Classification Models for Road Safety Prediction

Student: Pak Him LEUNG AH KANG

Student ID: 500866890

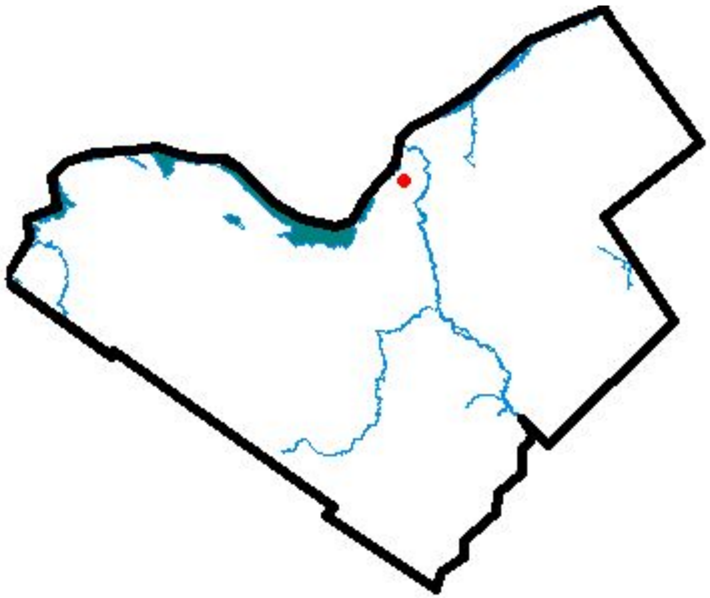
Supervisor: Tamer Abdou

Date: Dec 02, 2024

Ryerson University

Introduction

- This project focuses on the analysis of historical traffic collision data in Ottawa and identify patterns and trends that contribute to accidents and make predictions on accident hotspots



Ryerson University

Objectives and Goals

1. The motivation is to provide a useful tool for cyclists and drivers to learn about how safe their routes are.

2. The project aims to predict the chance of collision for a location accurately using historical traffic collision data

3. This initiative will help reduce road accidents and fatalities for cyclists in the light of the upcoming ban on bicycle lanes in Ontario.

Research Questions

1. What are the factors that contribute most significantly to the amount of road collisions at a specific road or location? Does the speed of the vehicle influence the collision rate?

2. Which model can best accurately predict the most dangerous locations for road traffic collisions in Ottawa

3. How do road surface conditions like wet or snowy or icy affect the probability of road collisions at various locations? Does time of the day and days of the week affect the number of vehicular collisions?

