NYC Crash Risk Prediction System: A Spatial-Temporal Machine Learning Approach with Uncertainty Quantification and Interpretable Analytics

Syed Abdul Ahad, Abdul Basit, M. Faozan
SEECS, NUST
syedahad171@gmail.com

Abstract

This paper presents a comprehensive machine learning system for predicting vehicle crash risk across New York City using spatial-temporal modeling. We develop an end-to-end pipeline incorporating H3 hexagonal spatial binning, weather data integration, advanced feature engineering with temporal autocorrelation features, and gradient boosting models. Our XGBoost model with Poisson objective achieves a 19.2% reduction in RMSE compared to baseline, with the 7-day rolling mean feature demonstrating the highest predictive importance due to strong temporal autocorrelation in accident patterns. We implement conformal prediction for uncertainty quantification, SHAP for model interpretability, and deploy an interactive dashboard supporting 6 analytical views. The system addresses the inherent challenge of zero-inflated count data, providing calibrated 90% prediction intervals for operational decision support. Experimental results demonstrate significant improvements over baseline across all models, with ensemble methods achieving the best performance at 20.4% RMSE reduction.

Keywords: crash prediction, machine learning, XGBoost, spatial analysis, H3 hexagons, temporal features, uncertainty quantification, SHAP explainability, zero-inflated data

1 Introduction

Traffic safety represents a critical public health concern in metropolitan areas, with New York City alone recording over 100,000 motor vehicle collisions annually. Accurate prediction of crash risk enables proactive resource allocation for emergency services, targeted infrastructure improvements, and data-driven traffic management policies.

Traditional crash prediction approaches rely on statistical models with limited capacity to capture complex non-linear relationships between environmental, temporal, and spatial factors. Recent advances in gradient boosting methods and spatial indexing systems provide new opportunities for developing high-accuracy prediction systems with interpretable outputs.

This paper presents the NYC Crash Risk Prediction System, an end-to-end machine learning pipeline that addresses three fundamental challenges in crash prediction: (1) efficient spatial aggregation at scale using Uber's H3 hexagonal indexing, (2) capturing temporal autocorrelation patterns through engineered lag features, and (3) providing calibrated uncertainty estimates for operational decision-making.

1.1 Contributions

The primary contributions of this work include:

- A complete spatial-temporal feature engineering pipeline using H3 resolution 7 hexagons with weather integration
- Comparative analysis of gradient boosting models (XGBoost, LightGBM, CatBoost) with Poisson objectives for count data
- Integration of conformal prediction for distribution-free uncertainty quantification
- SHAP-based interpretability analysis revealing the importance of temporal autocorrelation features
- An interactive multi-view dashboard for operational deployment

2 Related Work

2.1 Crash Prediction Models

Crash prediction has evolved from classical Poisson regression and negative binomial models to machine learning approaches. Abdel-Aty et al. demonstrated that neural networks outperform statistical models for real-time crash prediction. Recent work has focused on gradient boosting methods, with XGBoost achieving state-of-the-art performance on transportation datasets.

2.2 Spatial Analysis in Transportation

The H3 spatial indexing system, developed by Uber, provides a hierarchical hexagonal grid that addresses edge effects inherent in rectangular binning. Hexagonal grids ensure equidistant neighbors, making them ideal for spatial autocorrelation analysis.

2.3 Uncertainty Quantification

Standard machine learning models produce point predictions without calibrated confidence intervals. Conformal prediction provides distribution-free coverage guarantees, making it suitable for safety-critical applications where reliable uncertainty estimates are essential.

3 Methodology

3.1 Data Sources

Our dataset comprises over 2 million motor vehicle collision records from the NYC Open Data portal, spanning January 2018 to December 2023. Each record includes timestamp, geographic coordinates (latitude/longitude), contributing factors, and injury counts. Weather data was obtained via the Meteostat API, providing hourly temperature, precipitation, wind speed, and snow depth measurements for the NYC metropolitan area.

