MACHINE LEARNING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

To Grant or Not to Grant: Deciding on Compensation Benefits

Group 15

Gaspar Pereira, 20230984

Hugo Trigueiro, 20240577

Maria Quita, 20240749

Virgílio Ferreira, 20240689

Fall/Spring Semester 2024-2025

Table of Contents

1. Abstract

Since 2000, the New York Workers' Compensation Board (WCB) has handled more than 5 million claims involving occupational injuries, which is a very time-consuming process. In order to facilitate decision-making, this work attempts to create a machine learning model for multiclass classification of injury kinds. We addressed issues including data imbalances, missing values, and feature redundancy by concentrating on "Claim Injury Type" as the main target using a dataset that contained 593,471 entries and 30 features. Important preliminary results showed that middle-aged people make up the majority of claims, and that severe injuries are associated with higher average weekly salaries, longer reporting delays, and the presence of lawyers or IME-4 reports.

We used sophisticated preprocessing methods, such as frequency encoding, log transformations, and custom ordinal encoding, to engineer and choose features. In order to optimize model performance, the feature set was reduced to 24 variables using Recursive Feature Elimination (RFE) and Spearman correlation. Using a stratified cross-validation pipeline, models including Random Forest, Logistic Regression, and CatBoostClassifier were assessed; CatBoost performed better than the others, achieving the highest F1 macro score. Performance was significantly enhanced by threshold optimization and hyperparameter adjustment, especially for underrepresented classes.

With the exception of fatal injuries, when wages were less significant, "Average Weekly Wage_log" was identified as the most important predictor by feature importance analysis utilizing SHAP and CatBoost characteristics. "Days to First Hearing," "Attorney/Representative," and "IME-4 Count," particularly for serious cases, were further noteworthy elements. With possible applications in other domains requiring multiclass classification, this study highlights the significance of feature engineering in imbalanced datasets and offers a reliable, scalable strategy for automated injury categorization. To improve predicted performance even more, future research could investigate different models and improve encoding techniques.

2. Introduction

The New York Worker's Compensation Board (WCB) is responsible for deciding on injury claims whenever it becomes aware of a workplace injury. However, since 2000 the WCB has assembled and reviewed more than 5 million claims, which is very time-consuming. Therefore in this project, we aim to develop a machine learning algorithm for multiclass classification of the type of injury that should be given to WCB claims. We also aim to analyze relative feature importance for each Claim Injury Type

The study by Mathews (2016) explores the application of machine learning algorithms to predict the severity of workers' compensation claims however for binary classification. Research emphasizes the critical role of feature engineering in enhancing the predictive performance of machine learning models for workers' injury claims. However, class imbalance is a significant challenge in predicting worker injury claims, particularly when severe injuries are less frequent but carry higher stakes. Previous studies have demonstrated the effectiveness of strategies like class balancing, stratified sampling, and weighted algorithms (e.g., Support Vector Machines with class weights) in addressing this issue. These techniques ensure that predictive models are not biased towards the dominant class, resulting in improved sensitivity and specificity for identifying severe claim types. With this, we expect also to handle target class imbalances with some of these strategies.

3. Data Exploration

The dataset was composed on 30 features and 593,471 rows, and 3 target variables, but the main target was Claim Injury Type. The Dataset was form the New York Worker's compensation Board. Here we will present the features of the dataset that was proposed and analyse them individually.

Age at Injury: With middle-aged people being the majority of cases, is in line with the age range that is usually employed, when accidents at work are more common. Additionally, the study found that older people typically sustain more serious injuries, as evidenced by higher values of the target variable "Claim Injury Type." Since people in these age categories are unlikely to be employed, figures below 14 and over 70 years old were regarded as outliers to ensure data relevance. We can see the significant changes in "Figure 1" and "Figure 2". **Birth Year:** Outlier figures, which are probably the result of data entry errors, are among the possible problems with the data that are displayed. Most claimants were born in the middle to late 20th century, according to the median year of birth. 5.12% of the values were missing, but we didn't address them because we ended up removing this feature because it had the same information as the variable "Age at Injury". One outlier was birth year 0, which doesn't make sense, so we fixed it so that the minimum was 1900-01-01.

Average Weekly Wage: The significantly skewed right distribution of this statistic indicates that most claimants have lower incomes. To resolve the 4.99% of missing data for this variable, the

median value was imputed. Without altering the skewness of the variable, this technique maintains the data's central tendency.

IME-4 Count: The "IME-4 Count" variable shows a small positive skew, with higher values associated with more severe injuries. Due to missing values, imputation with a value of 0 was required, confirming missing IME-4 reports. As seen in the “Figure 3”, the study clearly shows a trend where greater IME-4 concentrations are associated with more severe injuries.

Number of Dependents: Since the "Number of Dependents" variable has an even distribution, extreme values are probably rare. It has a tiny negative correlation (-0.001%) with the target variable ("Claim Injury Type"), suggesting that it has little effect on injury type.

Industry Code Description: "HEALTH CARE AND SOCIAL ASSISTANCE," "PUBLIC ADMINISTRATION," and "RETAIL TRADE" are the industries that report the greatest injury frequencies. This trend may be explained by the inherent physical demands of these sectors. A total of 1.73% of the missing data were imputed using the category "Other". Interestingly, as the accompanying attachment “Figure 4” shows, the non-compensable injury category is the most common across all industries.

OIICS Nature of Injury Description: There is a substantial gap in the data collection process because this variable is completely absent from the dataset. Future analysis could be improved by addressing this missing.

WCIO Cause of Injury Description, WCIO Nature of Injury Description and WCIO Part Of Body Description: “LIFTING", "STRAIN OR TEAR", "CONTUSION" and "SPRAIN OR TEAR", “LOWER BACK AREA”, “KNEE”, “MULTIPLE AREAS” and “SHOULDERS” injuries are the most common, highlighting the dangers of hard labor. A total of 2.72% in “WCIO Cause of Injury Description”, 2.73% in “WCIO Nature of Injury Description” and 2.98% in “WCIO Part Of Body Description” of the variables were imputed as "Missing". We could conclude that injuries brought on by "PANDEMIC" are linked to higher severity as we can see in the attachment “Figure 5”. The visualization “Figure 6” shows that "STRAIN OR TEAR" is more common in injury category “6. PPD NSL” and according to the analysis “Figure 7”, FINGER(S) injuries are common in "2. NON-COMP " situations.

Carrier Name and Carrier Type: Due to its extensive use, the " STATE INSURANCE FUND" and “PRIVATE” are the insurance carrier that is most frequently reported. Interestingly, cases involving " NEW HAMPSHIRE INSURANCE CO" are more likely to involve the "2. NON-COMP" injury classification (attachment “Figure 8”). "2. NON-COMP " claims are more often when the carrier is "UNKNOWN” (attachment “Figure 9”).

County of Injury: “SUFFOLK” county has the most claims, which can be a sign of higher injury rates in this region.

District Name: Given the greater population and greater number of workplaces in the NYC district, it makes natural that the majority of claims originate from this area. As seen in “Figure 10” attachment, “ROCHESTER” had the most “2. NON-COMP” cases.

Gender: The majority of claimants are men, which is consistent with their propensity for high-risk jobs. Moreover, most claimants in situations that result in “8. DEATH” are men as we can see in the attachment “Figure 11”.

Medical Fee Region: “IV” is the most prevalent medical fee area. This could be because of regional medical service availability or local cost arrangements. The attachment “Figure 12” shows that the “UK” had more “2. NON-COMP” cases.

Zip Code: “11236” is the most prevalent zip code, which reflects local workplace danger and population density. 4.99% of the values were missing, and they were imputed with "missing" category.

Alternative Dispute Resolution: Most disputes are settled without the need for further legal processes, as evidenced by the fact that most claims do not include this method. As we can see in the “Figure 13” attachment, it was nearly entirely “NON-COMP” instances when it was “True”.

Attorney/Representative: The majority do not include an Attorney/Representative, indicating that these claims are usually resolved without the need for legal counsel. As we can see in the attachment “Figure 14”, it is nearly entirely “True” that there was an attorney or representative present when the injury type was “7. PTD”.

The COVID-19 Indicator: Most claims have nothing to do with COVID-19. On the other hand, injuries that belong to types “7. PTD” and “8. DEATH” are typically more severe when this signal is “True” as we can see in the attachment “Figure 15”.

Accident Date: The concentration of accident dates around 2021 indicates a rise in claims within this time frame. The lack of older records is probably the result of insufficient historical information. Since derived features were given priority, missing data (0.64%) were not handled. **Assembly Date:** Is mostly concentrated in the latter part of 2021, especially in July. This date most likely relates to the claims' processing or compilation. The latest assembly takes place by December 2022, with the earliest date beginning in January 2020. **C-2 Date and C-3 Date:** Most dates lie between 2021 and 2024, with most dates concentrated around mid-2021. Of the values, 2.54% were absent in “C-2 Date” and 67.38% in “C-3 Date”. Since the features won't be used and will be used to construct further features, we didn't treat the missing values. We can infer that few deliver C-3 dates and most deliver C-2 dates. **First Hearing Date:** This date indicates a focus in early 2022, specifically in March. The dates, which range from January 2020 to June 2024, show how claims have continued to develop over time. The 73.73% of missing values are justified by the fact that the initial hearing is only held in the most serious instances.

Claim Injury Type: this was our main target. A sizable percentage of claims fall under the “NON-COMP” category, suggesting that many injuries may not cause reimbursable losses or long-term

disability. More severy injury types are less present since they are rarers. Their relative distribution is this dataset can be seen on figure 16 in the annexes. In order to obtain the target variable for our prediction, we dropped the missing values because we are using a supervised learning method, and those values would not be useful for our project. As shown in “Figure 17”, COVID-19 had a minor impact on the total number of accidents, with its effects being more noticeable in the first trimester of 2020, peaking in April 2020. From the image, it is evident that COVID-19 had a greater impact on Claim Injury Types 4. Temporary and 2. Non-Comp. The data also clearly shows that accidents peaked early in the time frame, stabilized thereafter, and then declined sharply toward the end.

WCB Decision: this was a secondary target that happeared to be misslabeled since every entry said that the WCB decided that the injury was not work related. **Agreement Reached:** this was also a secondary target. Most claims do not lead to agreements, underscoring the intricacy and frequent disagreements surrounding workers' compensation claims.

For this project we created new features. We present the new feature and how they were created in table 1. The analysis of the new features revealed some insights about the features and their relationships to claim injury types. Starting with the **Accident Date Missing** where we could check that 99,35% of the cases had an accident date and 0,065 not. It this feature, it was observed that when this value is missing, the claim tends to be canceled as shown in “Figure 18”

The **Accident Date_weekend** feature indicated that most accidents occur on weekdays, with a ratio of 0.8 to 0.2, which is logical given the nature of work-related claims. When examining the distribution of accidents by month, it was evident that April had the highest proportion of deaths, a trend that may be linked to external events such as the COVID-19 pandemic. Seasonal analysis aligned well with established literature, showing that summer months are associated with a higher frequency of accidents, whereas winter sees fewer severe claims, including deaths. “Figure 19”, “Figure 20”, and “Figure 21”.

