Machine Learning Models for Calgary's Traffic Risks

DATA 607 - Project Report

Jincy Thomas
Joshua Quartey
Megha Radhakrishnan
Deepika Gollamandala

1. INTRODUCTION

Traffic is a part of everyday life in a growing city like Calgary. But with more vehicles on the road, the chances of accidents also go up affecting not just drivers, but also pedestrians, cyclists, and emergency workers. If we can understand where and when these incidents happen most often, we can help make the city safer for everyone.

In this project, we study Calgary's traffic incident data using machine learning. Our aim is to find patterns and build models that can predict the risk level of different areas and estimate incident rates based on various factors. Specifically, we aim to specifically answer the following questions:

Research Questions:

- Can we identify spatial hotspots of traffic collisions in Calgary using historical incident data?
- How do time-of-day and weather conditions influence the frequency and severity of incidents?
- Can we build a predictive model to flag high-risk locations ?
- What factors contribute to pedestrian-related incidents, and can we model a risk score for communities or intersections?

2. DATA SOURCE

To analyse how weather and time affect pedestrian involvement in traffic incidents, the following datasets were used.

1. Traffic Incidents Dataset (Open Calgary)

This dataset containing approximately 55,000 records is provided on the Open Calgary website by the City of Calgary and is updated every 10 minutes. Information regarding incident type, description, date/time, and geolocation (latitude and longitude) is included in the dataset in csv format. Data from the period July 2022 to July 2025 is used for analysis. The dataset can be accessed at

https://data.calgary.ca/Transportation-Transit/Traffic-Incidents/35ra-9556/about_data.

The dataset is made available under a free license, details of which can be found at: https://data.calgary.ca/d/Open-Data-Terms/u45n-7awa

2. Weather Data from Open-Meteo API

Weather data for the city of Calgary has been obtained from the Open-Meteo API, which provides free access to historical, current, and forecast weather information on a global scale. The API documentation is accessible at: https://open-meteo.com/en/docs.

Weather parameters known to influence traffic conditions were selected from the Open-Meteo API specifically for Calgary and are detailed as follows:

Variable	Unit	Description
temperature_2m	°C (°F)	Air temperature at 2 meters above ground
snowfall	cm (inch)	Snowfall amount of the preceding hour in centimeters.
snow_depth	meters	Snow depth on the ground
wind_gusts_10m	km/h	Gusts at 10 meters above ground as a maximum of the preceding hour
rain	mm (inch)	Rain from large scale weather systems of the preceding hour in millimeter
relative_humidity_2m	%	Relative humidity at 2 meters above ground
wind_speed_10m	km/h	Wind speed at 10 meters

Note: The variables from the datasets below will be explained in Part B.

3. Traffic Counts at Permanent stations

Traffic volumes for different stations in Calgary were obtained from the City of Calgary's Open Calgary website, along with information on location, direction, segment, and time. https://data.calgary.ca/Transportation-Transit/Traffic-Counts-at-Permanent-stations/vuyp-sbjp/about_data

4. Permanent station locations for Traffic counts

Permanent station location data from the City of Calgary was also obtained, providing precise geographic and descriptive details of each traffic counting station.

https://data.calgary.ca/Transportation-Transit/Permanent-station-locations-for-Traffic-counts/sqwx-tjsy/about_data

5. Major Road Network

The Major Road Network dataset from the City of Calgary contains the mapped layout of the city's primary transportation corridors, including major streets, expressways, and arterial roads.

https://data.calgary.ca/Transportation-Transit/Major-Road-Network/tqjs-vnhy/about_data

6. Street Centreline

The Street Centreline dataset from Open Calgary lists detailed information for each street segment in Calgary. It includes the street name, type, direction, address ranges, community name, road classification, and maintenance responsibility.

https://data.calgary.ca/Transportation-Transit/Street-Centreline/4dx8-rtm5/about_data

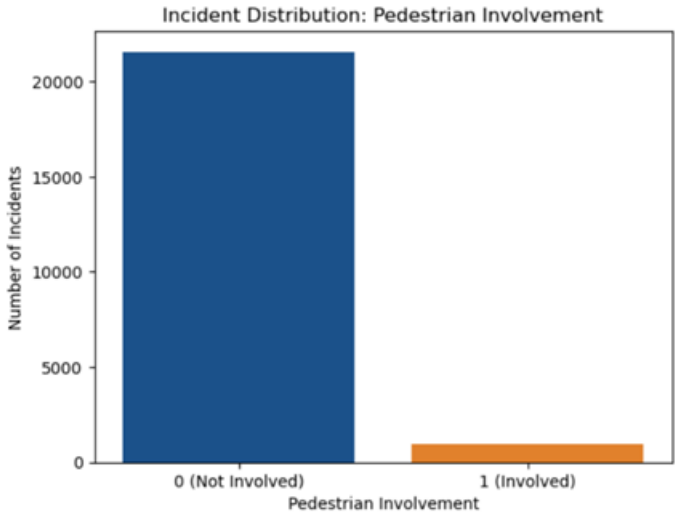
3. DATA PREPROCESSING

To prepare the traffic incident data for analysis, we followed a structured approach that involved cleaning, filtering, and enriching the dataset with additional features:

- **Loading and Cleaning the Data**: We began by loading the traffic incident CSV file and converting the START_DT column into a proper datetime format. This allowed us to perform time-based filtering and feature extraction.
- **Filtering by Date Range:** We restricted the dataset to incidents that occurred between July 1, 2022, and July 31, 2025. This helped us focus on recent and relevant data for our analysis.
- **Extracting Date Components**: From the incident timestamps, we extracted the year, month, and day. We also classified each incident as either a weekday or weekend and created a binary code to represent this (1 for weekday, 0 for weekend).
- **Seasonal Categorization**: Using the month of each incident, we assigned a season—Winter, Spring, Summer, or Fall—and added a corresponding numeric code to simplify future analysis.
- **Time-of-Day Classification:** We created a new column to categorize the time of day when each incident occurred. The categories included Rush Hour, Mid Morning, Midday, Evening, and Night/Early Morning, based on the hour of the incident.
- **Quadrant Coding**: Quadrants (NW, SW, NE, SE.) were coded numerically
- **Handling Missing Data:** We checked for missing values and removed any rows containing nulls to ensure the dataset was clean and consistent.

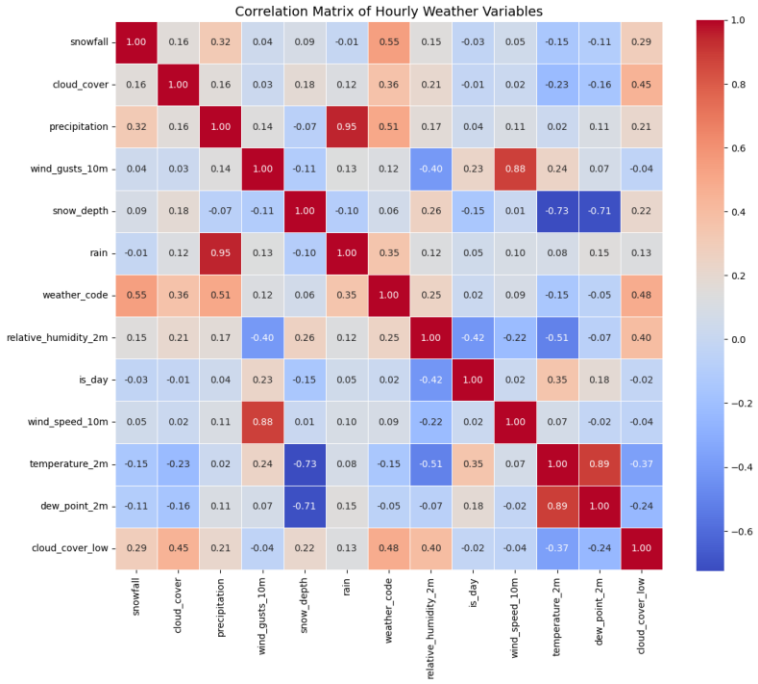
- **Identifying Pedestrian Involvement**: To flag incidents involving pedestrians, we added a binary column that checks whether the word “pedestrian” appears in the incident description.
- **Integrating Weather Data**: To enrich the dataset with environmental context, we matched each incident to hourly weather data. We rounded the incident time down to the nearest hour and merged it with a cleaned weather dataset using a timestamp-based join.

4. Exploratory Data Analysis (EDA)



Incident Distribution – Pedestrian involvement

It was observed that 4.2% of recorded incidents (950 out of 22,510) involved pedestrians, while the remaining 95.8% did not.



Correlation matrix – Weather data

To ensure model simplicity and avoid redundancy, we included raw and independent features such as rain, snowfall, cloud_cover, temperature_2m, snow_depth, relative_humidity_2m, wind_speed_10m, wind_gusts_10m(included to capture the impact of short-term, high-intensity wind events), and is_day, while excluding derived or highly correlated variables like precipitation, dew_point_2m, and weather_code.

5. MODELING

Part A: Predictive Modeling of Pedestrian Involvement in Traffic Incidents

The factors influencing pedestrian involvement in traffic incidents were investigated, and the potential to predict such involvement was evaluated. Initially, variables such as road conditions, weather, traffic volume, and driver behavior were planned to be incorporated by merging these datasets with the incident data. However, these datasets were not available in a format compatible for merging. Instead, weather data corresponding to the exact time of each incident was retrieved using a weather API to enrich the dataset. Additionally, time-related features including hour of the day, weekend indicator, month, and year were extracted from the incident timestamps. These weather and temporal features were then used in the development of predictive models for pedestrian involvement in traffic incidents.

At the first modeling attempt, the dataset suffered from extreme class imbalance — most incidents did not involve pedestrians, while pedestrian-involved incidents were much rarer. As a

result, the model mostly learned to predict the majority class (no pedestrian involvement) for every case. This led to what is known as the "trivial classifier" problem: the model achieves high overall accuracy simply by always guessing the dominant class but performs very poorly in identifying the minority class, which is pedestrian involvement in this case. Consequently, the model's recall for pedestrian incidents was very low, meaning it missed most of the actual pedestrian-related cases.

To handle class imbalance, class weights were manually calculated to give more emphasis to pedestrian incidents. These weights were assigned to each training sample, helping the model focus more on the minority class and improve detection despite its rarity.

