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

Zach 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”

1. Data Description(Project 1)

1.1 File Content

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.

1.2 Source

We pull data from New York City Department of Buildings. (n.d.). Incident Database [Data set].

1.3 Format and Size

The dataset includes approximately 958 rows, each representing an accident or incident record, and 20 columns containing attribute fields of these records.

2. Attributes

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

Proposal for attribute usage will be made, focusing on those with predictive relevance.

3. Proposal

3.1 Objectives

Our main objective is to analyze different types of construction incidents at New York City that happened within 1 or 2 years from now. For this project, we would mainly be examining the nature of construction related incidents and accidents as well as performing correlations with the data by examining the prevalence of each incident and accident at each of the five boroughs of New York City. We would want to see where each type of incident has the highest probability of occurring, and where specifically measures should be implemented to prevent these types of incidents. Finally, keeping track of when these incidents occurred will also be critical as the data could also be used to calculate the frequency of accidents over time.

3.2 Preprocessing

Filtering may be applied to eliminate less effective or redundant variables, such as postcode or latitude, to improve model performance. Tidying and cleaning the data is also necessary before analyzing and correlating the data. We would probably need to order our data in terms of when they happen as well as categorizing the incidents/accidents that happened at each borough.

3.3 Output

- Incident vs accident count over time (could be in spans of 1 month)
- How many construction incidents and accidents happened at each month
- Computed severity index at each borough (e.g., grade scale from 1 = less severe to 10 = highly severe)

3.4 Input

Date, record type, latitude, longitude, type of incident, and BIN (business effect case, combined with other data).

3.5 Significance

The purpose of the model is to help avoid incidents and accidents at New York City by identifying the dominant attributes influencing outcomes, thereby guiding proactive protection measures in construction management.

4. Exploratory Data Analysis(Project 2)

This section will mainly focus on introductory data analysis with some preliminary tables and plots describing critical aspects of the data. Visible patterns will be discussed with the data that has been

formulated and the coming sections below will describe how we plan on applying and modelling this data.

4.1 Data Integration and Cohort

For the data integration and cohort definition, we first filtered the dataset. We then applied groupby operations to extract and aggregate key information. This aggregation was performed by 'borough (Area)', month, and 'postcode'. The 'postcode' attribute serves as a critical key, as it is directly used to link and integrate the Heat Vulnerability Index data (Nayak et al., 2018).

4.2 Borough × Month Aggregation

To explore trends over time, the data were aggregated by borough and month. The aggregation reveals that incidents tend to cluster during the spring and summer months, aligning with increased construction activity.

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`.

4.3 Temperature, Precipitation, and HVI Added

In this data integration step, we enriched the dataset by incorporating external environmental and vulnerability factors. The Heat Vulnerability Index (HVI) was integrated by joining it with the dataset using 'postcode' as the linking key. We then performed a data cleaning step to ensure data quality by removing records with missing values. Following this, climate variables, specifically 'AvgTemp' (Average Temperature) and 'AvgPrecip' (Average Precipitation), were merged into the dataset. This process of joining HVI, merging climate data, and handling missing values resulted in a final, refined dataset containing 415 valid observations, which was then used for the subsequent correlation and regression analyses

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.

4.4 Injury & Fatality Plots

Four of the preliminary plots below count incidents such as fatalities and injuries that happened at each borough and district. The last four show the total injuries and fatalities that happened in New York City during each month, and then the accumulation of injuries and fatalities over time from January 2024 to October 2025. With this data, we can go even further with borough or district specific statistics regarding these construction incidents.