The feature **Accident Date_assembly_gap_days** revealed that more severe claim types, particularly **[7. PTD and 8. DEATH]**, tend to have larger gaps between assembly and accident dates. This may indicate delays in reporting or processing for these critical cases, reflecting their complexity or the nature of the claims. “Figure 22”. Patterns in missing data were also revealing, with the C2 feature having 2.54% of its values missing and 97.46% not missing. When the C2 feature is missing, claims tend to be canceled, suggesting the importance of this feature in the claims process. “Figure 23”. Similarly, the C3 feature has 67.38% of its values missing and 32.62% not missing; this feature is associated with a higher prevalence of non-comparticipated and death-related claims. The Hearing Date_missing feature is true in 73.73% of the cases, while 26.27% of cases have a hearing date. When the hearing date is not missing, claims are more likely to involve severe injury types, including [5. PPD SCH LOSS, 6. PPD NSL, 7. PTD, and 8. DEATH], suggesting that hearing dates are more frequently associated with critical cases. “Figure 24”.

Sin, cos and log transformations: We applied sine and cosine transformations to **Accident Date_month** to capture the cyclical nature of months, avoiding misleading relationships. The

cosine transformation ensures that December and January are treated as close neighbors, rather than being incorrectly perceived as one and twelve units apart. Additionally, we applied a logarithmic transformation to **Average Weekly Wage** to reduce skewness and improve model performance by stabilizing the variance.

For the **Accident Date_assembly_gap_months** feature, we updated the Accident Date column for rows where Accident Date_assembly_gap_months was negative by replacing it with the earliest date from C-2 Date and C-3 Date. If this new feature was greater than 0, the Accident Date remained unchanged. This process was conducted to gain better insights into the variable, as it originally contained 1,407 negative values**.** This variable was dropped because we decided to go for **Accident Date_gap_days** which had better results for our model

The **Accident Date_assembly_gap_days** feature exhibited highly skewed behavior, so we applied a threshold to cap its values at a maximum of 30 days. Additionally, it had 3,696 missing values, which were filled with the maximum value (30). Analysis of this feature revealed that when the interval is higher, claims are more likely to be either canceled or a death-related injury type.

To improve the analysis of **Accident Date_year**, we applied a threshold to reduce the skewness observed in the histogram. Specifically, if the **Accident Date_year** was earlier than 2019, it was replaced with 2019, and values for 2023 were replaced with 2022. The analysis of **Accident Date_year** highlighted that 2022, compared to 2020, experienced a decline in more severe claim injury types, potentially reflecting external factors such as changes in workplace policies or reduced exposure to risks during that year.

The **C3-C2_gap_days** feature exhibited extreme values on both positive and negative sides, which were sparsely distributed. To enable better analysis, we applied a threshold, capping the lower value at -60 days and the upper value at 60 days. When this feature is negative, it indicates that C2 was delivered first. After this process we see that the median dropped a lot and is negative in every claim injury type and in the first three claims (1. CANCELLED, 2. NON_COMP, 3. MED ONLY) as well as death the median drop to -60 so we can conclude after this transformation that on average the C2 form were delivered before the C3

The **C2_accident_gap_weeks** feature initially exhibited extreme values on both the positive and negative sides. To enable better analysis, we applied a threshold, at -4 weeks and at 24 weeks. When this feature is negative, it indicates that the C2 date occurred before the accident date. After this process, we could conclude that the majority of were in the first few weeks of the accidents and the bigger gaps were in the less extreme cases where the claim was mostly cancelled or non-comp, and the negative gaps were mostly on TEMPORARY and PPD SCH LOSS.

Similarly to the previous feature we used the same process for **C3_accident_gap_weeks**, applying the same threshold gap. We concluded that in the more extreme injury cases the C3 was delivered closer to the accident date as opposed to the others where it was more evenly distributed. And opposed to the C2_accident_gap_weeks where the median was 0 weeks for every claim except cancelled which was the cap, here the median was 4 weeks for every claim except PTD.

Similarly from what we done above **Hearing_C3 gap months** was once again with very extreme values to the right and left so we applied a threshold of -20 and 50 as lowest and highest values in order to have a better understanding of our variable.

After appling this threshold we were left with 464.118 missing values so we used the max value of this variable to fill missing values. With this being done we concluded that the hearing gap tends to increase with the severity of the claim type, indicating a correlation between claim complexity and the time required for hearings. Simpler cases such as non-compensated and medical-only claims tend to have smaller or more consistent gaps.

The **Hearing_C2 gap_months** feature was highly skewed to the right, with extreme values on both sides. To reduce the skewness, we applied a threshold, setting the lower value to -20 and the maximum value to 50. Additionally, due to the presence of 424,111 missing values, we decided to use the maximum value in this column as the imputation rule. After applying these adjustments, we concluded that the Hearing_C2 gap_months feature varies widely across claim types, with severe claims tending to have shorter hearing gaps, while canceled and temporary claims experience longer delays. This observation can be explained by the tendency for severe claim injury types to have a hearing date, whereas lighter claim types often do not, as also reflected in the Hearing_Date_missing variable.

For **Hearing_assembly_gap_months**, we began to remove every negative value by adding on year to every negative date, because there couldn't be a hearing before the injury. We still had 423,247 missing values so we chose to fill with the maximum value. After this transformation, we can see that it mostly after the first 4 injury claims and the more serious ones didn't have a big impact.

Days to Assembly feature boxplot highlights a strong relationship between claim severity and assembly time, where more severe claims are processed faster and with greater consistency, while less severe or administrative claims face longer delays and variability in processing If we observe this in terms of days, we can say that severe claim types have lower median days compared to lighter cases.

Days to First Hearing had 423,247 missing values, which we filled with the maximum value in this feature. With this feature, we concluded that severe claim injury types tend to have lower Days to First Hearing values, meaning that these cases are prioritized and processed more urgently compared to less severe or administrative claims, which tend to experience longer and more variable delays.

4. Preprocessing and Feature Selection

For feature selection and additional preprocessing, we used the following methods described in figure 25.

To note date we dropped beforehand datetime features, since there were already other features (new features) that captured the information of the datetime features. To ensure that the preprocessing and that the feature selection were not influenced by data leakage we did them inside a cross-validation loop.

For categorical encoding we created a function that calculates the mean of the target variable (Claim Injury Type encoded) for each category inside each categorical feature on the training set, then encodes the categories with like Label Encoder, but in an ordinal way, based on the mean

of the target variable for the training set. If a category has the same mean for the target variable it is encoded in alphabetical order. This way, we try to ensure that even if two categories have the same mean for the target variable, they are encoded differently, since they could have different distributions for the target (one variable that is associated with targets 3, 3, 3 as target, is different from another variable that has is associated with 1, 3, 5, even if the mean is the same). At the same time using the mean would imply that the target was continuous meaning that the distance between target 2 (Non comp) and target 3 (Med only) is the same as the distance between target 7 (Permanent total disability) and target 8 (Death), which probably is not the case, so using label ordered encoding we try to negate this problem. For categories in validation that were not presented in training we encoded them as 1, since we assume that infrequent categories are more likely to be canceled or non-compensated (low target mean value).

We are aware of at least one limitation of this approach: (1) in this "version" of target ordinal encoding we didnt use cross validation for encoding of the train set categories, making that the in sample predictions will be inflated (some data leakage). In future work we would like to implement a cross-validation strategy for the encoding of the training set. On the other hand, as we are using an outside loop cross-validation we ensure that each validation set is encoded blindly, so for the out of sample predictions, there is no data leakage. However we could confirm that this approach got us better results (on validation) than simple target encoding based on the mean using sklearn's TargetEncoder.

We also encoded categorical featues with frequency encoding, because the frequency of a category retains different information than the mean of the target variable, and could also be valuable for the model. In table 2 in the annexes we can see that different encodings results in different bivariate associations with targets. We used the normalized frequency of each category for the encoding in the train_set and imputed that value in validation. For categories in validation that were not presented in training we encoded them as 0 since we assumet that categories infrequent enough to not be present in the training could be considered as a category that is not present in the training set (0% present). In the end of the encoding we dropped the original categorical features, and got a total of 51 features.

For numerical imputation of the missing values we imputed as it was already we just present here a schema to summarize the imputations for the numerical features, done only during the cross-validation loop, to ensure that the mean, maximum and minumum are calculated only for the training set, and applied to the training set.

For feature selection we based on two methods we used only two methods - a filter (Spearman correlation) and a wrapper (Recursive feature elimination - RFE) - since the last is computationally expensive (RFE), but at the same time is a good estimator of feature importances. In spearman correlation if there were two pair of feature with a spearman correlation higher than 90% we would drop the feature with higher mean correlation with the other features. With redudant features droped we got 46 features. For RFE we used RandomForestClassifier (with max_depth of 6, n_estimators of 50, and max_samples of 0.8, to make it faster) as the estimator for feature importances, and used a loop to find the optimal number of features to keep recursivly in an increasing order (from low number of features untill number of columns) to find the one with the highest validation score. We used majority voting

for the 3 cross validation (feature choosen by the RFE estimator at least in two cross validation). In the end 24 features we selected. The features results are outlied in table 6. To consider that initially we used CatBoostClassifier (with 100 iterations) as the estimator for RFE, but since we made some last minute alterations to some encodings we had to change the estimator to RandomForestClassifier since it was faster. We believe that the results would be more precise if we used CatBoostClassifier as the estimator for RF, since we knew it was the model that we would use in the end. In the end the final features selected were (24): 'Attorney/Representative', 'IME-4 Count', 'Accident Date_year', 'Accident Date_assembly_gap_days', C3-C2_gap_days', 'C2_missing', 'C3_missing', 'C3_Accident_gap_weeks', 'Hearing_C3 gap_months', 'Hearing_C2 gap_months', 'Days to Assembly', 'Days to First Hearing', 'Average Weekly Wage_log', 'Carrier Name_encoded', 'Carrier Type_encoded', 'Industry Code Description_encoded', 'WCIO Cause of Injury Description_encoded', 'WCIO Nature of Injury Description_encoded', 'WCIO Part Of Body Description_encoded', 'Carrier Name_freq', 'Carrier Type_freq', 'Industry Code Description_freq', 'WCIO Nature of Injury Description_freq' and 'WCIO Part Of Body Description_freq'.

5. Multiclass Classification

We have done model assessment and optimization strategy in the same loop, with the stragies presented in figure 27.

For models we tried to cover different types of algorithms, selecting: Logistic Regression, Gaussian Naive Bayes, Multilayer Perceptron Classifier, Random Forest Classifier, CatBoostClassifier. To try to enhance the performance of the models we used OneVsRestClassifier for each estimator (altough we only present results for RandomForestClassifier which was the one who benefitted from this method), and StackingClassifier for defined combinations of two of the best performing estimators.

Inside a 3 fold stratified cross validation loop we preprocessed train and validation: encoded categoricals (with target ordinal encoding and frequency encoding), using training reference values for validation encoding, and then selected from all the generated features (51) only the best estimated features in the steps previously described (24). For numerical missing values we imputed with the same strategy done in feature selection, and for some models evaluation (Logistic Regression, MLPClassifier and GaussianNB) we also performed numerical normal standardization (StandardScaler).

After the preprocessing we trained each model with every possible combination (grid seach method) for the hyperparameters we defined presented in the following table 3.

For model performance assesment we used f1 macro score, to be in accordance with the competition metric. For performance metric (f1 macro) improvisation, we tried first to improve each class f1 score by (1) first finding the probability threshold for each classthat maximized the f1 score for that class (balancing precision and recall), and then (2) normalizing each sample class probability with the best probability threshold found. The sample class choosen was the one

with the highest normalized probability. For example if for sample 1 the initial probabilities were [0.1, 0.2, 0.7] (imagining that we only had 3 classes) the initial model predictions would be class 3, but if the best threshold for each class were [0.2, 0.1, 0.5] the normalized probabilities would be [0.1/0.2, 0.2/0.1, 0.7/0.5] = [0.5, 2, 1.4], and the class choosen would be class 2. This is a strategy to improve minority class predictions, since the model would be more prone to predict the minority for whom the probability best threshold is lower, and then the normalized probability would be higher.