Ryerson University

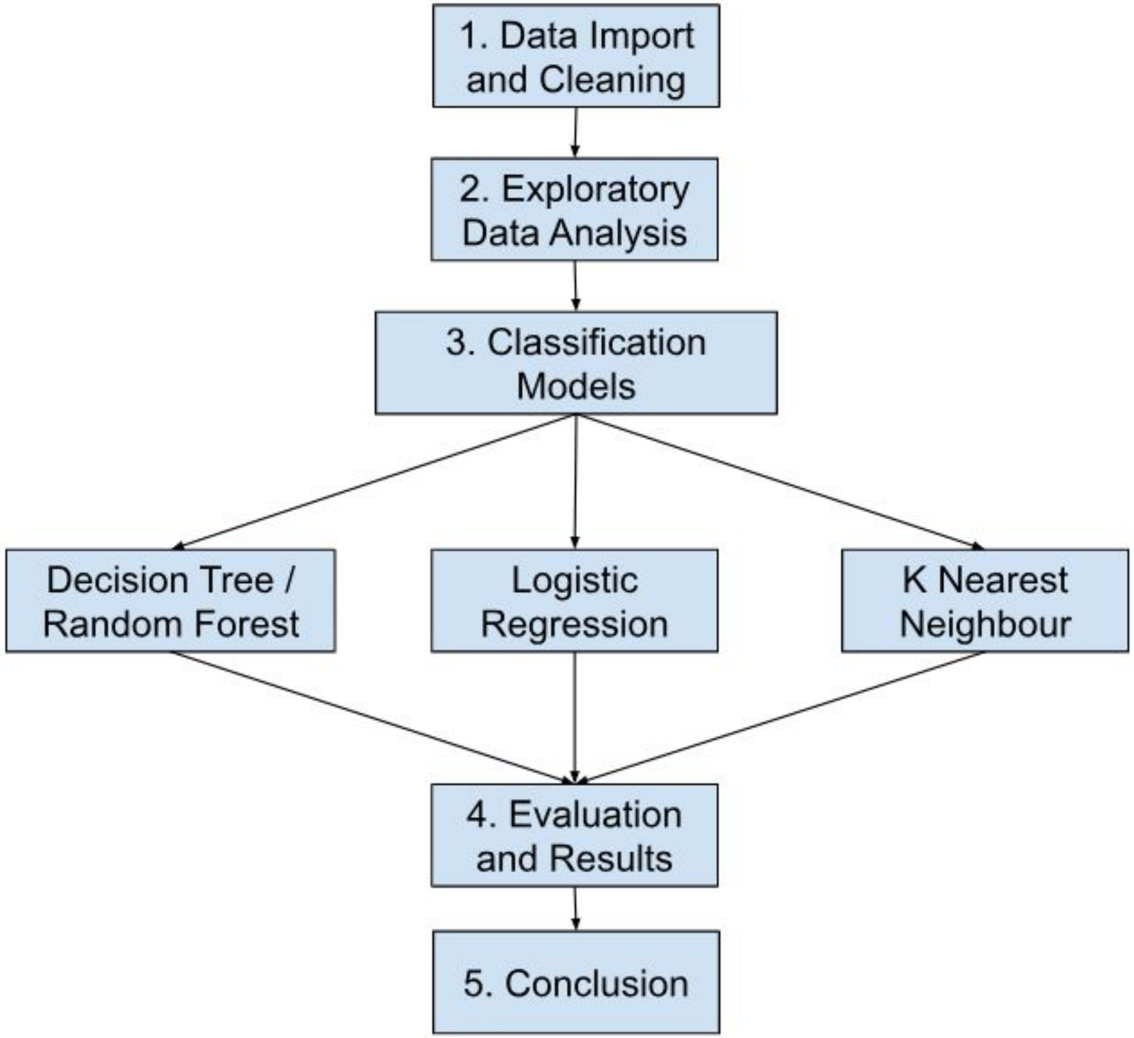
Literature Review

- Abohassan et Al. 2022 examines the positive relationship between pavement friction caused by snow and collision counts in Edmonton during inclement weather

- Mahshid Eltemasi et Al. 2024 looked into the relationships between different causes of road accidents and found that distracted driving is the most significant cause of traffic collision

- Perez et Al. 2007 determined that traffic cameras have an effect of traffic road collisions

Approach


1. Data Import and Cleaning
2. Exploratory Data Analysis
3. Classification Models
Decision Tree / Random Forest
Logistic Regression
K Nearest Neighbour
4. Evaluation and Results
5. Conclusion

Data Preparation

- The data is a csv file that can be downloaded from openottawa.ca
- The dataset contains 74,612 rows and 30 column initially and have been reduced to 74467 rows and 14 columns
- Columns have been dropped and columns that have less than 5 missing values have been replaced by a median value.
- Removed 10 duplicated rows

Data Description

Attribute	Attribute Type	Min	Max	Mean	Std	Unique
Road_Surface_Condition	Categorical / Nominal	-	-	-	-	11
Environment_Condition	Categorical / Nominal	-	-	-	-	9
Light	Categorical / Nominal	-	-	-	-	6
Traffic_Control	Categorical / Nominal	-	-	-	-	12
Num_of_Vehicles	Quantitative	1.0000	25.0000	1.841	0.58633	10
Num_of_Pedestrians	Quantitative	0.0000	3.000000	0.0223	0.15390	4
Num_of_Bicycles	Quantitative	0.0000	3.000000	0.0182	0.13535	4
Num_of_Motorcycles	Quantitative	0.0000	3.000000	0.0086	0.09431	4
Injury_Type	Categorical / Ordinal	-	-	-	-	5
Num_of_Injuries	Quantitative / Continuous	0.000	38.00000	0.2331	0.57171	10
Num_of_Fatal_Injuries	Quantitative / Continuous	0.0000	2.000000	0.0019	0.04688	3
Lat	Quantitative	0.0000	45.5249	45.292	1.83627	42382
Long	Quantitative	-79.23	-75.26158	-75.71	0.16750	42165
Accident_Timestamp	Ordinal	-	-	-	-	70045

