UNIVERSITY OF CALGARY

Machine Learning Models for Predicting Calgary's Traffic Risks

DATA 607 - Project

Jincy Thomas
Joshua Quartey
Megha Radhakrishnan
Deepika Gollamandala

Introduction

- Traffic is a major part of daily life in Calgary.
- As the city grows, the risk of accidents increases.
- In **2024**, Calgary recorded approximately **10,000 traffic collisions** of all severities
- Traffic incidents impact drivers, pedestrians, cyclists, and emergency responders.
- Our project uses machine learning to analyze traffic incident data.
- We aim to uncover patterns to help make Calgary's roads safer.
- Understanding when, where, and why incidents occur provides insights for safety improvements.

Project Tasks

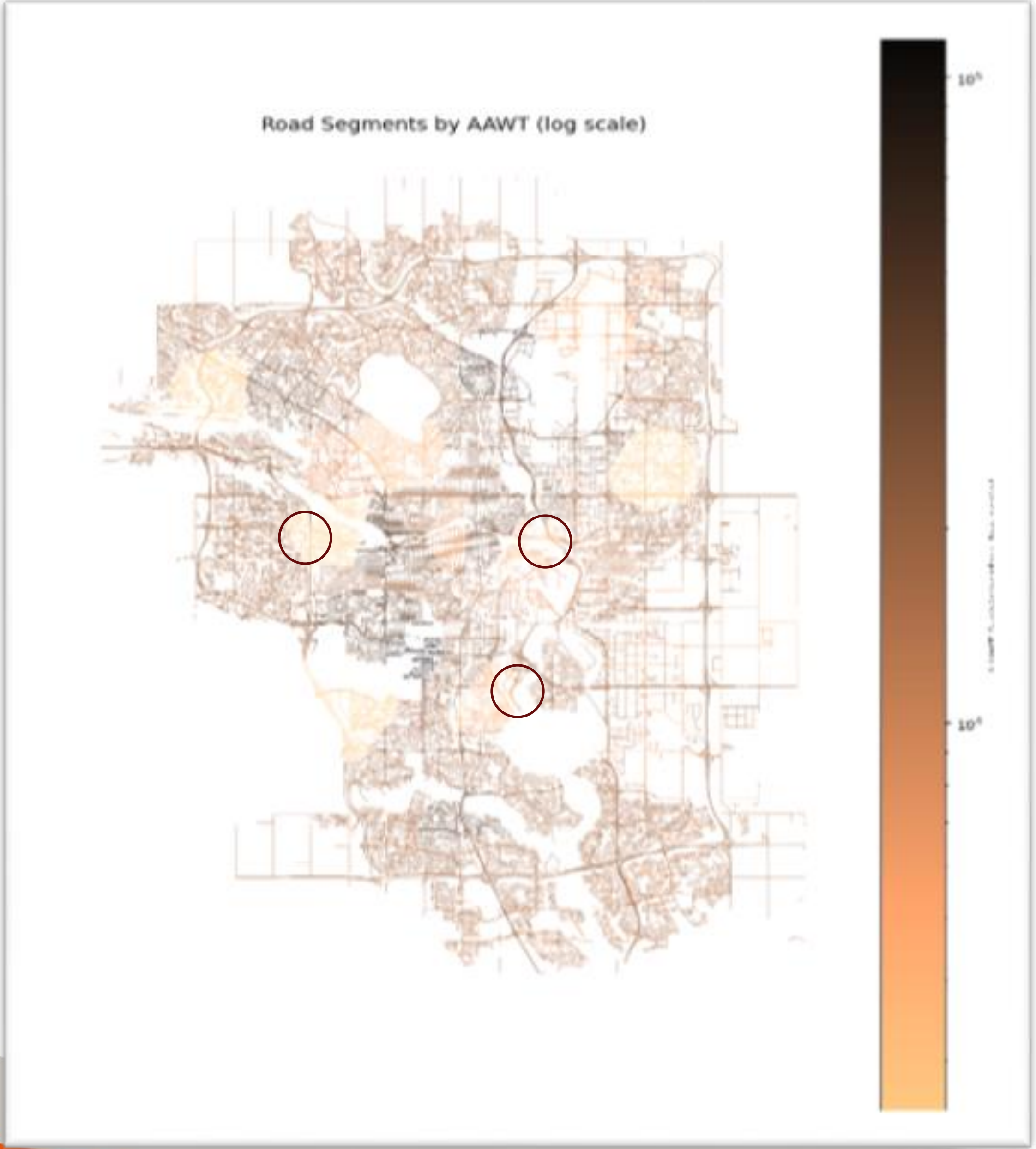
- Identify spatial hotspots of traffic incidents in Calgary using historical incident data.
- Build a predictive model to flag high-risk locations.
- Analyze factors contributing to pedestrian-related incidents and develop a model for predicting pedestrian involvement in the incidents.

