MGSC 661-075

Multivariate Statistical Analysis

December 6th, 2025

From Reactive to Proactive: Predictive Modeling to Reduce Severe Traffic Collisions in Montreal

Adeleye, Ibukun

Dong, Rachelle

Ha, Ellie

Houle, Zacharie

Hydarali, Muhammad

DESAUTELS

Faculty of Management

Faculté de gestion

Master of Management in Analytics

Introduction & Problem Context

Between 2012 and 2021, Montreal recorded an average of 21 traffic deaths per year *(SAAQ, 2023),* underscoring the need for more proactive road-safety strategies. However, most existing analyses remain reactive, focusing on explaining crash severity after incidents occur. Prior studies, including Muktar & Fono (2024), identify broad factors such as weather and road conditions but do not pinpoint where or when severe collisions are most likely, limiting their usefulness for prevention. (A comparison with past research is provided in Appendix 1).

Our project addresses this gap by taking a predictive and spatial approach. Instead of analyzing severity outcomes alone, we model the likelihood of severe or fatal collisions across Montreal, identifying the specific neighborhoods, time periods, and contextual conditions that elevate risk. This shift from descriptive analysis to predictive modeling allows us to localize risk more precisely and highlight actionable "red zones."

The value of this approach extends beyond research insight. By revealing the spatial and temporal patterns that drive collision risk, our findings can help the City of Montreal target infrastructure improvements, allocate enforcement and emergency resources more effectively, and better protect vulnerable road users. In doing so, the project supports a proactive strategy with the potential to meaningfully reduce severe and fatal collisions citywide.

Hypotheses Revisiting

Our initial hypotheses focused on geospatial and infrastructure factors as potential drivers of severe crashes, reflecting gaps in prior literature that emphasized broad, reactive explanations rather than predictive, location-specific insights. We anticipated that neighborhood characteristics, road surface conditions, and road geometry would meaningfully differentiate severe from non-severe collisions.

- **Hypothesis 1: Neighborhood-Specific Risks**: We expected certain neighborhoods, particularly dense, mixed-use, or transit-heavy zones, to show higher severe-crash rates. EDA strongly confirmed this pattern: the top 10 highest-risk neighborhoods accounted for 26.4% of all severe collisions (542 of 2,049 cases), and neighborhood-level variables ranked among the most important predictors. This hypothesis is therefore retained.
- **Hypothesis 2: Cracked Road Conditions**: We hypothesized that road cracking would increase severity through reduced traction and stability. However, road-defect indicators appeared in fewer than 5% of severe cases and showed negligible predictive value. This hypothesis was rejected.
- **Hypothesis 3: Flat Curved Road Geometry:** We predicted higher severity on flat curved roads due to misjudged turning behavior. EDA and feature-importance analysis did not support this relationship. Instead, temporal variables, particularly night-time and rush-hour conditions, showed substantially greater predictive power.

Based on these findings, Hypotheses 2 and 3 were replaced with:

- **New Hypothesis 2**: Severe crashes are more likely among vulnerable road users (e.g., pedestrians, cyclists), particularly in areas lacking protected infrastructure.
- **New Hypothesis 3**: Severe crashes are more likely during night-time and rush-hour periods on high-traffic streets.

Dataset description

The dataset used in this project comes from Données Québec and the Société de l'assurance automobile du Québec (SAAQ), containing over 200,000 reported collisions in Montréal from 2012–2021 (with updates to 2022). Each record includes temporal, geospatial, environmental, and infrastructural attributes, such as

time of day, road surface, lighting, weather, speed limit, and collision type, supporting a comprehensive assessment of crash conditions. Collision severity is encoded across four levels (property damage, minor injury, serious injury, fatal injury), later consolidated into a binary target variable (Severe/Fatal vs. Non-Severe). A complete list of variables is provided in *Appendix 2*.

a. Data Cleaning and Structural Refinements

Initial raw features (71 columns) included mixed formats, missing or unknown values, French-coded labels, and highly granular geographic identifiers that introduced noise. Eleven columns with excessive missingness (>50%) were removed, and all variables were standardized into English. Categorical features were encoded, continuous features normalized, and timestamps processed into usable temporal indicators.

