TRAFFIC COLLISION INJURY SEVERITY IDENTIFICATION

Weixiao Huang 40364300

BACKGROUND INFORMATION

- Approximately **1.3 million** people are killed because of traffic accidents around the world every year, and **50 million** people are injured, according to World Health Organization

- Only in Northern Ireland, **52 fatal collisions** and **5,116 injury collisions** caused **7,901 casualties** altogether in 2022, according to the investigation of NI Road Safety Partnership

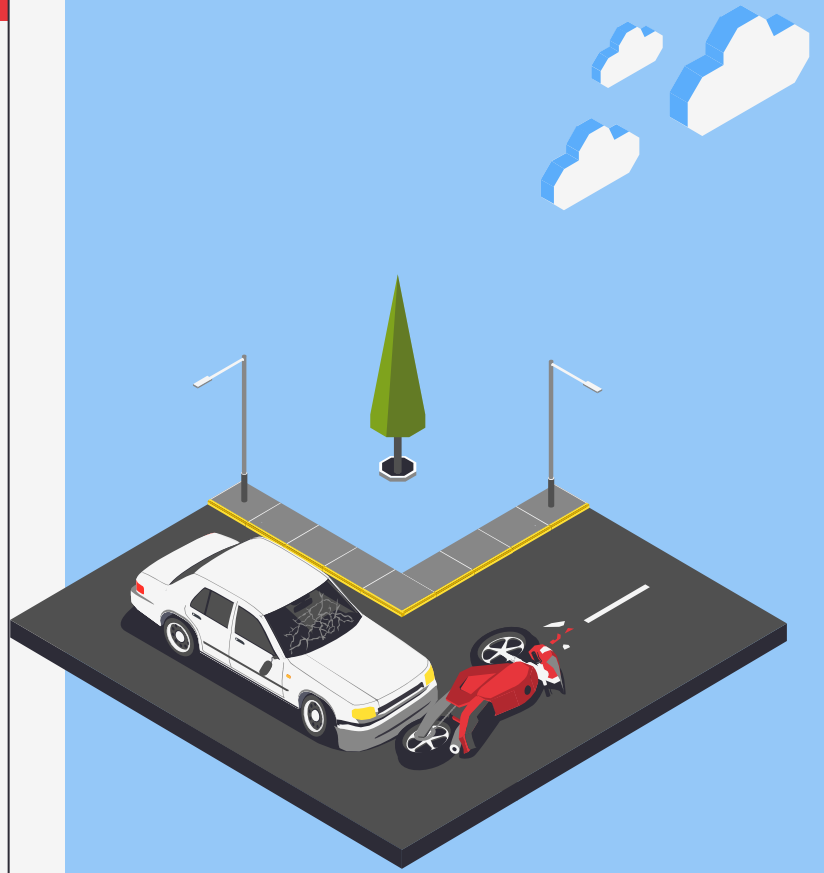
PROJECT OBJECTIVE

Aims:

- Reduce the number of traffic road collisions
- Deeply investigate accidents injuries
- Predict when and where and why they happen
- Give recommendations on how to prevent them.

What we will do:

- Explore the distribution and correlation of variables
- Build a machine learning classification model to predict how serious a person's injuries is in an accident



Data

Source Data

Police Service of Northern Ireland (PSNI)
Northern Ireland Police Recorded Injury Road Traffic Collision Data

