澳門大學

UNIVERSIDADE DE MACAU

FACULTY OF BUSINESS ADMINISTRATION

Data Analysis and Prediction of the UK Road Traffic Accidents Using Machine Learning

ISOM4008 – 001

Machine Learning for Business Intelligence

BB900537 IONG KA CHON, Karson

BB903426 WU SHIHAN, Skylar

BB906212 JI XU, Jason

BB907379 MA YIHANG, Estelle

Abstract

In recent years, the increasing pressure on traffic safety and the high number of road traffic accidents not only cause many casualties in various countries, but also bring a lot of medical expenses, administrative expenses, production losses and property losses. Therefore, how to prevent road traffic accidents scientifically and effectively has become the focus of traffic safety research. Many researchers use machine learning models to predict the severity of road traffic accidents, so as to explore the factors influencing the severity of road traffic accidents and make recommendations for prevention. However, the complexity of the elements involved in road traffic accidents, including road, environment, vehicles, people, and other related elements, in addition to the fact that most of the accidents in road traffic accidents are minor accidents and there is unbalanced data, these characteristics increase the difficulty of predicting the severity of road traffic accidents.

This paper integrates multiple sources of traffic accident data files to comprehensively consider basic accident conditions. For the unbalanced nature of the data, the data is resampled using random oversampling, and four models, logistic regression, artificial neural networks, XGBoost and decision tree, are built for the resampled training set. Then ANN and Decision Tree model had been optimized by changing activation, hidden layer and epochs and adjusting the Max-Depth respectively. Eventually, the optimal model is then obtained by comparing the eight models. This model can not only provide suggestions for the prevention of series road traffic accidents, but also provide a reference system for traffic enforcement departments to quickly determine the severity of accidents and for insurance companies to quickly accomplish compensation process. Although our model could achieve the objective mentioned at the beginning of the paper, there are some limitations when building a prediction model and improvement to do in the future.

Keywords: Machine Learning, Exploration Data Analysis, Decision Tree, Artificial Neural Network, XGBoost, Logistic Regression, Random Forest.

Contents

1. Introduction

The road safety is a very serious and fatal problem for many countries. The World Health Organization is running a campaign called the 'Decade of Action for Road Safety' (2021) [1]. Accidents on the road network are the second biggest killer of younger adults. According to the statistics of WHO, there are more than 1.3 million people are killed on the road each year, an average of just under 3,700 people each day. Another 20 to 50 million people are victims of non-fatal injuries, often resulting in some form of permanent or temporary disability. Furthermore, most of countries lose between 1% and 3% of their GDP, for instance: the loss of human life and personal suffering, in emergency services and health services such as funeral costs and medical costs, as well as repair of damaged roads and investigation of traffic accident-prone areas.

Driving is considered the most dangerous daily activity people do every day, which reflects the thousands of studies, campaigns, and programs developed to improve problems of traffic congestion, human death, health problems, environmental pollution, and economic losses [2]. So new technologies and ways need to devise to uncover accident-relate factors and to identify time and space that are risky, thus supporting traffic accident data analysis in decision-making processes.

In this paper, we proposed a novel end big data analytics system for UK traffic accident severity analysis using data mining and deep learning techniques. We first displayed the data on an interactive map that will effectively highlight disaster hotspots in time and space. We then visualized some of the features in the data to find the key factors that are high-relative with a road accident. Finally, machine learning algorithms are used to predict future traffic accidents severity. The goal of this project is to gain insight into the current climate of road traffic accidents in the UK, to gain additional understanding of what drives severity of road accidents, and to investigate the effectiveness of various machine learning techniques in the prediction of road accident casualty fatalities. Multiple datasets were collated and explored by using several machine learning models, and then be trained and tested to evaluate and compare the predictive accuracy.

2. Objectives

2.1 Business Objective

For insurance companies: The process of insurance compensation is cumbersome and takes a long time, causing great trouble to both the company and the beneficiary. Inferring the evaluation mode of the original dataset through machine learning, an automated car accident verification degree system is established. This can reduce the human and material resources consumed by the insurance company in the claim process and can also greatly improve work efficiency. The company places the predictive system in the car of policyholders who buy auto insurance. When an accident occurs, the predictive system will predict the severity of the accident based on real-time conditions and send the information back to the insurance company. After receiving the forecast results, the insurance company can draw up the required documents and prepare money in advance, and then pass the analysis. After the information is correct, the insurance payment can be settled in advance, thereby speeding up the compensation process, enabling car owners to obtain compensation in advance and improving customer satisfaction.

2.2 Other Objectives

Through exploratory research on multiple road traffic accident data sets, the current road traffic accident environment can be deeply understood, and its generalization can be explored to apply to road safety management in each country. This project can help the government comprehend the corresponding reasons and circumstances of accidents' occurrence that affect the severity of accidents and improve the roads to a certain extent within the feasible range according to these factors.

The traffic accident severity prediction system is installed on the road or on the vehicle. When an accident occurs, the local traffic bureau will assign work by observing the signals sent by the system. According to the severity of traffic accidents in different degrees, different personnel are allocated. Therefore, in addition to speeding up the overall rescue speed, an appropriate amount of rescue personnel can also be arranged to carry out rescue work.

3. Literature Review

The UK has one of the largest urban bus networks in the world resulting in a high number of vehicles on the road, which will lead to the development of road facilities and technology to ensure that road safety issues are always considered a priority by road users. Various studies have devoted to the aspects of road safety problem in UK, while most of which focus on using data mining and machine learning techniques to analyze traffic accidents. The research work carried out by Nour, M. K.et al. [3] used several classification models which are selected for the ability to deal with multi-class classifications, and ultimately found that tree-based techniques such as XGBoost outperform regression based one, such as ANN. In Feng et al. [4] system, several state-of-the-art deep learning and time series forecasting model to predict the number of road accidents in the future to assist decision making. Saini et al. [5] applied three machine learning algorithms (KNN, random forest and SVM), which are used in three classification problems. Yun [6] considered the non-balanced nature of data, so he used four methods of RUS, ROS, SMOTE and ADASYN to sample the data, and established three models of logistic regression, random forest, and ANN for sampled training set. Another research undertaken by Koh et al. [7] focused on how the logistics regression, random forest and SVM performed on predicting UK accidents based on their accuracy and ROC-AUC values.

There are some studies on similar topics using data from other countries. Sarkar et al. [8] in India classified the severity data of traffic accidents in Kolkata from 2008 to 2012 by using multiple logistic regression and artificial neural models, and ANN models has higher accuracy and less error. According to one such study conducted by Asare et al. [9], they used an ordered logistic regression model to analyze the traffic accident data in Ghana from 1989 to 2019. The article screened the variables through the Chi-square test and explained the coefficients of the ordered logistic regression model in detail: it shows that the nature of the car, whether it is on a national car, whether it is speeding and whether it is in a city, is an important indicator of the traffic severity. However, the research did not consider the unbalanced nature of the data. Kononen et al. [10] predicted the severity of accidents occurred in United States by using logistic regression model. They reported performance 40% and 98% for sensitivity and specificity respectively. They also identified the most important predictors for traffic accident severity level are change in velocity, seat belt use, and crash direction.

4.Data Analysis

4.1 Data Description

The Road Safety Data is soured from UK Government's Data Repository and is published by the Department for Transport, UK [11]. The dataset comprises following files, and this project will extract only the last 20 years of data for analysis in the subsequent data processing.

Dataset	Number of Records	Number of Attributes
Accidents_2002-2021.csv	3.12 million	36
Casualties_2002-2021.csv	4.19 million	19
Vehicles_2002-2021.csv	5.73 million	28

Table 1 Dataset used by the project

The additional details corresponding to the attributes in each dataset are showed in the appendix file.

4.2 Data Preprocessing

Data preprocessing is a component of data preparation, it describes any type of processing performed on raw data to prepare it for ensuring or enhancing performance in the following machine learning simulation procedure and data analysis. It involves several different tools and methods such as data cleaning, data reduction, data transformation and so on. This section will introduce the pre-processing methods that utilized to handle the dataset, and the remaining adjusted attributes are saved in a "Merged_data.csv" document for subsequent features selection.

- Data Reduction

Data reduction is the process of reducing the amount of data records by eliminating invalid data and produce summary dataset but keeps the quality of original data. Some attributes will be removed since they may not be meaningful or useful for the next analysis task. For example, in the 'accidents' dataset, the features such as 'LSOA_of_Accident_Location'. 'Location_Easting_OSGR' and 'Location_Northing_OSGR' etc. are dropped. With similar situation, 'Driver_IMD_Decile', 'Towing_and_Articulation' etc. in vehicles' dataset and 'LSOA_of_Casualty', 'Casualty_IMD_Decile', etc. are also removed. Final, the remaining features in three different datasets are merged into one dataset by using the column "Accident_Index", "Accident_ Reference", and "Vehicle_Reference" as the shared primary key.

- Data Cleaning

Data cleaning is the process of detecting and correcting corrupt or inaccurate data which can ensure the raw dataset is suitable for feature engineering and data modelling. Since most machine learning models cannot process the null value, missing data must be imputed or removed. And then analyze the entire merged dataset to identify missing (missing, null, NA, None) which is assigned a value of "-1" in dataset.

In addition to missing value, there are also several features contained the unknown value using the label of “99” or others, for example, label of “9” in “pedestrian_crossing_physical_facilities”, weather_conditions” and “road_surface_conditions” etc. After recognizing and imputing the number of missing and unknown value, and those who have high proportion in corresponding attributes will be dropped as these attributes could not be considered reliable. Then deleting the row of records containing the missing value and unknown value in the remaining features to obtain a complete dataset.

- Data Transformation

Data transformation is the process of converting, cleansing and structuring data into a usable format that can be easily analyzed and make the most sense for the goal. Changing the type of the classified features such as “Light_Conditions”, “Vehicle_Leaving_Carriageway” etc. from float into categories.

Handling time by changing the specific time into period (1-5) of a day: 1- Morning: Between 5 and 10; 2- Office Hours: Between 10 and 15; 3- Afternoon Rush: Between 15 and 19; 4: Evening: Between 19 and 23; 5: Night: Between 23 and 5.

