Predicting Customer Behavior Using a Portuguese Financial Institution Dataset

Antonio Flores

2024-07-12

# 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 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. This particular project will utilize Classification Prediction. Unlike Regression Prediction which will attempt to predict a continuous value (e.g., x amount of dollars, x amount of cells), Classification Prediction seeks to train a model that can correctly predict a classification (e.g., True/False, Success/Failure) based on the given predictors. In our case, our y variable or response variable is a binary variable, Yes/No, answering whether or not a customer subscribed to a term deposit.

## 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,17 columns/variables which include: age, marital status, job, education,details related to the phone call, as well as answers related to questions about past credit history. Additionally, as mentioned, the classification variable is whether or not the person subscribed to a term deposit.

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

## 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 | | --- | --- | --- | --- | --- | | 58 | management | married | tertiary | no | | 44 | technician | single | secondary | no | | 33 | entrepreneur | married | secondary | no | | 47 | blue-collar | married | unknown | no | | 33 | unknown | single | unknown | no | | 35 | management | married | tertiary | no |  | balance | housing | loan | contact | day | | --- | --- | --- | --- | --- | | 2143 | yes | no | unknown | 5 | | 29 | yes | no | unknown | 5 | | 2 | yes | yes | unknown | 5 | | 1506 | yes | no | unknown | 5 | | 1 | no | no | unknown | 5 | | 231 | yes | no | unknown | 5 | |

### 3.2.2 Dimensions:

Rows: 45, 211  
Columns: 17

### 3.2.3 Describing data

Summary of predictors

| x |
| --- |
| 'data.frame': 45211 obs. of 12 variables: |
| $ age : int 58 44 33 47 33 35 28 42 58 43 ... |
| $ job : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ... |
| $ marital : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ... |
| $ education : Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ... |
| $ default : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ... |
| $ balance : int 2143 29 2 1506 1 231 447 2 121 593 ... |
| $ housing : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ... |
| $ loan : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ... |
| $ contact : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ... |
| $ day : Factor w/ 31 levels "1","2","3","4",..: 5 5 5 5 5 5 5 5 5 5 ... |
| $ month : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ... |
| $ y\_termSubscribed: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ... |

### 3.2.4 Cleaning Data

There were several different data cleaning methods that were experimented with in order to find the best way to prepare the data for modeling. First, I converted several variables to factors as they had been read in as character variables, or strings. Next I tried to use the DummyVars tool to convert every categorical variable to a dummy value but this made modeling difficult due to a large amount of binary variables. Instead I proceeded with converting the categorical variables to numeric values while maintaining their factor status. I also created a dataset with these numeric values that were not factors. These were the two primary dataset I used for my modeling. Finally I created a dataset stripping the categorical values of their attributes, this was solely created to allow for my corrplot to work.

### 3.2.5 Removing Initial Predictors

I removed the ‘Duration’ variable due to a warning I discovered on the site hosting this dataset initially: “Important note: this attribute highly affects the output target (e.g., if duration=0 then y=‘no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.”(Moro & Cortez (2012)) Due to this new information I decided to remove all the variables related to this, which included ‘campaign’, ‘pdays’, ‘previous’, and ‘poutcome.’ Bringing our new total number of predictors to 11.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

In this section we will be exploring the data using different charts to identify outliers, abnormalities, relationships, and the general shape and feel of different variables.

[Figure 1](#fig-result1) shows the distribution of the ‘Age’ variable.

|  |
| --- |
| Figure 1: Age Histogram |

While this shows a slight skew to the left, if we account for the extreme value on the right, this data seems to be fairly normally distributed.

[Figure 2](#fig-result2) shows the distribution of the ‘Education’ variable

|  |
| --- |
| Figure 2: Education level Bar Chart |

[Figure 3](#fig-result3) shows the most common ‘Job’ types in our data

|  |
| --- |
| Figure 3: Most Common Job Types |

Interestingly, Blue-collar and Management are the two most common job types. This play an important role later in our analysis due to the percent of the data these two categories represent.

