Identification of Traffic Collisions Injury Severity in Northern Ireland

Weixiao Huang

MSc Data Analytics

School of Mathematics and Physics
Queen's University Belfast

September 2023

Abstract

This report forecasts a potentially life-threatening collision happening in Northern Ireland using collision data from the Police Service of Northern Ireland (PSNI). We will identify the injury severity in road traffic accidents by learning from past collision injury records and trends. Using the PSNI historic collision data and additional external data sources, we will determine what features of a collision can contribute the most to someone in an accident being fatally injured. Lastly, we provide suggestions on how to help reduce serious and fatal injuries in Northern Ireland (NI).

Acknowledgements

I would like to give my warmest appreciation to my supervisor Lucy Doyle and Dr. Aleksandar Novakovic for their support, guidance and mentorship during the whole internship. I would also like to thank Allstate NI for providing me with this valuable internship, as well as Darren Cheung, Rezwin Rafeek and others for their encouragement and support.

I would also like to give special thanks to my parents, for their continuous emotional support when undertaking my research and writing my project.

Contents

Chapter 1

Introduction

According to road traffic injuries statistics published by the World Health Organization, approximately 1.3 million people are killed because of traffic accidents around the world every year, and 50 million people are injured[1]. In the United Kingdom, 1,695 people lost their lives and 136,002 got injured in car collisions in 2022, as reported by the Department of Transportation[2]. And 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[3]. BBC reported a fatal six-vehicle crash happened on M1 in 2016, which caused one death, a number of minor injuries and several hours of motorway closure[4]. In early August of 2023, an eight-year-old girl died and a boy was taken to hospital by ambulance because of a serious single-vehicle crash that happened in County Antrim[5].

To address the outlined key road safety challenges from 2010 to 2020, the government published the Northern Ireland Road Safety Strategy(NIRSS) to 2020 in March 2011[6]. In recent years, although the number of road traffic collisions has decreased gradually potentially because of the implementation of the NIRSS to 2020, people still have concerns about the safety and well-being of individuals on the road. The severity of injuries resulting from these collisions can vary widely, ranging from minor bruises to life-threatening conditions.

Traffic accidents can be caused by various risk factors. The factors that impact injury severity can be broadly categorised into two groups, people involved and their personal characteristics, and the external circumstances. The people involved and their characteristics include information of both the driver and casualty, such as their age and gender. Followed by the external circumstances of a collision, which show general information such as the weather, time and location.

In this project, we use road traffic collision data collected by PSNI from 2016 to 2021, noting that the dataset does not include all the traffic collisions in Northern Ireland, only those where police officers are present at the scene. We will discuss the correlation between the two factors mentioned above and the injury severity in NI, and predict the likelihood of serious casualties caused by traffic accidents in the future based on the dataset.

We conduct an exploratory analysis of traffic collision data in NI to discover the distribution as well as the correlation between each feature in our data and the severity of injuries sustained by casualties. Significance tests are applied as well to find out if our features are dependent on the injury severity and level of dependence.

We also use Google API data to have access to further information. Based on the longitude and latitude of the collision from the PSNI dataset, we use Geocoded to help to get the postcode, address and location information from Google Map API, in order to allow us to make better classifications.

Finally, machine learning classification models will be performed on current traffic collision data to identify correlations between various features when people are injured in road collisions. This will help us to categorise whether an injured person in a traffic collision is slightly injured or seriously and fatally injured.

The next chapter of the report provides an explanation of the methodology used to complete the exploratory analysis of the traffic collision data, external Google API data, how significance tests were performed, and how classification models were built, then we present the results in a subsequent chapter. In the discussion chapter, we will discuss the relationship between features and injury severity and how we can forecast future traffic accident severity based on the current data from the results of the last chapter. In conclusion, this report will give a detailed description of what kind of person has more likelihood of being seriously or fatally injured in a traffic collision in Northern Ireland and try to give some suggestions on reducing the number of serious and fatal injuries.

Chapter 2

Methodology

In this chapter, we will explain how we divide this project into a few phases and solve them step by step. Figure 2.1 shows the workflow to build the binary classification model to identify injury severities: slight injury, serious and fatal injuries. In section 2.1, we discuss the patterns and main characteristics we discovered in the initial investigations and also employ data visualization methods on them. Section 2.2 introduces the external Google API location data we use to help.[7] In addition, significance tests are introduced to test the dependency of injury severity in section 2.3. The final section describes how all the features were combined with an external Google API and then placed as features into a classification model to predict injury severity for future road accidents in Northern Ireland.

2.1 Exploratory Analysis

The data collected by PSNI includes 71,964 collision records from January 1st 2016 to October 18th 2021, has 56 columns documenting features of each collision. The dataset includes information from accident reports, vehicle characteristics, road conditions, weather conditions, and injury outcomes.

During the preprocessing stage, the collected data goes through cleaning, transformation, and integration. In our data, some of the columns have more than 50% missing values, and some of the columns are required to be filled for serious and fatal collisions but not essential for slight injury, which means there is a high probability that this row is a severe casualty when the content of these columns is not empty, we consider this as target leakage, so these columns are removed since they are not appropriate to use to build the model. This step ensured that the data is consistent, accurate, and ready for analysis.

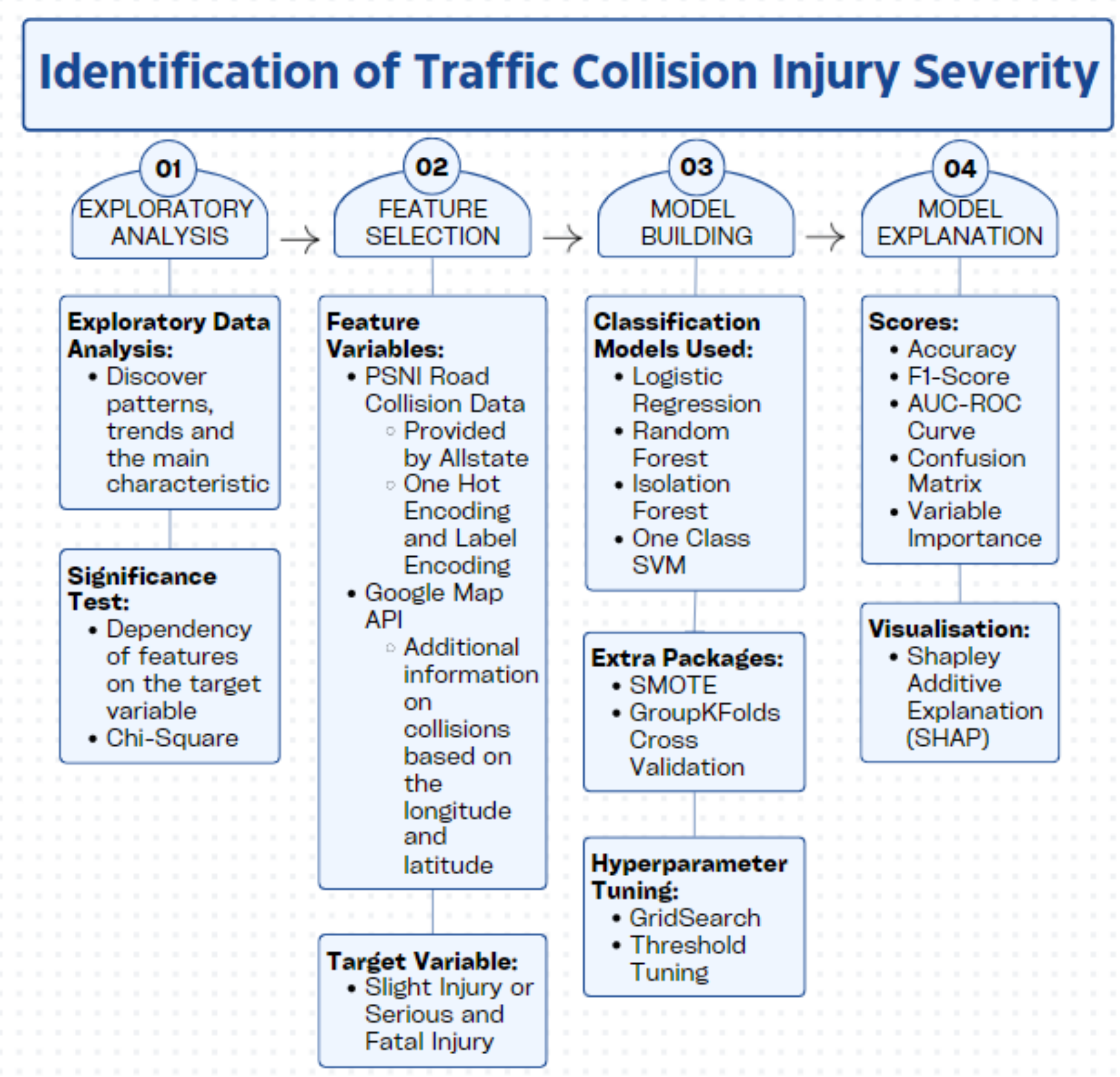
Identification of Traffic Collision Injury Severity
01
EXPLORATORY ANALYSIS
02
FEATURE SELECTION
03
MODEL BUILDING
04
MODEL EXPLANATION
Exploratory Data Analysis:
• Discover patterns, trends and the main characteristic
Significance Test:
• Dependency of features on the target variable
• Chi-Square
Feature 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)

