NOVA IMS Information Management School

Mestrado em Data Science and Advanced Analytics, com especialização em Data Science

TO GRANT OR NOT TO GRANT: DECIDING ON COMPENSATION BENEFITS

Project of Machine Learning

Ana Beatriz Farinha, 20211514
Francisco Capontes, 20211692
Laboni Raz, 20240943
Maria Cruz, 20230760
Sofia Gomes, 20240848

Fall/Spring Semester 2024-2025

Contents

1 Introduction

This work aims to create models to automate decision-making whenever a new claim is received.

2 Explore the dataset

To explore the dataset, we divided the work into 7 phases, namely:

1. **Initial Exploration:** In order to undestand the dataset in more detail, we use several functions such as head, info, shape and apply to check each row's length in the Claim Identifier column.

2. **Duplicates:** In this dataset, we first check for duplicated Claim Indentifier and then searching for duplicated rows.

3. **Missing Values:** A large proportion of the data for certain rows is missing, while others have relatively fewer missing values. Rows with more than 40% missing data may require special handling, such as imputation or removal, depending on the context and importance of the data. To understand better, we did a heatmap and a dendogram.

4. **Target Distribution:** To get to know our data, we explored better the target variable Claim Injury Type. To do so, we created a frequency bar chart and realize that there is a large discrepancy between the values. The most frequent is NON-COMP.

5. **Numerical variables:** We identify the numerical columns, saw the statistical, take conclusions about that columns and created some plots (histograms and boxplots).

6. **Categorical variables:** We identify the categorical columns, saw the statistics, take conclusions about that columns, like we did in numerical variables. Also we create some plots (plot_cases_by_county and barplots)

7. **Date variables:** We created a dataset only with date_features: Accident Date, Assembly Date, C-2 Date, C-3 Date, First Hearing Date

3 Clean and pre-process the dataset

To clean our dataset, we have to threat some types, incoherencies, missing values and outliers, that we saw in exploration part. To do so, we divided in that parts:

1. **Data Types:** We change some types in columns: date columns to data type, code columns to strings, column "Age at Injury" and "Birth Year" to int and column "Agreement Reached" to boolean.

2. **Incoherencies:** We find some incoherencies, like: some Birth date equal to 0, Person Age shoud be greather than "Age at injury", the min Age at injury shoudn´t be equal to 0, Average Weekly Wage shouldn´t be eaqual to 0 in min.

3. **Handling missing Values:** First, we identify the columns that can be dropped. After that, to distinguish some NAs from true missing values, we decided to replace them with -1.

4. **Remove outliers:** We use the method Z-score to identify the outliers based on the number of standard deviations from the mean. The threshold that we use is 3. That means that any data point with a Z-score beyond this range is considered an outlier.

4 Feature Engineering

Feature engineering is the process of transforming raw data into meaningful features that improve the performance of machine learning models.

Fist, we created new features about Date, age at injury, average weekly wage, accident date and accident timing indicator.

Then, we transform some existing features, like Alternative Dispute Resolution, Carrier Type and Gender

5 Variable encoding

To use the categorical variables, it is necessary to encode them as most models do not accept non-numerical values.

To the most categorical variables we used One Hot Encoder, because most of them are nominal and have a relative few unique categories.

In the target, Claim Injury Type, we use Label Encoder, because it is a ordinal data and it is better assign a numerical label.

6 Scaling

The technique we used to pre-processing the data was Min-Max Scaling, because the columns that we applied don´t have a normal distribution (have varying scales), so is better to normalize the data in a small range.

7 Feature Selection

To improve the model performance, by removing irrelevant or redundant features, we choose to use the filter method: Chi-square, in order to measures the dependence between categorical features and the target variable Claim Injury Type.