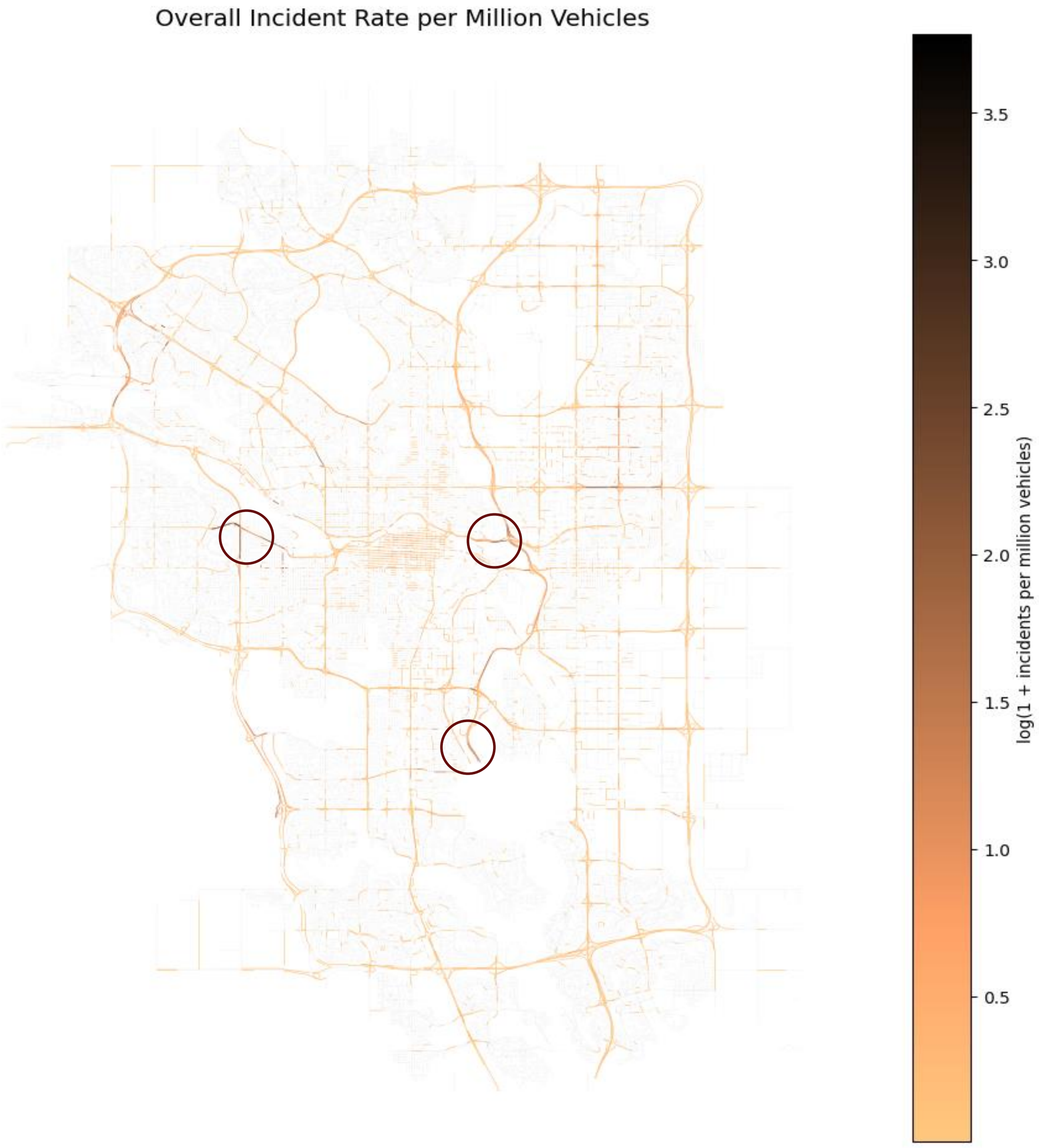
Datasets and normalizing metrics

- **Primary dataset:** Traffic Incidents Dataset 2018-2024 (Open Calgary)
- **Supplementary datasets:**
  - Weather data from Open Meteo API
  - Traffic Counts at Permanent stations
  - Permanent station locations for traffic counts
  - Major Road Network
  - Street Centerline

