Working Title: Predicting Customer Behavior Using a Portuguese Financial Institution Dataset

Antonio Flores

2024-06-05

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

# 1. Summary/Abstract

This project will seek to identify a model for predicting consumer financial behavior using a dataset from a Portuguese Bank.

# 2. Introduction

## 2.1 General Background Information

Providing businesses with a model that will allow prioritization of consumers could save time and allow for a more accurate use of resources.

## 2.2 Description of data and data source

The data was donated on 2/13/2012. It was collected from phone call marketing campaigns performed by a Portuguese banking institution.I have accessed this data from the UC Irvine Machine Learning Repository.

There are 45,212 records, and includes: age, marital status, job, education,details related to the phone call, as well as answers related to questions about past credit history. Additionally, the classification variable is whether or not the person subscribed to a term deposit. There are 16 features in total.

## 2.3 Questions/Hypotheses to be addressed

The research question I plan to address with my analysis is which features or combination of features are the best predictors of consumers making a deposit. The desired output of this analysis is a model which allows a financial institution to better prioritize/make decisions regarding future marketing campaigns. I don’t currently have specific predictors in mind to begin with, as I plan to test all of the available features while also looking at some combinations or concatenation of features.

*To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:*

*Examples of reproducible research projects can for instance be found in (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020).*

# 3. Methods

I want to first perform some EDA to identify if there are any variables with extreme variation or conversely if there some patterns that can be seen already. Then I will test for correllation between different features and proceed from there.

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). **?@fig-schematic** is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

## 3.2 Data aquisition

Data retrieved from UCI ML Repository

## 3.3 Data import and cleaning

### 3.3.1 Reading in the Data

raw = read.csv(“C:/Users/Client/Documents/Flores-P2-Project/bank-full.csv”,header = TRUE, sep=“;”) head(raw)

|  |
| --- |
| Table 1: Data Snapshot  age job marital education default balance housing loan contact day 1 58 management married tertiary no 2143 yes no unknown 5 2 44 technician single secondary no 29 yes no unknown 5 3 33 entrepreneur married secondary no 2 yes yes unknown 5 4 47 blue-collar married unknown no 1506 yes no unknown 5 5 33 unknown single unknown no 1 no no unknown 5 6 35 management married tertiary no 231 yes no unknown 5  month duration campaign pdays previous poutcome y 1 may 261 1 -1 0 unknown no 2 may 151 1 -1 0 unknown no 3 may 76 1 -1 0 unknown no 4 may 92 1 -1 0 unknown no 5 may 198 1 -1 0 unknown no 6 may 139 1 -1 0 unknown no |

### 3.3.2 Dimensions:

dim(raw)

|  |
| --- |
| Table 2: Data Dimensions  [1] 45211 17 |

### 3.3.3 Describing raw data

str(raw)

|  |
| --- |
| Table 3: Data Description  'data.frame': 45211 obs. of 17 variables:  $ age : int 58 44 33 47 33 35 28 42 58 43 ...  $ job : chr "management" "technician" "entrepreneur" "blue-collar" ...  $ marital : chr "married" "single" "married" "married" ...  $ education: chr "tertiary" "secondary" "secondary" "unknown" ...  $ default : chr "no" "no" "no" "no" ...  $ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...  $ housing : chr "yes" "yes" "yes" "yes" ...  $ loan : chr "no" "no" "yes" "no" ...  $ contact : chr "unknown" "unknown" "unknown" "unknown" ...  $ day : int 5 5 5 5 5 5 5 5 5 5 ...  $ month : chr "may" "may" "may" "may" ...  $ duration : int 261 151 76 92 198 139 217 380 50 55 ...  $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  $ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...  $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  $ poutcome : chr "unknown" "unknown" "unknown" "unknown" ...  $ y : chr "no" "no" "no" "no" ... |

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 4](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 1](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 1: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 5](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, et al., 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>