4.3 Exploratory Data Analysis

To better understand the data and gain look to gain insight into the figures, an exploratory analysis was carried out on the collated dataset using Python. Initially, the term of accidents severity has been explained.

Slight: An injury of a minor character such as a sprain (including neck whiplash injury), bruise or cut which are not judged to be severe, or slight shock requiring roadside attention. This definition includes injuries not requiring medical treatment.

Serious: An injury for which a person is detained in hospital as an 'in-patient', or any of the following injuries whether they are detained in hospital: fractures, concussion, internal injuries, crushing, burns (excluding friction burns), severe cuts, severe general shock requiring medical treatment and injuries causing death 30 or more days after the accident. An injured casualty is recorded as seriously or slightly injured by the police based on information available within a short time of the accident. This generally will not reflect the results of a medical examination but may be influenced according to whether the casualty is hospitalized or not. Hospitalization procedures will vary regionally.

Fatal: A casualty died at the scene, or within 30 days of the accident because of injuries from the accident.

- Basic Analysis

To investigate the location of most accidents, each accident was plotted using its longitude and latitude coordinates and a heat map was generated to visualize it. As expected, in Figure 1 below, we can see that major densely populated cities have the greatest number and density of traffic accidents and fatalities, particularly London and Birmingham, the area between Liverpool, Manchester, Leeds and Sheffield. The areas with the lowest densities of traffic accidents are in rural areas - primarily in Wales, as well as north close to the border with Scotland.

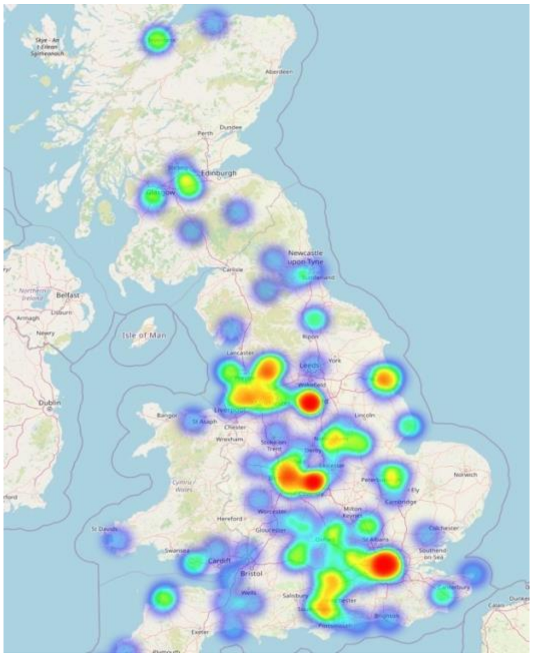


Figure1: Heatmap of Road Traffic Accidents

The pie plot in Figure 2 shows the proportion of accidents by severity, namely Slight, Serious & Fatal that took place in UK, over the time between 2002 to 2021. It clearly infers that a fairly large number of accidents i.e., 82.4% of the total accidents, were Slight Accidents. Whereas the proportion of Serious Accidents & Fatal Accidents comes out to be 15.4% & 2.23% respectively. That's the reason that the data need to be resampling after train the logistic model for the first time.

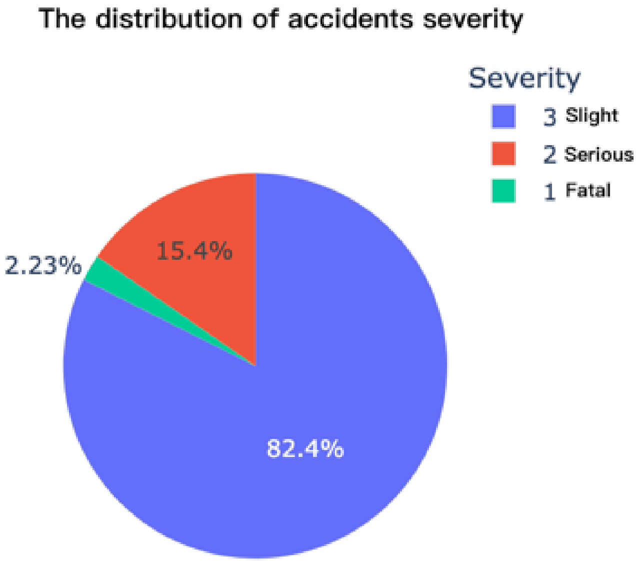


Figure 2: The distribution of accidents severity

- Time Analysis

A few points of interest a steady decline can be seen in the number of accidents occurring almost each year can be observed from Figure 3, while according to the RAC foundation, the UK is experiencing a some of its highest ever levels of traffic on motorways; reaching its highest ever level in 2021[12]. Interestingly, the average annual mileage per car is reported as decreasing but the number of vehicles has increased per household.

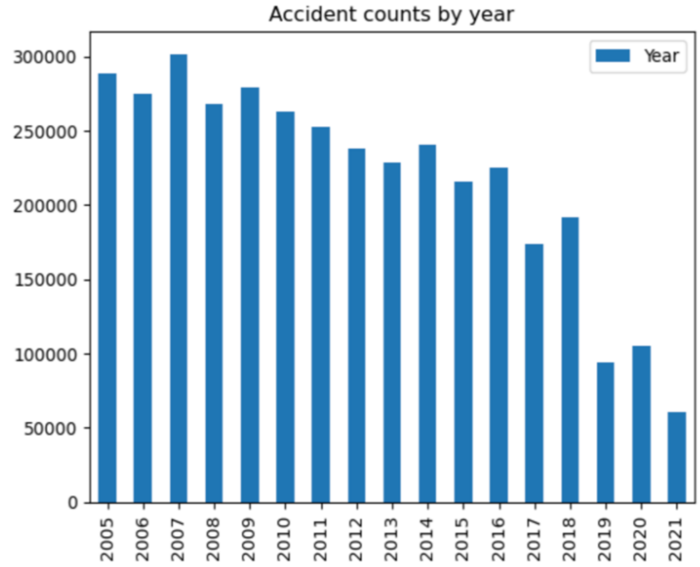


Figure 3: Accidents counts by Year

From the visualization, it can also be seen in Figure 4 that a lower volume of accidents can be seen as occurring over the weekends, with Sunday (reported as '1' in the data), shows the lowest volume of accidents since people are likely to take a break on Sunday before the workday. Meanwhile, there is a highest volume of accidents on Friday. Possibly they would drive to relax after hard weekdays.

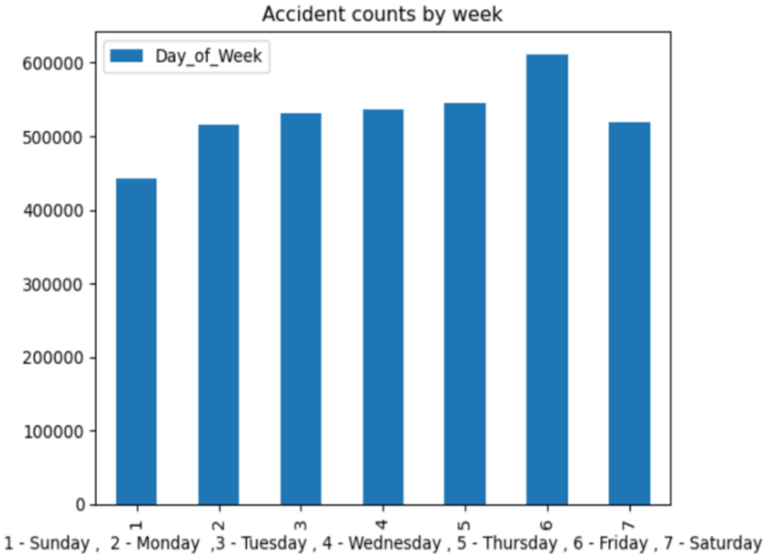
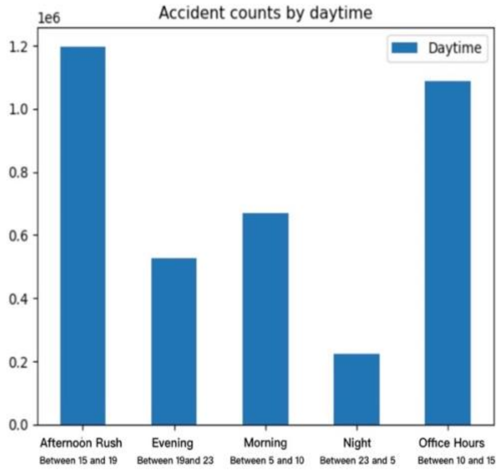


Figure 4: Accidents counts by week

The pattern seen in Figure 5(a) makes logical sense, showing more accidents in the usual commuting times (*Afternoon Rush*, *Office Hour*) fewer occurring at night and evening. Interestingly, we can see that in Figure 5(b), although less accidents are reported as occurring at night, there are more fatal accidents occurred during these hours



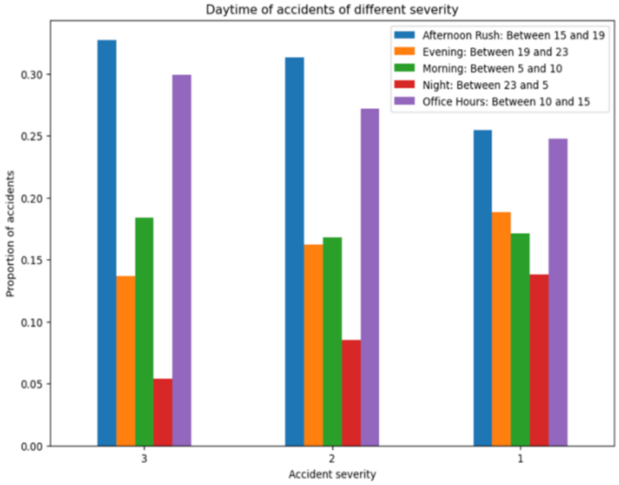
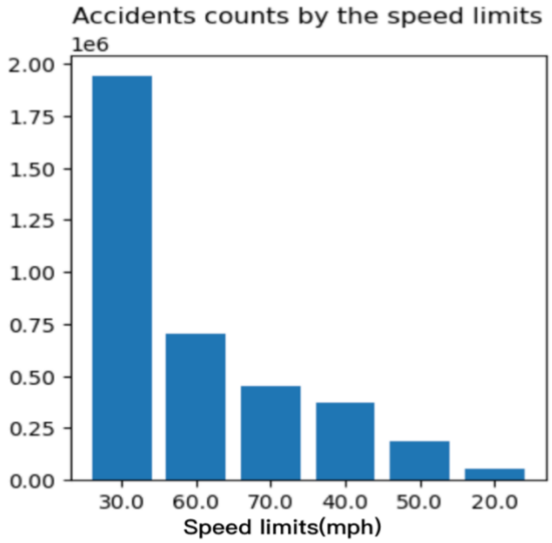


Figure5 (a): Accidents counts by daytime

Figure 5(b): Daytime with mean of the accidents severity

- Environment Analysis

Figure 6(a) shows that most accidents occur on roads with a speed limit of 30mph, followed by 60mph. Figure 6(b) confirm that fatal accidents tend to occur on faster roads. The median speed limit for fatal accidents is 50mph, compared to 30mph for slight and serious accidents. This is because the road speed limit is a proxy for the speed of the vehicles, and fatalities are more likely at higher speeds. Interestingly, this pattern does not hold for roads with a speed limit of 70mph, where there are a greater proportion of less fatal accidents than would be expected. Possibly these are accidents occurring at lower speeds on these high-speed roads (motorways), e.g., due to roadworks.



