Machine Learning

Master's in data science and advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

TO GRANT OR NOT TO GRANT:

DECIDING ON COMPENSATION BENEFITS

PROJECT REPORT

GROUP 50

Eldar Medvedev, r20181162

Guilherme Godinho, 20211552

Jéssica Cristas, 20240488

Joshua Wehr, 20240501

Umeima Mahomed, 20240543

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

ABSTRACT

Machine Learning is transforming industries by automating repetitive tasks and enabling data-driven decisions. This project aimed to streamline claims processing for the New York Workers' Compensation Board (WCB) by developing a multiclass classification model to predict Claim Injury Types, expediting decision-making and enhancing operational efficiency.

Using CRISP-DM methodology, we explored a dataset of over 590,000 claims, addressing challenges like missing values, outliers, and class imbalance. Feature engineering captured critical patterns, while advanced techniques like SMOTE and RUS dealt with class imbalance. We build an ensemble model using relative importance of the following models: XGBoost, LightGBM, Logistic Regression, Random Forest, and MLP. Hyperparameter tuning, resampling strategies and ensemble weighting were performed using Optuna, ensuring robust model performance.

The machine learning pipeline was implemented using Stratified K-Fold cross-validation for optimal performance. All steps were organized into Python functions and combined into a single function, which was executed three times to represent the three folds of cross-validation.

The model achieved an average F1 macro score of 0.497 with resampled data between 3 folds, a notable improvement in minority class predictions. The approach also introduced innovative features like Agreement Reached prediction and a user-friendly web app for real-time claim predictions, showcasing practical applications for WCB employees. This work demonstrates the potential of machine learning to revolutionize administrative processes, saving time, and improving service quality. Future work will focus on enhancing data quality, incorporating more advanced models, and refining feature engineering to further elevate performance and scalability.

KEYWORDS

Machine Learning; Classification; Multiclass; Ensemble Modeling; Class Imbalance

GROUP MEMBER CONTRIBUTION

Our group collaboration worked in the following way: while some members concentrated on examining variables and offering clear explanations, others focused on data preprocessing and model development. Ultimately, we collaborated to finalize the report, merging our separate inputs to produce a unified and thorough result. This method enabled us to utilize our strengths and guarantee that every facet of the project was completely covered.

INTRODUCTION

Machine learning is changing how businesses work by making it easier to make data-driven decisions and automate repetitive tasks. As these technologies become more available, companies are using machine learning to work more efficiently, simplify processes, and free up time for more creative work.

The New York Workers' Compensation Board (WCB) handles workers' compensation, disability claims, and cases involving volunteer firefighters and emergency medical personnel. Currently, their claims review process is manual, which takes up a lot of time and effort. This leads to delays, strains resources, and makes it hard to keep up with the growing demand.

To address these challenges, WCB partnered with NOVA IMS to explore the potential of automation. Together, we built a machine learning model that predicts whether a claim should be approved or denied.

This report outlines how we built the model, the challenges we faced, and the results. By automating claim processing, WCB can save time, cut costs, and operate more efficiently, providing better support for New York's workforce.

1.1. Objective and Approach

The main goal of this project is to develop a multiclass model that classifies new claims based on injury type, streamlining the WCB's decision-making process. The specific focus is to predict the target variable **Claim Injury Type**.

When comparing our objective to references, we can see that it is a problem that has been approached in diverse ways before. Researchers analysed 1.2 million Ohio workers' compensation claims using machine learning to identify high-risk industries for ergonomic and slip/trip/fall injuries. They found Skilled Nursing Facilities had the highest ergonomic injury risk while General Freight Trucking led in slip/trip/fall incidents. This created a new injury surveillance system that can help target prevention efforts and be replicated by other states (Meyers et al. (2018)).

Compared to the reference we do not expect to prevent injuries but to help the affected to get their claims solved faster so their mental and physical state increases faster after their injury and to provide more automation to this process for WCB.

Our main goal is derived from getting the most accurate **Claim Injury Type** prediction. To measure how good the prediction is, we used the F1 macro score. When uploading our model to the designated Kaggle competition, we got an F1 macro rating as feedback. Therefore, our decisions throughout the project were focused on maximizing cross-validation F1 macro score.

1.2. Methodology

We followed the **CRISP-DM** methodology to structure the report and Jupyter Notebooks for the code, implementing each step in Python 3.10.15.

In the Exploratory Data Analysis (EDA) we used **Missingno** to visualize missing data, **Tabulate** and custom-made functions to summarize statistics. Histograms, boxplots from **MatPlotLib**, and a **Seaborn** correlation heatmap highlighted distributions, outliers, and feature relationship.

In Preprocessing, **KNNImputer** was used to handle missing values, **MinMaxScaler** for scaling features, and custom Python functions to treat outliers. Categorical features were encoded with **Frequency Encoding**, **get_dummies**, and custom-made target encoding. Feature selection was done using **Mutual Information** (MI), **Recursive Feature Elimination** (RFE), and **Least Absolute Shrinkage and Selection Operator** (Lasso regression), improving model performance by identifying important features and removing less relevant ones.

For Modelling, we used ensemble with following models: **XGBoost**, **LightGBM**, **Logistic Regression**, **Random Forest** and **MLP**, chosen for their classification capabilities and diversity while parameter tuning was done with **Optuna** to optimize hyperparameters, resampling strategies, and ensemble model weights. To address class imbalance, **SMOTE** and **RUS** were applied. Model performance was evaluated using the **classification report** (F1 macro, F1 weighted, accuracy, and recall). **Stratified K-Fold cross-validation** (SKF) was used through all the steps after data preprocessing to prevent data leakage, ensuring balanced target classes across folds.

Note: All the algorithms that were not covered in class are presented in **Table B25.** The project repository is publicly available on **GitHub** and contains all inherent documents.

DATA EXPLORATION

2.1. Data Overview

The received data consists of two data sets: the training set, with claims from 2020-2022 (with complete information including decisions) - with 593,471 rows and 33 columns - and the test set with Claims from January 2023 onward (without post-decision information), with 387 975 rows and 30 columns. The datasets include various types of information, such as personal information (age, gender, birth year, zip code), incident details (accident date, county of injury, industry code), medical information (nature of injury, part of body affected), administrative details (attorney representation, carrier information) and information about the claim process (hearing dates, dispute resolution status). Details about the features can be found in **Table B1**.

While looking at **Figure B1,** we can find that the correlation between numerical variables is not high, but we see correlation of Average Weekly Wage and our target **Claim Injury Type**.

2.2. Data Exploration and Analysis

This chapter sums up the most important insights from the in-depth Data Exploration that was done in the project.

The dataset spans from January 2020 to December 2022, with notable impacts from COVID-19, including accident peaks in early 2020 and declines during lockdowns and holidays. The most common injuries are strains, tears, and lifting-related incidents, with the lower back and knees most frequently affected. Each time that the Accident Date is missing, the Age at Injury is equal to 0. Men represent 58.88% of cases, and injuries align with NYC workforce age profiles, predominantly affecting workers born between 1950–1990. Average Weekly Wage data shows high variability, with 58% earning $0 due to volunteer roles; the typical range for others is $750–$1200.

Claims are predominantly covered by the State Insurance Fund or private providers; public carriers are underrepresented due to fewer providers. COVID-19 claims account for 4.79% of cases, with most claims arising early in the pandemic. Severe or complex cases requiring attorneys or alternative dispute resolution are rare but notable indicators of case complexity. Medical Fee Regions I and IV dominate, reflecting population density and healthcare costs in urban and rural areas. The dataset includes inconsistencies and missing data in variables like zip codes, injury descriptions, and claim dates, highlighting data quality challenges. Claim resolution trends show efficiency, with Assembly and Accident Dates closely aligned, while agreements are reached in only 4% of cases.

DATA PREPROCESSING

The Flowchart below visualizes our preprocessing strategies and what took place in our overall work pipeline. All the preprocessing was done individually for every fold using **Stratified K-Fold** to prevent data leakage.

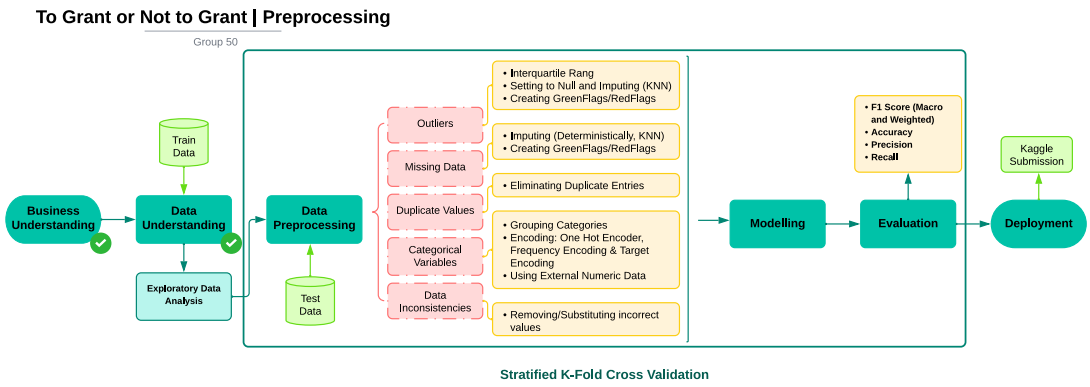

Figure 1 - Flowchart for Data Preprocessing

3.1. Incoherencies within and between datasets

The Test dataset has three extra columns compared to the training dataset: **Claim Injury Type**, Agreement Reached, and WCB Decision. Aside from these differences, most of the data types match between the two datasets, however, there are a couple of inconsistencies (for example, Age at Injury and Number of Dependents is of type *int6*4 in the test data and as *float64* in the training data). These discrepancies were addressed to keep the data types consistent.
As mentioned in **Table B1**, the WCB Decision only has one value, due to that, we decided to drop it. The prediction for the Agreement Reached was made for the test dataset in the open-ended section. OIICS Nature of Injury Description was also removed due to it being fully null.

In terms of duplicates, we had 2 observations with the same ID, but this observation falls into another problem in the training dataset. We have 19445 observations that have a unique trait: the **Claim Identifier** values in these rows were nine characters long, which differed from the standard length of eight characters observed in the rest of the dataset. This anomaly indicates data entry error which results in observation being fully null. These observations were removed.

3.2. Data Transformation

Throughout the project we converted all variables into numeric type so all the models can function smoothly with it. Date variables were transformed from *object* to *datetime* format. In COVID-19 Indicator and Attorney/Representative, values were mapped (Y -> 1, N -> 0). For Alternative Dispute Resolution and Gender, "Unknown" values and the rare "X" in Gender were replaced with the ZeroR value and then transformed into Boolean variables.

3.3. Handling Missing Values

Dealing with missing data was another crucial step in ensuring the dataset was both clean and reliable for analysis. The following strategies were applied:

1. **Red Flag (“Missing”) Column**: A Boolean column was created to flag missing values for every column that had it before any preprocessing.
2. **C-2 and Accident Date**: Missing values were replaced with the Assembly Date, the most reliable variable, in this case.
3. **IME-4 Count**: Missing values (76.9% missing) were replaced with 0, due to its high percentage we assume the absence of forms.
4. **Industry Code Description, WCIO Nature of Injury Description, WCIO Cause of Injury Description, WCIO Part of the Body Description**: Missing values were replaced with "Unknown".
5. **Age at Injury, Average Weekly Wage,** and **Industry Average Salary** (new variable): Missing values were imputed using KNN Imputer (Nulls in Age at Injury appeared after outlier treatment).

KNN imputation was performed on a 15% sample of the data due to its high computational cost. We used KNN for all variables simultaneously to efficiently capture feature relationships.

Features for imputation were selected based on the highest Pearson correlation with the target, excluding columns with null values, and limiting the selection to two Boolean columns per variable. KNN was fitted on scaled (MinMaxScaler) training data and applied to both training and test data, after which the data was unscaled.

3.4. Outlier Detection and Treatment

Outlier detection was another necessary step to ensure our model would not be biased by extreme values. To identify outliers, we used boxplots for visual reference and 1.5 IQR for outliers to identify thresholds. Each time we capped outliers based on specific rules or conditions (such as anomalies), we created corresponding red flag columns. These columns served as markers to preserve information about the original outlier status, ensuring that no valuable insights were lost during the capping process.

In the Age at Injury box plots, outliers were observed above the upper limit (88.5) but none below the lower limit. To address this, we manually defined the lower limit based on the legal working age. For these, we chose to replace them with '*Nan'* and impute missing value using K-Nearest Neighbors. This approach resolved errors associated with cases where the recorded age was 0. [Figure B11]

The IME-4 Count feature had several outliers, including one large value above 70. Instead of removing these outliers, we decided to keep them, as they represented rare but important cases that could provide valuable insights into complex claims.

To address the outliers in Average Weekly Wage, we applied a capping technique. Specifically, values exceeding the upper limit (2854.52) were replaced with the upper limit itself.

3.5. Categorical Variables

To ensure categorical data was ready for machine learning models we tested many different approaches and now we present our best solution.

The variables Carrier Name, District Name, and County of Injury were encoded with their frequency. These variables have many unique values and do not provide much information on their own. An important thing to mention is that the district 'Statewide' is the only one with a different target distribution, although it is underrepresented in our data set. Given this, frequency encoding is an appropriate approach for these features. **[Figure B13]**

The variables Carrier Type and Medical Fee Region we also encoded (one-hot) using Dummies (get_dummies). These variables have 8 and 5 unique values accordingly, suitable number for the encoding. After this, we joined different values of 'Special Fund' in Carrier Type in one single column and dropped one column from Medical Fee Region. The only meaningful information that we can see from the new variables is that the percentage of temporary claims is higher for Carrier 2A SIF. **[Figure B16]**

The Industry and WCIO descriptions were encoded (using the target[1]). After testing various techniques for these variables, this solution gave us better results. This is useful in our case because these features have many unique values, and this performed better than frequency encoder.

3.6. Specific Variable Incoherences Handling

Some variables needed special handling to ensure they were correctly formatted. The columns Birth Year and Accident Date were used to impute 0 values (when possible) for Age at Injury. Birth Year was removed afterward to avoid redundancy. We also cleaned up the Zip Code feature by replacing missing and invalid values with "Unknown". This standardization ensured all zip codes were consistent across the datasets.

MODELLING – MULTICLASS CLASSIFICATION

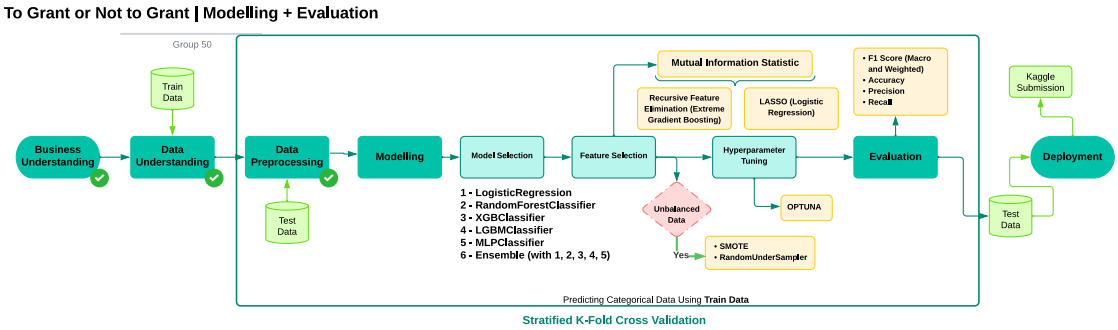

Figure 2 – Flowchart for Modelling, Evaluation and Deployment

We utilized a stratified k-fold strategy (3 folds) for splitting the data, as we believe it was the most effective method to ensure balanced representation across folds and to avoid data leakage. All preprocessing, feature selection, resampling, modeling, hyperparameter tuning, and model evaluation were implemented within a custom-made function called **Exodia**.

Throughout the entire process, we paid attention to mitigate any potential sources of data leakage. To ensure the validity of our results, all algorithms were trained only on the training data and subsequently applied to the validation data. Limits for outliers and frequency thresholds were determined based on the training set and then consistently applied to the validation set. Data scaling was performed only where necessary and was carefully managed to avoid any loss of valuable information or degradation in performance. To maintain interpretability and fidelity, the data was reverted to its unscaled form after scaling operations where appropriate. This systematic approach helped us maintain rigor and consistency across all steps of the pipeline, ensuring reliable and unbiased results.

[1] This function was created to encode categorical features using normalized frequency information, ensuring that no missing data disrupts the process.

4.1. Additional preprocessing

Here is a brief explanation of all the feature engineering processes that took place.

Sequence columns are created to validate every permutation of the Accident, Assembly and C-2 Dates variables in the sequence. Binary columns were added to indicate combinations of missing and non-missing values between the C-3 Date and First Hearing Date. A green flag feature was introduced to identify cases where dates followed the correct chronological order: Accident Date → C-2 Date → Assembly Date → C-3 Date → First Hearing Date. **Tables B6 to B16** show the usefulness of these new features with the target.

To analyse temporal relationships, we calculated the number of days between key dates, such as Assembly to C-2 and C-2 to Accident, revealing patterns in event timelines. We also created 2 features that ensure compliance with regulations, such as the requirement for C-2 submission within 10 days of the accident and C-3 submission within two years.

Additionally, features like Year, Is Weekend, and Season were mined from the Accident Date to enhance the model's predictive power. After that, all the date columns were removed as we took all the information that was relevant.

Two features were derived from Average Weekly Wage: Wage Indicator, which flags wages above $1,757.19 (New York's average salary), and Is Employed, which indicates if the wage is non-zero.

The Features Industry and Avg Salary was derived from the Industry Code Description by assigning the average salary for each industry in New York. The Injury Category feature grouped industries into broader categories based on work type and risk, with missing values filled as 'Unknown.' One-hot encoded columns for injury categories were added to classify work types and associated risks for each job.

As mentioned earlier, we target encoded the WCIO and Industry features. For these features, we used only the descriptions and excluded the codes, as the descriptions provide more meaningful context and better represent categorical information.

New binary columns were added to identify whether Permanent Total Disability (PTD) occurs in the categorical columns. These columns indicate the presence of PTD and calculate the normalized frequency of each value in the categorical variables related to PTD. The purpose of this technique is to improve the model's understanding of the data and its relationship to the target variable, which is particularly difficult to predict due to its small representation in the data.

The Zip Code (2 & 3) variable was created to indicate whether the Zip Code starts with the digits 2 or 3 as it is the only ones that have different distribution for the target. **[Figure B14]**

4.2. Feature Selection

Using 20% of the training data, we then applied three complementary feature selection methods. First, we implemented Mutual Information (MI) Mutual Information which is a measure of the dependence or shared information between each feature and target. MI ranks features based on their importance. Retaining features with importance above 0.02, which reduced our feature count almost in half. Next, we employed RFE for XGBClassifier optimized for F1 macro score and Lasso Logistic Regression with scaled data, using a threshold of 0.01, 1000 iterations, and a regularization strength of 0.1. Our final set of features is the result of intersection of these three feature selection approaches. This resulted in a set of features, preparing us for the next phase of introducing new features to enhance model performance.