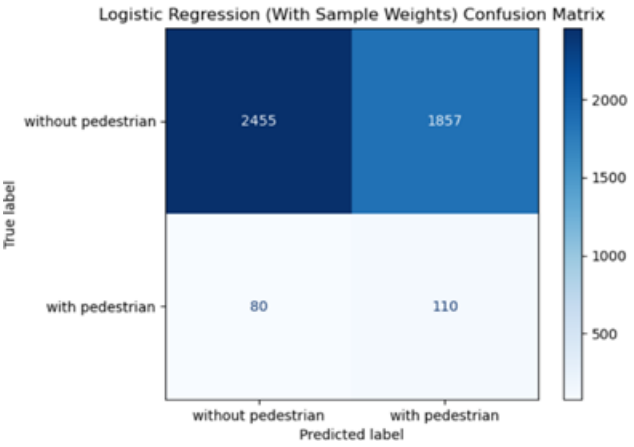
The majority class was assigned a weight of 0.52, while the minority class received a much higher weight of 11.85 to counterbalance its rarity during model training.

Sample weights were applied to different models, including Logistic Regression, Random Forest, and Boost, to address class imbalance. Subsequently, threshold tuning was performed for each model to optimize the balance between precision and recall.

A function was created to train classification models using sample weights to address the imbalance between approximately 21,560 non-pedestrian and 950 pedestrian incidents. The models were evaluated on test data through confusion matrices and classification reports, which included precision, recall, F1-score, and support. Precision was used to measure the accuracy of positive predictions, recall was employed to assess how many actual pedestrian incidents were detected, and the F1-score was calculated to balance both metrics, which was important given the class imbalance. Support indicated the number of instances per class. Together, these metrics provided a clear view of model performance in predicting pedestrian involvement.

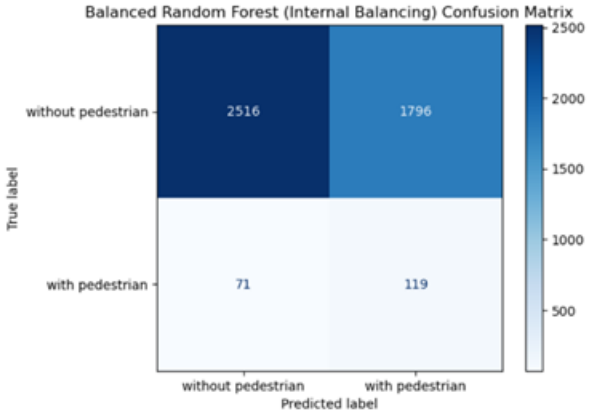
Modeling Choices

Four classification models were trained and evaluated to predict pedestrian involvement. Logistic Regression and Random Forest were trained using explicit sample weights to address class imbalance. A Balanced Random Forest was also tested, which handles imbalance internally without sample weights. Lastly, an XGBoost model was trained with a weighted scale parameter to emphasize the minority class. Each model's performance was assessed using confusion matrices and classification reports and the following results were obtained.



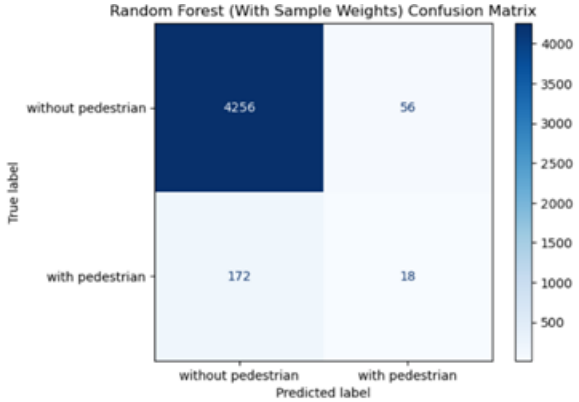
	precision	recall	f1-score	support
0	0.968	0.569	0.717	4312
1	0.056	0.579	0.102	190
accuracy			0.570	4502
macro avg	0.512	0.574	0.410	4502
weighted avg	0.930	0.570	0.691	4502

Logistic Regression – Confusion Matrix and Evaluation Metrics



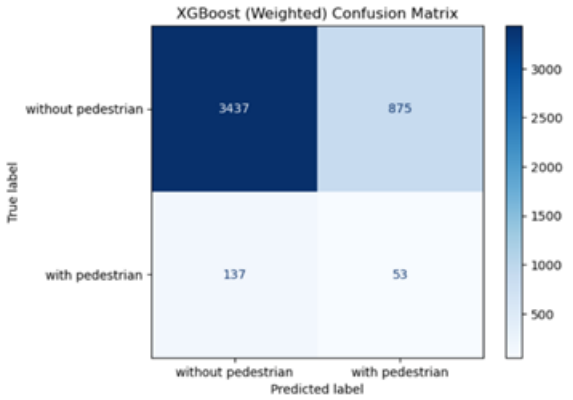
	precision	recall	f1-score	support
0	0.973	0.583	0.729	4312
1	0.062	0.626	0.113	190
accuracy			0.585	4502
macro avg	0.517	0.605	0.421	4502
weighted avg	0.934	0.585	0.703	4502

Balanced Random Forest – Confusion Matrix and Evaluation Metrics



	precision	recall	f1-score	support
0	0.961	0.987	0.974	4312
1	0.243	0.095	0.136	190
accuracy			0.949	4502
macro avg	0.602	0.541	0.555	4502
weighted avg	0.931	0.949	0.939	4502

Random Forest with sample weights – Confusion Matrix and Evaluation Metrics



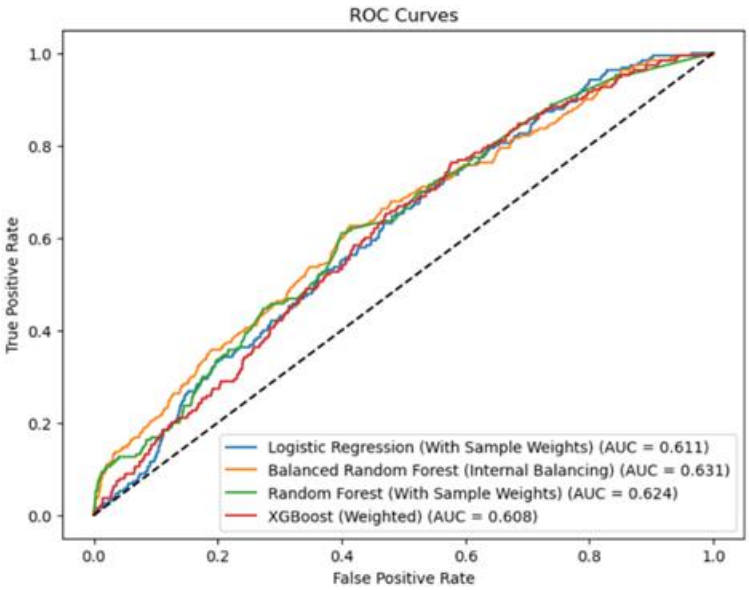
	precision	recall	f1-score	support
0	0.962	0.797	0.872	4312
1	0.057	0.279	0.095	190
accuracy			0.775	4502
macro avg	0.509	0.538	0.483	4502
weighted avg	0.923	0.775	0.839	4502

XGBoost (Weighted) – Confusion Matrix and Evaluation Metrics

When comparing the four models, the balanced random forest achieved the highest recall (62.6%), detecting nearly two-thirds of actual pedestrian cases. Logistic regression followed closely with 57.9% recall, while XGBoost weighted detected 27.9% and random forest with sample weights detected only 9.5%, missing most incidents. In terms of precision, all models struggled: the random forest with sample weights performed best at 24.3%, while the balanced random forest (6.2%), logistic regression (5.6%), and XGBoost weighted (5.7%) produced high false-positive rates. Accuracy was highest for the random forest with sample weights (94.9%), followed by XGBoost weighted (77.5%), balanced random forest (58.5%), and logistic regression (57.0%), though this metric is misleading given the severe class imbalance. Overall, the balanced random forest provided the best sensitivity to pedestrian incidents but at the cost of very low precision, the random forest with sample weights was the most precise but largely ineffective for detection, and logistic regression and XGBoost weighted offered moderate recall but similarly poor precision. None of the models achieved a practical balance between detecting and accurately identifying pedestrian-involved incidents.

ROC

ROC curves were plotted for all models to evaluate their ability to distinguish pedestrian involvement, with the AUC summarizing overall performance. ROC analysis was used because it effectively measures model discrimination across all classification thresholds, especially important in imbalanced datasets.



ROC curves for predicting pedestrian involvement in traffic incidents

AUC values just above 0.6 were observed for all models. An AUC of 0.5 is considered random guessing. It is shown that the models barely distinguish pedestrian from non-pedestrian incidents. The small improvement indicates that pedestrian involvement is not well predicted by the current features.

The highest AUC of 0.631 was achieved by the Balanced Random Forest, likely due to its internal handling of class imbalance. An AUC of 0.624 was obtained by the Random Forest with Sample Weights, performing similarly but slightly less well. Comparable AUCs of 0.611 and 0.608 were recorded for Logistic Regression and XGBoost, respectively, indicating weaker discrimination compared to the Random Forest variants.

Sample Weighting Methods used in the models

For models using manual sample weights, the weights were computed based on the inverse frequency of each class:

- **Minority class weight** = total samples / (2 × number of minority samples)
- **Majority class weight** = total samples / (2 × number of majority samples)

The sample weighting methods used by each model to address class imbalance are summarized in the table.

Model	Sample Weight Used	Why
Logistic Regression	Explicit sample weights passed in `.fit()`	Straightforward manual weighting to handle class imbalance; easy to implement and tune.
Balanced RF	Internal balancing via model's algorithm	Automatically balances classes by adjusting bootstrap samples; no need for manual weights.
RF (Sample Weights)	Explicit sample weights passed in `.fit()`	Manual weighting helps improve minority class detection in a standard RF.
XGBoost (Weighted)	`scale_pos_weight` parameter set at init	Built-in, simple class-level weighting to handle imbalance without extra complexity.

Threshold Tuning

A threshold tuning function was implemented to find the classification probability cutoff that maximizes the F1-score by balancing precision and recall. For each model, the best threshold was identified using the precision-recall curve on the test data. Models were then evaluated at these optimized thresholds, with confusion matrices and classification reports generated to assess performance beyond the default 0.5 cutoff.