In the table 4 we present the results of the models trained with the best hyperparameters found in the grid search method. We also present in the table 5 the f1 score of each best model for each class. As we can see CatBoost was the model with best mean f1 macro score in validation sets. This had to do probably due to the fact that it is an ensemble method, accounting the mistakes of the previous trees and learning the patterns for each class better. Another factor that could also account for this results is that we otimized more hyperparameters for CatBoost than for the other models. Having high number of iterations (1000) (and learning better decision borders for imbalanced classes) and an adjusted class weight of SqrtBalanced (adjusting to the class imbalance of the target) but not so much to overfit the model.

While MLPClassifier also works like an "ensemble" making it learn and adapt to the data patterns, maybe we couldn't find the best hyperparameters for it. Stacking diverse models (CatBoost + (NB or MLP)) lowered the f1 score for unbalanced classes, lowering the f1 macro score.

6. Open-Ended Section

In this section we focused on the analysis of the feature importances captured by the model for predicting the this multiclass target. We also analyzed the feature importances respective to each class on the target.

To capture the global feature importance we used CatBoost attribute get_feature_importance() resulting in the graph 28. We can see that the most important feature is by far the 'Average Weekly Wage_log'. We can also see that for the categoricals encoded Part Of Body Description was the most important categorical (more if target ordinal encoded, than frequency encoded). Generally the target ordinal encoded features were more important than the frequency encoded features, with exception of the "Carrier Type" and "Industry Code Description" that were more important when frequency encoded.

Regarding the analysis of the feature importances for each class we used the SHAP library to calculate the mean absolute shap values (all samples) for each feature for each class. Higher absolute shap values mean that the feature is more important for the prediction of that class. We present the results in the following figure 29. We can see that Average Weekly Wage_log is the most important for all classes with exception for Death claims. This could be due to the fact that that the Wage is important for supporting the claim costs (specially the more severe types as we saw on EDA). For death the wage was generally null (the person that died didnt had current wage) so the wage was imputed with the mean wage, turning this feature low variance for this class and therefore with lower predicting ability.

Days to First Hearing (from the first hearing date in the train data) was also an important feature for predicting the more severe claims (PPDNSL, PTD and Death) since as we saw in the EDA the more severe claims were first attended in hearings.

Attorney/Representative and IME-4 count was also important for predicting every class, specially the more severe ones. This could be due to the fact that the more severe claims are more likely to have an attorney or representative, and to have more IME-4 reports.

7. Conclusion

With this project, we created a robust model that could predict the claim injury types from WCB. We developed and created new features, and encoded the categorical features using two strategies (novel Target Ordinal Encoding and Frequency encoding). We used a CatBoostClassifier model that was the best performing model (with a mean validation f1 macro score of 48,5) in all validation sets and was the model that could handle better the imbalanced target. We also analyzed the feature importance for the model and found that the most important feature was the Average Weekly Wage was the single most important feature for almost every injury claim and also studied the relationship between each feature for respective claim types. For future work, we could maybe restrict the categorical encoding to only target ordinal that seems to be associated better with injury claim types. We could also improve on hyperparameter optimization.