It is important to note the presence of highly correlated features, such as IME-4 Count and Missing IME-4 Count. Despite their correlation, we chose to include both in the model, as they were selected through our feature selection process. A potential area for future work is comparing model performance with and without these on of the features to understand their impact. **[Table B26]**

4.3. Modelling and Optimization approach

Our main approach involved creating an ensemble of different models to maximize predictive power. The algorithms used in this ensemble included Multi-Layer Perceptron (MLP), Logistic Regression, LightGBM, XGBoost, and Random Forest. A key step in the modeling process was hyperparameter tuning, which was efficiently performed using Optuna. Once the models were fitted, our strategy for building the ensemble focused on determining the optimal combination of model weights, which was also facilitated by Optuna.

4.4. Resampling strategy

The primary challenge in our project was the significant class imbalance. To address this, we explored various solutions and combinations before settling on our final approach. The minority classes were oversampled using SMOTE, while the majority classes were under sampled with Random Under Sampling. The optimal sampling strategy was determined through hyperparameter tuning with Optuna, ensuring the best balance for model training.

4.5. Performance Assessment

To access performance, we compared metrics between folds with an average of 0.464 F1 macro, 0.762 F1 weighted and 0.801 accuracy for regular data and 0.497 F1 macro, 0.758 F1 weighted and 0.764 accuracy for resampled data (for details see **Table C6**). We can see a high improvement in macro f1 and small decrease in f1 weighted and accuracy which is expected, while we do resampling, we better predicted minority classes, but we lost performance on majority due to under sampling.

In the model comparison, XGBoost emerged as the best-performing model, while Random Forest was the least effective, even falling behind Logistic Regression.

Multiclass prediction results vary by fold, but a clear pattern emerges: **PPD NSL** and **PTD** are rarely predicted, while **Non-comp** and **Temporary** perform well with F1 scores of ≈ 0.9 and 0.75. **Canceled** and **PPD SCH LOSS** achieve around 0.6 and 0.66, respectively. However, **Medical** struggles with an F1 score of around 0.3 due to its similarity to **Temporary**, despite not being a minority class. The weights refer to the relative importance assigned to each model in the ensemble. These weights are determined through optimization (using Optuna) to maximize the macro F1-score on the validation set. We can see that Random Forest is not used in any fold and in any data. The most important model was XGBoost and in some of the folds the Logistic Regression, MLP and LightGBM was also used. **[Tables C8 to C10]**

Final average F1 macro between all the folds was **0.464** for regular data and **0.497** for resampled data. The Kaggle performance of the best fold (Fold 2) was 0.439 for the regular data and 0.404 for the resampled data. This discrepancy may be due to overfitting, which could be an area for future investigation. For detailed results see **Tables C6 to C10**.

OPEN-ENDED SECTION

5.1. Agreement Reached

One of the most effective solutions we implemented in the open-ended section was predicting the Agreement Reached feature for the test dataset. For this prediction task, we adopted a different preprocessing approach for the WCIO and Industry Description features, as target encoding was not the most suitable method in this case. Instead, we grouped these categories in an ordinal fashion, where higher values were considered more likely to lead to an agreement than lower values. The model achieved an F1 macro score of 0.71, which we considered a strong result. When comparing the model's performance with and without this feature, we observed an improvement of 1-1.5% in both regular and resampled datasets (for detailed information, see **Tables C1 to C5**). In the final modelling phase, we included this feature, and as anticipated, it emerged as one of the most important predictors in the model.

5.2. Lime

In this part of the open-ended section, we used the Lime library to identify key features for different target values. For **Canceled Claims**, the most important feature was whether Agreement Reached was 0, indicating the claim was canceled when no agreement was made. Other significant features included IME-4 count and Average Weekly Wage being 0, indicating no salary and no forms sent. For **Non-compensated Claims**, the patterns were similar. Medical-only Claims showed a negative correlation with Average Weekly Wage and IME-4 count, suggesting that working individuals who sent at least one IME-4 form are more likely to receive medical claims. For **Temporary Claims**, there was a strong negative correlation with Agreement Reached being 0 and employment status was also important. For **Permanent Partial Disability (PPD) Schedule Loss**, the key features were similar, but correlations with newly created target-

encoded features appeared. For **PPD NSL**, **PTD**, and **Death** claims, feature importance was low. For all visualizations, see **Figures C1 to C8**.

5.3. Interface

The other part of the open-ended section is a web app, in which the user can choose values for all different features in the pre-processed and engineered train data. After the user input, a pre trained model (in this case a XGBoost model optimized with Optuna) predicts **the Claim Injury type** for the given values. This shows how the solutions of our work could be applied in an everyday scenario of a WCB employee who is part of the decision making of claim injury types. Future improvements could include Excel data uploads and multiple prediction models to provide users with diverse insights for decision-making. The code is available in the "open ended" GitHub folder.

CONCLUSION

This project demonstrates the transformative potential of machine learning in streamlining claims processing for the New York Workers' Compensation Board. By leveraging a robust multiclass classification model and addressing challenges such as class imbalance and data quality issues, we achieved notable improvements in predicting minority claim injury types.

Our final Kaggle submission achieved an F1 macro score of 0.439, while the best score during research and experimentation was 0.515. Interestingly, the higher score was achieved on a less optimized model than our final solution, suggesting there may still be unexplored things to improve performance in future. These findings underscore the complexity of optimizing machine learning solutions for real-world applications.

Beyond performance metrics, this project introduced practical tools, including a user-friendly web application for real-time predictions, highlighting the benefits machine learning can offer in supporting WCB operations. We hope this work not only enhances the efficiency and accuracy of claims processing but also lays a foundation for future innovations in this domain.

6.1. Limitations and Future Work

This project faced several challenges, particularly in predicting certain classes such as **Medical**, **PTD**, and **PPD NSL**, where the model's performance was poor. These limitations highlight opportunities for future improvement. Future research could explore the development of more advanced models and refined feature engineering techniques to better capture complex patterns in the data, for example further exploring Keras and PyTorch libraries.

Hyperparameter tuning, while effective using Optuna, may have converged to a local optimum, as indicated by differences between Kaggle and cross-validation results. This suggests potential overfitting, which warrants further investigation. Alternative hyperparameter tuning methods, such as GridSearchCV or Bayesian optimization, could be explored for comparison. Additionally, testing the impact of outlier removal on model performance could provide insights into improving data quality. Comparative analysis of datasets with and without outliers might help enhance robustness and predictive accuracy. By addressing these limitations, future work can build upon this foundation to deliver even more effective solutions for claims processing.

BIBLIOGRAPHICAL REFERENCES

Meyers AR, Al-Tarawneh IS, Wurzelbacher SJ, Bushnell PT, Lampl MP, Bell JL, Bertke SJ, Robins DC, Tseng CY, Wei C, Raudabaugh JA, Schnorr TM. Applying Machine Learning to Workers' Compensation Data to Identify Industry-Specific Ergonomic and Safety Prevention Priorities: Ohio, 2001 to 2011. J Occup Environ Med. 2018 Jan;60(1):55-73. doi: 10.1097/JOM.0000000000001162. PMID: 28953071; PMCID: PMC5868484.

Adcroft, P., & Toor, F. (2021, March 11). *Timeline: How COVID-19 Changed NYC*. Ny1.com. https://ny1.com/nyc/all-boroughs/news/2021/03/10/timeline--how-covid-19-changed-nyc

Eastman, A. Q., Rous, B., Langford, E. L., Anne Louise Tatro, Heebner, N. R., Gribble, P. A., Lanphere, R., & Abel, M. G. (2023). Etiology of Exercise Injuries in Firefighters: A Healthcare Practitioners' Perspective. *Healthcare*, *11*(22), 2989–2989. https://doi.org/10.3390/healthcare11222989

LightGBM. (n.d.). *LightGBM 3.3.2.99 documentation*. Lightgbm.readthedocs.io. https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html

New. (2014, December 4). *Assembled Workers' Compensation Claims: Beginning 2000*. Ny.gov. https://data.ny.gov/Government-Finance/Assembled-Workers-Compensation-Claims-Beginning-20/jshw-gkgu/about_data

New York State. (2024). *The Official Website of New York State*. Welcome to the State of New York; New York State. https://www.ny.gov/counties

NYC DCAS. (2018). *Workforce Profile Report NYC Government FY 2018*. https://www.nyc.gov/assets/dcas/downloads/pdf/reports/workforce_profile_report_fy_2018.pdf

Statista Research Department. (2024, July 5). *New York average annual pay U.S. 2001-2023*. Statista. https://www.statista.com/statistics/305766/new-york-annual-pay/

Sydney Pereira. (2020). *Remembering March 2020 In NYC: When The "Before Times" Came To An End*. Gothamist. https://gothamist.com/news/remembering-march-2020-nyc-when-times-came-end

Walton, A., & Rogers, B. (2017). Workplace Hazards Faced by Nursing Assistants in the United States: A Focused Literature Review. International Journal of Environmental Research and Public Health, 14(5), 544. https://doi.org/10.3390/ijerph14050544

Yasobant, S., & Rajkumar, P. (2014). Work-related musculoskeletal disorders among health care professionals. *Indian Journal of Occupational and Environmental Medicine*, *18*(2), 75. https://doi.org/10.4103/0019-5278.1468

APPENDIX A - LITERATURE REVIEW TABLE

This table provides an overview of studies related to predictive modelling and machine learning in workers' compensation. We included this table to showcase different approaches to similar problems we faced.

Table A1 - Literature Review with similarly themed studies

Article Title	Publication Name	Objective	Main Techniques	Main Conclusions	Methods, Techniques, and Tools	Reference
Applying Machine Learning to Workers' Compensation Data to Target Injury Prevention by Industry and Injury Type	Journal of Occupational and Environmental Medicine	To utilize machine learning for targeting injury prevention efforts by analyzing workers' compensation data.	Machine learning, auto-coding methods	Machine learning can effectively identify high-risk industry groups for specific injury types, aiding in targeted prevention strategies.	Utilized auto-coding methods to classify over 1.2 million claims, followed by machine learning techniques to rank industry groups based on injury types.	Meyers, A.R., Wurzelbacher, S., et al. (2017). Applying Machine Learning to Workers' Compensation Data to Target Injury Prevention. Journal of Occupational and Environmental Medicine.
Using Predictive Modelling for Workers Compensation Ratemaking	Casualty Actuarial Society Presentation	To introduce a predictive model for calculating recommended schedule	Predictive modelling, statistical analysis	Predictive modelling enhances the accuracy of schedule rating factors, leading to	Developed predictive models to calculate schedule rating factors, analyzed deviations, and implemented tiering factors based on the models.	Scott, S., Borba, P.S. (2016). Using Predictive Modelling for Workers Compensation Ratemaking. Casualty

		rating factors in workers' compensation ratemaking.		better risk assessment and pricing in workers' compensation.		Actuarial Society Presentation.
Predictive Modelling for Workers Compensation	EMB America LLC Presentation	To discuss solutions to challenges in modelling workers' compensation insurance risks.	Predictive modelling, data analysis	A structured predictive modelling process can address challenges in workers' compensation risk assessment, improving decision-making.	Outlined a predictive modelling process including project planning, data gathering, model building, analysis, and implementation.	Otto, D. (2007). Predictive Modelling for Workers Compensation. EMB America LLC Presentation.

APPENDIX B – DATA EXPLORATION

Table B1 - Data Exploration about all train data set variables in depth

Features	Data Exploration
Accident Date **[datetime]**	While looking at the data of injury, we could see that the most common date was 01/03/2020. This variable value is highly correlated with **COVID-19**, since this was the date when the first COVID-19 case appeared in NYC. With a correlation of 73,25% (although it was only confirmed on the 2nd of March). This was the start of many claims filled due to the impact of COVID-19 on the NY labour force and the many cases of infection, hospitalizations, and consequential deaths, that followed the pandemic. Although this is not an overwhelming reality for this dataset, as stated by the co-authors of New York Workers' Compensation Handbook, "Because COVID-19 has broadly affected communities and not just workplaces, employers and carriers are controverting many COVID-19 claims.". (Sydney Pereira, 2020) Monthly, the accidents dates follow a similar trend, however, from march to June 2020, assuming that it is an effect of the lock down, there are fewer accidents registered. The same goes for December 2022 where there is a decrease in accidents, as compared to 2020 and 2021, probably due to holidays in December and data being restricted to the end of December 2022.
Age at Injury **[int]**	If we look at the mean and median, we can see that they are almost the same, indicating that this variable may not be swayed by extreme values. When looking at the first, second and third quartiles (value 31, 42, 54 respectively) we can see that the distribution for this variable is likely to follow a normal distribution. Additionally, when comparing the data that we have, we can see that it aligns with the Workforce Profile Report of 2018 for NYC, where the median age is 43. Furthermore, we can see the presence of some outliers. We have a minimum age of 0 and maximum of 117. In the boxplot there is a notable presence of outliers, particularly individuals aged 89 years or older, indicating some cases where there are older people involved or falsely documented data. (NYC DCAS, 2018)

Alternative Dispute Resolution **[str]**	This variable refers if a claim is subject to adjudication processes external to the Board, indicating that these are probably the most severe or complicated cases of injury. These cases make up around 0,45% of the total cases.
Assembly Date **[datetime]**	This variable represents when the WCB initiates a claim. This happens when an injured worker misses more than one week of work, suffers a serious injury with potential permanent disability, faces a dispute from the insurer or employer, or submits a claim form (Form C-3). The Assembly Date data spans from January 2020 to December 2022, with an average assembly date of July 2021. The data is relatively evenly distributed across this three-year period, suggesting consistent claim initiation. (New, 2014) The date that has the most frequency of assemblies is 2020-03-06, which is 5 days later than the most frequent **Accident Date**. When both assemblies and accident dates are compared, there is an increase of both accidents and assemblies' trends followed by a steep decrease in the beginning of 2020. This marks the beginning of the spread and contamination of COVID-19 on a wide scale. Followed by a lock down where many workers were working from home, decreasing the risk of work injury. On the following months there were many increases and decreases, where most decreases occurred during holidays (especially December), as expected, since workers were not on their working places before. This representation helps us to see that the assembly dates did not vary much from the accident date, being an indicative of efficiency from WCB. (Adcroft & Toor, 2021)
C-2 Date **[datetime]**	For data sets published prior to the first quarter of 2022, the "C-2 Date" is the date of receipt of the "Employer's Report of Work-Related Injury/Illness". For data sets created in the first quarter of 2022 and forward, the "C-2 Date" is set to the earliest received date of equivalent injury/illness filings, including electronic data filed though the Board's E-Claims process. If broken down in years, the date of receipt starts from 1996, having few older cases of receipts received, and most forms being received in the more recent years, especially around September (median). (New, 2014)
COVID-19 Indicator	As explored before in **Accident Date**, the most cases where there was a positive value for the **COVID-19 Indicator** were related to the first start of

[str]	the pandemic spread. In total, the positive values represent 4,79% of the cases**.**
Agreement Reached [str]	This feature has binary values, and it represents, when 1, an agreement and when 0, no agreement. The data is very imbalanced, since 547 239 instances show no agreement and only 26 787 reached an agreement.
WCB Decision [str]	The only value available is "Not Work Related". Since this feature is not present in the test dataset, and in the train dataset only has one possible value, it will be dropped during preprocessing.
Attorney/ Representative [str]	This variable represents more complicated cases, where, using Boolean values, it shows a positive value in 31,66% of the cases, indicating that attorneys are significantly more commonly involved as compared to a processes external to the Board (**Alternative Dispute Resolution**).
Gender [str]	As mentioned before, men are the most common in the database as compared to women or nonbinary, since they have a weight of 58,88%. While observing the accidents over time by gender, we can observe similar trends for both women and men, except for the quantity difference for each (since we have different proportions of each).
Average Weekly Wage [float]	The workers represented in the dataset's average weekly wage has a mean of 1886.77, and a standard deviation of 414,64, indicating a lot of volatility in this value. It has around 58% of cases with 0, as majority of people got injured while volunteering, therefore, naturally receiving no wage. If this category is skipped in the analysis, the next most common wage interval is 750-1200, which is a little bit below the 2023 average salary in NY according to Statista. (Statista Research Department, 2024)
Birth Year [datetime]	The Birth Year is a variable that can be considered redundant, in the sense that it can be auxiliary in calculating **Age at injury** if it has a missing value. If grouped, we get that the bigger portion of people, (41,2% + 29,5%), that were injured belong to the age group born in the range of 1950 – 1990.

C-3 Date **[datetime]**	For this Date, even though around 65% of data was unavailable, it followed a similar pattern to the previous variable, in terms of statistics. Given that it represents when the Form from the employee was received.
Carrier Name **[str]**	The Carrier Name is the name of primary insurance provider responsible for providing workers' compensation coverage to the injured worker's employer. The most common Carrier Name is the "State Insurance Fund", which makes sense according to the Average Weekly Wage. Corroborating the supposition that the majority of people do not have the means of obtaining a private insurance provider. (New, 2014)
Carrier Type **[str]**	While comparing Carrier Type and Name, we can be misled while looking at the percentages from private and public insurance providers when looking at the top Carrier Types. However, the category of "Private" seems much higher, since it holds the highest percentage, however, this is due to the high number of different private insurance providers existing, in comparison to the "Self-Public" that represents the state insurance funding category. Also, as expected, both variables are highly correlated.
Claim Identifier **[int]**	This variable represented an Identifier for each row, and naturally being one of the few variables without missing values.
Claim Injury Type **[str]**	This is our target variable, and these are the types of Claim Injury, where the most common is "NON-COMP" first, with 50,71% of cases being attributed as such. Interestingly, this category shows a higher prevalence on Tuesdays, suggesting a potential pattern worth further investigation.
County of Injury **[str]**	There are 62 Counties represented in this dataset, where the most common are Suffolk, Queens and Kings, which represent all the counties in New York. (New York State, 2024)
District Name **[str]**	This variable represents the district of the worker, where the most common district, as expected, is NY.