- **Final Dataset:** ~ 45000 road segments and their details
- **Normalized metrics:** AAWT (Annual Average Weekday Traffic), incident rate per million vehicles.

Traffic Volumes and High Risk Areas in Calgary





Pedestrian Risk

Can time, place, and weather predict pedestrian-involved incidents?

Datasets

- Traffic Incidents (July 2022 –July 2025) - Open Calgary
- Weather Data - Open Meteo API (Matched to incident's timestamp)

Class Imbalance

- Only 4.2% of incidents involved pedestrians.

Note: Pedestrian cases were **weighted ×11.85** to address class imbalance.

Modeling Pipeline

Feature extraction → Weather data merge → Weighting →Model training

Logistic Regression (Weighted), Random Forest(Weighted),Balanced Random Forest, XGBoost (Weighted)

Metrics & Tuning

- **Precision** - Accuracy of positive predictions
- **Recall** - Detection rate of actual pedestrian cases
- **F1-Score** - Balance between precision & recall

Threshold Tuning - Adjusted decision cutoff to maximize F1-score

Pedestrian Risk

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

Best Model: XGBoost (Weighted)

- Highest **F1-score**: 23.7%
- Best balance between precision and recall

Limitations

- **Low precision**: Many false positives/alarms
- **Moderate recall**: Missed many pedestrian cases
- **Limited features**: No road conditions, driver behavior, lighting and vehicle details

