24 Nexus Analytics

NOVA Information Management School
Master in Data Science and Advanced Analytics
Machine Learning

To Grant or not Grant: Deciding on Compensation Benefits

DATA SCIENTIST MANAGER: ANTÓNIO OLIVEIRA,20211595
DATA SCIENTIST SENIOR: TOMÁS RIBEIRO, 20240526
DATA SCIENTIST JUNIOR: GONÇALO PACHECO, 20240695
DATA ANALYST SENIOR: GONÇALO CUSTÓDIO, 20211643
DATA ANALYST JUNIOR: ANA CALEIRO, 20240696

Fall/Spring Semester 2024/2025

Table of Contents

Abstract

This project aims to assist the New York Workers' Compensation Board (WCB) in automating the decision-making process for classifying claims by predicting the Claim Injury Type using machine learning models built from historical claims data from 2020 to 2022. The initial phase involved data exploration and preprocessing of a dataset with 59,347 rows and 23 columns. Our team started by checking incoherencies, followed by feature engineering, which included data type conversions, encoding categorical variables, and analyzing pairs of features to reduce redundancy, and more.

The Train-Test split was done using the Holdout Method. For handling missing values and outliers, our team applied techniques like the median and Interquartile Range (IQR) to identify outliers, considering logarithmic transformation for specific variables.

For now, we have chosen to use RobustScaler to scale our data set, due to the presence of outliers. The feature selection strategy involved experimenting with Filter-Based, Wrapper, and Embedded methods, choosing features consistently identified as relevant.

Our team identified this type of problem as a classification one. We experimented with models like Logistic Regression, Random Forest Classifier and more, as well as combining models and using hyperparameter tuning with Random Search.

Our Open-Ended section of the project proposes developing an analytics interface for client use, featuring predictive capabilities based on user inputs, providing an intuitive experience.

1. Introduction

This project aims to assist the New York Workers' Compensation Board (WCB) in automating the decision-making process for classifying claims by predicting the Claim Injury Type. The task involves creating and optimizing machine learning models using historical data from claims assembled between 2020 and 2022.

2. Data Exploration and Preprocessing

To understand our initial dataset of 59,347 rows and 23 columns, we began by examining general statistical summaries of the features, as well as identifying any obvious incoherencies. For example, we found entries with ages over 100, which doesn't make sense for someone in the workforce, so we filtered them. We also explored the distributions of our data through visualizations, such as histograms.

Our team then moved on to data cleaning and preprocessing. The first step was to check for duplicates and remove them. In the feature engineering phase, we applied necessary transformations that we would also use on the features in the Test dataset. We started with data type conversions: transforming date features to date types, encoding categorical features into numeric ones, and so on. For some types of conversions, we have different options, and we will make final decisions on these during the feature selection phase. Pairs of features were also analyzed to reduce redundancy of the data set. For the Train-Test split we dropped the target variable and used the Holdout Method. For the future we pretend to experiment with the K-Fold Cross-Validation Method, specifically with K=10, but that is something that will be further examined since we already encountered some challenges after trying it. After the Train-Test split, we proceed to deal with the missing values: input calculations, filling with 0 and so on. All these alterations are going to be applied in X_train, X_validation and the Test Data in the same section. Regarding the outliers, we already knew that some variables stood out by looking at the box plots and histograms. To address this issue, we used the statistical measure of dispersion known as the Interquartile Range (IQR) and identified several columns with more than 10% outliers. Rather than automatically removing these columns and losing a significant amount of information, we examined them more deeply. We concluded that it might be beneficial to apply a logarithmic transformation to the IME-4 Count variable, however, we will evaluate this in feature selection to determine if keeping this transformed feature is useful. For binary, categorical, and code variables, we decided against this transformation, as it wouldn't make sense to make this type of transformation. In the next phase of our project, we plan to experiment by converting outliers into missing values and then try to fill them.