First Hearing Date **[datetime]**	This represents the date the first hearing was held on a claim at a WCB hearing location. A blank date means the claim has not yet had a hearing held. However, this variable had many missing values, almost 74%.
IME-4 Count **[num]**	Represents the number of forms received by the Medical Examiner, there is a minimum of 1 and maximum of 73 forms received. This variable had around 77% of missing values.
Industry Code and Description **[str]**	The industries that were most common were “Health Care And Social Assistance”,” Public Administration” and “Educational Services”, this goes in accordance to the rest of information with have.
Medical Fee Region **[str]**	The most common region where the worker is assigned to receive medical care is “IV”, followed by “I”, which represent the following counties: Rural areas outside of Buffalo, Albany, Syracuse, Rochester, Utica & Binghamton and New York City, Nassau, and Western Suffolk. The high prevalence of Medical Fee Regions I and IV aligns with expected geographic and population-driven variations in medical needs and costs. Region IV’s concentration in large, urban counties likely reflects higher healthcare costs and demand in densely populated areas, while Region I’s broader distribution aligns with a more standard medical fee region across both rural and urban counties.
WCIO Cause of Injury Code, WCIO Cause of Injury Description **[str]**	The most frequent cause of injury is "Lifting," accounting for 8.3% of cases, which can be associated with more manual tasks, particularly in professions like nursing, healthcare aides, and volunteers such as firefighters, where tasks like repositioning patients or handling heavy equipment lead to higher injury rates. Besides having around 2.6% of missing data, some inconsistencies were identified, such as multiple **WCIO Cause of Injury Codes** being associated with the same description. (Eastman et al., 2023) (Walton & Rogers, 2017)
WCIO Nature of Injury Code, WCIO Nature of	Following the logic before created, the most frequent nature of injury is “Strain or Tear” (27.5%), followed by Concussion (19.8%). These injuries are commonly associated with the causes mentioned before, where

Injury Description **[str]**	workers are at a higher risk for musculoskeletal injuries and head trauma. (Yasobant & Rajkumar, 2014)
WCIO Part Of Body Code, WCIO Part of Body Description **[str]**	There are around 3% of missing values in these features. Some inconsistencies were identified between the WCIO Part of Body Code and the associated descriptions, since there were found multiple codes for the same description again. The most frequently injured body part is the Lower Back Area (9.3%), followed by the Knee (8.6%) and Multiple Parts (7.5%). The least frequent injuries involve the Larynx, Trachea and Artificial Appliances.
Zip Code **[str]**	There are around 5% of missing values in this feature. Around 157 inconsistent zip codes have been identified during this step of exploring, such as zip codes with less than 5 characters and others with letters and numbers mixed. Attached to this document there is also a map that indicates the zip codes of the workers in this dataset. This can be useful for further geographical exploration.
WCB Decision **[str]**	The only value available is "Not Work Related". Since this feature is not present in the test dataset, and in the train dataset only has one possible value, it will be dropped during preprocessing.
Number of Dependents **[int]**	The distribution of the number of dependents is balanced across categories, ranging from 0 to 6 dependents, with a mean and median of 3, indicating low to no probability of extreme values. Each group contains about 81 000 to 82 000 individuals, which indicates that there is no tendency toward any specific number of dependents.

Table B2 – Train Before Preprocessing

Variable	Data Type	Count	Unique	Missing	Percentage Missing
Accident Date	object	593471	5540	23134	3.90
Age at Injury	float64	593471	109	19445	3.28
Alternative Dispute Resolution	object	593471	4	19445	3.28
Assembly Date	object	593471	1096	0	0.00
Attorney/Representative	object	593471	3	19445	3.28
Average Weekly Wage	float64	593471	120025	48096	8.10
Birth Year	float64	593471	108	48523	8.18
C-2 Date	object	593471	2476	34005	5.73
C-3 Date	object	593471	1649	406226	68.45
Carrier Name	object	593471	2047	19445	3.28
Carrier Type	object	593471	9	19445	3.28
Claim Identifier	int64	593471	593470	0	0.00
Claim Injury Type	object	593471	9	19445	3.28
County of Injury	object	593471	64	19445	3.28
COVID-19 Indicator	object	593471	3	19445	3.28
District Name	object	593471	9	19445	3.28
First Hearing Date	object	593471	1095	442673	74.59
Gender	object	593471	5	19445	3.28
IME-4 Count	float64	593471	42	460668	77.62
Industry Code	float64	593471	25	29403	4.95
Industry Code Description	object	593471	21	29403	4.95
Medical Fee Region	object	593471	6	19445	3.28
OIICS Nature of Injury Description	float64	593471	1	593471	100.00
WCIO Cause of Injury Code	float64	593471	78	35085	5.91

WCIO Cause of Injury Description	object	593471	75	35085	5.91
WCIO Nature of Injury Code	float64	593471	57	35102	5.91
WCIO Nature of Injury Description	object	593471	57	35102	5.91
WCIO Part Of Body Code	float64	593471	58	36527	6.15
WCIO Part Of Body Description	object	593471	55	36527	6.15
Zip Code	object	593471	10061	48082	8.10
Agreement Reached	float64	593471	3	19445	3.28
WCB Decision	object	593471	2	19445	3.28
Number of Dependents	float64	593471	8	19445	3.28

Table B3 – Test

Variable	Data Type	Count	Unique	Missing	Percentage Missing
Accident Date	object	387975	3439	2444	0.63
Age at Injury	float64	387975	102	0	0.00
Alternative Dispute Resolution	object	387975	3	0	0.00
Assembly Date	object	387975	434	0	0.00
Attorney/Representative	object	387975	2	0	0.00
Average Weekly Wage	float64	387975	39009	19204	4.95
Birth Year	float64	387975	103	19470	5.02
C-2 Date	object	387975	1049	9134	2.35
C-3 Date	object	387975	627	302759	78.04
Carrier Name	object	387975	1598	0	0.00
Carrier Type	object	387975	7	0	0.00
Claim Identifier	int64	387975	387975	0	0.00
County of Injury	object	387975	63	0	0.00
COVID-19 Indicator	object	387975	2	0	0.00
District Name	object	387975	8	0	0.00
First Hearing Date	object	387975	339	344947	88.91
Gender	object	387975	4	0	0.00
IME-4 Count	float64	387975	19	352726	90.91
Industry Code	float64	387975	25	7736	1.99
Industry Code Description	object	387975	21	7736	1.99
Medical Fee Region	object	387975	5	0	0.00
WCIO Cause of Injury Code	float64	387975	78	10348	2.67

WCIO Cause of Injury Description	object	387975	75	10348	2.67
WCIO Nature of Injury Code	float64	387975	57	10560	2.72
WCIO Nature of Injury Description	object	387975	57	10560	2.72
WCIO Part Of Body Code	float64	387975	55	9549	2.46
WCIO Part Of Body Description	object	387975	55	9549	2.46
Zip Code	object	387975	6277	19342	4.99
Number of Dependents	float64	387975	7	0	0.00

Table B4 – Train After Preprocessing

Variable	Data Type	Count	Unique	Missing	Percentage Missing
Accident Date	object	574026	5540	3689	0.64
Age at Injury	float64	574026	108	0	0.00
Alternative Dispute Resolution	object	574026	3	0	0.00
Assembly Date	object	574026	897	0	0.00
Attorney/Representative	object	574026	2	0	0.00
Average Weekly Wage	float64	574026	120025	28651	4.99
Birth Year	float64	574026	108	29078	5.07
C-2 Date	object	574026	2476	14560	2.54
C-3 Date	object	574026	1649	386781	67.38
Carrier Name	object	574026	2046	0	0.00
Carrier Type	object	574026	8	0	0.00
Claim Injury Type	object	574026	8	0	0.00
County of Injury	object	574026	63	0	0.00
COVID-19 Indicator	object	574026	2	0	0.00
District Name	object	574026	8	0	0.00
First Hearing Date	object	574026	1095	423228	73.73
Gender	object	574026	4	0	0.00
IME-4 Count	float64	574026	42	441223	76.86
Industry Code	float64	574026	25	9958	1.73
Industry Code Description	object	574026	21	9958	1.73
Medical Fee Region	object	574026	5	0	0.00
WCIO Cause of Injury Code	float64	574026	78	15640	2.72

WCIO Cause of Injury Description	object	574026	75	15640	2.72
WCIO Nature of Injury Code	float64	574026	57	15657	2.73
WCIO Nature of Injury Description	object	574026	57	15657	2.73
WCIO Part Of Body Code	float64	574026	58	17082	2.98
WCIO Part Of Body Description	object	574026	55	17082	2.98
Zip Code	object	574026	10061	28637	4.99
Agreement Reached	float64	574026	2	0	0
Number of Dependents	float64	574026	7	0	0

Figure B1 – Correlation Matrix of Numerical Features (Claim Injury Type Numeric)

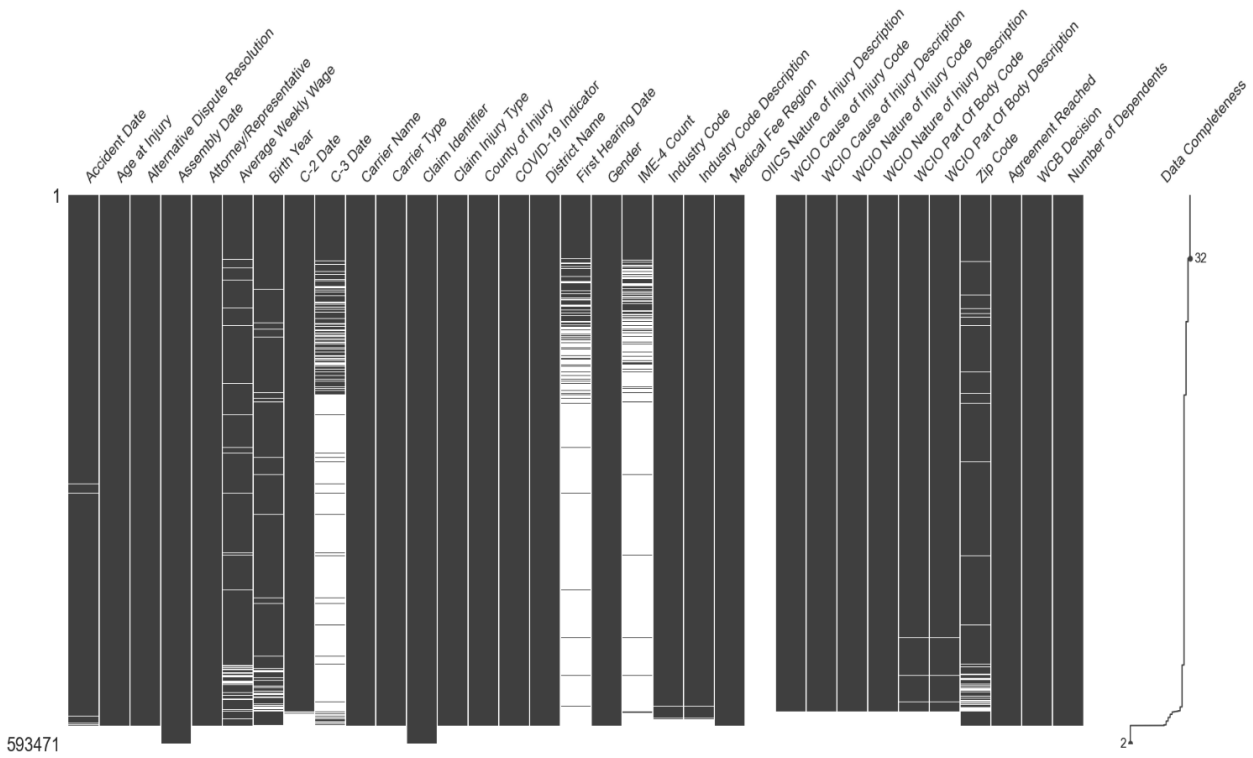

Figure B2 – Missing Values in all features

Table B5 – COVID –19 Correlation with March 1, 2020

COVID-19 Indicator on March 1, 2020	Count	Percentage
Y	912	73.25%
N	333	26.75%

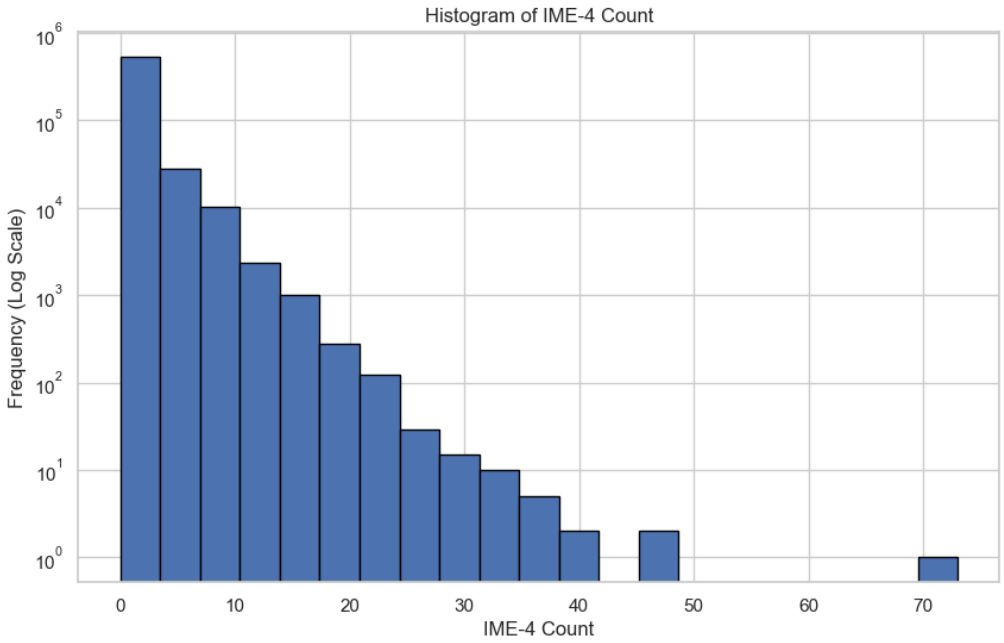

Figure B3 – Histogram of IME-4 Count

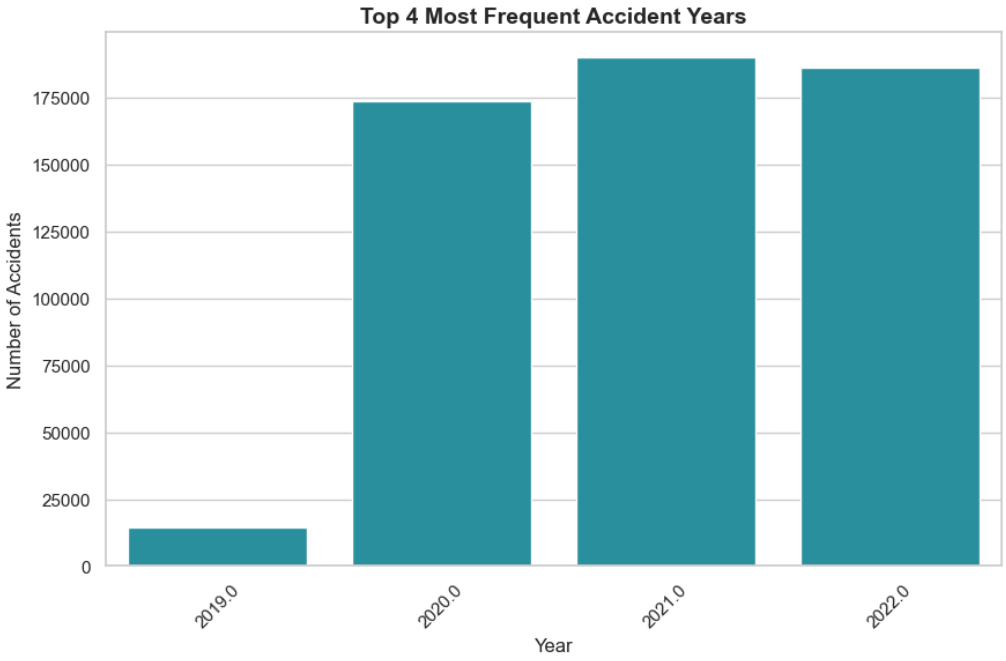

Figure B4 – Top 4 Most Frequent Accident Year

Table B6 – Comparing target with Green flag Dates

Green flag Dates	Claim Injury Type	Count	Percentage
0	0	12273	2.36
0	1	288772	55.57
0	2	62591	12.04
0	3	115587	22.24
0	4	37676	7.25
0	5	2245	0.43
0	6	55	0.01
0	7	454	0.09
1	0	204	0.38
1	1	2306	4.24
1	2	6315	11.61
1	3	32920	60.54
1	4	10604	19.50
1	5	1966	3.62
1	6	42	0.08
1	7	16	0.03

Table B7 – Comparing target with Missing First Hearing Date

Missing First Hearing	Claim Injury Type	Count	Percentage
0	0	2195	1.46
0	1	14241	9.44
0	2	22219	14.73
0	3	76695	50.86
0	4	30706	20.36
0	5	4175	2.77
0	6	97	0.06
0	7	470	0.31
1	0	10282	2.43
1	1	276837	65.41
1	2	46687	11.03
1	3	71812	16.97
1	4	17574	4.15
1	5	36	0.01

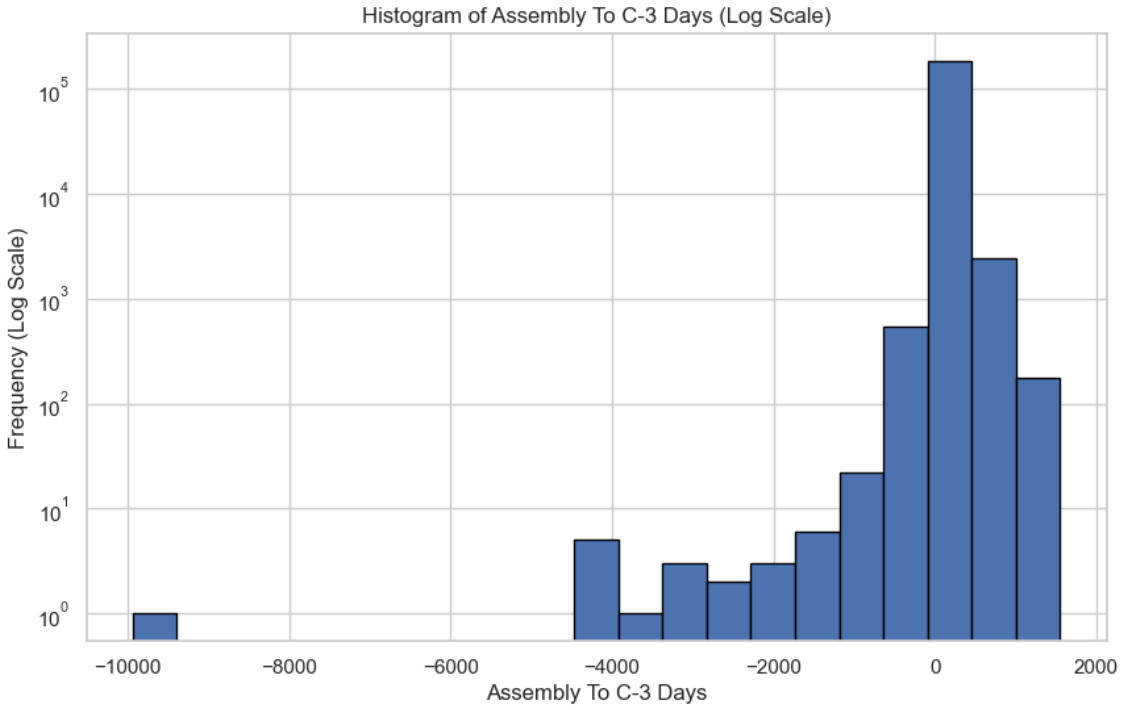

Figure B5 – Histogram of Assembly To C-3 Days (Log Scale)

