Risk Prediction and Assessment in the Construction Industry

Zhixing Wang Department of Civil and Environmental Engineering University of Illinois Urbana–Champaign Urbana, IL, USA
zw88@illinois.edu

Zain Sitabkhan Department of Civil and Environmental Engineering University of Illinois Urbana–Champaign Urbana, IL, USA
zsita@illinois.edu

Deago Sirenden Department of Civil and Environmental Engineering University of Illinois Urbana–Champaign Urbana, IL, USA
deagofs2@illinois.edu

Zhihui Da Department of Civil and Environmental Engineering University of Illinois Urbana–Champaign Urbana, IL, USA
zhihuid2@illinois.edu

Abstract

This project focuses on risk prediction and assessment in the construction industry using incident and accident data from New York City. By applying regression-based models, the objective is to predict fatality and injury outcomes, as well as generate a weighted index to evaluate the severity of such events. The study contributes to understanding which attributes most strongly influence construction-related incidents and provides insights that may improve safety measures in the industry.

keywords:“Construction Safety”, “Risk Prediction”, “Accident Reports”, “Regression Analysis”

Contents

Risk Prediction and Assessment in the Construction Industry 1
Abstract 1
1. Introduction 4
1.1 Background & Motivation 4
1.2 Objectives 4
1.3 Overview of Analytical Plan 4
2. Exploratory Data Analysis 5
2.1. Basic Information 5
2.2 Attributes 5
2.3 Correlation Mapping 6
2.3.1 Weighted HVI 6
2.3.2 Global Correlation 7
2.3.3 Log-scaled Correlation 7
2.4 Preprocessing 8
2.4.1 Data Integration and Cohort 8
2.4.2 Borough × Month Aggregation 8
2.5 Results for Preprocessing 8
2.5.1 Averaging the Data 12
2.6 Discussion 13
3.Methodology 13
3.1 Preliminary Predictive Modeling & Model Limitation 13
3.1.1 Result figures and Explaination 13
3.1.2 Poisson Model (Injury) 13
3.1.3 Negative Binomial Model (Fatality) 14
3.1.4 Logistic Model[10] 15
3.1.5 Visualization 15
3.1.6 Disccusion the Limitation of Data for Preliminary Regression Model 15
3.2 K-Means Classification Models Methodology 17
3.2.1 Data Preparation and Cleaning 17
3.2.2 Model Architecture 17
3.2.3 Weighted K-Means Function 17
3.2.4 Model Evaluation 18
3.3 Decision Tree Classification Models Methodology 18
3.3.1 Data Preparation 18
3.3.2 Model Architecture 18
3.3.3 Classification & Grouping of Data 18
3.3.4 Model Evaluation 19
3.4 Neural Network Classification Models Methodology 19
3.4.1 Data Preparation 19
3.4.2 Model Architecture 19
3.4.3 Loss Function and Optimization 19
3.4.4 Training and Validation 19
3.4.5 Model Evaluation 20
3.5 Neural Network Regression Models Methodology 20
3.5.1 Data Preparation and Cleaning 20
3.5.2 Feature Standardization 20
3.5.3 Model Architecture 20

3.5.4 Loss Function and Optimizer 21
3.5.5 Training and Validation 21
3.5.6 Model Improvements 21
4 Results and Discussion 22
4.1 Introduction 22
4.2 K-Means Classification Models 22
4.2.1 Models 22
4.2.2 Discussion 23
4.3 Decision Tree Classification Models 24
4.3.1 Models 24
4.3.2 Discussion 25
4.4 Neural Network Classification Models 26
4.4.1 Models 26
4.4.2 Discussion 30
4.5 Neural Network Regression Models 32
4.5.1 Models 32
4.5.2 Discussion 34
5. References 36

1. Introduction

1.1 Background & Motivation

Construction safety remains a critical and persistent challenge within the civil engineering and infrastructure development sectors, particularly in dense urban environments such as New York City. Despite advancements in regulation and monitoring, construction continues to account for a disproportionately high share of workplace fatalities. According to the NYCOSH Deadly Skyline report, construction incidents represent more than 20% of statewide workplace deaths, many of which are associated with preventable safety failures and oversight deficiencies [2]. Recent analyses of post-pandemic conditions indicate that construction activity in New York City has accelerated, increasing labor exposure and highlighting the urgency of strengthening safety intervention strategies [3].

Researchers have increasingly explored the use of analytical and machine-learning-based frameworks to improve the prediction and prevention of construction accidents. Prior studies demonstrate the potential of data-driven models to identify complex patterns in injury outcomes and enhance proactive safety decision-making, outperforming traditional retrospective analyses [1]. However, applying predictive modeling to real-world safety datasets presents substantial challenges due to issues such as extreme sparsity, zero-inflation, and uneven geographic distribution of incident data. Zero-inflated Poisson modeling, for example, has been shown to struggle under such imbalance, producing unstable parameter estimates and reduced inference reliability [4]. Machine learning offers an alternative pathway, with recent work highlighting advantages of algorithmic prediction and feature-driven risk assessment for improving site-level safety planning [5].

New York City offers a valuable context for evaluating predictive safety models due to the concentration of high-density development across its five boroughs and the strong geographic variability in project type, scale, and regulatory enforcement. Incidents are heavily clustered in regions such as Manhattan and the Bronx, while boroughs like Staten Island contribute minimal event frequency, intensifying statistical imbalance within the dataset. These characteristics motivate the development and comparison of complementary modeling approaches to determine their feasibility for real-world safety risk prediction.

1.2 Objectives

This study investigates construction incident and accident data from New York City to identify the factors associated with injury and fatality outcomes and evaluate the predictive capability of statistical and machine-learning-based models. The objectives are to:

- Examine spatial and temporal distributions of incidents across boroughs,
- Analyze the relationships among incident characteristics, contextual variables, and injury outcomes,
- Assess the feasibility of predictive models to support proactive construction safety strategies.

1.3 Overview of Analytical Plan

The study begins with exploratory data analysis (EDA) to characterize the structure of the dataset and identify preliminary patterns in incident distribution and variable relationships. Conventional regression models, including Poisson, Negative Binomial, and logistic regression, are first evaluated

to determine their suitability for predicting injury and fatality outcomes. These models provide a baseline for comparison and help identify the limitations arising from sparse and imbalanced data.

Following the regression analysis, the study implements machine learning approaches such as k-means clustering, decision tree classification, and neural network models to improve predictive performance and capture nonlinear relationships. The results from regression and machine learning models are synthesized to determine the most effective pathway for predictive incident modeling and to provide recommendations for future research and construction safety improvement.

2. Exploratory Data Analysis

This section presents a descriptive examination of the dataset and outlines the initial analytical procedures conducted to understand incident characteristics, spatial and temporal distribution patterns, and relationships between key variables. The exploratory findings provide the foundation for determining appropriate modeling strategies and evaluating the feasibility of predictive approaches.

2.1. Basic Information

The dataset consists of construction-related incidents and accidents at New York City in each of the five boroughs. It provides a large-scale CSV file suitable for predictive analysis.[6] The dataset includes approximately 958 rows, each representing an accident or incident record, and 20 columns containing attribute fields of these records.

2.2 Attributes

The dataset includes twenty primary attributes describing incident characteristics, spatial features, and safety outcomes. These variables encompass unique identifiers (Record ID), regulatory and administrative descriptors (BIN, block, lot), event classifications (Record Type, Check2), and severity indicators (Injury, Fatality). Spatial metadata such as Borough and Council District enables geographic aggregation and comparative analysis. This structure supports the development of both statistical and machine-learning-based models that assess relationships among incident attributes and risk outcomes.

Table 1. Attribute Definitions and Descriptions

Attribute Name	Unit/Type	Description
BIN	Integer	Building Identification Number (unique ID for each building)
Accident Report ID	Integer	Unique identifier of each accident report
Incident Date	Date	Date of the incident or accident
Record Type Description	Category (Text)	Record type, distinguishing Incident from Accident
Check2 Description	Category (Text)	Detailed category of the incident, e.g., Construction Related, Mechanical Equipment, Worker Fall

Fatality	Integer	Number of fatalities
Injury	Integer	Number of injuries
House Number	Text/Number	House number of the incident location
Street Name	Text	Street name of the incident location
Borough	Category	Administrative borough (e.g., Manhattan, Bronx, Brooklyn)
Block	Integer	Geographic block number
Lot	Integer	Lot number within the block
Postcode	Integer	Postal code of the location
Latitude	Float	Latitude coordinate of the incident location
Longitude	Float	Longitude coordinate of the incident location
Community Board	Integer	Community board identifier
Council District	Integer	City council district identifier
BBL	Integer	Borough-Block-Lot unique cadastral identifier
Census Tract (2020)	Integer	Census tract number from the 2020 census
Neighborhood Tabulation Area (NTA) (2020)	Text	Neighborhood Tabulation Area (NTA) code from 2020

2.3 Correlation Mapping

To enhance predictive capability and incorporate contextual influences, external variables were merged with the dataset. The Heat Vulnerability Index (HVI) and monthly climate variables such as average temperature and precipitation were integrated to examine potential environmental links to construction safety outcomes. A weighted averaging procedure was applied to HVI data to address uneven distribution across boroughs.

Table 3. Integrated Dataset with Climate and HVI Variables

Borough	Postcode	YearMonth	IncidentCount	Fatality	Injury	AvgTemp	AvgPrecip	HVI
Bronx	10451	Jun-24	3	1	2	71.7	4.4	5

Preliminary inspection indicates that higher-HVI areas (typically in the Bronx and parts of Brooklyn) correspond to marginally elevated injury counts, hinting at interactions between heat exposure and worker safety.

Some other parameter will be added such as Noncomplaint Count and IssueNumber (in section 6) in order to solve the regression model problems.

2.3.1 Weighted HVI

Weighted averaging is used when different observations contribute unequally to an aggregate measure. In another word it will directly contain the information about the borough.

2.3.2 Global Correlation

A global correlation analysis was conducted among key variables: TotalIncidents, Fatality, Injury, AvgTemp, AvgPrecip, and HVI.

Table 5. Correlation Matrix of Incident, Climate, and Vulnerability Variables

	TotalIncidents	Fatality	Injury	AvgTemp	AvgPrecip
TotalIncidents	1.000	0.120	0.958	0.025	0.023
Fatality	0.120	1.000	0.075	−0.007	−0.153

A global correlation analysis was conducted to examine linear relationships among the primary variables in the dataset. The results, summarized in Table 1, indicate several notable patterns. The strongest association occurred between TotalIncidents and Injury, reflected by a Pearson correlation coefficient of approximately 0.958. This relationship is expected, since the majority of reported incidents involved at least one injury. In contrast, Fatality exhibited very limited correlation with other variables. The correlation with TotalIncidents was low at approximately 0.120, and the relationship with Injury was even weaker at approximately 0.075, which aligns with the sparse distribution of fatal events.

Environmental variables exhibit similarly weak relationships with safety outcomes. Average temperature (r=-0.57) show negligible correlation with fatality counts. Analysis of the Heat Vulnerability Index (HVI) produces a moderate negative correlation with both TotalIncidents and Injury, with values ranging approximately from r = −0.57 to r=-0.606. This pattern suggests that areas with higher vulnerability scores may be associated with fewer reported incidents within this dataset. The counterintuitive relationship indicates the possibility of confounding factors and highlights the need for more comprehensive analysis in future work.