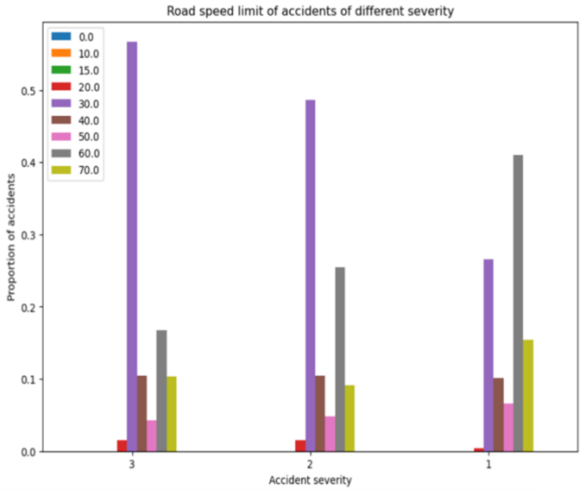


Figure 6(a): Accidents counts by speed limits

Figure 6(b): Accidents severity at different speed limits

Figure 7(a) shows that most accidents that occur in dark without lighting are fatal. Most of the injuries occurred during the day were slight. Figure 7(b) indicates that wet or damp road would cause the accidents to great extent. Interestingly, accidents in snowy conditions are less likely to result in fatalities - presumably because traffic tends to be travelling more slowly in the snow. From Figure 7(c), weather conditions have less of a notable effect on accident severity than light conditions and road surface conditions. Thus, the model does not select this feature to train prediction model.

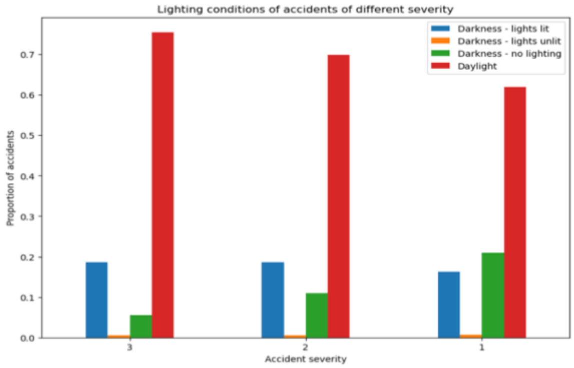


Figure 7(a): Lighting conditions of accidents severity

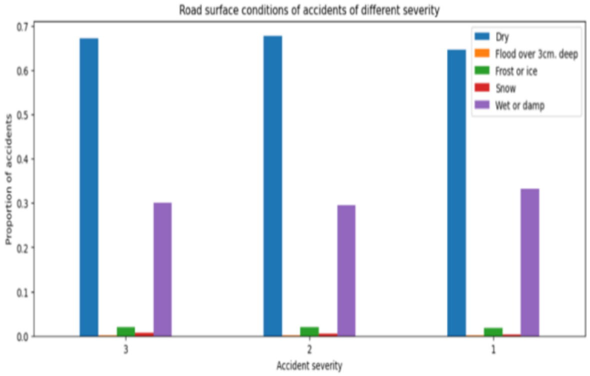


Figure 7(b): Road surface conditions of accidents severity

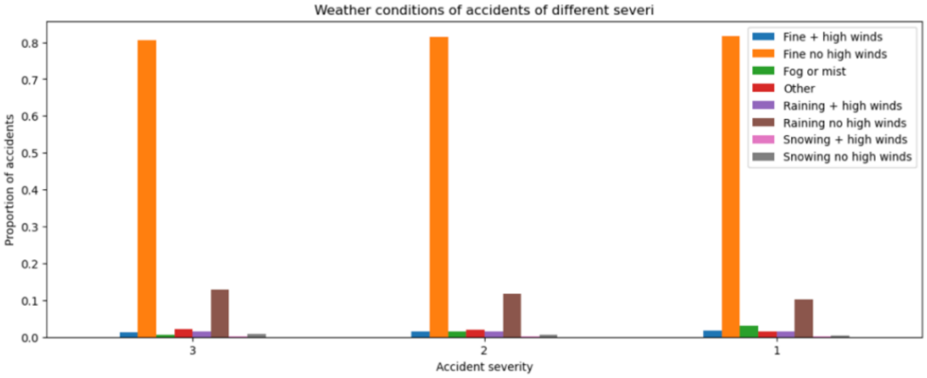


Figure 7(c): Weather conditions of accidents severity

Figure 8(a)(b) shows that fatal accidents occur in higher proportions on A roads, with 65% accidents and dual carriageways (larger, faster roads), and in lower proportions on A(M) and slip roads (smaller, slower roads) and roundabouts. Figure 8(c) shows that fatal accidents are much less likely to occur at junctions (presumably because this implies a lower speed). Accidents at junctions are more likely to only be slight.

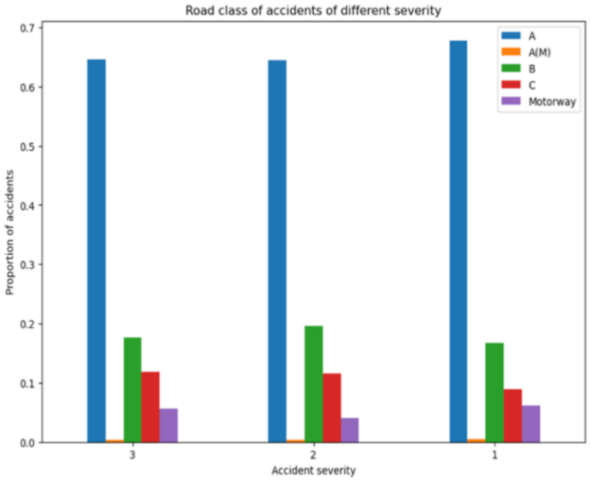


Figure 8(a): Road class of accidents of different severity

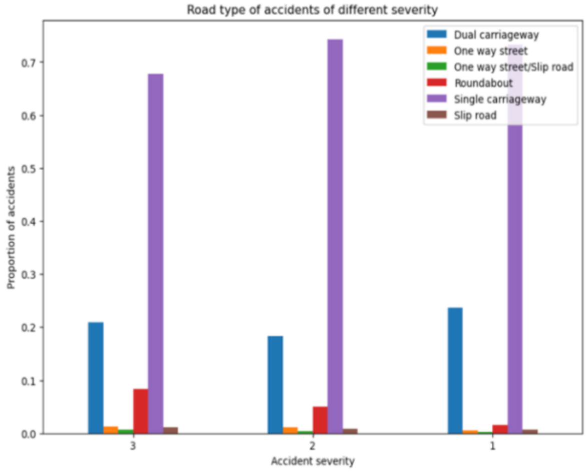


Figure 8(b): Road type of accidents of different severity

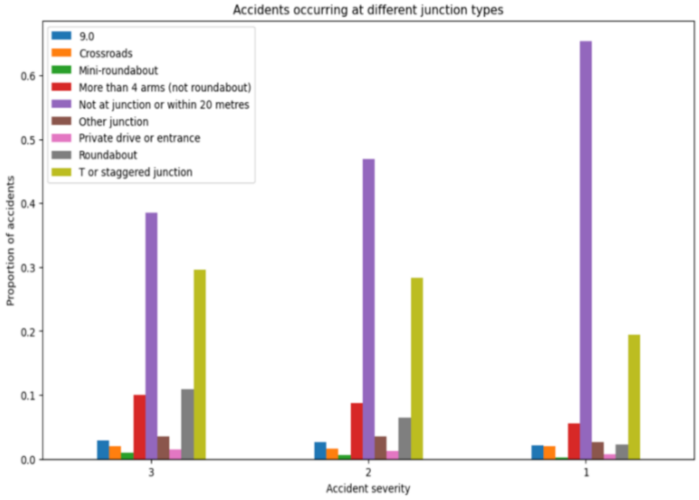


Figure 8(c): Accidents occurring at different junction types

- Casualty Analysis

Figure 9(a) that most of the people injured in accidents are between 26-55, as this age group has the highest number of drivers. However, according to Figure 9(b), it can be found that the injured masses of very high age and very low age are the most seriously injured. The age group over 75 has the most fatal injury.

