Toronto Traffic Collision Fatalities: *Machine Learning*

University of Toronto
ML 3253

Amrin Krauss, Paul Soong, Sean Harrigan, Desola Odumuye

Problem statement

Background

- Traffic collisions are one of the leading causes of injuries in Canada.
- Fortunately, injury and mortality counts have been steadily decreasing over time with the introduction of new technology contributing to safer vehicles.
- Despite these technological advances, traffic collisions still result in about 100,000 serious injuries per year, with over 1,500 deaths nation-wide (Statistics Canada).
- Beyond the obvious impact on human health, traffic collisions also result in a large burden to the healthcare system as well as increased demand to first responders, and large costs inflicted on insurance companies and the legal system.

Images courtesy of DALL-E

Problem statement

- Because of the serious impacts traffic collisions impose, cities would benefit from knowing which factors are associated with serious collisions as well as being able to predict if there will be a serious collision on a given day and if so how many.

- The City of Toronto is the largest and most populous city in Canada, and would greatly benefit from being able to understand what factors are associated with serious collisions and how to predict when collisions may be more common in order to better allocate resources and promote safer driving and ultimately reduce injuries and fatalities.

Images courtesy of DALL-E

Aim

Our project will aim to predict if a traffic collision results in a fatality using the Toronto Police Traffic Collision dataset.

Methodology

Data

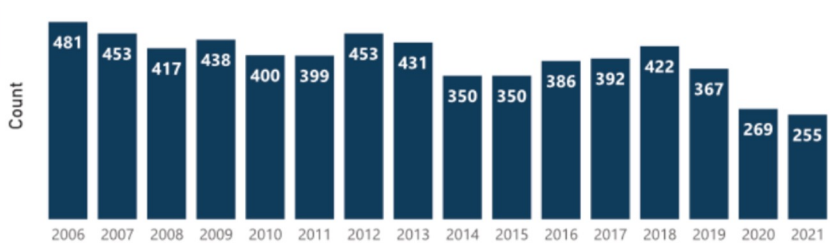
- The data used for this project comes from the Toronto Police data portal
- The Toronto police records all collisions that resulted in serious injury or fatalities
    - Killed or serious injury data (KSI)
- This dataset captures all KSI data from 2006 to present, including a variety of different variables concerning the collision conditions and passengers involved

KILLED/SERIOUSLY INJURED (KSI) COLLISIONS - OVERVIEW

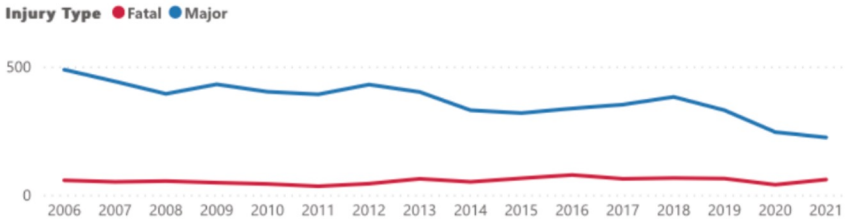
REPORTING PERIOD (2006-2021)

Prepared by
Analytics & Innovation

Collisions by Year



Injury Type Trends



Persons Involved by Involvement Category | Age Group

Filter by Year

2006	2007	2008	2009
2010	2011	2012	2013
2014	2015	2016	2017
2018	2019	2020	2021

Collisions	Persons Involved	Major Injuries	Fatal Injuries
255	284	224	60
-5.2%	-0.4%	-8.6%	50.0%

Previous Year % Change

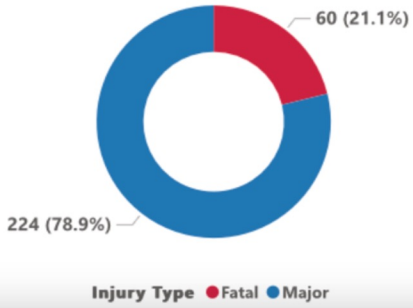
-14	-1	-21	20

Previous Year Absolute Change

Previous 5-Year Average of Total Collisions

367.2

Persons Involved by Injury Type



FILTERS

Date Range

01/01/2006 — 31/12/2021

MONTH

DAY OF WEEK

TIME RANGE

INJURY

DIVISION

NEIGHBOURHOOD

INVOLVEMENT TYP

Clear All Filters

Report Updated:
April 4, 2023

Data Available To:
December 31, 2021

Data cleaning

- The KSI data is collected at the individual-level data (an row for each individual involved in the collision)
- We were interested in the collision-level data, as our problem statement was whether or not there was a fatality in any given collision
    - We therefore grouped the data based on the the unique identifier for collision and removed all individual level data.
- The variables included in analysis are as follows:

Features, Instances and outcomes

Features: The features involved are collision-level features, which include data from the KSI dataset and a Ontario holidays dataset

→ Road type, visibility, impact type, pedestrian, cyclist, automobile, motorcycle, transit vehicle, truck, emergency vehicle, speeding, aggressive driving, red light, alcohol, disability, weekday, month

Instances: Each instance represents a collision. This could involve one or more cars as well as one more drivers/passengers/pedestrian or cyclist.

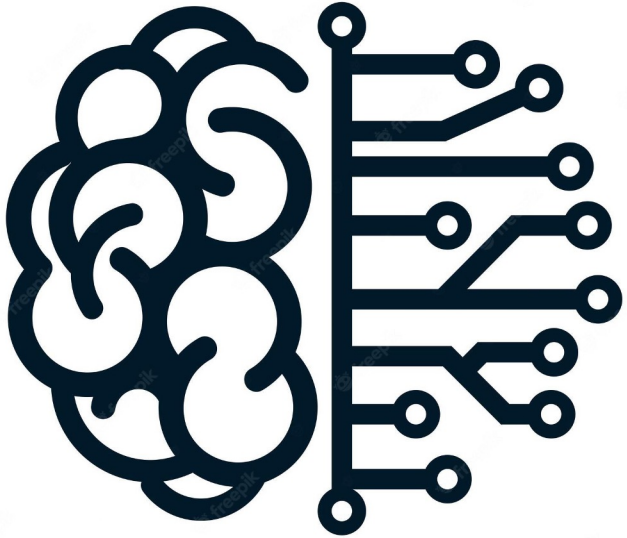
Outcome: Binary variable → whether or not there was a fatality a car collision.

ML Methodology

Train test split ratio of 70% to 30%.

Four ML algorithms:

1. Logistic regression
2. Decision Trees
3. Random Forest
4. KNN



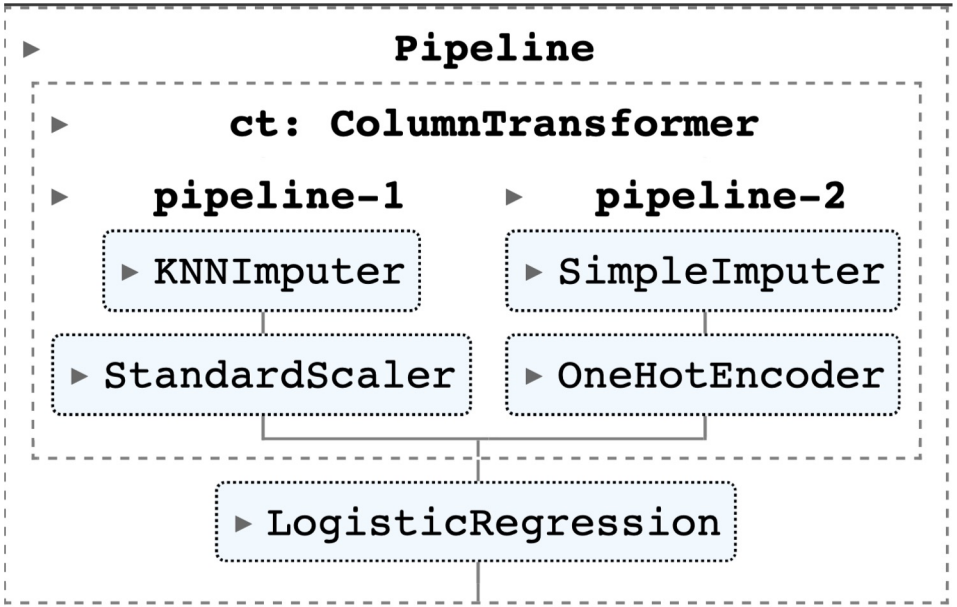
Pipeline

The image to the right is the pipeline used for all four ML algorithms

Pipeline-1: Numeric column transformers

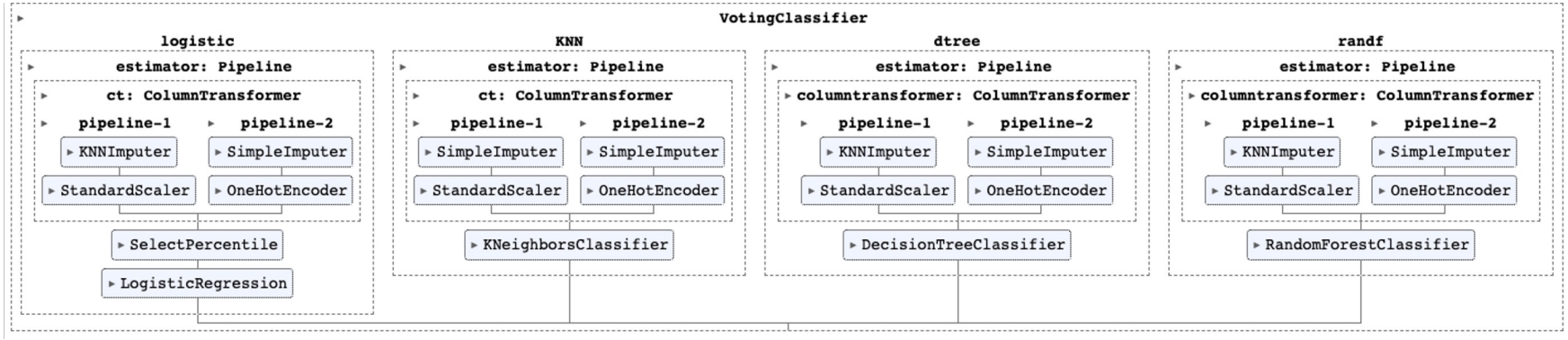
Pipeline-2: Categorical column transformers

Logistic regression replaced by appropriate ML algorithm



Voting classifier

Voting classifier using Soft voting, using all the previously mentioned algorithms and the best hyperparameters as per grid search



Hyperparameters

Hyperparameters were tuned for each pipeline via GridSearchCV 10-fold cross validation

Hyperparameter	Values
Logistics regression	
Penalty	None, L1, L2, Elastic net
C	100, 10, 1.0, 0.1, 0.01
KNN	
n neighbours	3, 5, 7, 9, 11, 13, 15, 17, 19
Weights	Uniform, Distance

Hyperparameter	Values
Decision Tree	
Criterion	Gini, Entropy
Max depth	2, 4, 6, 8, 10, 12
Random Forest	
N estimators	90, 100, 115, 130
Criterion	Gini, Entropy
Max depth	range(2, 20)
Min samples leaf	range(1, 10)
Min samples split	range(2, 10)
Max features	Auto, Log2

Scoring metrics

Our dataset was imbalance in terms of the outcome and we therefore used the following scoring metrics:

1. AUC
2. Precision
3. Recall
4. F1

We will also adjust the probability threshold cutoff of 0.50 to see if we can improve these metrics. In doing so we are aiming to better capture fatalities in the dataset, by lowering the cutoff