External Data

Google Map API Data
Further information on location of the collision

Collision

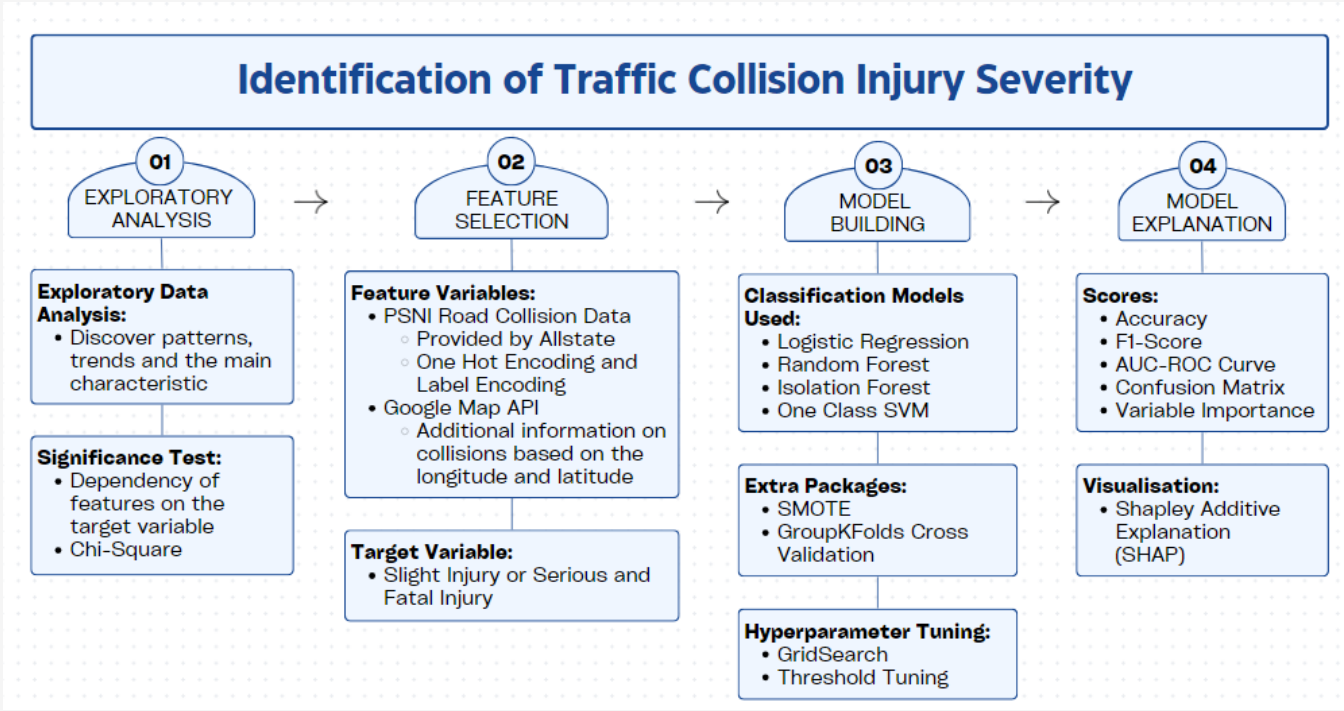
Vehicle

Casualty

Address

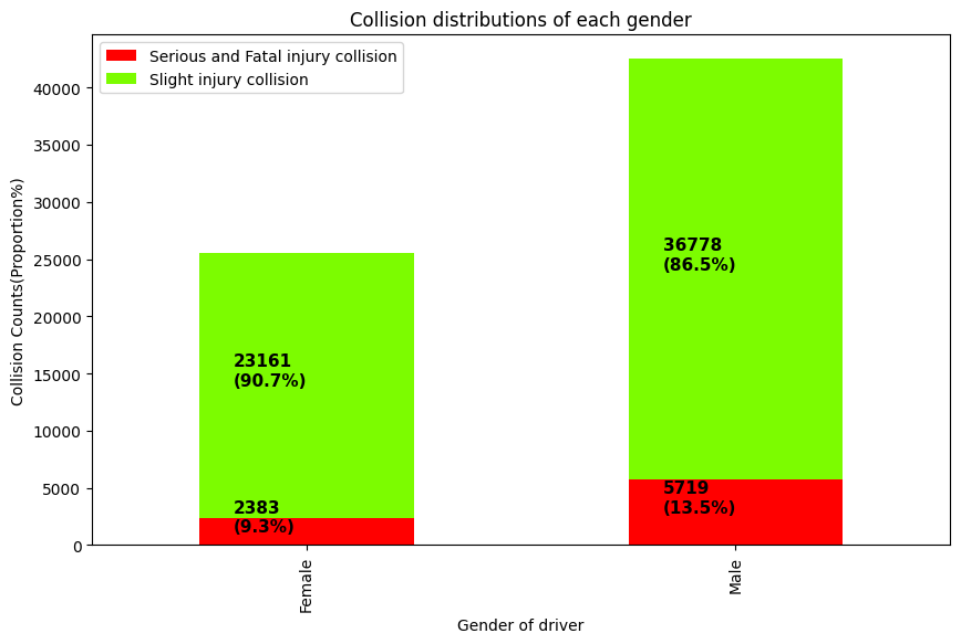
Location Info

PROGRESS


Identification of Traffic Collision Injury Severity
01
EXPLORATORY ANALYSIS
02
FEATURE SELECTION
03
MODEL BUILDING
04
MODEL EXPLANATION
Exploratory Data Analysis:
• Discover patterns, trends and the main characteristic
Significance Test:
• Dependency of features on the target variable
• Chi-Square
Feature Variables:
• PSNI Road Collision Data
◦ Provided by Allstate
◦ One Hot Encoding and Label Encoding
• Google Map API
◦ Additional information on collisions based on the longitude and latitude
Target Variable:
• Slight Injury or Serious and Fatal Injury
Classification Models Used:
• Logistic Regression
• Random Forest
• Isolation Forest
• One Class SVM
Extra Packages:
• SMOTE
• GroupKFolds Cross Validation
Hyperparameter Tuning:
• GridSearch
• Threshold Tuning
Scores:
• Accuracy
• F1-Score
• AUC-ROC Curve
• Confusion Matrix
• Variable Importance
Visualisation:
• Shapley Additive Explanation (SHAP)

SOME INTERESTING FINDINGS

Gender

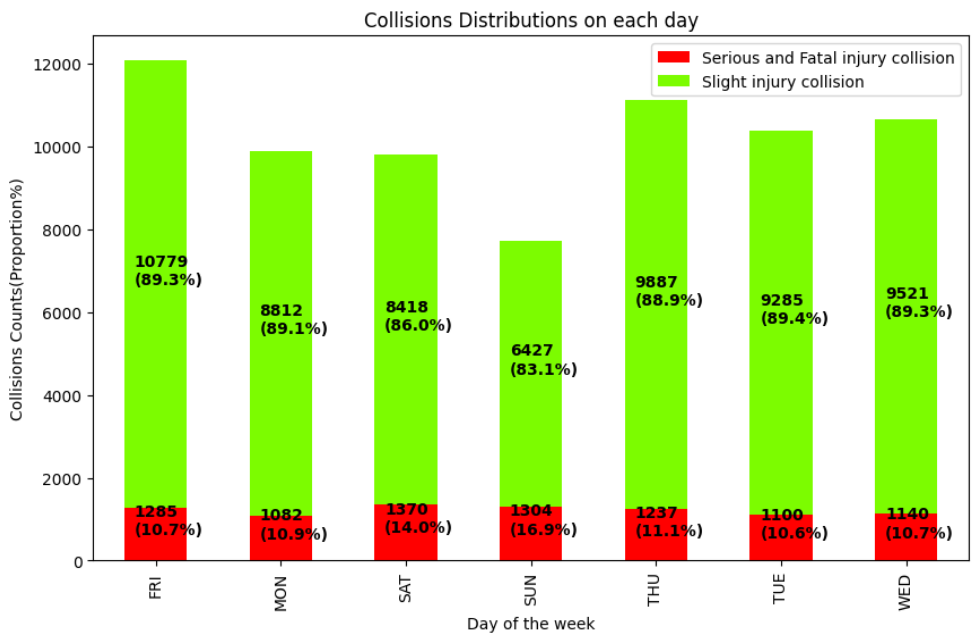


Not only male are more likely to cause a collision than female, they are also more likely to cause life-threatening crashes

The significant test has P-Value (8.5881e-59) ≤ 0.05, which means there is a significance difference between the number of collisions caused by male drivers and female drivers