Ryerson University

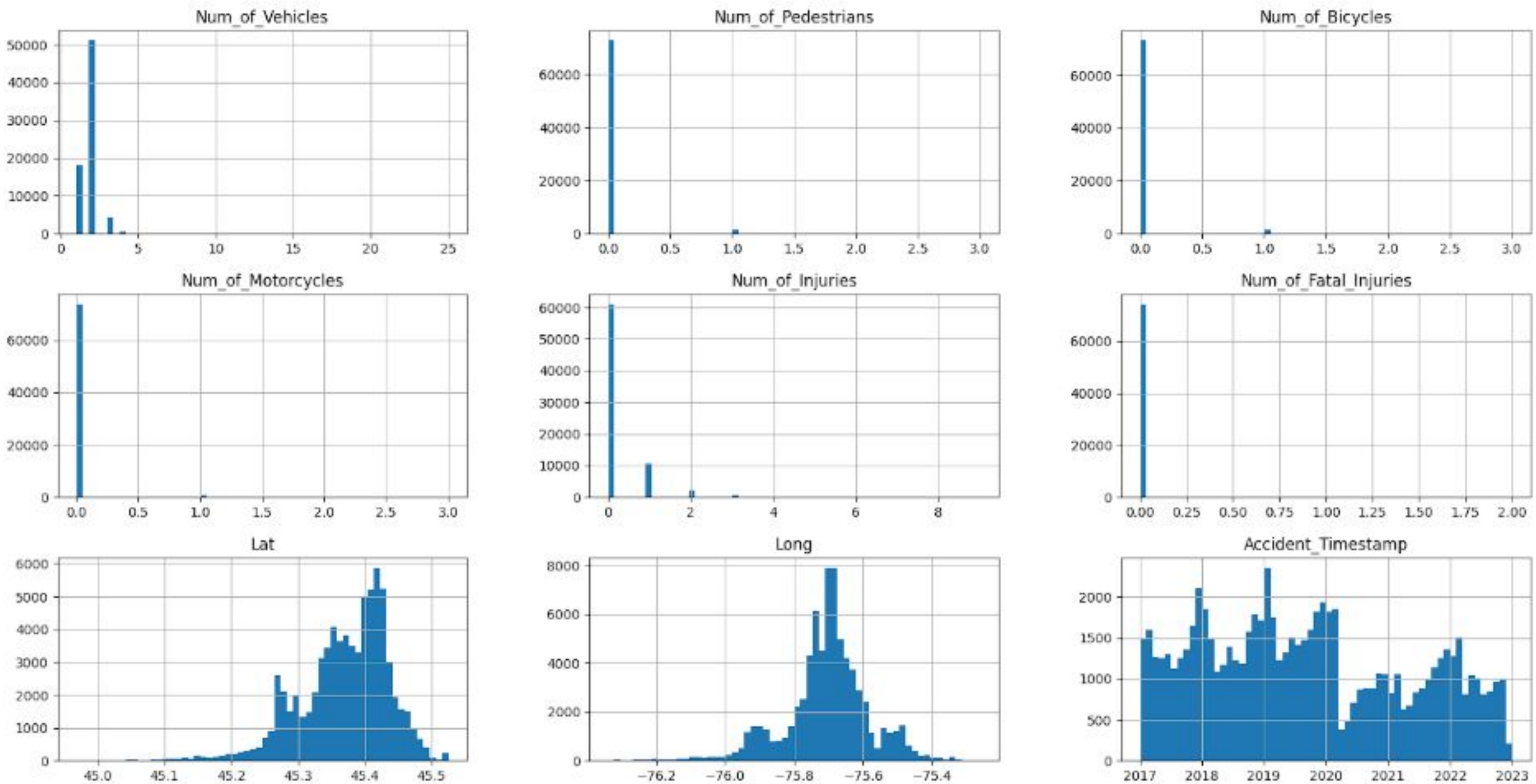
Feature Engineering

- **Collision** boolean column will be created using Num_of_Injuries > 0