3.2 Spatial Processing

We employ H3 hexagonal binning at resolution 7, yielding hexagons of approximately 5.16 km^2 area. This resolution balances spatial granularity with sufficient data density per hexagon. The study area contains 34 active hexagons covering the five NYC boroughs, aggregated to hourly time resolution.

3.3 Feature Engineering

We construct a comprehensive feature set capturing temporal, spatial, and weather-related patterns:

Temporal Features: Hour of day (0–23), day of week (0–6), month (1–12), and binary indicators for weekend and holiday.

Lag Features: accidents_1h_ago (count at same hexagon, previous hour), accidents_24h_ago (count at same hexagon, same hour previous day), and rolling_mean_7d (7-day moving average of hourly counts).

Weather Features: Temperature, precipitation, wind speed, and snow depth.

The total feature dimensionality is 13 features, with all missing values imputed as zero.

3.4 Model Architecture

We evaluate three gradient boosting frameworks optimized for count data prediction:

XGBoost: Extreme Gradient Boosting with Poisson objective, optimizing the negative log-likelihood:

$$\mathcal{L}_{\text{Poisson}} = -\sum_{i=1}^{n} (y_i \log(\hat{y}_i) - \hat{y}_i) \tag{1}$$

Hyperparameters were tuned via RandomizedSearchCV with 10 iterations and 3-fold cross-validation.

LightGBM and CatBoost: Alternative gradient boosting implementations with similar Poisson objectives, providing ensemble diversity.

Stacking Ensemble: A meta-learner combining base model predictions using ridge regression.

3.5 Uncertainty Quantification

We implement conformal prediction for calibrated prediction intervals. Given calibration residuals, the prediction interval at confidence level $1 - \alpha$ is computed using the quantile of calibration residuals.

3.6 Interpretability

SHAP (SHapley Additive exPlanations) values decompose each prediction into feature contributions, enabling both global importance rankings and local explanations.

4 Experimental Results

This section presents the empirical evaluation of our crash risk prediction system. We trained and evaluated models using a temporal split: data prior to September 2023 for training (500,000 samples) and subsequent data for testing (100,000 samples). All experiments were conducted on an Intel Core i7 system with 16GB RAM using Python 3.12 with scikit-learn 1.4, XGBoost 2.0, LightGBM 4.1, and CatBoost 1.2.

4.1 Baseline Comparison

Table 1 presents the comparative performance of all evaluated models. The baseline model predicts the training set mean for all test instances, establishing the lower bound for predictive performance.

Table 1: Model Performance Comparison on Test Set

Model	RMSE	MAE	vs Baseline
Baseline (Mean)	0.4521	0.1823	—
Random Forest	0.3847	0.1456	+14.9%
XGBoost (Tuned)	0.3654	0.1387	+19.2%
LightGBM	0.3689	0.1402	+18.4%
CatBoost	0.3712	0.1418	+17.9%
Stacking Ensemble	**0.3598**	**0.1354**	**+20.4%**

The XGBoost model with tuned hyperparameters achieves an RMSE of 0.3654, representing a 19.2% improvement over the baseline. The optimal hyperparameters obtained via RandomizedSearchCV were: learning rate = 0.1, max depth = 7, n_estimators = 200, subsample = 0.9, and colsample_bytree = 0.9. The stacking ensemble further improves performance to 0.3598 RMSE (20.4% improvement), demonstrating the value of model diversity in gradient boosting frameworks.

4.2 Feature Importance Analysis

Table 2 presents the SHAP-based global feature importance rankings. The rolling_mean_7d feature dominates with 28.3% importance, followed by short-term lag features.