Predictive Model to Flag High Risk Locations

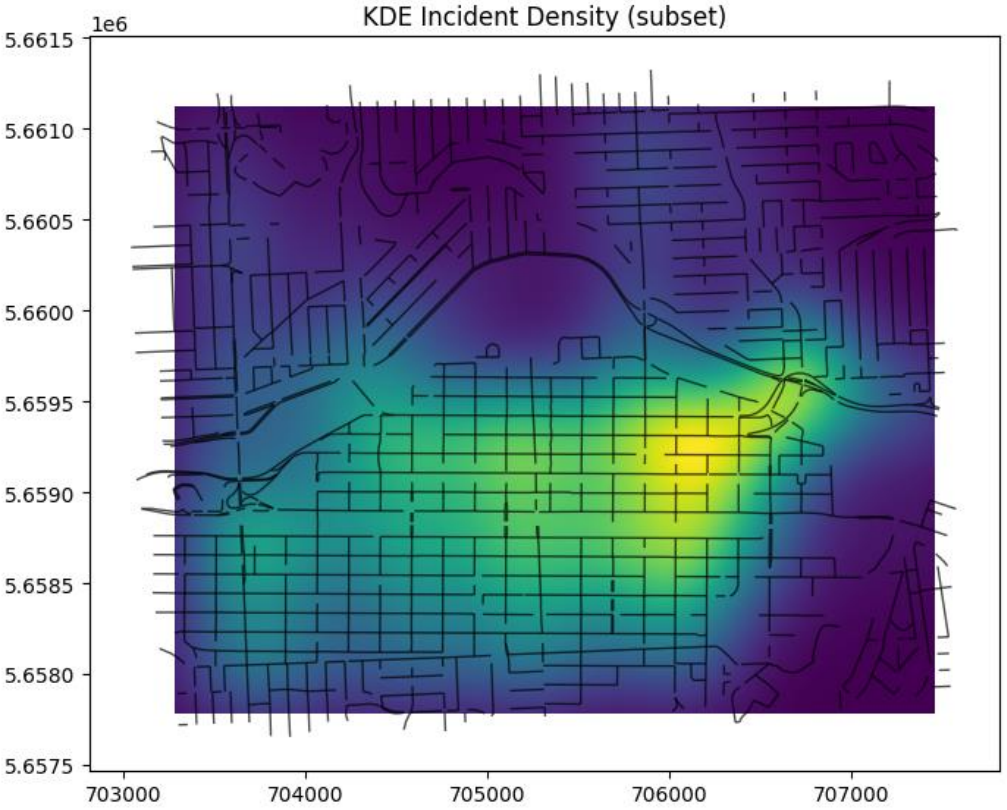
UNIVERSITY OF CALGARY

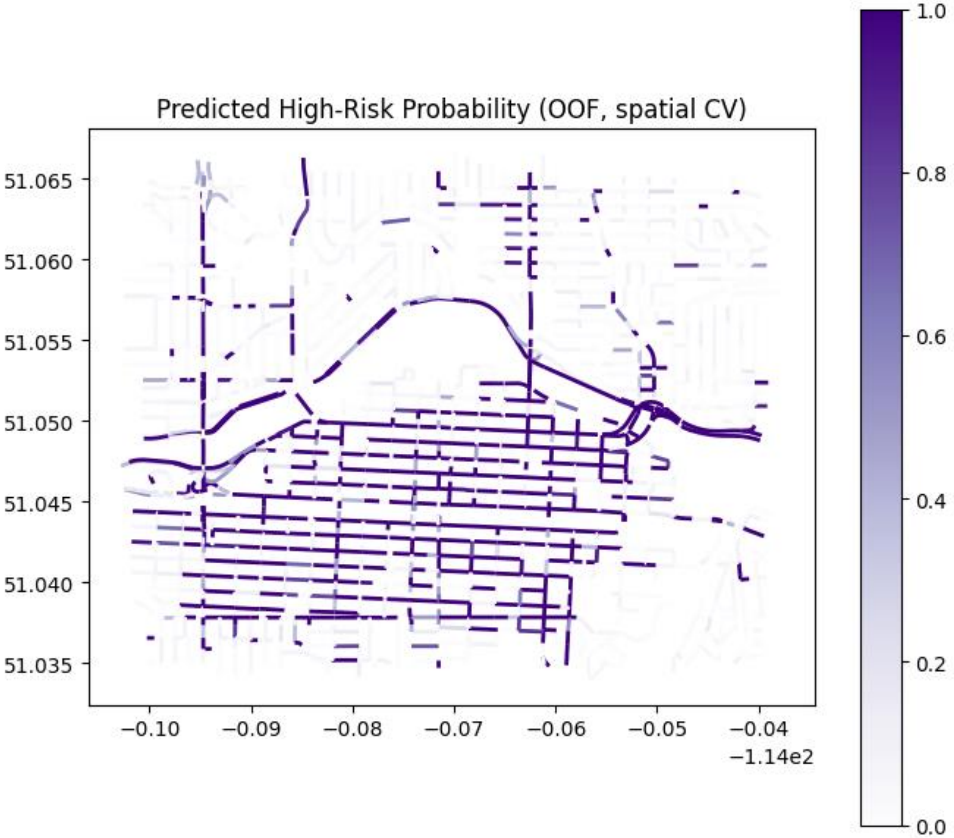
- Modeling pipeline:
  We compared logistic regression (L1/LASSO, L2/Ridge) and random forest classifiers using spatially aware cross-validation

  * Modeling results:
  Used 11 predictors across 3 categories(temporal, spatial context, and traffic exposure). Seasonal indicators consistently ranked highest in the LASSO and Ridge models while spatial context variables were ranked highest in the Random Forest model

  The random forest model emerged as the best model with PR AUC (0.497) and had a recall of 0.828 and precision 0.466 with F1 score of 0.597

KDE vs Random Forest





Limitations of Analysis

- * Incomplete traffic coverage – Missing AAWT values required propagation from nearby segments (< 2 km), potentially reducing precision in traffic volume estimates for risk scoring.
- * Class imbalance – Even with class weighting, the scarcity of high-risk segments may bias models toward lower recall or precision trade-offs in certain areas.

Future Work

- **Improve traffic volume estimates:**

Traffic volume estimates for individual road segments are not precise. These were assigned based on the nearest count station which was sometimes up to 2 kms away.

- **Find relevant datasets:**

Our current datasets lacked the information like driver behavior, speed details and road conditions

- **Modeling based on advanced methods.**

Because traffic systems are inherently stochastic and complex, models that explicitly capture uncertainty and spatio-temporal dynamics have to be used.