# Bank Loans – Solution

## Task 1 – Examine the target variable (5 points)

The target variable, y, is whether the person has subscribed to a term deposit. The predictors are client data, variables related to their last contact with the bank, the month and day of year, the number of times that they have been contacted in that campaign, as well as economic indicators.

The target is imbalanced because there are only 4,640 people who have a term deposit and 36,548 who do not. This creates problems with models because they tend to fit people who do not have loans well but fit the people who have loans poorly.

The most subscribers sign up in the middle of the week from Tuesday – Thursday. Signups are less common on Fridays, and there are no records from Saturdays and Sundays which makes sense because this is outside of business hours.

|  | **day\_of\_week** | **percent\_subscribed** |  |  |
| --- | --- | --- | --- | --- |
| 1 | fri | 10.81% |  |  |
| 2 | mon | 9.95% |  |  |
| 3 | thu | 12.12% |  |  |
| 4 | tue | 11.78% |  |  |
| 5 | wed | 11.67% |  |  |

The largest group of people who take out loans are people who have a university degree, with 13.72%. The unknown group is 14%, but this could include people with or without a degree, and so we discard this.

|  | **education** | **percent\_subscribed** |  |  |
| --- | --- | --- | --- | --- |
| 1 | basic.4y | 10.25% |  |  |
| 2 | basic.6y | 8.20% |  |  |
| 3 | basic.9y | 7.82% |  |  |
| 4 | high.school | 10.84% |  |  |
| 5 | illiterate | 22.22% |  |  |
| 6 | professional.course | 11.35% |  |  |
| 7 | university.degree | 13.72% |  |  |
| 8 | unknown | 14.50% |  |  |

There were only 18 observations in the “illiterate” category and so we do not give credibility to the 22% figure and these people were grouped together with “unknown” into an “other” category. The subscription rate for both for both groups is high.

## Task 3 – Decide on which variables to discard (5 points)

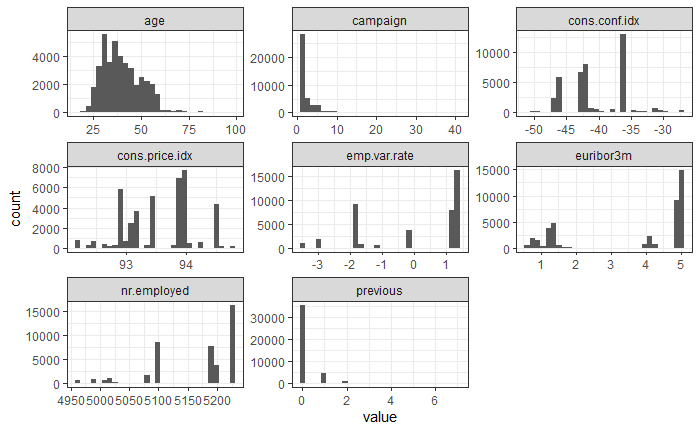
We want to predict which customers will subscribe before a phone call is made. The duration variable highly affects the output target but is not known before a call is performed. If we were to use this as a predictor then it would leak the signal. The resulting model would have duration as the highest importance predictor and it would appear to have good performance on the training and testing data, however; there would be no practical way of building this model in real life.

After inspecting the data dictionary, all the other variables were deemed to be appropriate to included.

The data dictionary says "Yet, the duration is not known before a call is performed"

## Task 4 – Examine the numeric variables (5 points)

The histograms show the distribution of each of the numeric variables.



**Age**

Most people in the data are under the age of 65. When building a predictive model, we recommend creating a separate indicator variable for whether the person is under 65. This will help linear models to take this information into consideration.

**Campaign**

The distribution is highly right skewed and positive. The data dictionary said that the max values are 40 and so six values which were 41 and above were removed.