Table 2: SHAP Feature Importance (Top 5)

Feature	Mean $|\phi|$	Importance %
rolling_mean_7d	0.0847	28.3%
accidents_1h_ago	0.0591	19.7%
accidents_24h_ago	0.0456	15.2%
hour_of_day	0.0363	12.1%
temperature	0.0252	8.4%

The predominance of the rolling_mean_7d feature (28.3%) is explained by strong temporal autocorrelation in accident occurrence patterns. Locations with historically elevated accident rates tend to maintain elevated risk, reflecting persistent factors such as road geometry, traffic volume, and infrastructure quality. This autocorrelation structure manifests as a strong linear relationship between past and future accident counts within each hexagon. The 7-day window captures weekly periodicity while smoothing daily fluctuations, making it the most informative predictor of current risk. The immediate lag (accidents_1h_ago, 19.7%) captures real-time risk escalation, such as cascading incidents or adverse weather conditions, while the 24-hour lag (15.2%) reflects daily commute patterns that repeat across consecutive days.

4.3 Limitations: Zero-Inflated Data

A notable characteristic of our dataset is its zero-inflated distribution: approximately 87% of hourly hexagon observations record zero accidents. This reflects the fundamental sparsity of accident events—most locations at most times experience no collisions. While the Poisson objective partially addresses count data properties, standard regression models do not explicitly model excess zeros.

The model exhibits slight systematic bias for zero-count observations, with residual analysis revealing a concentration of small positive errors when true values are zero. This occurs because the model predicts small positive values (the expected rate) for locations where zero accidents actually occurred. Mean residual for $y = 0$ observations is +0.043, while mean residual for $y > 0$ observations is -0.127, characteristic of zero-inflated data.

This limitation is well-documented in count regression literature. Alternative approaches such as zero-inflated Poisson (ZIP) models or hurdle models could potentially improve performance by separately modeling the probability of any accident occurring versus the count given occurrence. However, gradient boosting implementations for these specialized objectives remain limited.

4.4 Uncertainty Calibration

Our conformal prediction intervals achieve 89.7% empirical coverage at the 90% nominal level, demonstrating well-calibrated uncertainty estimates. The average interval width is ±0.187, providing operationally useful confidence bounds for risk assessment.

5 Dashboard Implementation

The operational deployment consists of a Streamlit-based dashboard with six analytical views: (1) Risk Prediction Map with 3D H3 hexagon visualization and confidence intervals, (2) SHAP Explainability with feature importance and local explanations, (3) 24-Hour Forecast for temporal risk projections, (4) Model Performance metrics and residual analysis, (5) Hotspot Cluster Analysis using K-Means, and (6) Scenario Comparison for weather impact assessment.

6 Conclusion and Future Work

This paper presented a comprehensive machine learning system for NYC crash risk prediction achieving 20.4% RMSE improvement over baseline through temporal feature engineering and gradient boosting ensembles. The dominance of the rolling_mean_7d feature confirms the strong temporal autocorrelation structure in accident patterns, suggesting that historical risk is highly predictive of future risk at the same location.

Key limitations include the zero-inflated nature of count data, which causes slight systematic bias in model predictions. Future work should explore zero-inflated regression models, spatial autocorrelation features with neighbor hexagon lag values, and deep learning approaches for sequence modeling.

The deployed dashboard provides actionable insights for traffic safety planners, with calibrated uncertainty intervals enabling risk-aware decision-making.

References

[1] NYC Open Data, “Motor Vehicle Collisions - Crashes,” 2024.

[2] D. Lord et al., “Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes,” Accident Analysis & Prevention, 2005.

[3] S.-P. Miaou, “Truck accidents and road geometry,” Accident Analysis & Prevention, 1994.

[4] M. Abdel-Aty et al., “Crash prediction models with real-time traffic,” Accident Analysis & Prevention, 2005.

[5] T. Chen and C. Guestrin, “XGBoost,” ACM SIGKDD, 2016.

[6] I. Brodsky, “H3: Uber’s Hexagonal Spatial Index,” 2018.

[7] V. Vovk et al., Algorithmic Learning in a Random World, Springer, 2005.