Figure 2.1: Workflow chart for binary classification model building

In our initial investigation, we explore road accidents of each feature with graphical packages named "matplotlib" and plot them into histograms and lines. We also add the count and proportion to the stacked histograms to give more information if needed.

2.2 External Data

We use external Google Map API data to aid further exploration of collision locations. Google Maps Platform offers access to data about maps, routes and locations.[7] Google Maps Geocoding API converts latitude/longitude coordinates to addresses or Place IDs and vice-versa.[8] We can use the address attained by this to explore which area has a higher probability of having life-threatening accidents. Also, Google Place guides us to explore the location information of the collision. We mainly check if the following location type is in the information:

- car_repair (indicates a shop/service that primarily repairs cars/automobiles)
- lodging (indicates a temporary accommodation)
- health (indicates a place related to health)
- park (indicates a named park)
- university (indicates a university)

2.3 Significance Test

We use significance tests to examine the correlation between features and our target variable [9]. The reason to have this test is to support the exploratory analysis findings. There are multiple types of significance tests, t-test, z-test, and chi-square test. Here we choose the Chi-square test in this project because our response variable is categorical.

We first state the hypothesis to perform a statistical significance test. The first hypothesis we made is null hypothesis H0, which assumes that there is no difference between the two means or that the recorded difference is not significant, in this case, the null hypothesis is casualty injury severity is independent of the feature. The other hypothesis is the alternative hypothesis, written as HA, which is opposite to the null hypothesis, and holds casualty injury severity is dependent on the feature.

Then we calculate the p-value to see if it falls into the Confidence Interval by comparing it with the alpha. The p-value is the probability when the null hypothesis is true. A confidence interval is a range of values within which the true value has a certain probability. The alpha is the significance level, which gives the probability of rejecting H0 in favour of HA. The smaller the p-value, the stronger the evidence against H0 provided by the data. The confidence interval for the test is 95%, which means if our p-value is greater than 0.05, we can accept the null hypothesis H0 and conclude there is no correlation.

2.4 Modelling

2.4.1 Feature Engineering

Feature engineering plays a critical role in enhancing the predictive capabilities of the model. Relevant features are selected based on their potential influence on injury severity. Additionally, new features are created through combinations, transformations, and aggregations of existing variables. This process aims to capture hidden relationships within the data that could improve the model's accuracy.

In this project, to improve the model performance, we mainly use the following methods:

- chooses the most prominent characteristic, and re-categorises the variable as having this characteristic and those that do not. For example, we know that Sunday has the highest probability of being seriously injured from previous exploratory analysis, we can group this feature as is_Sunday and not_Sunday.
- averagely divides features into several parts. We split months of the year into 4 parts and categorise them as seasons. Similarly, we split the hours of a day into four 6-hour periods and named them morning, afternoon, evening and night.
- combines those features that provide similar information, such as combining all goods vehicles with different weights as one type - Goods.
- checks if the information we are looking for exists. This only applies to location information and checking the existence of specific locations.

After careful selection and different combinations, we finally decide to include the following features:

- a_hour (Hour of Collision)
- a_month (Month of Collision)
- a_speed (Speed Limit of the road)
- a_wkday (Weekday of Collision)
- a_District (Policing Area - at council level)
- v_man (Vehicle Manoeuvre)
- v_loc (Vehicle Location at Time of Impact)
- v_impact (First Point of Impact)
- v_agegroup (Age of Driver)
- v_sex (Sex of Driver)
- v_type (Vehicle Type)
- c_sex (Sex of Casualty)

- c_class (Casualty Class)
- c_vtype (Casualty Vehicle Type)
- location_infos: see if the following locations exist
  - car_repair
  - lodging
  - health
  - park
  - university

Before we introduce these variables into the classification model we need to convert categorical variables into numerical variables with the help of One-Hot Encoder[10] and Label Encoder[11]. In One-Hot Encoding, we convert data into multi-dimensional binary vectors. The number of dimensions is decided by the number of categories, and for each category, a column filled with 0 or 1 is generated based on whether the item is the same as the category. In Label Encoding, each category is assigned a number starting from 0 and using the number to replace the category. This will not generate multiple columns and is usually used on the target variable.

2.4.2 Model Building

We build a model to classify injury severity with current collision records to predict how seriously people will injured in future road accidents. Since injuries that are fatal only take a very small percentage, we then combine them with serious injuries as one category.

Several classification algorithms are evaluated to determine the most suitable model for the task. Algorithms such as Logistic Regression[12], Random Forests[13], One-Class Support Vector Machines[14], and Isolation Forest[15] are considered. The dataset is split into training and testing subsets with a split proportion of 80/20 to train the selected model. Model performance is assessed using techniques like cross-validation and metrics such as accuracy, precision, recall, AUC-ROC score and F1-score.

Within our dataset, after removing all the rows that do not have casualty injury severity, we have 46246 rows remaining, 42359 of which belong to slight injury, with the rest 3887 rows being life-threatening injuries. These two types of injuries account for about 91.6% and 8.4% of the total data respectively. The size of the majority class is more than 10 times of the minor class. This indicates that the dataset is not balanced, with a ratio of 10:1. When we make predictions on imbalanced data [16], the model will be better at predicting the majority class since there is more data to learn from, and usually, the accuracy of the overall model is very high due to the high prediction accuracy of the majority class. We need to apply some techniques to improve the performance of this overall dataset, such as hyper-parameter tuning [17], threshold tuning [18], Group K Folds Cross-Validation[19] and Synthetic Minority Oversampling Technique (SMOTE)[20].