3. Multiclass Classification

Our feature selection strategy involved experimenting with Filter-Based, Wrapper, and Embedded methods, choosing features consistently identified as relevant. Despite the insights from these methods, we still had to rely on the old trial-and-error approach, which we plan to reduce in the future by sharpening our feature selection. Our final features for this stage are: 'Age at Injury', 'Average Weekly Wage', 'Assembly Year', 'C-2 Month', 'C-2 Year', 'First Hearing Year', 'IME-4 Count Log', 'Attorney/Representative', 'Carrier Name', 'Carrier Type', 'County of Injury', 'District Name', 'Gender', 'Industry Code', 'Medical Fee Region', 'WCIO Cause of Injury Code', 'WCIO Nature of Injury Code', 'WCIO Part Of Body Code', 'C-3 Date Binary'. Of course, before applying these methods, we had to scale our data. We experimented with the StandardScaler, the MinMaxScaler and RobustScaler. For now, we have decided to use the last option (RobustScaler), as our data set contains outliers, and we believe this is the best way to address them. However, we will keep in mind the StandardScaler, if we choose to use an algorithm that depends on normalized features. In respect to the type of problem we have in our hands, we defined it as a classification problem. We started by experimenting with models introduced in class, such as Logistic Regression and Decision Tree Classifier, and then moved on to try additional models we found through research. In fact, the one who got the best score was a simple Random Forest with 0.41072 Macro F1-Score on Kaggle. We also attempted to combine models and use hyperparameter tuning with Random Search, however, this is something we want to explore more deeply in the future. To assess the score, in our notebook, we used the Classification Report, since it includes precision, recall, F1-score, and support for each class.

4. Open-Ended Section

For the open-ended section, our team thought it would be interesting to develop an analytics interface that provides predictions based on new user inputs, as was stated on the project guidelines. The main goal is to deliver a quick, intuitive experience, allowing clients to access and understand the output generated by our model, without the need to reach out to us. To make the results user-friendly and easy to interpret, we are considering including both static and interactive visualizations. If feasible, we plan to incorporate a fraud detection and claim validity checker, or a feature to assess the likelihood of a claim being represented by an attorney, which could cause complications for our clients. Our team pretends to host this interface in a web application using python and the *streamlit library*. The “tabs” we want to create are the grant or not grant compensation model, another for visual exploration of the data and, if possible, an additional simple predictor.

5. Conclusion

The strategy we plan to implement in our project involves increasing the complexity and detail in all areas as we deepen our understanding of the subject, along with the classes. We intend to maintain the same order of the topics discussed, but in a lot more depth.

6. Annexes

Table 1 – Models comparison

Models

Model	Feature Selection	Parameters	Kaggle Score
Voting (sgd_rf_dt_gb_ab)	2	-	0.37300
Stacking (sgd_rf_dt_gb_ab)	2	-	0.40255
RF (agedrop_ime4drop_birthyear_drop_ime4log)	3	-	0.41072
Voting (sgd_rf_dt_gb_ab), (agedrop_ime4drop_birthyear_drop_ime4log)	3	-	0.34477
RF (all_scaled_new_encoding_agedrop_ime4drop_ime4log)	4	-	0.39087
RF (all_scaled)	5	-	0.38734
Voting (all_scaled)	5	-	0.37536
RF (all_scaled)	3	-	0.38814
Voting (all_scaled)	3	-	0.31799

Models K-Fold

Model	Feature Selection	Log	Parameters	Kaggle Score	Fold
LogReg	-	-	-	0.21122	5
RF	1	X	-	0.29078	5
XGB	1	X	-	0.20642	10
RF	-	-	-	0.26616	5

Models w/ Stratified K-Fold

Model	Feature Selection	Log	Parameters	Kaggle Score	Fold
RF	-	-	-	0.26912	10
DT	-	-	-	0.14236	10
DT	-	X	-	0.15589	10