[Figure 4](#fig-result4) shows a scatter plot figure produced by one of the R scripts.

|  |
| --- |
| Figure 4: Age and bank account balance stratified by marital status |

[Figure 5](#fig-result5) shows a barplot of most common days of the month to record a positive outcome.

|  |
| --- |
| Figure 5: Days of the month for positive outcomes |

While there doesn’t appear to be a significant trend, we can clearly see that the 30th of the month stands out as being a day of interest, especially compared to other days of the months.

[Figure 6](#fig-result6) shows potential correlation between different variales.

|  |
| --- |
| Figure 6: Correlation Matrix |

The biggest takeaways from our Exploratory Data Analysis is that most variables are fairly normally distributed due to the size of the data, and most variable are not correlated to each other, with one exception (‘Age’ and ‘Marital’)

## 4.2 Basic statistical analysis

Before we begin with the machine learning analysis, I sought to test a few variables of interest (‘Age’, ‘Job’) in their significance of affecting the response variable. As mentioned previously, our response variable is a binary value (Yes/No), which means that we need to use a logistic regression model instead of a linear regression model.

Below are the results for our first logistic model fit with Age as Predictor.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.3195063 | 0.0580829 | -39.934389 | 0e+00 |
| age | 0.0072291 | 0.0013522 | 5.346143 | 1e-07 |

Logistic Regression with Age as Predictor

Below are the results for our second logistic model fit with Job as Predictor.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.2306148 | 0.0290402 | -76.811284 | 0 |
| job | 0.0380809 | 0.0044355 | 8.585536 | 0 |

Logistic Regression with Job as Predictor

As expected, both predictors are very significant. We will proceed to the Machine Learning portion of this project.

## 4.3 Full analysis

So far, I have focused on the MARS model, KNN, and Logistic Regression to determine the prediction model for this project. As highlighted in the data cleaning section, there were two datasets I was focusing on. One that had numeric factors for the categorical predictors, and one that had just numeric variables for the categorical predictors. I decided to examine both so there are a total of 6 different runs recorded below, 3 runs for each dataset.

[Table 2](#tbl-resulttable1) displays the Accuracy and Kappa scores for the first three model runs using the predictors that are numeric and are not factors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Performance Results for Models using Numeric (unfactored) Predictors   |  | MARS | KNN | LogReg | | --- | --- | --- | --- | | Accuracy | 0.8817739 | 0.8811104 | 0.8815528 | | Kappa | 0.0172720 | 0.0465377 | 0.0000000 | |

[Table 3](#tbl-resulttable2) displays the Accuracy and Kappa scores for the first three model runs using the predictors that are numeric and are not factors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Performance Results for Models using Numeric Factored Predictors   |  | MARS | KNN | LogReg | | --- | --- | --- | --- | | Accuracy | 0.8815528 | 0.8813316 | 0.8811104 | | Kappa | 0.0290562 | 0.0339265 | 0.0019542 | |

In terms of accuracy, all 6 model runs seem to have almost identical accuracy. They do tend to differ quite a bit in terms of Kappa…

Finally, we can examine which predictors specifically improved prediction the most.

The following figures were taken from the Factored Numeric model runs.

[Figure 7](#fig-result7) shows the Predictors that had the most impact on the Mars Model

|  |
| --- |
| Figure 7: Mars Model Variables of Importance |

Housing2 indicates that the person does have a housing loan.  
Loan2 indicates that the person does have a personal loan for an unspecified purpose.  
Job9 indicates that the person is a Student.  
education 3 indicates that the person has attained Tertiary Education.

[Figure 8](#fig-result8) shows the Predictors that had the most impact on the KNN Model

|  |
| --- |
| Figure 8: KNN Variables of Importance |

# 5. Discussion

## 5.1 Summary and Interpretation

So far all of the models I have chosen have shown considerable predictive ability. I plan to tune the existing three models while also trying out Random Forest as well as potentially one more model. ## Strengths and Limitations

## 5.2 Conclusions

I will conclude with choosing the model that seems the most robust, accurate, and flexible, as well as determining which variables have the greatest predictive ability and which levels within those predictors stand out as particularly effective in classifying correctly.

# 6. References

Moro, R., S., & Cortez, P. (2012). *Bank Marketing*.