Introduction

Over the past five years, the city of Toronto has experienced a significant increase in auto theft, making it a growing concern for public safety. While the Toronto Police have created the Provincial Carjacking Joint Task Force (PCJTF) in partnership with the Ontario Provincial Police (OPP) in October 2023, the overall number of incidents continues to rise at an alarming rate. Addressing this problem is challenging due to limited resources, and it demands smarter approaches driven by careful planning, accurate data, and effective prediction models. As a lucrative crime that can involve organized groups and lead to serious insurance fraud, auto theft's impact extends beyond individual vehicle owners, raising important questions about safety and community well-being, Toronto.com (2024).

The dataset used in this analysis provides a detailed record of 63,633 confirmed auto theft occurrences in Toronto between 2014 and June 2024, retrieved from the Toronto Police Service's open data portal, Toronto Police Service (2024). It includes 31 columns with information on the nature of each offence, such as the date, day of the week, and month it occurred, along with the exact neighborhood, nearby intersections, and the responding police station. Although the location details are approximate to preserve privacy, these rich data points help paint a clearer picture of the factors driving the increase in auto theft.

To better understand the underlying patterns and trends, this report will focus on comparing two modeling approaches, a Generalized Linear Model (GLM) and a Random Forest model, to predict auto theft occurrences in the top five affected areas for the year 2025. This proactive, data-driven approach is essential, especially since current policies often focus on reacting to incidents rather than preventing them. Efforts at various levels—including federal and provincial initiatives—have shown some positive outcomes, such as a 17% reduction in early 2024 (as reported by the Equite Association), and the issue was highlighted at the National Summit on Auto Theft in February 2024. Ultimately, finding the best predictive model will support more intelligent resource allocation and planning, helping the city move towards proactive and effective solutions.

Exploratory Data Analysis

The dataset consists of 63,633 police reports detailing individual offenses, encompassing 31 variables. These include two datetime fields (Report Date and Date of Occurrence), seven float64 fields (such as coordinates, year, day of the year, day of the week, latitude, and longitude), eight int64 fields (e.g., unique ID, report year, day, hour, and police code), and 14 object fields (including neighborhood names, offense categories, and location details). Initially, neighborhoods were classified into 140 codes, but this number later increased to 158. As a result, columns "Hood 150" and "Hood 140" reflect this transition and were consolidated for clarity.

In preparing the data for analysis, seven columns—EVENT_UNIQUE_ID, OBJECTID, UCR_CODE, UCR_EXT, OFFENCE, MCI_CATEGORY, and the X and Y coordinates—were removed due to redundancy or irrelevance to neighborhood-focused analysis. Missing occurrence dates and times were replaced with the corresponding report dates to ensure data completeness. Additionally, 174 rows containing "NSA" values, signifying non-specific areas outside Toronto, were dropped to maintain a clear focus on the city's

neighborhoods.

Exploratory data analysis reveals a sharp rise in auto thefts over the last five years, from 3,617 cases in 2018 to 4,797 in recent years, indicating a loss of control despite recent interventions, such as the establishment of a specialized police task force in mid-2023. This trend underscores the importance of the project in identifying patterns and high-risk neighborhoods for more effective resource allocation, Min the World (2024).