Encoding

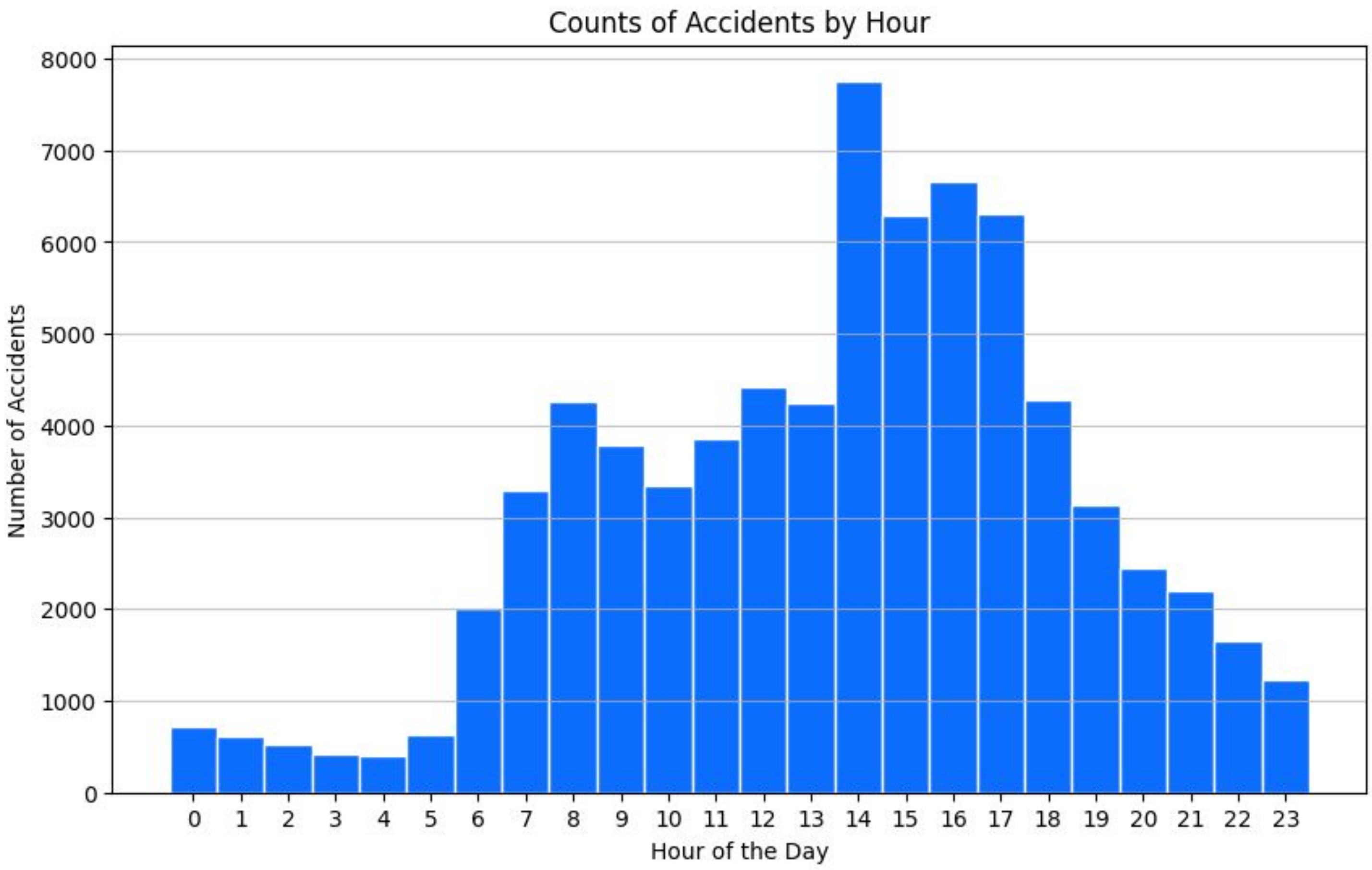
- **One Hot encoding** due to low cardinality

Exploratory Data Analysis (EDA)

