

# Predicting Auto Insurance Risk Using Gradient Boosting


---

Analyzing Socio-Economic Factors in Car Crashes for New York City

Author: AJ Strauman-Scott  
City University of New York Master's Capstone Project  
July 20



- 



# Research Questions

- Which socio-economic, transportation, and demographic factors are most predictive of crash risk?
- What quantity of risk can be predicted solely by these factors using a XGBoost model?



# Data and Sources

ACS 5-year estimates (2018–2023)

- Variables: Demographics, income, median gross rent, education, employment, transportation.
- Engineered features: Poverty × vehicle ownership, unemployment × vehicle ownership.

NYC Motor Vehicle Collisions (MVC) Open Data Portal

- Aggregated to 2020 Census Tracts.
- Normalized as **crashes per 1,000 residents** (proxy for claim frequency).

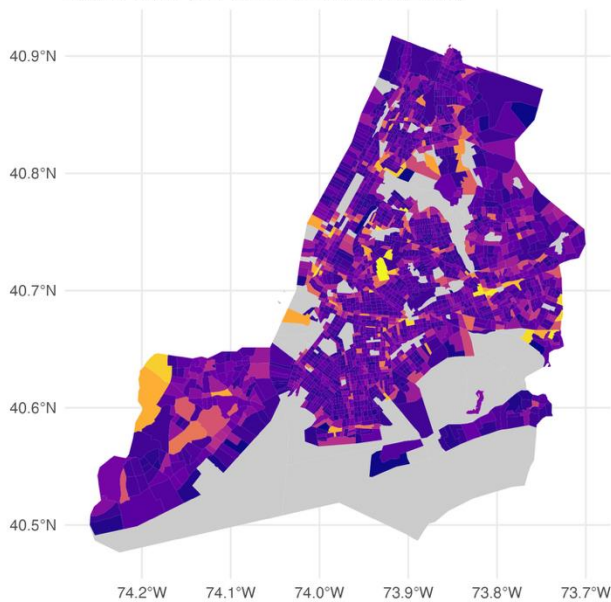
## PREPROCESSING:

Data cleaning: Removal of invalid coordinates, spatial join with 2020 census tracts.  
Harmonization of ACS variables, binning, and removal of highly correlated features.  
Log transformation of crash counts to reduce variance.

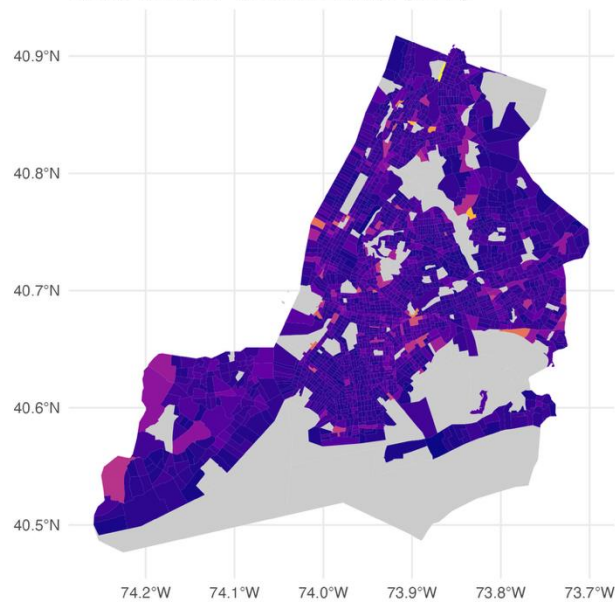




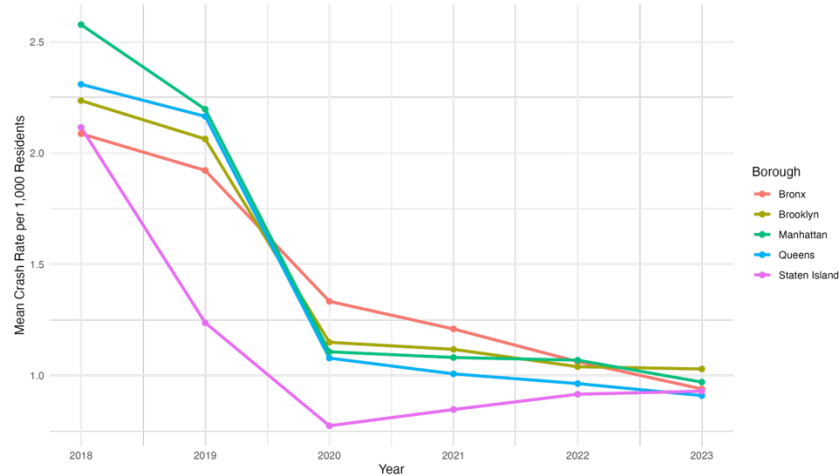
Crash Rate per 1,000 Residents (2018)



