**Capstone Project1: Banking Customer Recommendation**

The following describes the first Data Wrangling steps for the Capstone Project Banking Customer Recommendation. The data are stored as a csv-file with a size of approximately 2GB.

The goal of this project is to find a Machine Learning algorithm which is capable to predict whether a client is likely to add a specific service of the bank to his basket or not in the following month. The dataset comprises approximately 13 million rows with 22 features describing the customer and 24 features represent his current product basket in sparse form. Most features describing the bank’s client are categorical. Regarding the rows, the dataset is organized in time ordered fashion, thus encompassing about 18 month of customer readings. Consequently, the same customer is likely to appear a couple of times in this set i.e. once in each month.

* **Defining target variables**

As a first step, for each customer’s product basket (sparse matrix) a comparison to the respective customer’s future product basket was conducted. As a result, for each client changes in his product basket are assessed which is necessary to define the targets (new products added on a monthly basis) for the Machine Learning algorithm.

* **Cleaning the features**

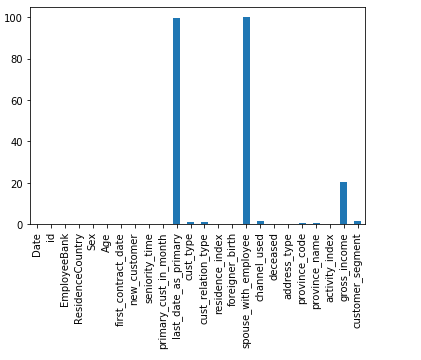
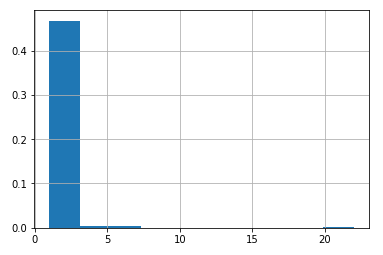
The dataset contains labels and feature names in Spanish; thus, translation and relabelling have been conducted. After assessing the targets in a sparse matrix format and translation to English, the datatypes of the respective columns have been controlled. Many features show mixed datatypes, as a result of mixing up float values with strings. An appropriate algorithm was applied to clean these datasets and to transform them to continuous types. Other features, mainly categorical ones, have inconsistent labels which are hard to understand logically. Accordingly, they have been replaced by more consistent and comprehensive labels. Features, which had integer values but are intended to represent categories are converted to categorical datatypes to enhance computational efficiency. In addition, one feature which comprises the dates of when a customer has become a client has been transformed to represent the time difference to the record’s dates in the set to enhance informational content.

* **Outlier detection**

Based on the fact that most features are categorical, an outlier is defined as percentage of occurrence. Thus, if a categorical label does not occur more often than 2 % in this feature it is signed as possible outlier. For continuous features, standard deviations from the mean are considered. Hence, outlier records of the features “Age” and “Gross Income” are marked in case of values departing more than three standard deviations from their respective mean. An inspection of possible outlier values in categorical features is not providing a real clue, whether these values are the consequence of typos during data collection or if they are a real print. For instance, if customers’ proveniences from a certain region are only about one percent, one cannot conclude this to represent a mistake. However, for continuous features a couple of outliers are detected that are dispersed widely from the feature’s average. But due to the shape of these features’ respective dispersion an additional step has been taken into account. Therefore, values of “Age” higher than 110 and lower than 12 have been replaced as missing values. Basically, if children are allowed to have accounts the control of these accounts is likely to be legally restricted to adults, thus age does not provide proper information. Moreover, “gross\_income” which reflect the household’s income of the customer has entries above 1.5 million Euro which is quite unlikely to occur that often. Therefore, beside upside deviations from the average, such values have been marked as missing values.