2.3.3 Log-scaled Correlation

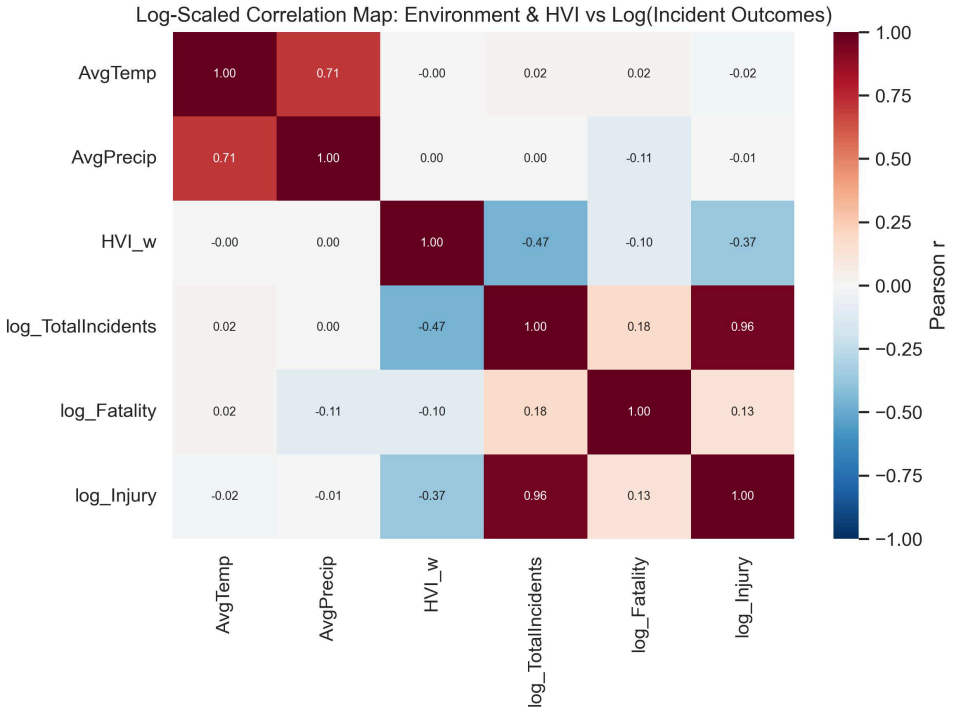

Figure 1: Correlation heatmap after log scaling

Results show a strong positive correlation between TotalIncidents and Injury ($r \approx 0.96$) and a negative correlation between HVI and Fatality ($r \approx -0.57$). Although counterintuitive at first glance, this may reflect underreporting or mitigation interventions in high-vulnerability areas. These relationships were visualized using a log-scaled correlation heatmap, emphasizing nonlinear dependencies that justify the use of both Poisson and Negative Binomial regression models in the next section.

2.4 Preprocessing

2.4.1 Data Integration and Cohort

The initial dataset of construction-related incidents was filtered to remove incomplete and inconsistent records to establish the analytic cohort. Group-by operations were then performed to extract and aggregate key attributes. Aggregation was conducted based on borough (Area), month, and postcode. The postcode variable functions as a critical spatial key because it enables direct linkage to external datasets, including the Heat Vulnerability Index (HVI) [7]. Integrating the HVI dataset provides an additional environmental and demographic dimension that enhances contextual interpretation of incident locations.

2.4.2 Borough × Month Aggregation

To examine temporal trends in incident activity, records were aggregated at the borough–month level. The results indicate distinct seasonal patterns, with incident frequency increasing during spring and summer periods, which generally align with intensified construction activity and workforce presence. This seasonal clustering provides an initial indication of time-dependent variation in construction risk that is relevant for subsequent modeling stages.

Table 2 presents an excerpt of the monthly aggregation output for selected borough–postcode combinations, including incident count, injury count, and fatality count. The full dataset is available in the file monthly_borough.csv.

Table 2. Monthly Aggregation of Incidents by Borough and Postcode

Borough	Postcode	YearMonth	IncidentCount	Fatality	Injury
Bronx	10451	Feb-24	1	0	1
Bronx	10451	Mar-24	1	0	1
Bronx	10451	Apr-24	1	0	1
Bronx	10451	Jun-24	3	1	2

Excerpt shown above; full panel saved as `monthly_borough.csv`.

2.5 Results for Preprocessing

The preliminary visualizations summarize the distribution of injuries and fatalities across boroughs and districts, as well as monthly and cumulative trends from January 2024 through October 2025. These results establish a foundational understanding of spatial and temporal variation in incident patterns and support further analysis through borough-specific and district-specific modeling approaches.