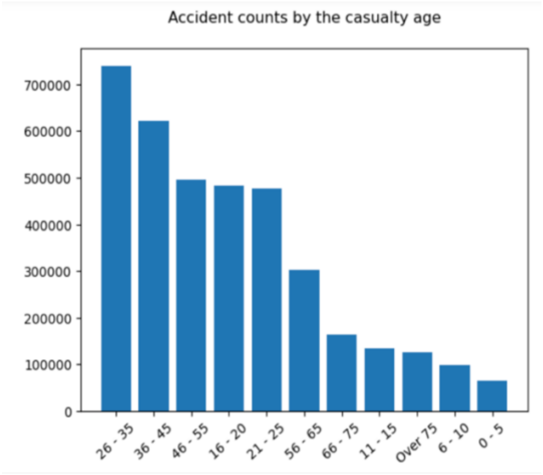


Figure 9(a): Accident counts by the casualty age

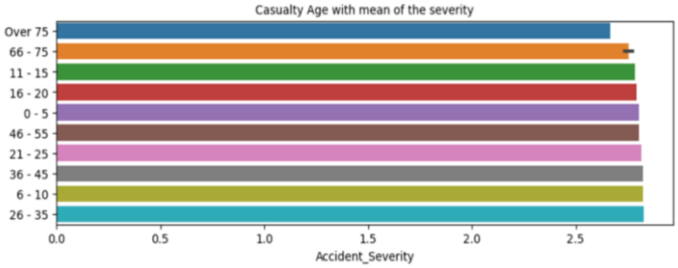


Figure 9(b): Casualty Age with mean of the severity

4.4 Feature Engineering

- Feature Deletion

In the datasets, there are some features to be highly correlated, meaning that certain features were linearly correlated with others. These features contribute very little to the predicted output but increase the computational cost. Therefore, we should remove these features. To achieve this step, we firstly removing the two target variables 'Accident_Severity', 'Casualty_Severity' and independent variables that cannot be predicted categorically including 'Accident_Index', 'Date', 'Latitude', 'Longitude'. We therefore identified highly correlated features greater than 0.5 (those with a dark red color) by using the upper triangle of the correlation coefficient matrix and removed one of each pair. Eventually, we dropped 10 features including "Police_Force", "Urban_or_Rural_Area", "Junction_Location", "Vehicle_Leaving_Carriageway", "Age_Band_of_Driver", "Bus_or_Coach_Passenger", "Sex_of_Casualty", "Casualty_Type", "Casualty_Class", "Pedestrian_Location".

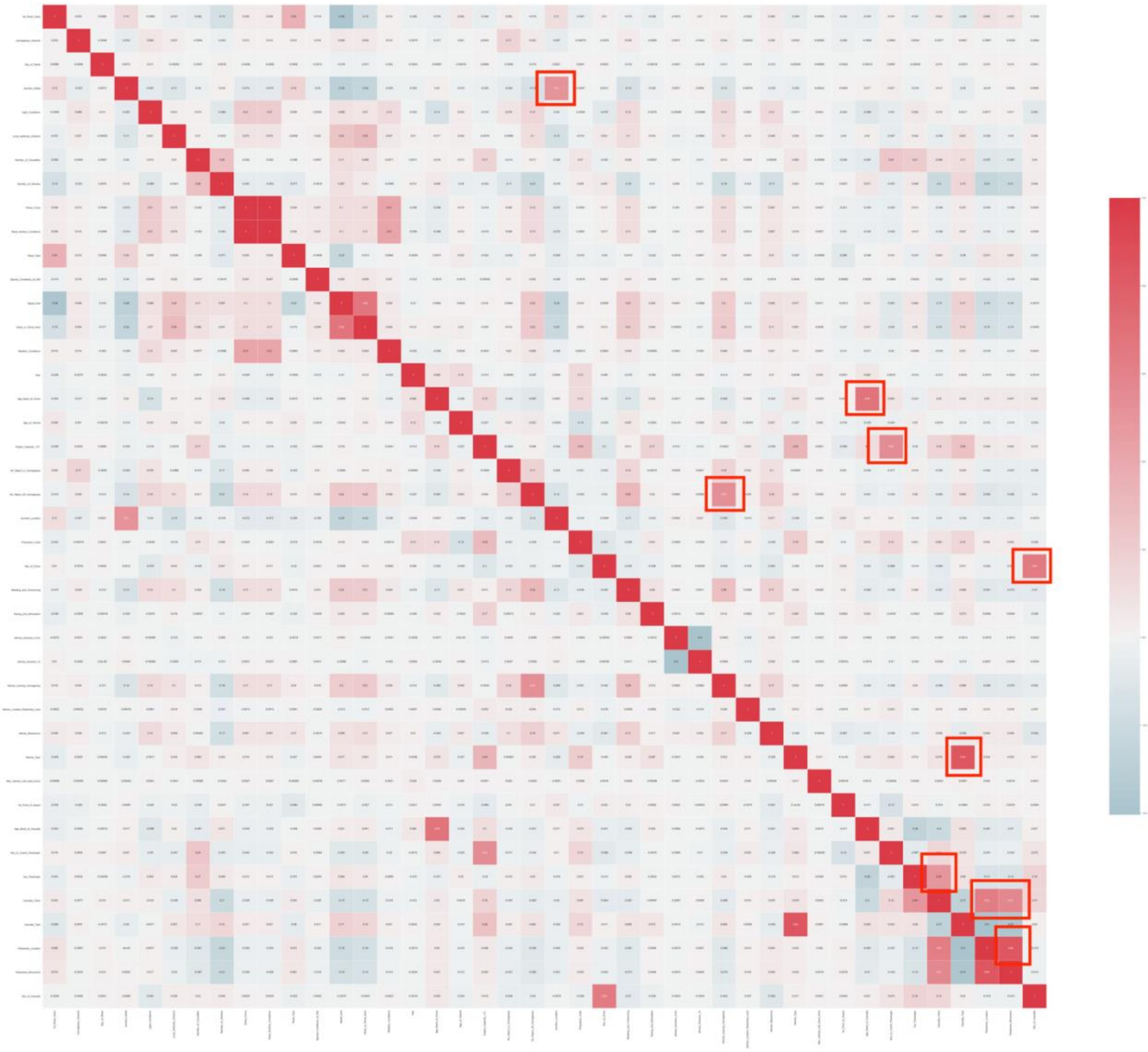


Figure 10: Correlation coefficient matrix of all features to accidents severity

- Feature Selection

Since there are still some low contributed features after the features deletion, it is necessary to drop the weak and redundant features. Random forest algorithm will be fitted to this task which can detect the weakly relevant features that can be completely obscured by the other features and discerning between weakly but truly relevant variables from those that are only seemingly relevant due to the random fluctuations. It is an ensemble of numerous weak classifiers (decision trees), each of these classifiers is constructed using different subset of variables and different subset of objects. During the construction process, each variable has numerous chances to be included in the classifier, so even weakly relevant attributes that are marginally related with the decision attribute will be used for construction of individual classifiers. Following horizontal histogram shows the proportion of features contributing to the severity of the accident, in descending order. We finally selected all features with a cumulative contribution greater than or equal to 70% in Figure 13

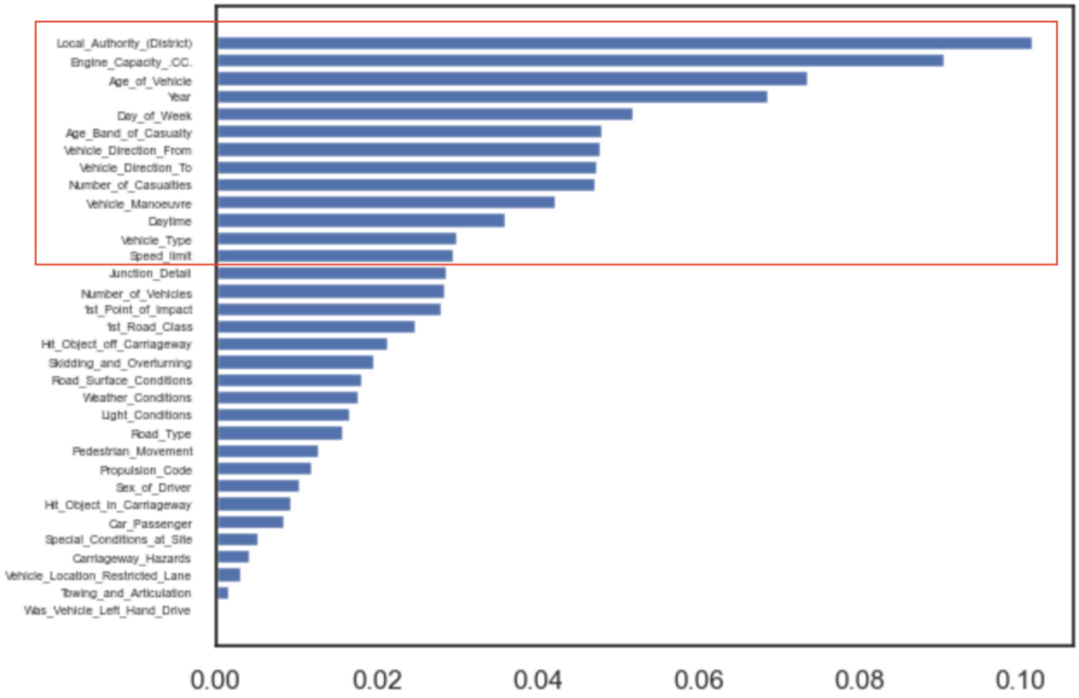


Figure 11: Proportion of features contributing to the accident severity

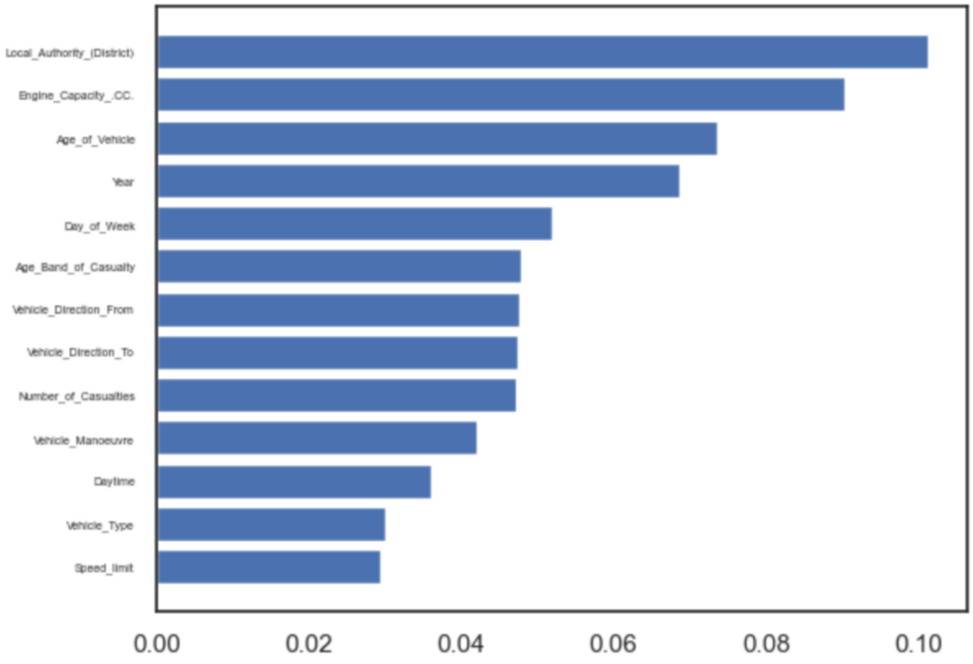


Figure 12: Features with a cumulative contribution ≥70% to the accidents severity