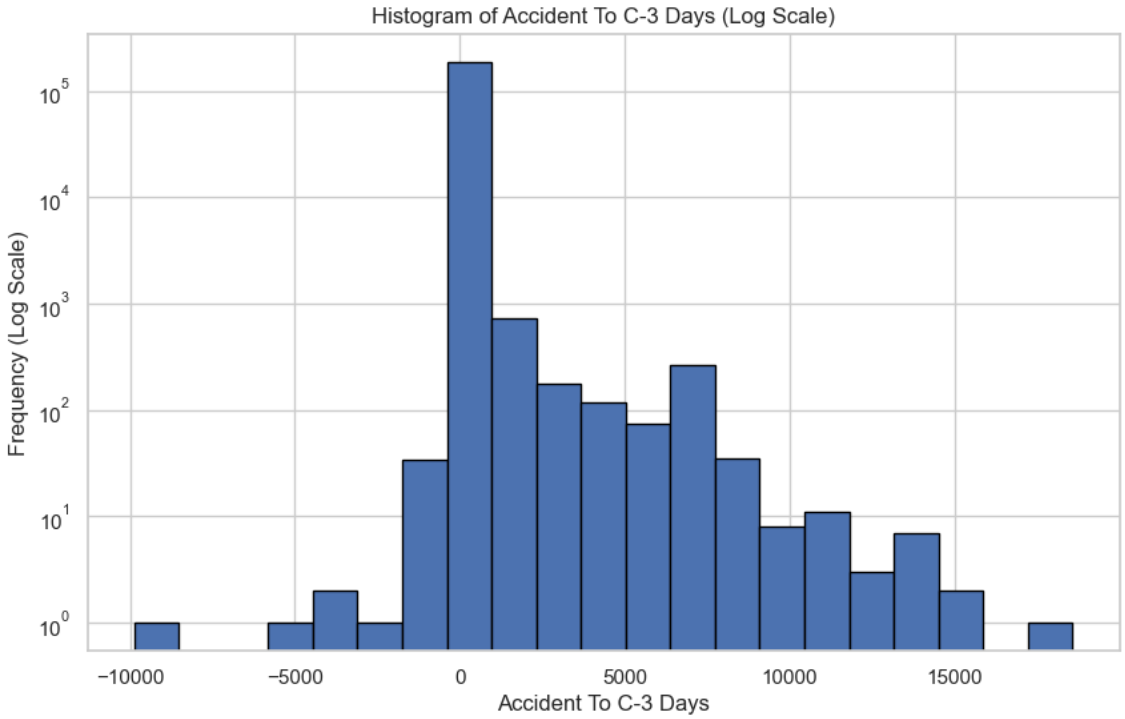

Figure B6 – Histogram of Accident To C-3 Days (Log Scale)

Table B8 – Comparing target with RedFlag_Accident_to_C-3

RedFlag_Accident_to_C-3	Claim Injury Type	Count	Percentage
0	0	11435	2.01
0	1	289047	50.77
0	2	68464	12.03
0	3	148023	26.00
0	4	47580	8.36
0	5	4197	0.74
0	6	95	0.02
0	7	470	0.08
1	0	1042	22.10
1	1	2031	43.08
1	2	442	9.37
1	3	484	10.27
1	4	700	14.85
1	5	14	0.30
1	6	2	0.04

Table B9 – Comparing target with RedFlag_Assembly_to_C-3

RedFlag_Assembly_to_C-3	Claim Injury Type	Count	Percentage
0	0	7370	1.50
0	1	271601	55.44
0	2	58456	11.93
0	3	120524	24.60
0	4	28469	5.81
0	5	2938	0.60
0	6	68	0.01
0	7	457	0.09
1	0	5107	6.07
1	1	19477	23.15
1	2	10450	12.42
1	3	27983	33.26
1	4	19811	23.54
1	5	1273	1.51
1	6	29	0.03
1	7	13	0.02

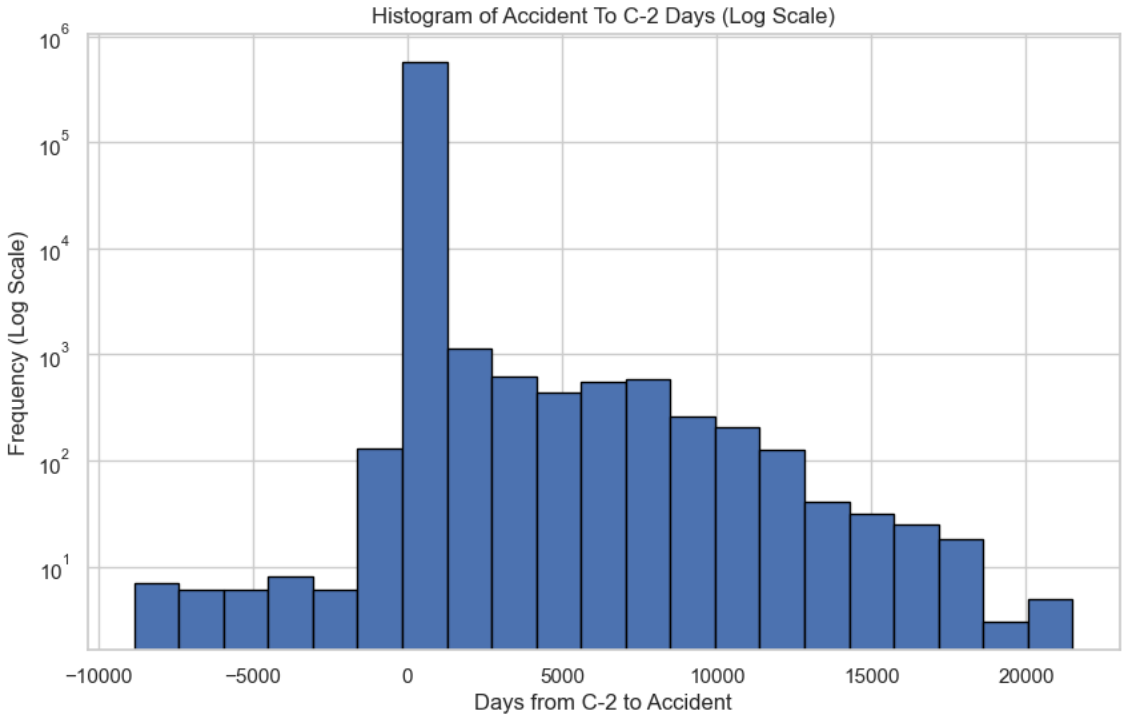

Figure B7 – Histogram of Accident To C-2 Days (Log Scale)

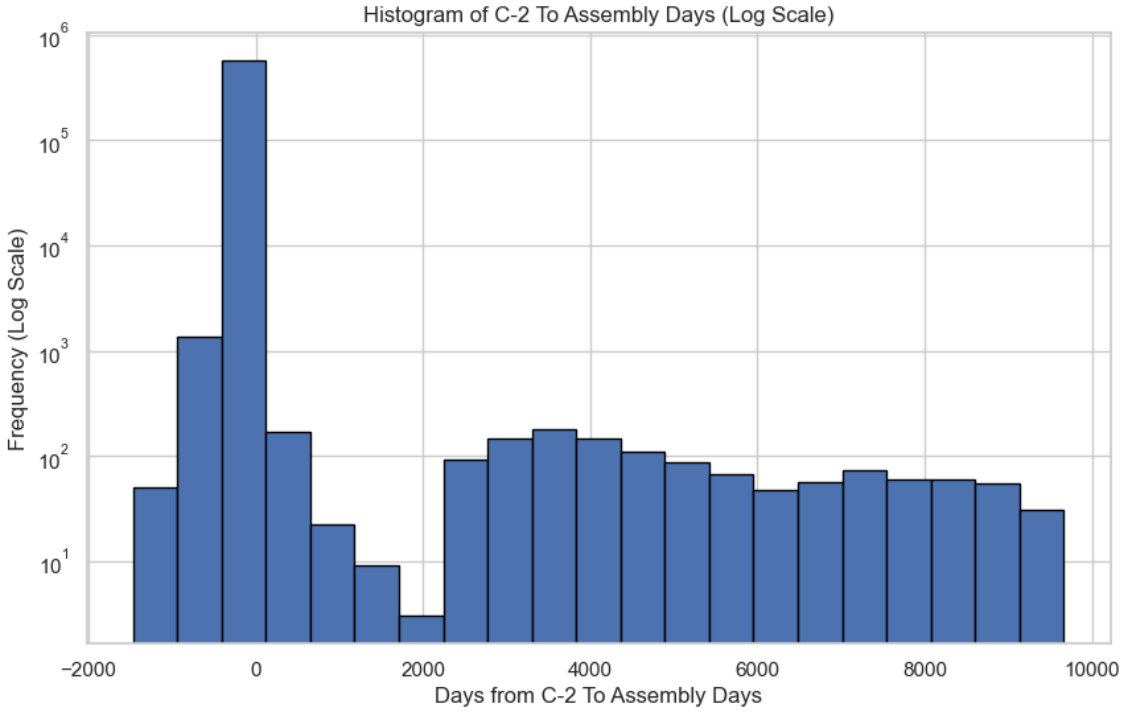

Figure B8 – Histogram of C-2 To Assembly Days (Log Scale)

Table B10 – Comparing target with Red_Flag_Accident to C_2

Red_Flag_Accident_To_C-2	Claim Injury Type	Count	Percentage
0	0	12464	2.18
0	1	291010	50.79
0	2	68652	11.98
0	3	148243	25.87
0	4	47874	8.35
0	5	4199	0.73
0	6	94	0.02
0	7	469	0.08
1	0	13	1.27
1	1	68	6.66
1	2	254	24.88
1	3	264	25.86
1	4	406	39.76
1	5	12	1.18
1	6	3	0.29
1	7	1	0.10

Table B11 – Comparing target with Red_Flag_C-2_To_Assembly

Red_Flag_C-2_To_Assembly	Claim Injury Type	Count	Percentage
0	0	11023	2.18
0	1	275917	54.62
0	2	59173	11.71
0	3	124845	24.71
0	4	30780	6.09
0	5	3198	0.63
0	6	76	0.02
0	7	169	0.03
1	0	1454	2.11
1	1	15161	22.02
1	2	9733	14.14
1	3	23662	34.37
1	4	17500	25.42
1	5	1013	1.47
1	6	21	0.03
1	7	301	0.44

Table B12 – Comparing target with C-3 Date_Missing_First Hearing Date_Missing

C-3 Date_Missing_First Hearing Date_Missing	Claim Injury Type	Count	Percentage
FALSE	0	6605	2.90
FALSE	1	43741	19.17
FALSE	2	31695	13.89
FALSE	3	97712	42.83
FALSE	4	43596	19.11
FALSE	5	4199	1.84
FALSE	6	97	0.04
FALSE	7	470	0.21
TRUE	0	5872	1.70
TRUE	1	247337	71.50
TRUE	2	37211	10.76
TRUE	3	50795	14.68
TRUE	4	4684	1.35
TRUE	5	12	0.00

Table B13 – Comparing target with C-3 Date_Missing_First Hearing Date_Not _Missing

First Hearing Date_Not_Missing_C-3	Claim Injury Type	Count	Percentage
FALSE	0	11482	2.15
FALSE	1	285464	53.54
FALSE	2	61052	11.45
FALSE	3	129769	24.34
FALSE	4	41899	7.86
FALSE	5	3364	0.63
FALSE	6	79	0.01
FALSE	7	47	0.01
TRUE	0	995	2.43
TRUE	1	5614	13.74
TRUE	2	7854	19.22
TRUE	3	18738	45.85
TRUE	4	6381	15.61
TRUE	5	847	2.07
TRUE	6	18	0.04
TRUE	7	423	1.03

Table B14 – Comparing Target with C-3 Date_Not_Missing_First Hearing Date_Not_Missing

C-3 Date_Not_Missing_First Hearing Date_Not_Missing	Claim Injury Type	Count	Percentage
FALSE	0	11277	2.43
FALSE	1	282451	60.86
FALSE	2	54541	11.75
FALSE	3	90550	19.51
FALSE	4	23955	5.16
FALSE	5	883	0.19
FALSE	6	18	0.00
FALSE	7	423	0.09
TRUE	0	1200	1.09
TRUE	1	8627	7.85
TRUE	2	14365	13.07
TRUE	3	57957	52.72
TRUE	4	24325	22.13
TRUE	5	3328	3.03
TRUE	6	79	0.07
TRUE	7	47	0.04

Table B15 – Comparing target with C-3 Date_Not_Missing_First Hearing Date_Missing

C-3 Date_Not_Missing_ First Hearing Date_Missing	Claim Injury Type	Count	Percentage
FALSE	0	8067	1.62
FALSE	1	261578	52.66
FALSE	2	59430	11.96
FALSE	3	127490	25.67
FALSE	4	35390	7.12
FALSE	5	4187	0.84
FALSE	6	97	0.02
FALSE	7	470	0.09
TRUE	0	4410	5.70
TRUE	1	29500	38.15
TRUE	2	9476	12.26
TRUE	3	21017	27.18
TRUE	4	12890	16.67
TRUE	5	24	0.03

Table B16 – Comparing target with Is Weekend

Is Weekend	Claim Injury Type	Count	Percentage
FALSE	0	10783	2.22
FALSE	1	246167	50.67
FALSE	2	58495	12.04
FALSE	3	125542	25.84
FALSE	4	40832	8.40
FALSE	5	3608	0.74
FALSE	6	90	0.02
FALSE	7	351	0.07
TRUE	0	1694	1.92
TRUE	1	44911	50.94
TRUE	2	10411	11.81
TRUE	3	22965	26.05
TRUE	4	7448	8.45
TRUE	5	603	0.68
TRUE	6	7	0.01
TRUE	7	119	0.13

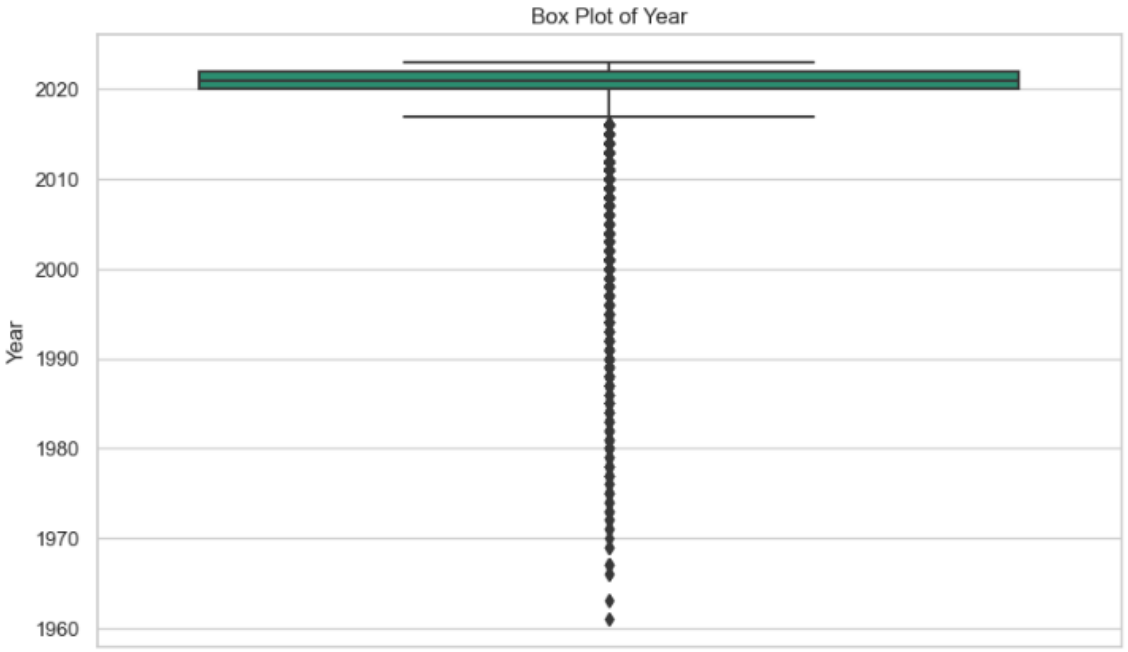

Figure B9 - Box Plot of Year

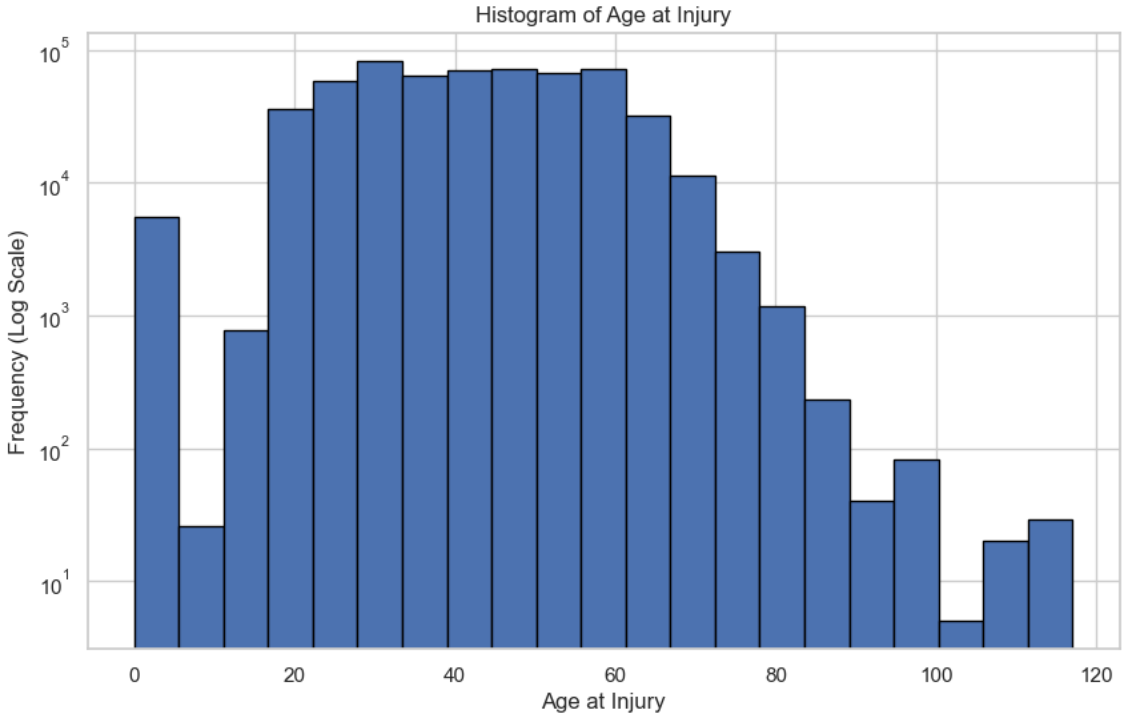

Figure B10 – Histogram of Age at Injury

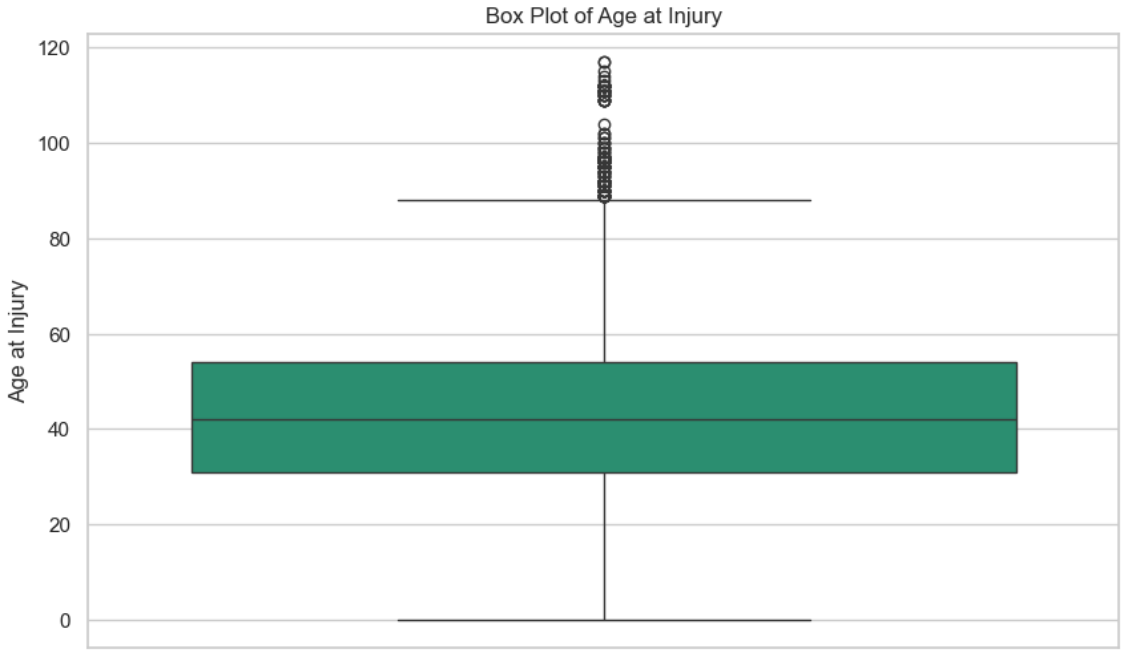

Figure B11 – Box Plot Age at Injury

Table B17 – Comparing target with Red_Flag_Age at Injury

Red_Flag_Age at Injury	Claim Injury Type	Count	Percentage
0	0	12450	2.18
0	1	289835	50.68
0	2	68399	11.96
0	3	148197	25.91
0	4	48248	8.44
0	5	4209	0.74
0	6	97	0.02
0	7	469	0.08
1	0	27	1.27
1	1	1243	58.58
1	2	507	23.89
1	3	310	14.61
1	4	32	1.51
1	5	2	0.09
1	7	1	0.05