Hyperparameters are parameters that could define the architecture of the model we built. To find parameters that exactly fit this project, we apply Grid-Search CV hyperparameter tuning. We put all the possible values of each parameter into the model iteration. The model runs through all possible combinations of parameters, just like each grid cross of the whole net. Each model has different parameters and the following are the parameters we used for tuning:

Logistic Regression:

- solver: the algorithm used in the optimization process, choose from 'newton-cg', 'lbfgs' and 'liblinear',
- penalty: the norm of penalty, choose from 'l1' and 'l2',
- C: regularization strength, choose from 100, 10, 1.0, 0.1, 0.01;[21]

Random Forest:

- n_estimators: the number of trees in the forest, choose from 10, 100 and 1000,
- max_features: number of the features of the best split, choose from 'sqrt' and 'log2';[22]

One-Class SVM:

- kernel: kernel type to be used in the algorithm, choose from 'poly', 'rbf' and 'sigmoid'. [23]

For all models, to give more concentration on the minor class, we also add class weight to all of them. The weight of each class is defined as the following function :

$$w_j = n\_samples/(n\_class * n\_samples_j)$$

Followed by the parameter of Grid Search, another parameter of hyperparameter tuning 'scoring' represents how we access the performance of the model, we focus on the AUC score here, so 'scoring' is set to 'roc_auc'.[24]

Each algorithm calculates the probability for positive classes which lies between 0 and 1. And we use the threshold to decide which class each input belongs to. A value above the threshold indicates a positive class and a value below indicates a negative class. The default value of the classification threshold is 0.5, but thresholds are problem-dependent. Therefore we apply the threshold tuning to find the threshold value for this problem and try to have the model with the best AUC score.

Group K-Folds Cross-Validation is another technique we use in this project and it is an enhancement of K-Folds Cross-Validation to help consider the groups in the data. K-Folds Cross-Validation is a resampling method that could help evaluate machine learning models on limited data. The single parameter k of this method refers to the folds that the data is divided into. Each fold is used for testing while the remaining folds are used for training. After all folds have been gone through, we take the average performance. Since for most of the collisions, there are multiple vehicles and casualties, we do not want the records in one collision to be split and appear in both test and train data, hence we use group K-Folds Cross-Validation. Groups are the unique collision ID to split on, and we want to achieve that the model is new to the test data for each iteration of folds.

The last technique we use is SMOTE. We use SMOTE to oversample the imbalanced classification dataset and make the amount of the minority class data equal to the data of the majority class. In this way, the dataset is balanced. The new data generated is based on the rows we got for each class.

Features that correlate the most to injury severity are found using permutation importance. We also use Shapley Additive Explanations (SHAP) [25] to visualise variable importance.

Chapter 3

Results

3.1 Exploratory Analysis

Figure 3.1 shows the number of collisions gradually decreased from the year 2016 to 2019, this may be the possible result of the implementation of the Northern Ireland Road Safety Strategy(NIRSS)to 2020. Due to COVID-19, the government announced the first national lockdown on 23rd March 2020 to encourage residents in Northern Ireland to stay at home rather than go outside except for limited purposes[26], people have fewer chances to go outside and the number of collisions falls sharply in 2020. Things started to go back to normal in 2021 and more collisions happened compared with 2020.

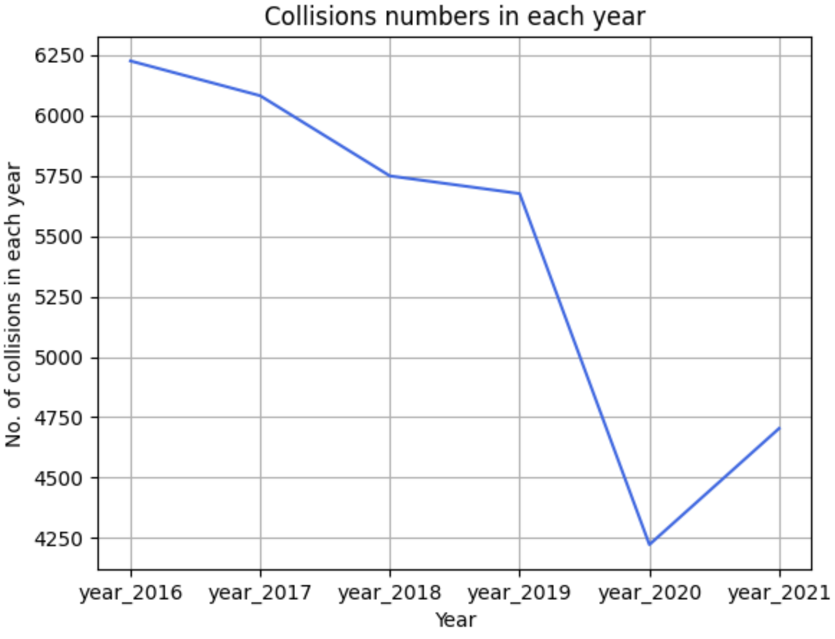

Figure 3.1: Line plot of collision number in each year from 2016 to 2021

Figure 3.2 shows the distribution of hours in which a car accident occurs. From the plot, the number of collisions has a huge increment from 7 am and roughly keeps increasing until 5 pm, because people leave their houses in the day, and after reaching the peak, the number starts decreasing and meets the bottom at 4 am, this is the time most of the people are sleeping. At 8 am and 5 pm, they are local maximum because usually around these two hours, it is morning and evening rush hour.

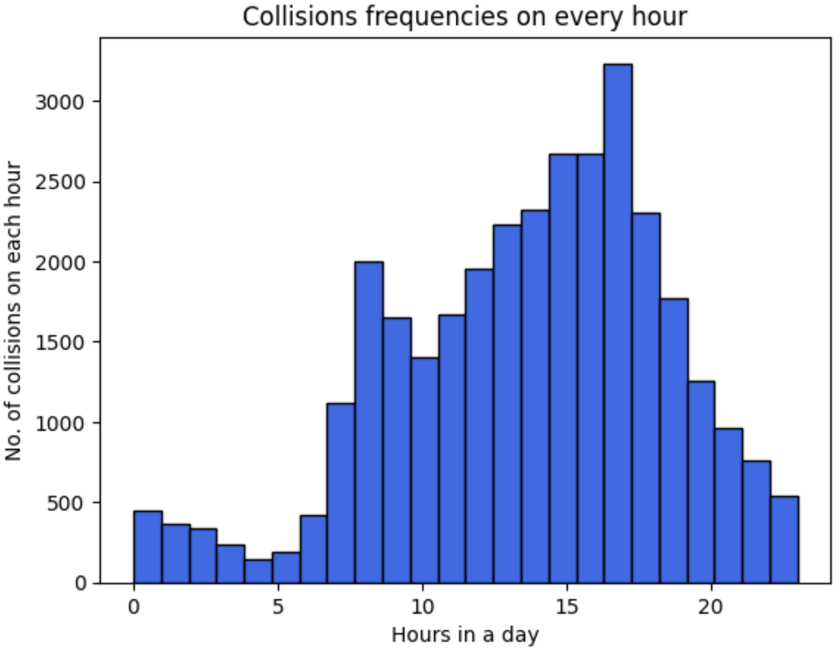

Figure 3.2: Histogram of collision distributions of every hour

Then we start to investigate the distribution of accidents in each month. From Figure 3.3, the number of collisions exceeds 2,500 in every month except April, with the last four months having almost 3,000 vehicle collisions. April has the least collisions,but still over 2,000, which means there are at least 67 collisions every day on average.