The best thresholds, precision, recall, and F1-scores for each model were determined and are summarized in the table. Among the models tested, the weighted XGBoost achieved the highest F1-score, indicating the best balance between precision and recall for predicting pedestrian involvement.

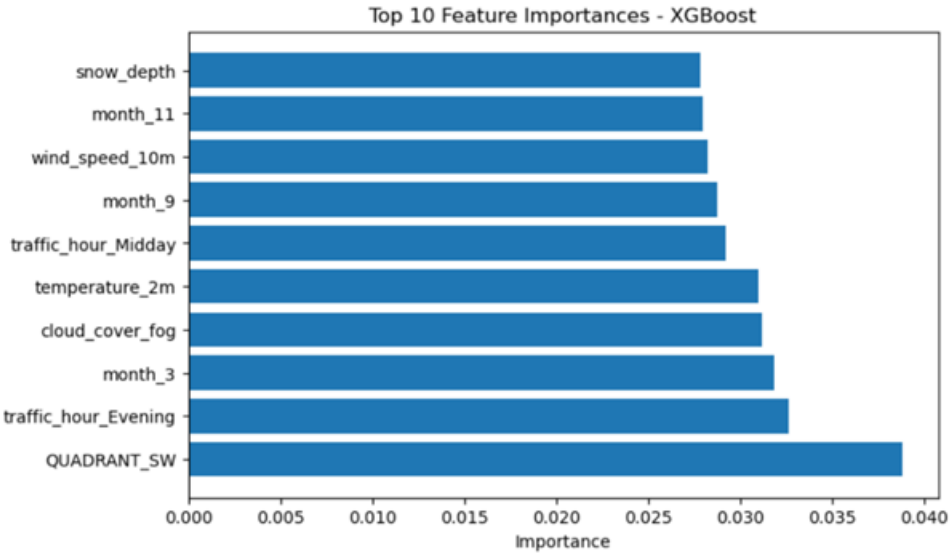
Model	Best Threshold	Precision	Recall	F1-Score
Logistic Regression (Weighted)	0.58	7.10%	36.30%	11.80%
Balanced Random Forest	0.605	20.20%	27.90%	23.50%
Random Forest (Weighted)	0.115	15.60%	26.30%	19.60%
XGBoost (Weighted)	0.62	17.40%	37.40%	23.70%

Performance Metrics at Optimal Thresholds

The models showed varied trade-offs between precision and recall at their best thresholds, with XGBoost and Balanced Random Forest achieving the highest F1-scores. However, overall precision remained low, indicating many false positives, while recall was moderate, reflecting challenges in reliably detecting pedestrian-involved incidents.

Feature Importance

The top 10 features for the best model were plotted to show the most important factors influencing its predictions.



Top features for predicting pedestrian involvement in traffic incidents

Given that the model's overall performance is poor and all feature importances are below 0.04, it indicates that the features provide weak signals for predicting pedestrian involvement. The model struggles to identify clear patterns, likely because no individual feature strongly correlates with the target, contributing to the limited predictive power observed.

Conclusion

It is suggested that strong predictive signals for pedestrian involvement are not contained in the existing features (such as time, location and weather). The prediction of rare events like pedestrian-involved crashes is hindered by class imbalance and complex, subtle contributing factors. Although operational deployment cannot be supported by the models alone, their predictions could be integrated into broader traffic safety risk frameworks as one of several inputs, pending further refinement and feature enhancement.

Part B: Modeling Incident Rates and High-Risk Road Segments

Since our primary dataset did not include many features that were necessary for predicting the risk of traffic incidents, we decided to merge a few supplementary datasets to draw meaningful insights.

Data Merging, Data Cleaning & Feature Extraction

To analyze traffic incidents in Calgary and build predictive models for estimating incident rate and high-risk road segments, we first cleaned, organized and then merged our supplementary datasets to prepare Calgary's traffic incident data for spatial analysis and predictive modelling. Each dataset contributed unique information such as roadway layout, traffic volume, or incident details which enabled a deeper and more meaningful analysis. Following are the details of the steps that we followed for cleaning and extraction of features from each of the datasets we used:

Primary Dataset: Calgary Traffic Incidents Data

Dataset Used: Calgary Traffic Incidents (Open Calgary)

- As described earlier our primary dataset has information of traffic incidents along with time stamp, geographical location details, incident description.
- The raw traffic incident dataset was cleaned by removing duplicate and empty rows to maintain data integrity.
- Column names were standardized for consistency, and timestamps such as start_dt and modified_dt were parsed into datetime format to facilitate time-based filtering and analysis.
- Geographic filtering was applied to exclude incidents with missing or invalid coordinates. Only records falling within Calgary's geographic bounds (latitude: 50.5–51.3, longitude: -114.3 to -113.8) were retained.
- The cleaned dataset was then converted into a GeoDataFrame to support spatial operations and mapping.

Street Centreline Data

Dataset Used: *Street Centreline (Open Calgary)*

- Purpose: We used this dataset to map and identify the exact street segment associated with each traffic incident.
- This dataset provided detailed information about Calgary's road network, including street id, street names, their geographical location, geographical shape, type (local road/ feeder roads/ arterial highway etc), ownership (private or maintained by city), directionality - One way?, CTP_Class – Calgary Transportation Planning route classification.
- Since our traffic incident dataset includes geographical coordinates for each incident, we used the Street Centreline dataset to determine which street segment each incident

occurred on. This spatial join allowed us to associate every incident with its corresponding road segment.

- By linking incidents to specific street segments, we were able to analyse risk and incident patterns at the segment level.

Traffic Volume Data

Dataset Used: (1) Traffic Counts at Permanent stations (Open Calgary) and (2) Permanent station locations for Traffic counts (Open Calgary)

- To effectively analyse traffic incidents, assess roadway risk, and identify high-risk traffic incident hotspots, it is essential to understand the volume of traffic on each street segment. For this purpose, we utilised two key datasets from Open Calgary:
    1. **Permanent Station Locations for Traffic Counts** – This dataset provides the geographic coordinates (latitude and longitude) of permanent traffic count stations, along with the street segment ID where each station is situated.
    2. **Traffic Counts at Permanent Stations** – This dataset contains detailed traffic volume records, including the station ID, date, and the number of vehicles (traffic volume) recorded at each station.
- We combined the station location dataset with the traffic counts dataset using station IDs. This enabled us to assign traffic volume data to specific road segments based on geographic coordinates.
- The traffic counts data was cleaned to remove duplicates and inconsistencies.
- Time-based features were extracted, including hour, day of the week, month, year, and weekend indicators.
- Flags for Alberta public holidays and school session days were added, assuming schools operate from September to June, excluding weekends and holidays. These features enabled analysis of traffic behaviour across different time frames and conditions.
- By combining these datasets, we were able to spatially link traffic volume data to specific road segments, forming the foundation for exposure estimation and risk-based traffic analysis.

Exposure Estimation (AAWT)

Dataset Used: *Traffic Counts (Open Calgary)*

- When analyzing crash or incident data, relying solely on raw counts can be misleading. Roads with higher traffic volumes naturally tend to have more incidents. To assess risk more accurately, we normalize incident data using **Average Annual Weekday Traffic (AAWT)** a measure of vehicle exposure.
- **AAWT** represents the **average number of vehicles** that travel on a specific road segment **per weekday over a year**.It's calculated by collecting traffic counts throughout the year and averaging them across all weekdays.

- AAWT gives a consistent measure of how busy a road is. This is crucial for understanding how often a road is used and by how many vehicles
- In our case, Traffic volumes were grouped by segment ID and date to calculate total daily traffic per segment and we computed separate averages for weekday and weekend traffic, resulting in two key metrics:
  - aawt_weekday: Average Annual Weekday Traffic
  - aawt_weekend: Average Annual Weekend Traffic

This approach will allow us to identify road segments with disproportionately high incident rates relative to their traffic volume, supporting more accurate and equitable risk-based modelling.

Refer to Appendix for a detailed explanation of the dataset merging process.

Dataset Validation

Python methods used: .info(), .describe(), .isnull(), .sum(), .value_counts()

We utilised a combination of inspection tools/ python methods to confirm the presence and completeness of key fields, validate data types, and ensure successful feature extraction. These checks also helped identify anomalies or missing values prior to modelling.

These preprocessing steps were essential for transforming raw data into a structured and analysable format. They laid the foundation for accurate spatial joins, exposure-based risk modelling, and time-aware traffic analysis.

Metrics Used for Modelling Decisions

To support volume-adjusted risk analysis and fair comparisons across Calgary's road network, we used the following key metrics:

Average Annual Weekday/Weekend Traffic (AAWT)

Definition: AAWT represents the average number of vehicles passing through a road segment per day, calculated separately for weekdays and weekends. It is measured in **vehicles/day.**

Purpose:

- Provides a snapshot of typical traffic volume.
- Describes how busy a road is on an average weekday or weekend.
- Serves as the foundation for exposure and incident rate calculations.

Exposure Vehicle Days

Definition: This metric measures how much traffic volume passes through a segment over the study period (1.1.2018 to 31.12.2024) It's calculated separately for weekdays and weekends because traffic patterns differ. This metric lets us quantify risk exposure/the total opportunity for a traffic incident to happen.

Calculation

road_segments["exp_vehicle_days_weekday"] = road_segments["aawt_weekday"] * num_wkdays

road_segments["exp_vehicle_days_weekend"] = road_segments["aawt_weekend"] * num_wkends

where:

* `aawt_weekday` / `aawt_weekend` = Average Annual Weekday/Weekend Traffic (vehicles/day)

* `num_wkdays` = number of weekdays in the analysis period (2018–2024)

* `num_wkends` = number of weekend days in the analysis period.

Interpretation:

"Over all weekdays from 2018–2024, X million vehicles traveled on a given road segment." This metric quantifies the total opportunity for a traffic incident to occur, based on vehicle exposure.

Incident Rate per Million Vehicles

Definition: This metric describes how many traffic incidents happen per million vehicles that travel on a given segment. We divide incident counts by traffic exposure to normalize for volume differences so, for example, busy arterials aren't automatically flagged as risky because more cars use them. This metric allows us to compare fairly across different road types (residential vs arterial), traffic volumes or locations because they are volume-adjusted.

Calculation:

road_segments.loc[wd, "rate_weekday_per_million"] = (road_segments.loc[wd, "incidents_weekday"] / road_segments.loc[wd, "exp_vehicle_days_weekday"]) * 1_000_000

Interpretation:

"For every million weekday vehicles using this road, about 1.25 crashes occurred."

