Master in Data Science and Advanced Analytics

Nova Information Management School
Universidade Nova de Lisboa

MACHINE LEARNING HANDOUT

Deciding on Compensation Benefits

November 2024

Group 51

André Lourenço – 20240743
Carolina Pinto – 20240494
Daniel Caridade – 20211588
Fábio dos Santos - 20240678
Gustavo Gomes– 20240657

Fall Semester 2024/2025

This document outlines the structure and content to be included in the final report, which will be organized as follows:

Abstract - This section will provide a concise summary of the report's objectives, methods, results, and conclusions, offering a quick overview of the project.

1. Introduction - The introduction will outline the project's aim to automate claim processing for the New York WCB by developing a model to predict *Claim Injury Type* based on historical data. It will also review relevant literature, stating the most relevant insights found and highlighting project objectives like model optimization and feature analysis.

2. Data Exploration and Preprocessing

2.1. Data Exploration - In this section of the final report, we will examine the provided dataset in detail, beginning with an analysis of the variable data types to identify any misalignments, such as variables assigned incorrect data types. Next, we will address missing values within the dataset, assessing their patterns and relationships to prepare an effective preprocessing strategy. Following this, we will provide descriptive statistics for key variables, explaining their relevance to the analysis (e.g. variable *Alternative Dispute Resolution* has 99.5% of its values as 'N', making it likely irrelevant for predicting the target).
This section will also highlight any inconsistencies discovered in the data, (e.g. accidents recorded with dates preceding assembly) and conduct univariate analysis of features that will include histograms and box plots to examine distributions and identify outliers within numerical features, as well as bar plots to show frequency distributions of categorical features. Additionally, a multivariate analysis will be done to explore relationships between variables, starting with pairwise relationships between numerical features, examining correlations and scatterplots to detect multivariate outliers. For categorical features, we will use bar plots to visualize relationships between variables, identifying and documenting key insights. This analysis will conclude with an assessment of each feature's predictive value by comparing all input features against the target variable.

2.2. Data Preprocessing - In this section our group will make some changes to the data to solve incoherences, handle missing values, explore and handle outliers, evaluate and delete redundant features and do some feature engineering to help our analysis. Some of the solutions will be used removing rows with missing values in all columns, attribute correct data types, substitute values in duplicates, attribute values to missing values based in justified rules (e.g. *C-2 Date* and *C-3 Date*), based in other features (e.g. *Age at Injury* and *Birth Year*) and using imputation methods (e.g. KNNImputer or DecisionTreeImputer).

3. Multiclass Classification

3.1. Additional Features - In this section, new features will be created based on the data exploration done previously. Firstly, we will analyse additional features that will allow us to ignore and/or input missing values, such as *Hearing,* a binary indicating whether *First Hearing Date* occurred or not. Regarding date type features included in the dataset, we will create and test, for example, *time_to_assembly* that is the difference between accident date and assembly date, '*C-2 delivered on time'* that indicates if the C-2 form was received within 10 days after the accident and *'C-3 delivered on time'* that indicates if the C-3 form was received within 2 years after the accident. We will explore further ways to deal with these features as we test different models, so the feature creation process will be an interactive process. After creating the additional features, they will be analysed through univariate and multivariate analysis to assess their predictive power towards the target. This will help identify their relevance and understand multicollinearity problems with other features.

3.2. Feature Selection

3.2.1. Filter Methods - In this section, we will apply two techniques for feature selection. First, we will perform a variance test on the numerical variables, analysing the resulting variances to identify and exclude variables with variance equal to zero. Next, we will conduct a correlation analysis among the numerical variables, analysing the resulting correlations to identify redundant features. Features with a correlation bigger than 0.8 and lower than -0.8 will be considered for exclusion.

3.2.2. Wrapper Methods - In this section, our group will apply the recursive feature elimination (RFE) method using a logistic regression model to identify the most relevant numerical features. The most relevant numerical features identified so far were *Accident Date*, *C-2 Date*, *C-3 Date*, *First Hearing Date* and *Average weekly Wage*.

3.2.3. Embedded Methods - In this section, we will apply two methods for feature selection. First, we'll use Lasso regression on the numerical features, analysing the resulting coefficients to identify and exclude redundant features from the final model. Second, we'll apply a random forest algorithm to the categorical features, using feature importance scores to predict the target variable. Features with an importance score below 0.01 will be considered for exclusion in the final model. (Note: We will explain in this section why we chose this approach over the Chi-squared test for categorical feature selection.)

3.3. Model Assessment - For model assessment, we will use cross validation with 5-folds to test Decision Tree, Random Forest, Logistic Regression and Extreme Gradient Boosting algorithms to check which one predicts better the target. As we are dealing with a classification problem, precision, recall and macro-f1 score will be the performance measures used to assess the model performance. In fact, macro-f1 score gives the same importance to each class regardless of its number of instances, so it is appropriate for the type of problem addressed. Accuracy is a more general measure not very insightful to deal with an imbalanced distribution for the target variable so it will not be used. After analysing the performance measures, the algorithm with better performance will be used for model optimization.

3.4. Model Optimization - For model optimization, we will perform hyperparameter tuning. We will use Grid Search and/or Random Search to find the best set of hyperparameters to reduce both overfitting and improve the validation scores of the default instance of the best algorithm identified in the model assessment phase.

4. Open-Ended Section - We will detail the development of an interactive interface that allows users to predict the *Claim Injury Type* based on input information about a worker who has experienced an injury. This interface will be implemented as a user-friendly website using HTML, designed to provide quick and accessible predictions based on the model we have trained. We will begin by describing the objectives for creating this interface. Serving as a practical demonstration of how the predictive model could function in a real-world application, providing users with a direct way to assess potential claim outcomes. Additionally, technical details behind the interface will be detailed like how the HTML frontend is designed to capture user inputs and format them for prediction, and explain how the interface connects to the predictive model, describing the process of sending input data to the backend for processing and returning the prediction result to the user.

5. Conclusion - This section will be used to summarize our final conclusions and insights taken during the entire project.

6. References - This section will present the references cited throughout the final report.

Appendix - In this section, we will add visualizations that aided us with our analysis and other supplement content.