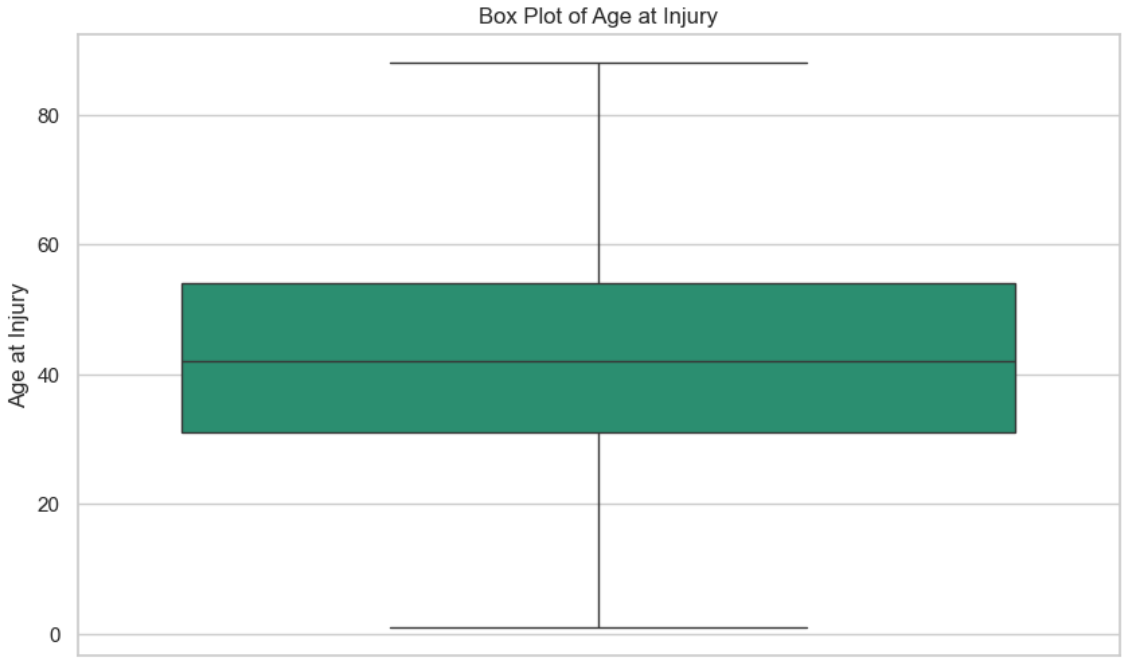

Figure B12 – Box Plot Age at Injury without outliers

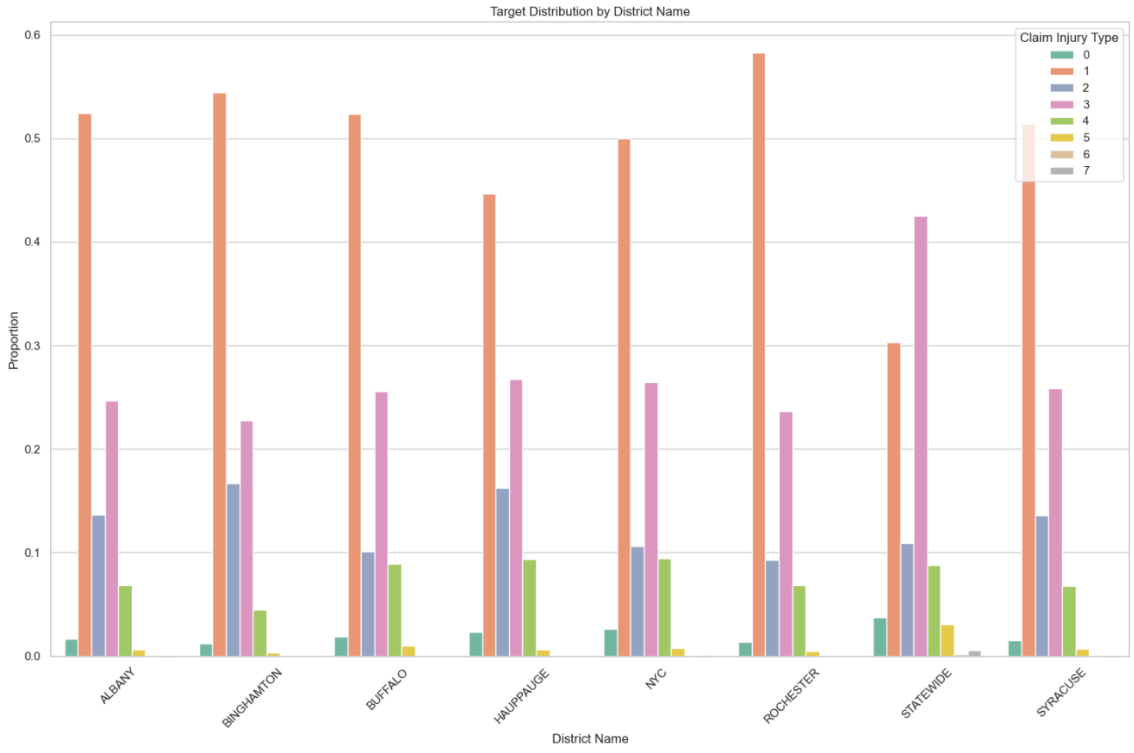

Figure B13 – Target Distribution by District Name

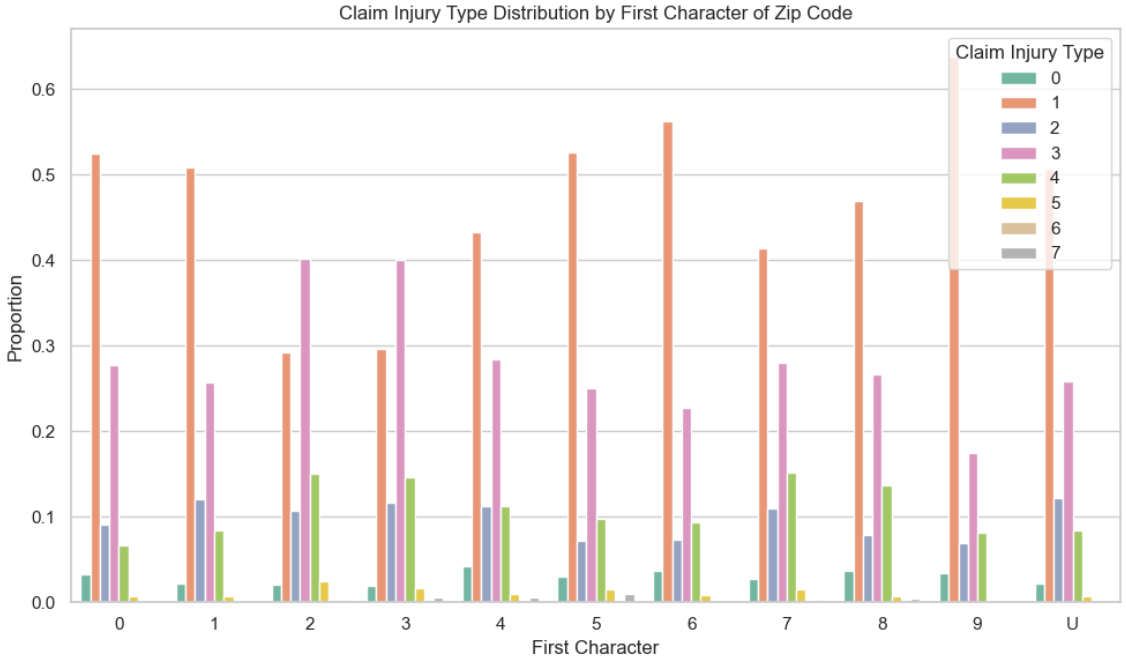

Figure B14 – Claim Injury Type Distribution by First Character of Zip Code

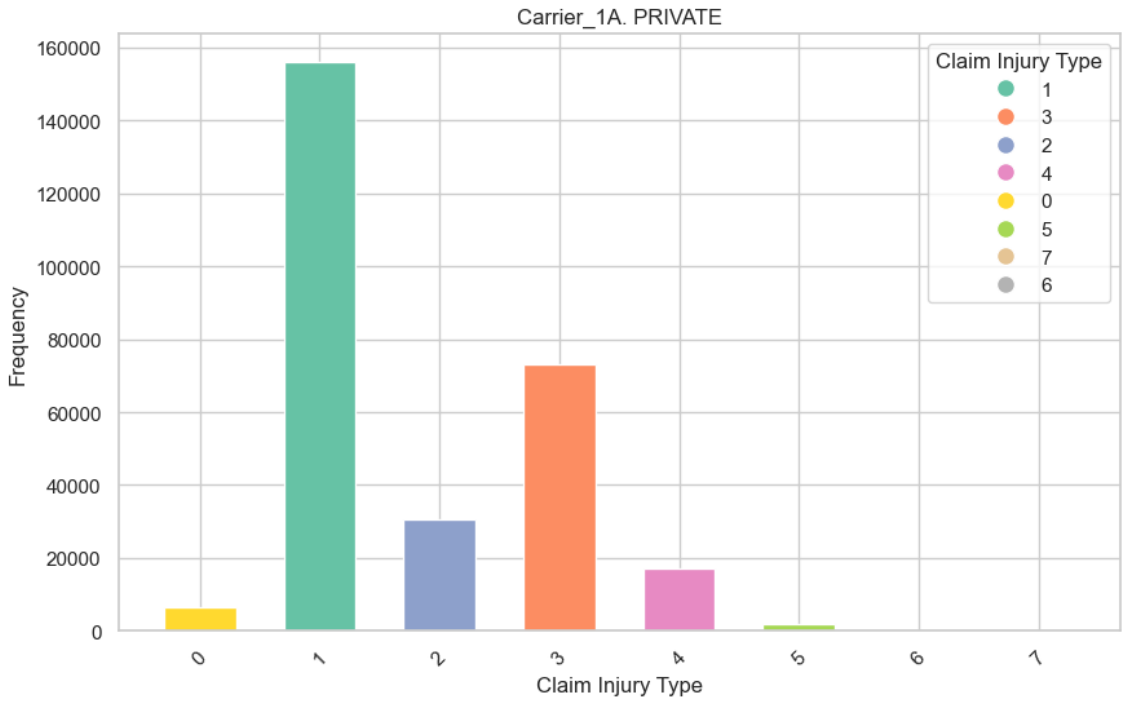

Figure B15 – Claim Injury Type Distribution for Carrier_1A. Private

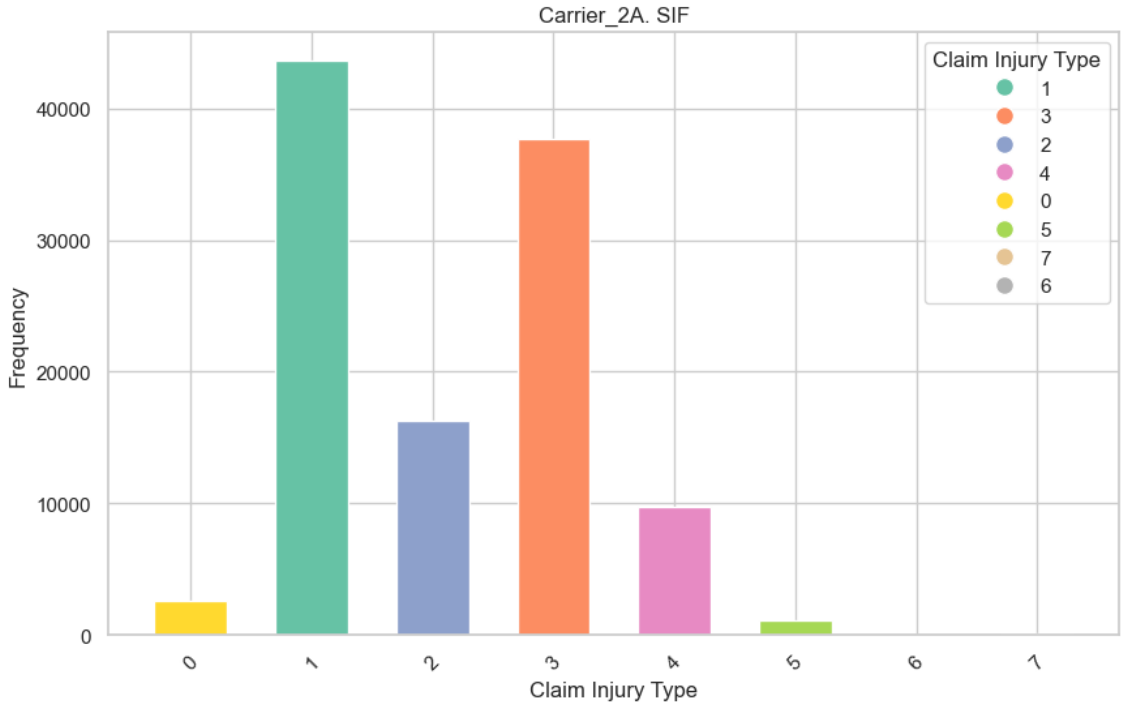

Figure B16 – Claim Injury Type Distribution for Carrier_2A. SIF

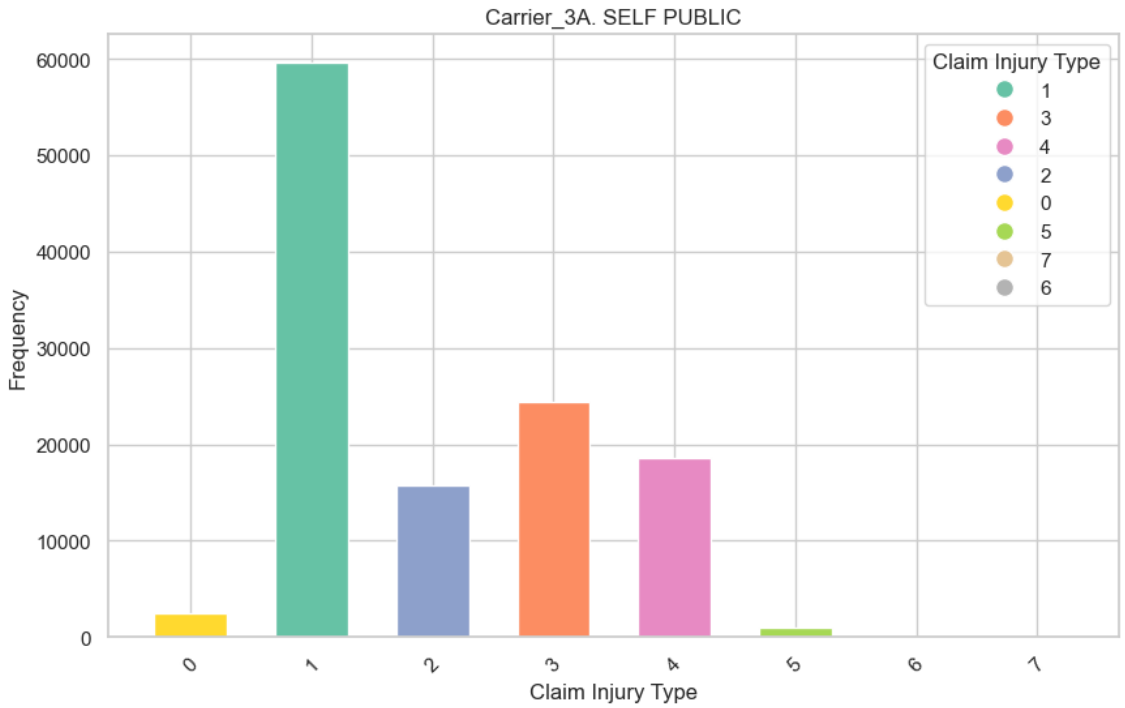

Figure B17 – Claim Injury Type Distribution for Carrier_3A. Public

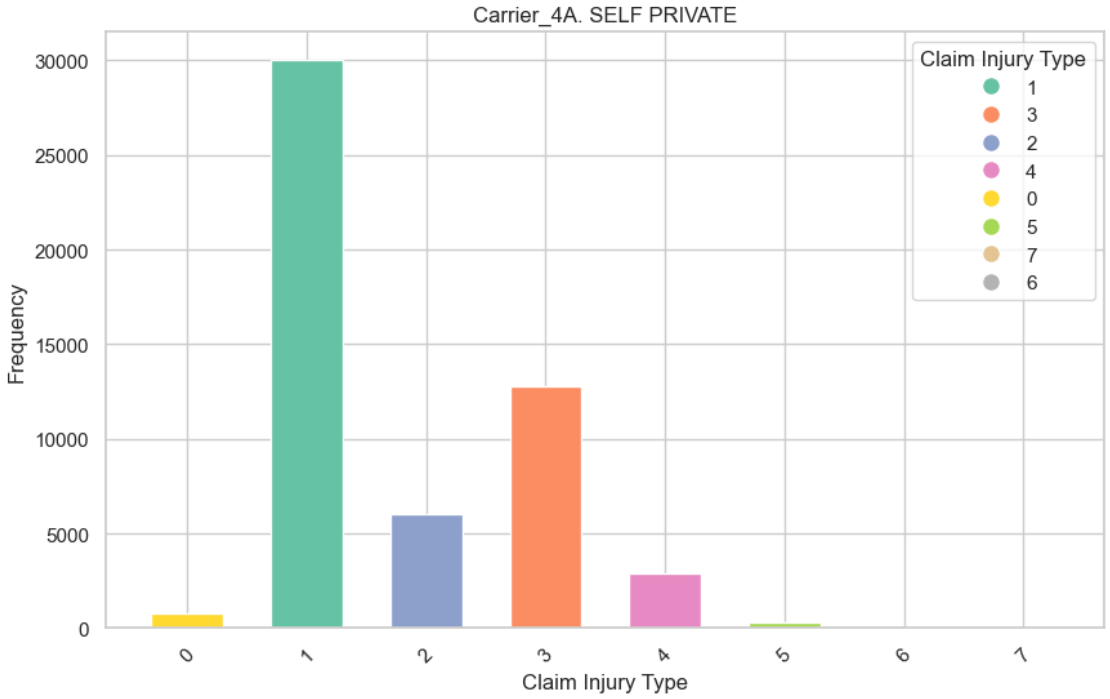

Figure B18 – Claim Injury Type Distribution for Carrier_4A. SELF Private

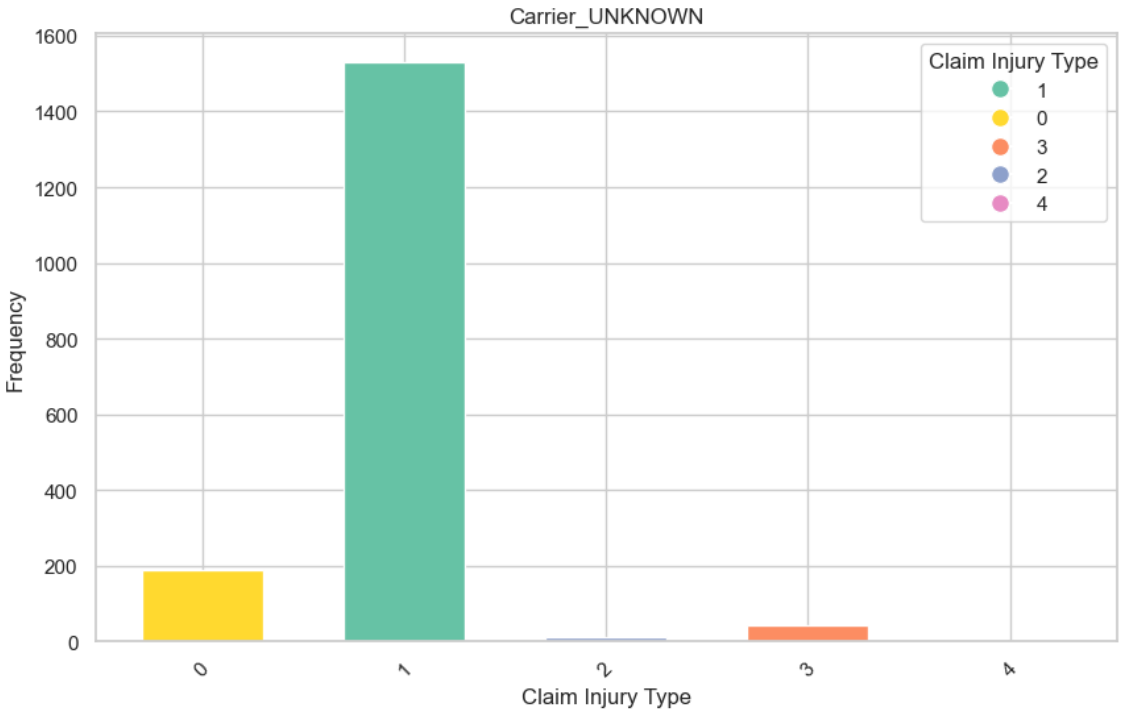

Figure B19 – Claim Injury Type Distribution for Carrier_Unknown

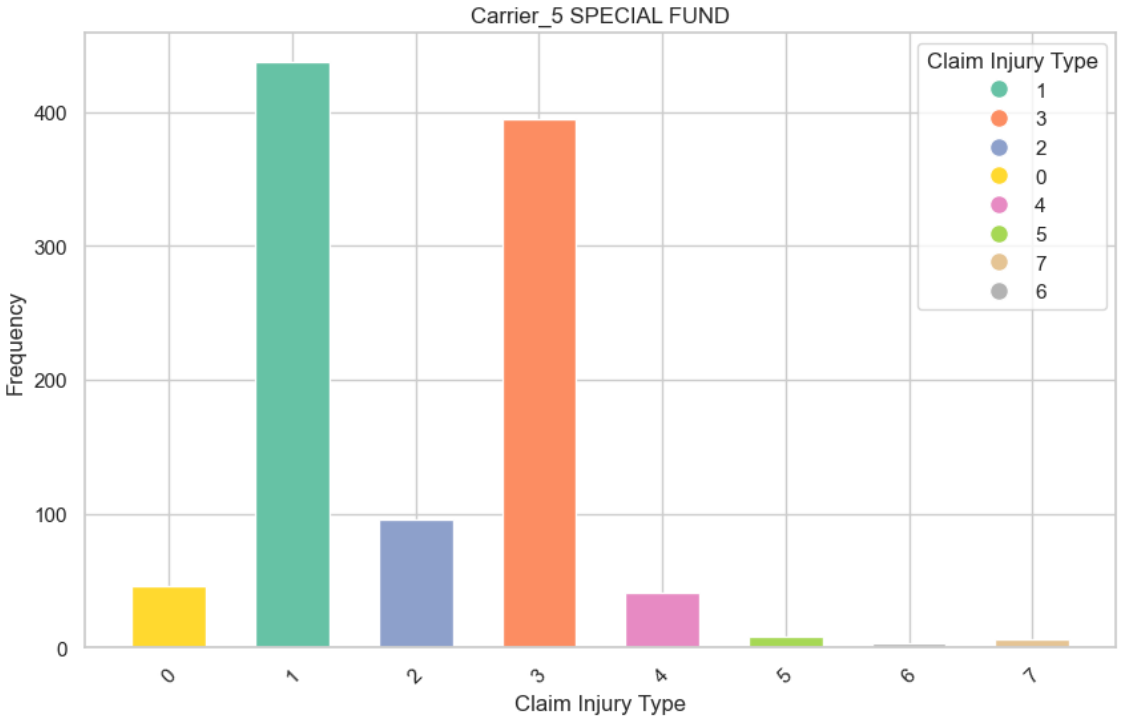

Figure B20 – Claim Injury Type Distribution for Carrier_5 Special Fund

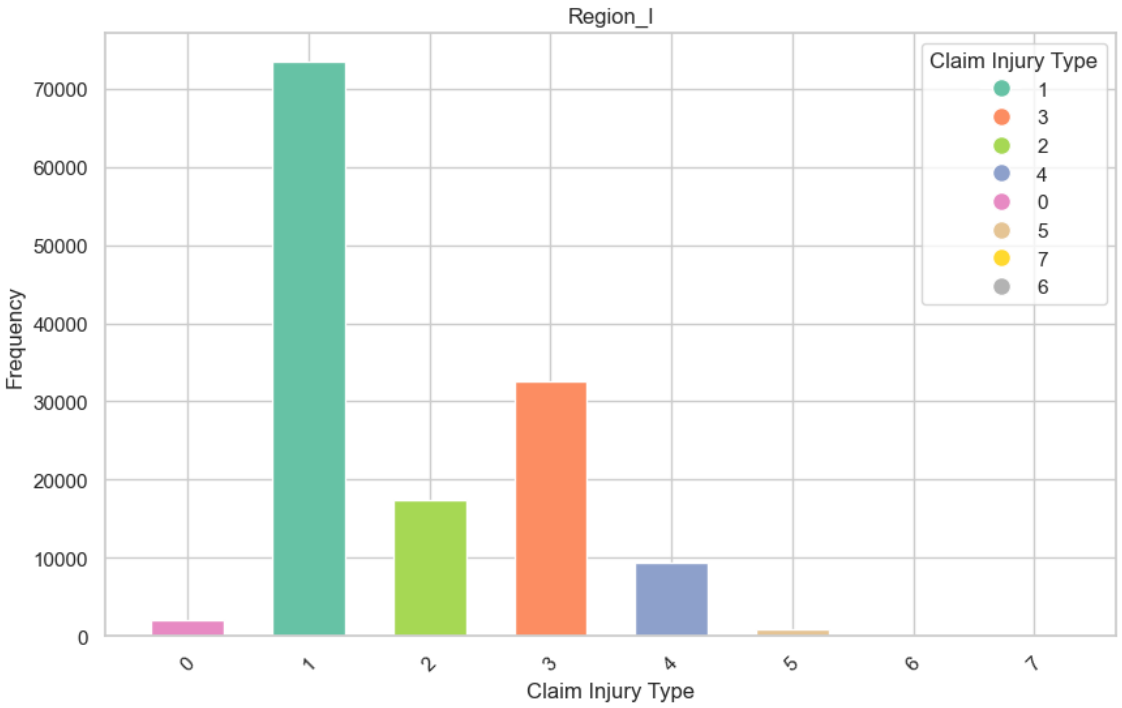

Figure B21 – Claim Injury Type Distribution for Region I

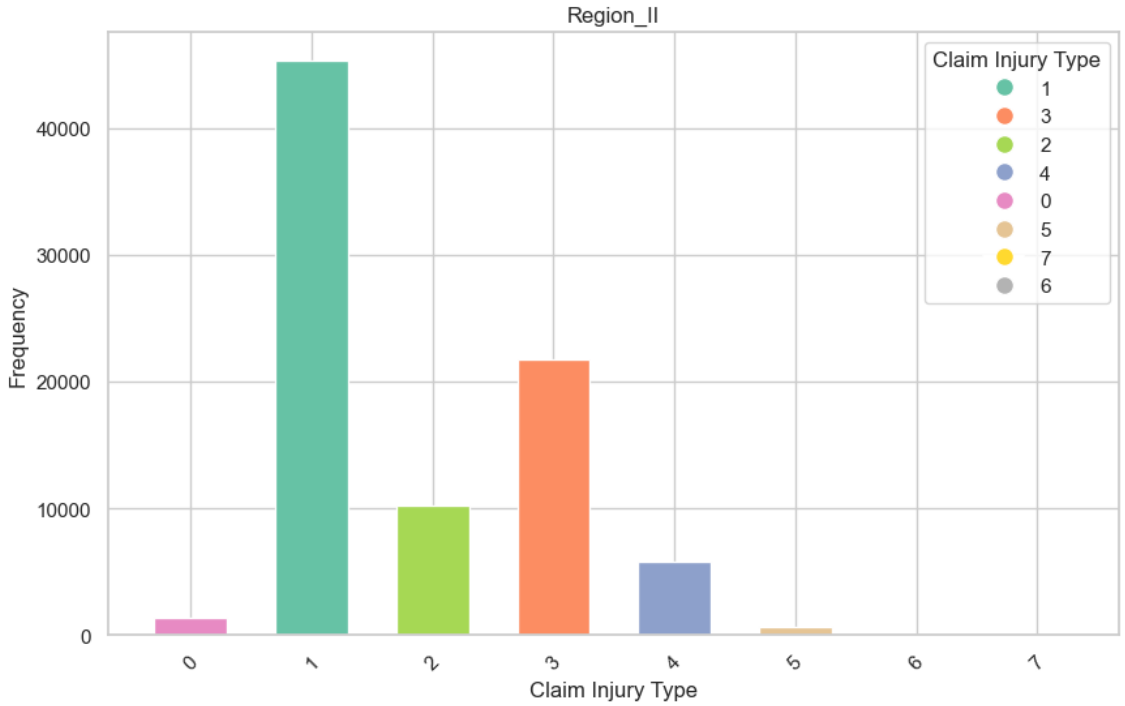

Figure B22 – Claim Injury Type Distribution for Region II

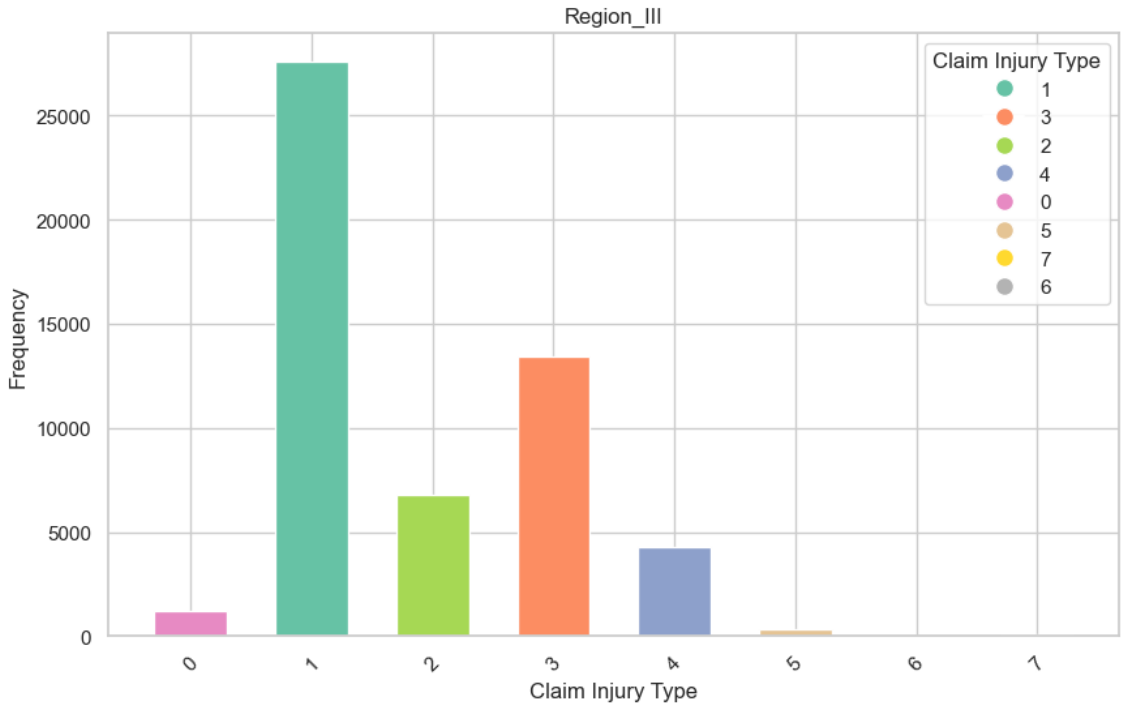

Figure B23 – Claim Injury Type Distribution for Region III

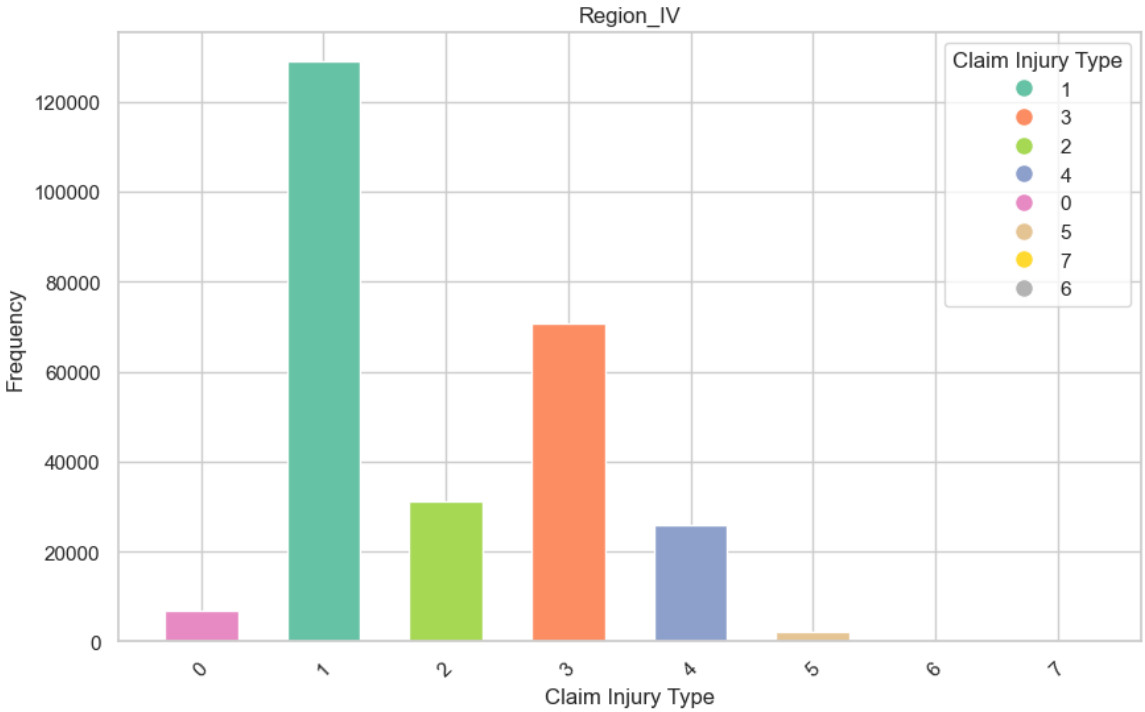

Figure B24 – Claim Injury Type Distribution for Region IV

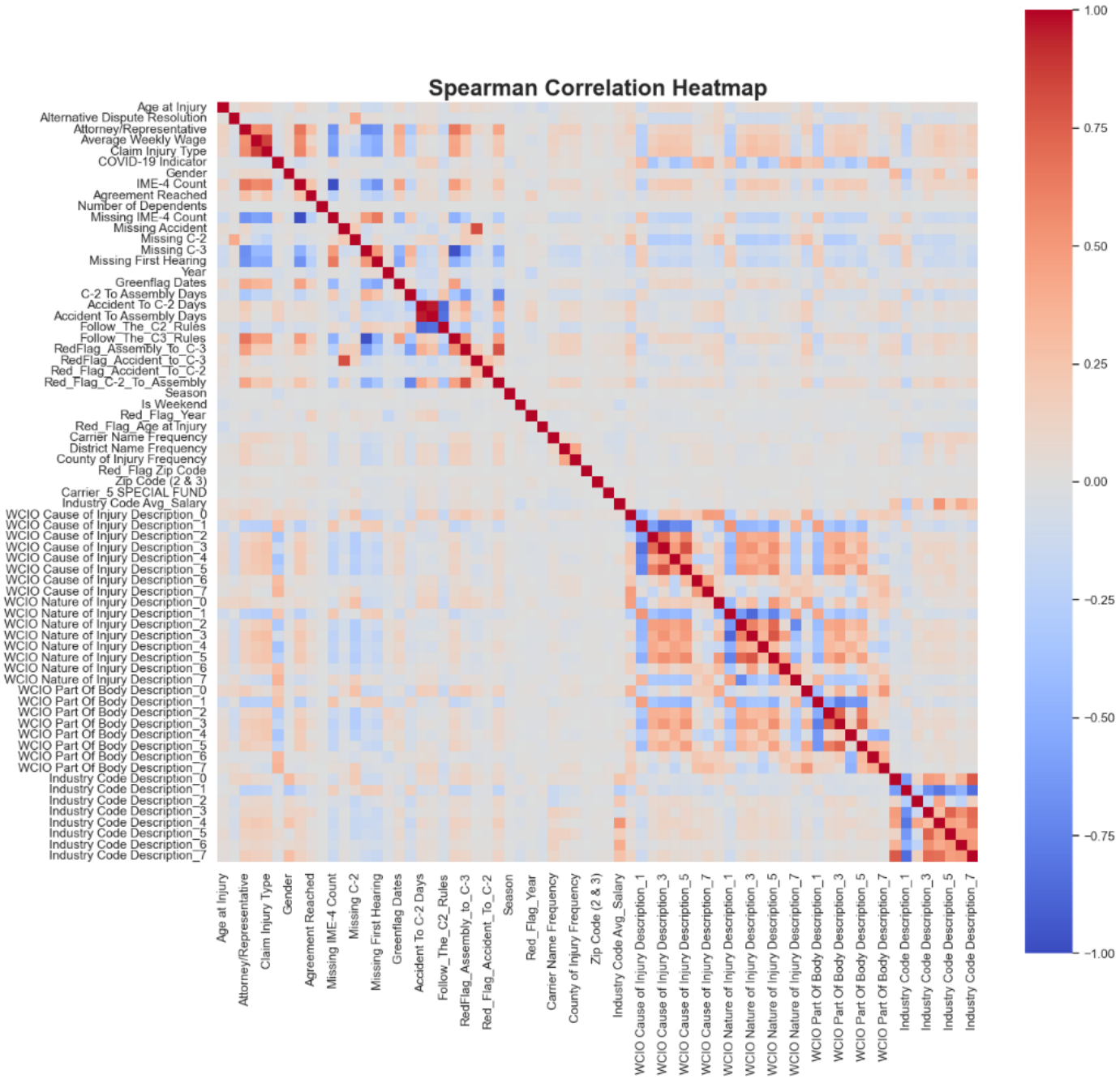

Figure B25 – Spearman Correlation Heatmap after Preprocessing

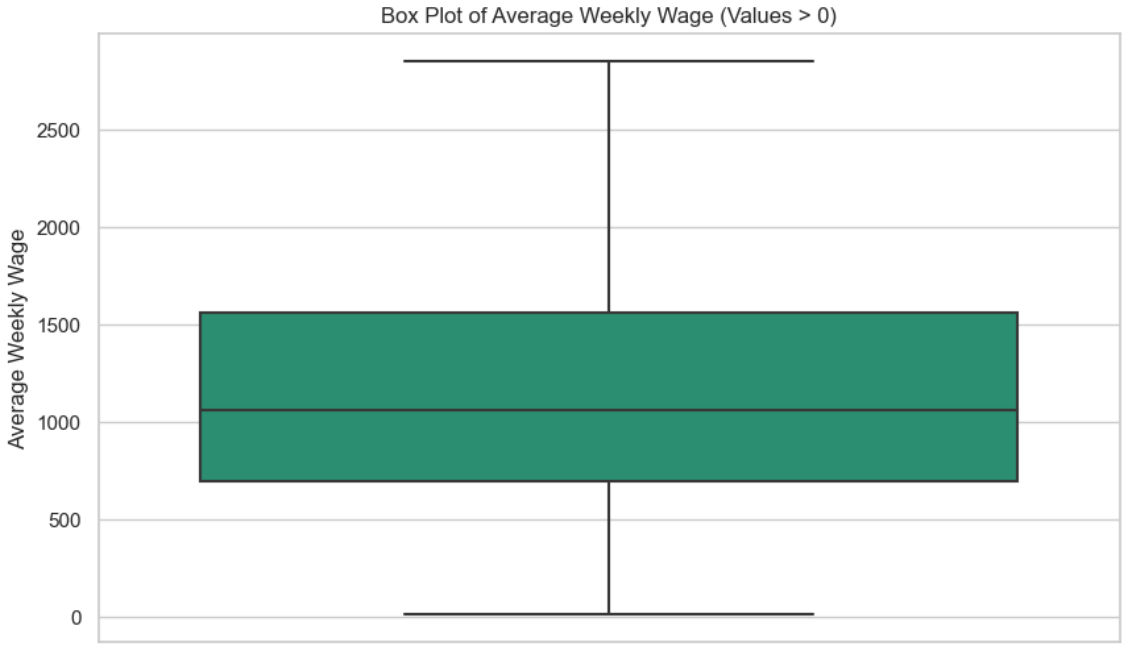

Figure B26 – Box Plot of Average Weekly Wage (Values > 0)

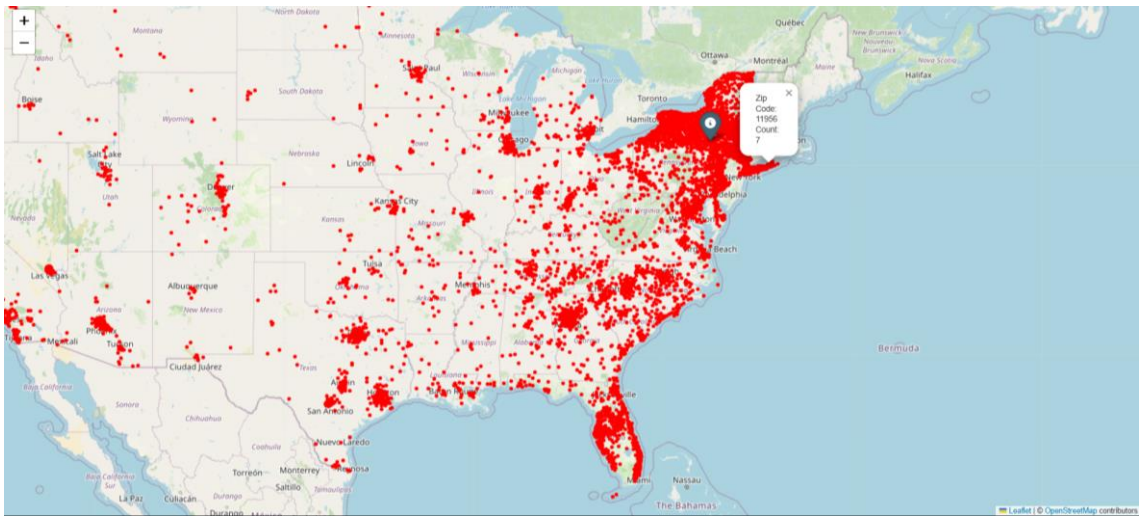

Figure B27 - Dynamic map with zip codes from workers (also attached with the rest of the deliverables for better visualization

Table B18 – Carrier Types

Carrier Type	Count	Percentage
1A. PRIVATE	285368	49.87%
3A. SELF PUBLIC	121920	21.31%
2A. SIF	111144	19.42%
4A. SELF PRIVATE	52788	9.22%
5D. SPECIAL FUND - UNKNOWN	1023	0.18%
5C. SPECIAL FUND - POI CARRIER WCB MENANDS	5	0.00%
5A. SPECIAL FUND - CONS. COMM. (SECT. 25-A)	4	0.00%

Table B19 – Claim Injury Type by descending order

Claim Injury Type	Count	%
2. NON-COMP	291078	50.71
4. TEMPORARY	148507	25.87
3. MED ONLY	68906	12.00
5. PPD SCH LOSS	48280	8.41
1. CANCELLED	12477	2.17
6. PPD NSL	4211	0.73
8. DEATH	470	0.08
7. PTD	97	0.02

Table B20 – COVID-19 Indicator representation

COVID-19 Indicator	Count	%
N	546505	95.21
Y	27521	4.79

Table B21 – Industry Code and Description

Industry Code	Industry Code Description	Count	%
62.0	Health Care And Social Assistance	114339	20.27
92.0	Public Administration	92240	16.35
44.0, 45.0	Retail Trade	61638	10.93
48.0, 49.0	Transportation And Warehousing	54023	9.58
61.0	Educational Services	44393	7.87
31.0, 32.0, 33.0	Manufacturing	38150	6.76
23.0	Construction	30903	5.48
72.0	Accommodation And Food Services	26456	4.69
56.0	Administrative And Support And Waste Management And Remediat	21027	3.73
42.0	Wholesale Trade	15236	2.70
81.0	Other Services (Except Public Administration)	13149	2.33
71.0	Arts, Entertainment, And Recreation	9967	1.77
54.0	Professional, Scientific, And Technical Services	9770	1.73
51.0	Information	9166	1.62
53.0	Real Estate And Rental And Leasing	8948	1.59
52.0	Finance And Insurance	8214	1.46
22.0	Utilities	2980	0.53
11.0	Agriculture, Forestry, Fishing And Hunting	2404	0.43
21.0	Mining	695	0.12
55.0	Management Of Companies And Enterprises	370	0.07

Table B22 – Medical Fee Region and Claim Injury Type

Medical Fee Region	1. Cancelled	2. Non-	3. Med	4. Temporary	5. Ppd	6. Ppd Nsl	7. Ptd	8. Death
I	2059	73391	17432	32607	9441	821	24	110
II	1339	45266	10217	21755	5754	648	18	36
III	1203	27573	6799	13410	4286	324	6	53
IV	6871	128904	31184	70843	25866	2089	38	186

Table B23 – Medical Fee Region Percentage

Medical Fee Region	Count	%
IV	265981	46.34
I	135885	23.67
II	85033	14.81
III	53654	9.35
NaN	33473	5.83

Table B24 – WCIO Cause of Injury Code for the same Description

WCIO Cause of Injury **Description**	**WCIO** Cause of Injury **Code**
OBJECT BEING LIFTED OR HANDLED	79.0, 17.0, 66.0
REPETITIVE MOTION	97.0, 94.0

Table B25 – Explanation of used algorithms throughout the project

Algorithm	Usage	Brief Explanation	Advantages and Disadvantages