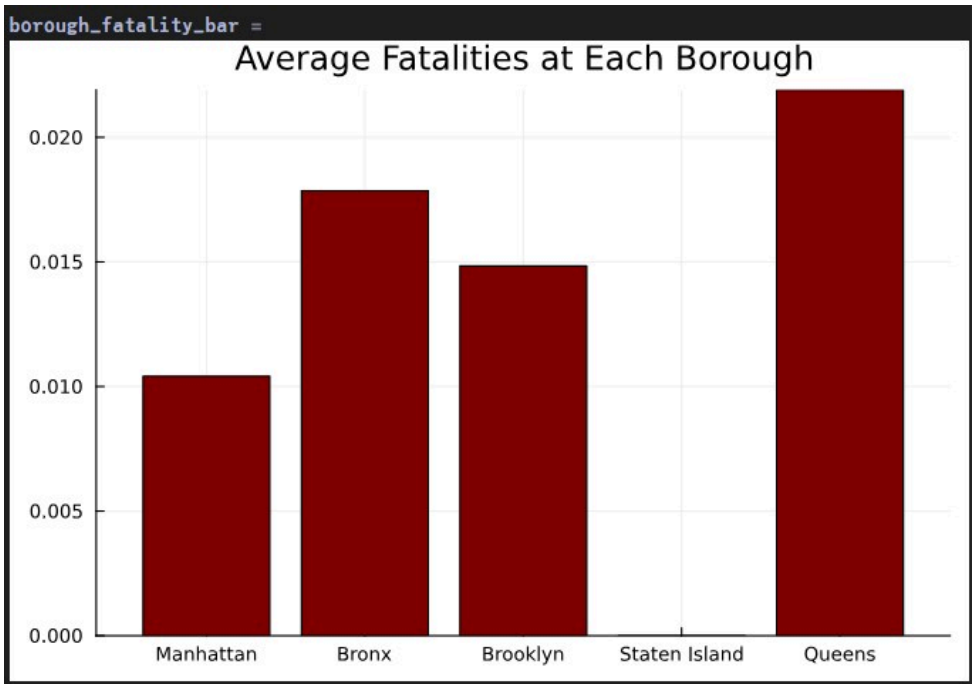

Figure 2: Average Fatalities at Each Borough

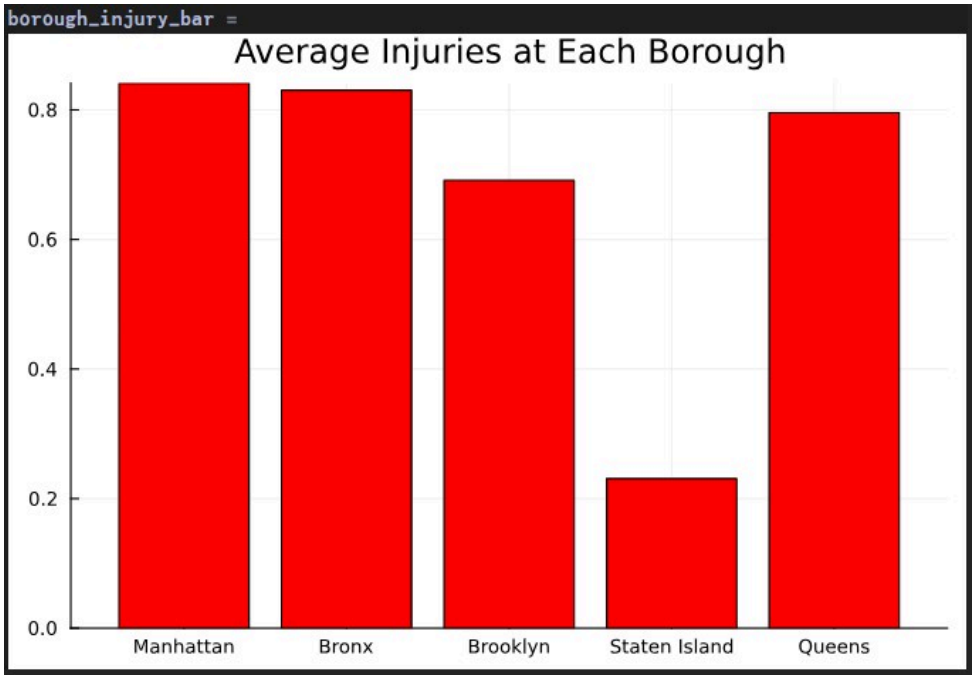

Figure 3: Average Injuries at Each Borough

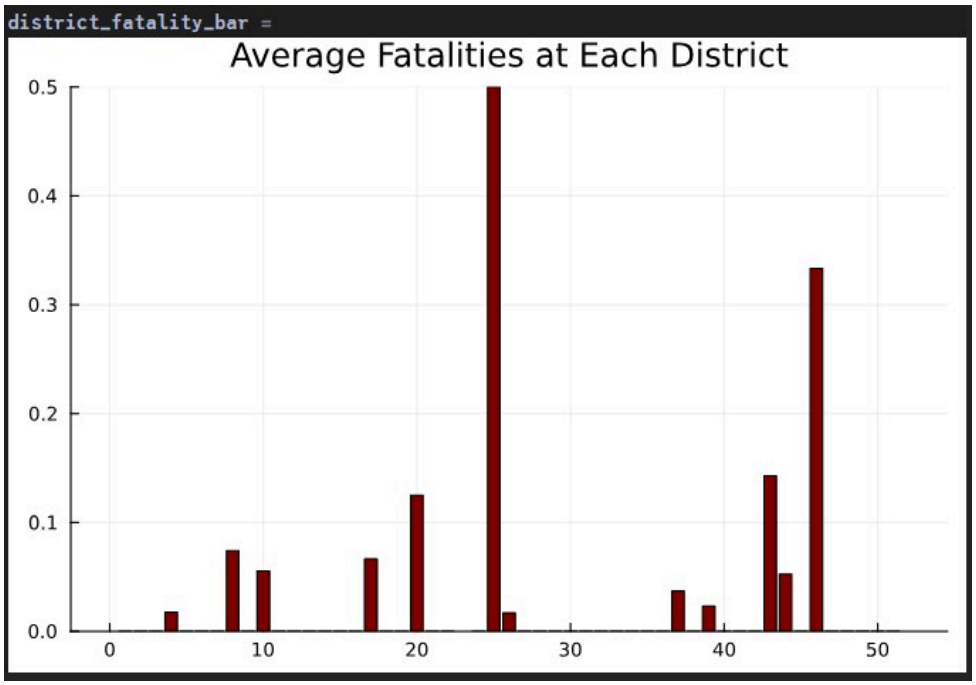

Figure 4: Average Fatalities at Each District

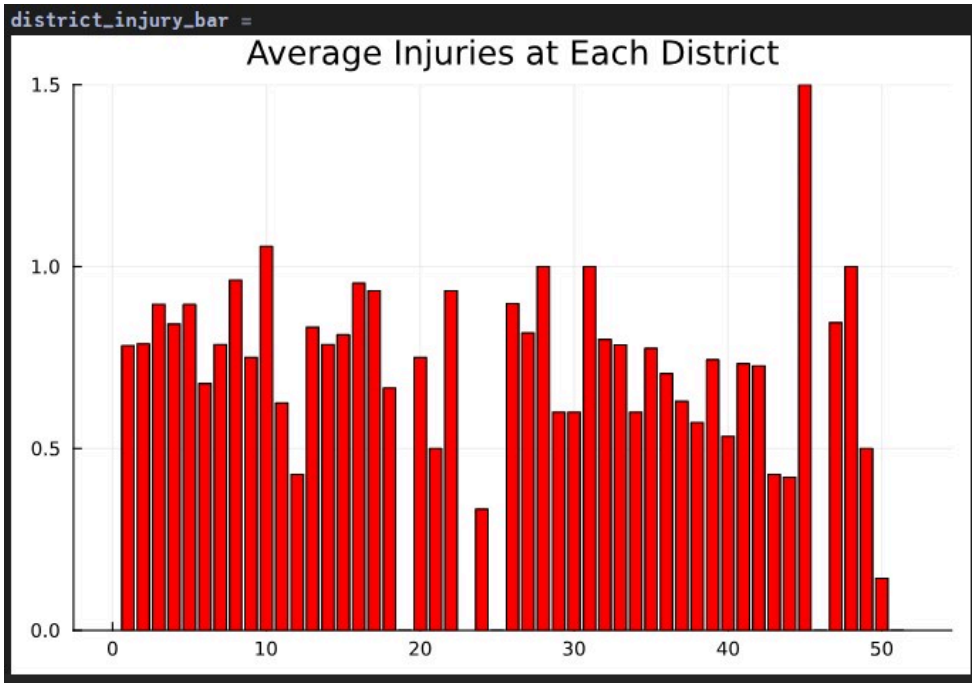

Figure 5: Average Injuries at Each District

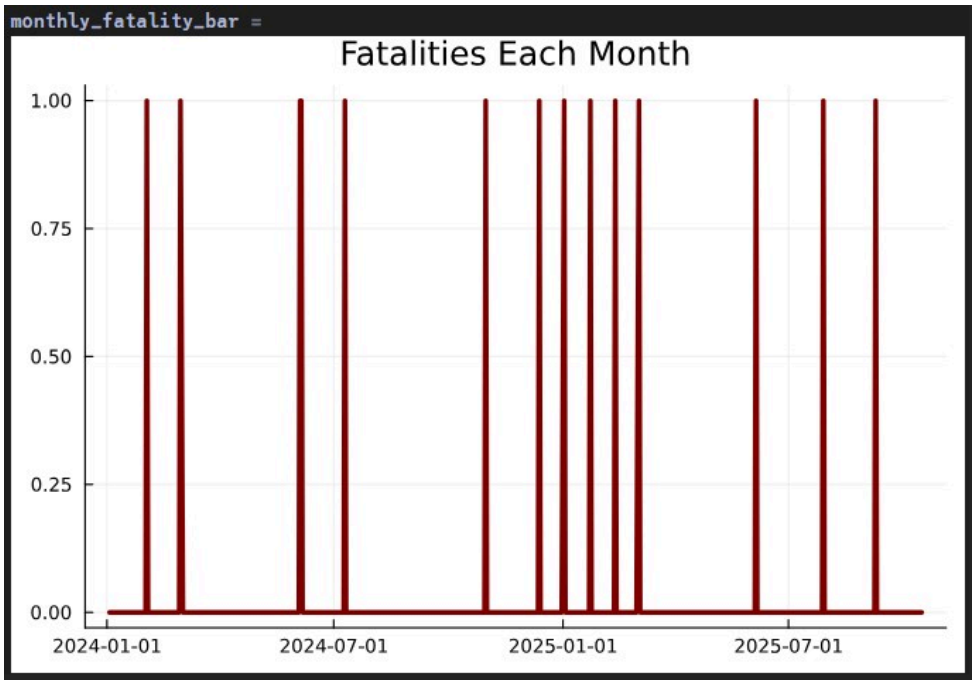

Figure 6: Cumulative Fatalities Overtime

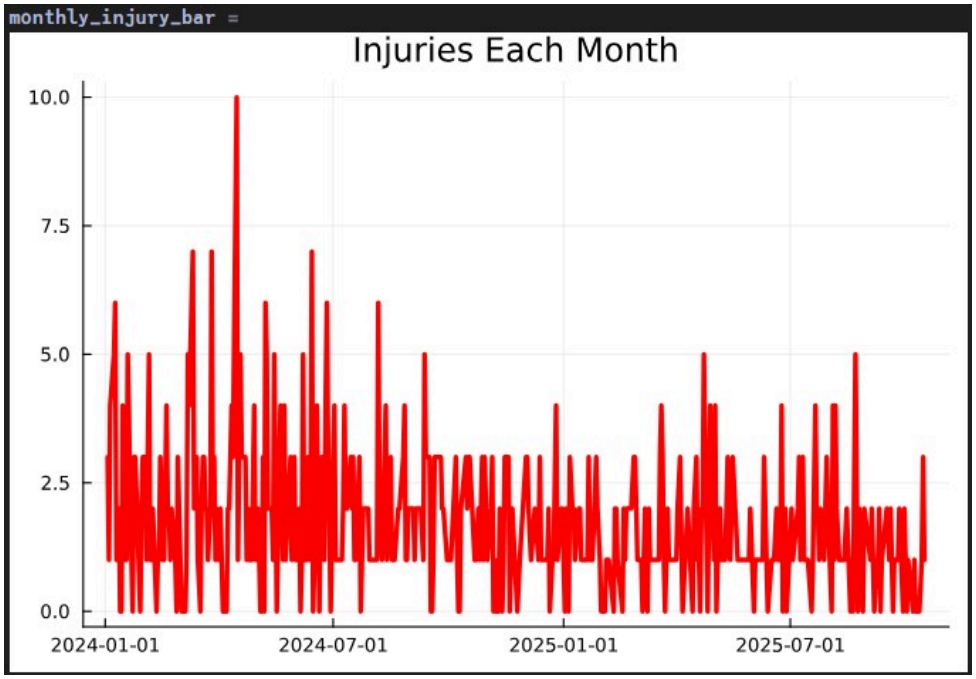

Figure 7: Cumulative Injuries Overtime

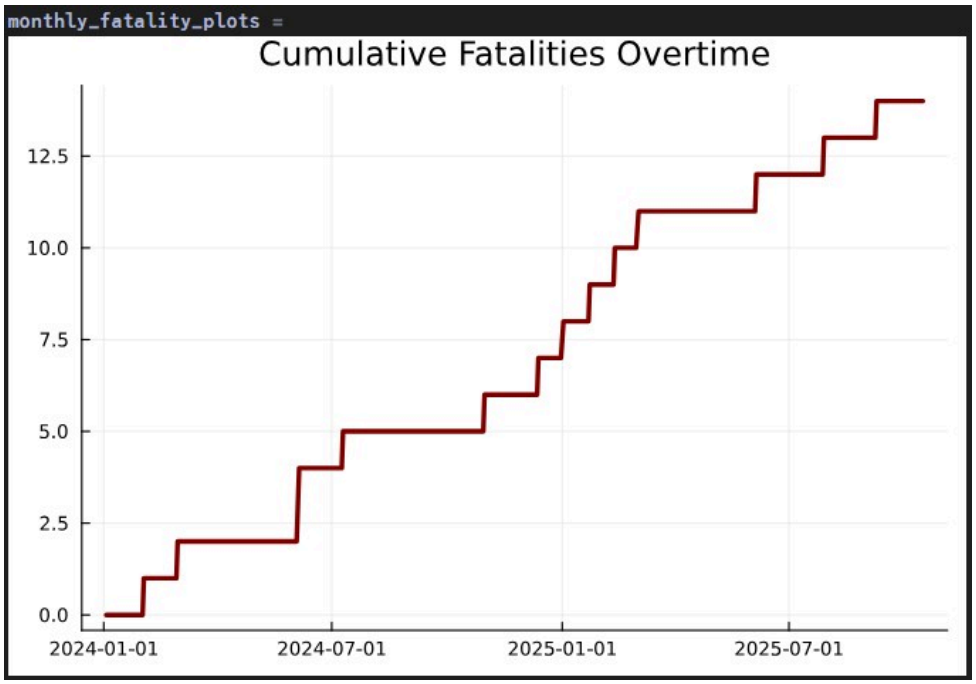

Figure 8: Fatalities Each Month

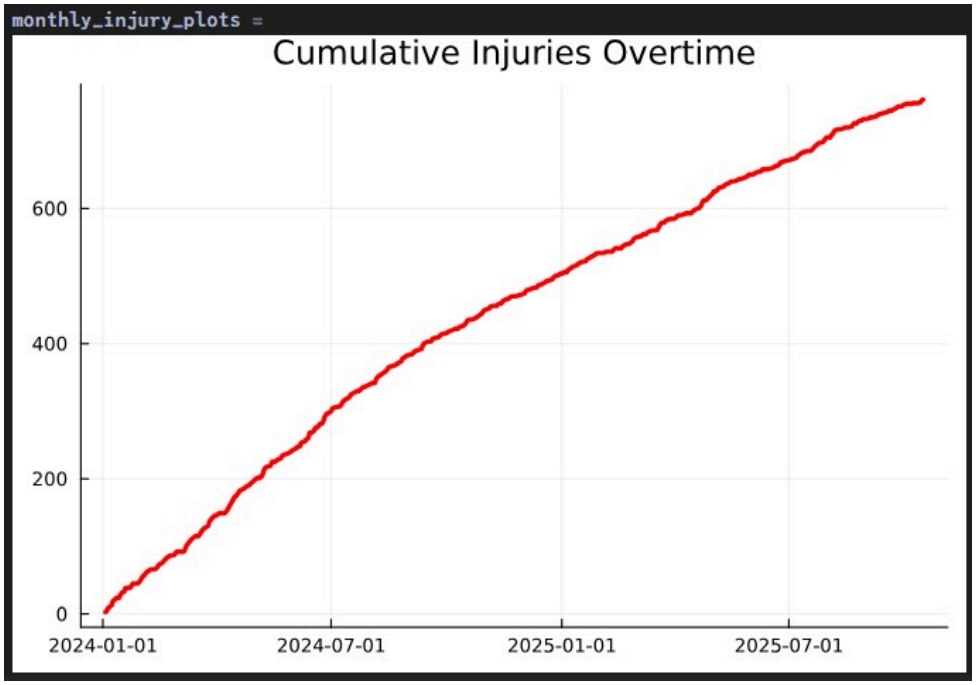

Figure 9: Fatalities Each Month

2.5.1 Averaging the Data

Table 4. Average Incident and Injury Rates by Borough

Borough	AvgFatality	AvgInjury	AvgIncident	FatalityRate%	InjuryRate%
Bronx	0.019	1.10	1.31	1.47	83.82

Table 4 summarizes borough-level averages for fatalities, injuries, and total incident counts. On average, approximately 83.8 percent of recorded incidents resulted in at least one reported injury, while fatal events represented only about 1.5 percent of total cases. The Bronx reported the highest

average injury rate, followed closely by Brooklyn, indicating a spatial concentration of elevated incident severity. These results suggest that incident frequency alone may not fully capture risk and that borough-level proportional measures provide a more informative perspective on relative safety outcomes.

2.6 Discussion

The exploratory analysis highlights several important characteristics of the dataset that inform subsequent modeling. Borough-level aggregation revealed substantial disparities in injury prevalence, with the Bronx and Brooklyn experiencing the highest proportional injury rates relative to total incident counts. Temporal analysis demonstrated clustering during warmer months, suggesting that seasonal workforce expansion and construction volume may influence incident trends.

Descriptive statistics further emphasized that injuries constitute the dominant outcome within the dataset, whereas fatalities occur infrequently and exhibit limited variability. This imbalance introduces challenges for predictive modeling based on traditional regression approaches, given the small number of positive fatality observations relative to the dataset size. The preliminary findings therefore underscore the importance of focusing on injury prediction and spatial-temporal risk differentiation rather than fatality forecasting.

The results also indicate that proportional measures and aggregated perspectives reveal patterns not visible through raw incident counts alone. These insights establish foundational context for the methodological approaches introduced in Section 3 and guide feature selection priorities for classification-based modeling frameworks.

3.Methodology

3.1 Preliminary Predictive Modeling & Model Limitation

3.1.1 Result figures and Explaination

Initial model development applied several traditional statistical prediction approaches to assess the suitability of the dataset and to diagnose structural limitations for downstream modeling. These preliminary models provide insight into sparsity, imbalance, and the distributional characteristics of the injury and fatality outcomes.

3.1.2 Poisson Model (Injury)

Table 6. Poisson Regression Model Results for Injury Counts

Variable	coef	std err	z	P>|z|	0.025	0.975
Intercept	0.1547	1.059	0.146	0.884	−1.921	2.230

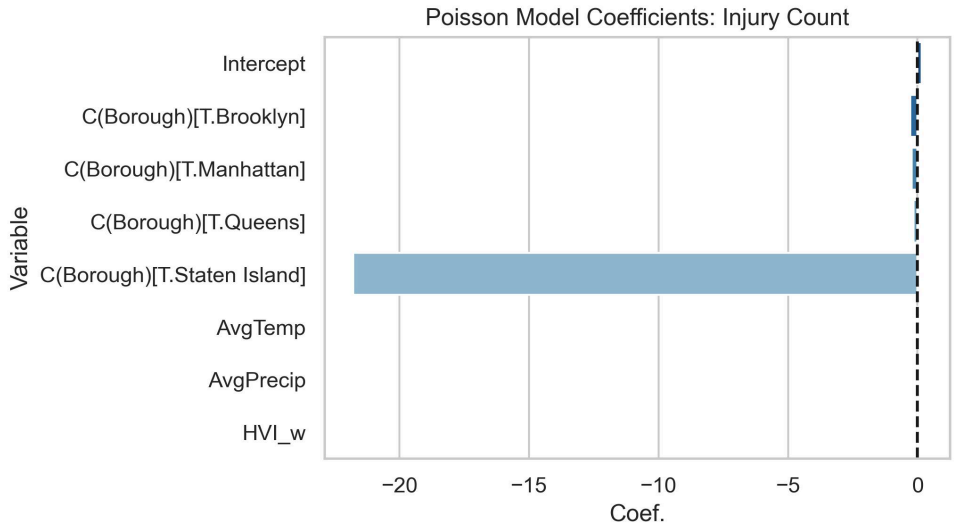

Figure 10: Poisson injury model coefficients

The Poisson model was selected due to the non-negative count nature of the injury variable and its suitability for modeling frequency-based outcomes. However, the regression output indicates weak statistical significance for most predictors, including borough dummy variables, meteorological features, and the Heat Vulnerability Index (HVI). Coefficient patterns also demonstrate instability, particularly for Staten Island, which exhibits wide confidence intervals and irregular magnitudes. These issues are consistent with the sparse distribution of non-zero values and the heavy concentration of zeros across the dataset.

