TO GRANT OR NOT TO GRANT: DECIDING ON COMPENSATION BENEFITS

REPORT STRUCTURE

Group 40: André Sousa (20240517), Hugo Fonseca (20240520), Miguel Azevedo (20240533), Rafael Bernardo (20240510), Tiago Soares (20240655)

1 Introduction

"To Grant or Not to Grant: Deciding on Compensation Benefits", is a project that aims to create a predictive model to support the New York Workers' Compensation Board (WCB) in automating decisions on compensation claims. This model is designed to classify claims into categories based on injury type, presenting a more refined option to substitute a manual, resource-intensive review process.

The report will describe the steps that group 40 took in order to achieve the desired outcome. The following article highlights how this article will be organized.

2 Exploratory Data Analysis (EDA)

This section will be divided in three parts:

Dataset Description: A brief description of the dataset will be given. How many records and features does the data have and which of those are input or output variables.

Handling Missing Values and Anomalies: The presence of missing values, outliers, or inconsistent data will be evaluated. A description of the steps that the group took in order to handle this anomalies will be displayed here.

Variable Distribution: Charts that show the distributions of the main variables will be mentioned here, and displayed in a referenced Appendix.

3 Data Cleaning and Preprocessing

This section will highlight the options made to organize, clean and setup the data to optimize the performance of a future predictive model that will be used.

Data Features: Any data type changes made to features related to dates will be described here.

Categorical Variable Processing: The way categorical variables were addressed converted into numerical variables will be explained in this section.

Outlier Treatment: Identification and handling of possible outliers in the data set will be acknowledged in this part of the report.

Normalization/Scaling of Data: The utilization of data scaling techniques to normalize the numerical features of the data set will be specified in this segment.

4 Feature Selection

This section will address the process of choosing the features that best contribute to the objective of predicting the target variable. Firstly all methods used will be described. Those are divided in three groups, **Filter Methods**, **Wrapper Methods** and **Embedded Methods**. Finally, a final subsection will cover the process of choosing the correct number of features that will be important for predicting the output variable, based on the results obtained from those previous methods. Any graphical visualizations will be displayed in Appendices at the end of the report.

5 Building a Simple Model

This section will mention the initial algorithm choices that were made. After that, the evaluation metrics selected to assess the model's performance will be described and justified. Those results will be key for selecting the best model to solve the problem. Finally the preliminary results of the model will be presented.

A brief description of the process of testing our algorithms on Kaggle will be shown here. By submitting incremental improvements and analyzing Kaggle's leaderboard feedback, the goal is to iteratively refine the model to climb the ranks and achieve superior performance.

6 Next Steps

In the next phase of this project, several enhancements are planned to further improve the model's performance and robustness. These improvements include:

Exploring Other Algorithms: After the first delivery, other algorithm options will be evaluateed and tested to see if they improve the performance of the model.

Tuning Hyper-parameters: Highlight any hyper-parameter manipulation that can further enhance the efficiency and results of the model.

Additional Feature Engineering: Any future features that can be created or transformed will be mentioned here.

7 Conclusion

In this final section, a brief summary of the key points will be done, going from the data processing techniques to the initial model construction.

The impact of the solution that the model can provide will be discussed in this part, specifically the benefits that it can bring to the decision-making problem that was initially introduced.