8. Appendix

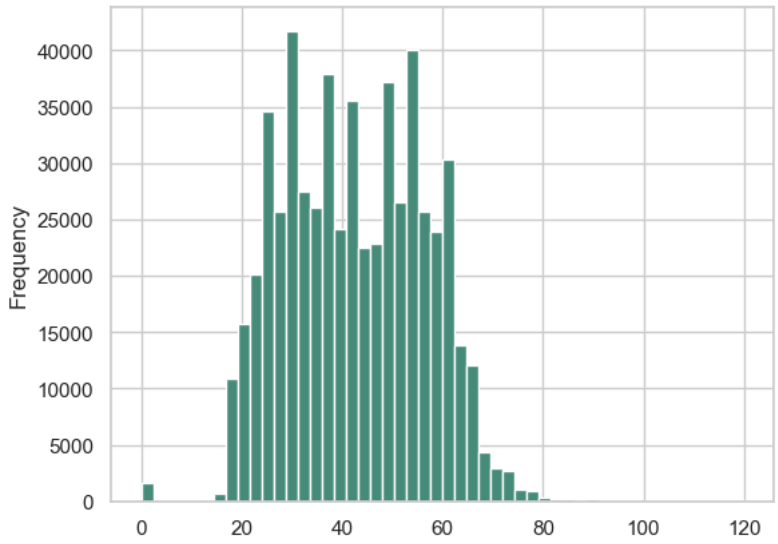

Figure 1

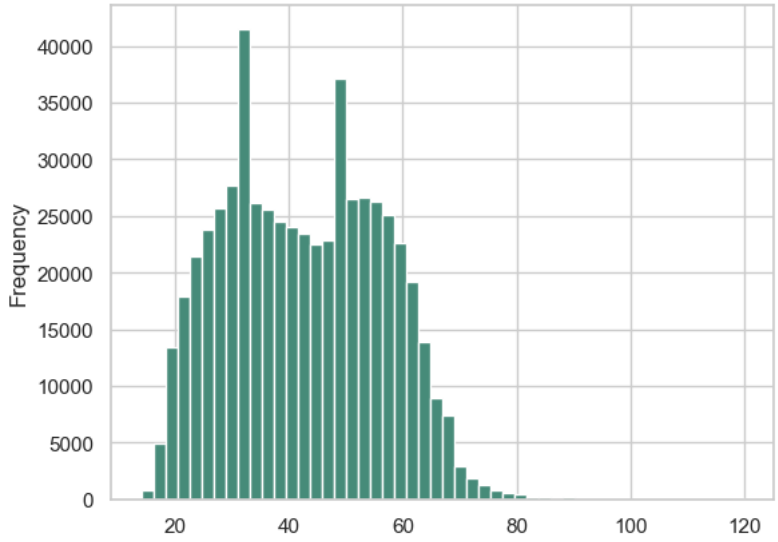

Figure 2

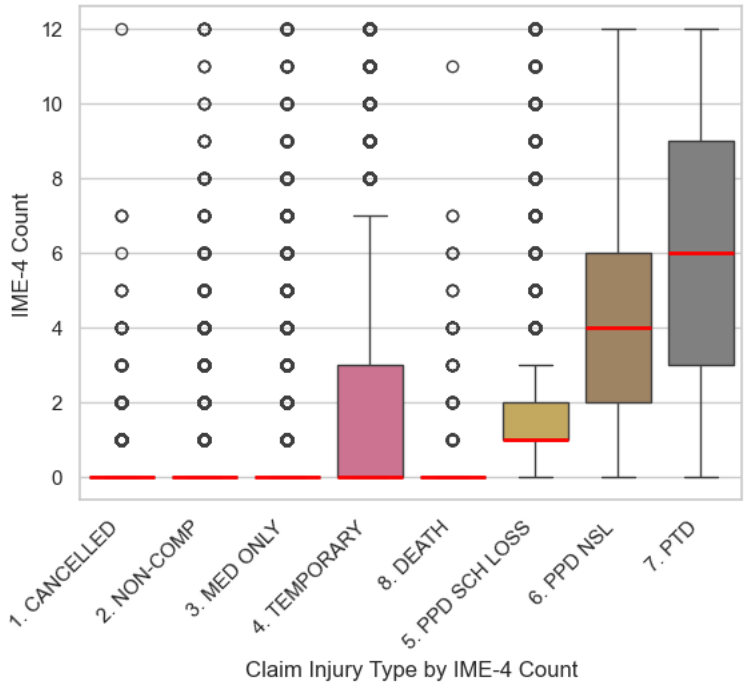

Figure 3

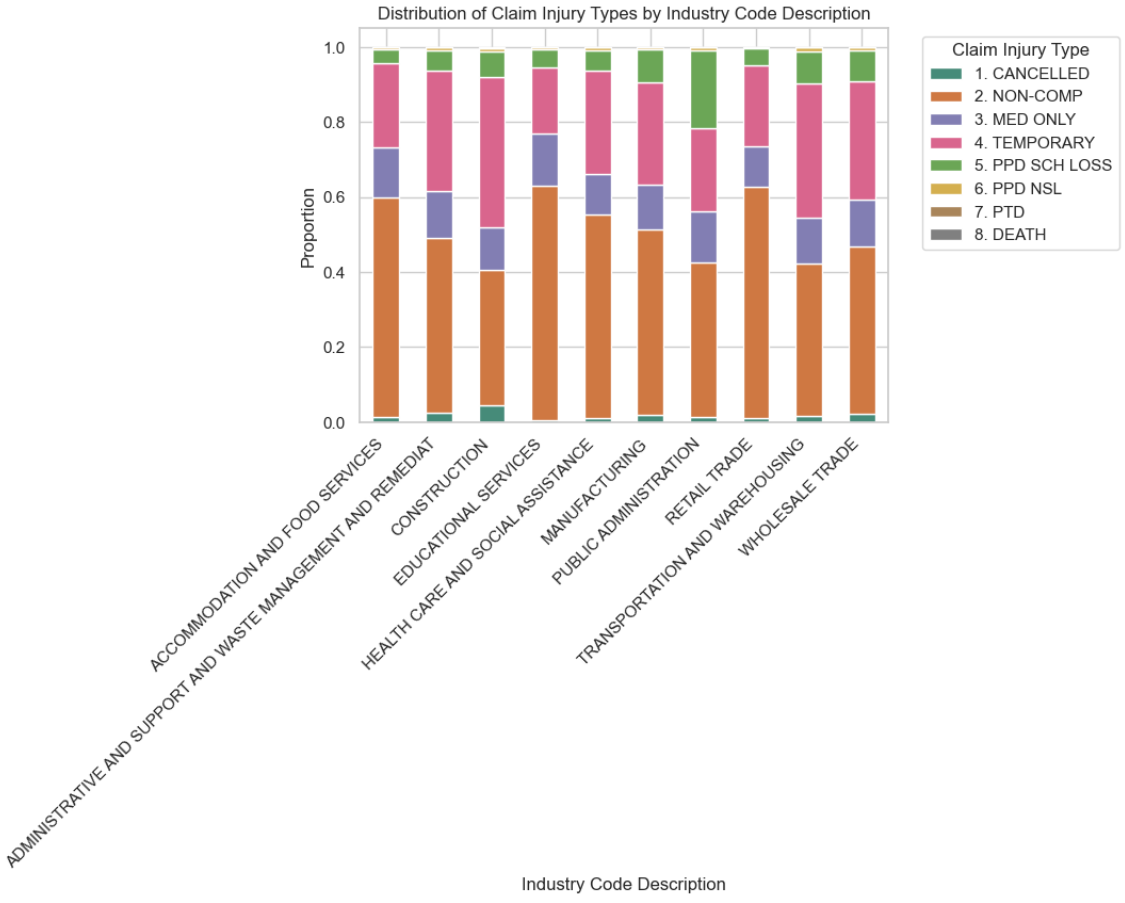

Figure 4

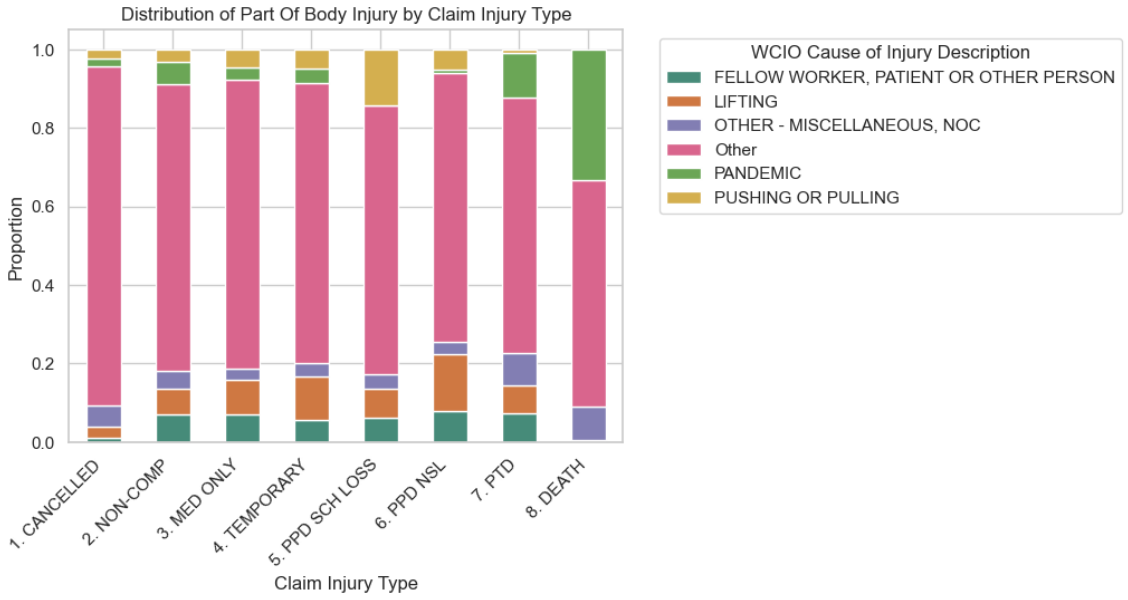

Figure 5

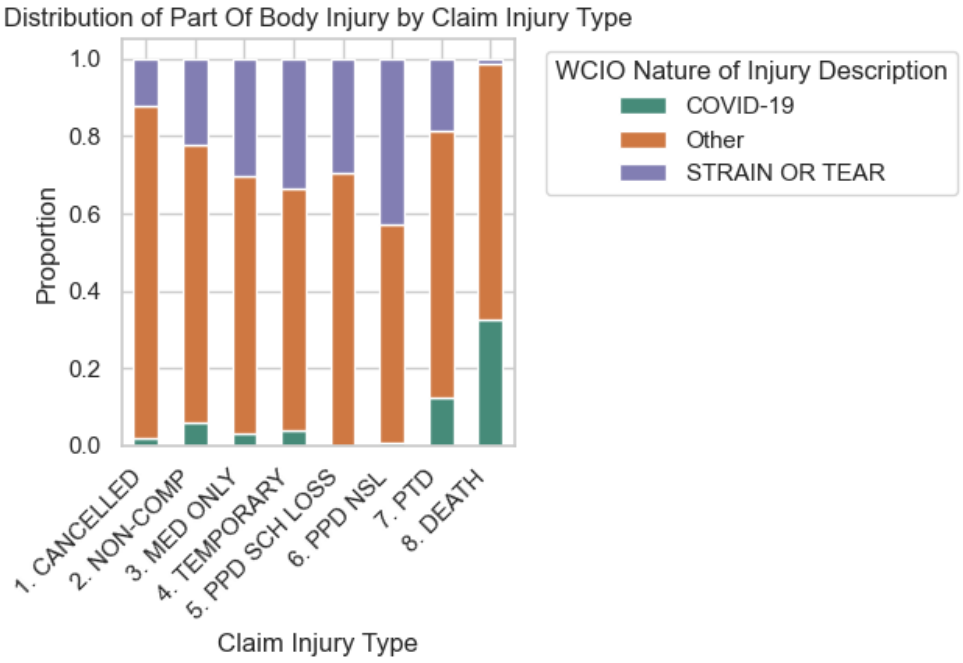

Figure 6

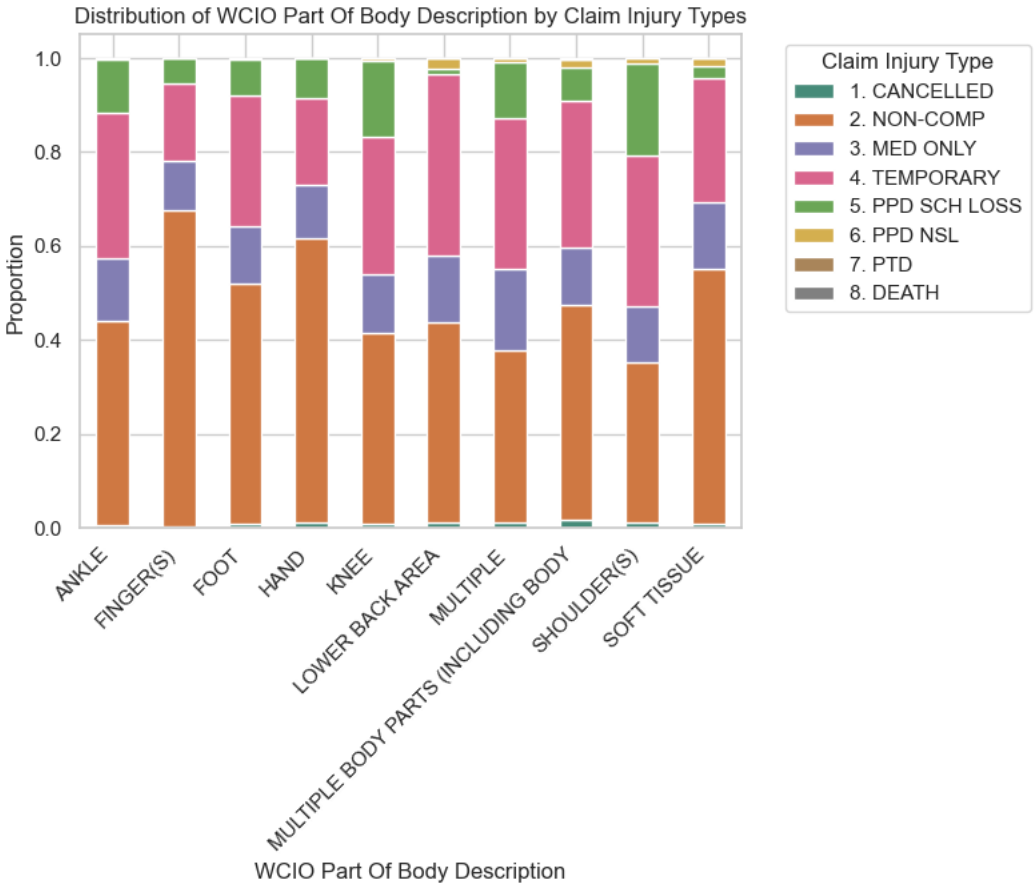
Distribution of WCIO Part Of Body Description by Claim Injury Types
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
ANKLE
FINGER(S)
FOOT
HAND
KNEE
LOWER BACK AREA
MULTIPLE
MULTIPLE BODY PARTS (INCLUDING BODY
SHOULDER(S)
SOFT TISSUE
WCIO Part Of Body Description
Claim Injury Type
1. CANCELLED
2. NON-COMP
3. MED ONLY
4. TEMPORARY
5. PPD SCH LOSS
6. PPD NSL
7. PTD
8. DEATH