SOME INTERESTING FINDINGS

Week day



Friday has the most collisions, and Sunday has the least, however Sunday has the highest probability of having serious injuries.

The significant test has P-Value (1.1897e-38) ≤ 0.05, which means there is a significance difference between the number of collisions of each week day

SOME INTERESTING FINDINGS

Proportion of each injured age group depending on class

c_class v_agegroup	Driver	Motorcyclist	Pedal cyclist	Pedestrian	Pillion passenger	Vehicle passenger - front	Vehicle passenger - rear
10-16	0.130435	0.072464	0.675362	0.000000	0.017391	0.060870	0.043478
17-24	0.627712	0.036570	0.020897	0.000000	0.001249	0.206133	0.107439
25-34	0.628108	0.033208	0.026603	0.000092	0.001101	0.177690	0.133199
35-44	0.588826	0.027281	0.033719	0.000000	0.001309	0.179289	0.169577
45-54	0.597625	0.044000	0.042375	0.000000	0.002125	0.177125	0.136750
55-64	0.619876	0.043478	0.042443	0.000207	0.002070	0.178054	0.113872
65+	0.689904	0.019623	0.022721	0.000000	0.000516	0.217919	0.049316
Under 10	0.046296	0.009259	0.935185	0.000000	0.000000	0.000000	0.009259
Unknown	0.072727	0.018182	0.068182	0.004545	0.013636	0.500000	0.322727

Over 90% of under 10 years old and around 70% of 10-16 years old teenagers are injured while they are riding bicycles, while for the other age group, injuries are basically inside the car, as a driver or a passenger, and front row passenger have relatively higher probabilities to be injured.