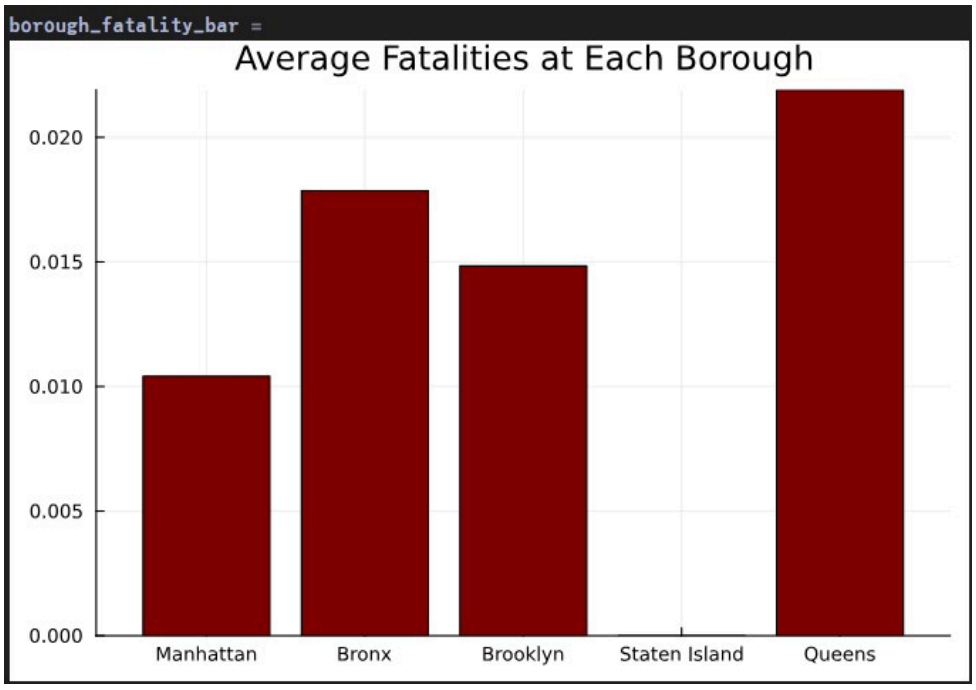

Figure 1: Average Fatalities at Each Borough

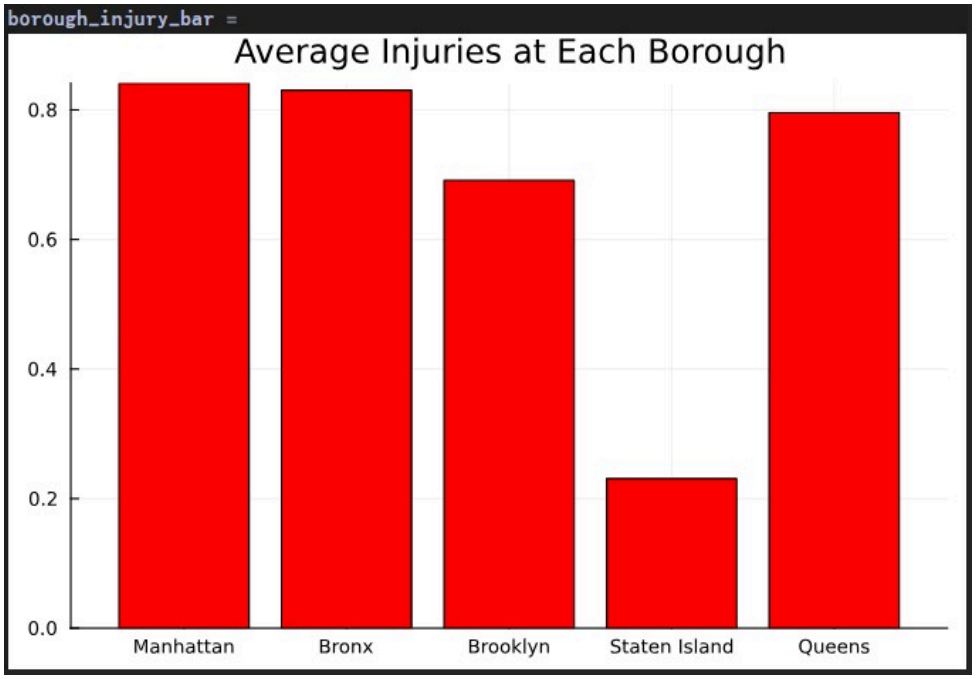

Figure 2: Average Injuries at Each Borough

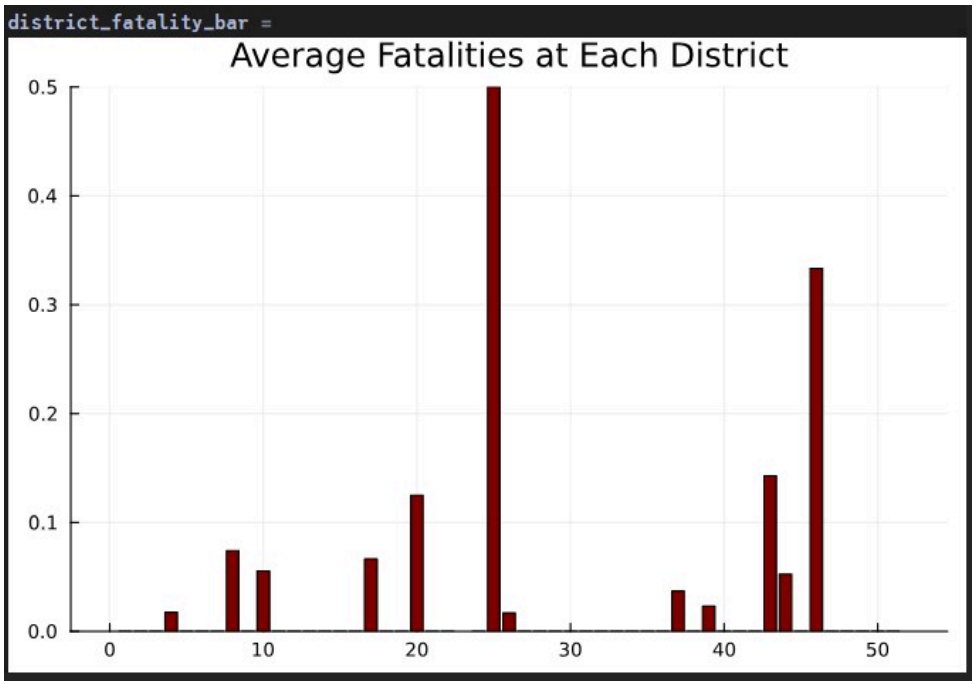

Figure 3: Average Fatalities at Each District

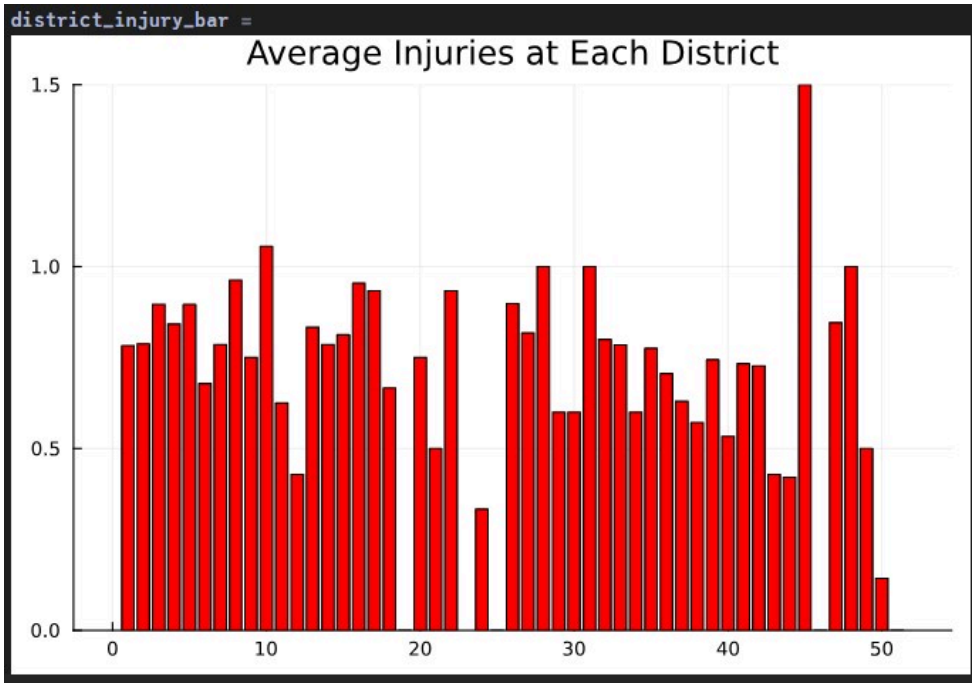

Figure 4: Average Injuries at Each District

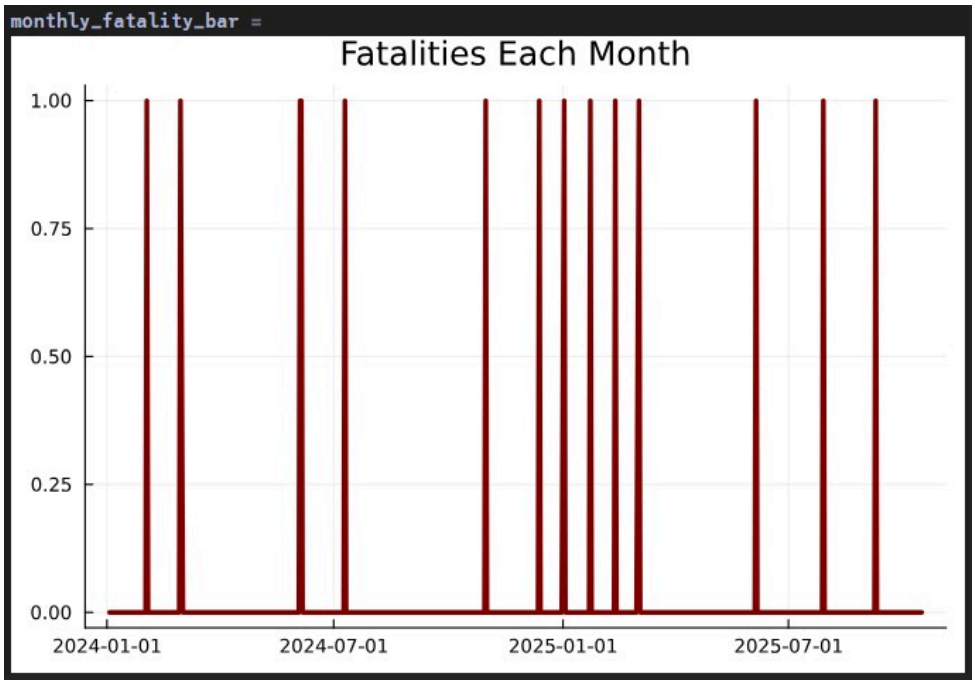

Figure 5: Cumulative Fatalities Overtime

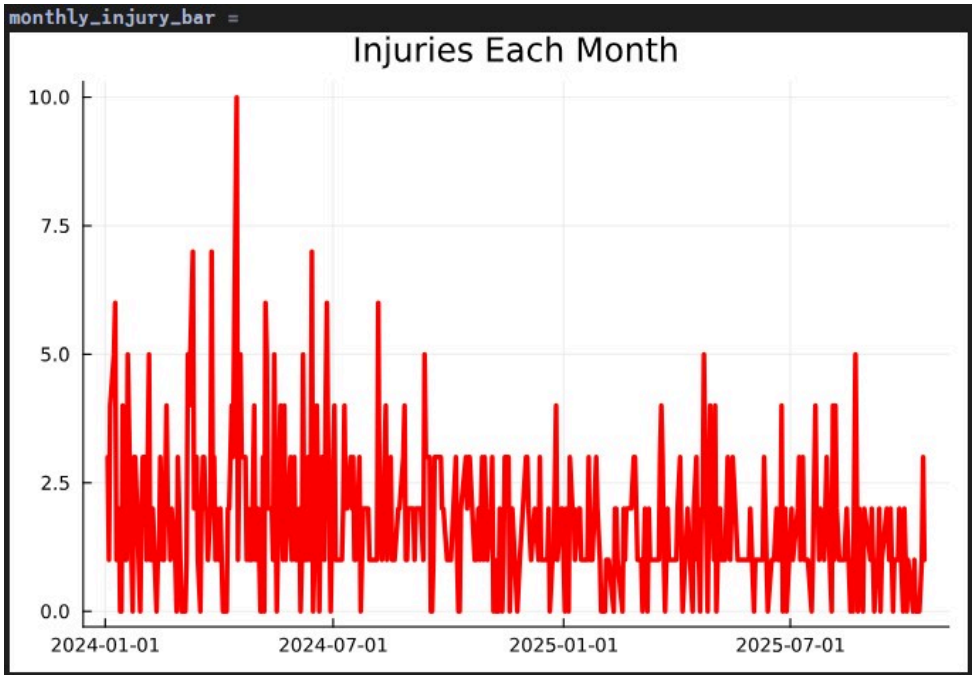

Figure 6: Cumulative Injuries Overtime

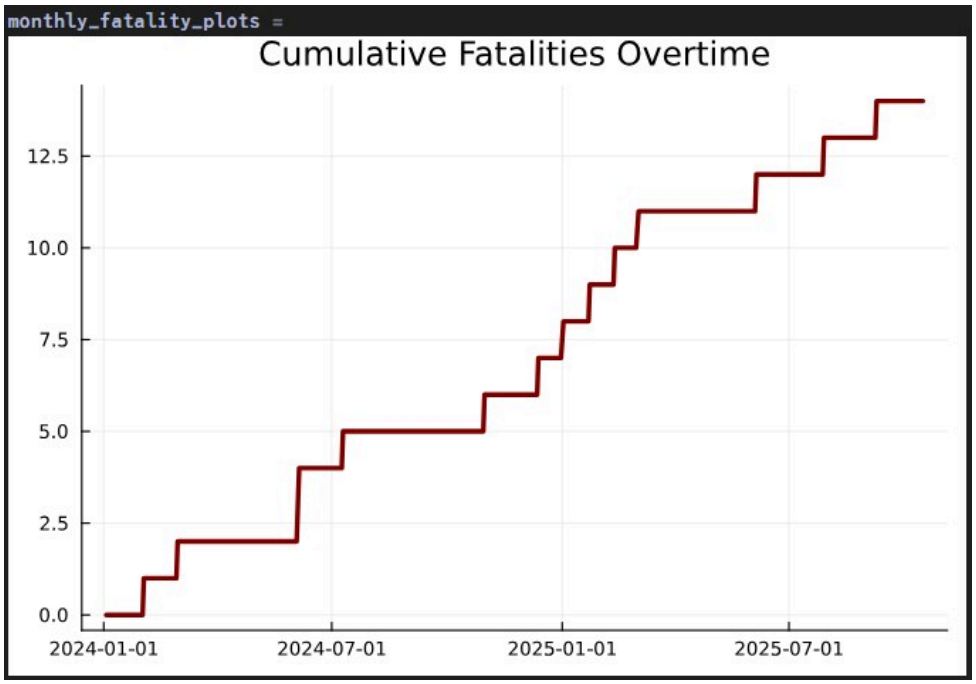

Figure 7: Fatalities Each Month

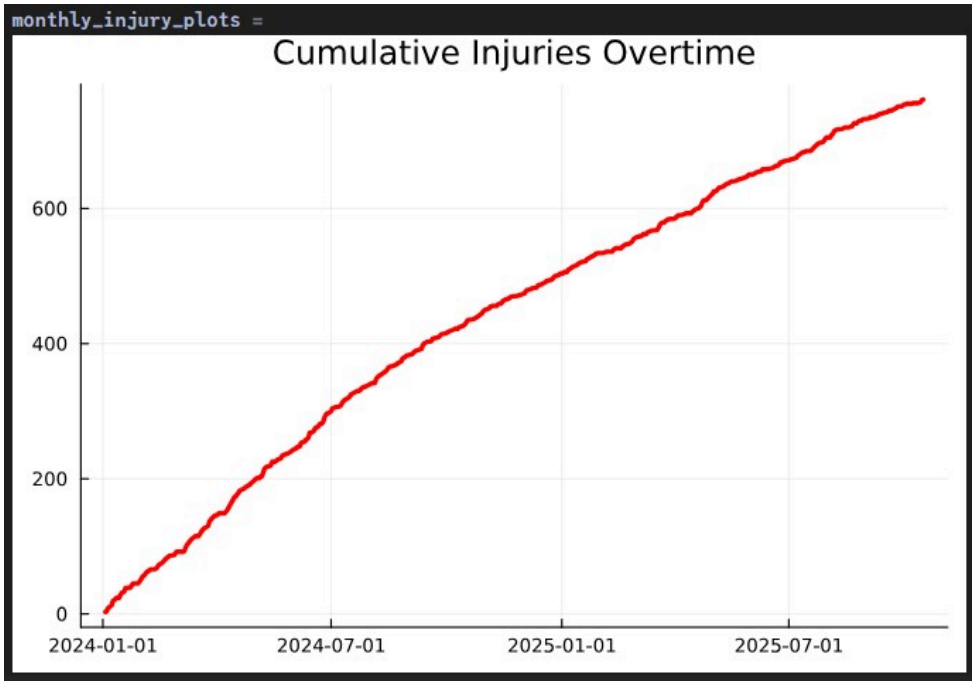

Figure 8: Fatalities Each Month

4.5 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

On average, 83.8% of incidents resulted in at least one reported injury, whereas fatalities were rare (around 1.5% of all cases). The Bronx recorded the highest injury rate, followed closely by Brooklyn.

4.6 Summary

This section synthesizes the primary findings from our exploratory data analysis. We aggregated the data to compute and examine key descriptive statistics. Specifically, we calculated the average number of fatalities, average number of injuries, and average incident counts for each of the five boroughs. From this, we also derived the fatality and injury rates (as percentages) per borough to better understand the proportional risk. Furthermore, our summary includes an analysis of temporal patterns. We investigated monthly trends by charting the frequency and cumulative totals of both fatalities and injuries over the study period. These initial summaries provide a foundational understanding of which areas are most affected and how incident severity fluctuates over time.

5. Correlation Plots

5.1 Weighted HVI

Weighted averaging is used when different observations contribute unequally to an aggregate measure.