As shown in 3.4, Friday has the most collisions and Sunday has the least, the remaining days have about the same number.

So we have a further investigation, we plot the stacked histogram and colour the bar by the injury severity, plotted in Figure 3.5. To help better reading, the red colour represents serious and fatal injuries and the green colour represents slight injury. We also add counts and their proportions each weekday as subtitles. The number of car accidents that result in serious injuries is similar every day, however, because Sunday has the minimum accidents, Sunday has the highest proportion of severe injuries, which is 16.9%, represents people are more likely to get serious and fatal injuries on Sunday. And comparatively, although Friday has the most collisions, most of them are slight injuries.

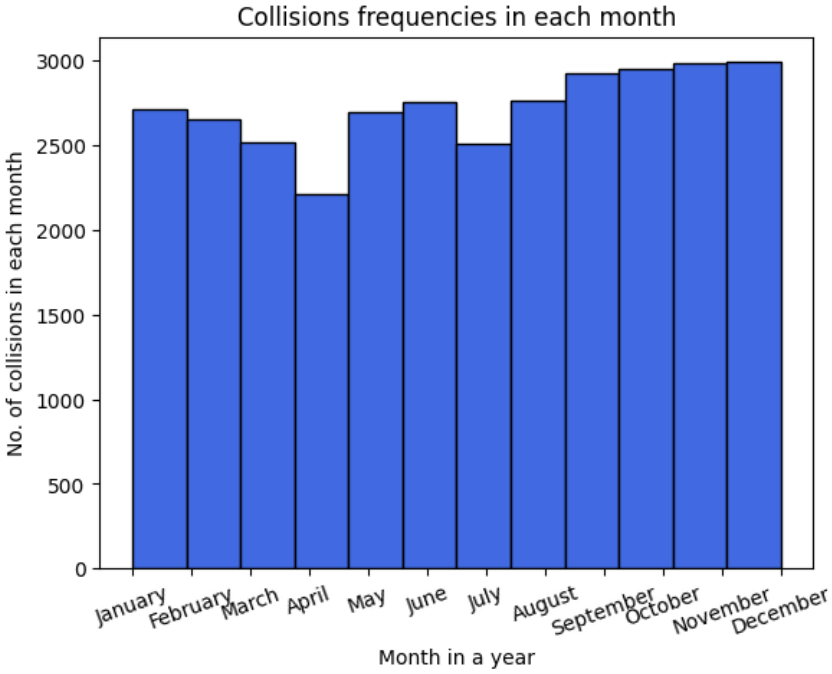

Figure 3.3: Histogram of collision distributions of every month

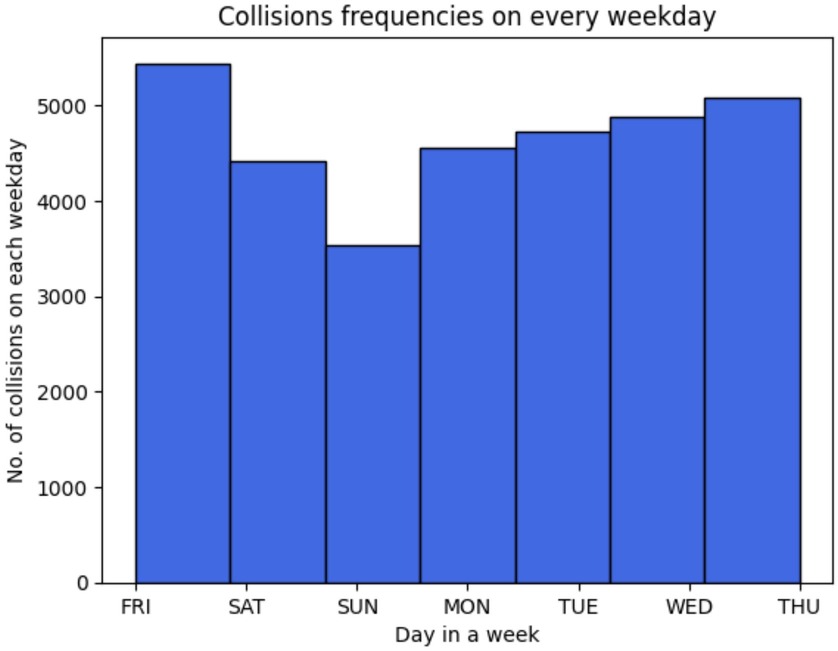

Figure 3.4: Histogram of collision distributions of every day of the week

The age of the driver is also a factor that will affect the injuries. As shown in Figure 3.6, young people aged from 25 to 34 cause the most crashes and then followed by young people aged 35 to 44 and adults aged 17 to 24. Traffic accidents do not necessarily include only those with a driver's licence. Teenagers under the age of 16 are the drivers of non-motorised vehicles such as scooters, roller skates and skateboards. They also cause a number of crashes each year, even if they don't have a driver's licence.

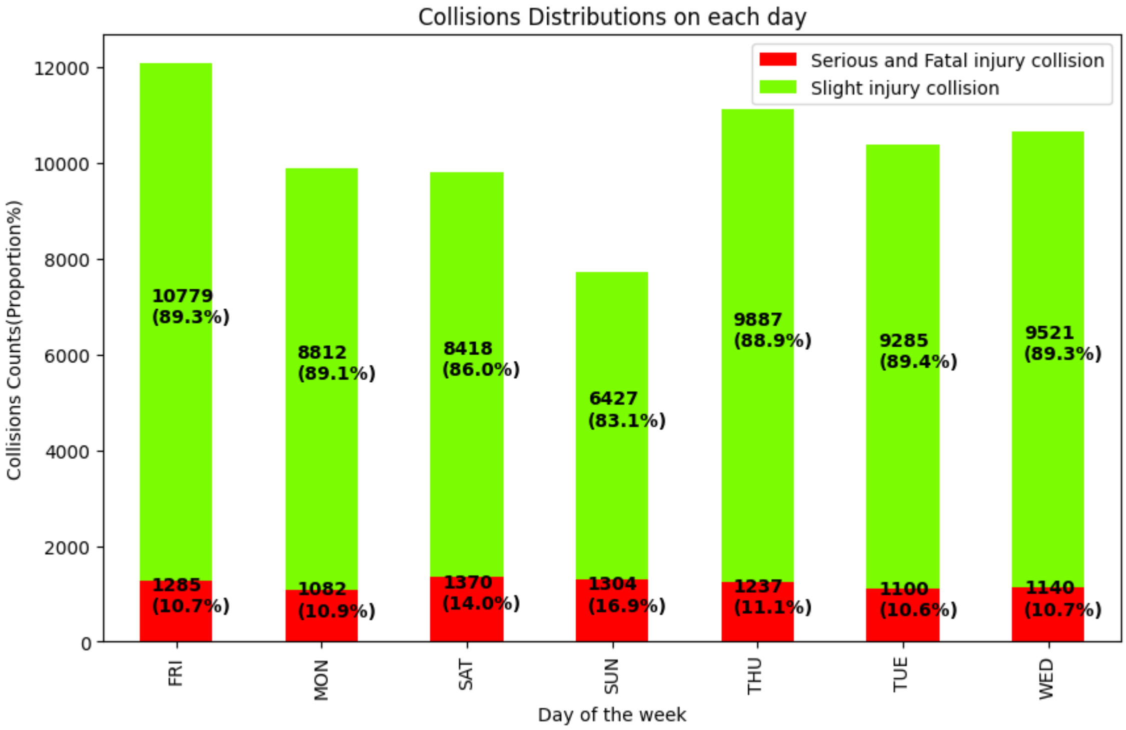

FIGURE 3.5: Stacked histogram of collision distributions of every day of the week coloured by injury severity

Variable	Results	P-value
Every hour	8.4347e-96	$P \leq 0.05$
Each month	0.0013	$P \leq 0.05$
Every weekday	1.1897e-38	$P \leq 0.05$
Agegroup of drivers	3.6703e-48	$P \leq 0.05$
Gender of driver	8.5881e-59	$P \leq 0.05$
Gender of casualty	0.000	$P \leq 0.05$

TABLE 3.1: Results of significance tests

By removing all the empty values and the "Unknown", we obtain the distribution of the driver's gender shown in Figure 3.7. Males cause more traffic accidents than females, with 1.7 times as many accidents caused by males as by females. Furthermore, male drivers result in 4.2% more life-threatening accidents than females.

