variables (based on results of STEPWISE);

**PROC** **LOGISTIC** data=STATS.BANKnew plots(only)=roc;

TITLE2 "Final Stepwise Model";

CLASS subscribed

job education month poutcome ;

FullModel: MODEL subscribed(event='ye') =

education month poutcome job\*education duration duration2 durationLog euribor3mLog

job\*education\*month

/ SCALE=none aggregate lackfit;

ROC 'Stepwise (no job\*education\*month)'

education month poutcome job\*education duration duration2 durationLog euribor3mLog

job\*education\*month

;

ROC 'Only job\*education\*month'

job\*education\*month

;

ROC 'Random Chance'

;

ROCCONTRAST / estimate e;

**RUN**;

\* Final variables in the more complex model:

education month poutcome duration

duration2 durationLog euribor3mLog

job\*education

job\*education\*month

## R Code

knitr::opts\_chunk$set(echo = TRUE)

knitr::opts\_chunk$set(**warning** = FALSE)

knitr::opts\_chunk$set(message = FALSE)

**library**(tidyverse) *# general data processing & plotting*

**library**(here) *# relative location references*

**library**(janitor) *# data cleanup tools*

**library**(naniar) *# dealing with missing values*

**library**(caret) *# misc functions for training and plotting classification and regression models*

**library**(tidymodels)

**library**(GGally) *# for ggpairs*

**library**(MASS) *# for LDA/QDA*

*# Read data*

bank <- read\_delim(here("data - raw", "bank-additional", "bank-additional-full.csv"),

";", escape\_double = FALSE, trim\_ws = TRUE)

*# Clean column names*

bank <- clean\_names(bank)

*# Quick examination of the data*

glimpse(bank)

head(bank)

*# Begin by replacing missing value placeholders with NA*

bank\_clean <- bank %>%

replace\_with\_na\_all(condition = ~ .x == "unknown") %>%

replace\_with\_na(replace = list(pdays = 999)) %>%

replace\_with\_na(replace = list(poutcome = "nonexistent"))

*# Replace each missing value in a categorical variable with the mode as*

*# determined during SAS EDA*

bank\_clean <- bank\_clean %>%

mutate(job = replace\_na(job, "admin."),

marital = replace\_na(marital, "married"),

housing = replace\_na(housing, "yes"),

loan = replace\_na(loan, "no"),

contact = replace\_na(contact, "cellular"),

education = replace\_na(education, "university.degree"),

month = replace\_na(month, "may"),

day\_of\_week = replace\_na(day\_of\_week, "thu"),

poutcome = replace\_na(poutcome, "failure"))

*# Drop `default` column since there are only 3 "Yes" values*

bank\_clean <- bank\_clean %>%

dplyr::select(-default)

*# Only Pdays is missing, and we should delete it, but for completeness, replace*

*# with median (6)*

bank\_clean <- bank\_clean %>%

mutate(pdays = replace\_na(pdays, 6))

*# Rename response variable*

bank\_clean <- bank\_clean %>%

rename(subscribed = y)

*# Convert all character variables to factors*

bank\_clean <- bank\_clean %>%

mutate(across(where(is\_character),as\_factor))

*# Re-run reports*

DataExplorer::create\_report(bank\_clean)

Hmisc::describe(bank\_clean)

psych::describe(bank\_clean)

skimr::skim(bank\_clean)

*# Reports from inspectdf (categorical variables and Pearson correlation*

*# coefficients)*

bank\_cat <- inspectdf::inspect\_cat(bank\_clean)

bank\_pear <- inspectdf::inspect\_cor(bank\_clean)

*# Set seed*

set.seed(123)

*# Save full bank\_clean data set*

*# write\_csv(bank\_clean, here("data - output", "bank\_clean.csv"))*

*# Create test/train data sets of full data*

inTraining <- createDataPartition(bank\_clean$subscribed, p = .75, list = FALSE)

training <- bank\_clean[ inTraining,]

testing <- bank\_clean[-inTraining,]

*# Save bank\_clean test/train sets*

*# write\_csv(training, here("data - output", "bank\_clean\_train.csv"))*

*# write\_csv(testing, here("data - output", "bank\_clean\_test.csv"))*

*# Down-sample the data to get an equal number of yes and no responses*

bank\_clean\_ds <- downSample(x = bank\_clean[, 1:19],

y = bank\_clean$subscribed)

*# Rename `Class` to `subscribed`*

bank\_clean\_ds <- bank\_clean\_ds %>%

rename(subscribed = Class)

*# Double-check the number of yes/no responses*

bank\_clean\_ds %>%

count(subscribed)