Frequency Encoder	Data Preprocessing	Attributes a code to a category based on its occurrence in the dataset.	Can be used to encode Categorical Data in a simple way, however, it is not able to distinguish between two features with the same frequency.
XGBClassifier (E**X**treme **G**radient **Boost**ing **Classifier**)	Modelling – Model Selection	This algorithm uses belongs to the ensemble category, specifically, gradient boosting framework- This means that it tries to minimize a loss function - using decision trees that are combined to form a strong predictive model.	It is highly flexible (has many parameters), robust due to regularization techniques (using LASSO or RIDGE) and can handle many types of data including missing values. However, it is hard to interpret, besides being computationally complex.
LGBMClassifier (**Li**ght **G**radient **B**oosting **M**achine **Classifier**)	Modelling – Model Selection	It is very similar to XGBoostClassifier, however, it is designed to be quicker and lighter (memory wise), however, instead of growing trees level-wise, it grows leaf-wise, focusing on the splits with the maximum loss reduction, leading to deeper trees. It also uses a histogram-based algorithm that bins continuous variables into intervals (makes training faster). Furthermore, it offers features that handle unbalanced data.	Its more appropriate for larger datasets, as compared to the previous algorithm, since it is faster and needs less memory for a similar performance. Due to the leaf growth, it can lead to overfitting, and it is hard to interpret, it also requires more attention in small datasets in order to prevent overfitting.
Mutual Information Statistic	Modelling – Feature Selection	It measures the independence between two variables, *xi* and *xj*, where: $MI_{i,j} = \sum_{i,j} P\left(x_i, x_j\right) \log \frac{P\left(x_i, x_j\right)}{P\left(x_i\right) P\left(x_j\right)}$ Where *p(xi)* and *p(xj)* are the marginal probability and *pxij* is the joint probability density function. The main	Captures Non-linear relationships, it is useful to select features, it is symmetric (which means it does not distinguish the direction of the relationship). However, it is computationally expensive since it needs to estimate probability densities. Besides

		difference with correlation is that correlation only measures linear relationships, while mutual information can measure all types of relationships, including non-linear ones.	not providing direction of the relationship, it also does not
SMOTE (**S**ynthetic **M**inority **O**versampling **Te**chnique)	Modelling – Feature Selection	Addresses categorical data that is imbalanced, (where the categories have very different proportions from each other), typically the minority class has synthetic samples added to it, avoiding bias towards the majority class. It uses the Nearest Neighbors, and one sample is created between a point and its neighbor.	This helps to balance datasets and prevents overfitting, consequently, improving the model's ability to learn patterns in the minority class. However, this can also amplify the noise, in case of outliers, contributing for the worsening of the model. Furthermore, when datasets have overlapping class distributions, this method can reduce distinction between classes.
Optuna	Modelling – Hyperparameter Tuning	Open-source optimization framework that helps to find the best hyperparameters to improve models' performance. Automates the process of searching for the best hyperparameters (such as learning rate, number of layers, etc.) by utilizing advanced optimization algorithms. It works by runs trials, evaluating different sets of hyperparameters and pruning poorly performing ones early to save resources, this is done many times until the set of best hyperparameters is found.	This allows to find the best hyperparameters without much trial and error and it supports parallel execution. It also reduces computational costs by stopping poorly performing trials early. However, it can lead to overfitting and can be computationally expensive for complex models, even though it uses pruning. Besides that, the quality of the optimization is highly dependent on the objective function.