- Feature Transformation

Through feature selection, we selected 13 features with a cumulative contribution ≥70% to the accidents' severity. Then we converted almost all numerical data to categorical data except "Engine_Capacity_.CC" (it belongs to continuous numerical variable) since its number represent different category actually. For example: 1 in "Day_of_Week" represent " represent "Sunday", 2 in "Day_of_Week" represent " represent "Monday" etc. Finally, because we need to use the regression model in the training model, in order to show the relationship more realistically between the independent variable and the dependent variable, we assign values to the category variables and convert them into dummy variables "0" and "1". For example, for the categorical variable "Day_of_Week", if the day is "Sunday", it represents 1000000 and if the day is "Sunday", it represents 0100000.

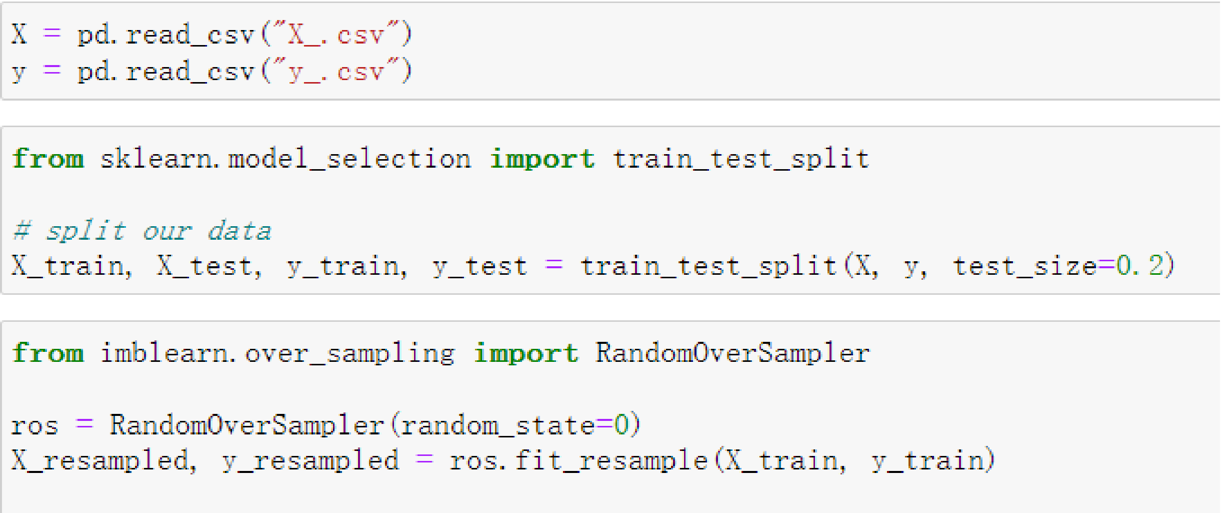
5. Data Modelling

5.1. Dataset Split and Scale

Before a dataset can be modelled for predictive analysis, it must be divided into training and testing subsets which can be used to train the model and evaluate its performance. We therefore split the dataset into 80% for training and 20% for testing. Meanwhile, because the range of variables in a dataset can vary widely, using the original scale may give more weight to variables with a larger range. To solve this problem, we will apply feature scaling techniques to the independent variables or features of the data so that each feature is equally important, which is known as "normalization".

5.2. Data Resampling

In this study, Random Oversampling is chosen. In this paper, samples from a small number of classes (hereafter referred to as "small number") are sampled to achieve the maximum number of samples from the largest class to achieve class balance. The essence of the random oversampling method is to randomly select a small number of sample classes and to increase the number of samples by replication.



```
X = pd.read_csv("X_.csv")
y = pd.read_csv("y_.csv")
```

```
from sklearn.model_selection import train_test_split

# split our data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
from imblearn.over_sampling import RandomOverSampler

ros = RandomOverSampler(random_state=0)
X_resampled, y_resampled = ros.fit_resample(X_train, y_train)
```

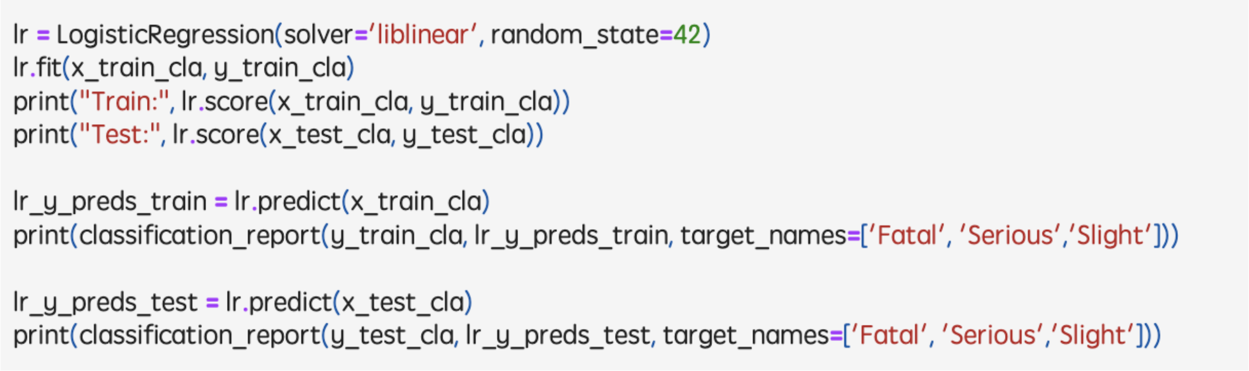
Code 1: Resampling the independent variables of training dataset

5.3. Methodology

- Logistic Regression

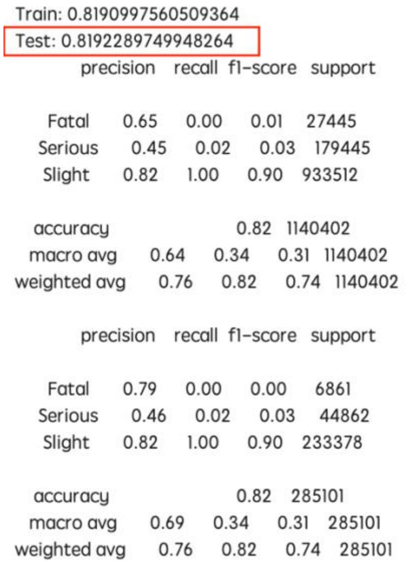
Logistic regression is a simple and widely used classification model, and it is also a special generalized linear regression model. It is also one of the most widely used models in the prediction of the severity of road traffic accidents.

Useful of Logistic Regression in accident severity prediction model: The model uses logistic regression to predict accidents severity. It considers road traffic accident as rare events and subjects the resulting probability values to correction steps to account for the rarity of positive samples (accidents). Our model has more than 30 features, and logistic regression can clearly explain the impact of parameters representing each feature on the output. In addition, the efficient operation of the model is also very useful for the big data model.



```
lr = LogisticRegression(solver='liblinear', random_state=42)
lr.fit(x_train_cla, y_train_cla)
print("Train:", lr.score(x_train_cla, y_train_cla))
print("Test:", lr.score(x_test_cla, y_test_cla))

lr_y_preds_train = lr.predict(x_train_cla)
print(classification_report(y_train_cla, lr_y_preds_train, target_names=['Fatal', 'Serious','Slight']))

lr_y_preds_test = lr.predict(x_test_cla)
print(classification_report(y_test_cla, lr_y_preds_test, target_names=['Fatal', 'Serious','Slight']))
```

Code 2: Construction of Logistic Regression



```
Train: 0.8190997560509364
Test: 0.8192289749948264
              precision    recall  f1-score   support

       Fatal       0.65      0.00      0.01     27445
     Serious       0.45      0.02      0.03    179445
      Slight       0.82      1.00      0.90    933512

    accuracy                           0.82   1140402
   macro avg       0.64      0.34      0.31   1140402
weighted avg       0.76      0.82      0.74   1140402

              precision    recall  f1-score   support

       Fatal       0.79      0.00      0.00      6861
     Serious       0.46      0.02      0.03     44862
      Slight       0.82      1.00      0.90    233378

    accuracy                           0.82    285101
   macro avg       0.69      0.34      0.31    285101
weighted avg       0.76      0.82      0.74    285101
```

