MACHINE LEARNING

MDSAA-BA

To Grant Or Not To Grant Project

Group 19

João Sampaio 20240748

Leonardo Caterina 20240485

Mara Mesquita 20241039

Maria Pereira 20240314

Mariana Neto 20211527

Chiel Groen

NOVA Information Management School

Universidade Nova de Lisboa

I. Abstract

In this part of the report, an overview of the work will be given. We will contextualize the problem at hand and define the goals expected. Additionally, a summary of the project itself and the results achieved will be provided. From that, some conclusions will be drawn so we can link the actual results to what was expected.

II.Introduction

In this section, we will provide an overview of the project, by describing the data used and the objectives we are trying to achieve with this project. Some similar projects will be referred to, so that we can compare the main differences between what we did and what was found.

III. Data Exploration

In this part, we will present the main insights taken from data exploration. We will present the number of features, the existing data types, the target variable and describe the main insights taken from the descriptive statistics for both numerical and categorical features. We will also identify some problems found, such as duplicated rows, and explain how we dealt with them. Other topics to be discussed will be the existence of missing values. We will identify the features that have missing data, explore the missingness between features and present results and, also, provide some solutions on how we fill some of it, such as imputing with a constant. We will check the coherence of the data by identifying some of the incoherences found. Moreover, we will identify some of the outliers found and what strategies we used to remove the impact of its presence. Finally, we will explain the lookup tables created and the removal of some columns.

IV. Preprocessing

Most of the preprocessing steps made will be explained in this part. Firstly, we will explain how we split the data into training and validation sets. After that, we will explain how we addressed the inconsistencies found during the data exploration. In this section, we will also explain the new features created and the transformations that were made to the existing ones. Besides that, we will explain how we dealt with the outliers found in other features. This will, also, lead to the explanation of the chosen scale. On top of that, we will explain how we filled the remaining missing values, which technique we used and what led us to the actual hyperparameters.

V. Multiclass Classification

In this section, we will talk about the feature selection strategies used, the result of each and the approaches taken. Additionally, a comparison for the different models tested with a focus on the performance in terms of macro F1 score will also be explained. We will also present the different combinations of data and hyperparameters used in each model to provide reasons for the final model choice.

VI. Open-Ended Section

At this stage of the report, the strategy of additional insights will be described aligning each aspect with the project’s original objectives. Additionally, it will be discussed if the initial objectives are aligned with the main findings of the project. The challenges faced during the realization of the project will be presented and will be ranked from easier to harder, based on the group's general opinion. Last but not least, we will provide and explain the results. We will also check if these results matched the objectives of the project.

VII. Conclusion

In this last section we will write briefly the initial objectives and evaluate if these objectives correspond to the final findings. We will also analyze what were the limitations on our work, such as different possible approaches, on data preprocessing, model selection, assessment, among other steps in our work. Another topic that will be discussed is, based on our work, what future work could be done.