Handout | Group 33

André Silvestre, 20240502 | João Henriques, 20240499 | Simone Genovese, 20241459
Steven Carlson, 20240554 | Vinícius Pinto, 20211682 | Zofia Wojcik, 20240654

ABSTRACT

The abstract will provide a brief description of the project, including context around the New York Workers' Compensation Board (WCB) and its challenge with automating claims review processes. In this part, we will summarise the objectives of this project including developing a projection model for injury type classification, optimising the model and exploring its insights. The main methods, key results, and conclusions will be briefly presented, along with the performance of the best model.

1. INTRODUCTION

The introduction will provide a background on the purpose of Workers' Compensation Boards (WCBs), specifically how it processes and determine workers' compensation claims. Next, we'll look for relevant information about the variables present in the dataset (wage, gender, type of injury...) that can explain some of the characteristics of the data that can lead to outliers, unusual values or missing values.

Additionally, we will discuss three important topics: First, how we leveraged the CRISP-DM methodology to implement the ML pipeline steps; Second, a more in depth explanation of the project objectives, including the development of multiple multiclass classification models and the selection of the best-performing one; Third, how we leveraged the knowledge obtained during the analysis of similar works in the area of claim prediction to improve our results.

2. DATA EXPLORATION AND PREPROCESSING

In this section, we will focus on exploring the dataset to understand its structure and quality and implementing an initial preprocessing stage. This process will include four main steps: univariate analysis, multivariate analysis, feature engineering and preliminary preprocessing.

The first step includes the examination of the distribution and summary statistics of each variable with the goal of detecting potential anomalies, errors, missing values and outliers.

Following the univariate findings, the second step will focus on identifying interactions, patterns, and dependencies between multiple variables. This analysis will help identify relationships and patterns that may have been missed during univariate analysis.

The third step will be focused on creating new variables to simplify complex data and reduce noise and anomalies within the dataset. This will help improve the data quality and alignment with the business needs.

In the last step, we will implement preliminary preprocessing tasks (e.g converting datatypes and dropping certain rows) and plan future preprocessing steps, such as strategies for handling missing values and outliers.

3. MULTICLASS CLASSIFICATION

This section will discuss building and evaluating the classification model. First, we will implement additional preprocessing steps, including the imputation of outliers, missing values treatment, and the encoding of the categorical features.

Second, we will explain our feature selection strategy, which involve *Filter Methods,* based on statistical measures, such as correlation coefficients and Chi-square tests; *Wrapper Methods* using recursive feature elimination (RFE); and *Embedded Methods* with regularization techniques like Lasso (L1) and Ridge (L2) regression, that need to be adapted to our classification problem.

Next, we will explain our model assessment strategy, which will cover the performance metrics used to evaluate our models (e.g., accuracy, precision, macro F1-score) and the cross-validation process employed for reliability and generalizability of the model.

Using this strategy, we will compare the performance of various algorithms, including *Logistic Regression*, *Decision Trees*, *Random Forest* and *Neural Networks* among others to evaluate their strengths and weaknesses and ultimately choose the best model.

Finally, we will discuss the hyperparameter tuning and complexity control used to optimize the selected model.

4. OPEN-ENDED SECTION

In this section, we plan to run a variable importance and variable interpretability analysis, to check which features are most predictive for each injury type. To do this, we would use the interpretability libraries LIME so that we can understand how features contribute to individual predictions.

Additionally, we want to create an interactive web application using Streamlit. Users will be able to provide data inputs, and based on the best performing model, the web app will output the prediction results. Additionally, users will be able to explore the dataset using the "*Data Exploration*" section of the web app, allowing them to visualise relevant insights in an interactive manner. This will enhance the accessibility of our model's results and showcase its potential for live claim prediction.

5. CONCLUSION

In the conclusion section, we will evaluate the three specific goals of the project, namely, multiclass classification benchmarking, model optimization and the open-ended segment. We will describe the decision-making process for selecting the final model and mention the trade-offs made. Additionally, we will evaluate the model's capabilities in predicting the different injury types. We will go over the model's limitations and their implications.

Furthermore, we'll consider how this model could impact the New York Workers' Compensation Board (WCB), specifically by streamlining the claim review process to improve efficiency, resource use, and consistency in decision-making. We'll assess whether the model meets the performance standards for real-world use, focusing on accuracy, reliability, and resource efficiency. If the model falls short, we will suggest ways to enhance its reliability for practical applications.