3.1.3 Negative Binomial Model (Fatality)

Table 7. Negative Binomial Regression Model Results for Fatalities

Variable	coef	std err	z	P>|z|	0.025	0.975
Intercept	−9.6771	10.237	−0.945	0.345	−29.742	10.387

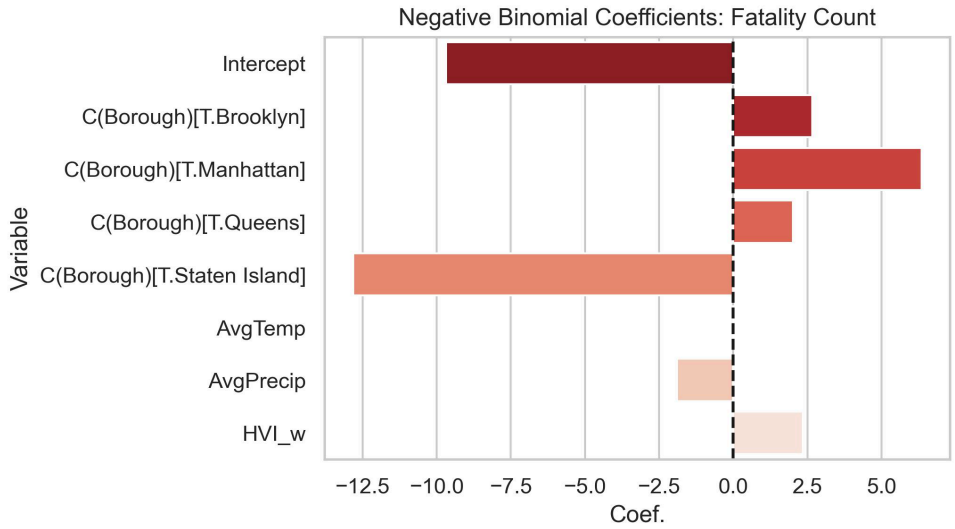

Figure 11: Negative binomial fatality model coefficients

The Negative Binomial approach was intended to address over-dispersion induced by the rarity of fatality events [8][9]. However, the resulting coefficients remain statistically insignificant, and the

intercept estimate is excessively large relative to expected outcome scales. Visual inspection of coefficient distributions further indicates minimal model learning, confirming that the available fatality data provide insufficient signal for reliable regression-based inference.

3.1.4 Logistic Model[10]

Table 8. Logistic Regression Results for Binary Fatality Events

Variable	coef	std err	z	P>|z|	0.025	0.975
Intercept	−8.1713	8.01e+06	−1e-06	1.000	−1.57e+07	1.57e+07

Fatality was modeled as a binary response due to the 0/1 nature of the variable. Logistic regression produced unstable and extreme coefficient values, particularly for borough and month indicators. The abnormal magnitude of coefficients and inflated confidence intervals reflect quasi-complete separation, a condition in which certain groups contain almost exclusively non-fatal outcomes. Under such circumstances, logistic regression attempts to push parameter values toward infinity in order to fit sparse patterns, producing unreliable predictions.

3.1.5 Visualization

Figure-based comparison of model coefficients illustrates that both the Poisson and Negative Binomial models generate coefficient estimates clustered closely around zero with limited variation. In contrast, the logistic model produces extreme values driven by sparsity and imbalance, reinforcing concerns regarding model validity.

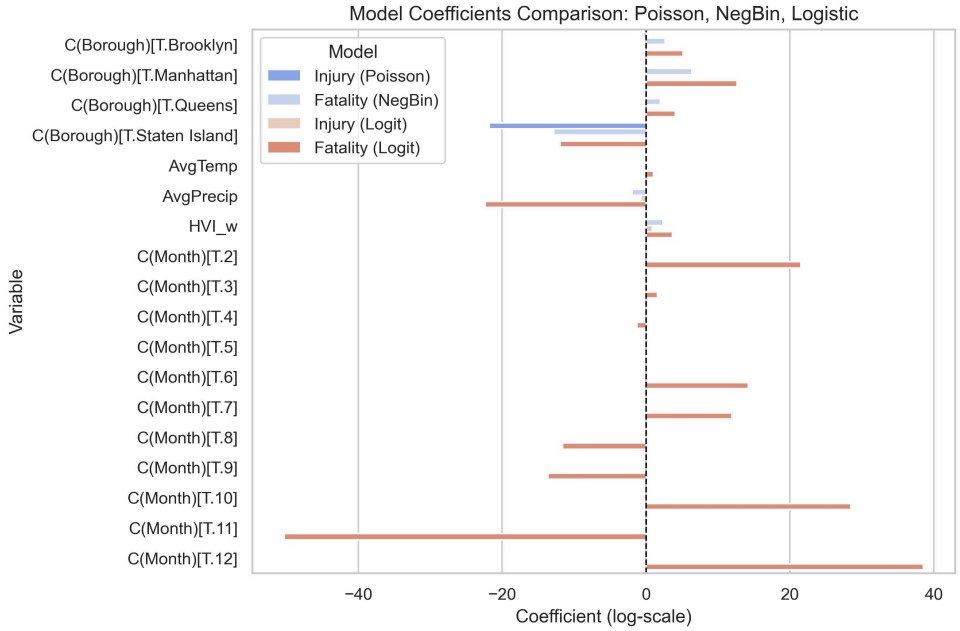

Figure 12: Coefficient comparison

3.1.6 Disccusion the Limitation of Data for Preliminary Regression Model

Poisson, Negative Binomial, and logistic regression models were selected based on the statistical characteristics of the dataset. The incident data contain a substantial number of zero values, which introduces pronounced sparsity. Although the injury variable includes non-zero counts, the fatality variable appears exclusively as 0 or 1 throughout the dataset, indicating a rare-event structure.

Under these conditions, Poisson regression is appropriate for modeling relationships between covariates and injury counts. To address over-dispersion associated with rare fatality events, the Negative Binomial model [8][9] relaxes the restrictive assumption that the variance must equal the mean. Additionally, because fatality is inherently binary, logistic regression provides a direct approach for modeling the probability of fatal events as a 0/1 response. The primary purpose of these preliminary regressions is to evaluate data sparsity and information content to guide subsequent methodological decisions and to determine a modeling direction suited to the underlying structure.

Poisson Model

Figure 10 illustrates that the significance levels of most predictors, including borough indicators, meteorological variables, and the Heat Vulnerability Index (HVI), are weak. Coefficient estimates demonstrate instability and wide confidence intervals, especially for Staten Island, which contributes very few observations. Although earlier correlation mapping indicated strong pairwise associations, these correlations rely solely on linear relationships and do not account for interdependencies among variables. For example, HVI correlates strongly with outcome variables but is also closely aligned with borough classification, suggesting that its apparent predictive strength may primarily reflect geographic clustering. Consequently, it is expected that multiple predictors fail to achieve statistical significance in Poisson and other count-based regression models. Sparse observations from Staten Island also intensify over-dispersion and reduce coefficient stability, limiting the interpretability and reliability of the model results.

Negative Binomial Model

Figure 11 demonstrates that the Negative Binomial model, although theoretically more suitable for sparse data, produces results with little inferential validity. The intercept is excessively large, and neither z-values nor p-values indicate meaningful statistical significance. Sparsity in borough-level fatality records continues to distort model behavior, preventing meaningful learning despite relaxed variance assumptions.

Three Model Comparison

Figure 12 compares coefficients across the three models. In both the Poisson and Negative Binomial models, coefficients cluster near zero. In contrast, the logistic model exhibits extreme coefficient magnitudes, particularly for borough variables with very limited fatality observations. This outcome is characteristic of quasi-complete separation, a condition in which particular borough–month combinations contain almost exclusively nonfatal outcomes. Under such conditions, logistic regression drives coefficients toward infinity in an attempt to discriminate between nearly homogeneous groups. This behavior indicates that fatality prediction functions more appropriately as a rare-event classification task rather than a count-based regression problem.

This kind of “coefficient blow-up” in the Logit model usually signals quasi-complete separation: within specific borough–month combinations, fatal events either almost never occur or never occur at all. In this casee, it is mostly the “almost never” situation. When that happens, logistic regression tends to push the associated coefficients toward $\pm\infty$ in an attempt to fit those extreme patterns. This indicates that fatality as a 0/1 outcome suffers from even stronger sparsity and imbalance than when

treated as a count, reinforcing the idea that fatal events resemble a rare-event classification problem rather than a conventional regression target.

Summary

Taken together, the sparsity and irregular structure of the data make standard regression models unsuitable unless additional, more informative predictors are introduced—such as the new parameters incorporated in Section 7.

3.2 K-Means Classification Models Methodology

3.2.1 Data Preparation and Cleaning

The K-means clustering analysis utilized four primary features from the dataset: longitude, latitude, number of injuries, and borough classification. The clustering procedure focused on four boroughs: Manhattan, Brooklyn, Queens, and the Bronx. Staten Island was excluded from the analysis because the available incident records were extremely limited, resulting in high sparsity that prevented reliable pattern identification.

To support spatial visualization and reduce coordinate noise, longitude and latitude values were rounded to three significant digits. Because injury counts frequently appeared as 0 or 1, values were aggregated by matched coordinate locations to consolidate observations representing nearby incident points. The cleaned dataset was then grouped by borough and combined into a unified file for plotting and cluster evaluation.

After preprocessing, the data were imported into Julia, and the selected variables were separated and formatted for clustering and spatial visualization. This preparation enabled borough-level comparison and supported the identification of geographically concentrated injury patterns.

3.2.2 Model Architecture

The K-means model architecture utilized four independent variables: longitude, latitude, number of injuries, and borough identification. These features represent spatial and contextual attributes necessary to form geographically meaningful clusters. City and neighborhood locations were not included directly within the dataset but were inferred based on interpolated coordinate values from longitude and latitude. This process enabled estimation of the highest concentration of incidents at the sub-borough level.

Input Selections: Longitude, latitude, injuries, and boroughs

Output Selection: Cities/counties with the highest concentration

3.2.3 Weighted K-Means Function

To enhance representation of incident density within the clustering results, a weighted K-means approach was implemented. Weighted centroids were constructed to emphasize cluster influence in areas with greater injury frequency. Prior to applying the weighted K-means algorithm, four supporting functions were developed to initialize the model and ensure iterative consistency: init_centroids, calc_distances, calc_groups, and update_centroids!. These functions collectively enabled the construction of a functional weighted clustering routine.

Rows containing missing values were removed to improve data reliability. Point locations and associated weights, defined by recorded injury counts, were then used to generate centroid positions. The resulting centroids were visualized within the clustering plot to illustrate spatial concentration patterns.

3.2.4 Model Evaluation

Model performance was evaluated through scatter map visualizations that display four cluster groups corresponding to the four boroughs included in the dataset. The spatial distribution of the clusters closely aligns with the geographic layout of New York City, demonstrating that the clustering structure is consistent with real-world urban boundaries. A supplementary visualization presents the weighted centroids as star markers, highlighting areas with the highest proportional burden of injury-related incidents.

These visual outcomes indicate that the weighted K-means method produces clusters that reflect spatial variation in incident concentration and support interpretation of localized construction safety patterns.

3.3 Decision Tree Classification Models Methodology

3.3.1 Data Preparation

The decision tree model was developed using three features from the dataset: Borough, Check Description, and Injuries. Consistent with the K-means clustering methodology, Staten Island was excluded due to insufficient sample size, which would otherwise limit interpretive reliability. The Check Description field identifies the reported cause of each construction incident. For this model, the categories included “Worker Fell,” “Mechanical Construction Equipment,” “Material Failure,” “Scaffold/Shoring Installations,” and “Excavation/Soil Work.” The category “Other Construction” was removed due to its ambiguous classification and limited analytical value.

3.3.2 Model Architecture

The model utilized two independent variables, Borough and Check Description, and one dependent variable, Injury Count. Five decision trees were produced: four representing each individual borough and one displaying the aggregated results across the four boroughs. The structure of the decision tree consists of a hierarchical arrangement that includes root nodes, internal nodes, branches, and leaf nodes. For each leaf node, both the number and percentage of injuries were calculated to show incident severity distribution.

Root Node: Borough

Internal Nodes: Check Description

Leaf Nodes: Injuries

3.3.3 Classification & Grouping of Data

The dataset was grouped by borough, and summary statistics for injury counts were calculated using Microsoft Excel. The structure and graphical representation of the decision tree models were constructed in Microsoft PowerPoint to support visual interpretability and comparative analysis across boroughs.