The gender of casualty also shows a similar trend in Figure 3.8. More males are injured in car crashes than females.

3.1.1 Significance Test

After calculation, the results of all the variables we mentioned above are shown in Table 3.1, all rounded up to four decimal places.

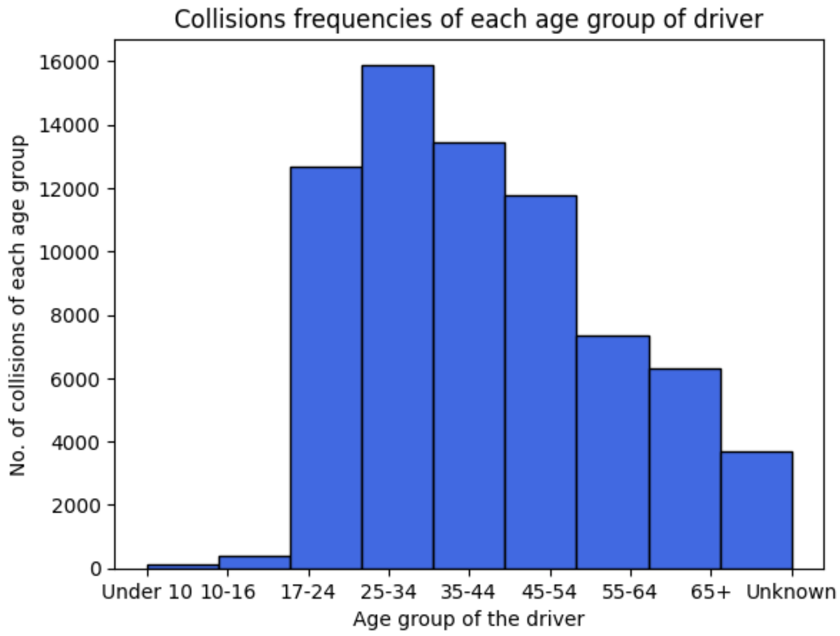

FIGURE 3.6: Histogram of collision distributions of every age group of the driver

3.2 Modelling

3.2.1 Model Building

All the model iterations described in section 2.4.2 with various models and techniques are shown in table 3.2 and 3.3. We choose the best model based on the AUC-ROC score and consider it as the optimal model. The closer the AUC-ROC score is to 1, the better the model is.

3.2.2 Optimized Model

From table 3.2, the Logistic Regression model after hyperparameter tuning that has the best AUC score is selected as our final model. We calculate the score metrics (Table 3.4) and confusion matrix(Table 3.5) to gain insights into the model's ability to classify injury severity levels correctly.

Figure 3.9 shows the permutation importance applied to the test set of this model. Because of the large number of variables, we have intercepted the first ten and the last ten important variables in order to better characterise the model. The permutation importance is usually applied after fitting the model. This will calculate the impact on prediction accuracy when we randomly re-ordered a single column. Shuffling a single column that is strongly correlated with the target value randomly will lead to a decrease in prediction accuracy[**?**].

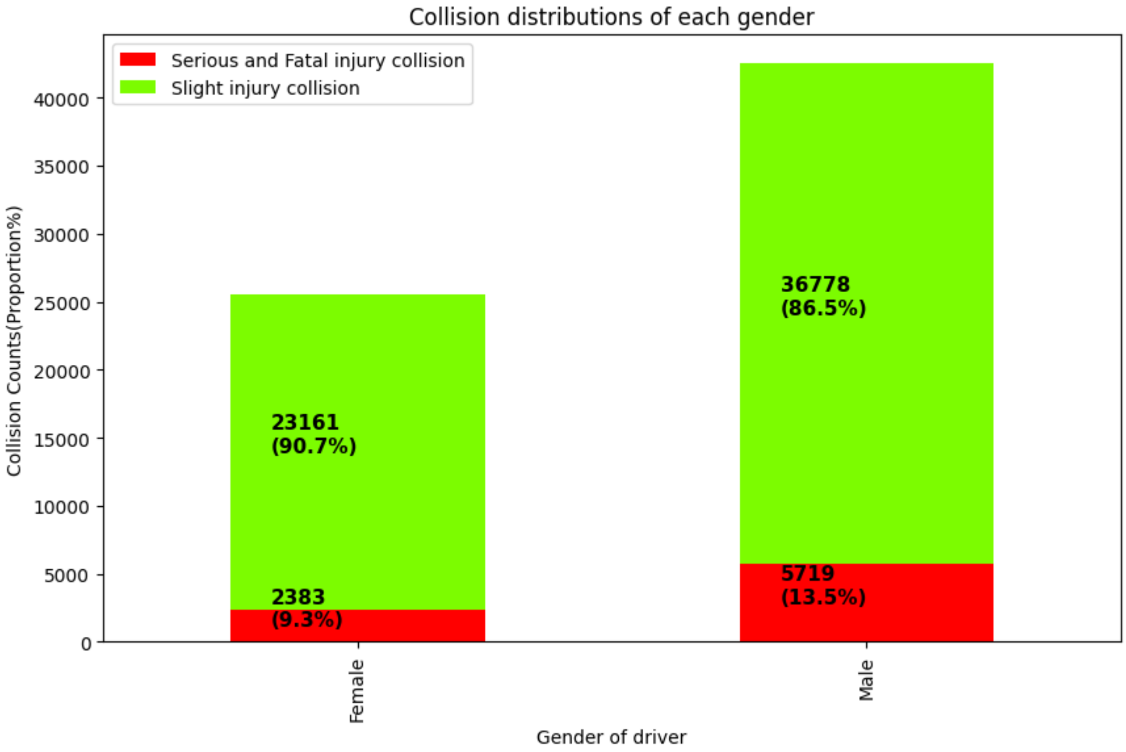

Figure 3.7: Histogram of collision distributions of every gender of the driver (without Unknown)

We use SHAP to have an explanation for the entire model, as seen in Figure 3.10. Similarly, the absolute SHAP value illustrates the influence of a single characteristic on the prediction, and the negativity shows us the contribution direction. A SHAP value less than 0 indicates that the model tends to predict minor injuries, while a value greater than 0 indicates that the model tends to predict severe injuries. The colour indicates the feature value from low to high from blue to pink as shown in the color bar at the right. The colour is closer to pink, indicating a greater value.

For those features that have a bigger value of getting serious injuries than slight injuries, choosing a motorcycle as transportation has the greatest value, which is around 2. This represents riding a motorcycle is very likely to be seriously injured in a car accident. Driving at night, and riding bicycles are also easier to be seriously injured.