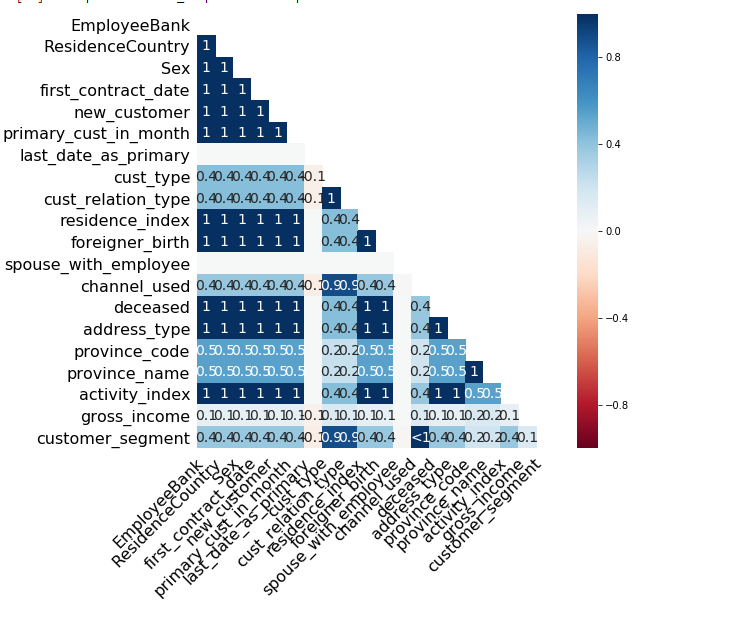
* **Missing values**

In this step, missing values are replaced and features which do not contain much informational content in terms of missing variance or its arbitrary content have been removed. The following graphs show the occurrence of missing values (NAs) only containing the matrix with client’s features: 1) per feature and 2) as frequency per row.

Obviously, the features “last\_date\_as\_primary” and “spouse\_with\_employee” have overly missing values, whereas record wise most have about 2 NAs and only a few are covered completely by missing values, i.e. values of 20. The feature “gross\_income” and “customer segment” are also covered quite by a remarkable percentage of missing values.

In order to detect some kind pattern of missing values between features a heatmap is shown below:



Here, red colour contour hints at dependence between the NA occurrence of feature X to feature Y. According to the heatmap not systematic dependence of missing values is being detected.

With regard to sparse product basket, two features had missing values and the following points describe the process of cleaning up NA values.

* + The feature “deceased” marks if a customer is actually leaving the bank or has died. Due to the fact that of deceasing accounts one cannot predict any future change all records with positive “deceased” entry are removed, then the whole column has been dropped from the set
  + Based on the fact, that a customer appears many times in the dataset some features are expected not to change, i.e. like “Sex”. Thus, a look-up procedure is applied to find missing values of record in another record of the same customer. For instance, missing values in gender or Age are not likely to change at all or over a couple of months. As a result, some missing values are being replaced by a proper value.
  + Records with missing entries in the product basket are removed completely.
  + The features “spouse\_with\_employee” and “last\_date\_as\_primary” which have entirely missing values are dropped. The features “address\_type” and “province\_code” are removed as well. These features do not provide any important content at all, i.e. province\_code is a mirror of the feature “province\_name”.
  + The continuous features “Age” and “gross\_income” a replacement procedure is designed. This takes other features which show appropriate variance and constructs clusters. According to the cluster to which a missing value record belongs the clusters’ median of the feature in question is taken as a replacement value. For categorical features the respective clusters’ most common value has been taken. This represents a common approach to replace missing values although the argumentation might be error prone.

Finally, two new features are being created. Following the first step, where the target variables have been calculated, which are resembled in sparse matrix format as well, the numbers of products purchased and deceased have been assessed. Thus, for customer i at datetime T the feature “new\_ones” shows the number of products the customer i has purchased in comparison to datetime T-1. The feature “leave\_ones” shows instead the number of products which have deceased from his customer i’s product basket between datetime T-1 to datetime T. Possibly, these two features might provide value in researching an appropriate machine-learning algorithm.