3.3.4 Model Evaluation

The decision tree model outputs illustrate the branching structure described in the model architecture. Nodes appear as rectangular blocks, while directional arrows indicate hierarchical decision paths. This visual format facilitates identification of incident categories associated with higher injury risk across different boroughs.

3.4 Neural Network Classification Models Methodology

3.4.1 Data Preparation

Two additional parameters were integrated to enhance model performance: **NoncompliantCount**, representing the frequency of non-compliant behaviors , [16] **IssueNumber**, representing the volume of active construction projects in a specific area and month, allowing for temporal lags[17]

Five input variables were selected from the final dataset: Average Temperature (AvgTemp), Average Precipitation (AvgPrecip), Weighted Heat Vulnerability Index (HVI_w), NoncompliantCount, and IssueNumber. The target variable, Injury, was binarized by assigning a label of 1 to observations with at least one injury and 0 otherwise. A StandardScaler was applied to standardize input features, and the dataset was divided into training (80 percent) and validation (20 percent) partitions.

3.4.2 Model Architecture

A three-layer Feedforward Neural Network (FNN) was utilized as the predictive framework. The model architecture included:

Input Layer: Corresponds to the five input features.

First Hidden Layer: 16 neurons utilizing the ReLU activation function.

Dropout Layer: Applied with a rate of 0.1 to mitigate overfitting.

Second Hidden Layer: 8 neurons utilizing the ReLU activation function.

Output Layer: A single neuron outputting unnormalized logit values, which are transformed into injury probabilities via a Sigmoid function.

This configuration provides nonlinear capacity required for representing complex interactions among multivariate inputs [12].

3.4.3 Loss Function and Optimization

Given the significant class imbalance (where “No Injury” cases far exceed “Injury” cases), the model utilizes a Weighted Binary Cross-Entropy with Logits Loss function. A positive weight, calculated as $\text{pos}_{\text{weight}} = \frac{N_{\text{neg}}}{N_{\text{pos}}}$, is automatically applied to balance the classes. The Adam optimizer[14], with a learning rate of $1 * 10^{-4}$, is employed to update network parameters and minimize the loss function during each iteration.

3.4.4 Training and Validation

The model was trained for 300 epochs. Loss and accuracy metrics for both training and validation sets were computed at each epoch. Dropout was applied during training to improve generalization. Convergence trends were monitored by logging performance metrics every 50 epochs, followed by visual inspection of loss and accuracy curves.

3.4.5 Model Evaluation

ROC Curve & AUC: The Receiver Operating Characteristic (ROC) curve was plotted[15] using validation results, and the Area Under the Curve (AUC) was calculated to quantify overall classification performance. Youden's J statistic ($\mathrm{TPR} - \mathrm{FPR}$) was utilized to determine the optimal classification threshold.

Confusion Matrix: Matrices were generated for both the default threshold (0.5) and the optimal threshold to visualize classification accuracy, false positive rates, and false negative rates.

Precision-Recall-F1 Analysis: Precision, Recall, and F1 scores were calculated across a threshold range of [0.1, 0.9] with a step size of 0.05. Curves were plotted to evaluate trade-offs under different judgment criteria, identifying the threshold that maximizes the F1 score.

3.5 Neural Network Regression Models Methodology

3.5.1 Data Preparation and Cleaning

The Neural Network Regression model was constructed to predict the count of construction-related injuries. To improve stability and robustness, several preprocessing procedures were applied, including denoising, feature engineering, and nonlinear feature refinement. Data cleaning and feature preparation included: **Missing Values**: Missing values in the Injury column were filled with zero and converted to floating-point format.

Feature Engineering: The Month variable was extracted from YearMonth, and Borough was processed using one-hot encoding.

Denoising: Extreme values (outside the 1st and 99th percentiles) were removed for Temperature, Precipitation, HVI, Noncompliant Count, Issue Number, and Injury counts. Samples exhibiting concurrent extreme heat and precipitation were excluded, as were records with negligible construction activity (low IssueNumber). HVI values were capped within a reasonable upper limit.

Log Smoothing: Logarithmic smoothing was applied to high-variance features (Noncompliant Count, Issue Number, Precipitation) to prevent dominance by single variables.

Additional regularization terms were excluded to avoid further underfitting, given the inherent sparsity of the dataset.

3.5.2 Feature Standardization

Input features consisted of Average Temperature, Average Precipitation, Heat Vulnerability Index, NoncompliantCount, IssueNumber, Month, and the one-hot encoded borough columns. All variables were standardized to improve gradient stability. The target variable (Injury) was normalized using mean and standard deviation scaling. The dataset was divided into training (80 percent) and validation (20 percent) subsets.

3.5.3 Model Architecture

The InjuryRegressor neural network used a multilayer nonlinear structure defined by:

Input Layer: Corresponds to all processed input features.

First Hidden Layer: 32 neurons using the LeakyReLU activation function.

Dropout Layer: Rate of 0.1, used to prevent overfitting.

Second Hidden Layer: 16 neurons using the LeakyReLU activation function.

Third Hidden Layer: 8 neurons using the LeakyReLU activation function.

Output Layer: Single neuron outputting the predicted injury count.

The LeakyReLU activation function supports non-zero gradient flow for negative inputs, suitable for sparse regression tasks with limited learning capacity.

Regularization mechanisms were omitted because the model did not show signs of overfitting and was unable to extract high-complexity patterns from the dataset.

3.5.4 Loss Function and Optimizer

The model uses Mean Squared Error (MSE) as the loss function. The Adam optimizer was selected with a learning rate of $3 * 10^{-4}$, balancing convergence speed and stability through automatic learning rate adjustment. Training and validation losses were recorded to monitor convergence trends and generalization performance. Note: Traditional nonlinear count models (e.g., Poisson and Negative Binomial) were tested but excluded due to convergence failures during training.

3.5.5 Training and Validation

erformance was evaluated using multiple error-based metrics: $\boldsymbol{R^2}$ **(Coefficient of Determination)**: Measures the proportion of variance explained by the model. An $R^2 < 0$ indicates the model failed to learn effectively.

RMSE (Root Mean Square Error): Reflects the average magnitude of prediction error.

MAE (Mean Absolute Error): Measures the average deviation between predicted and actual values.

Scatter plots comparing actual versus predicted values were generated to visually assess accuracy, where ideal performance would align predicted points along the diagonal.

3.5.6 Model Improvements

Two improvement strategies were explored: **Approach 1: Hybrid Lag and Group Bias Linear Model**

Concept: Incorporates time-lag features and borough-specific biases into linear regression to capture temporal inertia and regional disparities.

Feature Processing: Retained only samples with construction records; applied log smoothing to non-compliant counts, permits, and precipitation; generated one-period lag features by borough and month.

Structure: Includes global linear weights and regional bias terms to reflect baseline risks across boroughs.

Results: While the model demonstrated some capacity to explain regional differences, overall R^2, RMSE, and MAE metrics remained poor, indicating limited fit.

Approach 2: Two-Stage Hybrid Model (No Lag, Strict Denoising)

Concept: Adopts a “Classify-then-Regress” structure to improve stability under sparse data conditions.

Stage 1 (Classification): Uses a neural network to determine the probability of an injury occurring (Injury > 0).

Stage 2 (Regression): For confirmed injury samples, estimates the actual count using a linear model with borough biases.

Results: The classification stage achieved high accuracy (0.8–0.9), effectively identifying high-risk months. However, while the regression stage showed a slight improvement in R^2 over single-stage models, overall predictive capability remained unsatisfactory due to the limited volume of data.

4 Results and Discussion

4.1 Introduction

This section presents the performance and interpretive results of the four modeling methodologies introduced previously. The evaluation focuses on the predictive capabilities, structural characteristics, and limitations of each model. The discussion also considers the practical implications of the outputs in the context of construction safety analysis and risk assessment.

4.2 K-Means Classification Models

4.2.1 Models

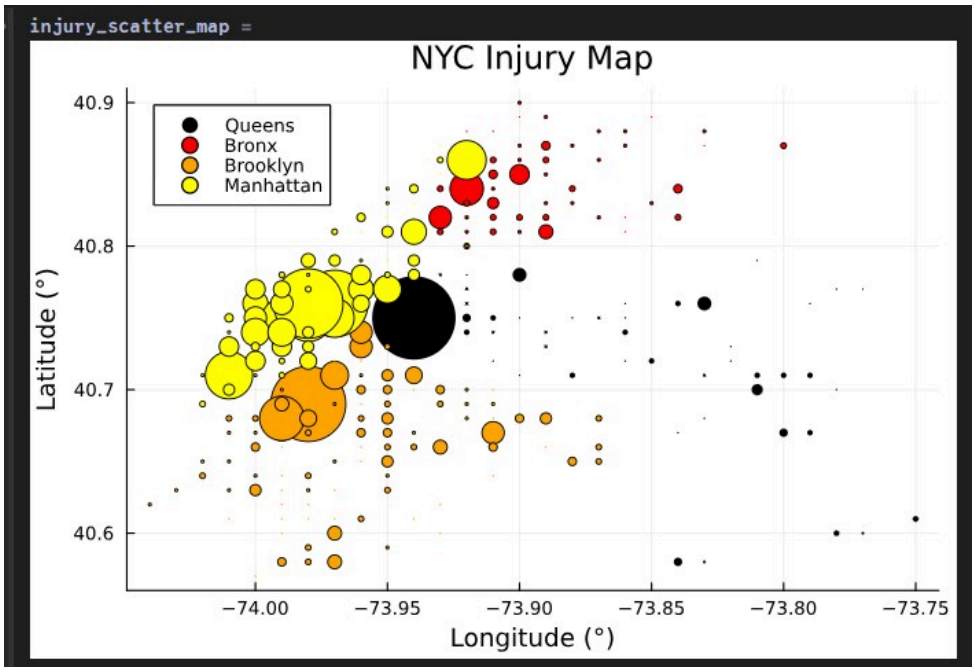

Figure 13: Spatial Distribution of Injuries from Each Borough

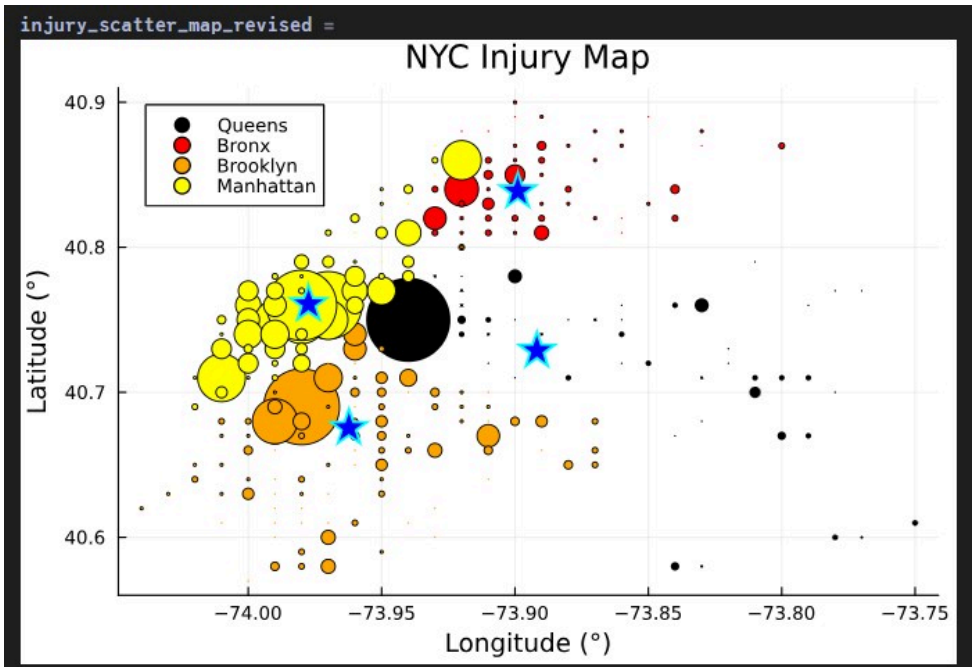

Figure 14: K-Means Model w/ Centroids

4.2.2 Discussion

The spatial distributions display varying concentrations of injuries across the four boroughs of New York City. The visual patterns indicate that incidents are heavily clustered in Manhattan, particularly around the Times Square area. As this region is densely populated and characterized by high real estate value, frequent construction activities are expected. The elevated concentration of injuries appearing in Brooklyn and Queens occurs near the same central region, suggesting that construction incidents are more prevalent within the most populated and commercially intensive zones of the city.