And for the value points to have slight injuries, the vehicle slowing or stopping is the highest. Followed by Belfast city policing district and the speed limit is below 30 mph. The one thing all these have in common is that the feature value of getting minor injuries is higher than getting fatal injuries.

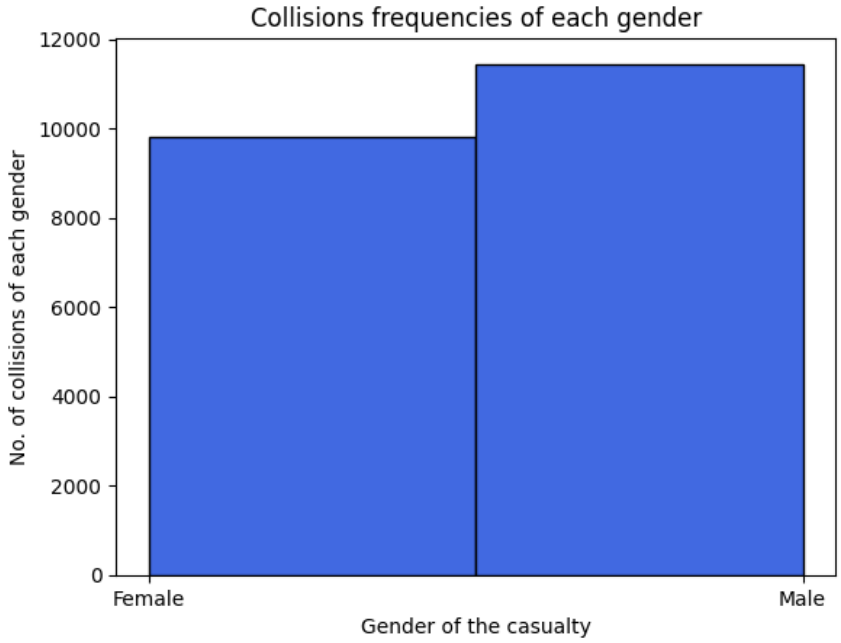

Figure 3.8: Histogram of collision distributions of every age group of the casualty (without Unknown)

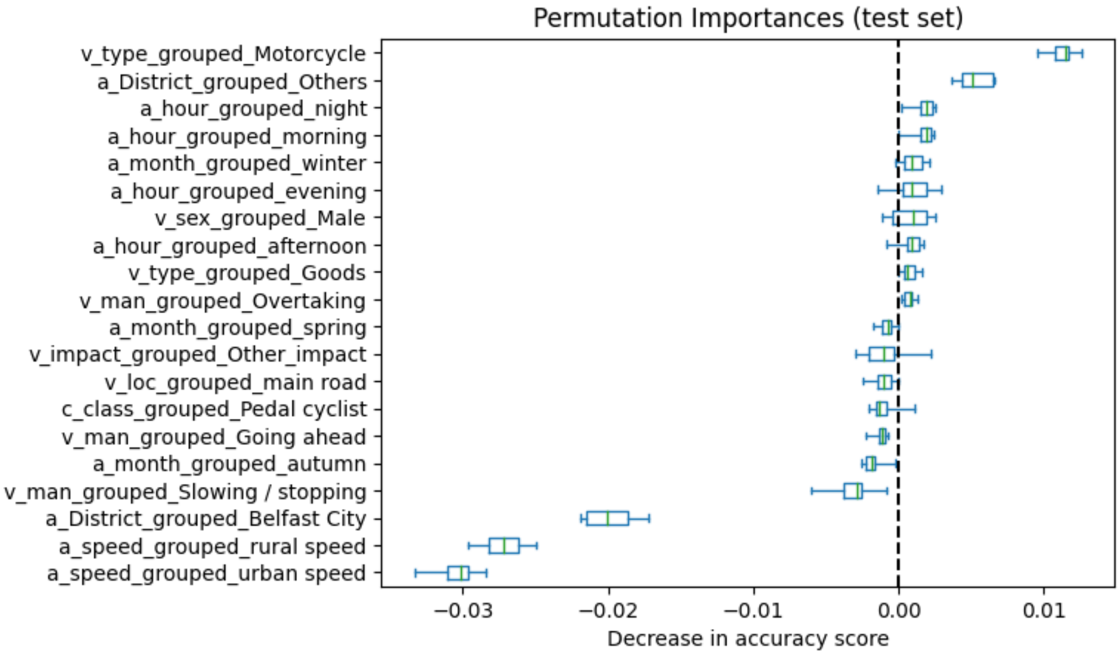

Figure 3.9: First 10 and last 10 feature importances

Model name	Techniques	Scores name	Scores
Logistic Regression	General model	Accuracy	0.9178
		f1_score	0.9569
		AUC	0.5290
		Average Precision	0.9153
		Confusion Matrix	[[8554, 47], [723, 49]]
Logistic Regression	Hyper-parameter Tuning	Accuracy	0.6897
		f1_score	0.8020
		AUC	**0.7154**
		Average Precision	0.8884
		Confusion Matrix	[[5882, 2719],[233, 539]]
Logistic Regression	Threshold Tuning	Accuracy	0.7138
		f1_score	0.2723
		AUC	0.7138
		Average Precision	0.1470
		Confusion Matrix	[[5528, 3073],[166, 606]]
Logistic Regression	SMOTE	Accuracy	0.5115
		f1_score	0.2264
		AUC	0.6737
		Average Precision	0.1239
		Confusion Matrix	[[4124, 4477],[102, 670]]
Logistic Regression	Group K-Folds CV	Accuracy	0.6894
		AUC	0.7091
Random Forest	General model	Accuracy	0.8236
		f1_score	0.2102
		AUC	0.5785
		Average Precision	0.1064
		Confusion Matrix	[[7500, 1101],[552, 220]]
Random Forest	Hyper-parameter Tuning	Accuracy	0.8224
		f1_score	0.8998
		AUC	0.5837
		Average Precision	0.9062
		Confusion Matrix	[[7478, 1123],[542, 230]]

Table 3.2: Model Iterations

Model name	Techniques	Scores name	Scores
One-Class SVM	General Model	Accuracy	0.9093
		f1_score	0.9524
		AUC	0.5049
		Average Precision	0.9184
		Confusion Matrix	[[16, 756], [94, 8507]]
One-Class SVM	Hyper-parameter Tuning	Accuracy	0.9063
		f1_score	0.9507
		AUC	0.5074
		Average Precision	0.9188
		Confusion Matrix	[[23, 749], [129, 8472]]
Isolation Forest	General model	Accuracy	0.8660
		f1_score	0.9271
		AUC	0.5479
		Average Precision	0.9250
		Confusion Matrix	[[129, 643],[613, 7988]]

Table 3.3: Model Iterations (continued)

Scores Matrix	
Accuracy	0.6897
F1-score	0.8020
AUC	0.7154
Average Precession	0.8884

Table 3.4: Matrix of scores of the optimal model

Confusion Matrix

		Prediction	
		Slight	Serious and Fatal
True Label	Slight	5889	2712
	Serious and Fatal	196	576