Purpose:

- Prevents high-traffic roads from being flagged as risky solely due to volume.
- Enables fair comparison across road types (e.g., residential vs arterial) and locations.
- Supports prioritization of segments with disproportionately high incident rates.

Feature engineering

In identifying useful features for modelling, our general strategy was to find three kinds of features:

- Spatial features Given we were working with modeling spatial data we thought these types of features would be highly/most relevant. They were also important for comparison with KDE. We compared both to see if learning from both spatial and non-spatial attributes would edge out some performance advantage.
- Volume and classification variables: Variables like AAWT and road class type(ctp-class) are obviously strong predictors when modeling traffic incident risk based on traffic volumes and are also very easily interpretable.
- Event-timing variables: e also wanted to investigate whether temporal patterns matter in differentiating high-risk roads so we engineered variabels like p_peak_hours and p_public_holiday.

Below is a list/breakdown of existing features in our data as well as engineered ones along with brief descriptions of them:

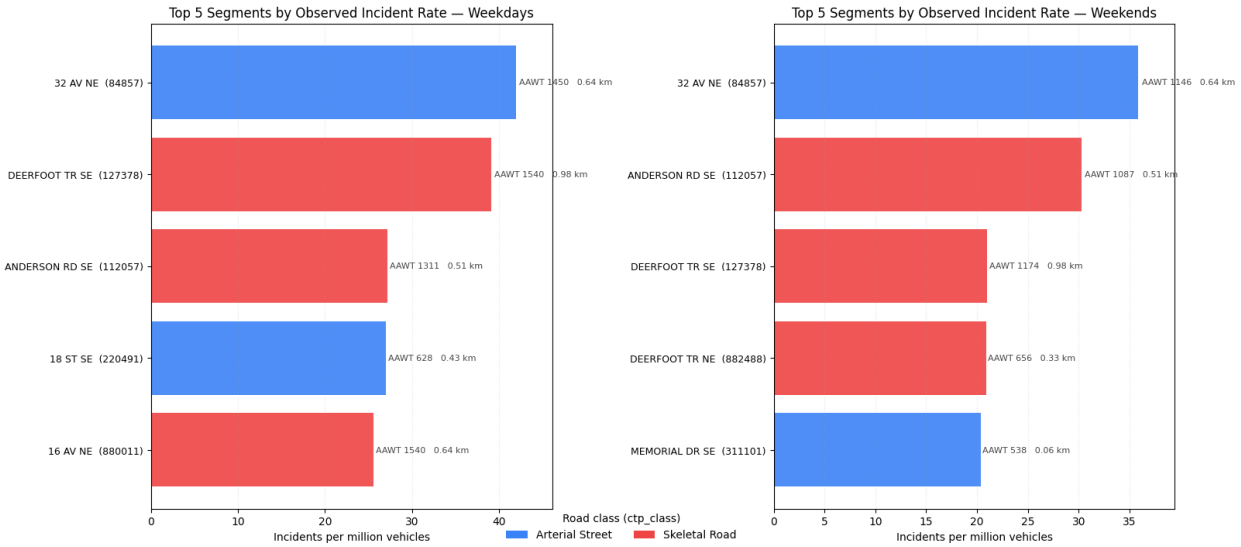
Existing features(from data sources)

Feature	Type	Description
`ctp_class`	Categorical	Road classification from city traffic plan (e.g., freeway, arterial, collector). Encodes functional role in the network.
`one_way`	Binary (0/1)	Whether the road segment is one-way. Impacts lane changes, flow, and potential for head-on collisions.
`aawt_weekday` , `aawt_weekend`	Numeric	Average Annual Weekday/Weekend Traffic volumes. Primary exposure measure for risk normalization and as a predictor.
`is_major_road`	Binary	Whether the segment is classified as a major road. Can correlate with both higher volumes and design features (e.g., more lanes, higher speeds).
`x_m` , `y_m`	Numeric	Projected centroid coordinates in meters (UTM). Captures location for spatial trends when modeling.
`octant`	Categorical	Orientation/direction of the road segment (e.g., N, NE, SE). A proxy for traffic patterns tied to directionality.

Feature	Type	Description
`p_peak_hours`	Numeric (0–1)	Proportion of incidents on this segment occurring during defined peak periods. Reflects congestion/vulnerability times.
`p_public_holiday`	Numeric (0–1)	Proportion of incidents on public holidays. Captures unusual traffic patterns or behaviors.
`incidents_per_km`	Numeric	Total incidents normalized by segment length. Measures raw incident density.
`rate_overall_per_million`	Numeric	Incidents normalized by vehicle exposure (per million vehicles). Used for regression targets or classification thresholds.
`high_risk`	Binary	Derived from top x% of `rate_overall_per_million` . Used as classification label.

Exploratory Data Analysis (EDA)

Most Dangerous roads/Roads with highest incident rates



Weekdays:

1. 32 AV NE (41.9 incidents/million vehicles, Arterial Street)
    - This is a short-to-medium length urban arterial serving both local businesses and through traffic.
    - High weekday risk could stem from mixed traffic flow (commuterand local traffic flows) and numerous intersections, increasing conflict points.
    - Arterials often have moderate speeds but have more direct pedestrian interaction than highways, which compounds risk.
2. DEERFOOT TR SE (39.1 incidents/million vehicles, Skeletal Road)
    - Deerfoot Trail is a Part of Calgary's primary freeway system. The Deerfoot Trail corridor is historically one of Alberta's busiest and most collision-prone roads, with over 150,000 AADT in some segments.
    - High risk is likely due to weekday peak-hour congestion
3. ANDERSON RD SE (27.1 incidents/million vehicles, Skeletal Road)
    - This is a connector route feeding traffic between residential zones and major arterials/freeways.
    - It has a shorter segment length but high flow, suggesting high collision density relative to exposure.
4. 18 ST SE (27.0 incidents/million vehicles, Arterial Street)
    - Serves both industrial and commercial areas. Traffic typically consists of a mix of trucks with other local light vehicles, the difference in driving skill between the types of drivers might explain weekday incident rate.

    - It has a lower AAWT than Deerfoot/Anderson but has a higher rate per million vehicles, which points to intrinsically hazardous geometry or traffic vehicle mix.
5. 16 AV NE (25.6 incidents/million vehicles, Skeletal Road)
    a. This segment represents a section of the Trans-Canada Highway passing through urban Calgary.
    b. It combines freeway speeds with frequent entry/exit ramps and intersections known nationally for collision clusters in urban highway stretches.

Weekends

1. 32 AV NE (35.9 incidents/million vehicles)
   32 AV is Still top-ranked for weekends as well, but has a reduced rate from weekdays, possibly due to lower commuter volume but steady local traffic and commercial access.

2. ANDERSON RD SE (30.2 incidents/million vehicles)
   Weekends are still busy, likely due to recreational and shopping trips. The higher weekday rate could suggest commuter-heavy weekday risk profile.

3. DEERFOOT TR SE (21.0 incidents/million vehicles)
   There seems to be a significant drop in numbers for the weekend. This could be due to lighter congestion and fewer peak-period chain-reaction collisions.

4. DEERFOOT TR NE (20.9 incidents/million vehicles)
   This is a Short segment with moderate weekend traffic. A large proportion of incidents are likely localized merge/diverge issues near interchanges.

5. MEMORIAL DR SE (20.4)
   This is a very short arterial stretch with a relatively high rate per million vehicles. This could be explained by high-conflict intersections or sudden merges.

- Freeways/Skeletal Roads like Deerfoot dominate absolute incident counts, but Arterial Streets like 32 AV NE have even higher per-vehicle rates. Possible contributing factors could be intersection density and mixed traffic.
- Weekday vs. Weekend Risks
    - Based on the incident rates calculated from the data, weekdays seem to have higher rates on commuter-heavy routes (Deerfoot, Anderson).

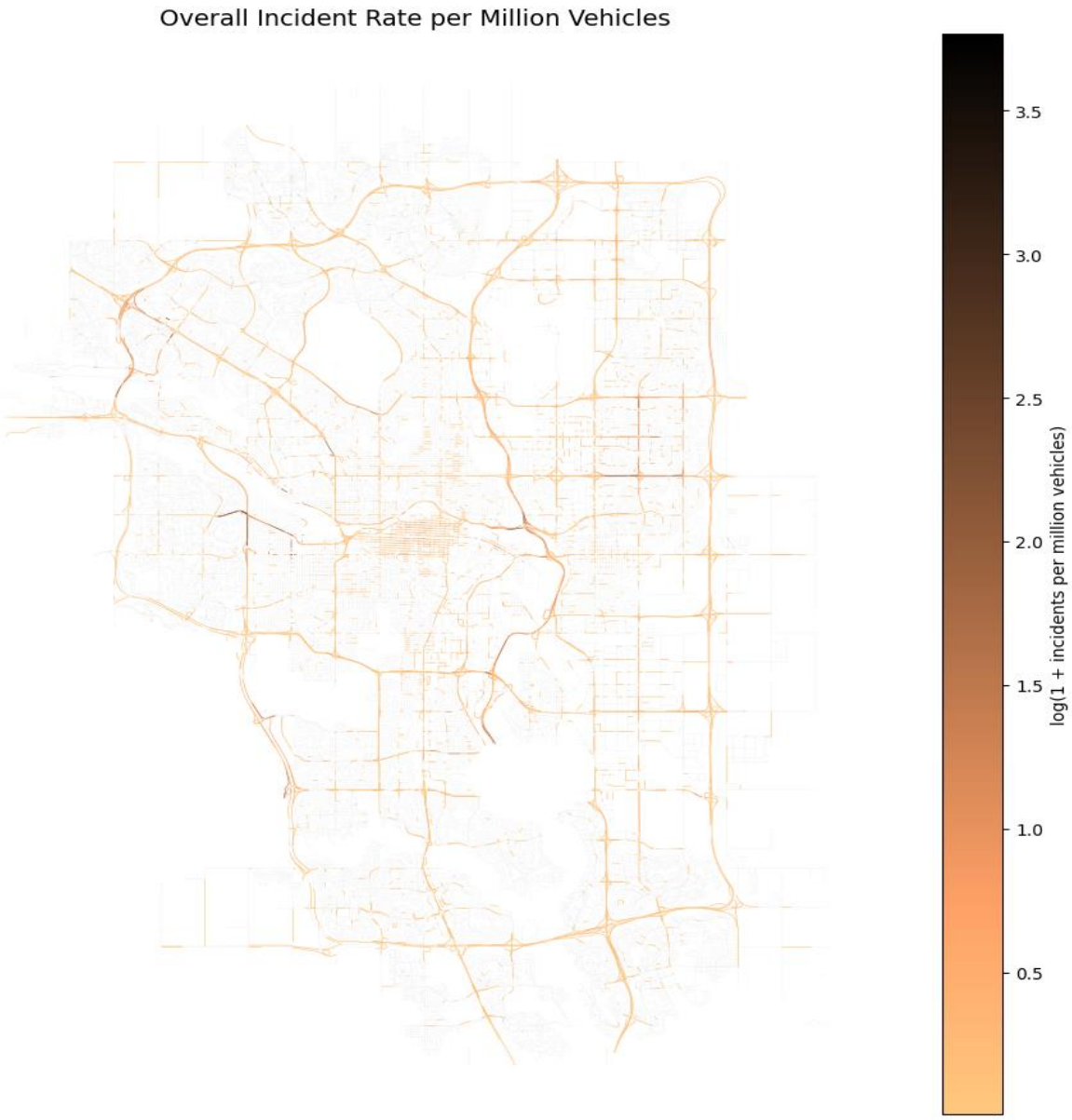
- Weekends have a relative rise in risk for shorter arterial road segments' risk ranking (Memorial Dr SE) but this is possibly due to recreational trips, distracted driving, and more varied driver skill levels(mixed vehicle types).