5.2 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

We conducted a global correlation analysis to understand the initial linear relationships between the primary variables. The results, partially shown in the table1, reveal several key patterns.Most notably, there is a very strong positive correlation between TotalIncidents and Injury (Pearson $r = 0.958$), which is expected as most incidents involve injuries.In contrast, Fatality demonstrates a weak correlation with the other variables in the table. The correlation with TotalIncidents is low ($r = 0.120$), and it is even weaker with Injury ($r = 0.075$). The environmental variables, AvgTemp ($r = -0.007$) and AvgPrecip ($r = -0.153$), also show negligible linear relationships with fatalities.Furthermore, analysis of the Heat Vulnerability Index (HVI) indicated a moderate negative correlation with both incidents and injuries, with correlation coefficients (r) observed in the range of approximately −0.57 to −0.606. This suggests that areas with higher vulnerability scores may, counterintuitively, be associated with fewer reported incidents in this dataset, prompting the need for further, more nuanced analysis.

5.3 Log-scaled Correlation

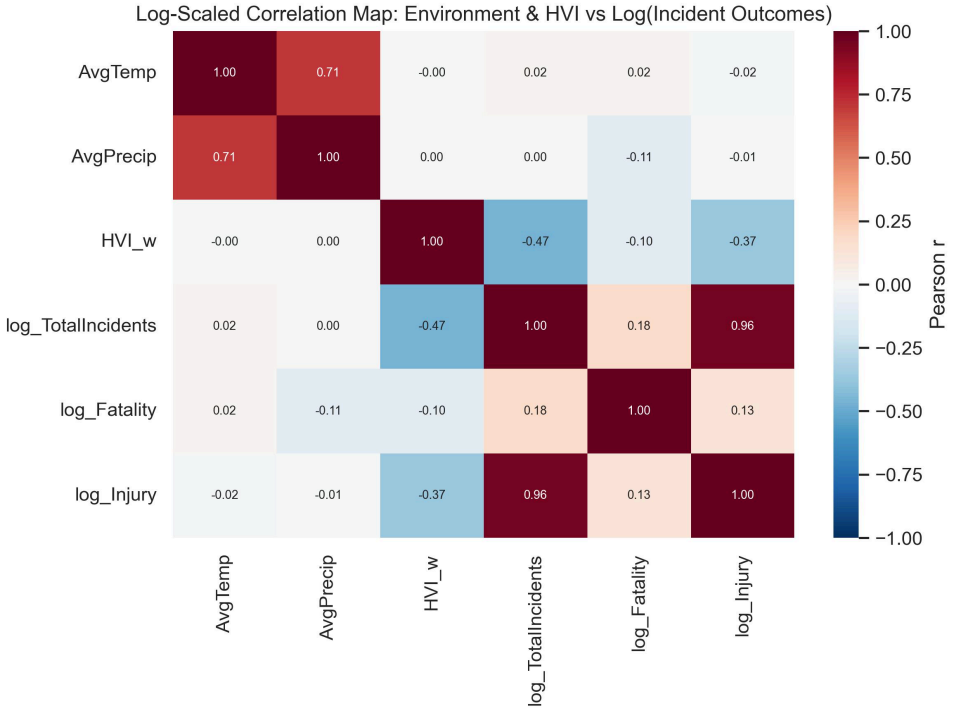

Figure 9: Correlation heatmap after log scaling

Results show a strong positive correlation between TotalIncidents and Injury (r ≈ 0.96) and a negative correlation between HVI and Fatality (r ≈ −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.

5.4 Summary

These visual representations are only preliminary and can be used as inspiration for the plots in the actual predictive modelling phase.

6. Regression Plots

Regression modeling was performed using Poisson, Negative Binomial, and Logistic regressions, which are standard approaches for count and binary outcomes in risk and safety studies.

