

**TASK**

**Exploratory Data Analysis on the Bank Marketing Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

# Summary of the Data Set

The data set was obtained by the UCI Machine Learning Repository <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>.

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The bank-additional-full.csv contains all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analysed in [Moro et al., 2014]. This dataset is publicly available for research. The details are described in [Moro et al., 2014].

Number of Instances: 41188 for bank-additional-full.csv

Number of Attributes: 20 + output attribute.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

**Attribute information:**

For more information, read [Moro et al., 2014].

**Input variables:**

**Bank client data:**

1. age (numeric)
2. job: type of job (categorical: "admin."," blue-collar", "entrepreneur", "housemaid", "management”, “retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
3. marital: marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)
4. education:(categorical:"basic.4y","basic.6y","basic.9y","high. school", "illiterate", "professional. course", "university. Degree", "unknown")
5. default: has credit in default? (Categorical: "no", "yes", "unknown")
6. housing: has housing loan? (Categorical: "no", "yes", "unknown")
7. loan: has personal loan? (Categorical: "no", "yes", "unknown")

**Related with the last contact of the current campaign:**

1. contact: contact communication type (categorical: "cellular", "telephone")
2. month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
3. day\_of\_week: last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")
4. duration: last contact duration, in seconds (numeric). 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.

**Other attributes:**

1. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
3. previous: number of contacts performed before this campaign and for this client (numeric)
4. poutcome: outcome of the previous marketing campaign (categorical: "failure", "non-existent", "success")

**Social and Economic context attributes**

1. emp.var.rate: employment variation rate - quarterly indicator (numeric)
2. cons.price.idx: consumer price index - monthly indicator (numeric)
3. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
4. euribor3m: euribor 3-month rate - daily indicator (numeric)
5. nr.employed: number of employees - quarterly indicator (numeric)

**Output variable (desired target):**

1. y - has the client subscribed a term deposit? (Binary: "yes", "no")

**DATA CLEANING**

SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

1. As missing values in the data set were labelled as ‘unknown’ the first step was to convert na\_values to unknown.
2. Checking for missing values - check to see the number of missing values in each column and the percentage of missing values they contribute to.
3. Find duplicate rows - the decision I took was to keep ‘last’ – as often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
4. Missing values exist in almost all datasets, and it is essential to handle them properly in order to construct reliable machine learning models with optimal statistical power. For the banking data set – I decided to replace them by using the K-Nearest Neighbors imputation method.
5. For k-Nearest Neighbor imputation, the values are obtained by using similarity-based methods that rely on distance metrics (Euclidean distance, Jaccard similarity, Minkowski norm etc). They can be used to predict both discrete and continuous attributes. KNN works by finding other observations that are almost similar to the observation with the missing value. For example, if the observation is 'Female' and 'Asian' we will find other users similar to her and get the mean or mode of the missing value.
6. For the imputation method I used the KNN Imputer supported by Scikit-Learn. This imputer utilizes the k-Nearest Neighbors method to replace the missing values in the datasets with the mean value from the parameter ‘n\_neighbors’ nearest neighbors found in the training set. By default, it uses a Euclidean distance metric to impute the missing values.
7. To see this imputer in action, we will import it from Scikit-Learns impute package.
8. One thing to note here is that the KNN Imputer does not recognize text data values. It will generate errors if we do not change these values to numerical values. For example, in our banking dataset, the categorical columns such as ‘job’ and ‘education’ have text data.
9. A good way to modify the text data is to perform one-hot encoding or create “dummy variables”. The idea is to convert each category into a binary data column by assigning a 1 or 0.
10. First, I will make a list of categorical variables with text data and generate dummy variables by using ‘.get\_dummies’ attribute of Pandas data frame package..
11. Another critical point here is that the KNN Imptuer is a distance-based imputation method, and it requires us to normalize our data. Otherwise, the different scales of our data will lead the KNN Imputer to generate biased replacements for the missing values. For simplicity, we will use Scikit-Learn’s MinMaxScaler which will scale our variables to have values between 0 and 1.
12. Now that our dataset has dummy variables and normalized, we can move on to the KNN Imputation. We import it from Scikit-Learn’s Impute package and apply it to our data. In this example, we are setting the parameter ‘n\_neighbors’ as 5. So, the missing values will be replaced by the mean value of 5 nearest neighbors measured by Euclidean distance.