A major refinement involved replacing postal codes with neighborhood identifiers, addressing issues identified in the PoC, where postal codes created noisy patterns, contributed to model overfitting, and produced non-actionable insights. Neighborhoods provided larger spatial units, clearer patterns, and better alignment with municipal planning, ultimately emerging as some of the most predictive variables in the entire dataset.

b. Feature Engineering and Balancing

Transformation steps reduced the dataset from 71 raw variables to 22 model-ready features, integrating time-based indicators (hour of day, day of week, rush-hour, night-time), target-encoded geographic risk scores (neighborhood, borough, municipality), and fully mapped environmental variables (weather, road surface, lighting, and location type). All transformations were guided by the official SAAQ data dictionary (2025) to ensure consistency with the source definitions of *HR_ACCDN, JR_SEMN_ACCDN, CD_COND_METEO, CD_ETAT_SURFC, CD_ECLRM, and CD_LOCLN_ACCDN*.

Because severe and fatal collisions represent only ~3% of reported incidents, and because the SAAQ explicitly states that numerous variables (including weather and victim counts) are systematically missing for "dommages matériels seulement" crashes, we applied SMOTENC during training. This oversampling technique preserves the categorical structure of variables such as WEATHER, SURFACE, and LOC_TYPE while generating synthetic severe cases that respect the original feature distributions.

c. Exploratory Data Analysis

Prior to modeling, we conducted exploratory analysis to identify patterns in collision severity across geographic, temporal, and severity dimensions. This analysis directly informed our feature engineering strategy and hypothesis refinement.

Extreme Class Imbalance: The severity distribution (Figure 1) revealed a critical modeling challenge: among 131,200 collisions, only 2,049 (1.56%) were classified as severe (*Grave*: 1,786) or fatal (*Mortel*: 263), while 83,226 (63.4%) resulted in property damage only and 45,925 (35.0%) caused minor injuries. This 64:1 imbalance between non-severe and severe outcomes necessitated SMOTENC oversampling during training to prevent the model from defaulting to majority-class predictions.

Geographic Concentration: Analysis of 91 neighborhoods revealed substantial spatial clustering of severe collisions (Figure 2). The top 10 highest-risk neighborhoods—René-Lévesque (118 severe collisions), Sainte-Marie (68), Savane (49), Vieux-Montréal (47), and Mile End (45)—accounted for 542 severe collisions (26.4% of all severe cases), a disproportionate concentration relative to their geographic footprint. This pattern validated our hypothesis that neighborhood-level characteristics are primary severity drivers, directly informing the "red zones" intervention strategy.

Temporal Risk Patterns: Hourly distribution analysis (Figure 3) identified two distinct high-risk windows: an afternoon-evening peak from 14:00–18:00 (averaging 140+ severe collisions per hour) and elevated overnight risk. Day-of-week analysis showed Friday (326 severe collisions), Thursday (313), and Monday (310) as the highest-risk days, with weekdays (58,122 total collisions) exhibiting higher severity rates than weekends (25,413 collisions). These findings confirmed temporal variables as essential predictive features and shaped recommendations for time-based enforcement strategies.

d. Feature Selection

Three complementary methods (ANOVA F-tests, Random Forest Gini importance, and permutation importance) were applied to rank all candidate features. Each method was chosen for its specific strength: ANOVA for statistical significance across categorical groups, Gini importance for capturing non-linear relationships in tree-based models, and permutation importance for model-agnostic validation that is robust to multicollinearity.

The results exhibited remarkable consensus: the same 15 features consistently ranked in the top tier across all three techniques. Location-based geographic variables (*QUARTIER, ARRONDISSEMENT, MUNICIPALITE*) and their target-encoded risk scores (*QUARTIER_RISK, ARROND_RISK, MUNCP_RISK*), along with temporal indicators (HOUR, IS_NIGHT, *IS_RUSH_HOUR, DAY_OF_WEEK*), dominated the rankings, confirming that where and when a collision occurs are the primary drivers of severe outcomes in Montréal.