6.1 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

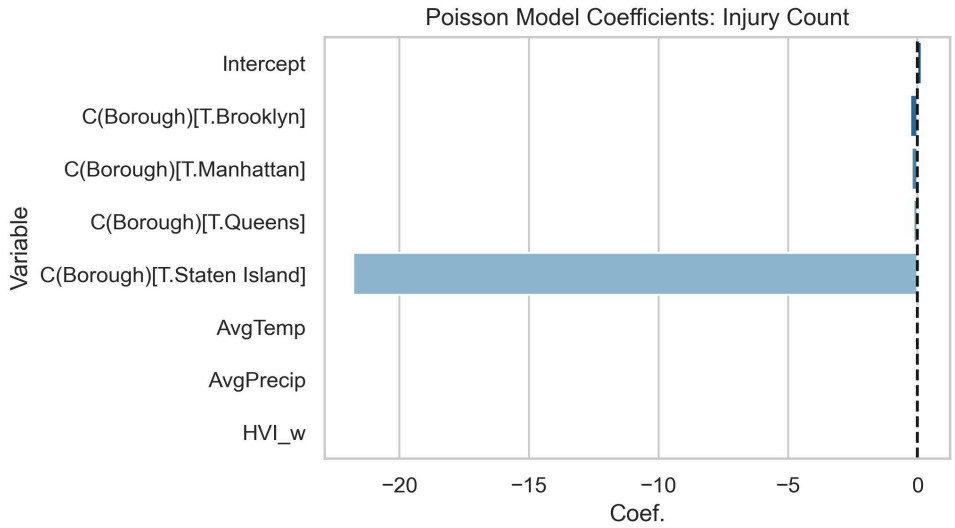

Figure 10: Poisson injury model coefficients

6.2 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

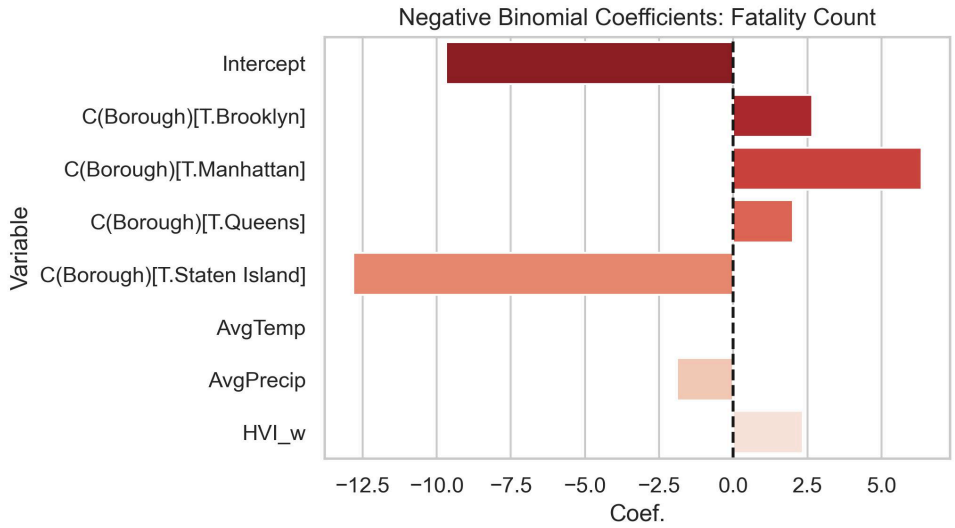

Figure 11: Negative binomial fatality model coefficients

6.3 Logistic Model

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

6.4 Visual Comparisons

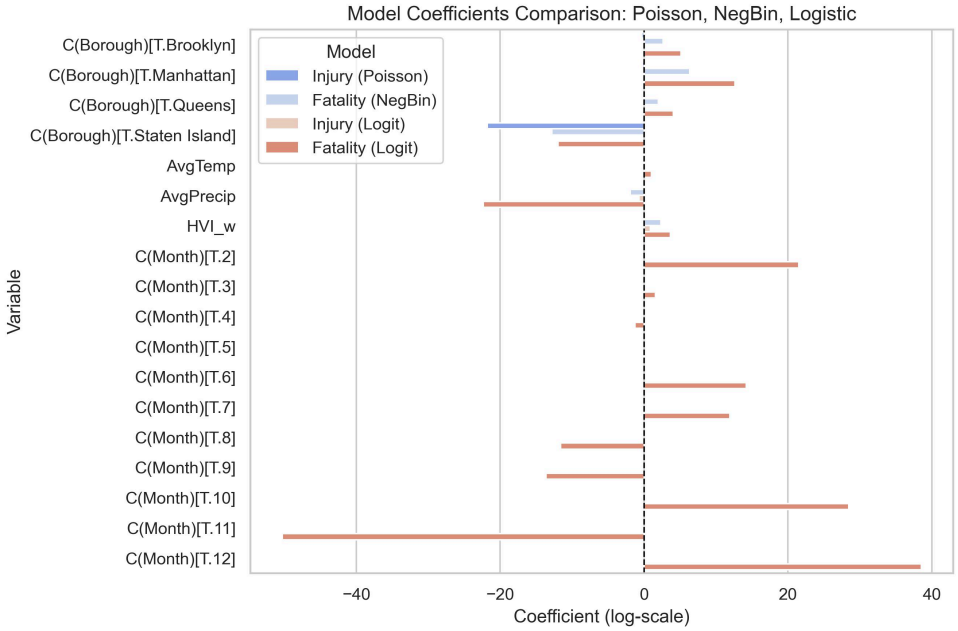

Figure 12: Coefficient comparison

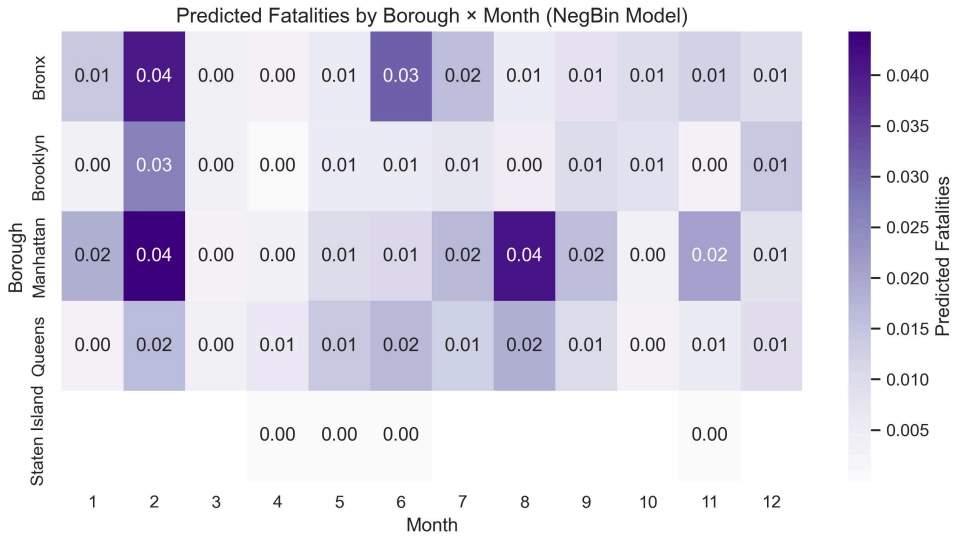

Figure 13: Predicted fatality heatmap

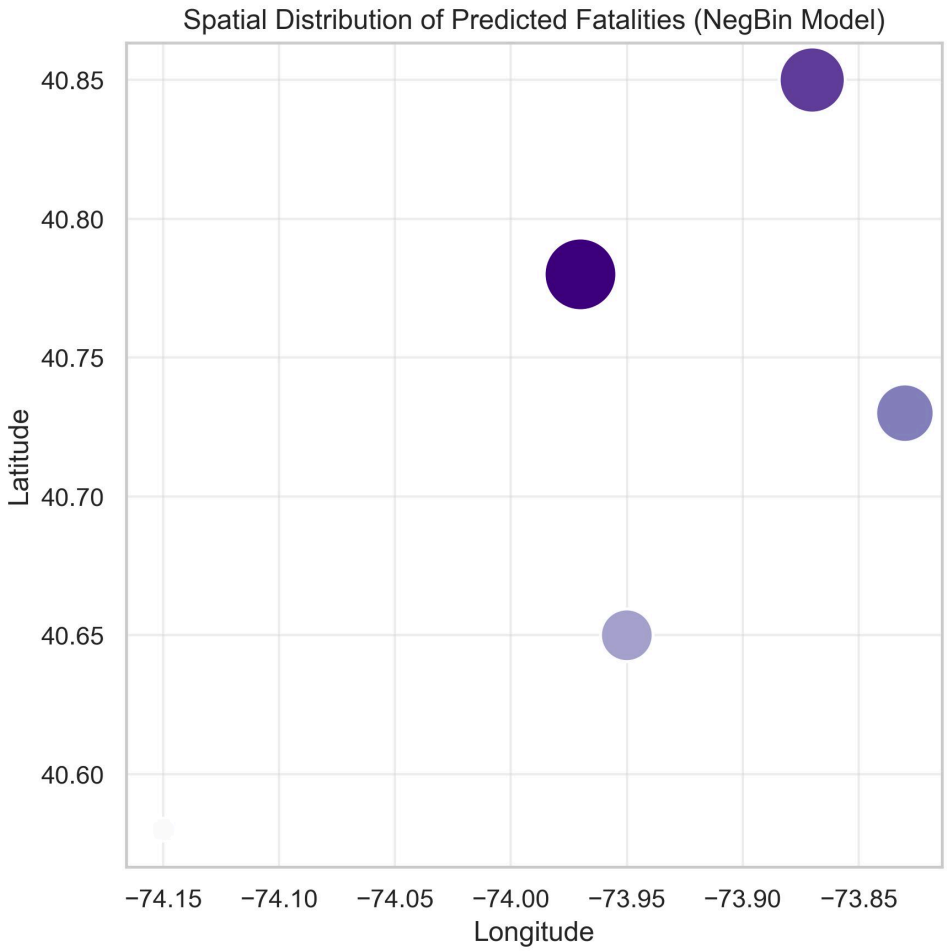

Figure 14: Spatial fatality prediction map

6.5 Summary

These visual representations are only preliminary and can be used as inspiration for the plots in the actual predictive modelling phase.

7. Preliminary Predictive Modeling

This section will describe how we plan on creating more complex models pertaining to our data, such as predictive modelling. This aspect is necessary in terms of understanding important patterns and information that cannot be simply derived by just correlating two variables.

7.1 Objectives

Here are objectives we want to achieve if predictive modelling would be implemented:

1. Know when and where injuries and fatalities are the most severe around New York City
2. Know the most prevalent underlying causes with injuries and fatalities through check descriptions
3. Thoroughly investigate each borough or districts if we want to take this further to thoroughly detect where these incidents are occuring
4. Figure out how much weather (e.g. precipitation, temperature) affects the prevalence of construction incidents
5. Know how create more complex models with this data

7.2 Models

Here are some examples of models we may implement:

1. Binary injury risk per event
2. Count severity per borough-month (injury counts or SeverityIndex)
3. Poisson vs. Negative Binomial
4. Regularized logistic for rare events
5. Heatmap or spatial distribution (longitude and latitude) detailing fatalities or injuries around New York City (may need to include detail with which borough)

7.3 Interpretability & Fairness

This would include inspecting borough effects and error parity across high-HVI ZIPs.

7.4 Next Steps

Steps we could eventually take are adding exposure controls (permits, active sites), extending years, and considering hierarchical models if time allows.

8. K-Means Models & Results

8.1 Hypothesis

Given the specific location of each incident along with the number of construction projects happening at that location, we can figure out the severity of construction incidents at a given location within the boroughs of New York City. We can then use that information to determine which areas need better protocol with their construction projects. Therefore, we can use K-Means to determine which area within each borough has the highest concentration of incidents.