Figure 13: Classification Report of Logistic Regression

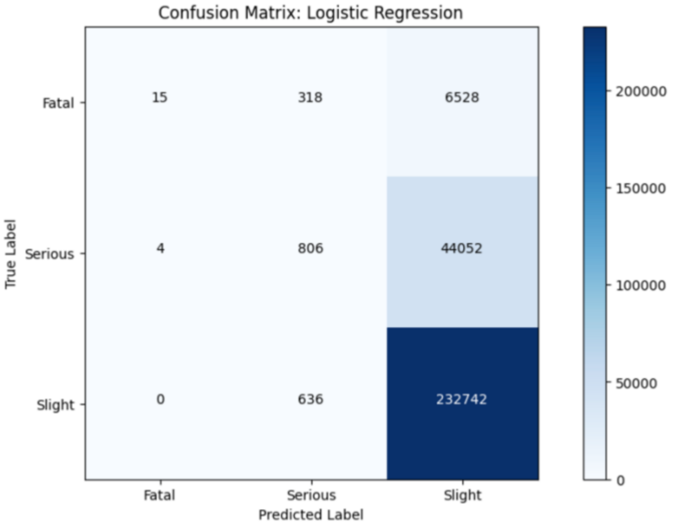
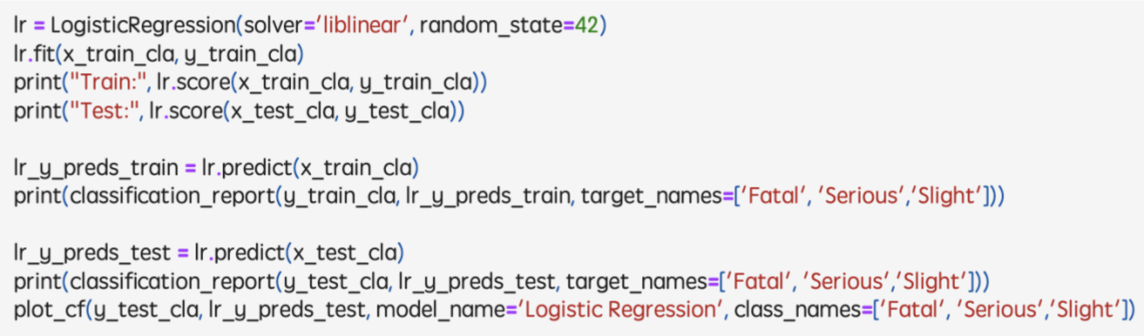


Figure 14: Confusion Matrix of Logistic Regression

Accuracy of Logistic Regression: 0.8192

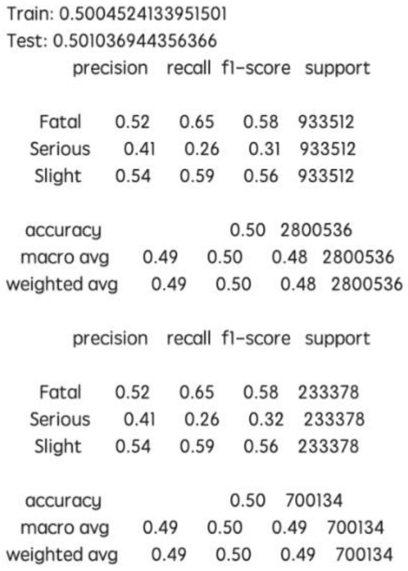
F1-Score of Logistic Regression: 0.00

From Confusion Matrix combined the visualization of distribution of accidents severity, this dataset is quite imbalanced. Also, although the accuracy of test set arrives 0.8193, the value of recall and f1-score is quite low, which indicates that this model is not enough good. Thus, we decided to resample the dataset and use **Random Over Sampling Methods** to handle data, which is to randomly select a small number of classes and replicate them several times, thus increasing the number of classes to balance the training sample set. Though the accuracy of LR model after resampling, the whole performance is better than original model. Also, we apply the resampling dataset in the rest of model.



```
lr = LogisticRegression(solver=’liblinear’, random_state=42)
lr.fit(x_train_cla, y_train_cla)
print("Train:", lr.score(x_train_cla, y_train_cla))
print("Test:", lr.score(x_test_cla, y_test_cla))

lr_y_preds_train = lr.predict(x_train_cla)
print(classification_report(y_train_cla, lr_y_preds_train, target_names=[’Fatal’, ’Serious’,’Slight’]))

lr_y_preds_test = lr.predict(x_test_cla)
print(classification_report(y_test_cla, lr_y_preds_test, target_names=[’Fatal’, ’Serious’,’Slight’]))
plot_cf(y_test_cla, lr_y_preds_test, model_name=’Logistic Regression’, class_names=[’Fatal’, ’Serious’,’Slight’])
```

Code 3: Construction of Logistic Regression after Resampling



```
Train: 0.5004524133951501
Test: 0.501036944356366
              precision    recall  f1-score   support

       Fatal       0.52      0.65      0.58    933512
     Serious       0.41      0.26      0.31    933512
      Slight       0.54      0.59      0.56    933512

    accuracy                           0.50   2800536
   macro avg       0.49      0.50      0.48   2800536
weighted avg       0.49      0.50      0.48   2800536

              precision    recall  f1-score   support

       Fatal       0.52      0.65      0.58    233378
     Serious       0.41      0.26      0.32    233378
      Slight       0.54      0.59      0.56    233378

    accuracy                           0.50    700134
   macro avg       0.49      0.50      0.49    700134
weighted avg       0.49      0.50      0.49    700134
```

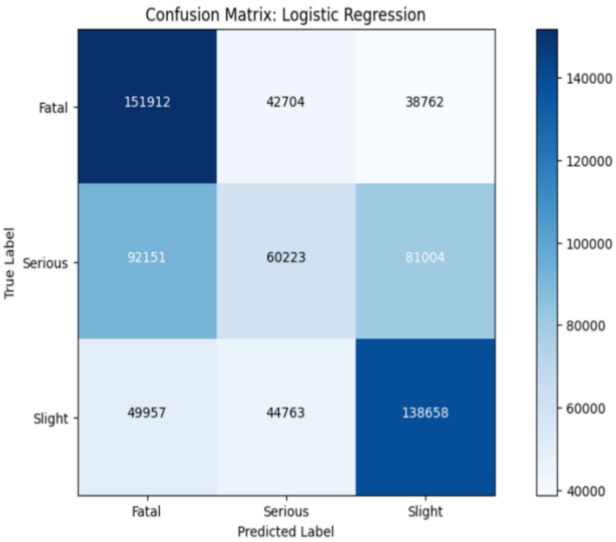


Figure 15: Classification Report of LR after resampling

Figure 16: Confusion Matrix of LR after resampling

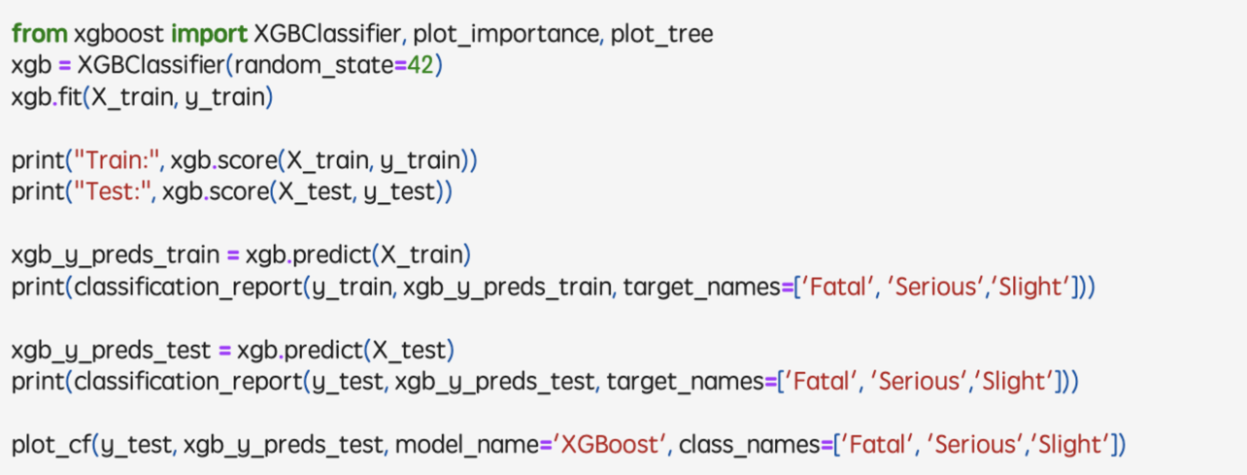
Accuracy of Logistic Regression model after resampling: 0.5010

F1-Score of Logistic Regression: 0.58

- XGBOOST

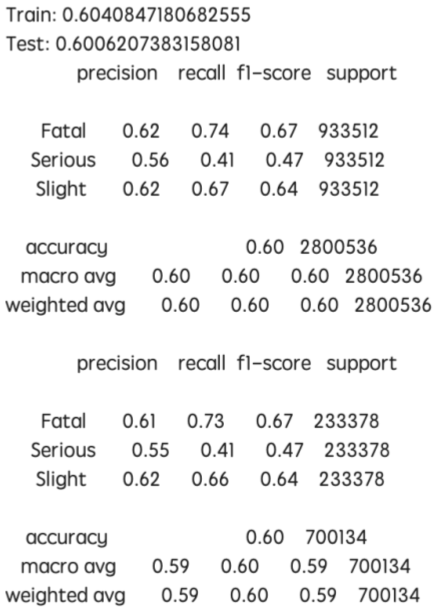
XGBOOST is a software package that allows users to easily solve classification, regression, or ranking problems. It implements the Gradient Boosting Tree (GBDT) model internally and optimizes the algorithms in the model to achieve high precision while maintaining extremely fast speed. The scalability, portability, and accuracy offered by XGBOOST push the upper limit of machine learning computing constraints.