MISSING DATA

Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

DATA STORIES AND VISUALIZATIONS

* The first graph is a heatmap to find the dependent variables - One of the best ways to find the correlation between the features can be done using heat maps. From the heatmap generated, there does not seem to be any strong correlations between any of the features.

Chart, scatter chart

Description automatically generated

* Chart

  Description automatically generatedNext, is a boxplot showing the distribution of various ages in the dataset. Due to the ages being normalized, the exact ages were not able to be extracted. Although, one can see that there are a great deal of outliers regarding the age distribution.
* Followed after this is bar chart depicting individuals who subscribed to a term deposit VS those who did not. From the chart it is clear that more people did not subscribe VS those who did.

**Chart, bar chart

Description automatically generated**

* Below this, is also a bar chart depicting individuals who subscribed to a term deposit VS those who did not (**previous campaign**). The result is the same, more individuals chose not to subscribe VS those who did. Although, in comparison, the latest campaign was more successful.

**Chart, bar chart

Description automatically generated**

* Chart, pie chart

  Description automatically generatedFollowed after this, are 5 pie charts. The first dealing with the job categories, the second dealing with education, third dealing with marital status, fourth dealing with housing loans and lastly one dealing with personal loans.

**Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated**

**Chart, pie chart

Description automatically generated**

* **Observations**: Regarding the first pie chart – the largest group in terms of occupation seems to be those in Admin jobs.
* Regarding the second pie chart - the largest group in terms of education seems to be those with university degrees.
* Regarding the third pie chart - the largest group in terms of marital status seems to be those who are married.
* Regarding the fourth pie chart – more than half had housing loans at the bank.
* Lastly, regarding the fifth pie chart – majority of individuals (84%) did not have a personal loan at the bank.
* Next, an analysis was performed regarding job type and education in relation to those who did subscribe to a term deposit VS those who did not.
* Regarding occupation – those in admin jobs subscribed more and second were technicians.
* Regarding education – those with university degrees subscribed more and second were those with high school education.

Graphical user interface, application, table, Excel, calendar

Description automatically generated

**Graphical user interface

Description automatically generated with medium confidence**

* The same analysis was done regarding month and day of the week. From the data – May was the busiest month in terms of subscriptions and Thursday was slightly higher than the other days of the week.

Graphical user interface, table, Excel

Description automatically generated

**Graphical user interface, chart, application, bar chart

Description automatically generated**

* Lastly, an analysis was done to view the relationship between previous campaign success, consumer price index, consumer confidence index and the 3-month Euribor interest rate in relation to y\_yes == 1.0.
* For **poutcome success and y\_yes == 1.0** – there is no relationship between X and Y.
* For **consumer price index and y\_yes == 1.0** – there seems to be a positive relationship between X and Y, although the data points are dispersed with a few outliers.
* For **consumer confidence index and y\_yes == 1.0** - there also seems to be a positive relationship between X and Y, although the data points are dispersed with a few outliers.
* For **the 3-month Euribor interest rate and y\_yes == 1.0** - there is no relationship between X and Y.

Chart, scatter chart

Description automatically generated

* Conclusion: The binary classification goal is to predict if the client will subscribe a bank term deposit (variable y).
* Regarding the input features, there does not seem to be a clear cut in what features most impact a client’s decision. Although, from the data, those in administrative jobs seemed to subscribe more. Those with university degrees seemed to subscribe more and those who were married seemed to subscribe more. Clients who also subscribed to the previous campaign were most likely to subscribe again. Lastly, the consumer price index and consumer confidence index seemed to impact a client’s decision.

**THIS REPORT WAS WRITTEN BY: DEVAN BOOY**