Input: Longitude, Latitude, Injuries, Boroughs

Output: Cities/Counties with Highest Concentration

8.2 Models

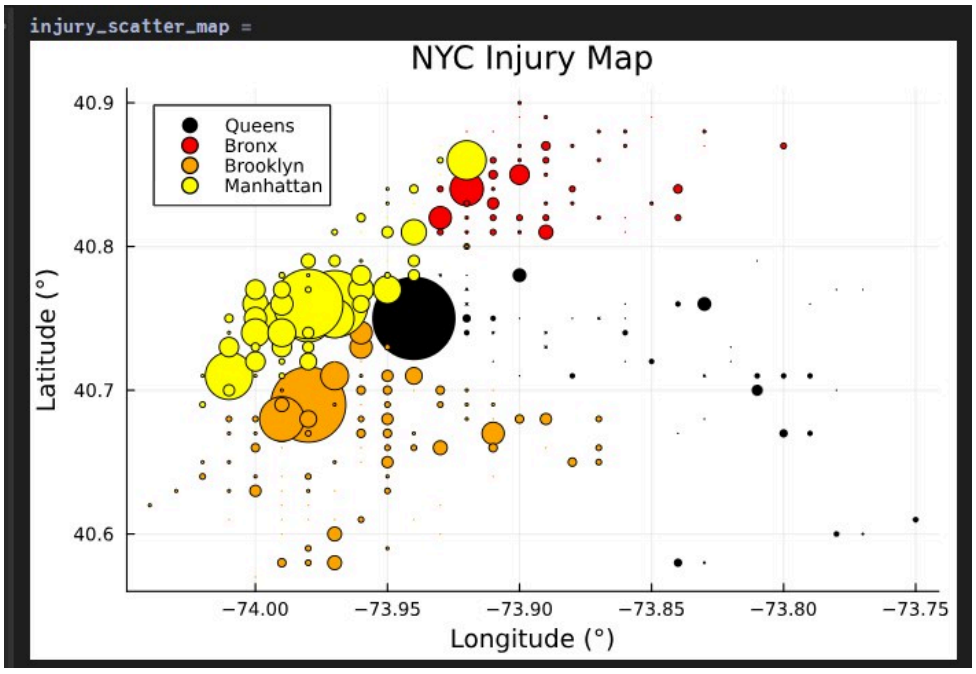

Figure 15: Spatial Distribution of Injuries from Each Borough

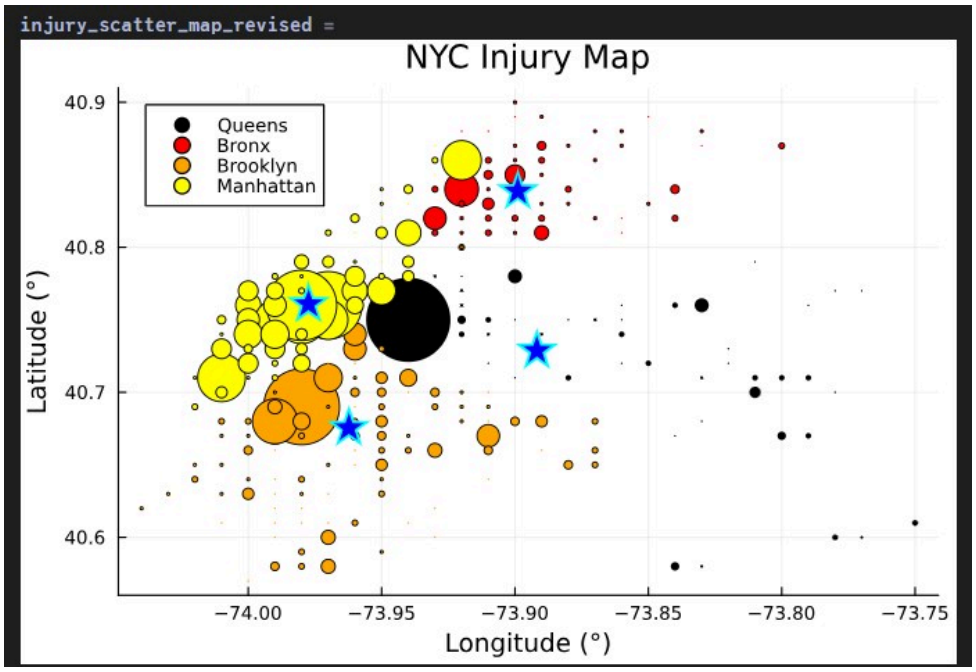

Figure 16: K-Means Model w/ Centroids

8.3 Summary

The K-Means model shows that construction-related injuries in New York City form clear spatial clusters, with higher concentrations appearing within parts of the Bronx and Brooklyn, where repeated incident points are densely grouped. By using longitude, latitude, injury counts, and borough information, the clustering results highlight these boroughs as priority areas for strengthened safety protocols and resource allocation.

9. Classification Tree Models & Results

9.1 Hypothesis

Through the 4 major boroughs of New York City, we can determine common factors pertaining to construction incidents/accidents that lead to injuries in New York City. From this data, we can determine which factors should primarily be examined in terms of implementing new state OSHA regulations.

Input: Borough, Incident Type

Output: Injuries

9.2 Model

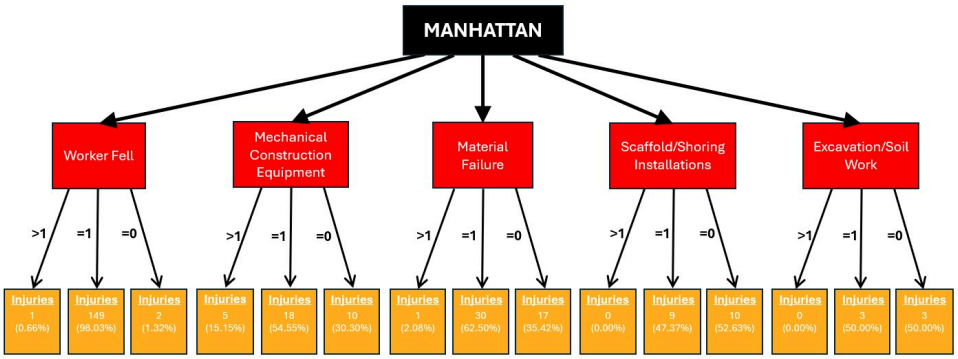

Figure 17: Manhattan Injury Classification Tree

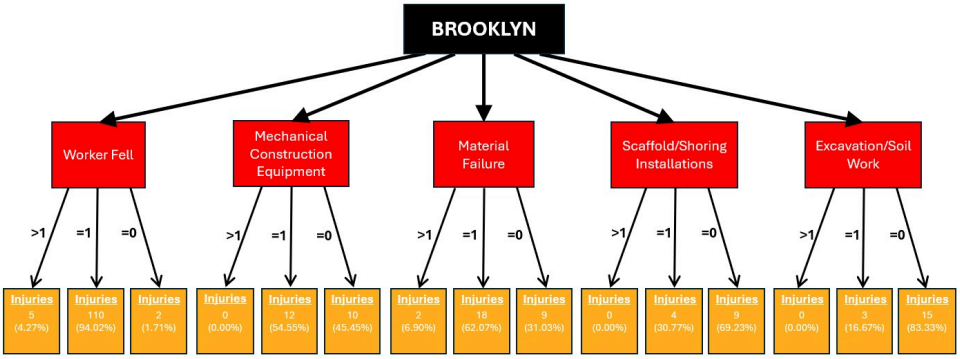

Figure 18: Brooklyn Injury Classification Tree

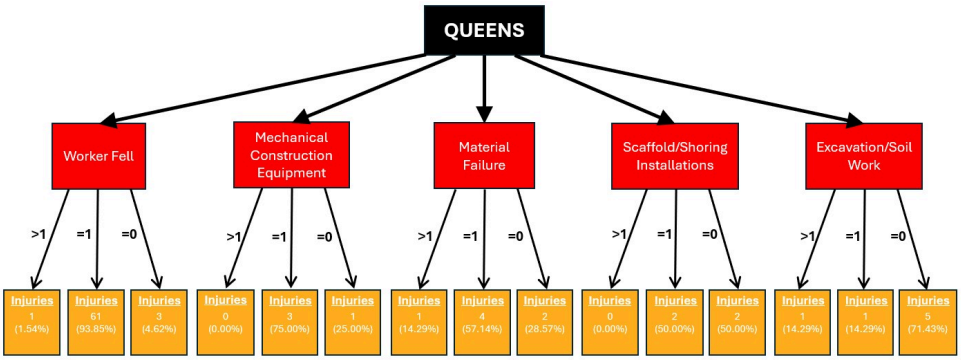

Figure 19: Queens Injury Classification Tree

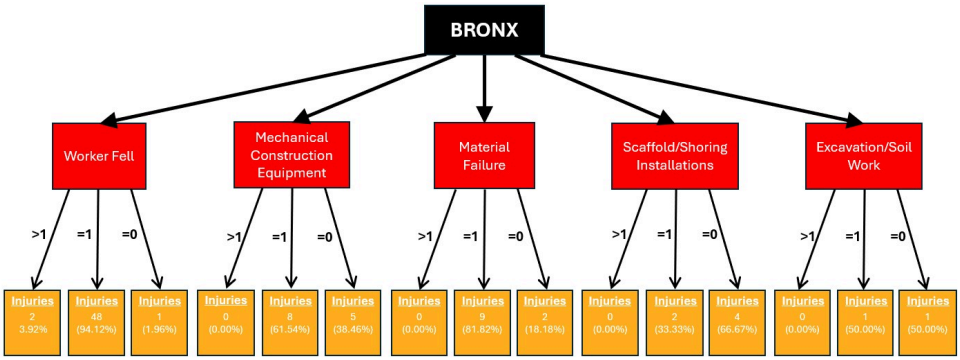

Figure 20: Bronx Injury Classification Tree

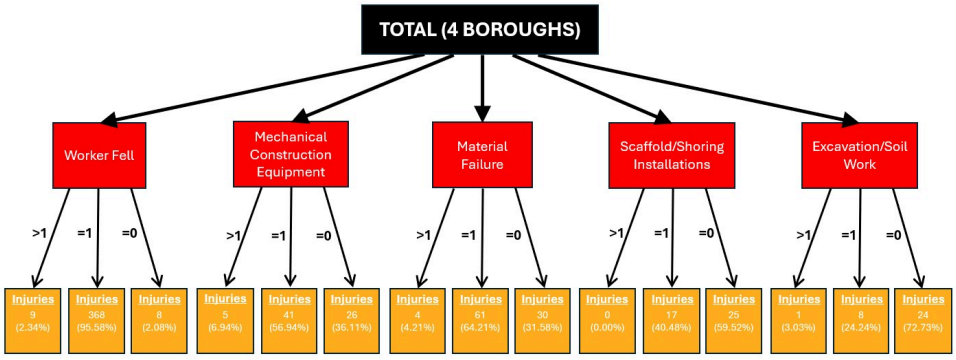

Figure 21: Four Boroughs Total Injury Classification Tree

9.3 Summary

The Classification Tree model identifies which incident types and borough characteristics are most strongly associated with injury outcomes. The results show that Manhattan, Brooklyn, Queens, and

the Bronx each display different dominant contributing factors, indicating that incident type plays a key role in predicting injury likelihood. These insights support targeted regulatory and safety strategies tailored to the needs of each borough.