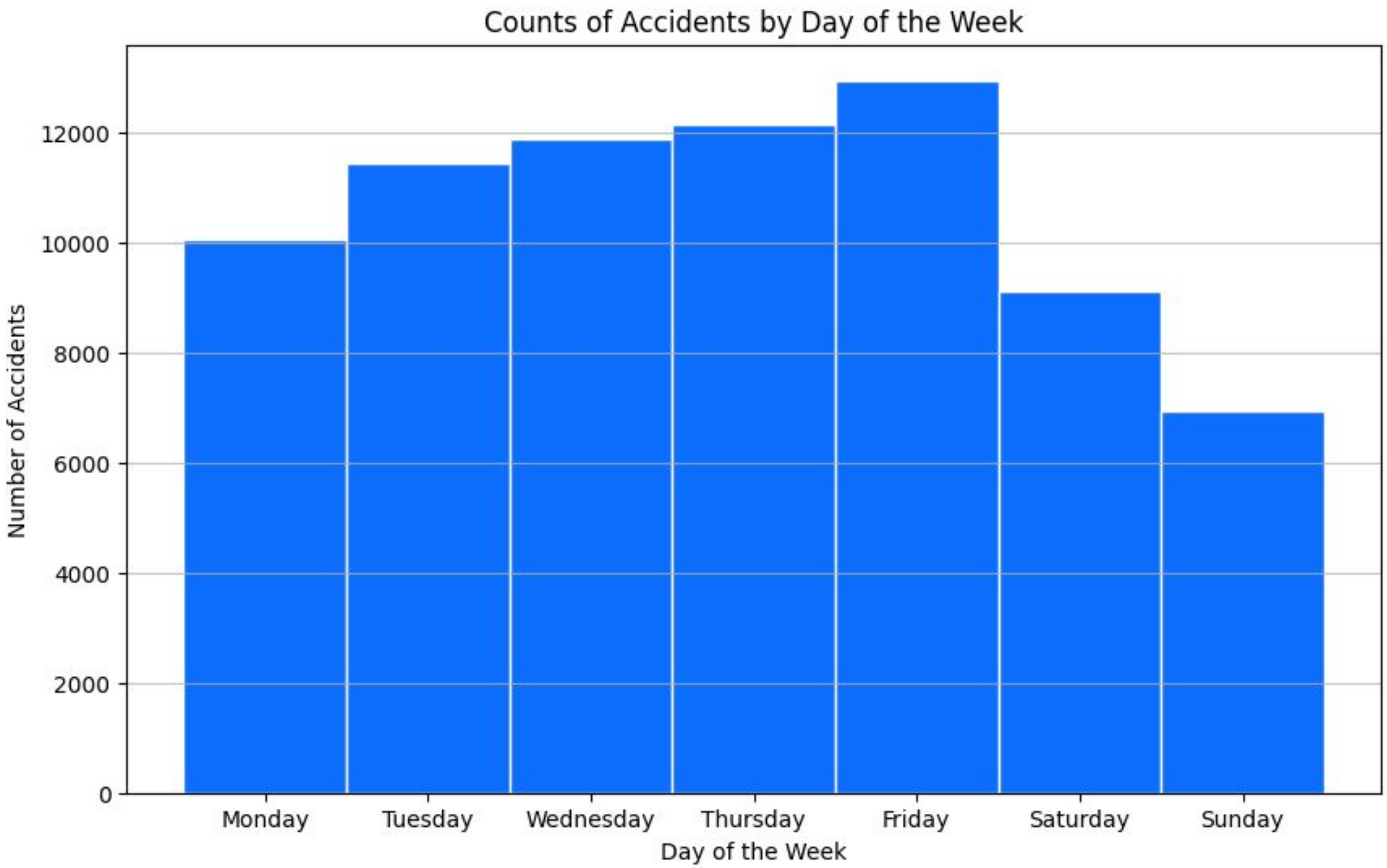


Very Few outliers

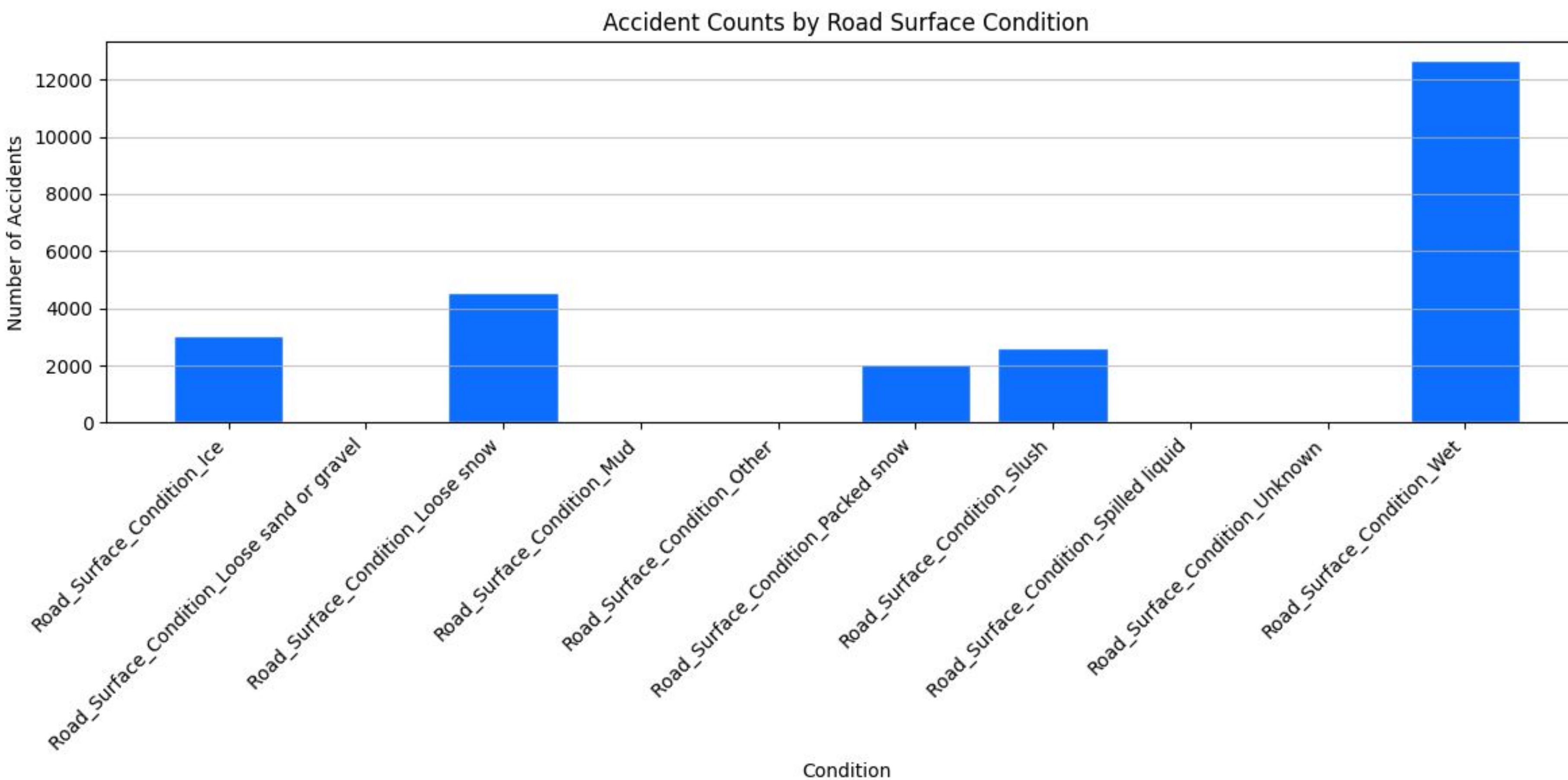
Trends: Accident count by hour



Trends: Accident Count by Day of the week

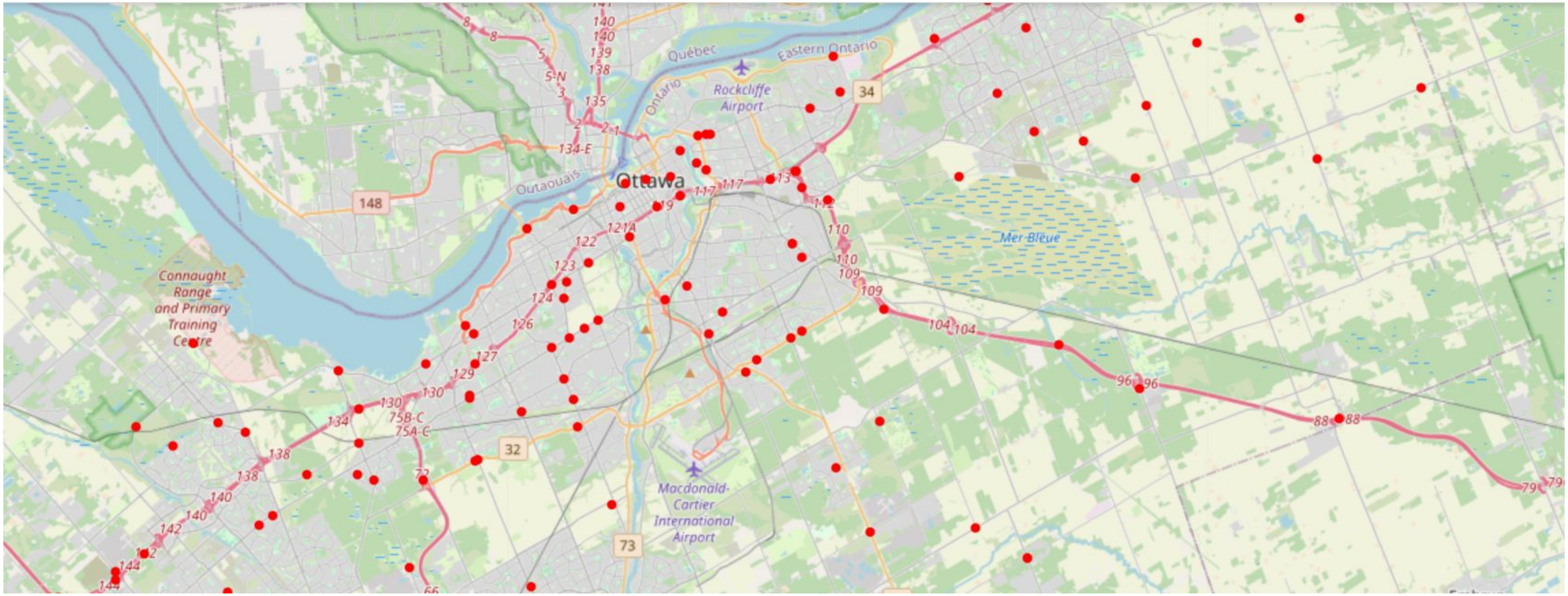


Trend: Accident Count by Road Surface Condition



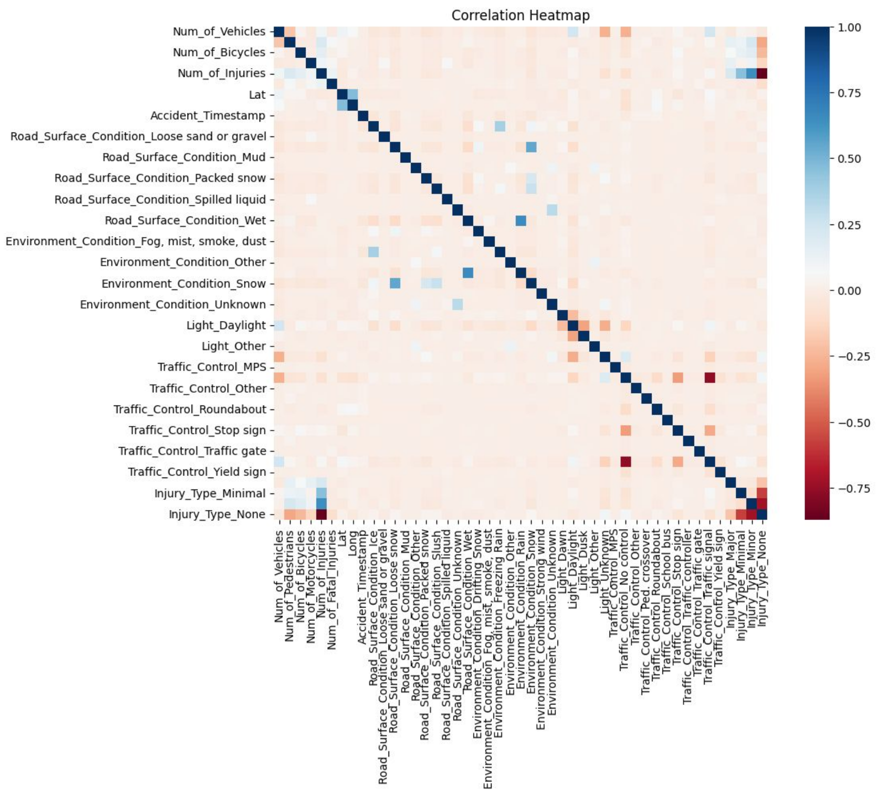
Ryerson University

Fatal Injury Geospatial Map



Correlation Heatmap

- Strongest correlations between road surface condition and rainy weather



Cardinality

- The categories in this dataset have Low Cardinality

```
Traffic_Control: 12 unique values out of 74467 rows.
Light: 6 unique values out of 74467 rows.
Road_Surface_Condition: 11 unique values out of 74467 rows.
Environment_Condition: 9 unique values out of 74467 rows.
```

Chi Square Test

- Examines relationship between categorical values

```
Traffic_Control: Chi-square p-value = 0.0000
Light: Chi-square p-value = 0.0000
Road_Surface_Condition: Chi-square p-value = 0.0000
Environment_Condition: Chi-square p-value = 0.0000
```

Ryerson University

Classification Models

1. Split Data into 80% training and 20% test set
