Belvo - Test [Data Intern]

Candidate: *Maria Júlia Cristofoletti de Souza*

Date: *June 18th, 2021.*

Categorizing Bank Transactions - Report

Here is a brief report on how banking transaction data was categorized into 14 categories previously given by the test.

Initial insights

First, an overview of the data was made, investigating if there were missing values, seeing the distribution of values and the pattern of each column of data.

From this it was seen that it was a short period of time, about 11 months, with 75% of transaction values below U$ 685.00 with a previous classification regarding the type of movement varying with G and N, being that the values of transactions within these two categories are distributed similarly among themselves and in relation to all data. These two types of movement were quickly recognized as being N related to incoming transactions and G related to outgoing transactions.

The most important information from the dataset was considered to be related to the descriptions of each transition, where we chose to make the classification. The description of each transaction provided general information about the one it was related to, but there was no pattern between them and it was felt that it would have to go through prior preparation before applying classification filters.

Categorizing

As stated before, the classification was based entirely on information referring to the description of each transaction and also using the type of movement. Thus, for each category, a series of keywords was selected that would give us a first way to classify the data.

Each transaction was then initially related to some categories and then a prioritization was made in which certain word patterns were given as having a higher weight so that the right category was chosen.

The classification process was made in such a way that certain transactions, mainly related to small purchases in stores and other establishments, which did not present any keywords and, thus, were not included in any of the previous categories, were included in the category "Income & Payments", due to lack of information. This process took place mainly because the transactions were only related to the name of stores and establishments and could be resolved if there was a better description of the transaction or additional information about the origin of that transaction. This information would definitely help in the sorting process for all the other data.

The lack of this information is the main cause for the drop in the accuracy of the applied model which, being 20 of the 179 transactions, leads to an estimate of 88.9% of the model's accuracy. However, it is worth complimenting that the use of keywords and the category prioritization is also a major source of uncertainty, estimating that keywords are only correct for 90% of cases, value close to data that were not in any of the categories, we get a final estimate of 78.9%.

Conclusion

While performing this categorization, it was very important to carry out a good research on common transactions within a bank statement, in addition to familiarizing yourself with taxes and other types of transfers and investments within a bank.

During the task, there was a question about the application or not of a classification model with Machine Learning, as this task becomes more difficult as the number of data increases. However, for that, there would have to be more data. In this context, this first preparation, standardization and classification of data could be used as a baseline model to apply more sophisticated techniques.