*# Save downsampled data set*

*# write\_csv(bank\_clean\_ds, here("data - output", "bank\_clean\_downsampled.csv"))*

*# Create test/train data sets of downsampled data*

inTraining <- createDataPartition(bank\_clean\_ds$subscribed, p = .75, list = FALSE)

training <- bank\_clean\_ds[ inTraining,]

testing <- bank\_clean\_ds[-inTraining,]

*# Save downsampled data test/train sets*

*# write\_csv(training, here("data - output", "bank\_clean\_ds\_train.csv"))*

*# write\_csv(testing, here("data - output", "bank\_clean\_ds\_test.csv"))*

set.seed(123)

*# Create LDA data set*

lda\_data <- bank\_clean\_ds %>%

dplyr::select(c(1, 10:13, 15:17, 19:20)) *# class + numeric columns + drop euribor3m*

*# Check for collinearity*

view(inspectdf::inspect\_cor(lda\_data))

*# Create test/train data sets of downsampled data*

inTraining <- createDataPartition(lda\_data$subscribed, p = .75, list = FALSE)

training <- lda\_data[ inTraining,]

testing <- lda\_data[-inTraining,]

*# Estimate preprocessing parameters*

preproc\_parameter <- training %>%

preProcess(method = c("center", "scale"))

*# Transform the data using the estimated parameters*

train\_transform <- preproc\_parameter %>% predict(training)

test\_transform <- preproc\_parameter %>% predict(testing)

*# Fit the model*

lda\_model <- lda(subscribed ~ ., data = train\_transform)

*# Look at the model*

lda\_model

*# Make predictions*

lda\_predictions <- lda\_model %>% predict(test\_transform)

*# Model accuracy*

mean(lda\_predictions$class == test\_transform$subscribed)

*# Generate confusion matrix*

confusionMatrix(predict(lda\_model, newdata = test\_transform)$class,

test\_transform$subscribed)

*# Get the posteriors as a tibble*

lda\_posteriors <- tibble(predict(lda\_model, newdata = test\_transform)$posterior)

*# Create prediction and performance values*

**library**(ROCR)

lda\_predict <- prediction(lda\_predictions$posterior[,2], test\_transform$subscribed)

lda\_perf <- performance(lda\_predict, "tpr", "fpr")

*# Calculate AUC*

lda\_auc <- performance(lda\_predict, measure = "auc")

lda\_auc <- lda\_auc@y.values[[1]]

lda\_auc

*# Create and plot ROC curve for LDA model*

**library**(ggfortify)

autoplot(lda\_perf) +

labs(title = "LDA Model",

subtitle = "True Positive Rate vs. False Positive Rate",

x = "False Positive Rate",

y = "True Positive Rate") +

annotate(

geom = "text", x = 0.50, y = 0.50,

label = paste("AUC =", round(lda\_auc, 4)), size = 6

) +

theme\_minimal()

*# QDA*

qda\_model <- qda(subscribed ~ ., data = train\_transform)

*# Make predictions*

qda\_predictions <- qda\_model %>% predict(test\_transform)

*# Model accuracy*

mean(qda\_predictions$class == test\_transform$subscribed)

*# Generate confusion matrix*

confusionMatrix(predict(qda\_model, newdata = test\_transform)$class,

test\_transform$subscribed)

*# Get the posteriors as a tibble*

qda\_posteriors <- tibble(predict(qda\_model, newdata = test\_transform)$posterior)

*# Create prediction and performance values*

qda\_predict <- prediction(qda\_predictions$posterior[,2], test\_transform$subscribed)

qda\_perf <- performance(qda\_predict, "tpr", "fpr")

*# Calculate AUC*

qda\_auc <- performance(qda\_predict, measure = "auc")

qda\_auc <- qda\_auc@y.values[[1]]

qda\_auc

*# Create and plot ROC curve for LDA model*

autoplot(qda\_perf) +

labs(title = "QDA Model",

subtitle = "True Positive Rate vs. False Positive Rate",

x = "False Positive Rate",

y = "True Positive Rate") +

annotate(

geom = "text", x = 0.50, y = 0.50,

label = paste("AUC =", round(qda\_auc, 4)), size = 6

) +

theme\_minimal()

1. Identified by *“Log”* at the end of the original variable’s name. [↑](#footnote-ref-1)
2. Identified by *“2”* at the end of the original variable’s name. [↑](#footnote-ref-2)
3. The methods in this case came from the `caret` package in R. *“Center”* subtracts the mean of the predictor’s data from the predictor values, and *“scale”* divides by the standard deviation. [↑](#footnote-ref-3)