- Length-adjusted interpretation is important . Shorter road segments with high rates indicate concentrated safety issues rather than network-wide problems.

Road Segments by AAWT (Busiest Roads)

The above plot shows the traffic volume on Calgary's roads. The lighter colored segments indicate lesser volumes of traffic where as the darker colored segments indicate higher volumes of traffic. The unit of traffic volume is Average Annual Weekday Traffic vehicles/day. We used log transformation to reduce skewness in data.

Incident rate map



Road segments with zero incidents per million vehicles (or missing data) are drawn very faintly in light gray making them visually distinct but subdued.

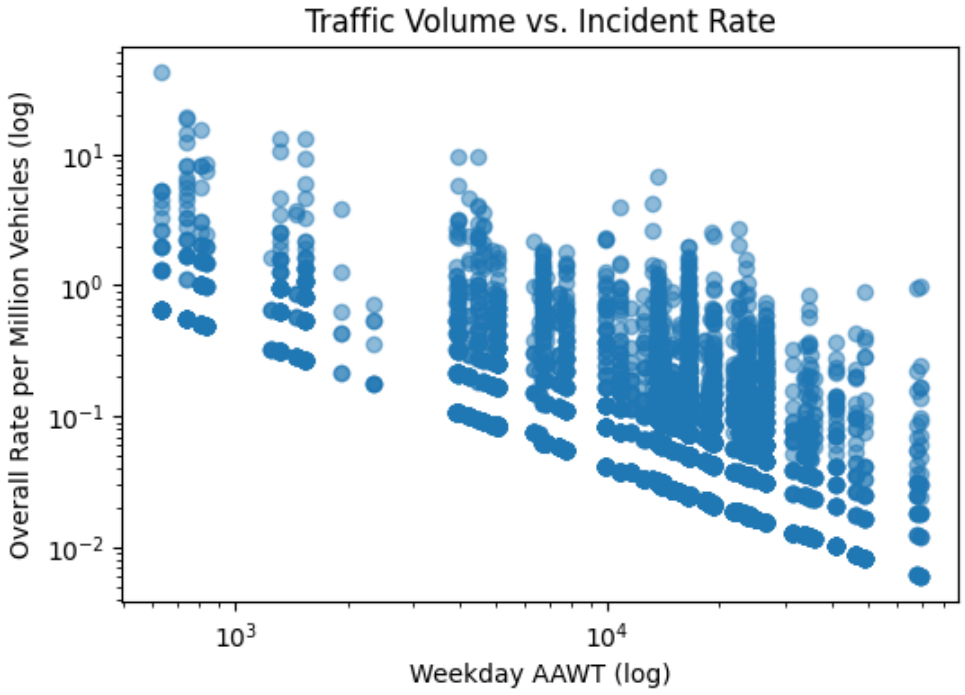
Road segments with nonzero incident rates are colored using a continuous copper colormap where colors represent the log-transformed incident rate (log(1 + rate_overall_per_million)). The log transform is applied to reduce skewness in the data (since incident rates can vary widely and be heavily right-skewed), improving contrast and making differences between low and high values easier to see visually.

The scale shows lighter colours (yellow/orange tones) = lower incident rate and darker colours = higher incident rate. The darker segments are the *hotter* areas in terms of rate, and the lighter lines are the low-rate areas. The road segments with zero incidents per million vehicles are drawn very faintly in light grey.

Some stand out areas in the city from the map are:

- Downtown core stands out as one of the brightest areas, meaning a lot of segments with higher incident rates per million vehicles. However, the incident rates are lesser compared to the volume of traffic on these roads.
- South-central & south-east arterial corridors show several bright stretches heading out of downtown (near Calgary Zoo) likely high-risk arterials with more incidents relative to traffic.
- The interchange between Sarcee Trail SW and Bow Trail SW, near Strathcona park is one of the hotspots for traffic incidents
- Major interchanges on the east side of the ring road show several bright arcs at key junctions.

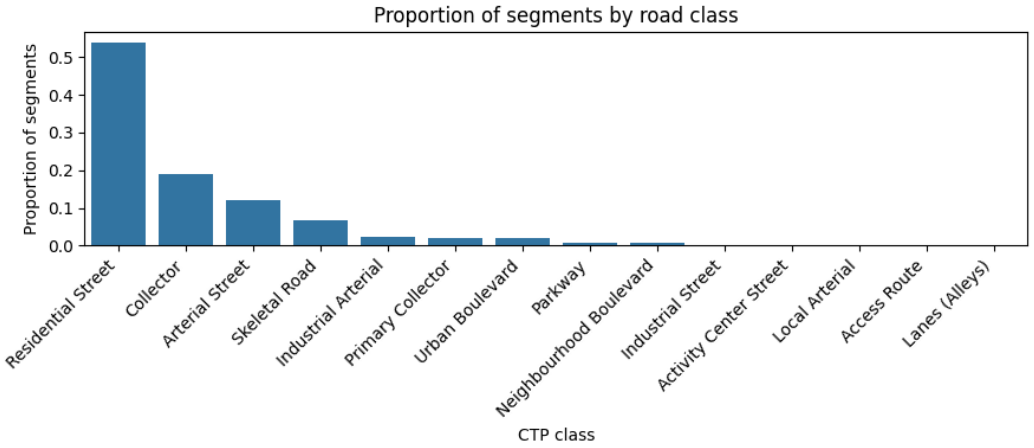
Traffic volumes vs incident rate(incidents per million vehicles)



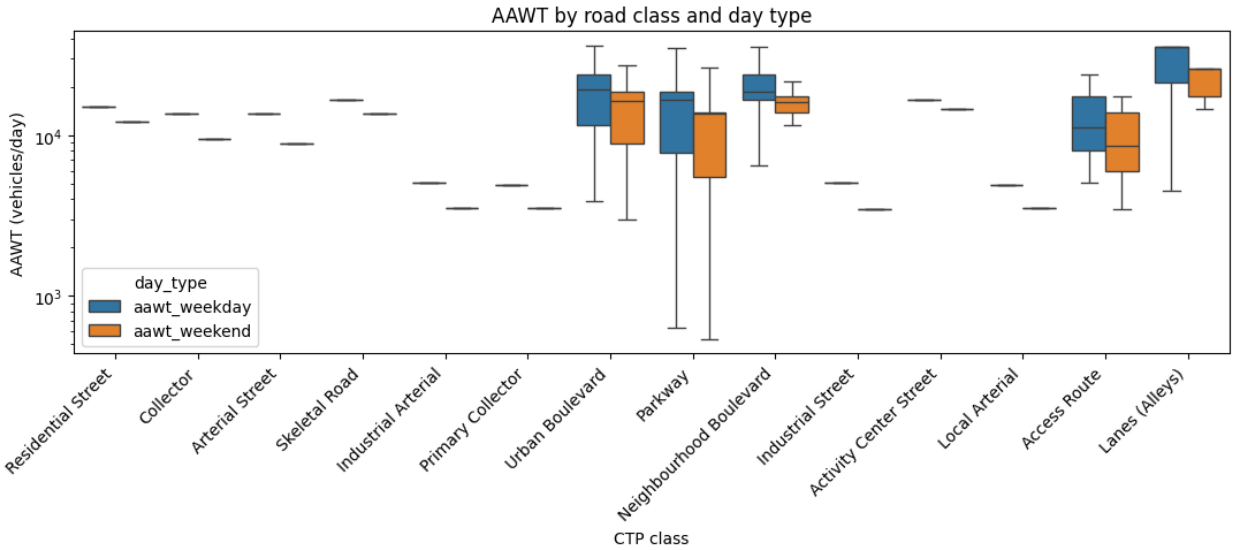
The downward trend shows that high-volume segments tend to have lower rates (incidents per million vehicles). The vertical stacking of points shows how some aawt_weekday values are reused across multiple segments (due to our nearest station-based imputation).

The wide breaks in the low-volume end (<10**3 weekday AAWT) could be indicative of many low-volume road types like residential roads.