Useful of XGBOOST in accident severity prediction model: The dataset we use is UK car accident data from the past 20 years, so the data size is very large. XGBOOST is efficient and scalable. It is fast and effective when processing large-scale data sets and does not require high hardware resources such as memory. In addition, XGBOOST internally implements a boosted tree model, which can automatically handle missing values.



```
from xgboost import XGBClassifier, plot_importance, plot_tree
xgb = XGBClassifier(random_state=42)
xgb.fit(X_train, y_train)

print("Train:", xgb.score(X_train, y_train))
print("Test:", xgb.score(X_test, y_test))

xgb_y_preds_train = xgb.predict(X_train)
print(classification_report(y_train, xgb_y_preds_train, target_names=['Fatal', 'Serious','Slight']))

xgb_y_preds_test = xgb.predict(X_test)
print(classification_report(y_test, xgb_y_preds_test, target_names=['Fatal', 'Serious','Slight']))

plot_cf(y_test, xgb_y_preds_test, model_name='XGBoost', class_names=['Fatal', 'Serious','Slight'])
```

Code 4: Construction of XGBoost



```
Train: 0.6040847180682555
Test: 0.6006207383158081
              precision    recall  f1-score   support

       Fatal       0.62      0.74      0.67    933512
     Serious       0.56      0.41      0.47    933512
      Slight       0.62      0.67      0.64    933512

    accuracy                           0.60   2800536
   macro avg       0.60      0.60      0.60   2800536
weighted avg       0.60      0.60      0.60   2800536

              precision    recall  f1-score   support

       Fatal       0.61      0.73      0.67    233378
     Serious       0.55      0.41      0.47    233378
      Slight       0.62      0.66      0.64    233378

    accuracy                           0.60    700134
   macro avg       0.59      0.60      0.59    700134
weighted avg       0.59      0.60      0.59    700134
```

Figure 17: Classification Report of XGBOOST after resampling

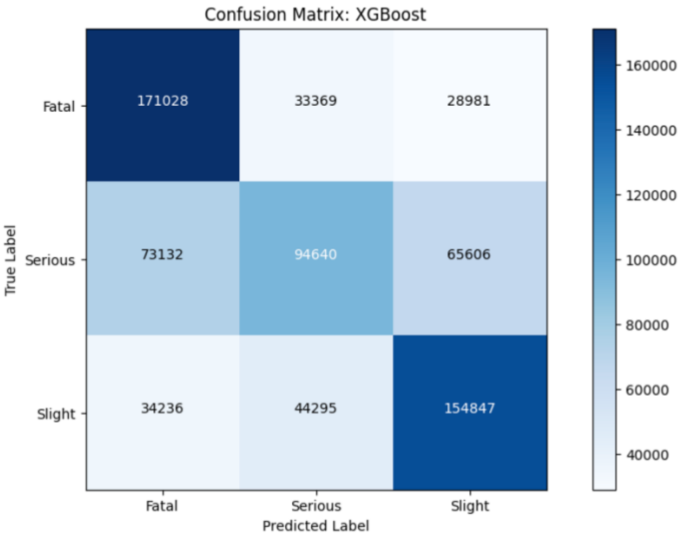


Figure 18: Confusion Matrix: XGBoost after resampling

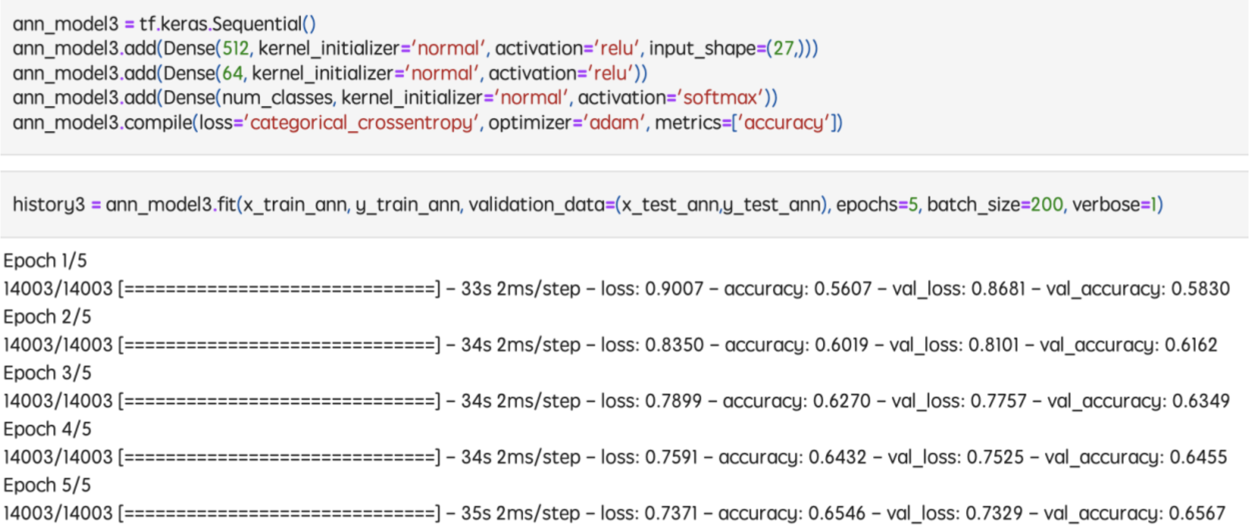
Accuracy of XGBOOST model after resampling: 0.6006

F1-Score of XGBOOST: 0.67

- Artificial Neural Network

Artificial neural network (ANN) models are often used to predict cases. By using an ANN model, these factors can be determined by collecting input data from key roads or highways. The input data can be processed by artificial neural network application software to obtain prediction results for road or highway prediction purposes. ANN application software can implement or analyze all parameters and accident data for future prediction.

Useful of Artificial Neural Network in accident severity prediction model: Artificial Neural Network can process in parallel and divided, so that many calculations can be performed to process data. With a self-learning function, it can input highly relevant data into the neural network to improve the accuracy of future predictions.

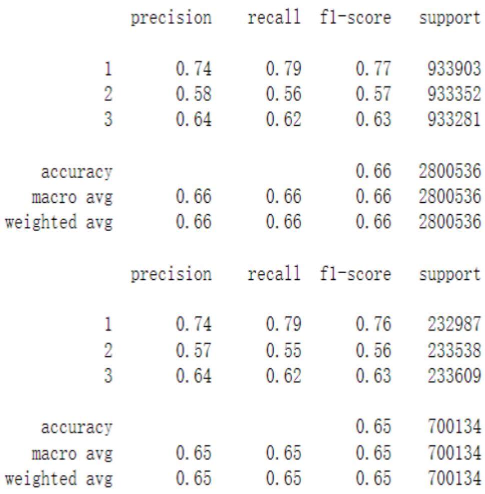


```
ann_model3 = tf.keras.Sequential()
ann_model3.add(Dense(512, kernel_initializer='normal', activation='relu', input_shape=(27,)))
ann_model3.add(Dense(64, kernel_initializer='normal', activation='relu'))
ann_model3.add(Dense(num_classes, kernel_initializer='normal', activation='softmax'))
ann_model3.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
history3 = ann_model3.fit(x_train_ann, y_train_ann, validation_data=(x_test_ann,y_test_ann), epochs=5, batch_size=200, verbose=1)
```

```
Epoch 1/5
14003/14003 [==============================] - 33s 2ms/step - loss: 0.9007 - accuracy: 0.5607 - val_loss: 0.8681 - val_accuracy: 0.5830
Epoch 2/5
14003/14003 [==============================] - 34s 2ms/step - loss: 0.8350 - accuracy: 0.6019 - val_loss: 0.8101 - val_accuracy: 0.6162
Epoch 3/5
14003/14003 [==============================] - 34s 2ms/step - loss: 0.7899 - accuracy: 0.6270 - val_loss: 0.7757 - val_accuracy: 0.6349
Epoch 4/5
14003/14003 [==============================] - 34s 2ms/step - loss: 0.7591 - accuracy: 0.6432 - val_loss: 0.7525 - val_accuracy: 0.6455
Epoch 5/5
14003/14003 [==============================] - 35s 2ms/step - loss: 0.7371 - accuracy: 0.6546 - val_loss: 0.7329 - val_accuracy: 0.6567
```

Code 5: Construction of ANN after resampling



```
              precision    recall  f1-score   support

           1       0.74      0.79      0.77    933903
           2       0.58      0.56      0.57    933352
           3       0.64      0.62      0.63    933281

    accuracy                           0.66   2800536
   macro avg       0.66      0.66      0.66   2800536
weighted avg       0.66      0.66      0.66   2800536
              precision    recall  f1-score   support

           1       0.74      0.79      0.76    232987
           2       0.57      0.55      0.56    233538
           3       0.64      0.62      0.63    233609

    accuracy                           0.65    700134
   macro avg       0.65      0.65      0.65    700134
weighted avg       0.65      0.65      0.65    700134
```

Figure 19: Classification Report of ANN after resampling

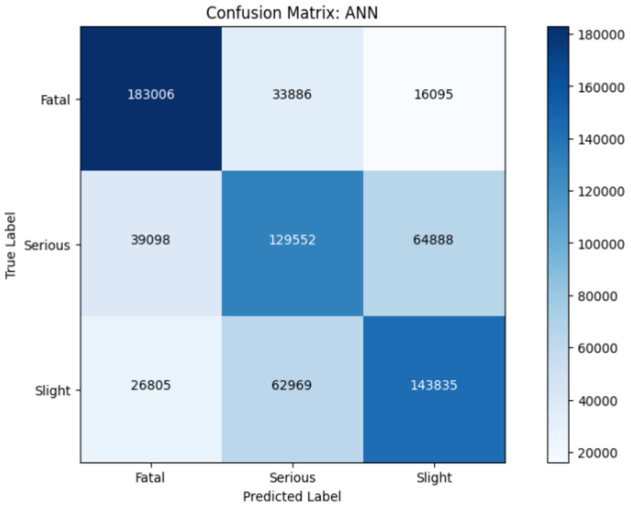


Figure 20: Confusion Matrix: ANN after resampling

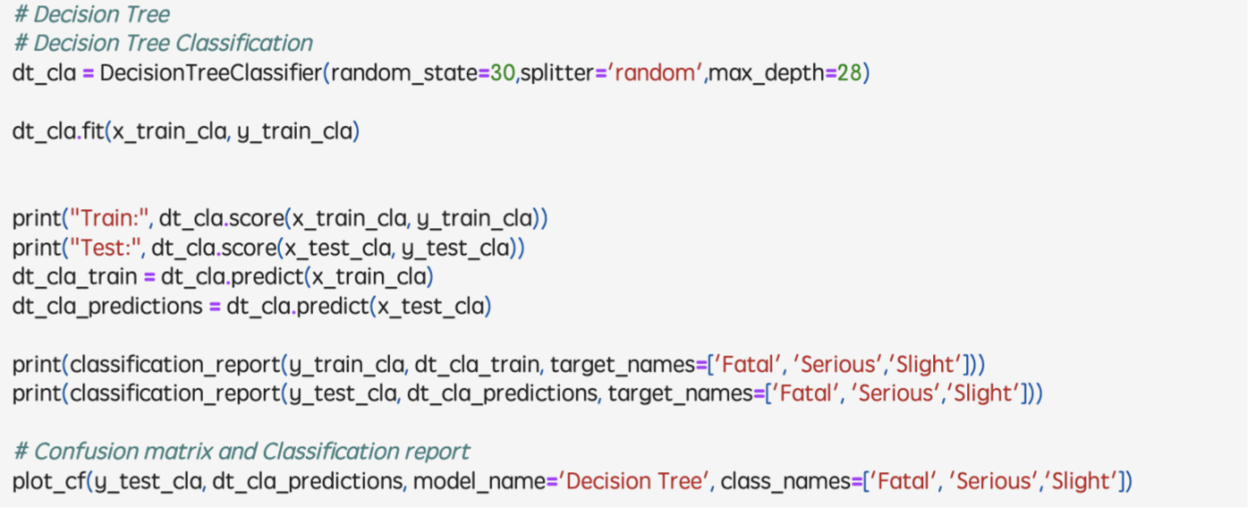
Accuracy of ANN model after resampling: 0.6567