Figure 7

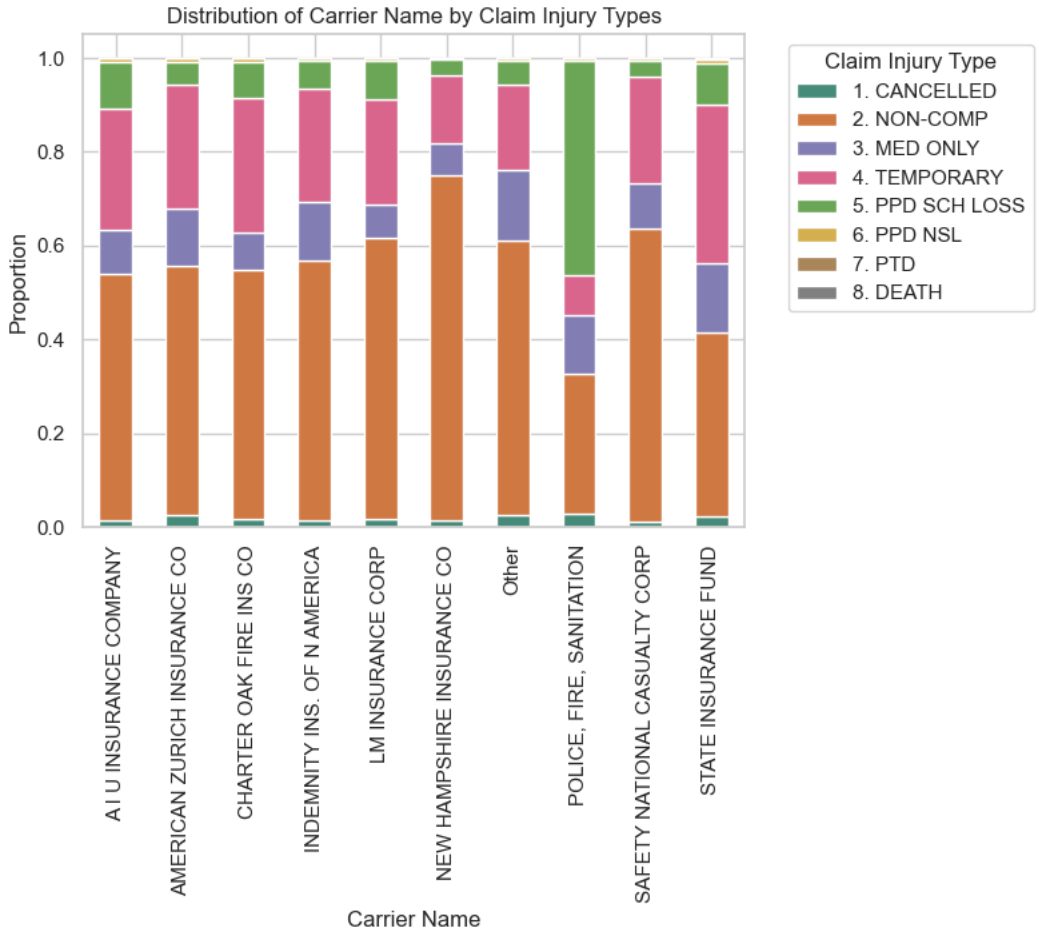

Figure 8

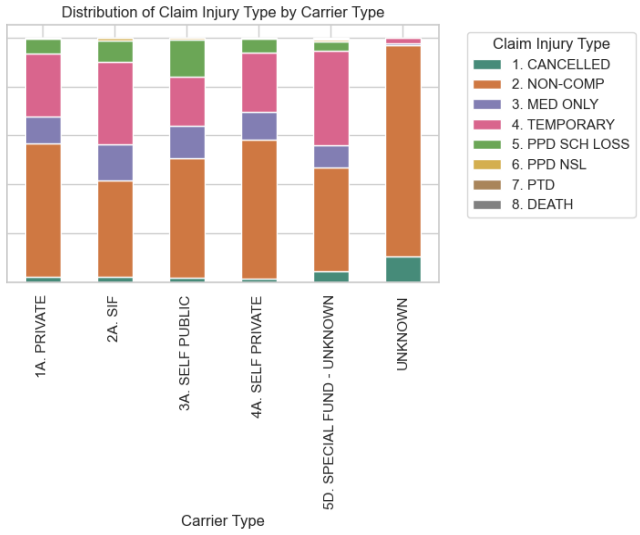

Figure 9

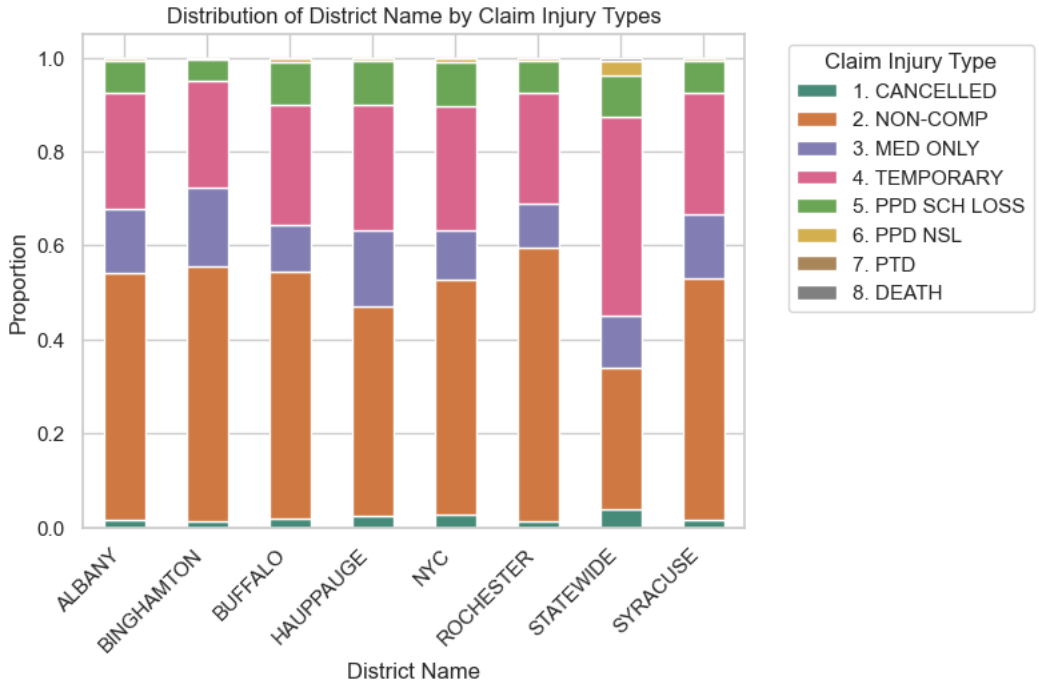

Figure 10

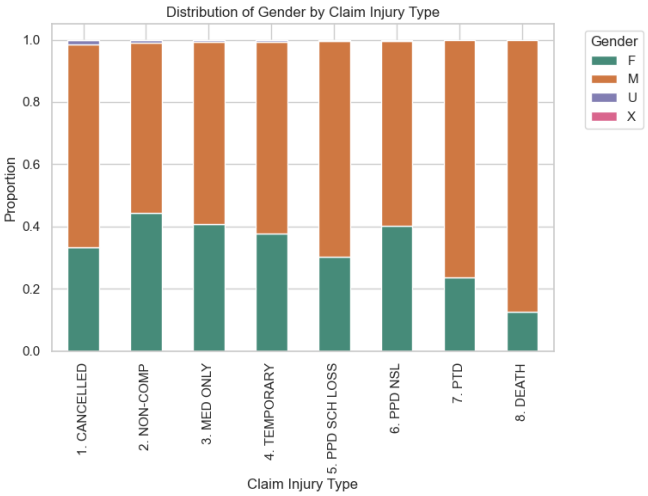

Figure 11

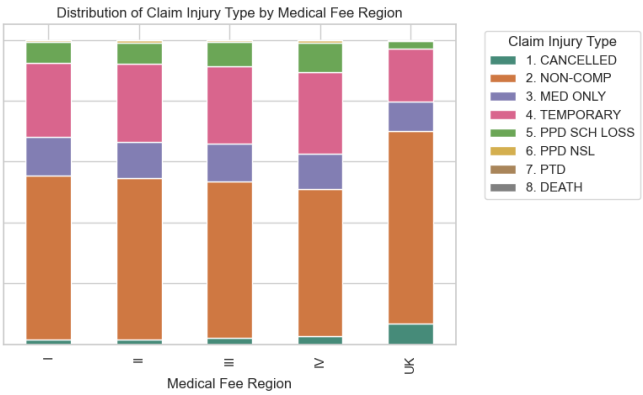

Figure 12

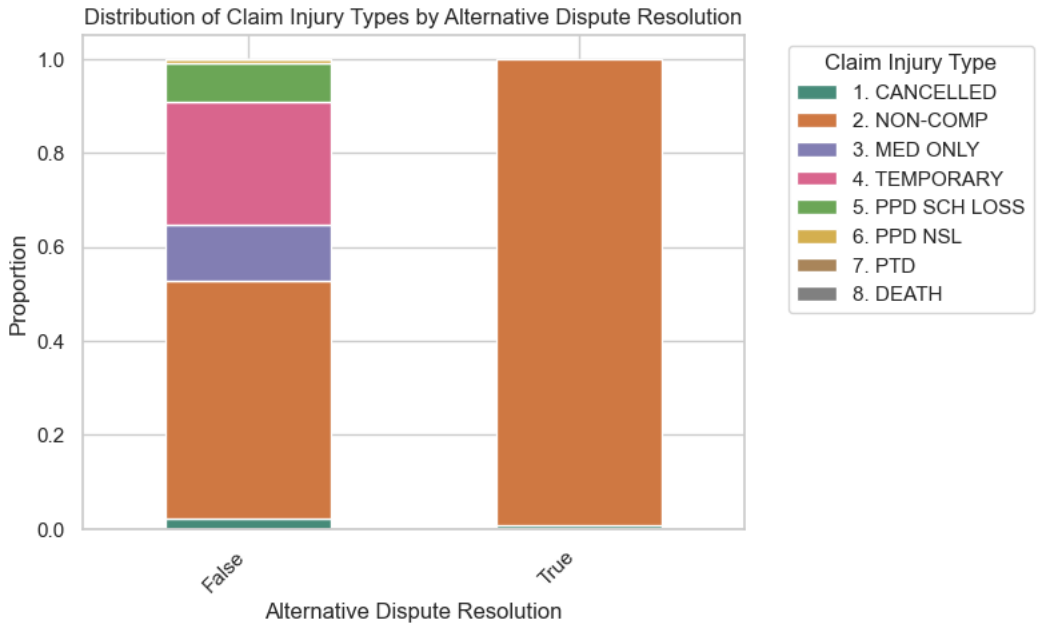

Figure 13

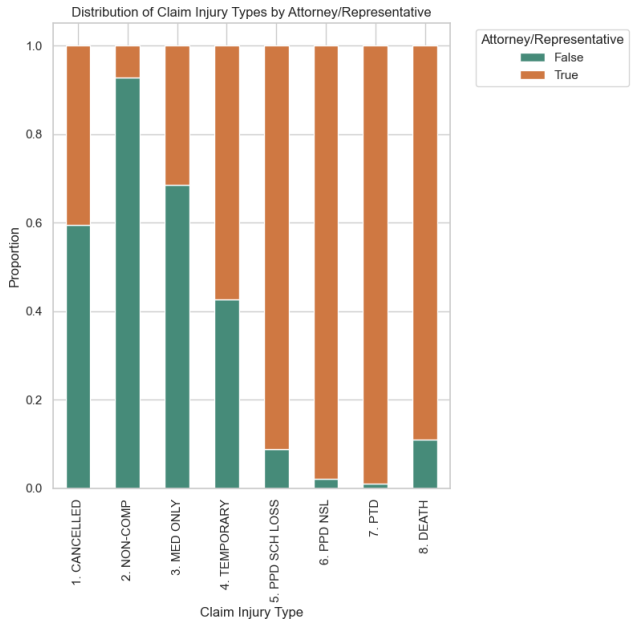

Figure 14

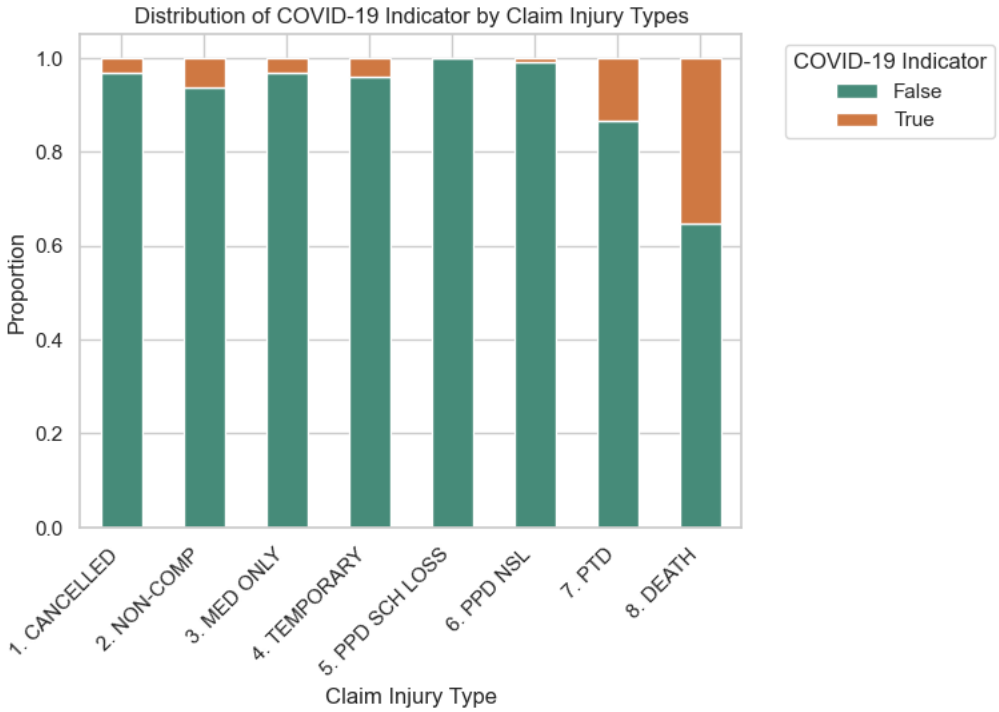

Figure 15

NUMERICAL FEATURES	MEANING
Average Weekly Wage_log*	Logarithmic transformation
Days from COVID	Days since start of COVID
Days to first Hearing	Number of days from the accident date to the first hearing
Days to Assembly	Number of days from the accident date to the assembly date
Hearing_assembly_gap_month	Difference between hearing and assembly date measured in weeks
C3-C2_gap_days	Days between C3 and C2
C2_Accident_gap_weeks	Difference between C2 and accident date measured in weeks
C3_Accident_gap_week	Difference between C3 and accident date measured in weeks
Hearing_C3_gap_months	Difference between Hearing and C3 measured in months
Hearing_C2 gap_months	Difference between Hearing and C2 measured in months

CATEGORICAL/BOOL. FEATURES	
Accident Date_year	Returns the accident year
Accident Date_missing	Checks if the accident date is missing
Accident Date_weekend	Checks if the accident was on weekend
Accident Date_month_cos*	Returns the accident month
Accident Date_month_sin*	Returns the accident month
AccidentDate_quarter_cos*	Returns the accident quarter
Accident Date_quarter_sin*	Returns the accident quarter
Accident Date_assembly_gap_days	calculates the gap in days between the Assembly Date and the Accident Date
C2_missing	Checks if C2 is missing
C3_missing	Checks if C3 is missing
Hearing Date_missing	Checks if Hearing date is missing
Work_on_distance	Checks if the injury occurred in the worker's county
Accident Date_season	[1:Winter, 2:Spring,3:Summer, 4:Fall]
Combined_injury	Cause, Nature and part of injury
Cause_nature_injury	Cause and nature of injury
Nature_part_injury	Nature and part of injury
Cause_part_injury	Cause and part of injury

* we applied sin and cos transformations in order to capture the cyclical nature of months avoiding misleading relationships. Cosine transformations ensure that december and january are treated as close neighbors and not being 1 far from 12. We applied Logarithmic transformation to average weekly wage to reduce the skewness and improve model performance by stabilizing variance.