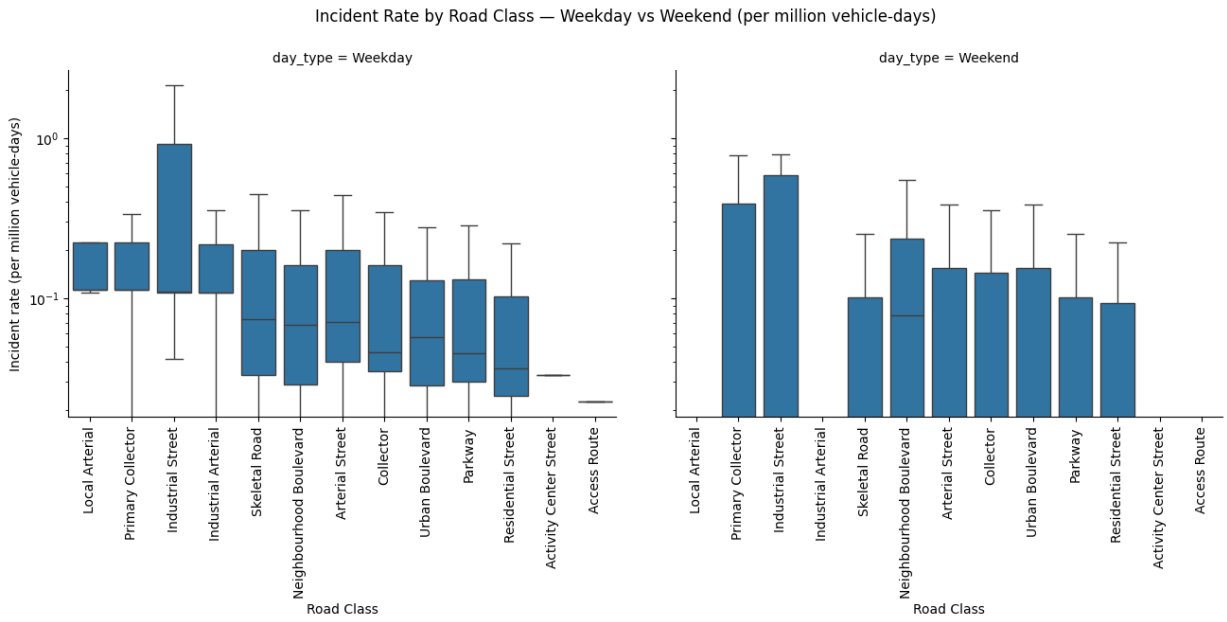
Proportion of Segments by Road class



The horizontal axis represents the various street classifications defined by Calgary's Transportation Planning Route Classification system, which categorizes roads based on their intended function and design within the city's transportation network. Residential streets form the highest proportion of streets in Calgary (~ 0.53)



The plot shows traffic volumes across various road types. Week days have more traffic volumes than weekends for all types of roads. There is greater variability in traffic volumes on road types like urban boulevard, parkway and neighbourhood boulevard.



Incident rates vary notably by road class, with industrial streets and primary collectors showing the highest median rates, especially on weekdays. Weekday rates also exhibit greater variability, suggesting location-specific factors influence pedestrian risk more during working days than weekends

Model Development and Evaluation

Data Preparation

We merged our primary incidents dataset with the supplementary datasets to create a combined dataset containing road segments identified by segment ids and assigned other features representing their expected volume, incident counts, historical incident rates, road structure and spatial properties, etc. We filtered our data to the window 2018 to 2024.

To account for spatial effects, we projected the dataset to EPSG:32611 and calculated the centroid coordinates for each segment. We then assigned each segment to a 2 km by 2 km grid cell, creating a grid_id that we later used in GroupKFold cross-validation so that training and validation data came from different locations. Finally, we removed any rows that did not have a target value for the modelling task.

For classification, we defined a ‘high-risk’ label which was assigned to any segment whose rate_overall_per_million value was in the top 10% of all segments recorded in our dataset of ~

46,000 segments, and created a binary column high_risk where these segments were marked as 1 and the rest as 0

Predictor Selection Strategy

Effective predictor selection is crucial to almost every modeling process ensuring we build interpretable and well-performing models. Given the variety of available features(both existing and extracted/engineered) in our data including spatial location variables (eg. coordinates, octant), temporal context variables (eg. season, peak hours, public holiday indicators), and traffic volume measures (eg. AAWT, expected number of vehicles-days), we came up with a systematic approach to identify the most relevant inputs while avoiding redundancy and overfitting.

Feature Set	Predictors
baseline	"aawt_weekday","aawt_weekend","length_km","is_major_road","x_m","y_m"
baseline + temporal	"aawt_weekday","aawt_weekend","length_km","is_major_road","x_m","y_m",
baseline + categorical	"aawt_weekday","aawt_weekend","length_km","is_major_road","x_m","y_m","ctp_class","one_way","octant","ownership"
all	"aawt_weekday","aawt_weekend","length_km","is_major_road","x_m","y_m","p_peak_hours","p_school_day","p_public_holiday","season_Winter","season_Spring","season_Summer","season_Fall","ctp_class","one_way","octant","ownership"

We came up with a multi-model predictor selection framework using three complementary models:

- LASSO Logistic Regression (L1 regularization)
    - LASSO applies a penalty on the absolute size of coefficients, forcing some to shrink exactly to zero.
    - This effectively performs embedded feature selection, retaining only those predictors with non-zero coefficients.
    - The advantage is that LASSO can isolate a sparse set of highly informative predictors, and is particularly useful when multicollinearity is present.

- Ridge Logistic Regression (L2 regularization)

- Ridge regression penalizes the squared magnitude of coefficients, shrinking all towards zero but never fully eliminating them.
- This tends to retain correlated predictors, distributing weights among them rather than discarding them outright.
- By comparing Ridge and LASSO results, it is possible to identify core essential predictors(selected by LASSO) and contextually important correlated predictors(identified by Ridge).

- Random Forest Feature Importance
  - As a tree-based ensemble method, Random Forest estimates variable importance by averaging the decrease in impurity (or accuracy) when a variable is used for splitting across many trees.
  - Each tree is trained on a random subset of predictors and observations
  - Random Forest is not affected by predictor scaling, making it complementary to regularized logistic regression.

  The final predictor set was determined by examining the overlap between the top-ranked variables from each method:

Features appearing in all three methods were considered the strongest candidates for modeling. Features that appeared in at least two of the three models were noted as likely to add value, especially if they represented different dimensions of the problem (spatial, temporal, volume).

Modeling Strategy Overview

Our modeling strategy was designed to address two goals:

1. Understanding the relationships between predictors and collision risk (inference)
2. Building an accurate predictive system for high-risk road segments(prediction)

To achieve this, the project compared supervised and unsupervised modeling approaches.

- Supervised Modeling
  Supervised models were chosen for their ability to leverage labeled data (in this case, the binary high-risk target) to directly learn the mapping from predictors to risk classification.

  Three algorithms were selected:

- Logistic Regression(Ridge and Lasso)
    - Serves as a baseline interpretable model for binary classification.
    - Supports regularization (L1/L2) for feature selection and coefficient shrinkage.
    - Produces probability estimates, enabling threshold tuning to optimize for recall, precision, or F1-score.
    - To address the class imbalance, class weighting was used and set to “balanced” which means the class weights are automatically adjusted inversely proportional to the class frequencies in the input data
- Random Forest Classifier
    - A non-parametric ensemble method capable of modeling complex nonlinear interactions between features.
    - Inherently handles mixed data types and is robust to irrelevant predictors.
    - Provides variable importance rankings for feature interpretability.
    - To help deal with class imbalance, the class weight parameter was set to “balanced_subsample” which adjusts class weights inversely proportional to the class frequencies in each bootstrap sample used for tree growth.

- Unsupervised Modeling

  Unsupervised techniques were introduced to explore whether any latent structure in the predictor space could align with observed collision risk patterns without explicit labels.

  The main method used was:

  - Kernel Density Estimation (KDE)
    - KDE estimates the underlying probability density of collisions in geographic space.
    - It is particularly useful for identifying hotspots of collision risk purely from spatial distribution, independent of the supervised classification process.
    - It provides an alternative view of the problem, focusing more on identifying risk emerging from where incidents cluster rather than why.

Rationale for Model Selection

The specific models were chosen for their individual and complementary strengths to help maximize our modeling goals:

- Logistic Regression (interpretability, feature selection)
- Random Forest (nonlinear relationships, robustness)
- KDE (spatial hotspot detection without target labels)

The diversity in our models ensures that both explanatory and predictive perspectives are captured. By comparing supervised and unsupervised methods, the project can assess how well collision risk is explained by engineered predictors versus how much it emerges purely from spatial clustering.

Training & Evaluation Strategy

Our training and evaluation framework was made to model the spatially clustered nature of road network collision risk data, while ensuring that model selection and hyperparameter tuning remained unbiased and reproducible.

Spatial Cross-Validation (CV)

Standard k-fold CV assumes independence between samples, which is violated in spatial data where neighboring segments often share traffic patterns, geometry, and environmental context. To mitigate this, the project implemented a *custom Spatial Block CV strategy:

- The study area was partitioned into 2 km by 2 km spatial blocks using road segment centroids.
- Folds were created by holding out entire spatial blocks during validation, ensuring no spatial leakage between training and validation sets.
- This helped simulate model generalization to unseen geographic areas.

Metrics for Model Selection

Our evaluation prioritized metrics that balance precision and recall, given the problem's imbalanced nature (relatively few high-risk segments(10%)):

- F1-score:
    - This is our primary model selection metric during initial predictor selection.
    - The F1 score balances the trade-off between precision (avoiding false alarms) and recall (capturing actual high-risk segments).

- Precision & Recall:
    - Precision measures whether identified high-risk segments are truly dangerous.
    - Recall asseses coverage of most high-risk areas, which is critical for public safety planning.
- ROC AUC:
    - The ROC AUC measures the model's ability to discriminate between classes over all thresholds.
- PR AUC (Precision-Recall AUC):
    - The PR AUC is more informative than ROC AUC in imbalanced settings, as it focuses on the positive (high-risk) class.

Model Tuning Strategy

The tuning process was structured in three phases:

1. Predictor selection
   Involved searching hyperparameter grids for L1 Logistic Regression, L2 Logistic Regression, and Random Forest.
   In this phase, we aim to identify high-value predictors via regularization (L1/L2 shrinkage) and RF feature importances.
2. Best Model Family Selection
   The models from the previous phase were tuned with a broader parameter search using the reduced predictor set.
   We used the average precision(PR AUC) as the refit criterion, due to the goal of addressing the class imbalance.
3. Final Model Training & Threshold Optimization
   The best-performing model from the previous phase was retrained on the full training data and then probability threshold tuned to maximize F1-score while monitoring precision and recall trade-offs.

The KDE model was not tuned along with the other supervised models.

Predictor Selection Results

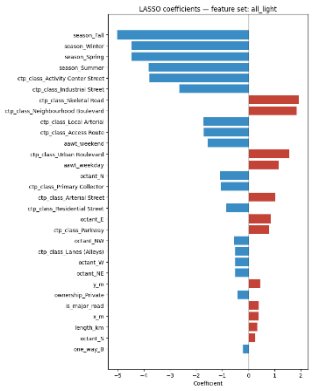
The feature selection process compared four candidate feature sets across three models using 5-fold spatial cross-validation (2 km grid) and multiple metrics: PR AUC, ROC AUC, precision, recall, F1, and accuracy.