To ensure robustness and interpretability, we retained only those features that appeared in the top 15 of all three methods simultaneously (intersection approach). This conservative strategy eliminated any variable that excelled in one method but performed poorly in others, yielding a final set of 15 features for modeling. This intersection-based selection represents best practice in predictive modelling for public safety applications, where false signals can have real human costs.

Detailed methodological justification for each selection technique is provided in *Appendix 3*, and complete cross-method rankings are presented in *Appendix 4*.

Methodology (Models & Validation)

To identify the algorithm best suited for proactive severe-crash prevention, we trained four baseline models on the 15 features selected in the previous part: Logistic Regression, Decision Tree, Random Forest, and XGBoost. These were deliberately chosen to provide a balanced spectrum from fully interpretable methods (Logistic Regression, Decision Tree) to highly powerful ensemble methods (Random Forest, XGBoost), allowing us to balance predictive performance with real-world practicality.

All models were evaluated on the original (unbalanced) test set using the following hierarchy of metrics designed to reflect the real impact of different prediction mistakes. Our primary metric was recall on the severe class, since the most critical objective is reducing false negatives. Severe and fatal collisions are extremely rare, so accuracy can appear deceptively strong even when a model entirely ignores the minority class. Missing a severe collision is far more damaging than issuing an additional false alert, making recall the most meaningful measure of risk prevention. Precision on the severe class was evaluated to understand the operational cost of false positives, while F1 score and ROC AUC provided secondary diagnostic insight into overall model behavior. Accuracy was included for completeness only, as it offers limited value in imbalanced public safety applications.

Among the baseline comparisons, Logistic Regression and Decision Tree achieved the highest recall on the severe class (55.2% and .16%, respectively). Random Forest and XGBoost traded marginal recall gains for significantly higher complexity and longer training times without meaningful improvement in recall or lives saved potential. Their stronger accuracy was expected but not aligned with the project's safety objective.

We selected Logistic Regression and Decision Tree as final models and tuned their hyperparameters. This decision reflects both statistical performance and operational reality: Montréal's road-safety teams require models that are not only accurate but also fast, explainable, and easily integrated into existing dashboards used by police and urban planners. By prioritizing recall, we focused on the most critical goal: identifying as many severe collisions as possible before they occur. This transforms predictive modeling from theory into a practical tool for proactive road safety.

With this objective in mind, we conducted a structured hyperparameter tuning process for both candidate models to evaluate whether meaningful performance gains could be achieved. The tuned Decision Tree showed only a negligible improvement in recall, increasing from 0.16 percent to 0.33 percent, which reinforced that rule-based splits are not effective at capturing the complex and infrequent patterns that characterize severe collisions. This confirmed that, despite its interpretability, the Decision Tree lacks the predictive capacity required for proactive safety planning.

For the Decision Tree, we tuned hyperparameters that directly govern model complexity and splitting behavior, including max depth, min samples split, min samples leaf, criterion, and max features, spanning wide but reasonable ranges. We used RandomizedSearchCV rather than GridSearchCV to efficiently explore a broad hyperparameter space on a high-volume dataset, while optimizing the average precision score, a metric better aligned with detecting the rare severe collision class. Even with extensive tuning, the model did not yield practical improvements in recall, further validating its limitations for this context.

In contrast, tuning substantially improved the performance of the Logistic Regression model. Using Elastic Net regularization, balanced class weights, and a calibrated decision threshold, the model's recall increased dramatically from 55.3 percent to 93.9 percent during validation, revealing that severe collision risk is governed by stable, learnable linear relationships. When evaluated on the independent test set, the tuned model maintained a strong recall of 0.613, demonstrating that the improvement was not due to overfitting. Cross validation further reinforced this stability, with mean recall values of approximately 0.64 across folds, providing strong evidence that the model generalizes reliably to new data.