These patterns imply the potential necessity for strengthened Occupational Safety and Health Administration (OSHA) oversight within these high-risk regions. Consideration may also be given to subdividing boroughs into smaller administrative districts, enabling more localized regulatory strategies intended to reduce the likelihood of construction-related injuries and fatalities.

The weighted centroids illustrated by the star markers highlight areas with the highest proportional density of injuries. For example, the centroid in Manhattan aligns with the Times Square vicinity, confirming that this region represents a concentrated zone of risk where increased construction activity is likely to lead to more reported injuries. A similar pattern is observed in Brooklyn, where the centroid lies within a densely populated development corridor.

In contrast, the centroid locations in Queens and the Bronx provide less definitive insight due to broader spatial dispersion. Injury cases in these boroughs are more evenly distributed, rather than concentrated around a central location, reflecting the larger geographical size and more dispersed construction activities. This also emphasizes a primary limitation within this model structure: differences in borough area introduce bias that affects centroid interpretation. Manhattan registers the highest incident count largely because it represents the most densely populated and development-intensive area. Overall, these results suggest that additional contextual features, such

as the number of active construction projects over time, would strengthen the predictive capacity of this model and support more accurate clustering interpretations.

4.3 Decision Tree Classification Models

4.3.1 Models

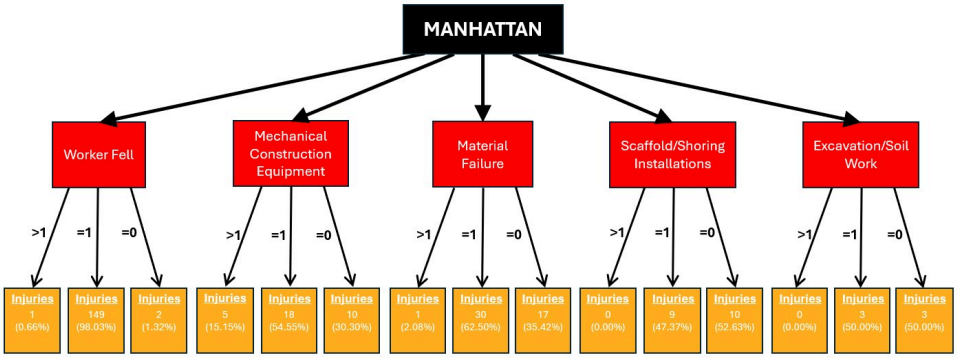

Figure 15: Manhattan Injury Classification Tree

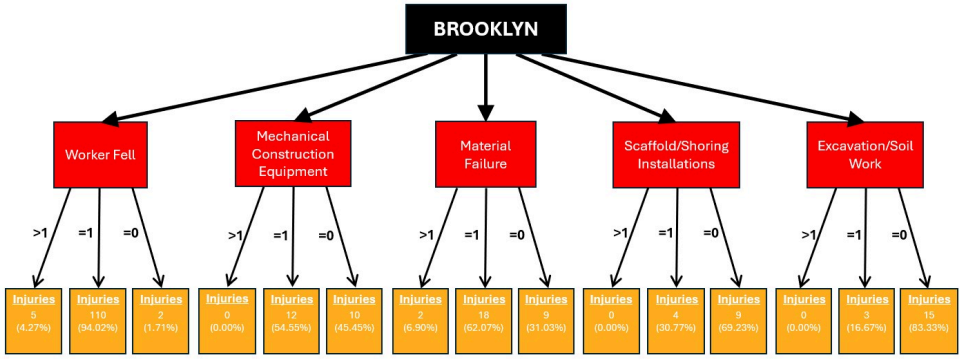

Figure 16: Brooklyn Injury Classification Tree

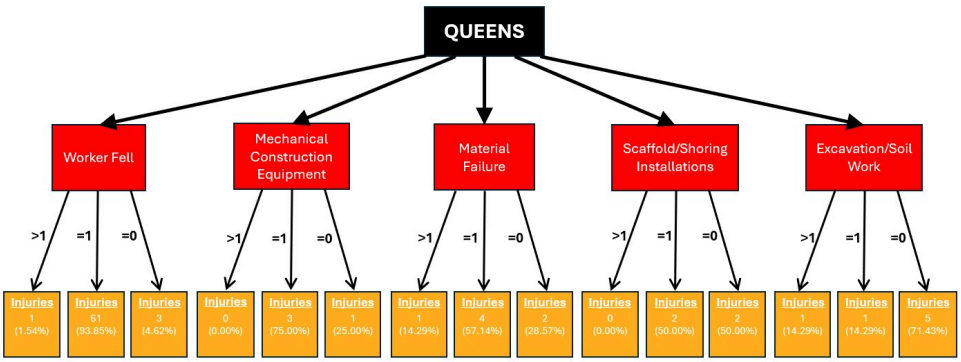

Figure 17: Queens Injury Classification Tree

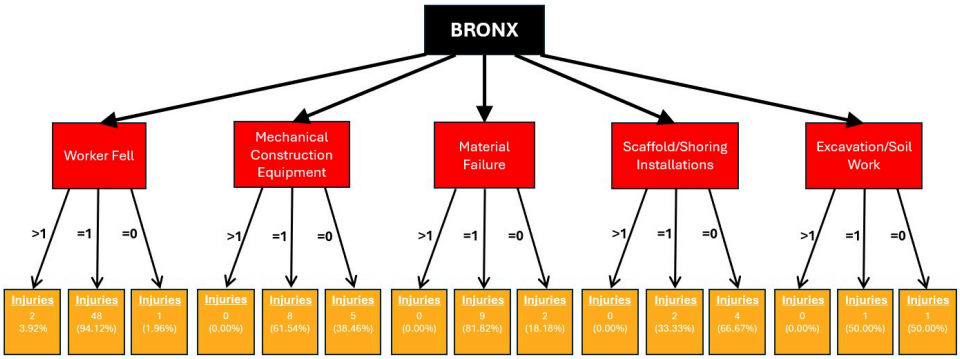

Figure 18: Bronx Injury Classification Tree

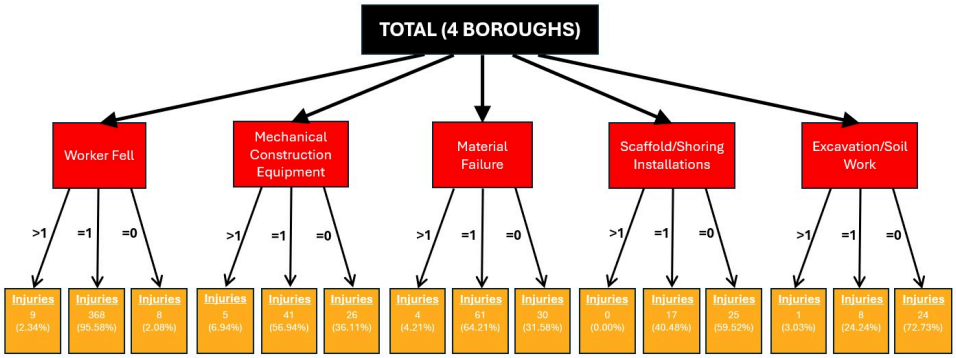

Figure 19: Four Boroughs Total Injury Classification Tree

4.3.2 Discussion

Model Interpretability and Structure: Unlike the Neural Network models discussed in Section 4.4, which provide probabilistic outputs, the Decision Tree Classification models offer a transparent, white-box visualization of risk factors. As shown in Figures 15 through 19, the models successfully hierarchized the data, utilizing Incident Type as the primary splitting criterion at the internal nodes. This structural arrangement confirms that while the borough determines the baseline environment, the specific nature of the incident—such as is the most critical determinant of injury severity.

Borough-Specific Risk Pathways: The topology of the decision trees reveals distinct risk profiles across the boroughs. The Manhattan tree (Figure 15) and Brooklyn tree (Figure 16) exhibit more complex branching structures compared to Queens and the Bronx. This complexity likely reflects the higher density and variety of construction projects in these core boroughs, where risks are multifaceted. Specifically, in Manhattan, branches related to **Material Failure** and **Mechanical Construction Equipment** are prominent, suggesting that infrastructure-heavy projects in dense urban environments carry specific equipment-related risks. In contrast, the trees for Queens and the Bronx (Figures 17-18) show a more streamlined structure, where **Worker Fall** remains the dominant predictor of injury outcomes.

Risk Hierarchy and Classification Logic: A critical insight from the leaf nodes is the identification of high-probability injury pathways. Across the aggregate model (Figure 19), incidents categorized as **Worker Fall** and **Scaffold/Shoring Installations** consistently lead to leaf nodes associated with injury occurrences. This aligns with the finding in Section 4.2 that certain incident types are inherently more dangerous regardless of location. The decision trees effectively function as a rule-based classifier: if an accident involves a fall or excavation, the probability of it being an injury-causing event is statistically maximized.

Operational Implications: While the Neural Network provided higher predictive accuracy through non-linear feature combination, the Decision Trees provide actionable safety rules. The **If-Then** logic derived from these trees supports the development of targeted regulatory checklists. This complements the **Conservative Prediction Strategy** discussed in Section 4.4.2 by providing clear interpretability for on-site safety officers, allowing them to identify high-risk scenarios before they escalate into actual injuries.

4.4 Neural Network Classification Models

4.4.1 Models

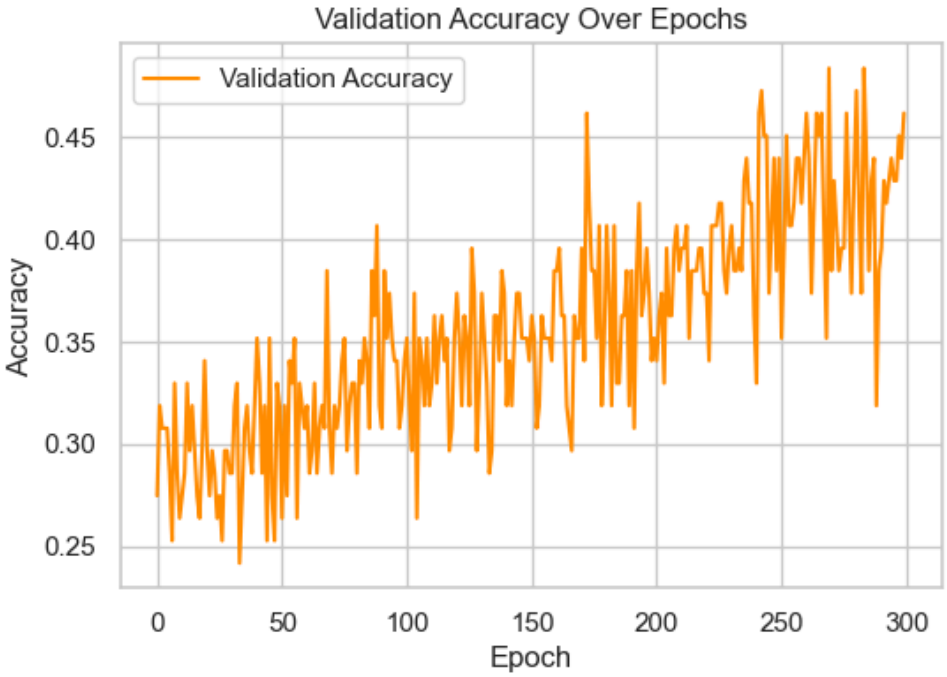

Figure 20: Validation Accuracy Over Training Epochs

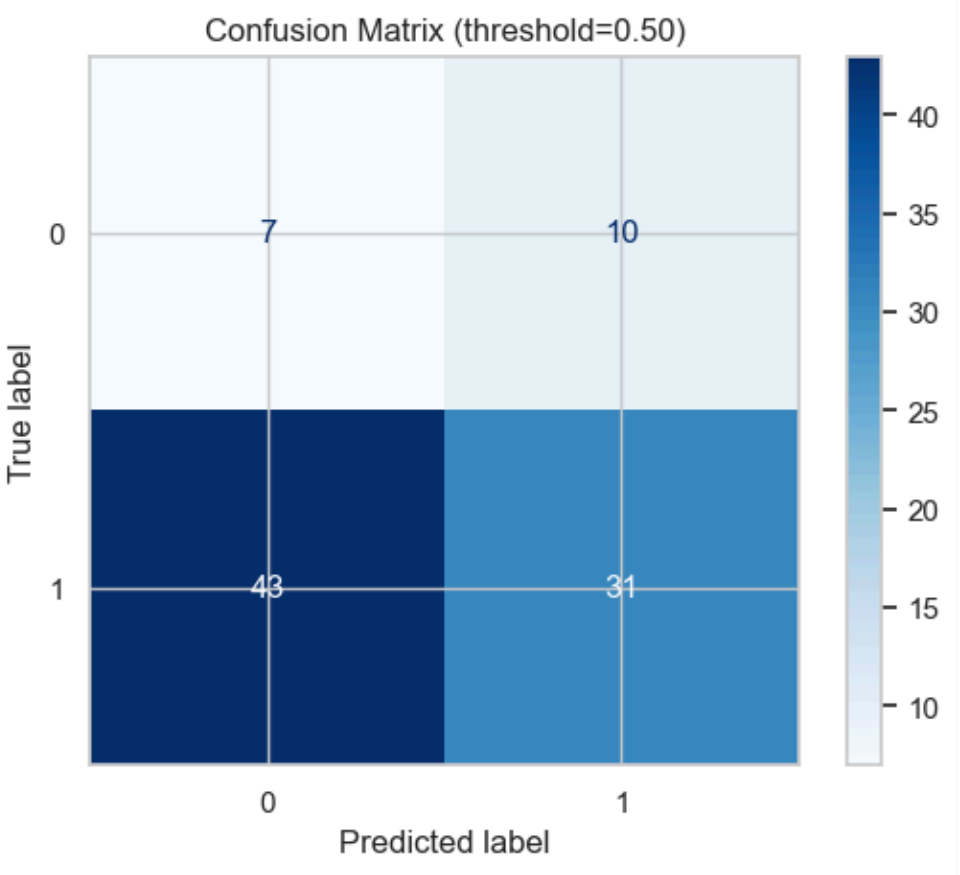

Figure 21: Confusion matrix of the neural network classifier (threshold = 0.50)