10. Classification & Neural Network Models & Results

10.1 Methodolgy

To enhance model performance, two additional parameters were integrated:

NoncompliantCount, representing the frequency of non-compliant behaviors ,

IssueNumber, representing the volume of active construction projects in a specific area and month, allowing for temporal lags

10.1.1 Data Preparation

Five input features were selected from the final dataset **(df_final)**: Average Temperature **(AvgTemp)**, Average Precipitation **(AvgPrecip)**, Weighted Heat Vulnerability Index **(HVI_w)**, NoncompliantCount, and IssueNumber. The target variable, Injury, was binarized: samples with an injury count greater than zero were labeled as "Injury Occurred" (1), while others were labeled as 0. To address dimensional discrepancies, a **StandardScaler** was applied to standardize all input features. The dataset was randomly partitioned into a training set (80%) and a validation set (20%).

10.1.2 Model Architecture

A three-layer Feedforward Neural Network (FNN) was adopted as the predictive model. The architecture is defined as follows:

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 structure provides the necessary nonlinear expressive capacity to capture complex interactions among the multivariate inputs.

10.1.3 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, with a learning rate of $1 * 10^{-4}$, is employed to update network parameters and minimize the loss function during each iteration.

10.1.4 Training and Validation

The model was trained for 300 epochs. Loss and Accuracy for both training and validation sets were calculated in each epoch to monitor convergence trends. Dropout was active during training to enhance generalization. Key metrics were logged every 50 epochs. Finally, training/validation loss curves and validation accuracy curves were plotted to assess the stability of model convergence.

10.1.5 Model Evaluation

ROC Curve & AUC: The Receiver Operating Characteristic (ROC) curve was plotted using validation results, and the Area Under the Curve (AUC) was calculated to quantify overall classification performance. Youden's J statistic (TPR − 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.

10.2 Model

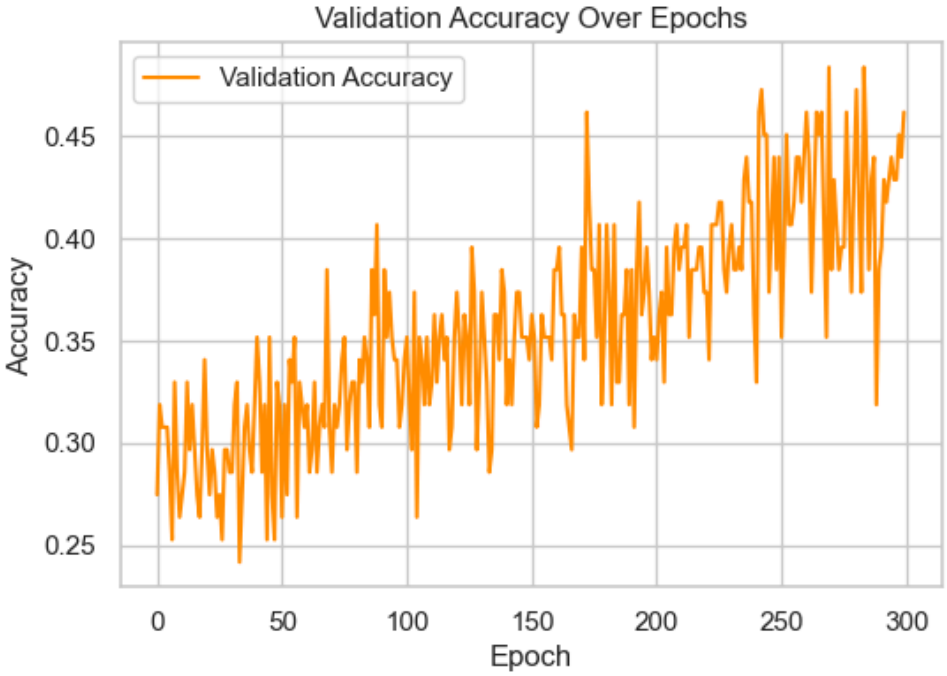

Figure 22:

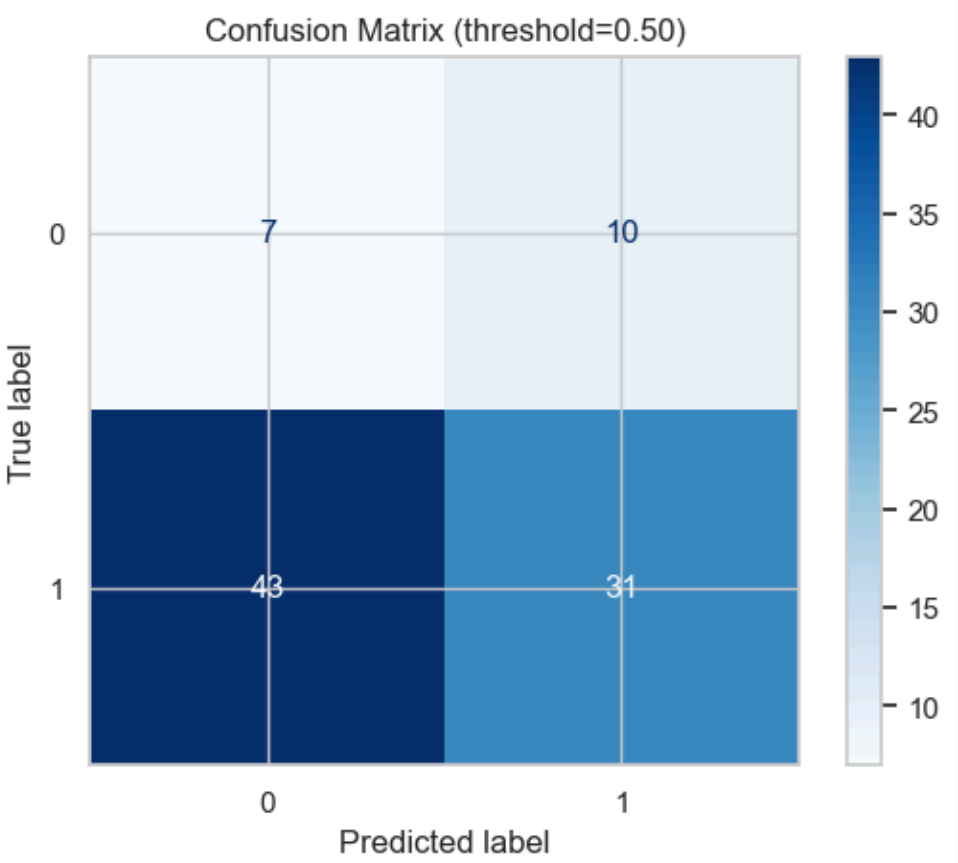
Confusion Matrix (threshold=0.50)
True label
Predicted label
0
1
0
1
7
10
43
31
40
35
30
25
20
15
10

Figure 23:

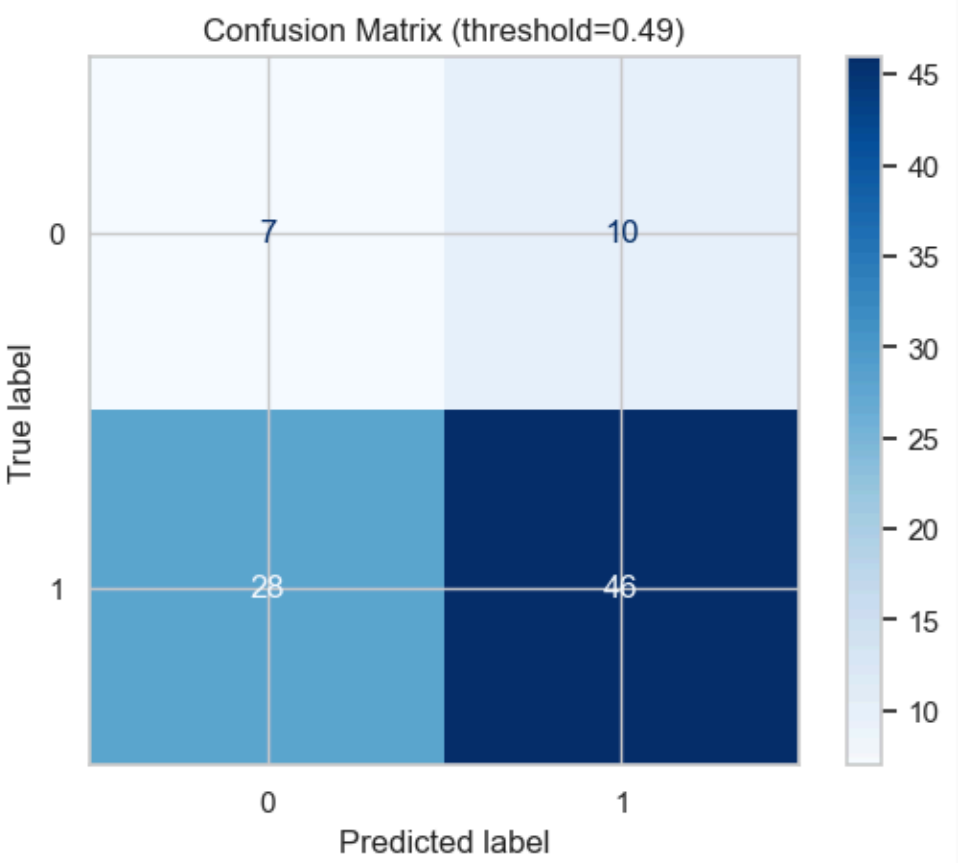
Confusion Matrix (threshold=0.49)
True label
Predicted label
0
1
0
1
7
10
28
46
45
40
35
30
25
20
15
10

Figure 24:

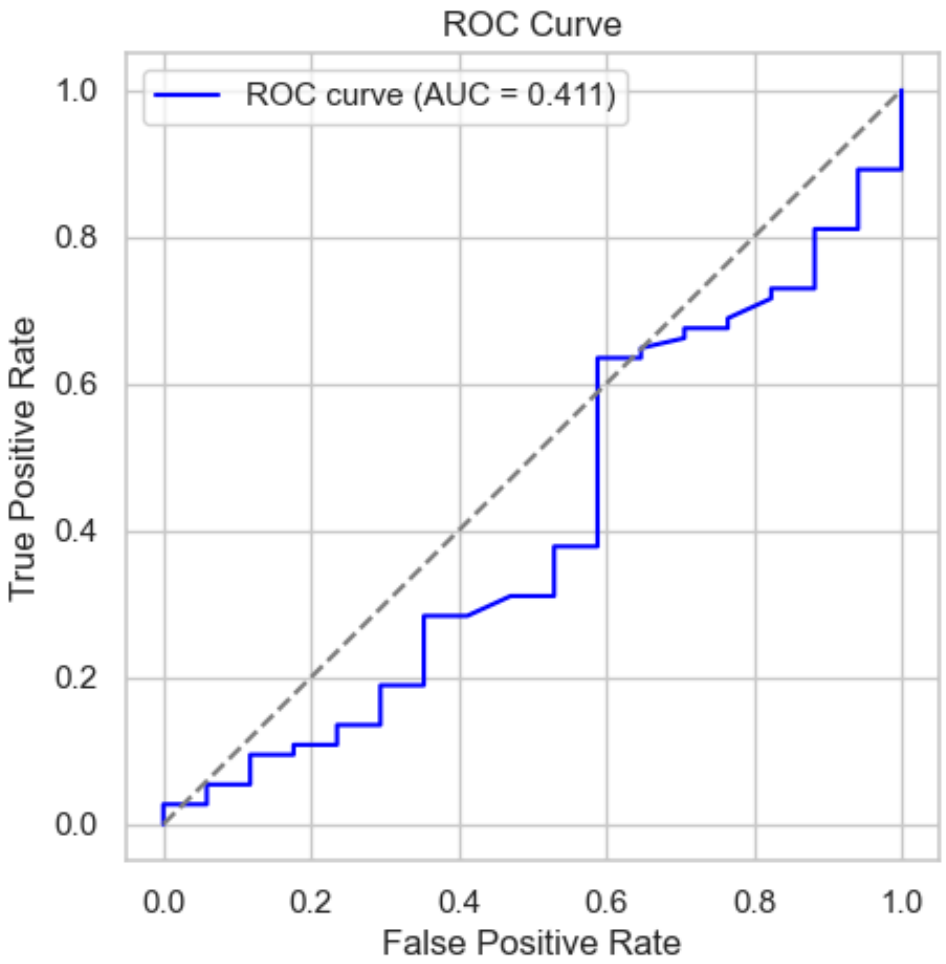
ROC Curve
ROC curve (AUC = 0.411)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 25:

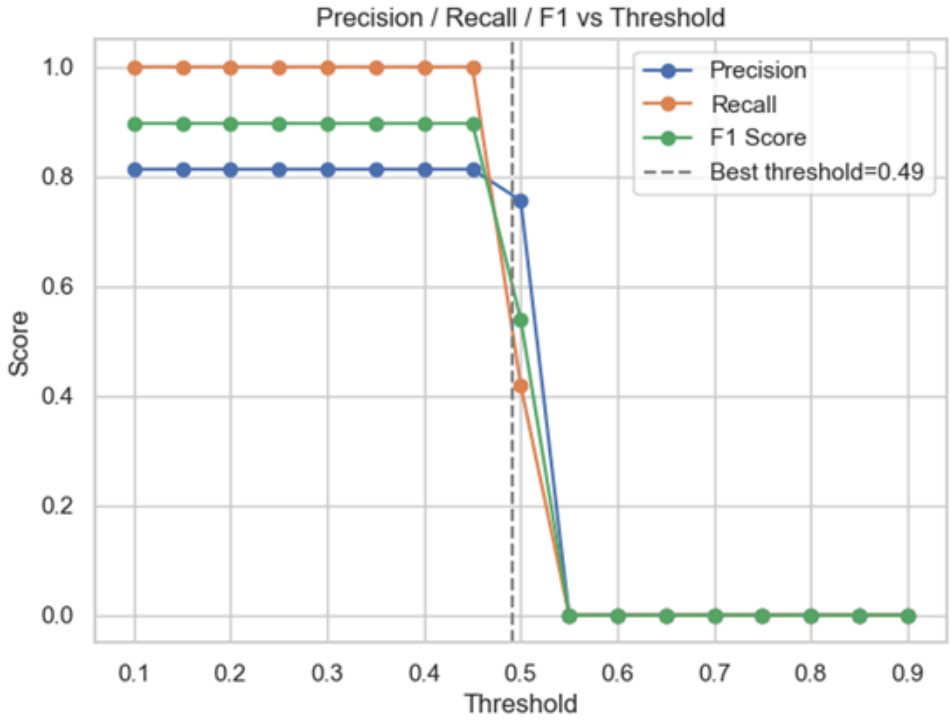

Figure 26:

10.3 Summary & Discussion

10.3.1 Validation Accuracy Over Epochs

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.

10.3.2 Confusion Matrix (Threshold = 0.50)

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.

10.3.3 Confusion Matrix (Threshold = 0.49, Optimal by Youden's J)

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.

10.3.4 Precision/Recall/F1 vs. Threshold

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.

11. Regression and Neural Network Models and Results

11.1 Hypothesis

This study employs an improved Neural Network Regression model to predict the count of construction-related injuries. To enhance prediction stability and robustness, the model incorporates mechanisms for data denoising, standardization, and nonlinear feature extraction.

11.1.1 Data Preparation and Cleaning

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.

Due to the inherent sparsity of the data, additional regularization terms were omitted to avoid further underfitting or gradient convergence issues.

11.1.2 Feature Standardization

Input features comprised Average Temperature, Average Precipitation, Heat Vulnerability Index, Noncompliant Count, Issue Number, Month, and borough encoding columns. All input variables

were standardized. To ensure numerical stability and facilitate gradient convergence, the target variable **(Injury)** was normalized using its mean and standard deviation. The dataset was subsequently partitioned into an 80% training set and a 20% validation set.

11.1.3 Model Architecture

A Neural Network model named InjuryRegressor was defined with a multi-layer nonlinear structure:

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.

LeakyReLU was selected because it maintains non-zero gradients in the negative interval, avoiding the vanishing gradient problem, which is particularly suitable for regression tasks involving sparse data.

11.1.4 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.

11.1.5 Training and Validation

The following metrics were calculated using validation predictions:

$\boldsymbol{R^2}$ **(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.

Additionally, scatter plots of predicted vs. actual values were generated to assess fit; a point cloud clustering near the diagonal indicates good predictive performance.

11.1.6 Model Improvements

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.