**Pdays**

This was converted to a categorical variable and so the histogram is not shown above. Most customers had not been contacted in a prior campaign and so they had a “999” value. If this had been kept as a numeric value there would have been problems in linear models because of these non-linear values. To fix this, a binned value was created. The original variable was removed so that there will not be any linear dependency. The base factor level was set to that which had the most observations. The bin sizes were chosen so that all groups have more than 50 observations.

**cons.price.idx**

The consumer price index has several repeated values for different people because it reflects the state of the economy on that particular day. The distribution is approximately uniform and ranges between 92.2 and 94.8.

**cons.conf.idx**

The consumer confidence index ranges from -50 to -30. The data dictionary does not provide an interpretation of these values and so we are not sure how they should be interpreted.

**emp.var.rate**

Similarly to the other economic indices for the price index and confidence index, there is no clear interpretation for this beyond that it is the “employment variation rate.”

**euribor3m**

This is a bimodial distribution that has centers near 1.5 and 4.5.

**nr.employed**

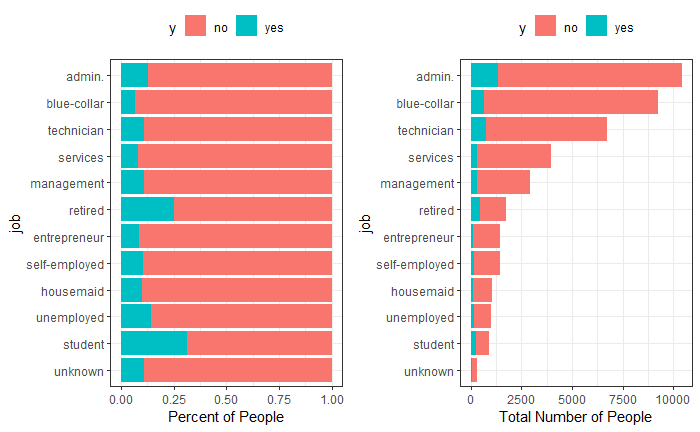
The number of employees ranges from about 5,000 to about 5,230 and has a mean of 5,167. The data dictionary does not specify whether this is the number of employees of the firm at the time of the call, which would be an economic indicator for the firm issuing the loans. We will follow up with the client to clarify that this is the case.

**Previous**

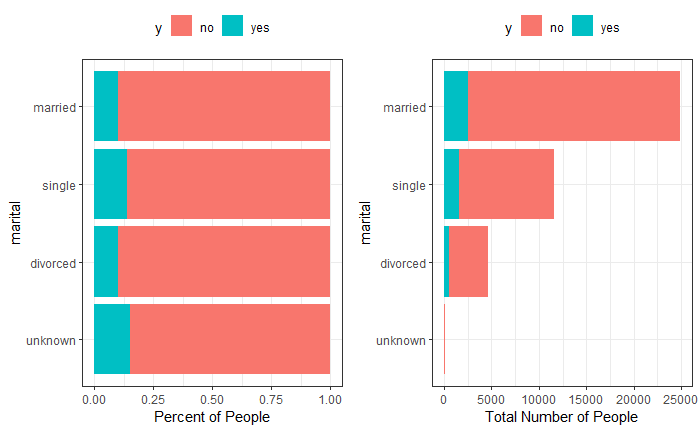
The distribution is inflated at zero because most people have never been contacted before.

## Task 5 – Examine the factor variables (10 points)

The bar graph below shows both the percentage of people (left) as well as the number of people (right). The most common jobs are admins, blue collar workers, and technicians, but the highest subscription rates are for students and retired people. These people do not have a consistent source of income and so it makes sense that they would require a loan.



The same graph for marital status shows that divorced people are less likely to have a loan than married or single people. The subscription rate for the “unknown” group is highest (left) but this does not represent a significant number of people (right). The largest group is for married people.



Most people have a mortgage (housing loan) and the subscription rate is highest for this group. People who not have a housing loan are also less likely to have one of these bank loans.