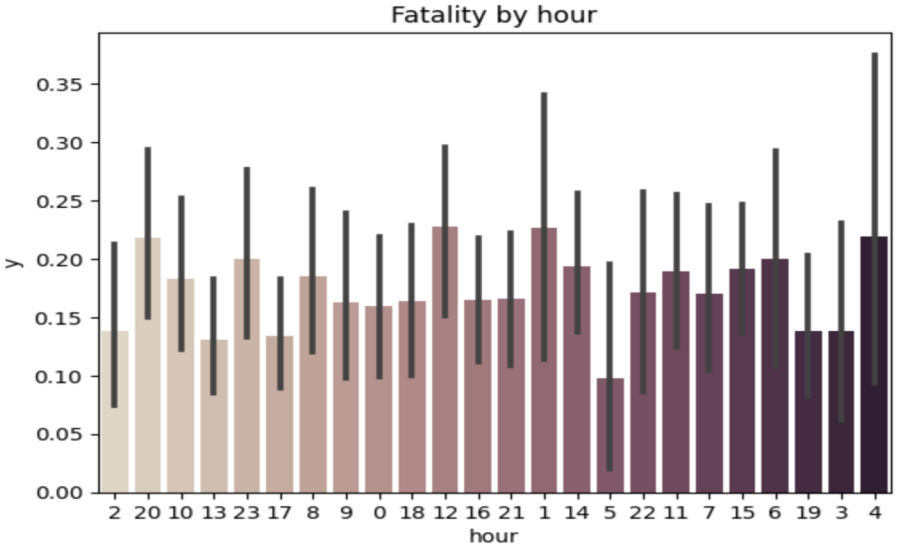
Results

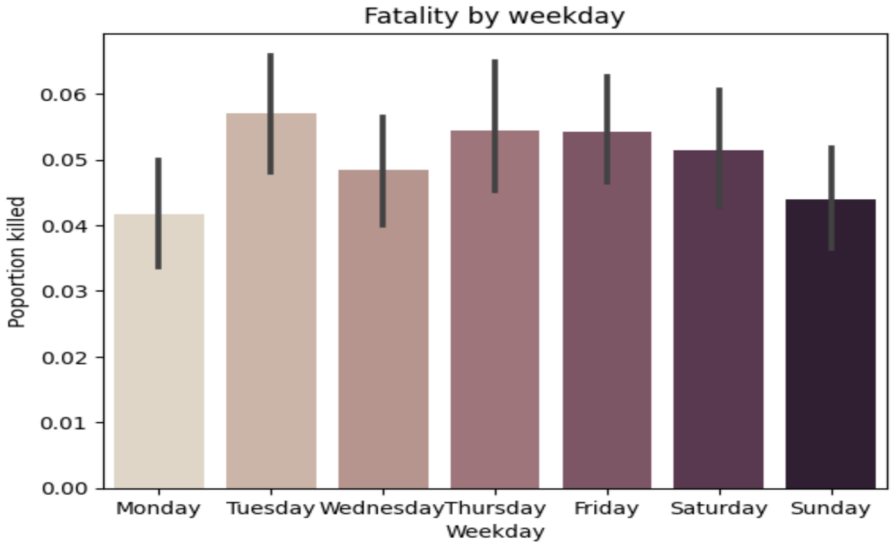
Exploratory analysis

Most collisions were centred around the downtown core of Toronto

Exploratory analysis

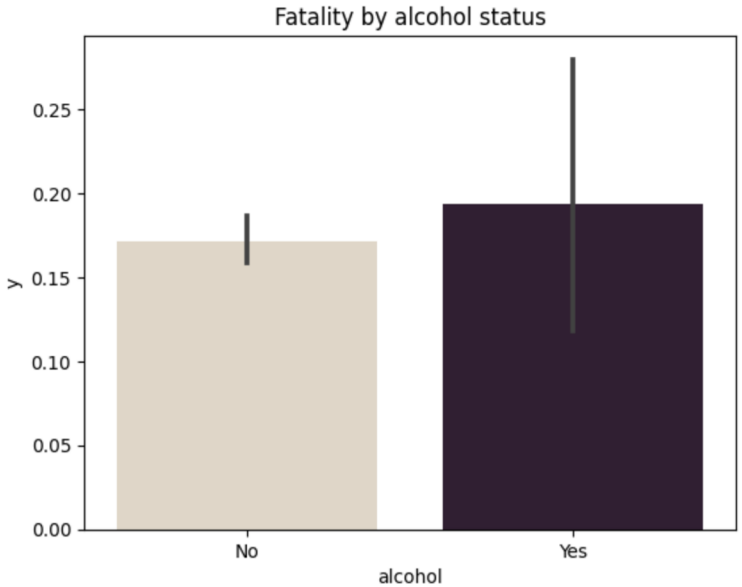
Surprisingly, there was not obvious trends across the time and the day of the week