2. Choose target variable **collision** which is a boolean.
3. Run the following algorithms
    a. Logistic Regression
    b. Decision Tree
    c. K Nearest Neighbor
    d. Random Forest (Optional)

Ryerson University

Logistic Regression

- Performs well for “No collision” with 0.86 precision and 0.99 recall
- Fails to identify Collisions with 0 precision and recall
- Model is not effective at predicting collisions as it classifies everything as no collision

```
Accuracy: 0.8583993554451457
Confusion Matrix:
[[12147    88]
 [ 2021   638]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.99      0.92     12235
           1       0.88      0.24      0.38      2659

    accuracy                           0.86     14894
   macro avg       0.87      0.62      0.65     14894
weighted avg       0.86      0.86      0.82     14894
```

Decision Tree

- Accuracy 86%, high precision means Model good at identifying non-collisions
- but bad at detecting actual collisions (recall = 0.99)
- Because of Support imbalance.
- Solution: oversampling, undersampling, ensemble

	precision	recall	f1-score	support
False	0.86	0.99	0.92	18333
True	0.86	0.25	0.38	4008
accuracy			0.86	22341
macro avg	0.86	0.62	0.65	22341
weighted avg	0.86	0.86	0.82	22341

Ryerson University

Random Forest

- Same accuracy, precision and recall as decision tree

```
Accuracy: 0.8580188890380914
Confusion Matrix:
[[18169   164]
 [ 3008  1000]]
Classification Report:
              precision    recall  f1-score   support

       False       0.86      0.99      0.92     18333
        True       0.86      0.25      0.39      4008

    accuracy                           0.86     22341
   macro avg       0.86      0.62      0.65     22341
weighted avg       0.86      0.86      0.82     22341
```

Ryerson University

K-Nearest Neighbour

```
Accuracy: 0.8400698267758829
Confusion Matrix:
[[11788   447]
 [ 1935   724]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.96      0.91     12235
           1       0.62      0.27      0.38      2659

    accuracy                           0.84     14894
   macro avg       0.74      0.62      0.64     14894
weighted avg       0.82      0.84      0.81     14894
```

Ryerson University

Algorithm comparison

Algorithm	Accuracy	Precision	Recall
Logistic Regression	85.84%	0 - 86% 1 - 88%	0 - 99% 1 - 24%
Decision Tree	85.73%	0 - 86% 1 - 86%	0 - 99% 1 - 25%
Random Forest	85.80%	0 - 86% 1 - 86%	0 - 99% 1 - 25%
K-Nearest Neighbor	84%	0 - 86% 1 - 62%	0 - 96% 1 - 27%

Ryerson University

Result

- Logistic Regression is the most accurate
- Requires hyperparameter tuning

Research Question

Which locations or roads are the deadliest with the highest amount of fatal injuries? Are accidents more likely to occur where there are a lot of accidents that happened in close proximity?

- K Nearest Neighbor can identify deadliest concentrations of collisions

Ryerson University

Conclusion

In this project, several machine learning classification models have been employed to predict collision probability based on historical road collision data. Decision tree and logistic regression exhibit the highest performance at 85% in accuracy score and the decision tree outscore the other by a little bit.

Ryerson University

Questions

Ryerson University