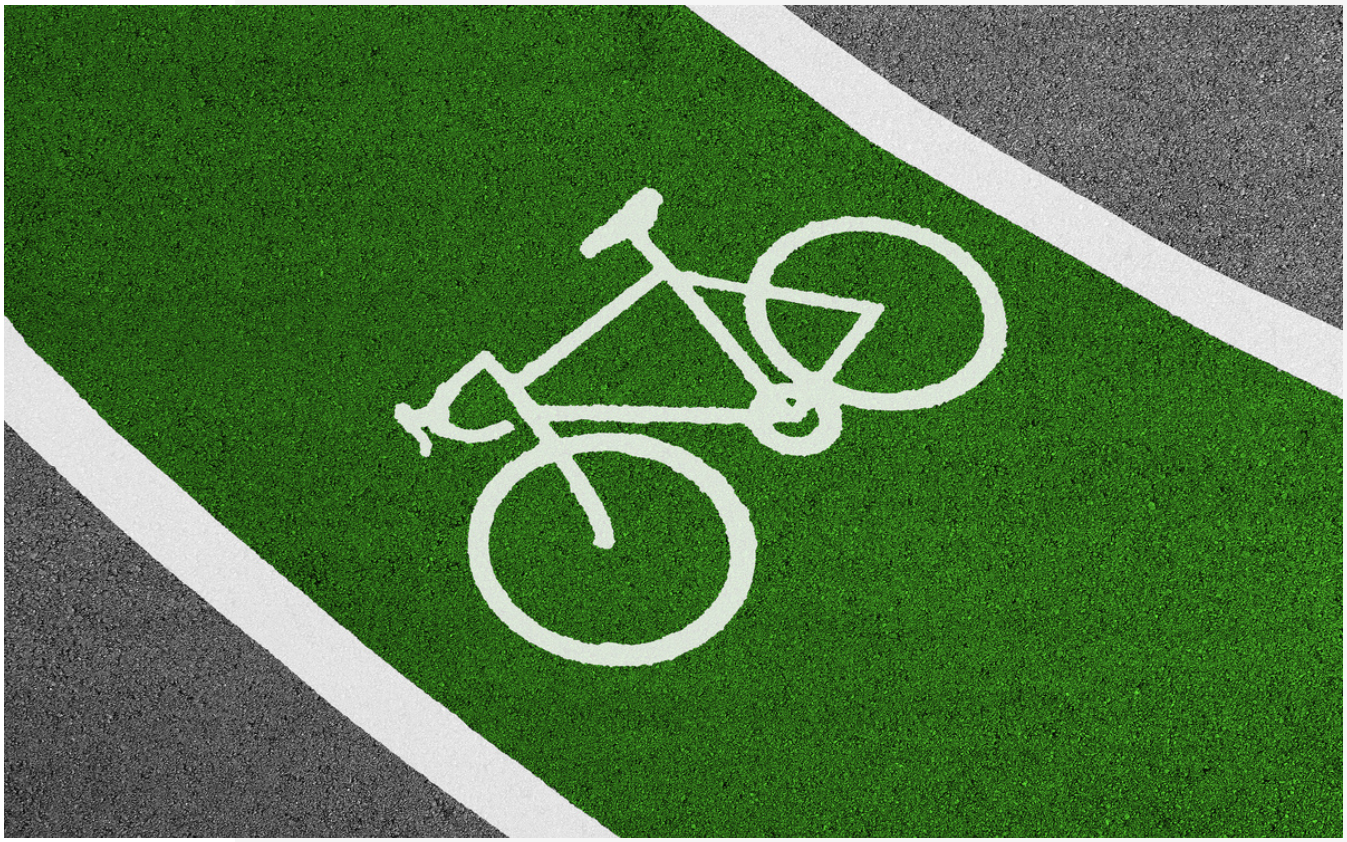
JANUARY 2023

PREDICTING CYCLISTS' TRAFFIC INJURIES BY USING MACHINE LEARNING MODELS



PREPARED AND PRESENTED BY

LEE PHAM
PROJECT LEADER

INTRODUCTION

This report summarizes the methods and findings from using machine learning algorithms to identify the trends and important factors which lead to traffic accidents involving cyclists in the city of Toronto.

RATIONALE

According to the Ontario Provincial Police, there is a 300% increase in the number of cyclists involved in fatal traffic accidents in 2022[1]. At the same time, the demand for cycling in Toronto have been growing quickly. Between February 2018 and January 2019, nearly one million cyclists travelled on the Bloor bike lane[2]. Despite how there are many environmental and health benefits that associate with riding bicycles, being a cyclist in a metropolitan city comes with considerable threats from traffic accidents.