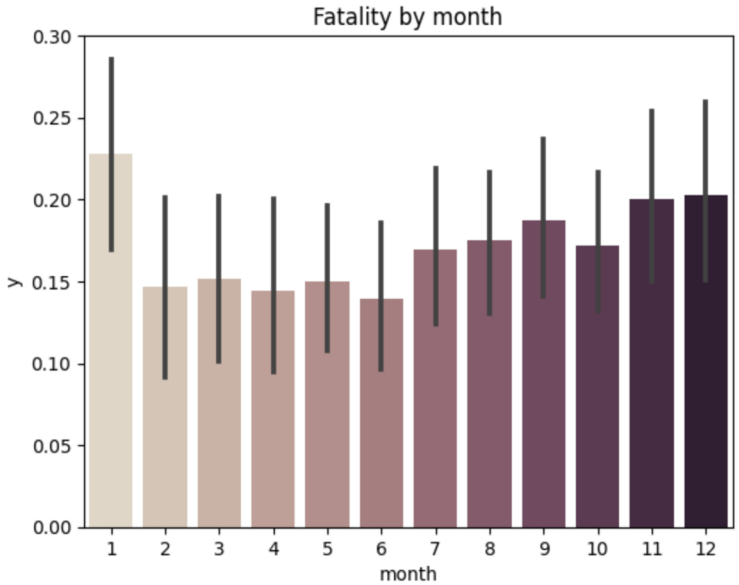




Exploratory analysis

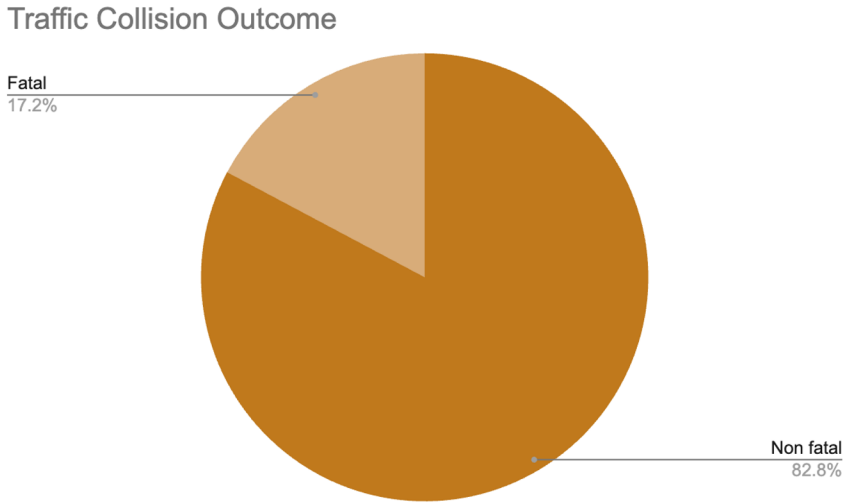
This was also the case for other variables which we anticipated to be highly correlated with fatalities





Machine Learning results

- We filtered the dataset to only include collisions after 2014
  - There were large differences in reporting before 2014
- 2,791 instances
  - 481 fatalities
  - Imbalanced dataset
- Train size = 1,953
- Test size = 838



Hyperparameters

After using grid search to tune hyperparameters, the following are the parameters chosen for each of the corresponding algorithms

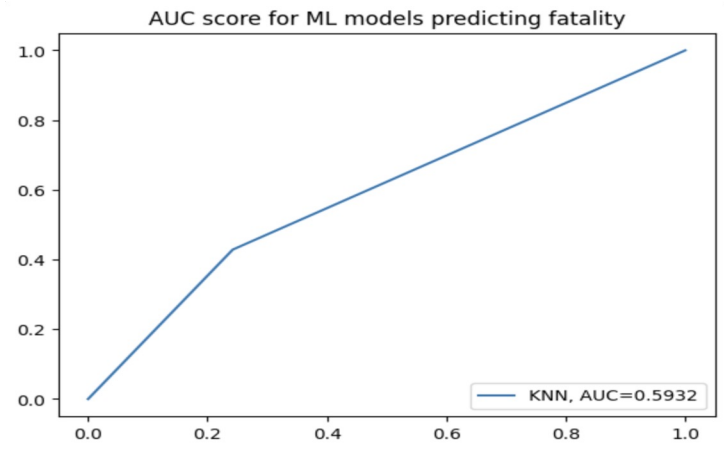
Hyperparameter	Values
Logistics regression	
Penalty	L2
C	0.1
KNN	
n neighbours	15
Weights	Uniform