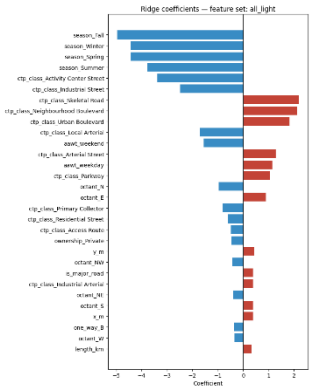
Feature Set Selection Results:

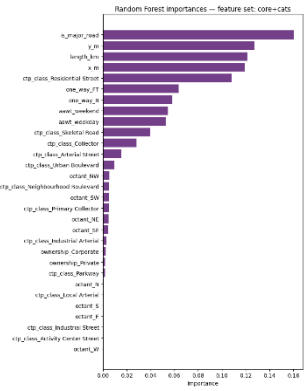
- Random Forest (RF) with the baseline & categorical feature set achieved the highest F1 score(0.552) PR AUC (0.568) and strong ROC AUC (0.892), with relatively high recall (0.646) but moderate precision (0.482).
- Logistic Regression models (ridge/lasso) showed lower PR AUC (~0.525) but still maintained competitive recall (>0.80) for the positive class, indicating they are more recall-oriented but less precise.
- Results by Model:

Model	Feature Set	PR AUC	ROC AUC	F1	Precision	Recall	Accuracy
Random Forest	baseline + categorical	0.568	0.892	0.552	0.482	0.646	0.883
lasso Logistic Regression	all	0.525	0.876	0.499	0.362	0.806	0.819
ridge Logistic Regression	all	0.525	0.876	0.499	0.362	0.806	0.819

Feature importance/rankings:







Interpretation:

The baseline + categorical predictor set (traffic volume, geometry, and categorical road attributes) appears to carry the most predictive signal. Temporal variables did not improve early-model performance, possibly because the target variable was an aggregate rate rather than time-resolved incidents.

Model Tuning Results

Next, the three models were tuned using the top feature set identified in the precious step, tuning with expanded hyperparameter grids. The models were evaluated with the same spatial CV setup and scoring metrics:

rf	{'clf__max_depth': 8, 'clf__max_features': 0.5, 'clf__n_estimators': 400}
ridge	{'clf__C': 2.2758459260747887}
lasso	{'clf__C': 0.6812920690579615}

Results Summary:

- The Random Forest model with the baseline + categorical predictor set model remained the top model by PR AUC and F1 after further tuning.
- Ridge and lasso LR retained high recall but still lagged in PR AUC due to low precision, a typical pattern for linear classifiers in imbalanced spatial datasets.

Threshold Optimization & Final Model

Model	PR AUC	ROC AUC	F1	Precision	Recall	Accuracy
Random forest	0.547	0.892	0.512	0.379	0.788	0.832
Ridge Logistic Regression	0.529	0.875	0.50	0.362	0.807	0.820
Lasso Logistic Regression	0.528	0.875	0.499	0.362	0.807	0.819

```
PR AUC=0.554 | ROC AUC=0.890 | best_F1=0.541 @ thr≈0.735
              precision    recall  f1-score   support

           0      0.949     0.920     0.934     40253
           1      0.487     0.607     0.541      5069

    accuracy                          0.885     45322
   macro avg      0.718     0.763     0.737     45322
weighted avg      0.897     0.885     0.890     45322
```

The next step used cross-validated predicted probabilities to search for the classification threshold that maximized F1 score.

- The optimal threshold for the Random Forest model was about 0.735, yielding an OOF PR AUC of 0.59 and ROC AUC of 0.89.
- At this threshold, the Random Forest model maintained a recall of 0.61 with precision 0.49, prioritizing recall over precision

The final tuned Random Forest model was retrained on the full dataset and used to generate per-segment predicted probabilities and labels for downstream mapping and risk analysis.

Model performance results summary:

- Best model: Random Forest
- Best predictor set: Baseline & categorical
- Best F1 score: Random Forest
- Best recall: Logistic Regression
- Predictive signal is concentrated in traffic volume metrics, spatial coordinates, road class, and categorical layout attributes.

Model Predictions

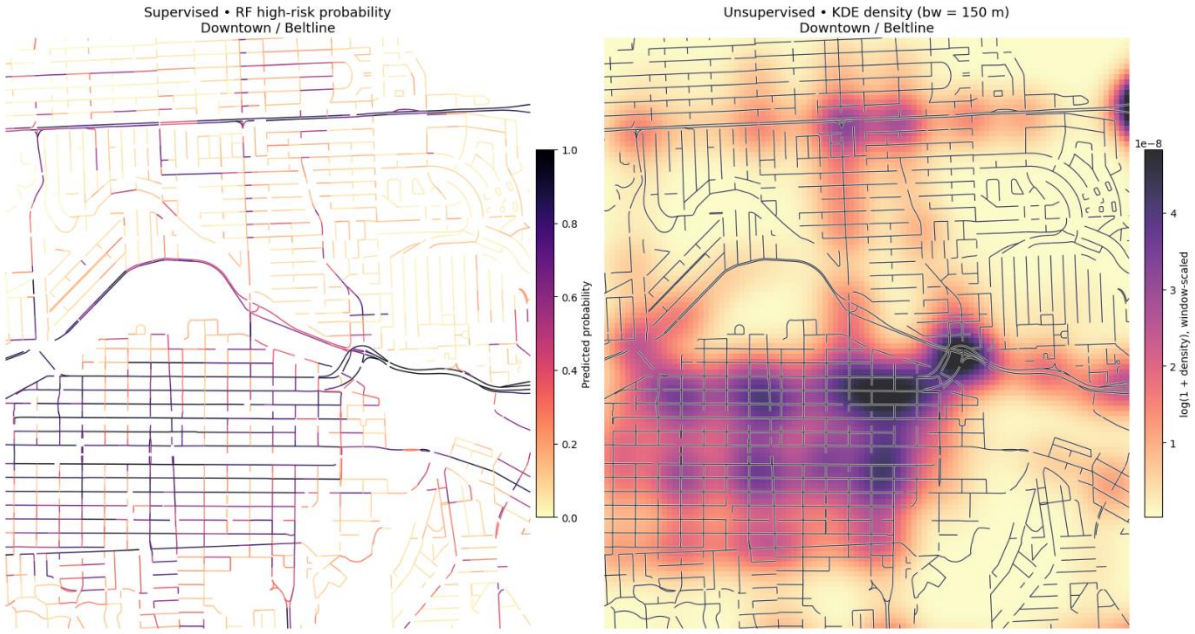
Comparing Random Forest Predictions to KDE Density Estimates

In the Random Forest maps, some arterial and connector roads between hot spots still receive moderate-to-high probability scores despite not being at the absolute centers of past incident clusters.

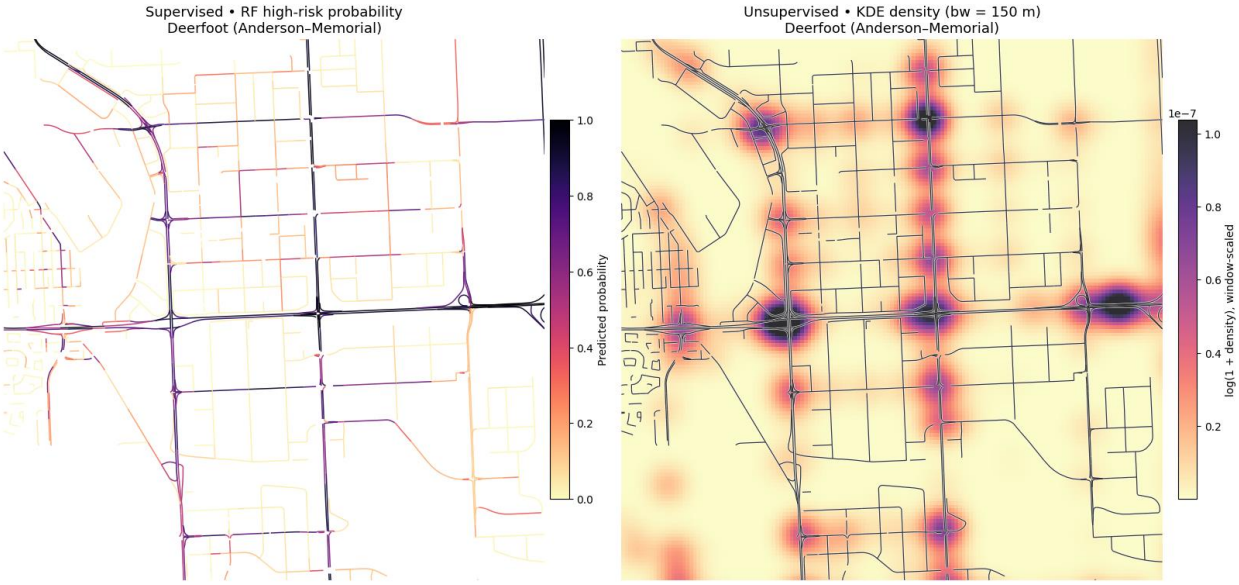
The Kernel Density Estimation (KDE) maps are spatial and unsupervised, showing the smoothed spatial distribution of past incident points. KDE highlights only where incidents have typically occurred, with intensity decaying smoothly away from the event locations. It does not consider road attributes or other predictors, meaning that high-density areas always correspond to historical hot spots, but it may underrepresent emerging risk areas when past data is limited.