[1] https://dailyhive.com/toronto/cycling-fatalities-ontario
[2] https://www.cycleto.ca/news/one-million-cyclists

PROBLEM STATEMENT

The aim of this project is to use machine learning algorithms to gain more insights into traffic accidents that involve cyclists and predict potentially dangerous locations or situations for cyclists on the road.

DETAILS ON DATASET

The 'Killed or Seriously Injured' dataset which is analyzed and used to train the machine learning algorithm in this project is retrieved from the Toronto Police Service's Public Safety Data Portal[1] which contain all records of traffic accidents from January 1st, 2006 to December 30th, 2020. The dataset contains 57 attributes and 16,680 records which represent each person who involved in traffic accidents.

[1]https://data.torontopolice.on.ca/datasets/TorontoPS::ksi/about

EARLY DATA ANALYSIS AND PROCESSING

The dataset is thoroughly analyzed and processed before it can be use to train the machine learning models. The dataset contains some ambiguous and missing values which was filled in by using information obtained from other features. An additional feature which reports the total precipitation in Toronto at the time of the accident is added to the processed dataset to replace the road surface conditions and visibility categorical columns which contain a consider number of ambiguous records. Since the dataset contains a large number of attributes, the attributes are grouped with other attributes sharing similar information into five groups which contain the following information regarding the traffic accidents:

1. When did the accident occurred?
2. Where did the accident took place at?
3. What were the environmental conditions at the time and site of the accident?
4. Who were the people involved in the accident?
5. What were the human factors that contribute to the causes of the accident?

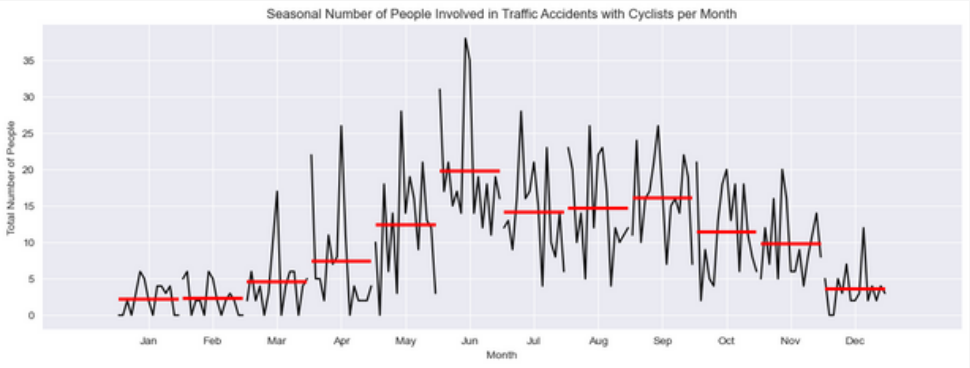


Figure 1: Graph showing the seasonality of traffic accidents involving cyclists each month from 2006 to 2020.

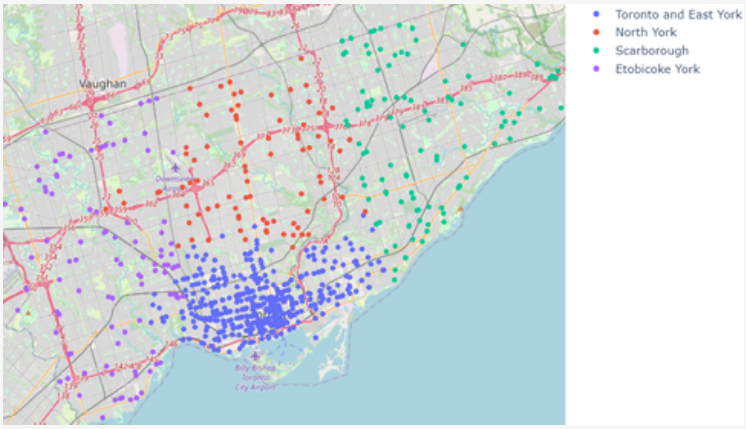


Figure 2: Map showing the locations in Toronto where the traffic accidents involving cyclists occurred which is color coded by the district name.

The findings gathered from analyzing the dataset shows that out of 16,680 people who were reported in the dataset, only 10.6% of them were involved in an accident with a cyclist. Moreover, a higher count of people involved in traffic accidents with cyclists is observed during the warm summer months with clear visibility and on a dry road surface. The location with the highest number of cyclist traffic accidents is the Waterfront Communities-The Island neighborhood which is located at the Toronto and East York district's downtown area.