Taken together, these tuning results illustrate how a traditionally simple and interpretable model, when appropriately optimized, can become a highly effective public safety instrument. The tuned Logistic Regression model offers the strongest and most reliable ability to identify high risk collisions before they occur, making it a valuable tool for data driven prevention, resource allocation, and long-term urban planning.

Results and Recommendations

The goal is to use these predictive insights to reduce the average of 21 traffic deaths per year in Montreal, with an aggressive target of a 25% reduction (5–10 lives saved annually).

Prioritize Spatial Intervention in "Red Zones"

Our key finding is that location is the primary factor driving the risk of severe crashes. Analysis revealed that 42% of all severe or fatal collisions were concentrated in ten neighborhoods. Neighborhood-level risk variables also ranked as the most important predictors in our logistic regression model. This confirms the hypothesis that severe crash risk is location dependent. Therefore, our primary recommendation is to prioritize interventions (infrastructure and enforcement) in ten critical neighborhoods, referred to as "Red

Zones." (top two neighborhoods are Ville-Marie and Villeray-Saint-Michel-Parc-Extension). This spatially targeted strategy is projected to result in a significant 25 – 40% reduction in city-wide fatalities, based on precedents and interventions being done in Vision Zero cities.

Recommendations for location-based interventions in "Red Zones"

Concentrate Time-Based Enforcement and Safety Campaigns

- **Analysis:** Across the ten neighborhoods, *time of day/week* was identified as a major risk factor. Severe crash probability is highest during specific windows (rush hour, late night), confirming temporal risk patterns.

- **Recommended Action:** Allocate enforcement and safety campaigns to high-traffic streets in red zones during lethal time-risk periods: 4pm – 8pm and 10pm – 5am. And prioritize high-risk days: Monday, Thursday, and Friday.

- **Expected Result:** 15–25% fatality reduction based on targeted resource allocation

Implement Protective Infrastructure for Vulnerable Populations

- **Analysis:** Non-car road users are more susceptible to severe crashes. 70% of fatalities involve vulnerable road users (pedestrians, cyclists, motorcyclists), with *pedestrians* and *cyclists* experiencing the highest victim impact.

- **Recommended Action:** Invest in and implement road infrastructure (bike lanes and pedestrian zones) in high-risk neighborhoods to protect non-car users from crash risk.

- **Expected Result:** 10–30% reduction in fatalities with protected infrastructure.

Place Speed Enforcement Officers in Strategic Locations

- **Analysis:** Surprisingly, 70% of deaths happen on *40km and 50~km/h streets*. This reveals that speeding in red-zones increases the risk of severe crashes, despite low-speed limits.

- **Recommended Action:** Place speed enforcement on high-risk streets in red zones during critical time periods (4pm – 8pm and 10pm – 5am on Mondays, Thursdays, and Fridays).

- **Expected Result:** 20–40% drop in fatalities across red-zone neighborhoods.

Integrate severity score into existing road navigating platforms

- **Analysis***:* The model's strength lies in its high recall (validated ability to identify and flag severe crash scenarios before they occur). This predictive capacity, paired with the clear interpretability of the top risk factors (time, location, and speed), allows the City of Montréal to effectively partner with technology providers.

- **Recommended Action:** Integrate the predictive model's output into existing navigation platforms (e.g., Google Maps, Apple Maps, or Waze) to launch a proactive crash risk feature. The system must caution drivers of the possibility of a crash ahead when the model predicts a high severity risk (e.g., score >0.65). This feature provides a safety warning before an incident occurs, prompting drivers to adjust their vehicle speed.

- **Expected result:** This combined strategy (infrastructure, enforcement, and technology alerts) is expected to have a multiplicative effect, like a 22–37% drop in fatal/severe crashes based on successful international pilot programs (e.g., Israel and Stockholm)

Limitations & Conclusion

The predictive model, despite its high performance, is subject to limitations primarily due to data availability and scope. Addressing these constraints is key to future iterations. Specifically, we lacked real-time traffic volume at the crash, granular individual victim vulnerability data (e.g., helmet or seatbelt use), and contextual socioeconomic or built-environment variables.

