A poster of a machine learning

Description automatically generated

Table of Contents

[1. Introduction 2](#_Toc182087877)

[2. Exploratory Data Analysis 2](#_Toc182087878)

[3. Data Pre-Processing 2](#_Toc182087879)

[4. Feature Selection 2](#_Toc182087880)

[5. Modelling, Assessment and Deployment 3](#_Toc182087881)

[6. Open-Ended Section 3](#_Toc182087882)

1. Introduction

This project aims to develop a model to assist the New York Workers’ Compensation Board in automating decisions on new workplace injury claims. We will begin with an in-depth Exploratory Data Analysis (EDA) to gain insights that guide the model development phases. Next, we’ll conduct Data Preprocessing to prepare the dataset, followed by Feature Selection to identify the most relevant variables to train our model. Finally, in the Modeling, Assessment, and Deployment phases, we’ll evaluate several predictive models to choose the best one for unseen data. Additionally, we will include an Open-Ended Section to apply advanced techniques to further enrich the project.

2. Exploratory Data Analysis

The Data Exploration phase begins with **Univariate Analysis**, examining the main statistics and distributions of numeric features, as well as the frequency of each categorical variable. We then move to **Bivariate Analysis** to explore relationships between variables, including correlations and potential nonlinear associations among numeric features. Next, we analyze the distribution of numeric features across categorical feature groups and examine interaction effects between categorical features. As this is a supervised learning project, we will focus on each feature’s discriminative power with respect to the target variable.

3. Data Pre-Processing

Using insights from the EDA, we’ll proceed with Data Pre-Processing to clean and transform the dataset, improving data quality for model development. We’ll address missing values and outliers, filling some based on logical inference or provided information and imputing others based on data type and percentage of NaNs. High-cardinality categorical features will be grouped, and irrelevant columns removed to simplify the dataset. We’ll also apply Feature Engineering to create new features that capture meaningful relationships and scale numeric features where needed. This preprocessing stage is crucial to model performance, and if results are unsatisfactory, we may revisit and refine our approach to ensure optimal data quality.

4. Feature Selection

To select the most relevant features for our model, we will use two Feature Selection methods:

* **Filter Methods**

We’ll start by identifying **Constant** and **Quasi-Constant** **features** through variance analysis and apply **Spearman’s correlation** for numeric variable relationships. Although direct correlation with the categorical target isn’t suitable, we’ll use normalized heatmaps, stacked bar plots, and **Chi-squared tests** to assess feature impacts on the target variable.

* **Wrapper Methods**

We will also apply wrapper methods, including **Recursive Feature Elimination** (RFE) to iteratively remove the least important features and optimize the feature subset. Additionally, **Recursive Feature Elimination with Cross-Validation** (RFECV) will be used to provide a more stable feature ranking by incorporating cross-validation.

* **Embedded Methods**

We will also use embedded methods, such as **Lasso Regression**, which select important features during model training by applying regularization to reduce complexity.

5. Modelling, Assessment and Deployment

In this stage, we will experiment with different types of models, such as logistic regression and decision trees, to identify which one best captures the patterns in the dataset. To avoid overfitting and ensure a fair evaluation, we plan to use cross-validation to assess the models across different data subsets. Given the unbalanced nature of our data, where certain target categories are more frequent and others have a high number of missing values, we will use a macro-based approach for evaluation. To obtain a comprehensive understanding of each model’s performance, we will utilize multiple metrics, including accuracy and the F1 score, as these are more suitable for imbalanced datasets than accuracy alone.

6. Open-Ended Section