Random**U**nder**S**ampler	Modelling – Resampling Strategy	It randomly selects observations from the majority class, reducing it, without changing the minority class.	It works quickly, balancing the dataset and allowing for better performances, however, it may lead to loss of information and might overfit the minority class, SMOTE is a good alternative.

Table B26 – Stratified K-Fold Preprocessing's Feature Selection

Fold Number	Number of selected features	Number of selected features	Number of selected features	Best average F1 macro
1	65	52	43	0.497
2	65	54	37	0.496
3	68	54	34	0.498

Table B27 – Numeric Variables Statistical Information

Variable	Min	Q1	Median	Q3	Max
Age at Injury	0.0	31.0	42.0	54.0	117.0
Average Weekly Wage	0.0	0.0	0.0	841.0	2828079.0
Birth Year	0.0	1965.0	1977.0	1989.0	2018.0
IME-4 Count	1.0	1.0	2.0	4.0	73.0
Industry Code	11.0	45.0	61.0	71.0	92.0
WCIO Cause of Injury Code	1.0	31.0	56.0	75.0	99.0
WCIO Nature of Injury Code	1.0	16.0	49.0	52.0	91.0
WCIO Part Of Body Code	-9.0	33.0	38.0	53.0	99.0
Agreement Reached	0.0	0.0	0.0	0.0	1.1
Number of Dependents	0.0	1.0	3.0	5.0	6.0

APPENDIX C - RESULTS

Table C1 – Agreement Reached prediction results

	Precision	Recall	F1-score	Support
0	0.98	0.95	0.97	109449
1	0.38	0.59	0.46	5357
Accuracy			0.94	114806
Macro Avg	0.68	0.77	0.71	114806
Weighted Avg	0.95	0.94	0.94	114806

Table C2 – Classification Report of Non-Resampled data **WITHOUT** Agreement Reached

Class	Precision	Recall	F1-Score	Support
1	0.737	0.525	0.613	2496
2	0.854	0.968	0.907	58216
3	0.539	0.080	0.139	13781
4	0.712	0.882	0.788	29702
5	0.694	0.583	0.634	9656
6	0.667	0.002	0.005	842
7	0.000	0.000	0.000	19
8	0.661	0.436	0.526	94
Accuracy			0.789	114806
Macro Avg	0.608	0.435	0.451	114806
Weighted Avg	0.762	0.789	0.748	114806

Table C3 – Classification Report of Resampled data **WITHOUT** Agreement Reached

Class	Precision	Recall	F1-score	Support
1	0,575	0,652	0,611	2496
2	0,881	0,899	0,890	58216
3	0,359	0,246	0,292	13781
4	0,751	0,774	0,762	29702
5	0,581	0,746	0,653	9656
6	0,216	0,010	0,018	842
7	0,250	0,053	0,087	19
8	0,542	0,553	0,547	94
Accuracy			0,763	114806
Macro Avg	0,519	0,492	0,483	114806
Weighted Avg	0,748	0,763	0,752	114806

Table C4 – Classification Report of Non-Resampled data **WITH** Agreement Reached

Class	Precision	Recall	F1-score	Support
1	0,747	0,517	0,611	2496
2	0,863	0,971	0,914	58216
3	0,600	0,089	0,155	13781
4	0,742	0,909	0,817	29702
5	0,685	0,648	0,666	9656
6	0,143	0,001	0,002	842
7	0,000	0,000	0,000	19
8	0,695	0,436	0,536	94
Accuracy			0,804	114806
Macro Avg	0,559	0,446	0,463	114806
Weighted Avg	0,777	0,804	0,763	114806

Table C5 – Classification Report of Resampled data **WITH** Agreement Reached

Class	Precision	Recall	F1-score	Support
1	0,656	0,599	0,626	2496
2	0,873	0,948	0,909	58216
3	0,376	0,256	0,305	13781
4	0,836	0,683	0,751	29702
5	0,540	0,861	0,664	9656
6	0,161	0,044	0,069	842
7	0,333	0,053	0,091	19
8	0,560	0,596	0,577	94
Accuracy			0,774	114806
Macro Avg	0,542	0,505	0,499	114806
Weighted Avg	0,766	0,774	0,762	114806

Table C6 – Metrics of Non-Resampled Data vs Resampled Data for each fold

	Fold		
Metric	**1**	**2**	**3**
Non-Resampled Data			
F1 Macro	0.462	0.465	0.466
F1 Weighted	0.762	0.762	0.761
Accuracy	0.802	0.802	0.800
Resampled Data			
F1 Macro	0.497	0.496	0.498
F1 Weighted	0.759	0.757	0.757
Accuracy	0.768	0.759	0.766

Table C7 – Best Weights for Ensemble in Resampled and Non-Resampled Data for each Fold

Algorithm	1	2	3
Resampled Data			
RF	0,00	0,0000	0,0000
XBG	1,00	0,5714	0,8571
LGBM	0,00	0,0000	0,0000
LogREG	0,00	0,2857	0,0000
MLP	0,00	0,1429	0,1429
Non-Resampled Data			
RF	0,00	0,00	0,00
XBG	1,00	1,00	1,00
LGBM	0,00	0,00	0,00
LogREG	0,00	0,00	0,00
MLP	0,00	0,00	0,00

Table C8 – Performance Metrics for the Models tried, for Resampled and Non-Resampled data in Fold 1

Fold 1					
Algorithm	**Accuracy**	**F1 Macro**	**F1 Weighted**	**Precision**	**Recall**
Non – Resampled					
Logistic Regression	0,7808	0,3889	0,7303	0,7436	0,7808
Random Forest	0,7914	0,3605	0,7393	0,7781	0,7914
XGBoost	0,8018	0,4622	0,7623	0,7729	0,8081
MLP Classifier	0,7936	0,4304	0,7495	0,7617	0,7936
Light GBM	0,8020	0,4458	0,7609	0,7745	0,8020
Resampled					
Logistic Regression	0.6663	0.4016	0.6809	0.7582	0.6663
Random Forest	0.7370	0.4441	0.7343	0.7505	0.7370
XGBoost	0,7675	0,4973	0,7589	0,7632	0,7675
MLP Classifier	0,7188	0,4446	0,7265	0,7525	0,7188
Light GBM	0,7700	0,4868	0,7603	0,7642	0,7700