Table 1

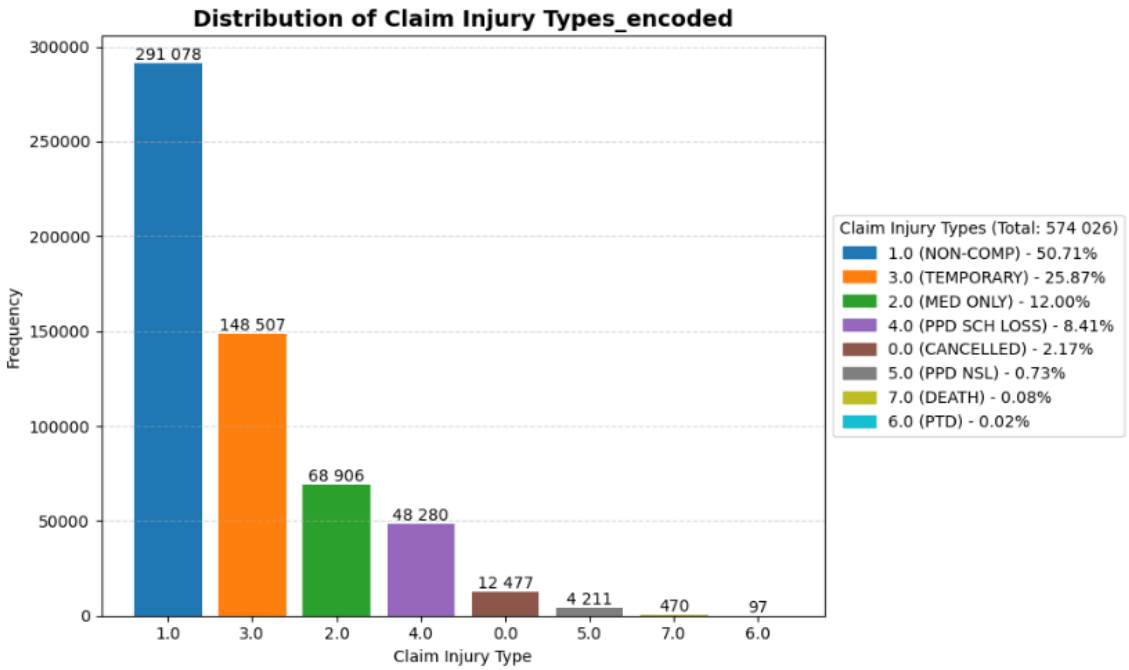

Figure 16

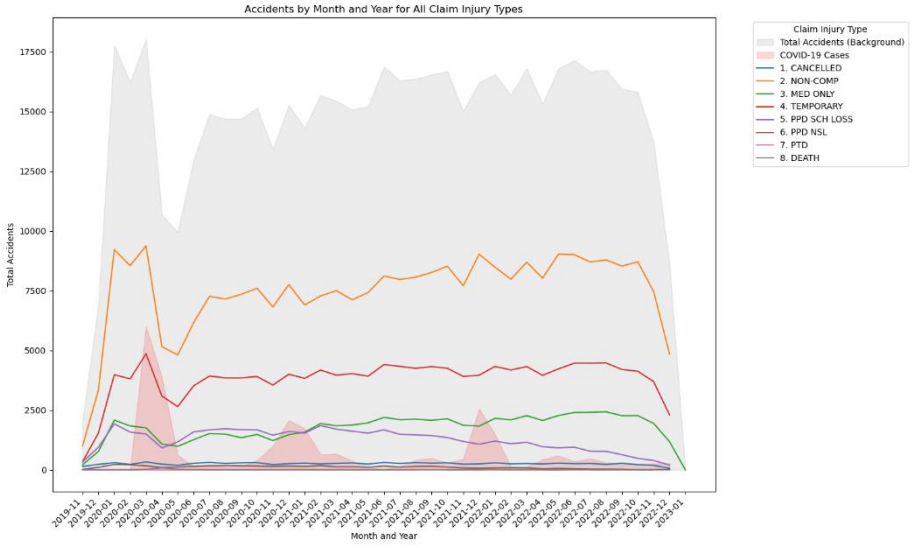

Figure 17

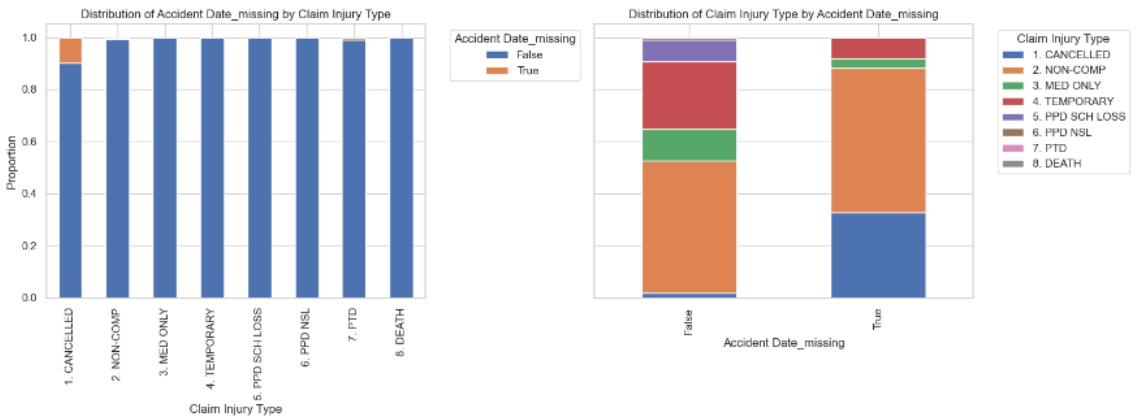

Figure 18

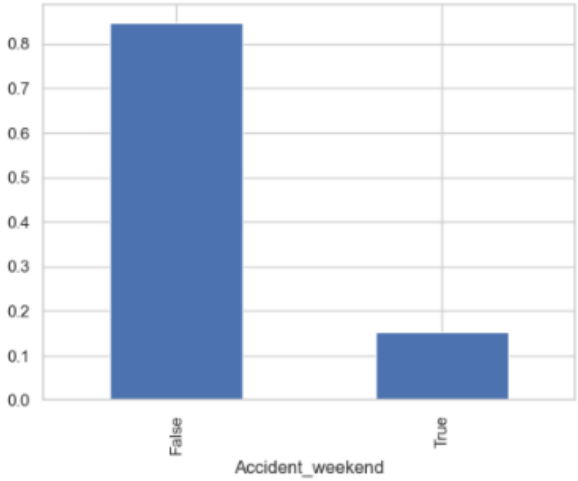

Figure 19

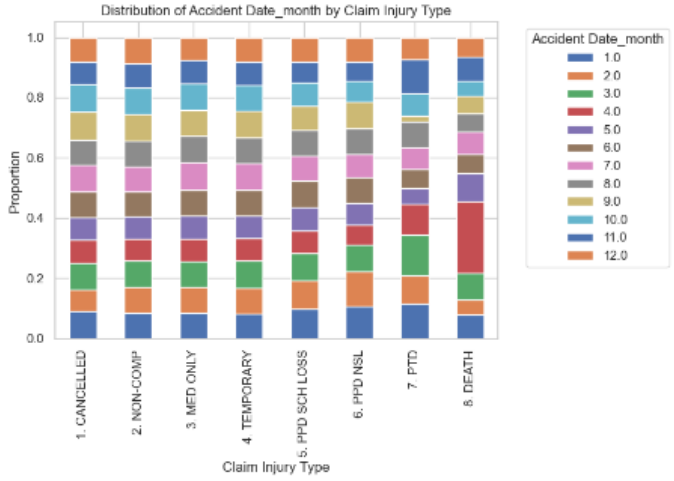

Figure 20

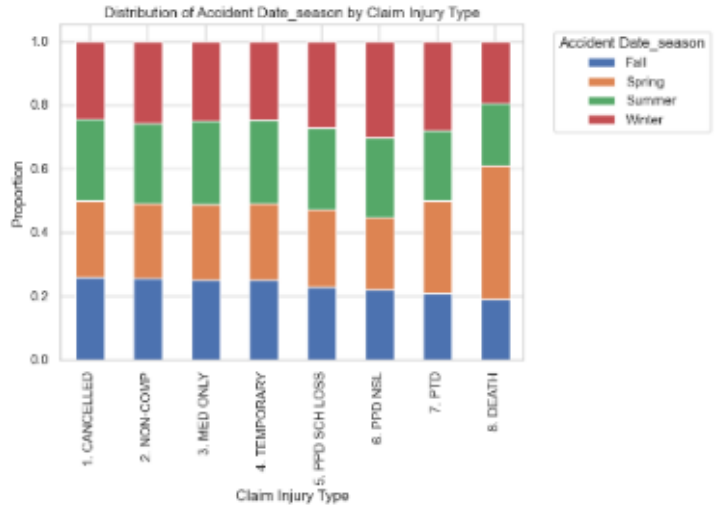

Figure 21

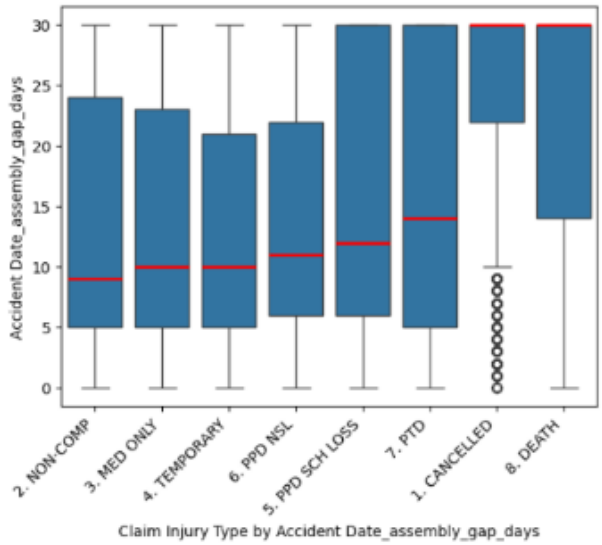

Figure 22

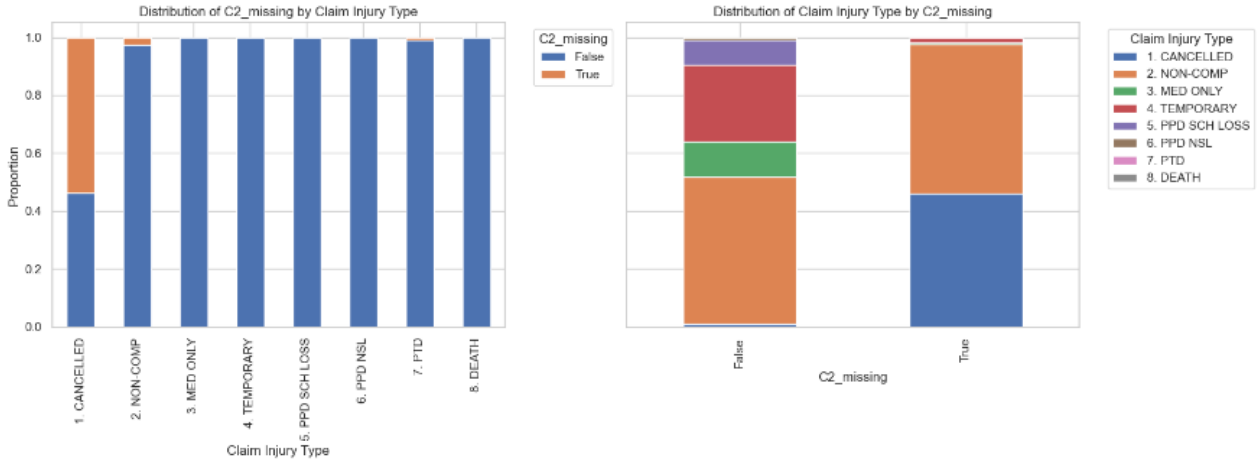

Figure 23

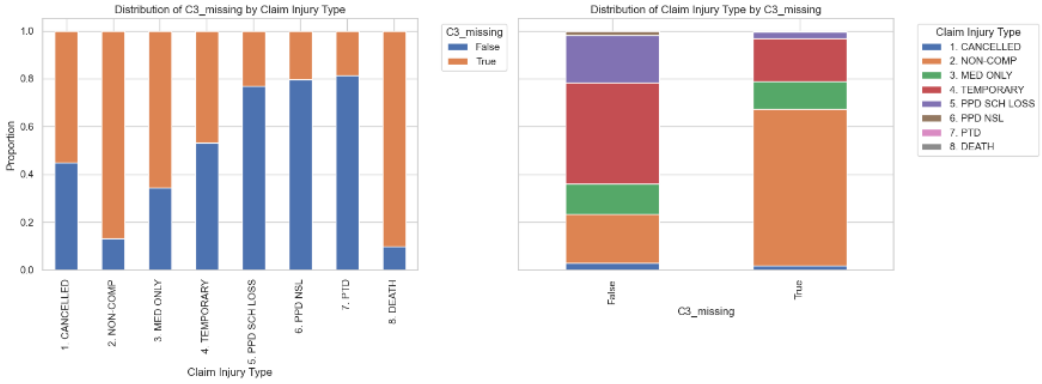

Figure 24

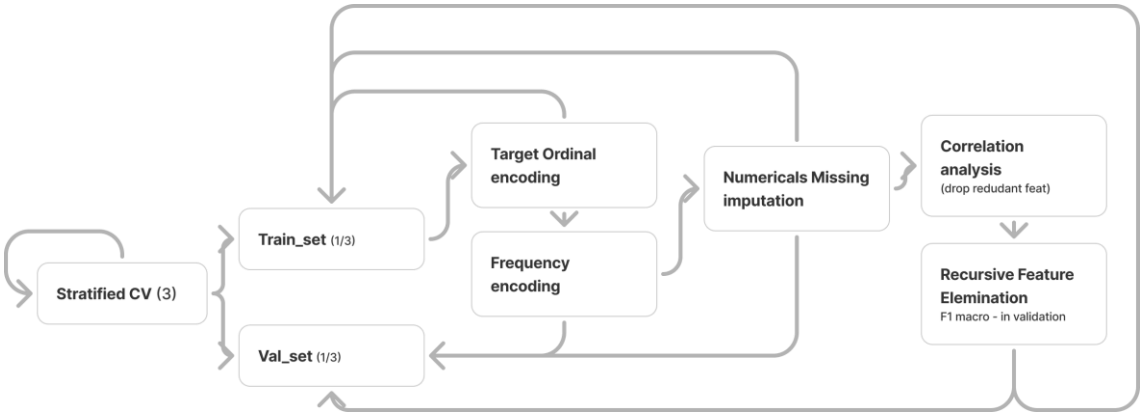

Figure 25

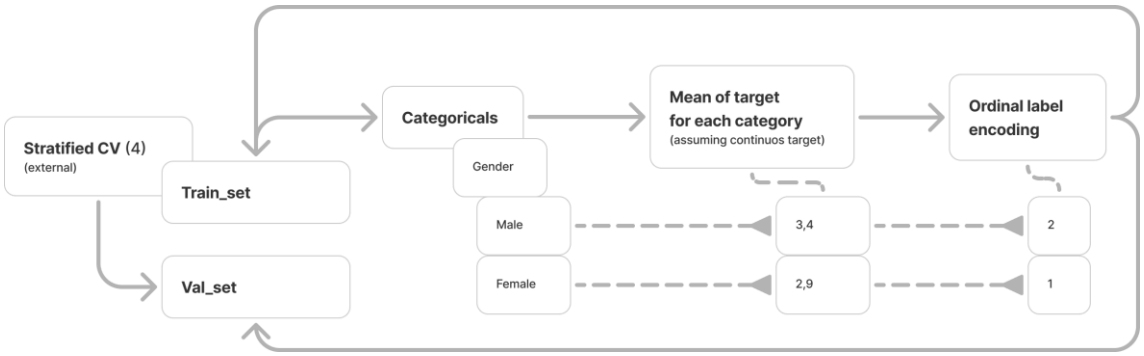

Figure 26

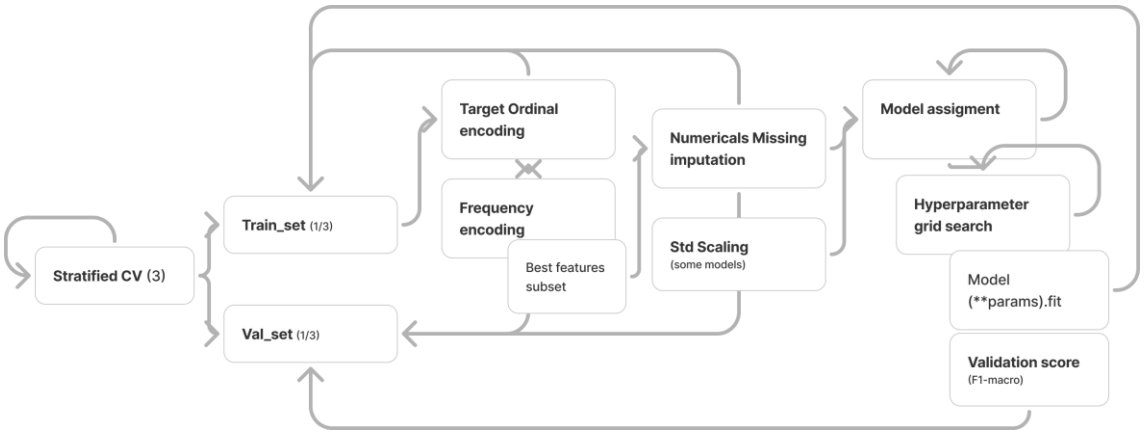

Figure 27

Categorical	Ordinal Target Encoding	Frequency Encoding
Carrier Name	Target 4 (PPD), 5 (PPD NSL), 6 (PTD) associated with carriers with higher mean for target	Target 6 (PTD) associated with slightly more frequent carrier names
Carrier Type	Target 4 (PPD), 5 (PPD NSL), 6 (PTD) associated with carrier types with higher mean for target	Target 0 (canceled) and 1 (non comp) associated with Carrier Types with higher frequency
County of Injury	For each target each county is evenly distributed	Target 0 (canceled), 4 (PPD), 5 (PPD NSL) slightly associated with county of injury with higher frequency
Distric Name	For each target each district is evenly distributed	Target 0 (canceled), 4 (PPD), 5 (PPD NSL) associated with district names with higher frequency
Gender	For each target the gender is evenly distributed, except death (7) which is more predominant in a gender with higher mean for target (probably males)	For each target, gender is evenly distributed, except for 7 (death) where it occurs more in males
Industry Code Description	Target 4 (PPD) associated with industries with higher mean for target	Target 0 (canceled) associated with Industries with lower frequency, and target 4 (PPD) and 6 (PTD) associated with industries with higher frequency
Medical Fee Region	Target 1 (non comp) and 2 (med only) are more associated medical fee regions with low mean for target	Target 1 (non comp) and 2 (med only) associated with medical region with lower frequency