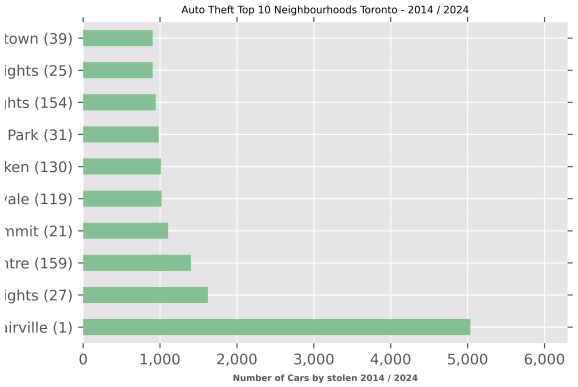

Figure 1: Auto Theft Top 10 Neighbourhoods

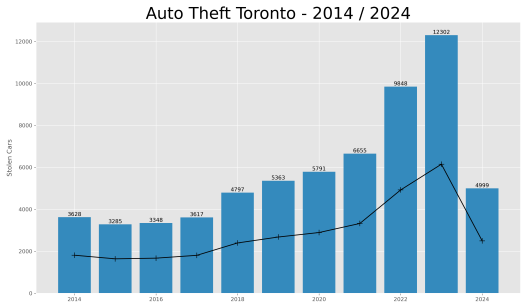

Figure 2: Auto Theft in Toronto by Year

Data Engineering

After streamlining the dataset by removing irrelevant variables and addressing missing values, the next step involved grouping the data. Records were aggregated by occurrence year, occurrence month, police division, premises type, and neighborhood. This process created a new response variable, Count, representing the total number of thefts for each unique combination of these factors.

At this stage, the dataset consisted of six variables: two numeric and four categorical. The numeric variables were OCC_YEAR and Count, while the categorical variables included OCC_MONTH, DIVISION (identifying police stations), PREMISES_TYPE (classifying incident locations, such as houses or commercial properties), and NEIGHBOURHOOD_158.

To prepare the data for modeling, the categorical variables were converted into dummy variables. This step transformed each categorical category into a binary indicator (0 or 1), making the data suitable for a wide range of predictive techniques. After these data engineering procedures, the final dataset consisted of 190 integer variables and 28,027 observations, providing a structured foundation for subsequent model fitting.

Methods

To achieve the analysis goals, two modeling approaches were chosen for comparison: a generalized linear model (GLM) with a negative binomial link function and a random forest regression model. Both methods offer distinct mathematical frameworks and are well-suited to uncovering patterns in count-based data such as auto theft incidents.

1. Generalized Linear Model with a negative binomial link function
   The GLM with a negative binomial link function is particularly adept at modeling overdispersed count data, where the variance exceeds the mean. Let Y represent the number of theft incidents, and let $\mathbf{X}$ be the vector of predictors. The expected value of Y, denoted by μ, is related to the predictors through a linear combination:

   $$\log(\mu) = \mathbf{X}\boldsymbol{\beta}.$$

   Here, $\boldsymbol{\beta}$ is a vector of coefficients, and $\mu = \exp(\mathbf{X}\boldsymbol{\beta})$. Unlike a Poisson model, the Negative Binomial model introduces an extra dispersion parameter θ to account for overdispersion, where the variance of Y is given by:

   $$\text{Var}(Y) = \mu + \frac{\mu^2}{\theta}.$$

   This form allows the model to better capture variability and growth patterns. By including a quadratic term for the OCC_YEAR predictor, OCC_YEAR^2, the model can approximate exponential growth in theft incidents, improving its ability to reflect rapid changes in crime trends.

2. Random Forest Regression Model
   In contrast, the Random Forest regression model approaches the problem by aggregating the predictions of multiple decision trees. Let $\{\hat{y}_b(\mathbf{X})\}_{b=1}^{B}$ be the predictions from B independent trees, each built on a bootstrap sample of the training data and a random subset of features. The final prediction is the average of all tree predictions:

   $$\hat{y}(\mathbf{X}) = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_b(\mathbf{X}).$$

   This ensemble approach naturally models complex, nonlinear relationships and reduces the risk of overfitting, as the averaging process stabilizes the overall prediction. The inclusion of diverse categorical features as dummy variables is straightforward, and the model's robustness to outliers and noise makes it well-suited for this scenario.

   A 10-fold cross-validation grid search was performed to identify the optimal hyperparameters for the Random Forest model. These included the number of trees (n_estimators $= 200$), the depth of the trees (max_depth $=$ None), the minimum number of samples per leaf (min_samples_leaf $= 4$), the

minimum number of samples required to split an internal node (min_samples_split $= 10$), and a fixed random seed (random_state $= 42$) for reproducibility. These selections help the Random Forest model efficiently navigate the high-dimensional feature space and produce accurate predictions.