Table C9 – Performance Metrics for the Models tried, for Resampled and Non-Resampled data in Fold 2

Fold 2					
Algorithm	**Accuracy**	**F1 Macro**	**F1 Weighted**	**Precision**	**Recall**
Non-Resampled					
Logistic Regression	0.7814	0.3914	0.7317	0.7424	0.7814
Random Forest	0.7910	0.3584	0.7385	0.7761	9.7910
XGBoost	0.8018	0.4641	0.7621	0.7723	0.8018
MLP Classifier	0.7936	0.4277	0.7493	0.7595	0.7936
Light GBM	0,8000	0,4496	0,7564	0,7743	0,8000
Resampled					
Logistic Regression	0,6858	0,4393	0,7032	0,7523	0,6858
Random Forest	0,7487	0,4259	0,7445	0,7526	0,7487
XGBoost	0,7701	0,4913	0,7623	0,7633	0,7701
MLP Classifier	0,7263	0,4624	0,7340	0,7522	0,7263
Light GBM	0,7680	0,4815	0,7593	0,7607	0,7680

Table C10 – Performance Metrics for the Models tried, for Resampled and Non-Resampled data in Fold 3

Fold 3					
Algorithm	**Accuracy**	**F1 Macro**	**F1 Weighted**	**Precision**	**Recall**
Non-Resampled					
Logistic Regression	0.7807	0.3854	0.7308	0.7466	0.7807
Random Forest	0.7917	0.3628	0.7396	0.7783	0.7917
XGBoost	0.7998	0.4623	0.7613	0.7689	0.7998
MLP Classifier	0.7925	0.4161	0.7476	0.7595	0.7925
Light GBM	0,8017	0,4485	0,7611	0,7751	0,8017

			Resampled		
Logistic Regression	0,6531	0,4087	0,6664	0,7588	0,6531
Random Forest	0,7325	0,4436	0,7333	0,7529	0,7325
XGBoost	0,7663	0,4929	0,7565	0,7576	0,7663
MLP Classifier	0,7117	0,4509	0,7217	0,7506	0,7117
Light GBM	0,7742	0,4774	0,7608	0,7609	0,7742

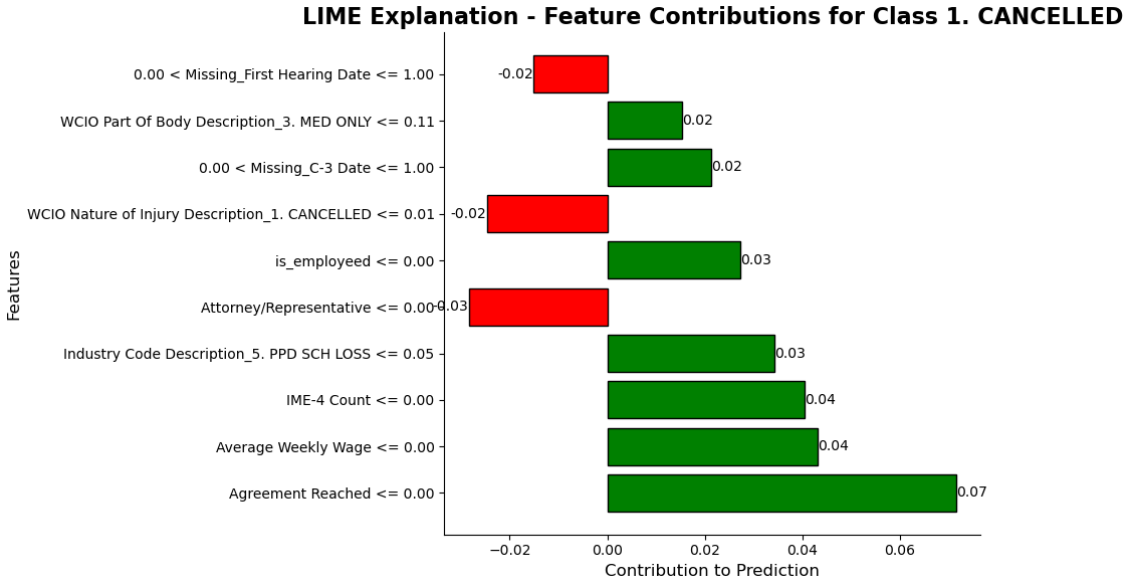

Figure C1 – LIME Explanation for Fold 3: 1. Cancelled

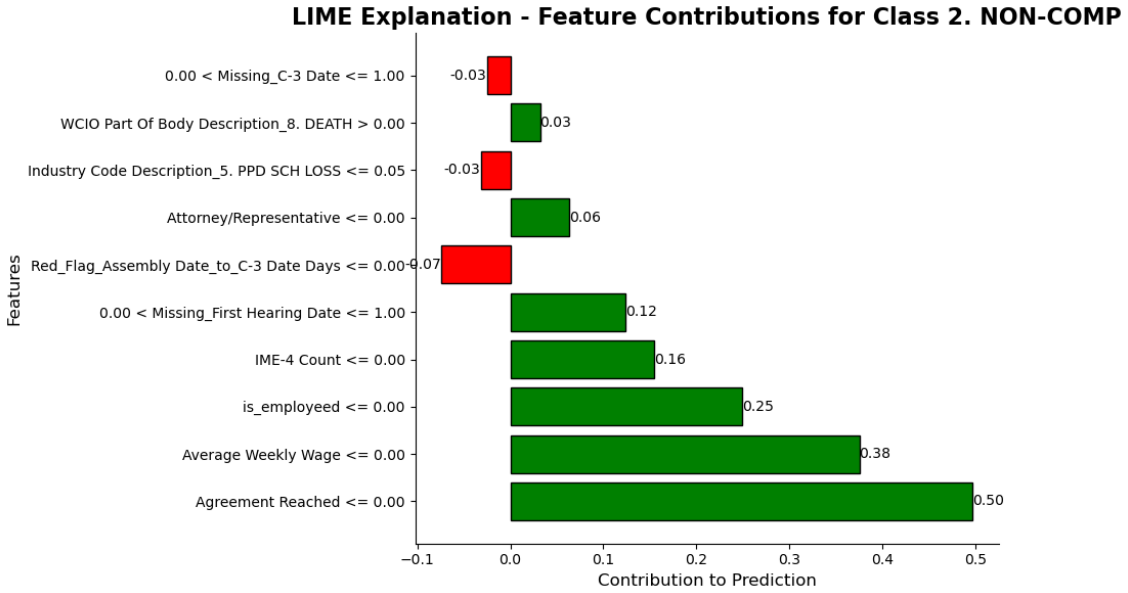

Figure C2 – LIME Explanation for Fold 3: 2. Non-Comp

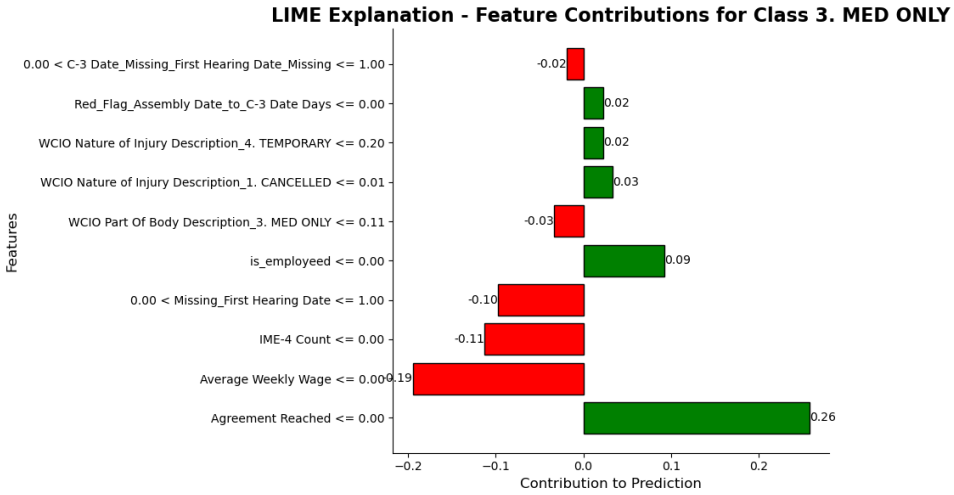

Figure C3 – LIME Explanation for Fold 3: 3. Med Only

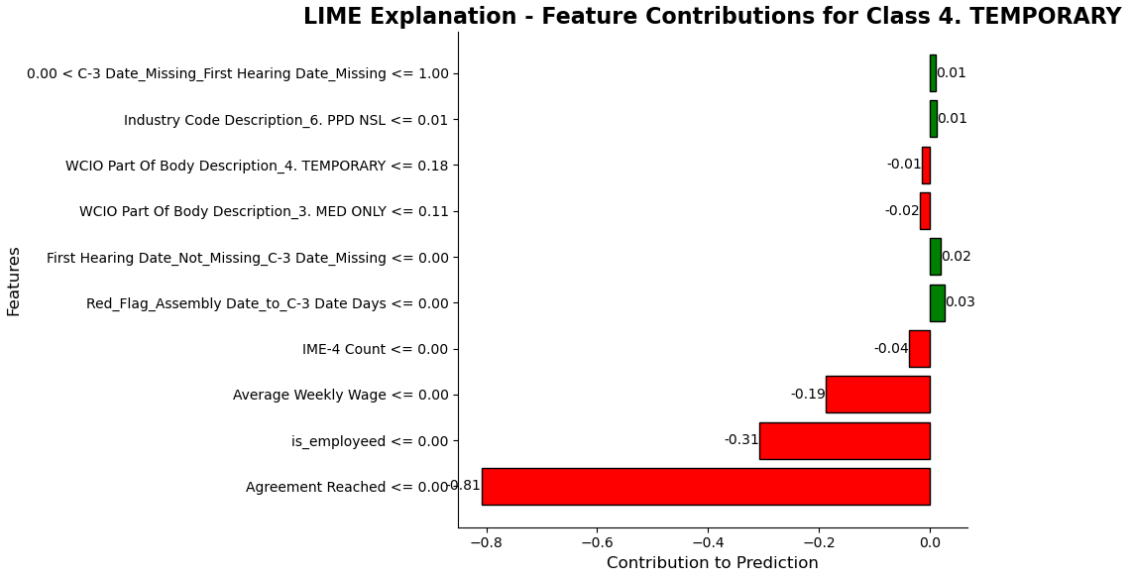

Figure C4– LIME Explanation for Fold 3: 4. Temporary

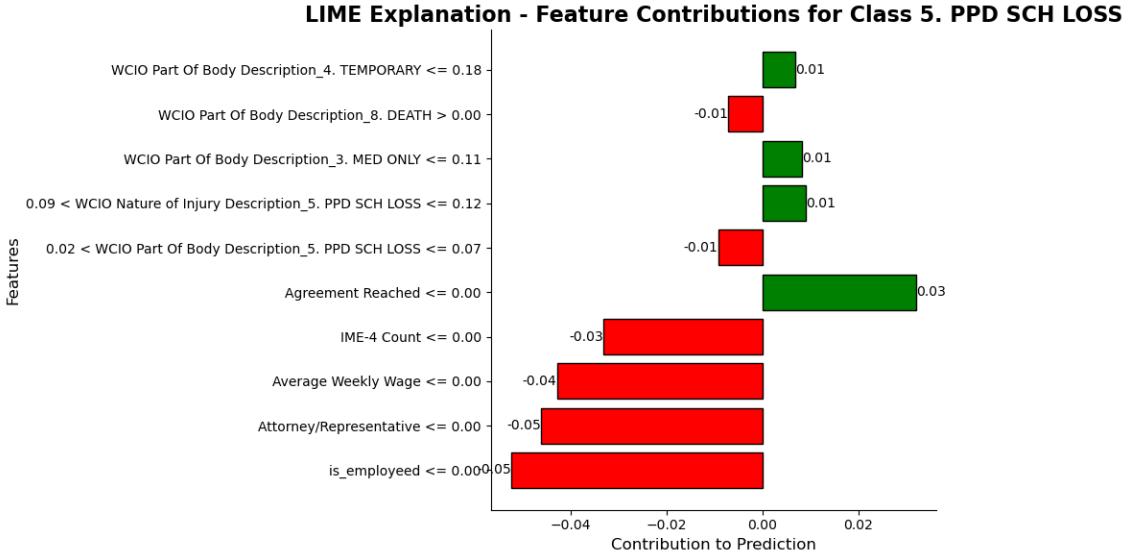

Figure C5 – LIME Explanation for Fold 3: 5. PPD SCH LOSS

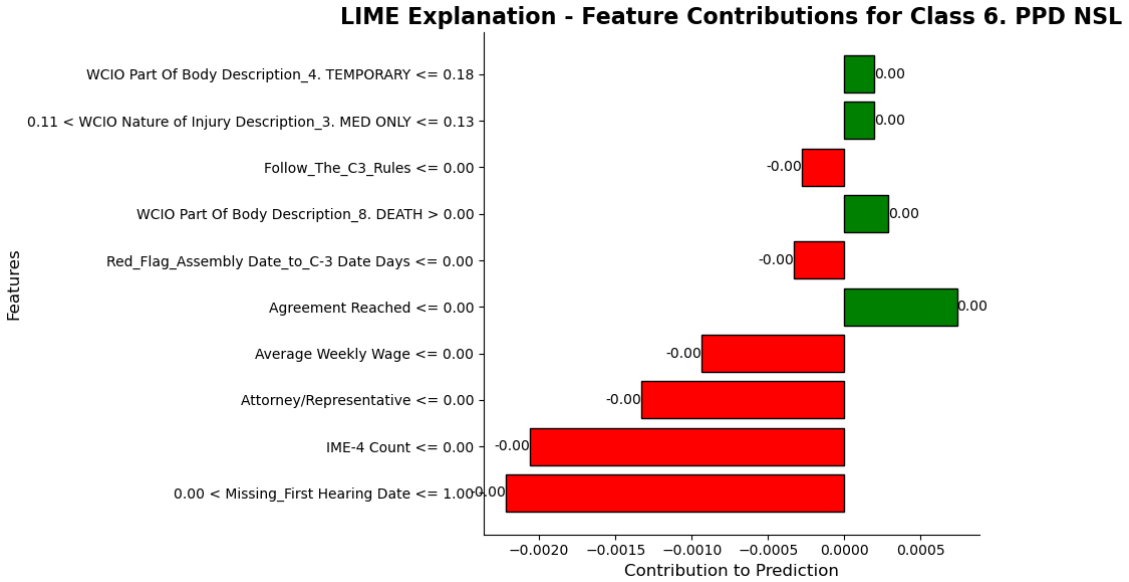

Figure C6 – LIME Explanation for Fold 3: 6. PPD NSL

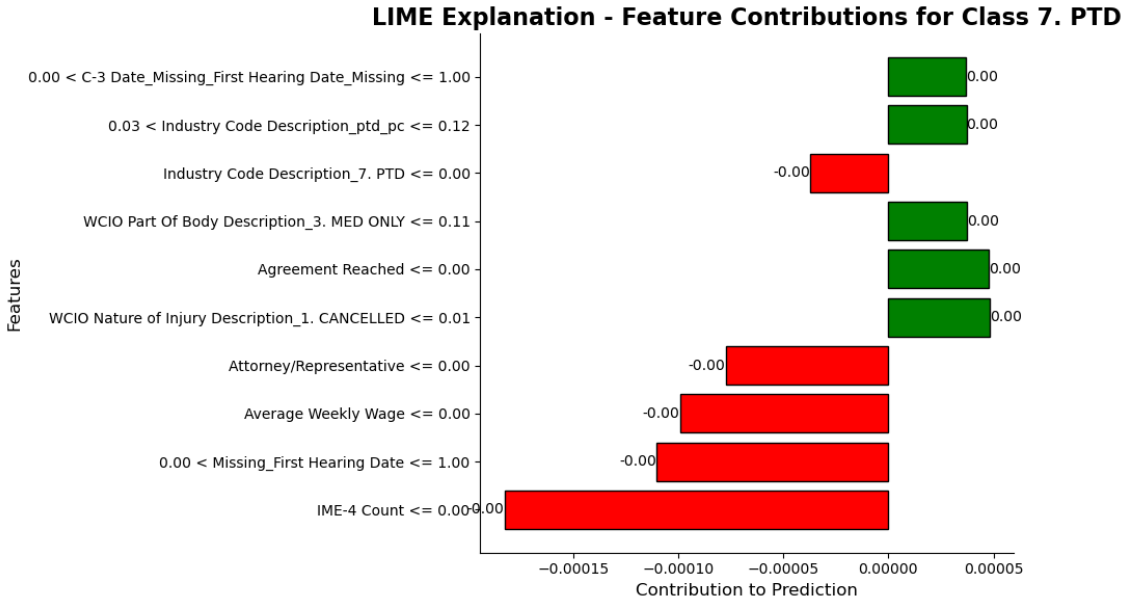

Figure C7 – LIME Explanation for Fold 3: 7. PTD

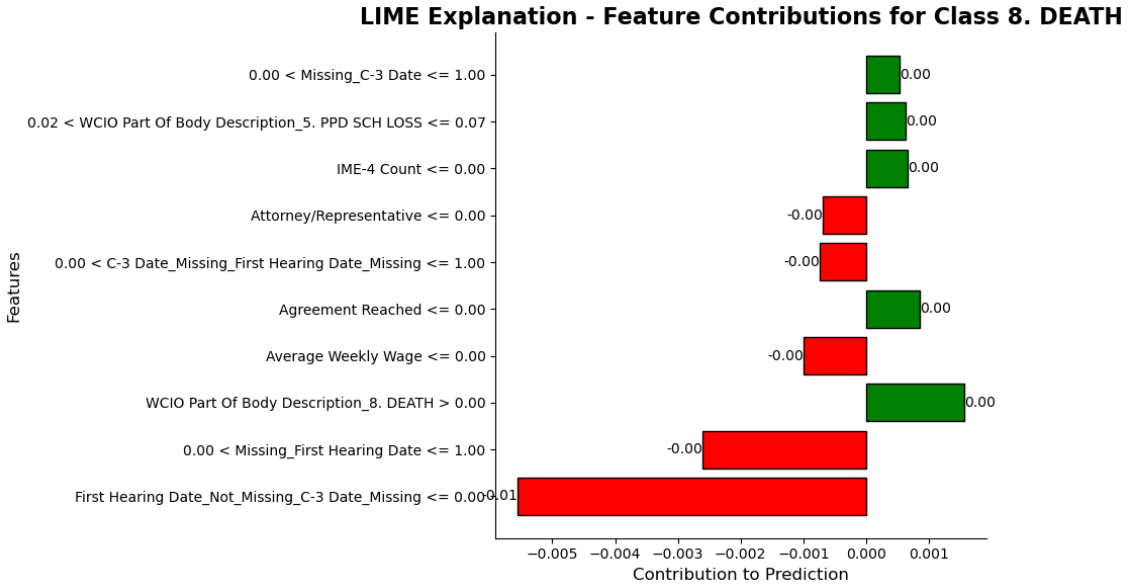

Figure C8 – LIME Explanation for Fold 3: 8. Death