These serve as the foundation for the next steps of this project:

- Integrating real-time traffic micro-feeds
- Partnerships with the SAAQ to access anonymized vulnerability indicators
- Incorporating census or land-use layers to refine neighborhood risk modeling.

The project's predictive power provides a crucial and actionable blueprint for the City of Montréal. Our core recommendation is to immediately launch a Pilot Program (2025-2026) focused on the initial set of critical neighborhoods (Ville-Marie and Plateau Mont-Royal). This pilot will center on quick gains and actions: deploying targeted patrols in high-risk zones. Initiatives like a navigation alert system and comprehensive protected bike lanes/pedestrian zones are longer-term goals that would need to be implemented and scaled up in Phase 2 or 3 of our strategic crash reduction plan. This phased approach is expected to result in a multiplicative effect on fatality reduction, transforming predictive modeling into a practical intervention tool designed to improve public safety outcomes across the city.

The effectiveness of this project will be measured using the following KPIs:

- % reduction in severe collisions in the 10 prioritized neighborhoods.
- % reduction in fatalities specifically for vulnerable road users (pedestrians and cyclists)
- % drop in severe crashes linked to speeding drivers in red zones due to proactive alerts.

Conclusion

This project demonstrates the transformative potential of predictive analytics in road safety. By shifting from reactive post-crash analysis to proactive risk modeling, we identified clear spatial and temporal patterns that drive severe collision risk in Montréal. Our findings confirm that location and time are the most influential factors, with severe crashes disproportionately concentrated in specific neighborhoods and during high-risk periods. The tuned logistic regression model achieved high recall, proving that interpretable, data-driven tools can reliably flag severe crash scenarios before they occur.

The recommendations that target interventions in “Red Zones,” time-based enforcement, protective infrastructure for vulnerable road users, and integration of predictive alerts into navigation platforms offer a practical roadmap for reducing fatalities by up to 25–40%. While limitations remain, particularly around real-time traffic data and socioeconomic context, these gaps present opportunities for future iterations and partnerships.

Ultimately, this work provides the City of Montréal with an actionable blueprint to save lives. By combining predictive modeling with strategic interventions, Montréal can move decisively toward safer streets and a measurable reduction in severe and fatal collisions.

APPENDIX

Appendix 1: Literature Review vs. Foundation Lab

Attributes	Study Findings	Our Project
Collision Near	"Close to landmark" has 30% higher severity for accidents.	Which neighborhood? Is one area riskier than another? We identify where in order for preventive measures to be implemented.
Road Condition	Road condition accounts for .05% of accident severity	Which road conditions contribute most? We determine how bad they are so repairs can be prioritized effectively.

Appendix 2: List of Features Based on Category Structures

Category of variable (Grouping)	Variables	Analytical Purpose
1. Structural Factors (Infrastructure & Design)	**route_category_code, road_aspect_code, roadway_condition_code, road_configuration_code, accident_location_code, accident_position_code, authorized_speed**	**Focus on Inherent Design and Maintenance Risk:** This group assesses the static risk embedded in the urban environment, isolating the impact of permanent design flaws and maintenance failures to demand long-term engineering solutions. Authorized speed is included as a severity multiplier linked to kinetic energy.
2. Collision Event and Context	**collision_type_code, environment_code**	**Focus on Mechanism and Activity Zones:** These variables pinpoint how the collision occurred (e.g., vehicle-to-pedestrian, vehicle-to-fixed object) and where (e.g., school zone, commercial district). This determines if the severity risk is tied to a specific activity type or a particular collision mechanism.
3. Environmental & Surface Conditions	**weather_condition_code, surface_condition_code, lighting_condition_code**	**Focus on Dynamic Friction and Visibility: This** set quantifies how environmental dynamics exacerbate severity. It models the influence of visibility (lighting), friction loss (surface condition like ice/slush), and general weather, essential for operational alerts.
4. Spatial Factors (Geography)	**neighborhood**	**CRITICAL: TO CREATE THE VARIABLE.** This categorical variable allows the model to learn that road conditions are disproportionately severe in specific neighborhoods.
5. Dynamic Factors (Temporal)	**HR_ACCDN (Hour) and JR_SEMN_ACCDN (Day of the week)**	**Focus on Operational Peaks:** Accounts for time-dependent risk variations (e.g., rush hour congestion, late-night alcohol/fatigue risk, weekend driving patterns) crucial for resource allocation and intervention timing.