3. Evaluating Categorical Variables

   To assess the importance of categorical details, three scenarios were tested. The first included all dummy variables, fully preserving data complexity. The second scenario removed dummy variables derived from the DIVISION attribute, slightly increasing the Mean Squared Error (MSE) and indicating that division-level information contributes predictive value. The third scenario excluded dummy variables associated with the PREMISES_TYPE attribute, resulting in a substantial MSE increase. This outcome underscored the importance of maintaining comprehensive categorical detail.

   Ultimately, retaining all dummy variables proved beneficial. The dataset's size was sufficient to accommodate them without risking overfitting, and removing these variables did not improve model performance. This balanced approach ensures that both the Negative Binomial GLM and the Random Forest model have ample information to reveal meaningful patterns in auto theft incidents.

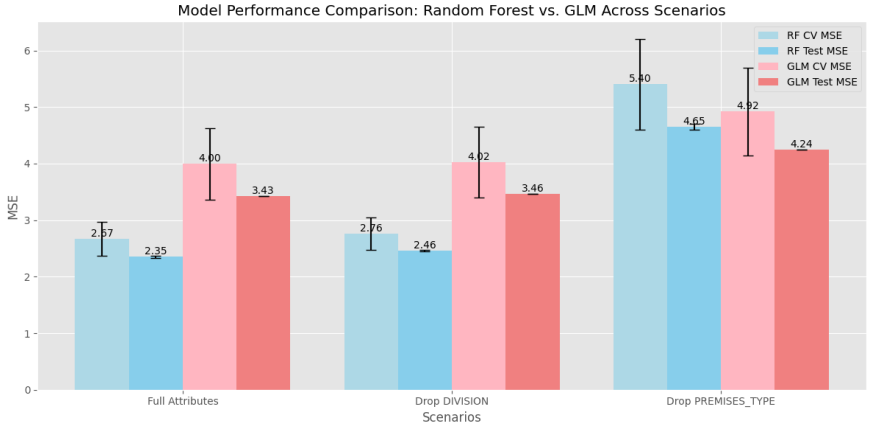

Figure 3: Comparison of the Importance of Categorical Variables Across Models

Results

To evaluate the performance of the Generalized Linear Model (GLM) and the Random Forest regression model, a 10-fold cross-validation approach was applied. The results demonstrated that the Random Forest regression consistently outperformed the GLM model in terms of Mean Squared Error. The Random Forest model achieved the lowest Mean Squared Error (MSE) and showed a smaller performance gap between the training and test datasets, highlighting its robustness and reliability for this analysis.

The Random Forest model showed superior performance with a cross-validation MSE of 2.6689 (standard deviation: 0.2985) and a test set MSE of 2.3517 (standard deviation: 0.0180). In contrast, the GLM model had a higher cross-validation MSE of 3.9967 (standard deviation: 0.6313) and a test set MSE of 3.4302 with no observed variance. These results confirm that the Random Forest model is more effective at

capturing the underlying patterns in the dataset.

In the Prediction Scatter Plot, the Random Forest model demonstrates predictions that closely align with the reference line ($\mathbf{y} = \mathbf{x}$), indicating a higher degree of accuracy. Conversely, the GLM model struggles to predict higher counts of thefts, as evidenced by its deviations from the reference line for larger values.

In the Residuals Plot, the Random Forest model exhibits a tighter clustering of residuals around the zero line, reflecting its superior prediction accuracy compared to the GLM model. The GLM residuals display larger deviations, particularly for higher counts, suggesting that the GLM fails to fully capture the exponential growth trends in the data.