WCIO Cause of Injury	Target 0 (canceled) and 7 (death) associated with specifc causes with lower mean score for target	Targets are associated with cause of injury with even proportions
WCIO Nature of Injury	Target 0 (canceled) associated with categories with lower mean for target	Target 0 (canceled), 6 (PTD) and 7 (death) associated with nature of injury more infrequent; vs target 2 (med only), 4 (PPD), 5 (PPD NSL) associated with nature of injuries of higher frequency
WCIO Part of Body	Target 0 (canceled) and 7 (death) associated with part of body categories with lower mean for the target; vs target 4 associated with part of body categories with higher mean for target	Target 7 (death) associated with part of body with lower frequency, vs target 4 (PPD) with part of body with higher frequency

Zip code	Target 6 (PTD) slighty more associated with zip code categories with higher mean for target	For each target zip codes are evenly distributed regarding their frequency (always low)

Table 2

Model	Parameters	Values
LogisticRegression	C	1, 10, 0.1
	solver	“lbfgs”, “linear”
	class_weight	None, “balanced”
	multi_class	“multinomial”, “ovr”
GaussianNB	var_smoothing	1e-9, 0,1
MLPClassifier	hidden_layer_sizes	(100,), (25,8)
	learning_rate_init	0.001, 0.01
RandomForestClassifier	max_depth	6
	class_weight	None, “balanced”
CatBoostClassifier	iterations	1000, 300, 500
	depth	6, 7,8,9
	boosting_type	“Ordered”, “Plain”
	auto_class_weights	None, “SqrtBalanced”, “Balanced”
	loss_function	“MultiClass”, “MultiClassOneVsAll”

Table 3

Models	**Parameters**	**Train Score**	**Validation**
Logistic Regression	{‘C’:1,’solver’:lbfgs’,’class_weight’:None)	0.426	0.419
Gaussian NB	{‘var_smoothing’:0.1}	0.351	0.350
OVR_Random Forest	{‘max_depth’:6,’class_weight’:’balanced’}	0.369	0.350
Random Forest	{‘max_depth’:6,’class_weight’:’balanced’}	0.332	0.313
Catboost	{‘iterarions’:1000,’depth’:6,’boosting_type’:’Ordered’,’auto_class_weights’:’SqrtBalanced’, loss_function:”MultiClassOneVsAll”}	0.560	0.485
Neural Network	{‘hidden_layer_sizes’: (25,8), ‘learning_rate_init’:0.01}	0.451	0.446
Stacking (CatBoost, MLP)	{'C': 1, 'multi_class': 'ovr', 'class_weight': 'balanced'}	0.456	0.432
Stacking (Catboost, NB)	{'C': 1, 'multi_class': 'ovr', 'class_weight': 'balanced'}	0.455	0.434