Crash Rate per 1,000 Residents (2022)



Annual Mean Crash Rate per 1K Residents by Borough



## Descriptive Statistics

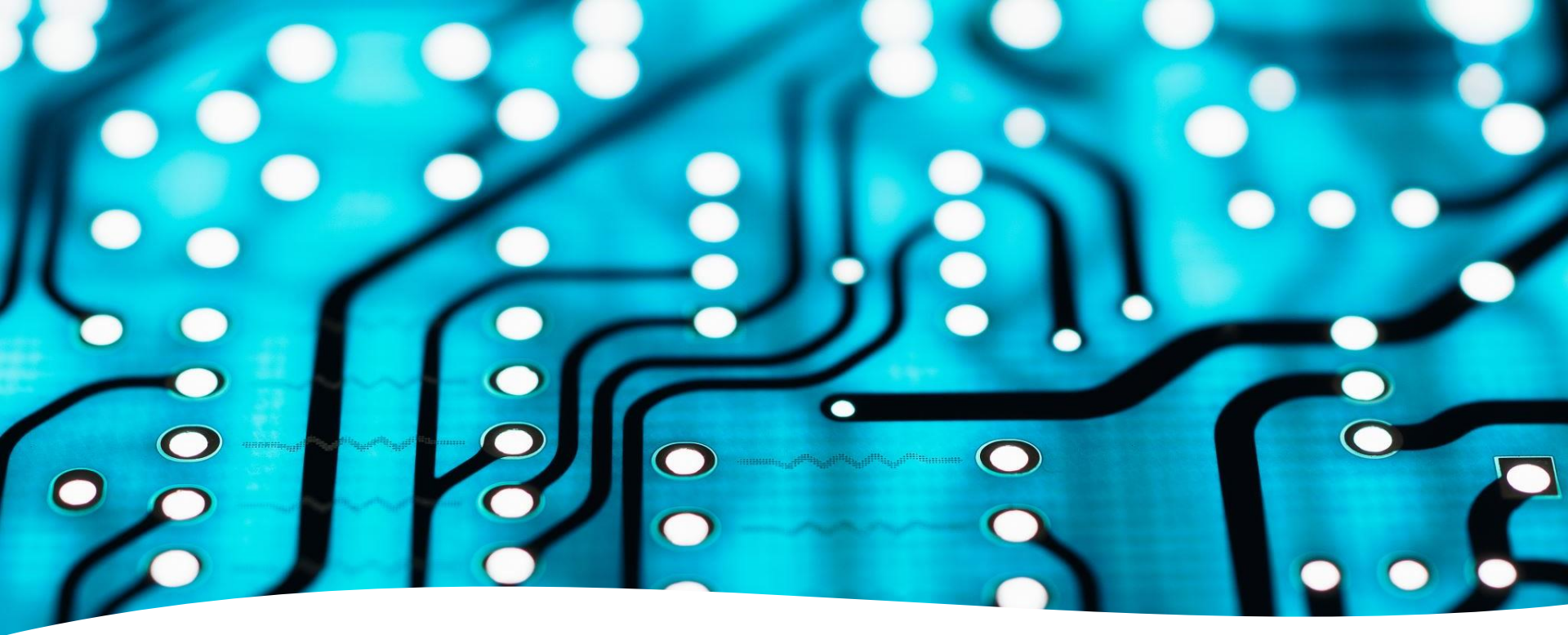
**Average crash rate:** 1.43 crashes/1,000 residents (max > 23/1,000)

- Crash rates declined during the pandemic (2020) with gradual rebound.
- Hotspots: South Bronx, central Brooklyn, northern Manhattan.

**Median household income:** \$74,681

**Median gross rent:** \$1,589





# Model Framework

**Algorithm:** XGBoost, chosen for non-linear modeling and strong predictive performance in insurance contexts.

- SHAP values used to interpret variable contributions and interactions.