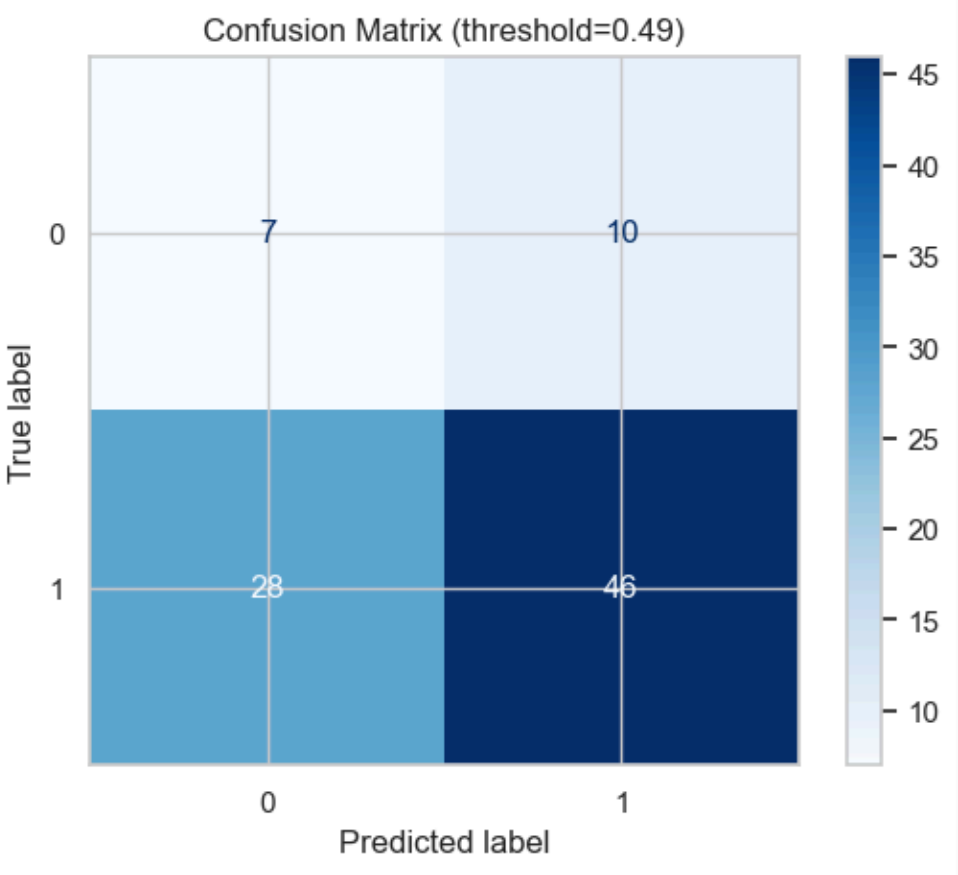

Figure 22: Confusion matrix of the neural network classifier (threshold = 0.49)

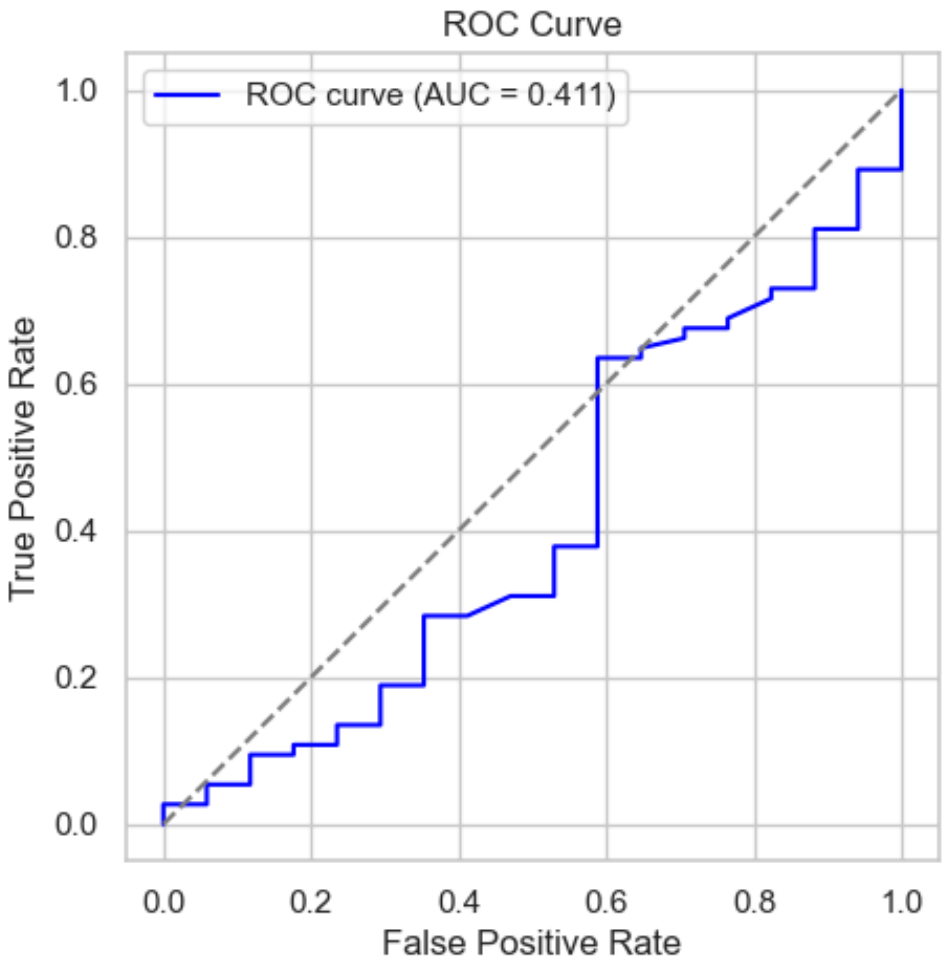

Figure 23: ROC curve with AUC = 0.411

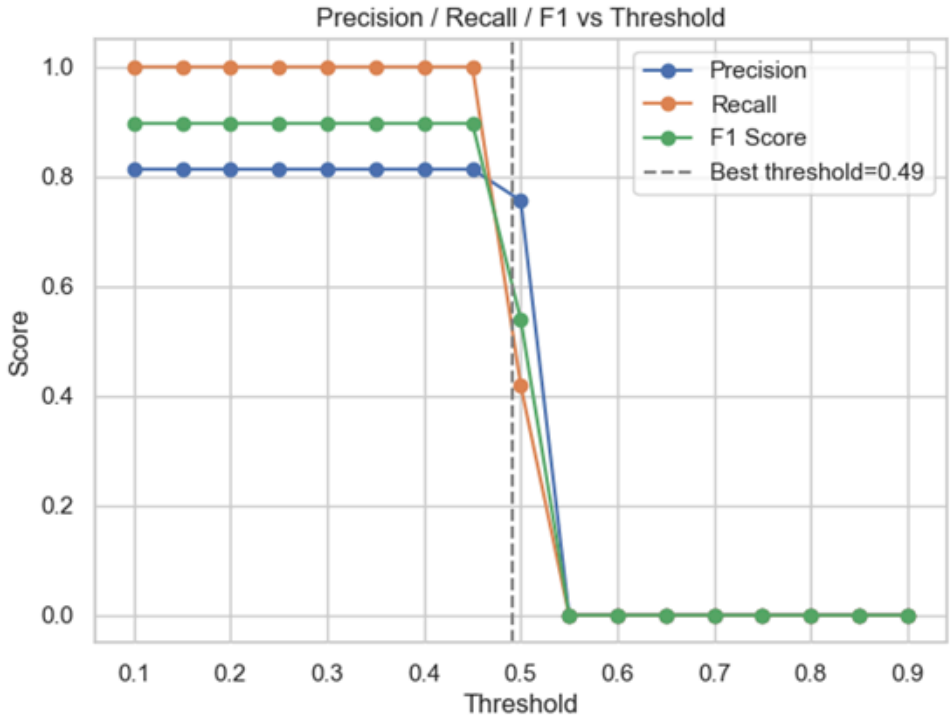

Figure 24: Precision, recall, and F1 score across decision thresholds

4.4.2 Discussion

As illustrated in the results, the validation accuracy exhibited significant fluctuation during the initial training phase but demonstrated a steady upward trend overall, rising from approximately 0.28 to nearly 0.45. This indicates that the model progressively learned effective relationships between features, leading to improved validation performance. While there is room for further accuracy improvement, the absence of significant overfitting suggests that the network architecture and regularization settings (Dropout=0.1) are reasonable and provide good generalization capability.

At the default threshold of 0.50, the model's identification of "Injury" (positive class) showed high recall but slightly lower precision. The confusion matrix results are as follows:

True Positives (TP) = 31 (Correctly identified injuries)

False Positives (FP) = 10 (Non-injuries incorrectly predicted as injuries)

True Negatives (TN) = 7

False Negatives (FN) = 43

These results suggest a conservative prediction strategy (preferring false alarms over missed detections). In the context of accident analysis, this bias is acceptable, as false negatives (missed injury predictions) typically carry a higher safety cost than false positives.

Applying the optimal threshold of 0.49, determined by Youden's J statistic, significantly improved the model's recognition capability:

TP increased to 46, and **FN** decreased to 28.

TN remained at 7, with a slight increase in **FP** to 10.

This adjustment achieved a better balance, enhancing overall classification accuracy while maintaining high recall. The significant reduction in missed detections (FN) compared to the default threshold highlights that threshold optimization is a critical step in tasks involving imbalanced datasets. Analysis of the metrics is defined as follows:

Precision: The proportion of true injuries among predicted injuries. High precision implies high confidence in positive predictions (few false alarms).

Recall: The proportion of actual injuries correctly identified. High recall implies comprehensive coverage of safety risks (few missed incidents).

F1 Score: The harmonic mean of Precision and Recall, providing a balanced metric for imbalanced datasets.

The plotted curves show the relationship between these metrics and the threshold. In the 0.1–0.49 range, all three metrics remain high: Recall stays near 1.0, Precision stabilizes around 0.8, and the F1 score approaches 0.9. However, beyond the 0.5 threshold, all metrics decline rapidly, indicating that an excessively high threshold makes the model overly conservative, resulting in missed positive samples. Consequently, 0.49 was selected as the optimal threshold, achieving an ideal balance between Recall and Precision and maximizing the F1 score.

