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

Antonio Flores

2024-06-22

# 1. Summary/Abstract

This project will seek to identify a model for predicting consumer financial behavior using a dataset from a Portuguese Bank. After the data has been cleaned and prepared for analysis, exploratory data analysis will be performed to gather more information regarding the shape, size, and behavior of different variables, to gauge their usefulness in a prediction model. Additionally, several different statistical tests will be implemented to identify which classes within different variables affect the outcome variable. Finally (*haven’t started on this yet*), with some additional data preprocessing, this project will test different machine learning models and conclude with the resulting findings.

# 2. Introduction

## 2.1 General Background Information

Providing businesses with a model that will allow prioritization of consumers and/or demographics has great potential in improving resource management and future marketing campaigns, as well as increasing efficient spending.

## 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 17 features in total.

Among the variables are a handful of features relating to the marketing campaign itself. For example, included are the day, month, and duration of the call, the number of contacts performed during the campaign (campaign), the number of days since the client was last contacted (pdays), the number of contacts performed before this campaign (previous), and finally the outcome of precious marketing campaigns (poutcome).

## 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. Currently, I plan to investigate all variables, but I am specifically interested in both job type, education and age.

# 3. Methods

Cleaning -> This includes converting numeric variables to factors, changing column names, and converting the positive outcome varible to 1 and 0.

Exploring -> Using different charts to identify outliers, abnormalities, relationships, and the general shape and feel of different variables.

## 3.1 Data aquisition

The dataset for this project was retrieved from UCI ML Repository in CSV form. Additionally, I created a codebook based on data from the same source.

## 3.2 Data import and cleaning

### 3.2.1 Reading in the Data

|  |
| --- |
| 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 |

### 3.2.2 Dimensions:

dim(raw)

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

### 3.2.3 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" ... |

# 4. Results

## 4.1 Exploratory/Descriptive analysis

[Figure 1](#fig-result) shows a boxplot figure comparing balance levels across job types.

|  |
| --- |
| Figure 1: Job type and bank account balance |

[Figure 2](#fig-result1) shows a bar chart showing education levels of the data.

|  |
| --- |
| Figure 2: Education level |

[Figure 3](#fig-result2) shows a scatter plot figure produced by one of the R scripts.

|  |
| --- |
| Figure 3: Age and bank account balance stratified by marital status |

[Figure 4](#fig-result4) shows barplot of age data for the full dataset

|  |
| --- |
| Figure 4: Age data for those with positive classifcation |

[Figure 5](#fig-result3) shows barplot of age data for only those who subscribed to a term deposit

|  |
| --- |
| Figure 5: Age data for those with positive classifcation |

[Figure 6](#fig-result5) shows barplot of most common days of the month to record a positive outcome.

|  |
| --- |
| Figure 6: Days of the month for positive outcomes |

[Figure 7](#fig-result6) Job type for count of positive outcomes.

|  |
| --- |
| Figure 7: Job type for count of positive outcomes. |

[Figure 8](#fig-result7) Job type for percent of positive outcomes per job type.

|  |
| --- |
| Figure 8: Job type for percent of positive outcomes per job type. |

## 4.2 Basic statistical analysis

Example [Table 4](#tbl-resulttable4) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Logistic regression model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | -2.2210117 | 0.0653189 | -34.0026066 | 0.0000000 | | jobblue-collar | -0.4679972 | 0.0599444 | -7.8071921 | 0.0000000 | | jobentrepreneur | -0.5663637 | 0.1050096 | -5.3934488 | 0.0000001 | | jobhousemaid | -0.2947269 | 0.1112776 | -2.6485722 | 0.0080833 | | jobmanagement | -0.1378963 | 0.0599581 | -2.2998766 | 0.0214552 | | jobretired | 0.7890177 | 0.0676837 | 11.6574254 | 0.0000000 | | jobself-employed | -0.1939794 | 0.0910067 | -2.1314853 | 0.0330492 | | jobservices | -0.3210169 | 0.0692643 | -4.6346685 | 0.0000036 | | jobstudent | 0.9916078 | 0.0849382 | 11.6744663 | 0.0000000 | | jobtechnician | -0.1771822 | 0.0566066 | -3.1300619 | 0.0017477 | | jobunemployed | 0.2617484 | 0.0882630 | 2.9655515 | 0.0030214 | | jobunknown | -0.0941673 | 0.1907228 | -0.4937388 | 0.6214907 | | educationsecondary | 0.2011099 | 0.0524294 | 3.8358211 | 0.0001251 | | educationtertiary | 0.5836750 | 0.0601011 | 9.7115569 | 0.0000000 | | educationunknown | 0.3711903 | 0.0845751 | 4.3888848 | 0.0000114 | |

#WORK IN PROGRESS

## 4.3 Full analysis

*Pending Data Preprocessing and ML Modeling* *I plan to use recipies and some of the tinymodels for this section*

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References