Deliverable 3

Numeric and Binary targets Forecasting Models

Júlia Gasull i Claudia Sánchez

December 4, 2020

Contents

1 First setups

```
if(!is.null(dev.list())) dev.off()  # Clear plots
rm(list=ls())                       # Clean workspace
```

1.1 Load Required Packages for this deliverable

We load the necessary packages and set working directory

```
setwd("~/Github Repositories/FIB-ADEI-LAB/deliverable3")
filepath<-"~/Github Repositories/FIB-ADEI-LAB/deliverable3"
#setwd("C:/Users/Claudia Sánchez/Desktop/FIB/TARDOR 2020-2021/ADEI/DELIVERABLE1/FIB-ADEI-LAB/deliverabl
#filepath<-"C:/Users/Claudia Sánchez/Desktop/FIB/TARDOR 2020-2021/ADEI/DELIVERABLE1/FIB-ADEI-LAB/delive

# Load Required Packages
options(contrasts=c("contr.treatment","contr.treatment"))
requiredPackages <- c("missMDA","chemometrics","mvoutlier","effects","FactoMineR","car", "factoextra","
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]
if(length(missingPackages)) install.packages(missingPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

1.2 Load processed data from last deliverable

```
load(paste0(filepath,"/Taxi5000_del2.RData"))
```

1.3 Some useful information

1.3.1 Y (Numeric Target).

This variable will be the target for linear model building (connected to blocks Statistical Modeling I and II).

1.3.2 Explicative Variables for modeling purposes are generally (not all in this dataset):

- Socioeconomic variables: gender, age, education, type of work, etc

- Trip characteristics, etc.
- Bank marketing history
- Economic vars

1.3.3 Multivariant Analysis conducted in previous deliverables has to be used to select the initial model. Students have some degrees in freedom in model building, but the following conditions are requested:

- At least two numerical variables have to be considered as explicative variables for initial steps in model building, called covariates. Non-linear models have to be checked for consistency.
- Select the most significant factors found in Multivariant Data Analysis as initial model factors. Put some reasonable limits to initial model complexity.
- **You have to consider at least one interaction between a couple of factors and one interaction between factor and covariate.**
- Diagnostics of the final model have to be undertaken. Lack of fit observations and influence data have to be selected and discussed (connections to multidimensional outliers in Multivariant Data Analysis is highly valuable)

1.3.4 Outcome/Target : A binary response variable (Binary Target) will be the response variable for Binary Regression Models included in Statistical Modeling Part III.

- Explicative Variables for modeling purposes are those available in dataset, exceptions will be indicated, if any.

- Multivariant Analysis conducted in previous deliverables has to be used to select the initial model. Students have some degrees in freedom in model building, but the following conditions are requested:

  - Split the sample in work and test samples (consisting on a 80-20 split). Working data frame has to be used for model building purposes.
  - At least two numerical variables have to be considered as explicative variables for initial steps in model building.
  - Select the most significant factors according to feature selection as initial model factors. Put some reasonable limits to initial model complexity.
  - **You have to consider at least one interaction between a couple of factors and one interaction between factor and covariate.**
  - Diagnostics of the final model have to be undertaken. Lack of fit observations and influence data have to be selected and discussed (connections to multidimensional outliers in Multivariant Data Analysis is highly valuable).
  - You have to predict Y (Binary Target) in the Working Data Frame vs the rest according to the best validated model that you can find and make a confusion matrix.
  - Make a confusion matrix in the Testing Data Frame for **Y (Binary Target)** according to the best validated model found.

1.3.5 Confusion Matrix:

When referring to the performance of a classification model, we are interested in the model's ability to correctly predict or separate the classes. When looking at the errors made by a classification model, the confusion matrix gives the full picture. Consider e.g. a three class problem with the classes A, and B. The confusion matrix shows how the predictions are made by the model. The rows correspond to the known class of the data, i.e. the labels in the data. The columns correspond to the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column for examples with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.