## Optimization and Validation

- **Hyperparameter Tuning:** Conducted via Optuna (Bayesian optimization) for optimal learning rate, tree depth, and regularization.
- **Cross-Validation:** Spatial cross-validation by borough to prevent geographic overfitting.



# Results

## Performance:

$R^2 = 0.43$ , RMSE = 0.34, MAE = 0.24 – strong accuracy for socio-economic and transportation predictors alone (no behavior or weather).

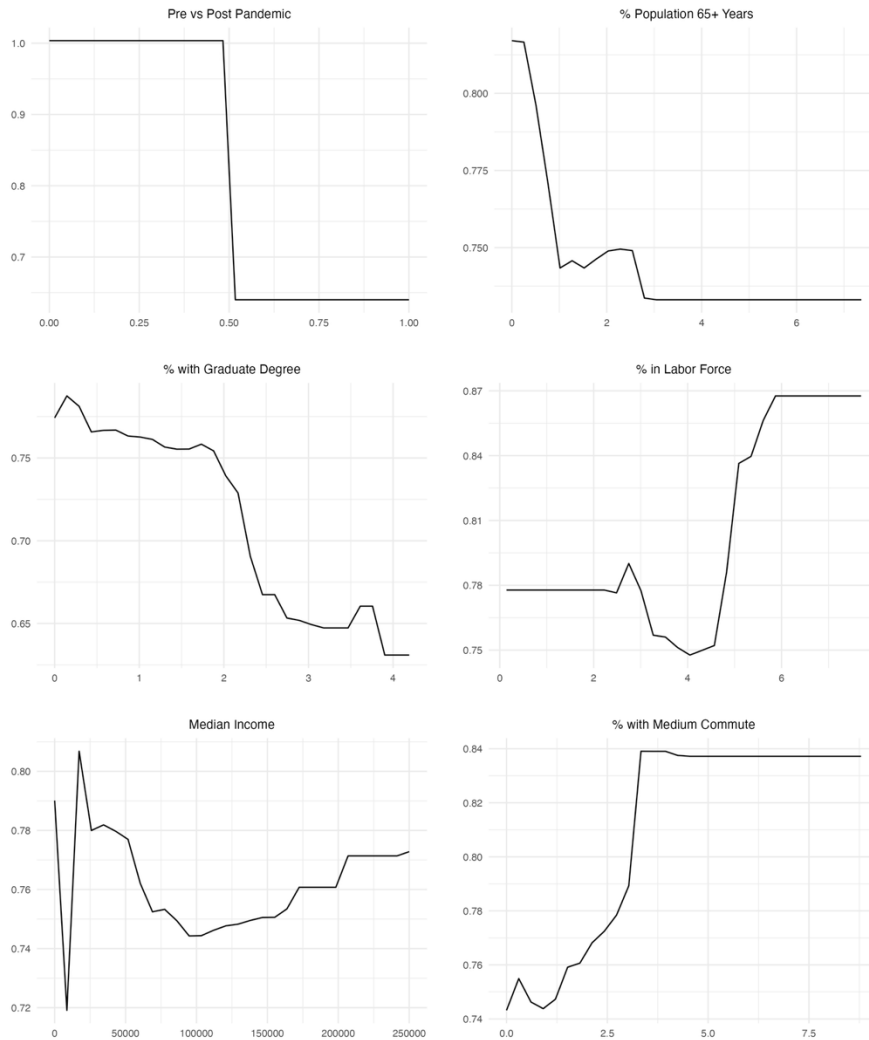
## Crash Trends:

- Citywide crash rates declined sharply during 2020 (COVID-19) and stabilized afterward.
- High-risk hotspots persist in central Brooklyn, South Bronx, and northern Manhattan.

## Top Predictors (SHAP):

- Post-pandemic indicator (sharp drop in predicted risk after 2020).
- Median income (U-shaped risk relationship: higher risk in very low/high-income areas).
- Labor force participation & medium commute share (higher commuter exposure = higher risk).
- Graduate degree share (negative association with crash risk).

Partial Dependence Plots for Key Features





## Conclusions

- Crash risk is strongly tied to socio-economic context and mobility behaviors – can predict ~40% of crashes without driver or weather data with XGBoost.
- Commuting intensity, neighborhood income distributions, wealth distribution and Covid-19 effects reflect structural differences in traffic volume and pedestrian safety.
- The model is suitable for regional risk analysis and safety planning, but *not individual underwriting*, due to fairness and ethical concerns.