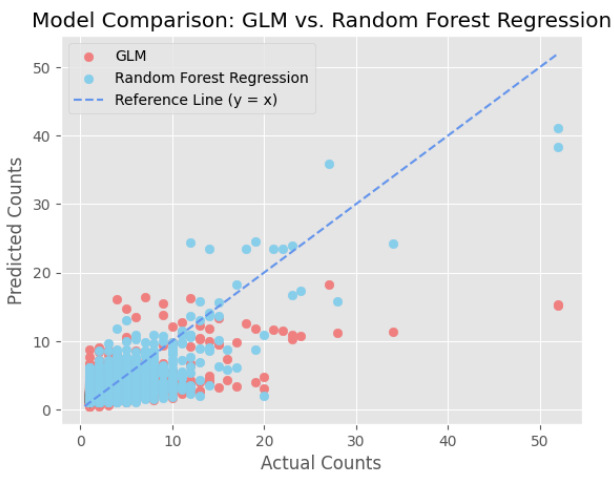

Figure 4: QQ Plot: GLM vs. Random Forest

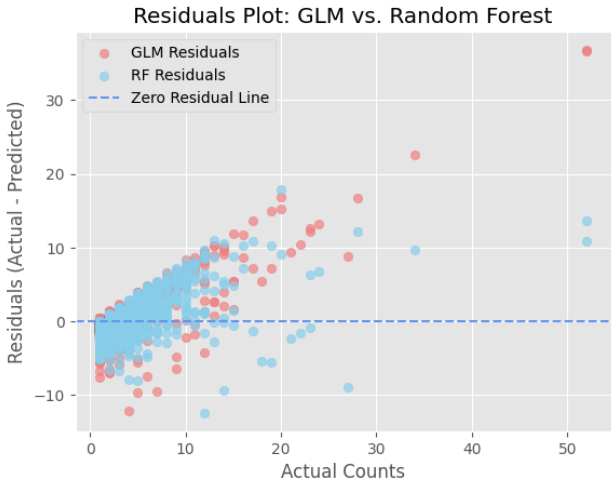

Figure 5: Residual Plot: GLM vs. Random Forest

Most predictions for both models are clustered around lower values, which is expected given the count distribution of theft incidents. However, the GLM model struggles to account for the nonlinear and exponential growth patterns observed in theft incidents, making it less effective at predicting higher counts. In contrast, the Random Forest model more effectively handles the nonlinearity in the data, providing more reliable predictions, as evidenced by its tighter residual distribution and lower MSE values. These results highlight the Random Forest model's robustness and suitability for this analysis.

Discussion

The analysis reveals that the Random Forest regression model is better suited for predicting auto theft trends in Toronto compared to the Generalized Linear Model (GLM). This superiority stems from its ability to effectively handle nonlinear relationships and provide more precise predictions, especially for lower theft counts, where variability in the data is significant. Its robust performance highlights its reliability for capturing complex patterns in auto theft incidents and underscores its potential for guiding data-driven decision-making. The model identified the top five most affected areas as West Humber-Clairville, York University Heights, Etobicoke City Centre, Humber Summit, and Wexford/Maryvale, which represent the neighborhoods experiencing the highest rates of auto theft.

To further enhance the predictive capabilities of the models, incorporating additional variables could prove beneficial. Socioeconomic indicators, economic trends, or seasonal factors that influence auto theft rates could provide deeper insights into the drivers of theft incidents and improve model accuracy. Additionally, developing real-time prediction systems would enable immediate analysis of theft trends, allowing law enforcement to adopt proactive strategies and reduce response times, ultimately curbing the occurrence of thefts.

The insights generated by the Random Forest model also carry important policy implications. By identifying high-risk areas such as the top five neighborhoods, law enforcement agencies can allocate resources more effectively, targeting the regions most affected by auto theft.

In conclusion, the Random Forest regression model is the recommended approach for predicting auto theft trends in 2025. Its superior performance ensures smarter resource allocation and more effective preventive strategies.