Looking at the zoomed-in areas:

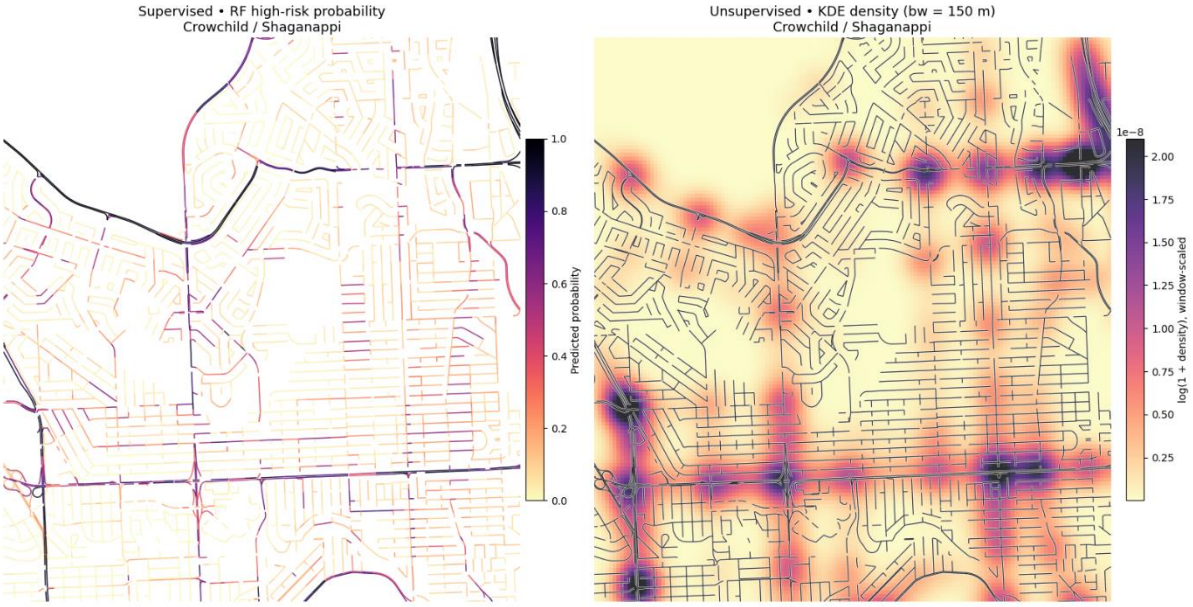
- Downtown/Beltline: Both models identify the dense grid network as high-risk, but KDE shows a very localized core around the heaviest clusters, while Random Forest spreads risk to nearby streets with similar patterns of connectivity and volume.



- Deerfoot (Anderson–Memorial): KDE strongly isolates the interchange ramps and crossing points where incidents have historically clustered. Radnom Forest, however, assigns elevated probabilities along larger stretches of Deerfoot Trail, suggesting that it infers risk from the road type even beyond the exact historical hot spots.



- Crowchild/Shaganappi: The KDE map captures high densities at major intersections and merges, while the Random Forest model expands elevated risk along the corridors feeding into them, reflecting the influence of road classification and traffic patterns.



Overall:

- KDE seemed to be better for descriptive hotspot mapping
- Random Forest is better for predictive risk scoring
- Using them together can help differentiate between current, historically proven high-risk zones (KDE) and potentially emerging high-risk zones (Random Forest).

Data Insights

What the models found

From the supervised learning approach, the Random Forest model consistently surfaced predictors tied to traffic exposure and road function as dominant risk indicators:

- Traffic volume metrics (AAWT, estimated daily traffic) were among the strongest predictors in the tree-based model.
- Road class and functional category (eg. major arterial vs. residential) played a major role in separating high and low-risk segments, especially when interacting with geometry features such as road segment length and octant orientation.
- Categorical infrastructure attributes (road classification, direction) emerged as important refinements. While they didn't dominate overall, they were often critical splits in high-risk subregions.

For linear models (ridge/lasso logistic regression), the dominant features were similar but with a slightly stronger emphasis on the geometry-normalized traffic measures and speed-related variables, due to the shrinkage-based feature weighting that penalizes collinear raw volume measures.

From the unsupervised perspective, clustering on the predictor space without incident labels revealed:

- High-traffic arterial segments naturally cluster together, and many of these clusters aligned with supervised high-risk predictions.
- Clusters driven by low-traffic, complex road intersections occasionally matched high supervised risk even when historical incident rates were modest which suggests that road design complexity might be an early risk indicator even before crash counts rise.

Agreement and divergence between modelling approaches

- Both approaches point to traffic exposure (volume and road type) as the primary driver of relative risk.
- Certain spatial pockets (eg. dense downtown grids, high-speed suburban connectors) were consistently marked high-risk by both supervised models and unsupervised clusters.

- Supervised models, tuned for precision, often down-weighted rare but plausible risk patterns. For example, isolated intersections with unusual layouts but low historical incident counts.
- Unsupervised clustering sometimes elevated these rare patterns, treating them as structurally similar to high-risk areas elsewhere.
- This divergence between approaches suggests that incident history and design risk capture a fuller safety picture than either on its own.

Real-world implications for traffic safety planning

- For targeted interventions, the supervised models' precision bias makes them strong tools for idenitfying where to focus near-term engineering reviews and as a result minimizing false alarms for resource-limited safety audits.
- For proactive design checks, the unsupervised groupings can flag atypical road designs that have not yet produced high crash counts but may share risky characteristics with known high-risk areas.
- For policy and monitoring, combining both approaches creates a reactive (based on incident history) and proactive tool (based on structural risk similarity) for idenitfying risk which gives traffic engineers a richer evidence base for decision making and planning.

Limitations

1. Traffic volume estimates for individual road segments were derived from the nearest permanent count stations. In some cases, these stations were located up to 2 kilometers away from the segment in question. As a result, the estimates may lack precision for those segments. Traffic volumes can vary significantly over short distances due to factors such as intersections, road classification, and surrounding land use. Relying on data from distant count stations does not account for this local variability, potentially affecting the accuracy of segment-level traffic analysis.

2. Our regression model for predicting incident rates per million vehicles yielded poor performance, likely due to the predictors being either irrelevant or insufficient to capture the true variability across segments. Incident rates are influenced by a complex interplay of factors such as road geometry, traffic control features, driver behavior, land use, and environmental conditions that may not have been adequately represented in the model.

3. We learnt that traffic systems are inherently stochastic and models struggle to quantify uncertainity or predict rare events accurately

6. FUTURE WORKS

1) Come up with a better technique to get the traffic volumes for segments.
2) We would like to improve our model by incorporating additional datasets such as road surface quality, driver behaviour, granular road type details

7. REFERENCES

I. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://www.statlearning.com/

II. Ledesma, C. (2021, April 15). *Spatial cross-validation using scikit-learn: Implementing cross-validation for datasets with spatial autocorrelation using scikit-learn.* Medium. TDS Archive. Medium

III. IBM. (n.d.). Feature engineering. IBM. https://www.ibm.com/think/topics/feature-engineering

IV. DataCamp. (n.d.). Feature engineering. DataCamp. https://www.datacamp.com/tutorial/feature-engineering

V. GeeksforGeeks. (n.d.). What is feature engineering? GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/what-is-feature-engineering/

VI. GeeksforGeeks. (n.d.). What is model validation and why is it important? GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/what-is-model-validation-and-why-is-it-important/

VII. GeeksforGeeks. (n.d.). Data preprocessing in machine learning using Python. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/data-preprocessing-machine-learning-python/

VIII. Scalable Path. (n.d.). Data preprocessing phase. Scalable Path. https://www.scalablepath.com/data-science/data-preprocessing-phase

IX. Federal Highway Administration. (n.d.). Road safety audits (RSA). U.S. Department of Transportation. https://highways.dot.gov/safety/data-analysis-tools/rsa/road-safety-audits-rsa

X. imbalanced-learn. (n.d.). imbalanced-learn: A Python package to tackle the curse of imbalanced datasets in machine learning. https://imbalanced-learn.org/stable/

XI. Esri. (n.d.). Kernel density (spatial analyst). ArcGIS Pro documentation. https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/kernel-density.htm

XII. Shang, L., & Li, X. (2023). Kernel density estimation for spatial data analysis. *Spatial Statistics*, *56*, 100776. https://www.tandfonline.com/doi/full/10.1080/19439962.2023.2178566

XIII. GeoPandas. (n.d.). Merging data. GeoPandas documentation. https://geopandas.org/en/stable/docs/user_guide/mergingdata.html

8. APPENDIX

Merging pipeline

1. Spatially join incidents to street segments to get incident count per street segment_id

2. Join with:

   a. street attributes (ctp_class, street_label, major_road etc.)

   b. traffic counts & AAWT volume data (weekday & weekend)

   c. major road tagging (optional)

3. Compute normalized risk metrics:

   - Incidents per million vehicles (weekday & weekend)

   - incidents per km

To merge the datasets, we aggregate counts per segment, merge them back to your street centreline dataset. Next, we merge with your AAWT data before finally computing normalized risk metrics

Results:

Joining the incidents data with the street centreline dataset involved aligning their different geommetries(point vs multiline). An exact match will always return 0/mismatch unless the point lies exactly on the polyline. Because a point rarely lies exactly on a centerline vertex/segment, we used the following strategy: a nearest join with a distance tolerance of 25 m.

The join matched 38,543 out of 55,293 incidents to a matching road segment. The missing ~200 are probably extreme edge cases (just outside the max distance of 25 m or outside the centreline dataset's coverage). We thought the spatial alignment between incidents and the street centreline data was satisfactory/very good.

With the Average Annual Weekday/Weekend Traffic per segment join, we had 38 matches out of 71 rows. This is because some permanent traffic monitoring stations have segment_id values that don't appear in the street centreline dataset.

Between our datasets, there seems to be a coverage gap between:

- The traffic volume count coverage (38 segments)
- The incident occurrence locations (tens of thousands of other segments)

Essentially, the traffic volume dataset contains roads that just didn't happen to have incidents in the dataset window. If permanent stations cover only a tiny slice of the network. Strategies we considered to expand coverage were:

- Nearest-station propagation:
  For each segment, find the nearest permanent station with a valid AAWT value. Assign its AAWT to that segment if it's within a reasonable distance (eg. < 1–2 km) and ideally on the same street_label or road class.
- Road-class average AAWT:
  Compute median AAWT for each CTP_CLASS in the segments that have data. Assign that to all segments in the same class.
- Hybrid approach
  First try nearest-station within a distance. If none found, fall back to class median. We used a 2-tier approach in the nearest station propagation. We used a 500m distance threshold as the first pass followed by a 2km one. Our motivation was to derive as representative a sample as possible and we felt this strategy for imputation would best fill in gaps in traffic volume data using nearby traffic patterns

Final Dataset

After filtering out noisy low volume segments, the final dataset had 45,298 records.Mean distance of incident coordinates to road segment geommetry is ~1.89 m, incident points are generally very close to the road centerlines. Median distance is ~0.79 m, half the matches are within a meter of the centreline.