Additionally, cyclists are more likely to involve in accidents with car drivers than other types of traffic. Out of 90.8% of accidents that involve car drivers, 17% of them also involve a cyclist. While a majority of 93.4% of accidents involving cyclists is not fatal, 40.4% of those accidents report a minor level injury. One of the main human factors which lead to the traffic accidents that involve cyclists is failure to yield to right of way.

MODELLING INSIGHTS AND RESULTS

Both linear and non-linear supervised machine algorithms were used to explore the predictive trends found in the dataset which can be used to predict instances of traffic accidents that involve cyclists. The models included in the modelling process were: Logistic Regression Classifier, Non-linear Support Vector Machines Classifier, K-Nearest Neighbors Classifier, Decision Tree Classifier, and Random Forest Classifier. The Logistic Regression model shows that whether the accident was located in Toronto and East York district or an involver failed to yield to right of way play an important role in predicting the chances of someone getting involved in a traffic accident with a cyclist. On the other hand, the instances of a traffic accident that involve pedestrian or have a high number of people involved are least likely to involve a cyclist.

However, the non-linear model's performances suggest that there are non-linear trends that the Logistic Regression failed to capture. Moreover, the current dataset also lacks features which can distinguish accidents involving cyclists from other types of traffic accidents. For instance, locational features such as the presence of bike lanes, volume of traffic at certain time and locations, and the number of lanes a road have may help to gain a better understanding of the road designs which are dangerous for cyclists.

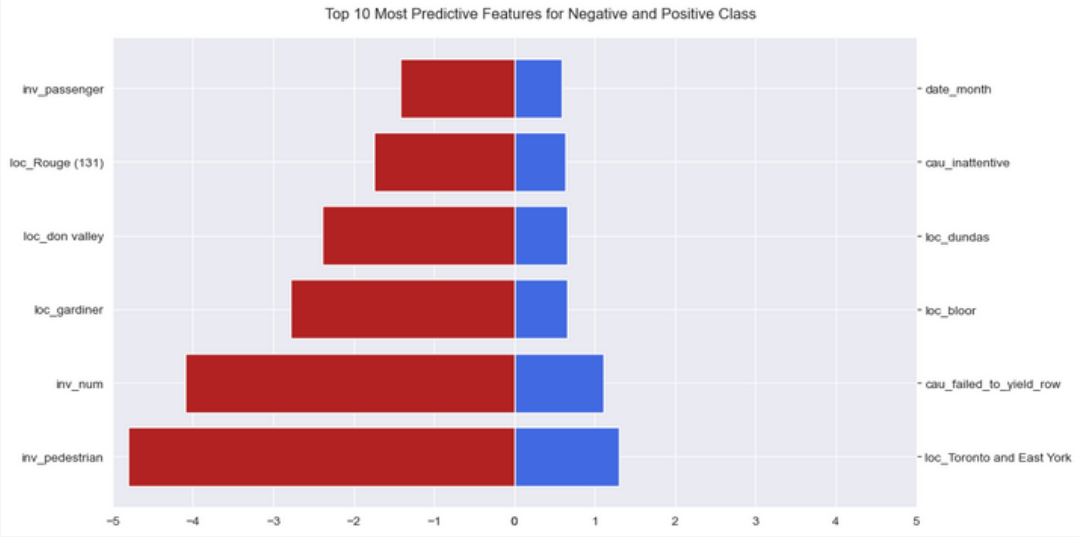


Figure 3: Graph showing the most important dependent variables used by the logistic regression model to distinguish between accidents involving cyclists (positive class) and those that do not involve cyclists (negative class).

FINDINGS AND CONCLUSIONS

Out of the best models for each classifier, the non-linear SVM model have both a high precision and a high F1 harmony score which indicate there might be clear boundaries presence between the accidents involving cyclists and other types of traffic accidents. Similarly, the K-Nearest Neighbors model also achieve a high level of precision by grouping the data points together based on their distances. As expected, the Logistic Regression did not achieve a high performance due to its limitations in processing non-linear data. The Random Forest model has the lowest performance out of all models since it couldn't make distinctions between the accidents involving cyclists and other accidents without overfitting the training data.

Figure 4: Graph showing the precision, recall and F1 harmonic scores of four different machine learning classifiers: Logistic Regression, Support Vector Machines, K-Nearest Neighbors, and Random Forest.

NEXT STEPS

The next steps for this project can be incorporate more locational attributes into the dataset to further study how the road designs can be used to predict where cyclists are more likely to get into traffic accidents. The models can also be trained again by using an updated Toronto traffic accident dataset to capture emerging trends or improvements in road safeties. In addition, the outputs from the machine learning models created in this project can be incorporated into GPS or mapping applications to help cyclists avoid potentially dangerous streets, congestions, and situations when it may not be safe to ride bikes.