Models

Basic Models:
- Logistic Regression
- Random Forest
- One Class SVM
- Isolation Forest

Extra Packages:
-SMOTE
- Group K-Folds Cross Validation

Tuning:
- Hyperparameter Tuning
- Threshold Tuning

Model Metrics

Model Name	Best AUC Score
Logistic Regression	**0.7154**
Random Forest	0.5837
One-Class SVM	0.5074
Isolation Forest	0.5479

Optimal Model

Score Metrics	
Accuracy	0.6897
F1 – score	0.8020
AUC	0.7154
Average Precession	0.8884

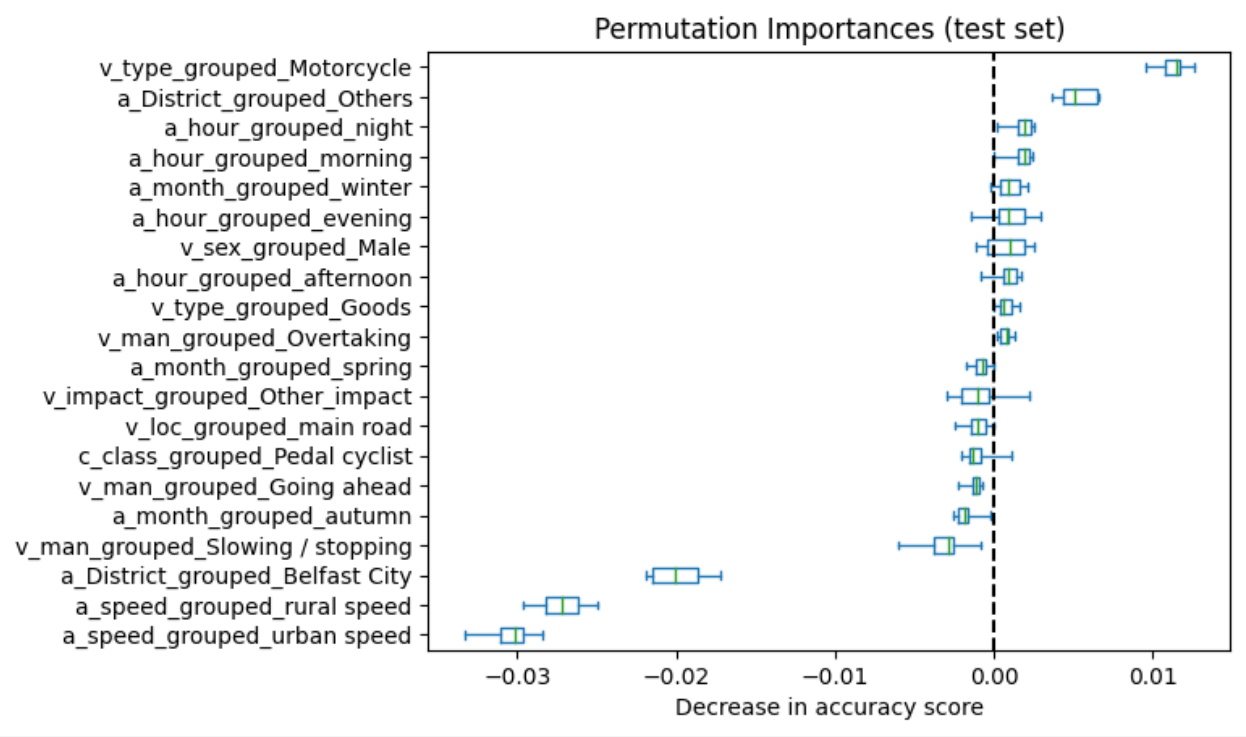
Confusion Matrix

Actual \ Predict	Slight	Serious & Fatal
Slight	5889	2712
Serious & Fatal	196	576

Features:

- Hour of Collision
- Month of Collision
- Speed Limit of the road
- Weekday of Collision
- Policing Area - at council level
- Vehicle Maneuver
- Vehicle Location at Time of Impact
- First Point of Impact
- Age of Driver
- Sex of Driver
- Vehicle Type
- Sex of Casualty
- Casualty Class
- Casualty Vehicle Type
- location info: see if the following locations exist
  - car repair, lodging, health, park, university

Optimal Model



Calculates the impact of prediction accuracy when we randomly re-order a single column

Stronger correlation
=> decrease in accuracy

Absolute value
=> extent of the impact

The most important feature is choosing a motorbike as a way of transportation, reducing the accuracy by over 0.01

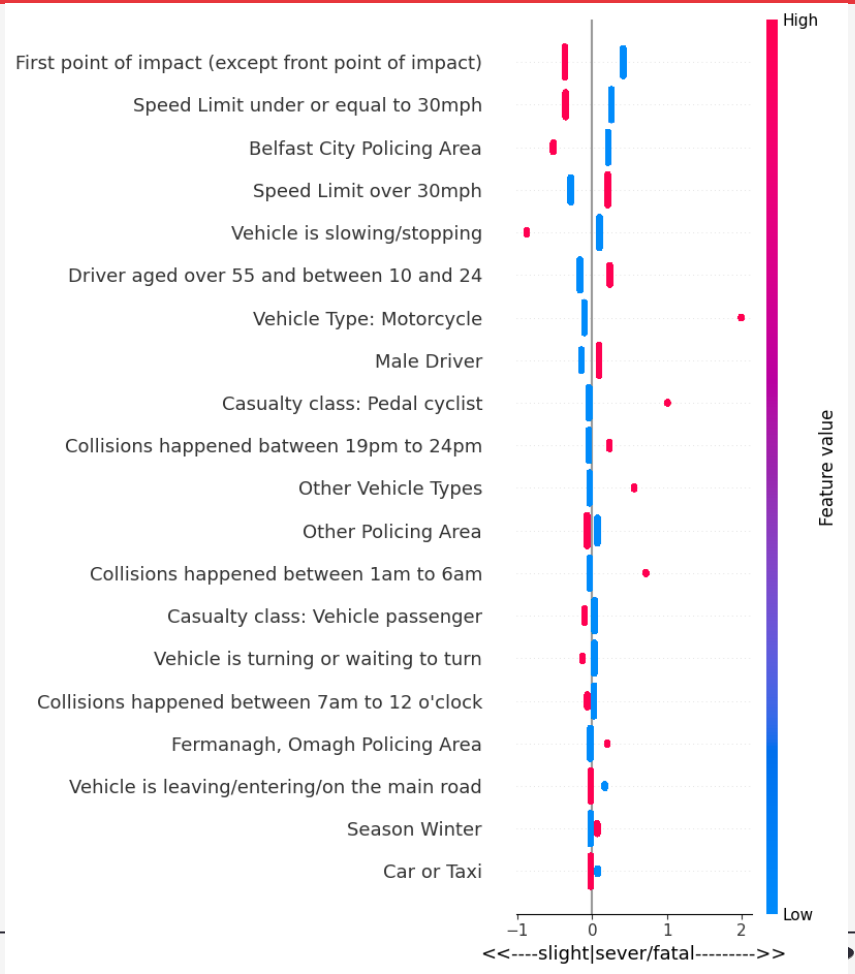
worth noting that accuracy has significantly increased as a result of shifting the speed and Belfast City columns

Model Explanations

Shapley Additive Explainations (SHAP)

The closer the color is to red, the larger the feature value is, and conversely, the closer the color is to blue, the smaller the feature value is

Similar as variable importance, the negativity show the correlation with injury severity



Conclusions

- Ride motorcycles or bicycles as a way of transportation
- Travel at night
- Drive on a road whose speed limit is over 30 mph

Serious and Fatal Injury

- The vehicle is slowing or stopping
- The location of crash is not at the front
- In Belfast city
- Drive on a road whose speed limit is under or equal to 30 mph

Slight Injury

Recommendations

Policy Makers	• Improve road infrastructure • Hold awareness campaigns
Insurance Company & Car Rental Company	• Based on the driver's driving patterns and records, give guidance while charging customers
Hospital	• Estimate the extent of injuries • Have the preparation accordingly

Future Work

Collect more
post pandemic
collision data

Extra External data
on weather and
public transportation

Apply Principle Component Analysis
(PCA) to reduce dimensionality

THANKS!

ANY QUESTIONS?