# **Project Report for**

# **Data Mining Group Project**

# 

## **Master’s in Data Science and Advanced Analytics at NOVA IMS, Lisbon**

## **Group Information**

Group:

Group name

Group members:

Henrique Vaz

Philipp Metzger

Link to GitHub repository:

<https://github.com/ph1001/NOVA-Data-Mining-Project>

**Abstract**

**I. Introduction**

The goal of this project is to develop a Customer Segmentation based on a dataset provided by the non-profit organization Paralyzed Veterans of America (PVA). PVA provide programs and services for US veterans with spinal cord injuries or disease. The dataset contains information on individuals that have donated to PVA and that are classified as “Lapsed” donors[[1]](#footnote-1), meaning that they made their last donation to PVA 13 to 24 months ago.

**II. Background**

**III. Methodology**

The purpose of this chapter is to explain the steps that were taken in this project. The explanations will be oriented by the Jupyter notebook that has been handed in alongside with this report, meaning that the order of the explanations of each step will be respective to their order in the notebook. The results and findings as well as their implications will be presented and discussed in the following chapter IV.

**III.1 Importing and organising the data**

As a first step, the necessary libraries are imported, and the dataset is loaded into main memory from the csv file provided.

Next, using the file “pva\_metadata.txt”, the features obtained from the imported dataset are split into metric and non-metric features.

Then, a few

**III.2 Data cleaning**

In this step, again using the metadata file, the existence of cells is assessed, where spaces (“ “) carry a meaning such as for the feature ‘MAILCODE’, where a space means that the address of this individual “is OK”, whereas the value “B” means that the address “is bad”. Other features where something like this or similar was the case are ‘NOEXCH’, 'RECINHSE', 'RECP3', 'RECPGVG', 'RECSWEEP' and ‘MAJOR’. For these features the spaces are replaced by a meaningful string, such as “Address is OK”.

In the next step, all remaining spaces in the dataset are replaced by NaN values[[2]](#footnote-2). It is also checked if the dataset contains any empty strings.

In the next part of the code, the percentage of missing values contained in each feature is assessed. Features that have more then 40 % missing values are discarded.

**III.3 Data transformation**

In this section of the Jupyter notebook, all features containing time related data such as dates are transformed from their original string format to datetime.date[[3]](#footnote-3) format.

**III.4 More data cleaning**

Removing values?

In the following part of the code, some inconsistencies are removed from the dataset. In this case, an observation is seen as inconsistent, if it does not comply with the definition of a “Lapsed” donor, as it is explained in chapter I. Observations like this are discarded.

**III.5 Feature Selection**

As a first feature selection step, correlations between all metric features are assessed. Of feature pairs that are highly correlated, one is discarded, and one is kept in the dataset. Also, metric features that only contain a very small number of distinct values and thus carry little information for our analysis as well as features containing mostly zeros are discarded.

III.6 Further Data Transformation

In this step, the date features are converted to integers, representing their distance in days to each observation’s reference date, which was chosen to be the date on which the last promotion was sent to each individual. The in this way obtained features are names in the following format: “<<original name>>>\_rel\_in\_days”

**IV. Results** **and Discussion**

**V. Conclusion**

**VI. References**

|  |  |
| --- | --- |
| [1] | [Online]. Available: https://docs.python.org/3/library/datetime.html. [Accessed 30 Dec. 2020]. |

1. In the dataset’s metadata file also denoted as “Lapsing” donor [↑](#footnote-ref-1)
2. NaN, short for “not a number”, commonly denotes a missing value in a dataset. [↑](#footnote-ref-2)
3. For the documentation of the Python library “datetime”, see [1] [↑](#footnote-ref-3)