MACHINE LEARNING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

To Grant or Not to Grant: Deciding on Compensation Benefits

Expected Report Structure and Content

Fall/Spring Semester 2024-2025

Group 38

Ana Marta Azinheira | 20240496

Braulio Damba | 20240007

Henry Tirla | 20221016

Marco Galão | r20201545

Rodrigo Sardinha | 20211627

Table of Contents

1. Abstract

This report presents an overview of our approach, detailing the context in which the problem was framed, the specific objectives we set, and the initial hypotheses that guided our analysis. We highlight key insights derived from the data, setting the stage for understanding the methodologies and results achieved. For further exploration, you can access the full project via our GitHub Repository. Feel free to fork it!

2. Introduction

The primary focus of this project is to address a significant real-world problem with potential impactful solutions. Here, we outline the key goals, emphasizing the practical implications of solving this issue. We will also review the existing literature to see if similar problems have been tackled by others, identifying gaps that our analysis may fill.

3. Data Exploration and Preprocessing

Our initial analysis begins with a comprehensive exploration of the dataset to understand its structure, feature distribution, and overall quality. This involves assessing data types, identifying missing values, and verifying data integrity. The goal is to derive meaningful insights that are critical for building robust predictive models.

3.1. Missing Values

Handling missing values is essential to ensure data quality. This step includes categorizing missing data, checking for columns with high percentages of missing entries, dropping irrelevant columns, and imputing or handling missing values as needed to maintain data consistency.

3.2. Inconsistencies

We will also examine data inconsistencies, such as unrealistic values (e.g., a birth year of zero or dates that do not follow a logical sequence). Also, given the employment context, are ages below 18 relevant to our analysis? Probably not! These types of questions will be addressed to ensure that our data is not only accurate but also contextually valid for analysis.

3.3. Outliers

To avoid data leakage, we will split the dataset into training and validation sets before analyzing outliers. Outliers will be examined to determine if they represent genuine anomalies or data entry errors. Our approach includes visualizing distributions to assess the impact of these outliers on model performance. Extensive use of graphs and tables will enhance the clarity of our exploratory analysis.

4. Multiclass Classification

4.1. Feature Engineering

At this stage, we will refine our dataset by creating new, informative features that may improve model accuracy. This process involves analyzing existing attributes to derive new ones that capture underlying patterns, enhancing the ability to predict the target variable.

4.2. Feature Selection

Here, we will apply feature selection techniques (e.g., Chi-Square, Recursive Feature Elimination, Lasso Regression) to identify the most significant predictors. Any necessary additional preprocessing, such as encoding categorical variables or scaling numerical features, will be performed before modeling.

4.3. Modelling and Assessment

We will develop and compare at least three distinct models, such as Logistic Regression, Random Forest, and XGBoost. These models will be evaluated using performance metrics like accuracy, precision, recall, and the F1 score, with a particular focus on the weighted F1 score to account for class imbalances. Hyperparameter tuning will be conducted to enhance model performance, ensuring the most effective model is selected for deployment.

5. Open-Ended Section

In this section, we will extend our analysis by adding two variables, specifically "Agreement Reached" and "WCB Decision," to evaluate their impact on model performance. Our goal is to determine if these new variables can improve the model's predictive accuracy for the primary target variable (“Claim Injury Type”). Additionally, we will explore how these variables interact with existing ones, potentially uncovering new insights.

6. Conclusion

This section summarizes the key findings, limitations (e.g., data quality issues or model constraints), and propose recommendations for future research.

7. Bibliography

References for relevant literature, research papers, and sources that informed our analysis.

8. Annexes

Tables, graphs, and visualizations will be included in the annexes for detailed analysis.