Trends for Characteristics of Subscribing Customers in Response to Bank Marketing Campaigns

Mikaela Taylor

**Introduction**

With this data set, the goal is to be able to predict which customers will subscribe to a term deposit based upon specific qualities and characteristics of the customer and methods used by the campaign. The data set is made up of 45,212 observations and contains 17 variables. Among all of the variables, there were no missing values or null values found. The variables and their descriptions are listed below:

1 - age of customer(numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar", "self-employed", "retired", "technician", "services")

3 - marital : marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown", "secondary", "primary", "tertiary")

5 - default: has credit in default? (binary: "yes", "no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes", "no")

8 - loan: has personal loan? (binary: "yes", "no")

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown", "telephone", "cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable:

17 – y - has the client subscribed a term deposit? (binary: "yes", "no")

**Exploratory Data Analysis for Numeric Variables**

After finding no missing values or null values, the variables distributions were individually looked at to find outliers. After creating boxplots for the numerical variables, outliers were found in age, balance, duration, campaign, pdays, and previous. With each variable, the normality was checked by looking at a histogram and the p-values for normality tests. For the variable age, the histogram was skewed right, so a log transformation was computed to create a more normal distribution. The variable balance also had a right skew to its histogram, but as some of the values were negative or zero, the log transformation could not be used. Therefore, the cube root was taken of each value to get a more normal histogram. After the cube root transformation, the distribution was still skewed right, but was improved from before. The duration variable was found to be right skewed as well, but after the log transformation was performed, the distribution came to look more normal. The campaign variable was more difficult because none of the transformations like the log, square root, and cube root transformations normalized the distribution. So, for now, this variable is being left as is. Both pdays and previous were 80% made up of customers that were not previously contacted, so I grouped both variables as not contacted or contacted.

Next, pairwise comparison was performed with a pairwise scatterplot for each numeric variable, excluding the newly grouped variables previous and pdays. The scatterplots all showed the output variable to be dependent on each numeric variable, meaning they should all be included in further subsequential models and algorithms.

**Exploratory Data Analysis for Categorical Variables**

None of the categorical variables included any missing values or null values, so next a pairwise comparison was performed. Each variable was compared to the output variable y in a mosaic plot. The new grouped variables previous and pdays were included in this comparison due to their newly grouped values. Most of the plots showed that the outcome of whether the customer subscribed to the deposit was dependent on the different variables because the plots individual categories were not equal. The only variable that showed itself to be independent was the variable default, so this means this variable should not be involved in any subsequential models and algorithms.