

**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

# Data Description

The data was provided by the UCI Machine Learning Repository and 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.

|  |  |  |
| --- | --- | --- |
| 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\_of\_week** | *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 |

# 

# Exploratory Data Analysis

In reviewing the original data, we see that there is a fairly even split of continuous and categorical variables in addition to the dependent variable *y*. This dependent variable has two levels, *yes* and *no*, that correspond to the whether or not a bank customer signed up for a subscription. Because the original variable name was not very descriptive, we made the decision to rename it to *subscribed*.

# Data Cleanup

## 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 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

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

# Model 2 – Complicated Model

Someone, INSERT TEXT HERE. Use Styles for Headings (next level down is Heading 2).*As a reminder, this model should contain interactions, feature engineering, transformations, and/or polynomials.*

# Model 3 – Discriminant Analysis Model

The Linear Discriminant Analysis (LDA) model began with selecting the numeric variables along with the *subscribed* variable from the down-sampled data and preprocessing the data using the *“center”* and *“scale”* methods. The results of the LDA model showed an accuracy of 82.63% (95% confidence interval: 81.03%, 84.15%) (Figure 5). The ROC curve showed an AUC value of 0.9126 (Figure 6).

The QDA model showed an accuracy of 78.32% (95% confidence interval: 76.59%, 79.98%) (Figure 7). The ROC curve showed an AUC value of 0.8997 (Figure 8).

# Model 4 – Non-parameterized Model

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

# Overall Conclusion

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

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

# Appendix

Table : 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 : skimr data summary

Table

Description automatically generated

Figure : skimr factor variable report

Table

Description automatically generated

Figure : skimr numeric variable report

Graphical user interface, text

Description automatically generated

Figure : LDA model summary

Graphical user interface, text, application

Description automatically generated

Figure : LDA confusion matrix

## Chart, scatter chart Description automatically generated

Figure : LDA ROC curve and AUC

Graphical user interface, text, application

Description automatically generated

Figure : QDA confusion matrix

Chart, scatter chart

Description automatically generated

Figure : 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";

\*/ Matt's code to run the file on his machine /\*

\*PROC IMPORT REPLACE DATAFILE="Y:\R\6372-project-2\data - \bank\_clean.csv"

\* DBMS=CSV

\* OUT=Bank;

\* GETNAMES=YES;

\*RUN;

\* 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**;

## SAS or R Code – Exploratory Data Analysis

|  |
| --- |
|  |
|  |

## R Code – Data Cleanup

|  |
| --- |
|  |

## SAS Code – Model 1 (Interpretable Model)

|  |
| --- |
|  |

## R Code – Model 2 (Complicated model)

|  |
| --- |
|  |

## R Code – Model 3 (Discriminant Analysis Model)

|  |
| --- |
| R Code – Model 4 (Non-parametric Model) |