NOVA IMS Information Management School

Table 2 – Feature Selection

Numeric Variables

Variable	Variance	Correlation	RFE LR	RFE RF	Lasso	Extra Trees	Decision
Accident Day	keep	keep	**discard**	keep	**discard**	keep	try
Accident Month	keep	keep	keep	keep	**discard**	keep	keep
Accident Year	keep	keep	keep	keep	keep	keep	keep
Age at Injury	keep	?	keep	keep	**discard**	keep	**discard**
Assembly Day	keep	?	keep	keep	**discard**	keep	try
Assembly Month	keep	?	keep	keep	**discard**	keep	**discard**
Assembly Year	keep	keep	keep	**discard**	**discard**	keep	**discard**
Average Weekly Wage	keep	keep	keep	keep	keep	keep	keep
Birth Year	keep	?	keep	keep	keep	keep	keep
C-2 Day	keep	?	keep	keep	**discard**	keep	try
C-2 Month	keep	?	keep	keep	keep	keep	keep
C-2 Year	keep	keep	keep	keep	keep	keep	keep
First Hearing Year	keep	keep	keep	keep	keep	keep	keep
IME-4 Count	keep	?	keep	keep	keep	keep	try
IME-4 Count Log	keep	?	keep	keep	keep	keep	try
Number of Dependents	keep	keep	**discard**	keep	**discard**	keep	**discard**

Categorical Variables

Variable	Chi-Squared	MI	Extra trees	Decision
Alternative Dispute Resolution Enc	keep	**discard**	**discard**	**discard**
Alternative Dispute Resolution freq	**discard**	**discard**	**discard**	**discard**
Attorney/Representative Bin	keep	keep	keep	keep
C-3 Date Binary	keep	keep	keep	keep
Carrier Name freq	keep	keep	keep	keep
Carrier Type freq	keep	**discard**	**discard**	**discard**
Carrier Type_1A. PRIVATE	keep	**discard**	**discard**	**discard**
Carrier Type_2A. SIF	keep	**discard**	**discard**	**discard**
Carrier Type_3A. SELF PUBLIC	keep	**discard**	**discard**	**discard**
Carrier Type_4A. SELF PRIVATE	keep	**discard**	**discard**	**discard**
Carrier Type_5. SPECIAL FUND	keep	**discard**	**discard**	**discard**
County of Injury freq	keep	**discard**	keep	try
District Name freq	keep	**discard**	keep	try
District Name_ALBANY	keep	**discard**	**discard**	**discard**
District Name_BINGHAMTON	keep	**discard**	**discard**	**discard**
District Name_BUFFALO	keep	**discard**	**discard**	**discard**
District Name_HAUPPAUGE	keep	**discard**	**discard**	**discard**
District Name_NYC	keep	**discard**	**discard**	**discard**
District Name_ROCHESTER	keep	**discard**	**discard**	**discard**
District Name_STATEWIDE	keep	**discard**	**discard**	**discard**
Gender Enc	keep	**discard**	**discard**	**discard**
Gender_F	keep	**discard**	**discard**	**discard**
Gender_M	keep	**discard**	**discard**	**discard**
Gender_U	keep	**discard**	**discard**	**discard**
Industry Code	keep	keep	keep	keep
Medical Fee Region freq	keep	**discard**	keep	try
WCIO Cause of Injury Code	keep	keep	keep	keep
WCIO Nature of Injury Code	keep	keep	keep	keep
WCIO Part Of Body Code	keep	keep	keep	keep
Zip Code Numbers	**discard**	**discard**	keep	**discard**
Zip Code Valid	**discard**	**discard**	**discard**	**discard**

Table 3 – Final Features

`Final Features (Numeric & Categorical)`

Variable
Accident Month
Accident Year
Average Weekly Wage
Birth Year
First Hearing Year
--
Attorney/Representative Bin
C-3 Date Binary
Carrier Name freq
Industry Code
WCIO Cause of Injury Code
WCIO Nature of Injury Code
WCIO Part Of Body Code

Try
Accident Day
Assembly Day
C-2 Day
IME-4 Count
IME-4 Count Log

County of Injury freq
District Name freq
Medical Fee Region freq