4.5 Neural Network Regression Models

4.5.1 Models

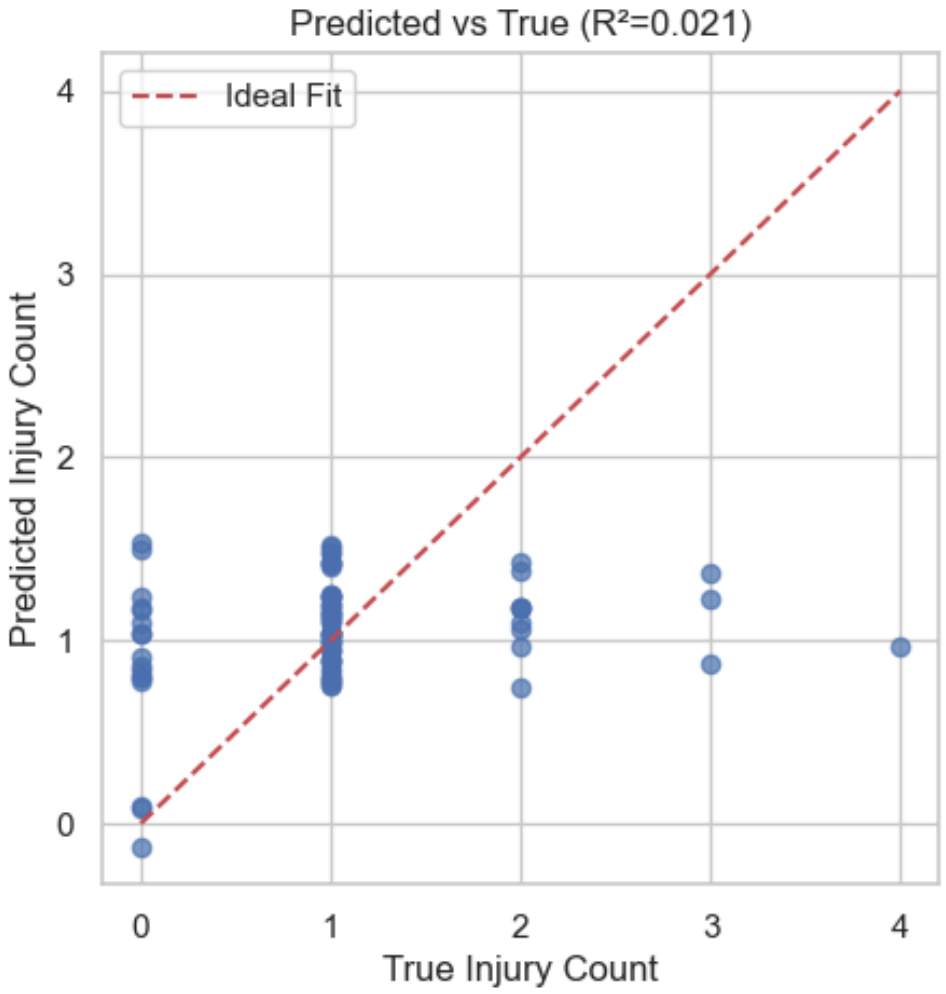

Figure 25: Baseline model performance—predicted vs. true injury counts

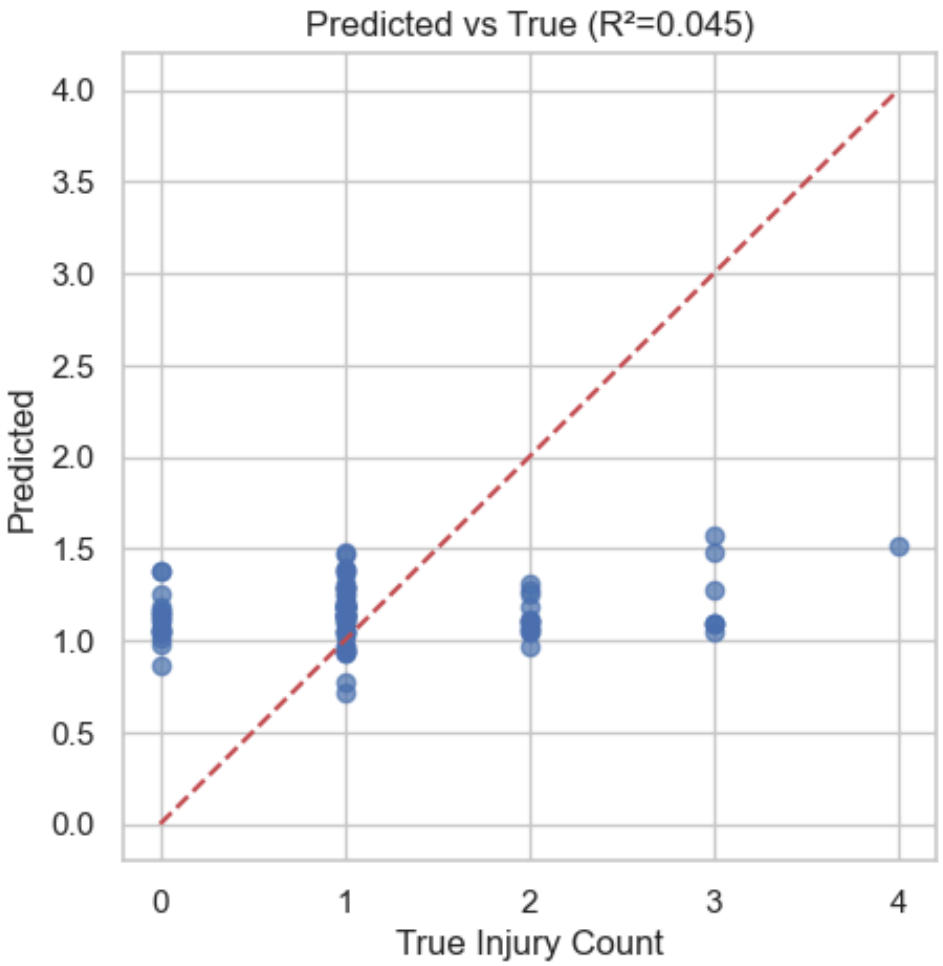

Figure 26: Hybrid Lag and Group Bias Linear Model—predicted vs. true injury counts

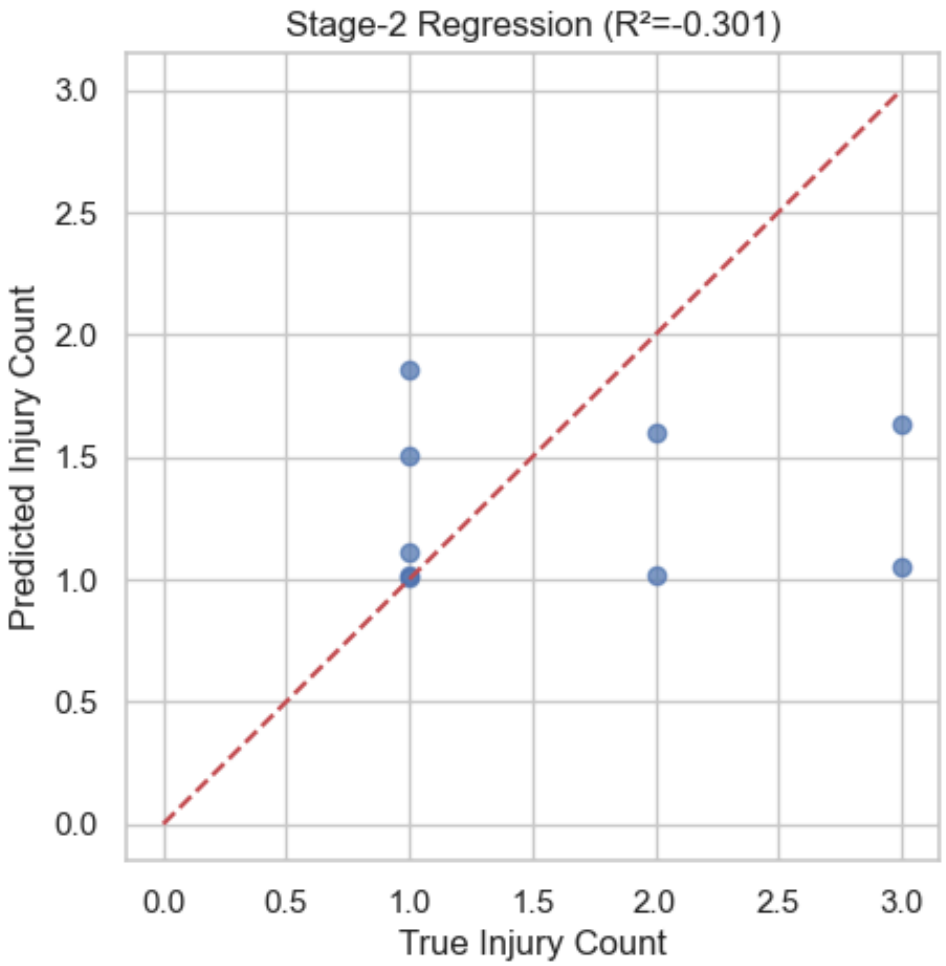

Figure 27: Two-Stage Hybrid Model—predicted vs. true injury counts

4.5.2 Discussion

As illustrated in the regression results, the predictive performance of the count-based models stood in stark contrast to the classification models. While the classification approach successfully identified high-risk scenarios, the regression models struggled to quantify the exact number of injuries, yielding suboptimal R^2 values across all three experimental setups.

Baseline and Hybrid Models Performance: The Baseline Neural Network Regressor (Figure 25) achieved an R^2 of only 0.021. A visual inspection of the scatter plot reveals a distinct regression to the mean phenomenon: while the true injury counts range from 0 to 4, the model's predictions cluster tightly in the narrow range of 0.5 to 1.5. This indicates that the model, unable to find strong signal patterns in the sparse data, defaulted to predicting the average injury rate to minimize the global loss function.

The introduction of temporal features in the Hybrid Lag and Group Bias Linear Model (Figure 26) resulted in a marginal improvement, raising the R^2 to 0.045. Although the scatter points show a slightly wider distribution compared to the baseline, the improvement is negligible. This suggests that construction injury events in this dataset lack significant temporal autocorrelation; the occurrence of an injury in the previous month does not strongly predict the magnitude of injuries in the subsequent month.

Two-Stage Model Limitations: Most notably, the Two-Stage Hybrid Model (Figure 27), which was designed to mitigate sparsity by first classifying injury occurrence and then regressing the count,

resulted in a negative R^2 of −0.301. While the classification stage (Stage 1) showed promise in filtering out zero-event months, the subsequent regression stage (Stage 2) suffered critically from data scarcity. After filtering for only positive-injury samples and applying strict denoising, the effective sample size became too small for the regressor to generalize. The negative R^2 implies that the model's predictions were worse than simply using the horizontal mean of the test data, highlighting the dangers of aggressive data segmentation on small datasets.

Summary of Regression Challenges: The poor performance across these models reinforces the findings from the preliminary methodology section: predicting the exact magnitude of construction accidents is significantly harder than predicting their probability. The high sparsity (zero-inflation) and the stochastic nature of accidents mean that the signal-to-noise ratio is too low for standard regression loss functions to converge effectively. Future improvements would likely require a significantly larger dataset to support complex architectures or the use of specialized loss functions like Zero-Inflated Poisson loss, provided that convergence issues can be overcome.

5. References

[1]Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. Automation in Construction, 69, 102-114. https://doi.org/10.1016/j.autcon.2016.05.016
[2]NYCOSH. (2022, February 10). Deadly Skyline: An annual report on construction fatalities in New York State (2022 ed.). https://nycosh.org/wp-content/uploads/2022/02/NYCOSH_Deadly-Skyline-Report_2022.pdf
[3]Office of the New York State Comptroller. (2025, July). The construction sector in New York City: Post-pandemic trends (Report No. 8-2026). https://www.osc.ny.gov/files/reports/pdf/report-8-2026.pdf
[4]Carrivick, P. J. W., Lee, A. H., & Yau, K. K. W. (2003). Zero-inflated Poisson modeling to evaluate occupational safety interventions. Safety Science, 41(1), 53–63. https://doi.org/10.1016/S0925-7535(01)00057-1
[5]Junjia, Y., Alias, A. H., Haron, N. A., & Abu Bakar, N. (2024). Machine learning algorithms for safer construction sites: Critical review. Building Engineering, 2(1), 544. https://doi.org/10.59400/be.v2i1.544
[6] New York City Department of Buildings. (n.d.). *Incident Database* [Data set].
[7] Nayak, S. G., Shrestha, S., Kinney, P. L., Ross, Z., Sheridan, S. C., Pantea, C. I., Hsu, W. H., Muscatiello, N., & Hwang, S. A. (2018). *Development of a heat vulnerability index for New York State.* Public Health, 161, 127–137.
[8] Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.
[9] Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.
[10] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
[11] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
[12] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
[14] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.
[15] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
[16] City of New York. (2025). *Official website of the City of New York*. Retrieved November 11, 2025, from https://www.nyc.gov/main
[17] City of New York. (n.d.). *DOB Job Application Filings* [Data set]. NYC Open Data. Retrieved November 11, 2025, from https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2 [1NYC Maps. Maps of World. https://www.mapsofworld.com/usa/new-york-city-map.html#google_vignette