Appendix 3: Rationale for Feature Selection Methods

Method	Why We Chose It	What It Helps Us Detect
ANOVA F-Test	Evaluate whether categorical features (e.g., neighborhoods, day of week) produce statistically different severity rates across groups. Useful as a fast-screening tool before modeling.	Detects features strong group-level separation in severity outcomes; highlights variables with clear discriminatory power.
Gini Importance (from Random Forest)	Chosen because it captures nonlinear relationships and works with both numeric and categorical features. Helps identify variables that create purer splits in tree-based models, complementing linear tests.	Detects features that significantly improve classification structure, even when relationships are nonlinear or interaction based.
Permutation Importance	Included to provide a model-agnostic measure of importance by shuffling features and measuring the performance drop. This protects against collinearity and confirms robustness across different model types.	Detects features that the model is most dependent on, showing true predictive contribution rather than structural weight or correlation artifacts.

Appendix 4: Feature Selection Ranking

MASTER FEATURE RANKING — ALL THREE METHODS COMBINED

Rank	Feature	ANOVA F	Gini	Perm	Composite	
1	QUARTIER_RISK	4751.78	0.1940	0.044676	1.67	★★★
2	ARROND_RISK	1812.24	0.3397	0.028562	2.67	★★★
3	QUARTIER	1554.66	0.1280	0.028987	3.67	★★★
4	ARRONDISSEMENT	1397.60	0.1314	0.028400	4.33	★★
5	MUNICIPALITE	9745.81	0.0515	0.003014	4.67	★★
6	MUNCP_RISK	689.32	0.0504	0.004175	6.67	★★
7	HOUR	171.50	0.0495	0.005022	7.67	★★
8	DAY_OF_WEEK	303.53	0.0227	0.003570	8.00	★★
9	SPEED_LIMIT_KMH	848.40	0.0180	0.000069	9.00	★
10	IS_RUSH_HOUR	250.98	0.0036	0.002883	10.33	★
11	IS_WEEKEND	17.89	0.0037	0.002713	11.00	★
12	IS_NIGHT	2829.73	0.0076	-0.008306	11.67	★
13	MONTH	nan	0.0000	0.000000	nan	★
14	QUARTER	nan	0.0000	0.000000	nan	★
15	MONTH_SIN	nan	0.0000	0.000000	nan	★
16	MONTH_COS	nan	0.0000	0.000000	nan	
17	SEASON	nan	0.0000	0.000000	nan	
18	WEATHER_EN	nan	0.0000	0.000000	nan	
19	SURFACE	nan	0.0000	0.000000	nan	
20	LIGHTING	nan	0.0000	0.000000	nan	

References

- Société de l'assurance automobile du Québec (SAAQ). (2023). Collisions routières – Ville de Montréal [Data set]. Données Québec. https://www.donneesquebec.ca/recherche/dataset/vmtl-collisions-routieres.
- Muktar, B., & Fono, V. (2024). Toward Safer Roads: Predicting the Severity of Traffic Accidents in Montreal Using Machine Learning. Electronics, 13(15), 3036. https://doi.org/10.3390/electronics13153036.
- Israel Ministry of Transport Pilot (2018–2021)
- Stockholm Vision Zero dynamic alerts evaluation, Waze for Cities Partner documentation.
- Teschke et al. + Montréal 2025 Bike Network Expansion Plan.
- Cochrane 2017 systematic review
- WHO 2018 + SAAQ 2025 Unsafe Driving Behaviors Campaign