Table 3.5: Confusion Matrix of the optimal model

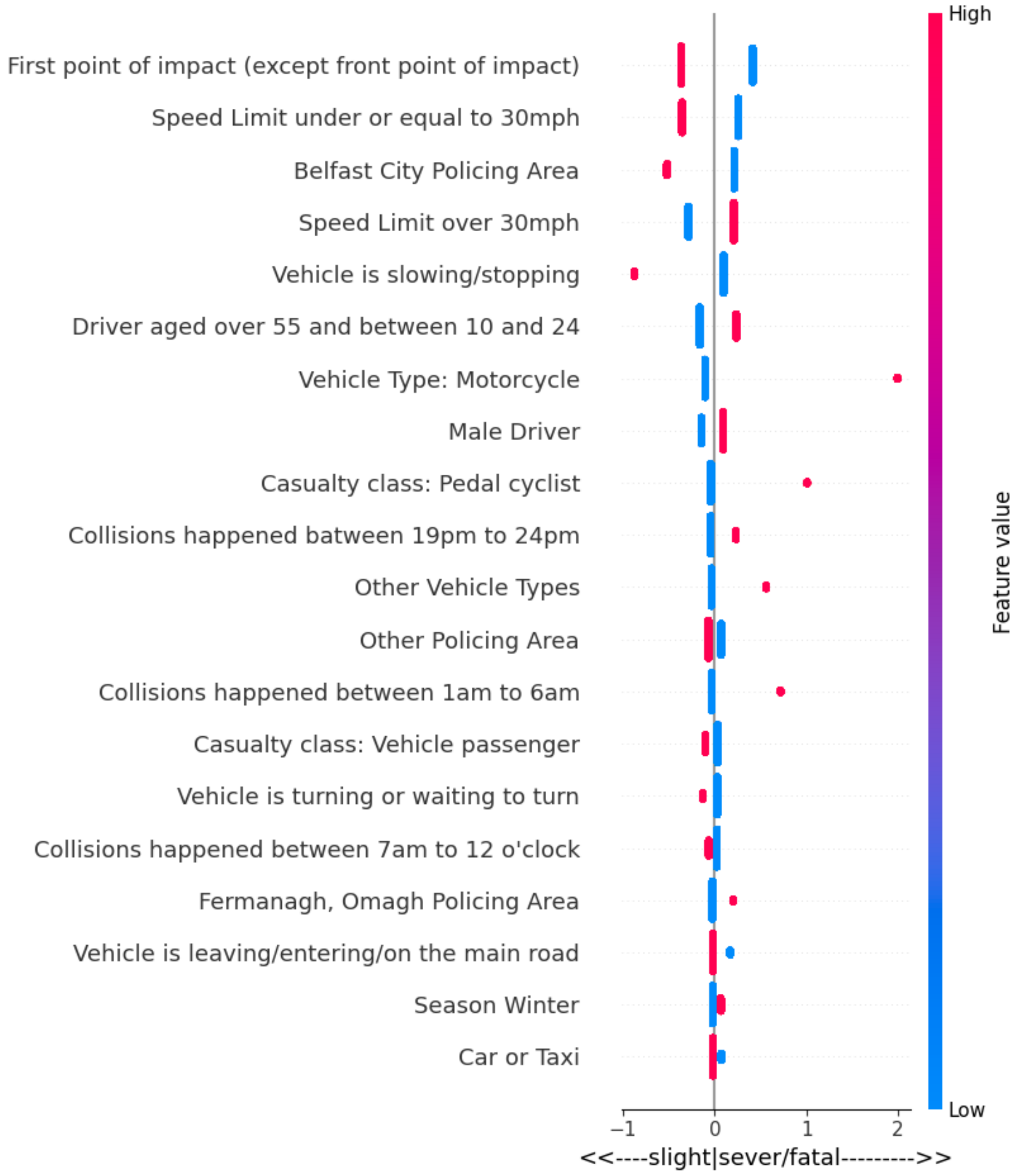

FIGURE 3.10: Contribution of each feature on predicting injury severity for the entire model

Chapter 4

Discussion

4.1 Distributions and correlations of features

An initial objective of this project is to identify injury severity of traffic road collisions in Northern Ireland. The first part of this project reveals the distribution and correlation of each variable.

From Figure 3.7, male drivers not only cause more crashes than females, but they also cause more fatal crashes. Through the article published by Al-Balbissi in 2010 [27], we can obtain a similar conclusion. Gender plays an important role in impacting accident occurrences. The male accident rates are consistently and significantly higher than female accident rates and accidents caused by males are more harmful.

Friday has the most crashes, but Sunday has the highest probability of being seriously or fatally injured in road accidents(Figure 3.5). In our expectation, weekends are supposed to have more collisions than weekdays. However, as shown in Figure 3.4, weekends are the two days with the fewest accidents, while Friday has the most. A possible explanation for this might because people need some immediate relaxation on Friday after a whole week's work. And to prepare for next Monday's work, most of the people prefer to stay at home on Sunday. Although Sunday has the least accidents, Sunday has a 16.9% chance of getting life-threatening injuries. Those having a short trip to somewhere close will get back to their accommodation before Monday. This result might be explained by the fact that long hours of drowsy driving at night.

Young drivers do not have too much experience in driving, furthermore, some young people may engage in risky driving behaviours in order to show off to their peers, so relatively young drivers have a higher likelihood of causing car accidents, as shown in Figure 3.6.

Due to the bad weather in winter, people tend to choose driving as their transportation. Slippery roads due to rain and snow have caused the number of car accidents to gradually approach 3,000 per month since September, as shown in Figure 3.3. Data was collected from 2016 to 2021 and includes COVID-19. On March 23, the prime minister announced the lockdown policy to slow down the spread of the pandemic [28]. Residents in the UK can only go outside for limited reasons. The constraints were announced not to be lifted on the 14th of April [29]. Eventually, the prime minister said to lift restrictions on enterprises that are best able to introduce social distancing in the workplace in the first place at the end of April [30]. Since the entire month of April 2020 was spent in lockdown, the drastic reduction in the number of people going outside had a direct impact on the number of car accidents. This is also the reason for the lowest number of car accidents in April.

The p-values of the significance test are all smaller than 0.05, demonstrating that the results are statistically significant and reject the null hypothesis. Hence all factors have a correlation with our target variable.

4.2 Predict serious and fatal injuries in a road collision

The classification model described in section 3.2.2 can identify whether traffic collisions will result in slight injuries or life-threatening injuries, by inputting the characteristics data of drivers and casualties as well as general information like weather and collision location.

The AUC score of the optimal model is 0.7154. The score is not ideal for a classification model and we would still prefer this value to be closer to one. However, this is probably the best result we can have here, due to the limitation of the data. The data only records collisions where PSNI was present, so in a way, the data we have is incomplete. And because some of the crashes are very minor and may not have resulted in injuries, or may have only resulted in injuries to some of them, we have had to delete a lot of rows from this data. These are all possible reasons for the low AUC score.

From the permutation importance plot in Figure 3.9, the most important feature is choosing a motorbike as a way of transportation, reducing the accuracy by over 0.01. Followed by the policing districts grouped as Others, reducing the accuracy to around 0.005. In other words, the vehicle type and policing districts need to be filled in to ensure the classification results.

It is worth noting that accuracy has significantly increased as a result of shifting the speed and Belfast City columns. It is vital to take into account that the model may have discovered some spurious relationship in the data set, comparing this improvement with the rest of the model. In this scenario, it's crucial to verify whether the model's performance changes significantly when these two features are removed.

The SHAP figure can give us guidance on which characters can contribute to severe accidents, and others can contribute to slight accidents.

The driver of a motorcycle has less protection than the vehicle driver comparatively. Also, driving at night usually comes with poor lighting and low visibility compared to daytime. There is a thin layer of ice on the road due to the low temperature in winter. This indicates riding a motorcycle or bicycle, travelling at night and driving in winter tends to cause severe collisions.