Table 4

	1	2	3	4	5	6	7	8
LogisticRegression	0.52	0.90	0.10	0.74	0.46	0.11	0.02	0.16
GaussianNB	0.47	0.88	0.26	0.43	0.44	0.08	0.04	0.26
OVR_Random Forest	0.01	0.90	0.11	0.77	0.48	0.00	0.00	0.23
RandomForest	0.12	0.90	0.15	0.78	0.47	0.00	0.00	0.16
CatBoost	0.61	0.91	0.24	0.79	0.65	0.15	0.04	0.44
Neural Network	0.55	0.90	0.28	0.76	0.60	0.06	0.00	0.41
StackingClassifier (CatBoost, MLP)	0.59	0.91	0.22	0.79	0.66	0.01	0.03	0.25
StackingClassifier (CatBoost, NB)	0.60	0.91	0.20	0.79	0.65	0.01	0.03	0.27

Table 5

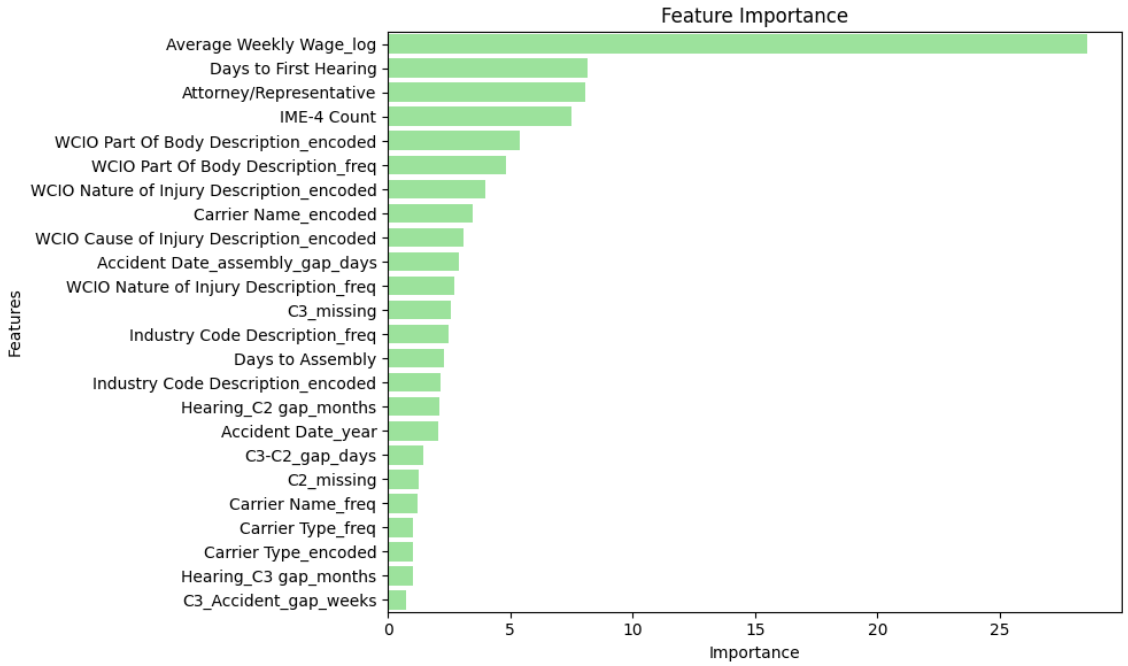

Figure 28

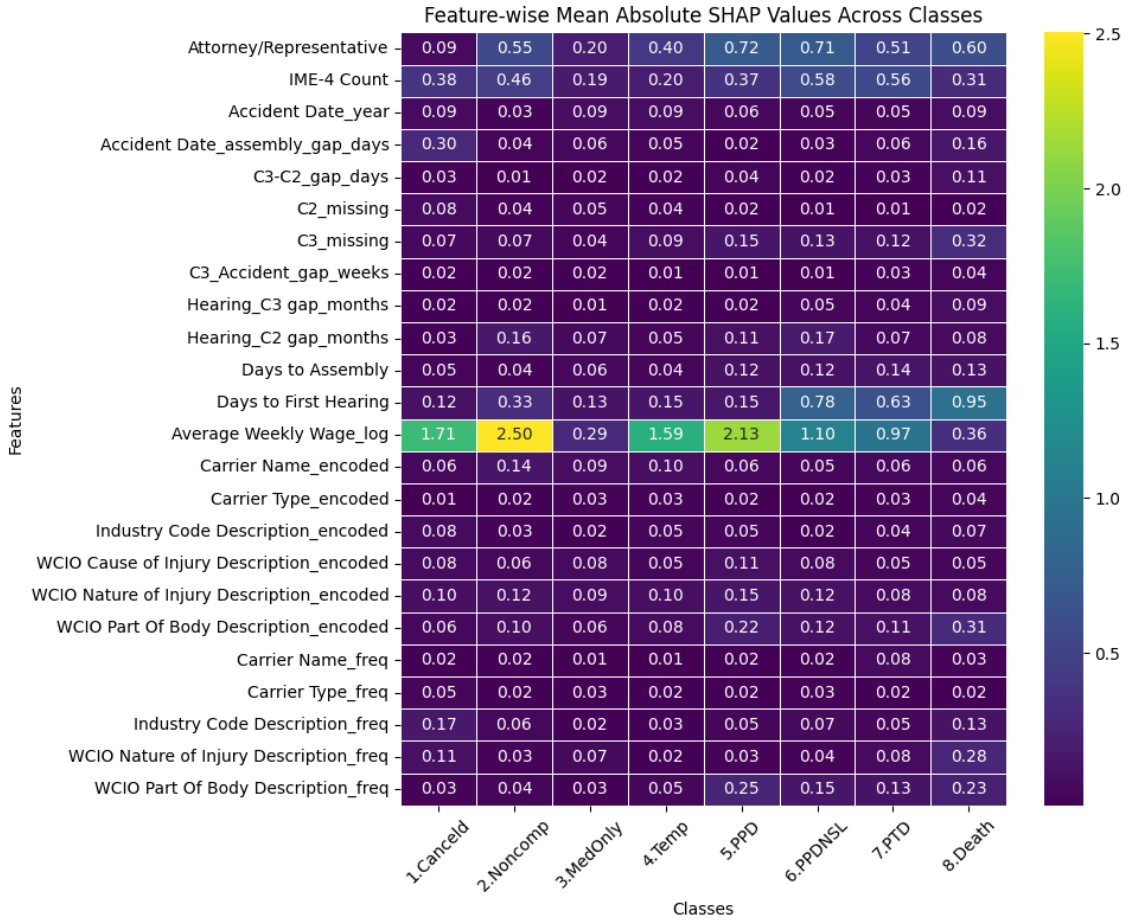
Feature-wise Mean Absolute SHAP Values Across Classes
Features
Attorney/Representative 0.09 0.55 0.20 0.40 0.72 0.71 0.51 0.60
IME-4 Count 0.38 0.46 0.19 0.20 0.37 0.58 0.56 0.31
Accident Date_year 0.09 0.03 0.09 0.09 0.06 0.05 0.05 0.09
Accident Date_assembly_gap_days 0.30 0.04 0.06 0.05 0.02 0.03 0.06 0.16
C3-C2_gap_days 0.03 0.01 0.02 0.02 0.04 0.02 0.03 0.11
C2_missing 0.08 0.04 0.05 0.04 0.02 0.01 0.01 0.02
C3_missing 0.07 0.07 0.04 0.09 0.15 0.13 0.12 0.32
C3_Accident_gap_weeks 0.02 0.02 0.02 0.01 0.01 0.01 0.03 0.04
Hearing_C3 gap_months 0.02 0.02 0.01 0.02 0.02 0.05 0.04 0.09
Hearing_C2 gap_months 0.03 0.16 0.07 0.05 0.11 0.17 0.07 0.08
Days to Assembly 0.05 0.04 0.06 0.04 0.12 0.12 0.14 0.13
Days to First Hearing 0.12 0.33 0.13 0.15 0.15 0.78 0.63 0.95
Average Weekly Wage_log 1.71 2.50 0.29 1.59 2.13 1.10 0.97 0.36
Carrier Name_encoded 0.06 0.14 0.09 0.10 0.06 0.05 0.06 0.06
Carrier Type_encoded 0.01 0.02 0.03 0.03 0.02 0.02 0.03 0.04
Industry Code Description_encoded 0.08 0.03 0.02 0.05 0.05 0.02 0.04 0.07
WCIO Cause of Injury Description_encoded 0.08 0.06 0.08 0.05 0.11 0.08 0.05 0.05
WCIO Nature of Injury Description_encoded 0.10 0.12 0.09 0.10 0.15 0.12 0.08 0.08
WCIO Part Of Body Description_encoded 0.06 0.10 0.06 0.08 0.22 0.12 0.11 0.31
Carrier Name_freq 0.02 0.02 0.01 0.01 0.02 0.02 0.08 0.03
Carrier Type_freq 0.05 0.02 0.03 0.02 0.02 0.03 0.02 0.02
Industry Code Description_freq 0.17 0.06 0.02 0.03 0.05 0.07 0.05 0.13
WCIO Nature of Injury Description_freq 0.11 0.03 0.07 0.02 0.03 0.04 0.08 0.28
WCIO Part Of Body Description_freq 0.03 0.04 0.03 0.05 0.25 0.15 0.13 0.23
1.Canceld
2.Noncomp
3.MedOnly
4.Temp
5.PPD
6.PPDNSL
7.PTD
8.Death
Classes
2.5
2.0
1.5
1.0
0.5

Figure 29

	Corr			RFE		
	CV1	CV2	CV3	CV1	CV2	CV3
Age at Injury						
Alternative Dispute Resolution						
Attorney/Representative						
Carrier Name_encoded						
Carrier Type_encoded						
County of Injury_encoded						
COVID-19 Indicator						
District Name_encoded						
Gender_encoded						
IME-4 Count						
Industry Code Description_encoded						
Medical Fee Region_encoded						
WCIO Cause of Injury Description_encoded						
WCIO Nature of Injury Description_encoded						
WCIO Part Of Body Description_encoded						
Zip Code_encoded						
Number of Dependents',						
Accident Date_year						
Accident Date_missing						
Accident_weekend						
Accident Date_month_cos						
Accident Date_month_sin						
Accident Date_quarter_cos						
Accident Date_quarter_sin						
Accident Date_assembly_gap_days						
C3-C2_gap_days						
C2_missing						
C3_missing						
C2_Accident_gap_weeks						
C3_Accident_gap_weeks						
Hearing Date_missing						
Hearing_C3 gap_months						
Hearing_C2 gap_months						
Hearing_assembly_gap_months						
Days to Assembly						
Days to First Hearing						
Days from COVID						
Age_not_correct						
Average Weekly Wage_log						
Work_on_distance						
Carrier Name_freq						
Carrier Type_freq						
County of Injury_freq						
District Name_freq						
Gender_freq						
Industry Code Description_freq						
Medical Fee Region_freq						
WCIO Cause of Injury Description_freq						
WCIO Nature of Injury Description_freq						
WCIO Part Of Body Description_freq						
Zip Code_freq						
Number of Features	46	46	46	26	25	25

Table 6