11.2 Model

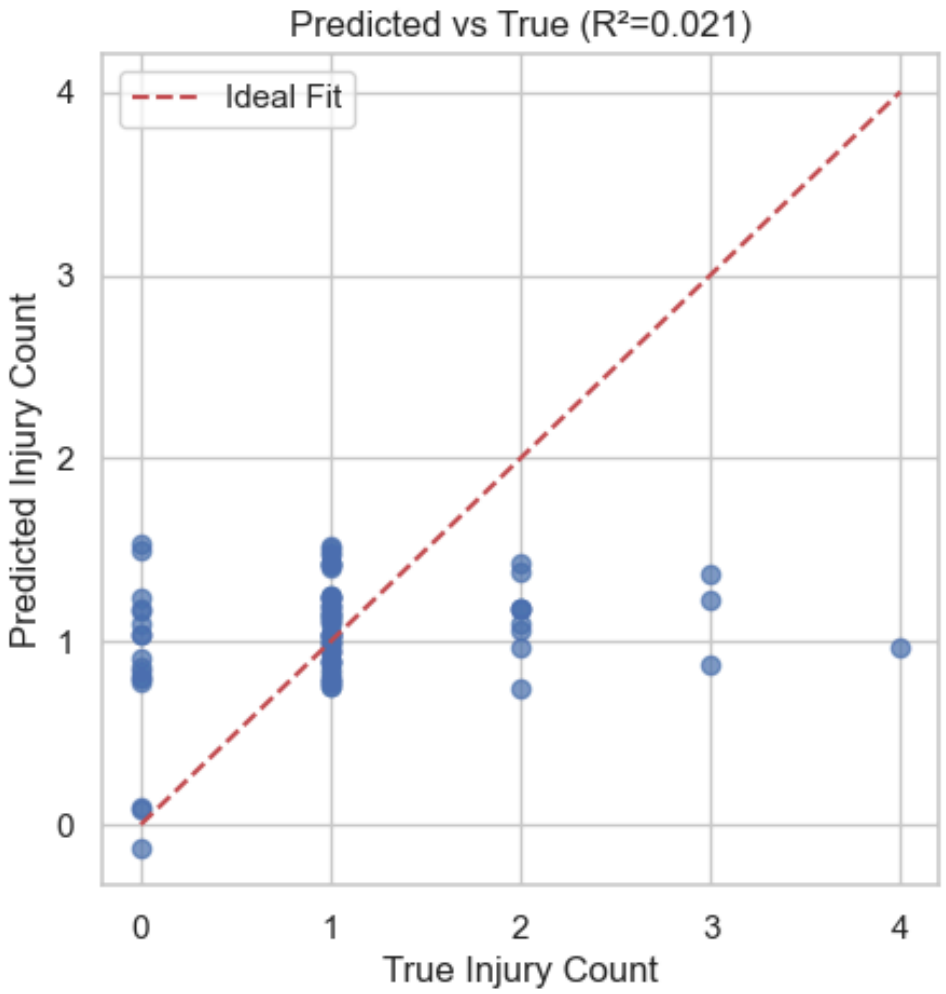

Figure 27:

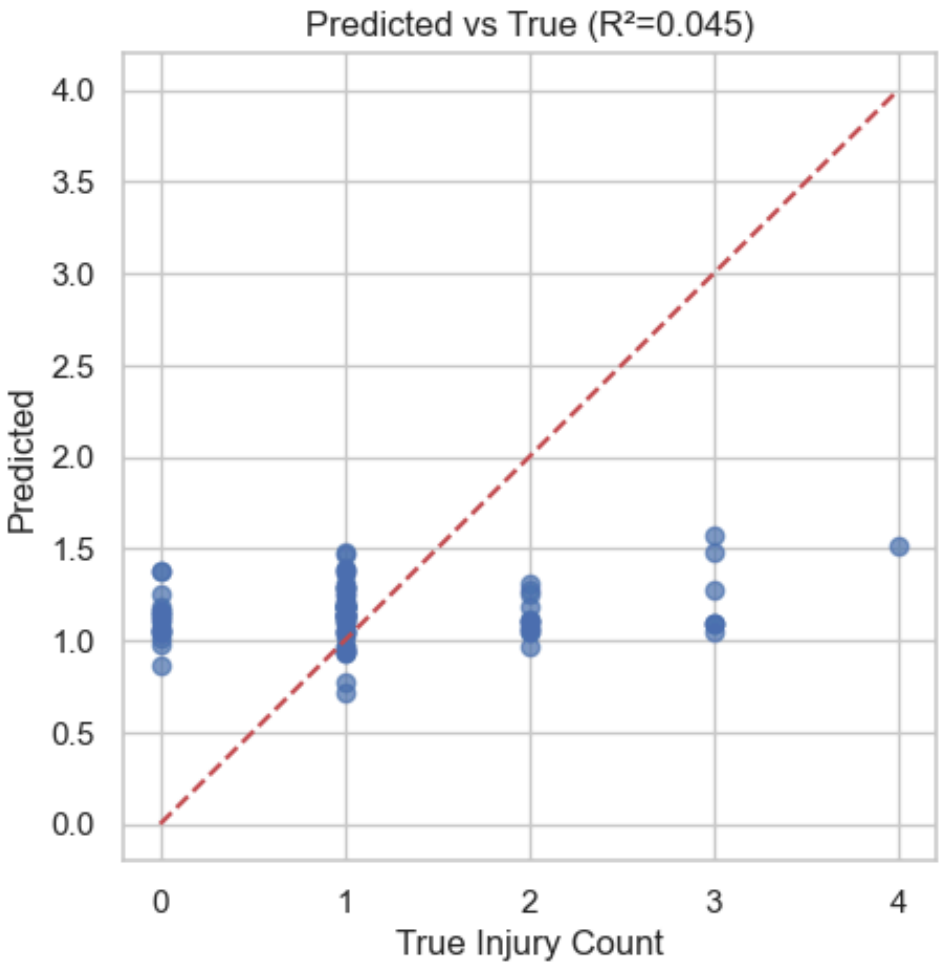
Predicted vs True (R²=0.045)
Predicted
True Injury Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
1
2
3
4

Figure 28:

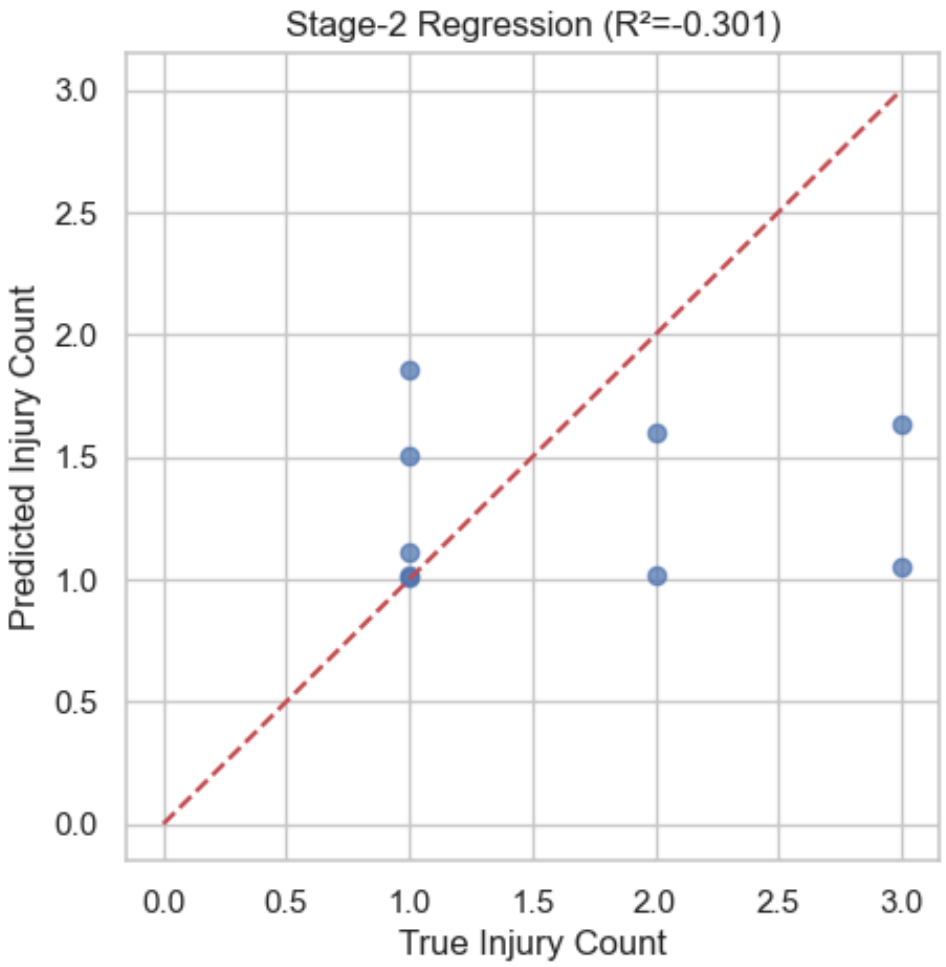

Figure 29:

11.3 Summary and Discussion

This study evaluated three regression approaches: the Neural Network Regressor, the Hybrid Lag and Group Bias Model, and the Two-Stage Hybrid Model. Overall, all three yielded suboptimal results, characterized by negative or near-zero R^2 values, indicating predictive performance weaker than simple baseline averages.

First, the Neural Network Regressor struggled to learn stable patterns. The construction injury data is highly sparse (most months have 0 or 1 injury). Consequently, the network tended to predict near-mean values to minimize error, resulting in negative R^2.

Second, the Hybrid Lag and Group Bias Model failed to improve performance despite introducing lag features. This suggests a lack of significant temporal autocorrelation in construction injury events; lag variables likely introduced additional noise rather than signal.

Finally, while the Two-Stage Hybrid Model performed well in classification, the regression stage suffered from severe data scarcity. After filtering for only positive-injury samples and applying strict denoising, the effective sample size was insufficient for the model to generalize, leaving R^2 negative.

In conclusion, the poor performance is attributed to: (1) high data sparsity and discreteness; (2) a low signal-to-noise ratio; and (3) excessive reduction in sample size due to aggressive cleaning. Future

research should focus on expanding the dataset (more years/regions) and incorporating smoother temporal features or risk indices.

12. References

[1] New York City Department of Buildings. (n.d.). *Incident Database* [Data set].

[2] 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.

[3] Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

[4] Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

[5] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

[6] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[8] 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.

[10] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[11] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

[12] City of New York. (2025). *Official website of the City of New York*. Retrieved November 11, 2025, from https://www.nyc.gov/main

[13] 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