On the contrary, when driving in Belfast and the first point of the impact is not at the front usually leads to slight injuries. Belfast City has more population than other rural areas so it is relatively likely that there will be some minor traffic accidents with small scratches and bumps. The first point of the impact is not at the front can avoid the driver's position, and the driver is the person most likely to be injured in the vehicle, so usually leads to slight injuries.

Chapter 5

Conclusion

In conclusion, we develop a Logistic Regression classification model to identify the injury severity of road traffic collisions using data from 2016 to 2021. Through a systematic methodology of preprocessing the data, feature engineering, model selection, and evaluation, a predictive tool was established.

From the model, we can tell that people who ride motorcycles or bicycles as a way of transportation, travel at night, drive on a road whose speed limit is over 30 mph, are under 24 or above 55, are male and in winter are very likely to get severe injuries.

For those crashes that cause slight injuries, they may have the following characteristics, happen while slowing or stopping, the location of the crash is not at the front, in Belfast city, drive on a road whose speed limit is under or equal to 30 mph and as a vehicle passenger.

The application of this predictive model is not limited to their direct use in injury severity classification. Policymakers can leverage model insights to develop targeted safety interventions, such as road infrastructure improvements or awareness campaigns. By predicting the potential consequences of accidents, proactive measures can be taken to reduce severe injuries and save lives. In addition, insurance companies can use this model to give guidance while quoting car insurance. Customers who usually ride a motorcycle at night, are more likely to be seriously injured in a car crash, and they are charged more than those who are likely to have slight injury. Furthermore, the car rental company can track a driver's driving patterns and inform customers that they will be charged a higher rate if they are categorised as likely to be seriously injured. Doctors and nurses in hospitals can also use this model to estimate the extent of injuries sustained by the injured in a car accident and prepare accordingly.

5.1 Future outlook

The COVID-19 pandemic restricted residents from leaving their houses and changed their driving habits on the road. Since there were fewer vehicles on the road at that time, drivers tended to do more dangerous driving behaviours like speeding, which are very likely to cause more life-threatening collisions. These driving habits are not the same as in post-pandemic times. Therefore, having the data in the pandemic as a part of the training data may lead to inaccurate predictions.

To help better train the model, we can import extra external datasets that may be correlated to injury severity in future analyses. The data on weather and public transportation can be helpful.

Finally, we regroup features in various ways to improve the AUC score during feature engineering. This is based on the values of each variable. Principle component analysis (PCA)[31] can be applied later to reduce the dimensionality of the large dataset and may improve the AUC score to gain a better model.

Bibliography

[1] Conor Ashleigh. Road traffic injuries. URL https://www.who.int/health-topics/road-safety#tab=tab_1.

[2] May 2023. URL https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-provisional-results-2022/reported-road-casualties-great-britain-provisional-results-2022#headline-figures.

[3] Dec 2015. URL https://www.nidirect.gov.uk/articles/ni-road-safety-partnership.

[4] *BBC News*, Apr 2016. URL https://www.bbc.co.uk/news/uk-northern-ireland-36001607.

[5] *BBC News*, Aug 2023. URL https://www.bbc.co.uk/news/uk-northern-ireland-66450277.

[6] Statistics Analysis and Research Branch. Northern ireland road safety strategy to 2020 statistics — department for infrastructure, May 2016. URL https://www.infrastructure-ni.gov.uk/articles/northern-ireland-road-safety-strategy-2020-statistics.

[7] Google Developers. Google maps platform documentation places api, . URL https://developers.google.com/maps/documentation/places/web-service.

[8] Google Developers. Google maps platform documentation geocoding api, . URL https://developers.google.com/maps/documentation/geocoding.

[9] J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009.

[10] Liang Jie, Chen Jiahao, Zhang Xueqin, ZHOU Yue, and LIN Jiajun. One-hot encoding and convolutional neural network based anomaly detection. *Journal of*

Tsinghua University (Science and Technology), 59(7):523–529, 2019. doi: 10.16511/j.cnki.qhdxxb.2018.25.061.

[11] Eric Jackson and Rajeev Agrawal. *Performance evaluation of different feature encoding schemes on cybersecurity logs.* IEEE, 2019. doi: 10.1109/SoutheastCon42311.2019.9020560.

[12] Jan Salomon Cramer. The origins of logistic regression. 2002. URL http://dx.doi.org/10.2139/ssrn.360300.

[13] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001. URL https://doi.org/10.1023/A:1010933404324.

[14] Hyun Joon Shin, Dong-Hwan Eom, and Sung-Shick Kim. One-class support vector machines—an application in machine fault detection and classification. *Computers & Industrial Engineering*, 48(2):395–408, 2005. URL https://doi.org/10.1016/j.cie.2005.01.009.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. URL https://doi.org/10.1109/ICDM.2008.17.

[16] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009. URL https://doi.org/10.1142/S0218001409007326.

[17] Schratz Patrick, Muenchow Jannes, Eugenial Iturritxa, Richter Jakob, and Brenning Alexander. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120, 2019. ISSN 0304-3800. URL https://doi.org/10.1016/j.ecolmodel.2019.06.002.

[18] Muhammad Qasim Ali, Ehab Al-Shaer, Hassan Khan, and Syed Ali Khayam. Automated anomaly detector adaptation using adaptive threshold tuning. *ACM Trans. Inf. Syst. Secur.*, 15(4), apr 2013. ISSN 1094-9224. doi: 10.1145/2445566.2445569. URL https://doi.org/10.1145/2445566.2445569.

[19] Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, Sandro Ridella, et al. The'k'in k-fold cross validation. In *ESANN*, pages 441–446, 2012.

[20] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the

15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018. URL https://doi.org/10.1613/jair.1.11192.

[21] scikit-learn developers. sklearn.linear_model.logisticregression — scikit-learn 0.21.2 documentation, 2014. URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[22] Scikit learn developer. sklearn.ensemble.randomforestclassifier — scikit-learn 0.20.3 documentation, 2018. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[23] Scikit learn developer. sklearn.svm.oneclasssvm — scikit-learn 0.24.2 documentation. URL https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html.

[24] scikit-learn developers. sklearn.model_selection.gridsearchcv — scikit-learn 0.22 documentation, 2019. URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[25] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2019.

[26] Cabinet Office. Staying at home and away from others (social distancing), Mar 2020. URL https://www.gov.uk/government/publications/full-guidance-on-staying-at-home-and-away-from-others.

[27] Adli H. Al-Balbissi. Role of gender in road accidents. *Traffic Injury Prevention*, 4 (1):64–73, 2003. doi: 10.1080/15389580309857. URL https://doi.org/10.1080/15389580309857. PMID: 14522664.

[28] Prime Minister's Office. Pm address to the nation on coronavirus: 23 march 2020, Mar 2020. URL https://www.gov.uk/government/speeches/pm-address-to-the-nation-on-coronavirus-23-march-2020.

[29] BRONWEN MADDOX. The hard choices in lifting the coronavirus lockdown, Apr 2020. URL https://www.instituteforgovernment.org.uk/article/comment/hard-choices-lifting-coronavirus-lockdown.

[30] The government must be straight with the public: there can be no single grand exit plan to release the coronavirus lockdown, Apr 2020. URL https://www.instituteforgovernment.org.uk/article/press-release/lifting-lockdown-coronavirus-exit-strategy.

[31] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. URL https://doi.org/10.1098/rsta.2015.0202.