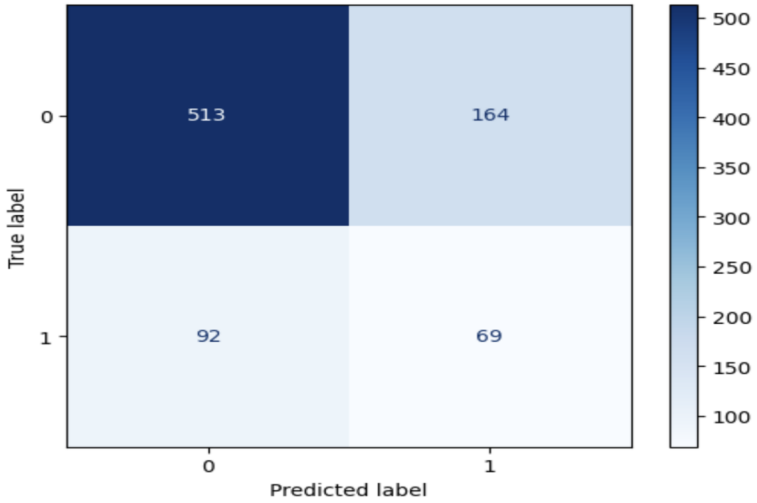
Hyperparameter	Values
Decision Tree	
Criterion	Gini
Max depth	2
Random Forest	
N estimators	115
Criterion	Entropy
Max depth	8
Min samples leaf	1
Min samples split	7
Max features	Auto

Scoring metrics

- We used a probability threshold cut-off of 0.25 to try to improve out precision and recall. We also felt that it was important to capture more potential fatalities

- Our best scoring model was the KNN model
  - AUC = 0.59
  - This was compared to the train AUC of 0.66, suggesting our model is slightly overfit





Precision, recall, F1

- Our model had difficulty in being able to predict positive cases (fatalities)
- Overall metrics:
  - Precision = 0.296
  - Recall = 0.429
  - F1 = 0.35
- More detailed precision-recall scores can be seen to the right by outcome class

	Precision	Recall	F1-score
Non-fatal	0.85	0.76	0.80
Fatal	0.30	0.43	0.35

Recall:

The number of fatality predictions made out of all fatalities

Precision:

The number of fatality predictions that actually belong to the fatality class

Discussion

Interpretation

Unfortunately we were not able to achieve relatively high precision and recall

This may be due to a number of reason:

- This dataset only included collisions involving death or serious injuries
  - The difference between the two of these may be marginal in many cases and it may be difficult to predict this
  - For instance, what may have resulted in fatality for one individual may not for another
- Since we grouped by car incident, we lose person-level data
  - This means we lose variables like age, which may be a very important variable in terms of predicting fatality, as younger children or older adults may have higher probability of fatality

Interpretation

Unfortunately we were unable to achieve relatively high precision and recall

This may be due to a number of reason:

- We also noticed in our exploratory analysis that there were not any obvious trends in variables like month, day, alcohol etc.
  - We had hypothesised that these would be large factors (i.e. weekends, snowy months, etc.)
- Ultimately, the smallest of things could affect accident severity, which the dataset was not able to pick up
  - Seatbelt use, angle of collision, age of car, age of airbags, whether breaks were applied before collision, underlying injuries to people involved
- The time it takes for first responders to arrive on seen also would have a large impact
  - Since these are all serious accidents, the amount of time it takes for paramedics to arrive on seen could be the difference between life or death

Lessons learned

What we learned

- Car incidents are by nature often random events and although severity of injuries can be related to certain factors, the severity of the incidents is hard to predict
- Real world data is much harder to find patterns in than the toy datasets often practice with
- Ultimately, being able to find patterns in data is linked directly to the availability of data recorded.
    - In the case of our project, there were not enough features that were tightly linked to severity as anticipated
    - The features in this dataset may have likely been better suited to predicting the number of incidents (serious or not) as opposed to incident severity.

Thank you!