F1-Score of ANN model after resampling: 0.76

- Decision Tree

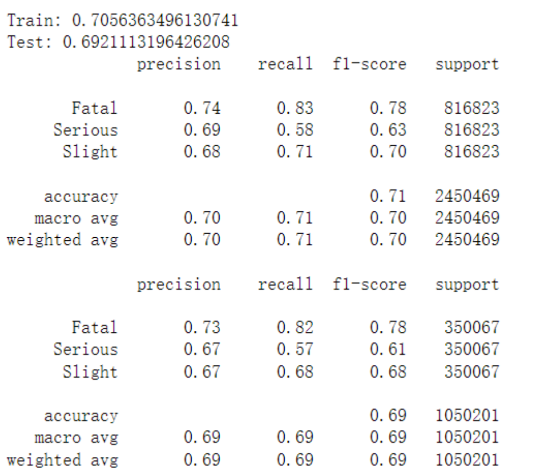
Decision trees are an example of a supervised machine learning method which can be used for both regression and classification problems. Decision trees operate by developing a set of decision rules based on the input predictor variables, which it will follow when make predictions about the target variable. With each row of predictor variables and corresponding target label that is passed to the decision tree, the model gradually improves and fine-tunes its set of decision rules. Decision trees are very versatile and can accept both categorical and numeric datatypes, continuous and ordinal. In addition to the robustness of the Decision Tree model, one of the benefits of using decision trees is they are quite intuitive to interpret, the most important attributes in the decision tree can be seen in at the top of the tree, as these will be the first decisions made on which branches the decision will follow.

Useful of Decision Tree in accident severity prediction model: Decision trees are easy to understand and interpret and can be analyzed visually. There are many model features, and the decision tree can handle irrelevant features to make the overall model more accurate. In addition, when testing the data set, the running speed is relatively fast, and it can make feasible and effective results for large data sources.



```
# Decision Tree
# Decision Tree Classification
dt_cla = DecisionTreeClassifier(random_state=30,splitter='random',max_depth=28)

dt_cla.fit(x_train_cla, y_train_cla)


print("Train:", dt_cla.score(x_train_cla, y_train_cla))
print("Test:", dt_cla.score(x_test_cla, y_test_cla))
dt_cla_train = dt_cla.predict(x_train_cla)
dt_cla_predictions = dt_cla.predict(x_test_cla)

print(classification_report(y_train_cla, dt_cla_train, target_names=['Fatal', 'Serious','Slight']))
print(classification_report(y_test_cla, dt_cla_predictions, target_names=['Fatal', 'Serious','Slight']))

# Confusion matrix and Classification report
plot_cf(y_test_cla, dt_cla_predictions, model_name='Decision Tree', class_names=['Fatal', 'Serious','Slight'])
```

Code 6: Construction of Decision Tree after resampling



```
Train: 0.7056363496130741
Test: 0.6921113196426208
              precision    recall  f1-score   support

       Fatal       0.74      0.83      0.78    816823
     Serious       0.69      0.58      0.63    816823
      Slight       0.68      0.71      0.70    816823

    accuracy                           0.71   2450469
   macro avg       0.70      0.71      0.70   2450469
weighted avg       0.70      0.71      0.70   2450469

              precision    recall  f1-score   support

       Fatal       0.73      0.82      0.78    350067
     Serious       0.67      0.57      0.61    350067
      Slight       0.67      0.68      0.68    350067

    accuracy                           0.69   1050201
   macro avg       0.69      0.69      0.69   1050201
weighted avg       0.69      0.69      0.69   1050201
```

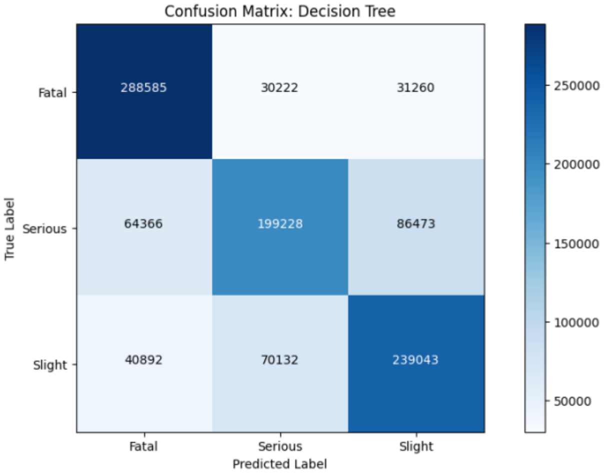


Figure 21: Classification Report of Decision tree after resampling

Figure 22: Confusion Matrix of Decision tree after resampling

Accuracy of Decision Tree model after resampling: 0.6921

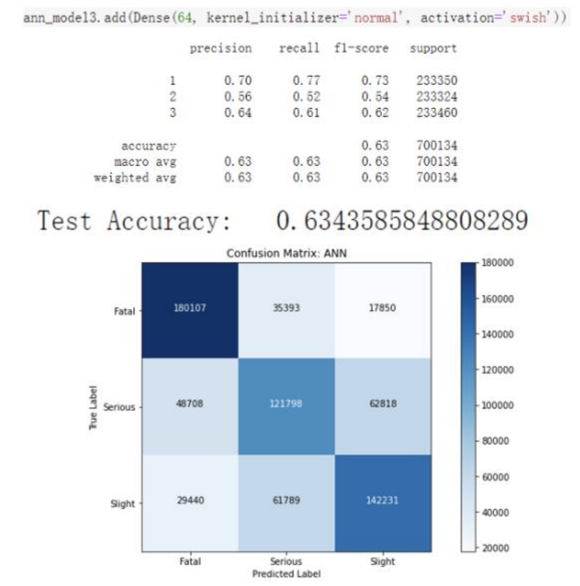
F1-Score of Decision Tree model after resampling: 0.78

5.4 Model Optimization

According to this accuracy, it can be seen that these models are all not enough good. Thus, we optimized two models ANN and Decision Tree.

- ANN Optimization

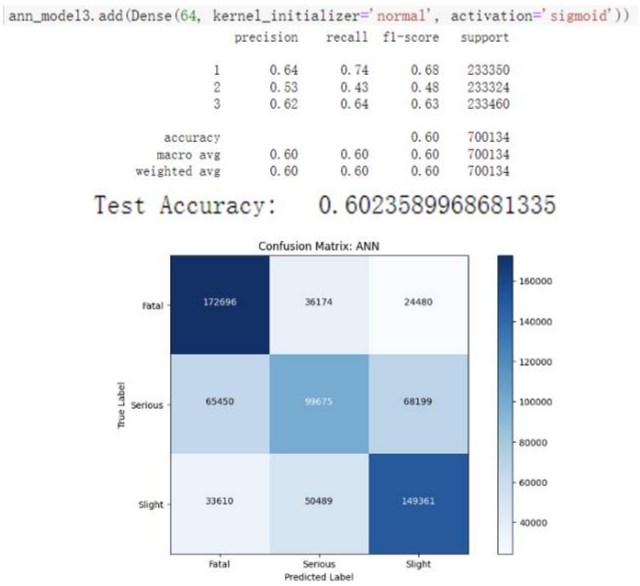
We use two methods to optimize our original method including changing the activation function and adding hidden layers and epochs Firstly, we converted Relu to swish and sigmoid and we found both lower than Relu's accuracy, with 63.43% and 60.24% respectively. Secondly, we increase



```
ann_model3.add(Dense(64, kernel_initializer='normal', activation='swish'))
```

```
              precision    recall  f1-score   support

           1       0.70      0.77      0.73    233350
           2       0.56      0.52      0.54    233324
           3       0.64      0.61      0.62    233460

    accuracy                           0.63    700134
   macro avg       0.63      0.63      0.63    700134
weighted avg       0.63      0.63      0.63    700134

Test Accuracy:  0.6343585848808289
```

Figure23: Confusion Matrix of ANN Optimization of Relu



```
ann_model3.add(Dense(64, kernel_initializer='normal', activation='sigmoid'))
```

```
              precision    recall  f1-score   support

           1       0.64      0.74      0.68    233350
           2       0.53      0.43      0.48    233324
           3       0.62      0.64      0.63    233460

    accuracy                           0.60    700134
   macro avg       0.60      0.60      0.60    700134
weighted avg       0.60      0.60      0.60    700134

Test Accuracy:  0.6023589968681335
```

Figure 24: Confusion Matrix of ANN Optimization of Sigmoid

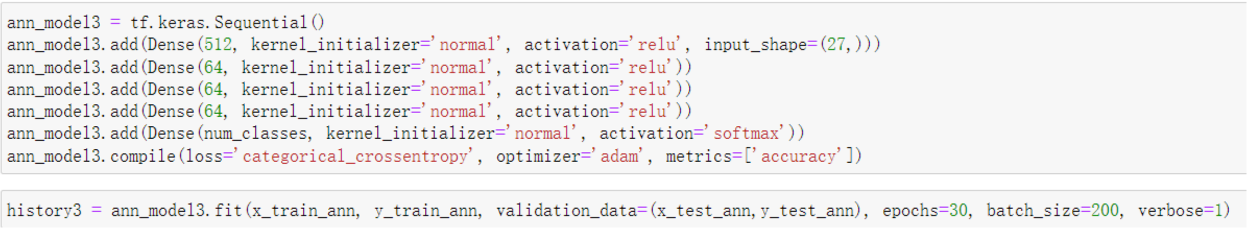
Accuracy of ANN of swish: 0.6343

F1-Score of ANN of swish: 0.73

Accuracy of ANN of sigmoid: 0.6024

F1-Score of ANN of sigmoid: 0.68

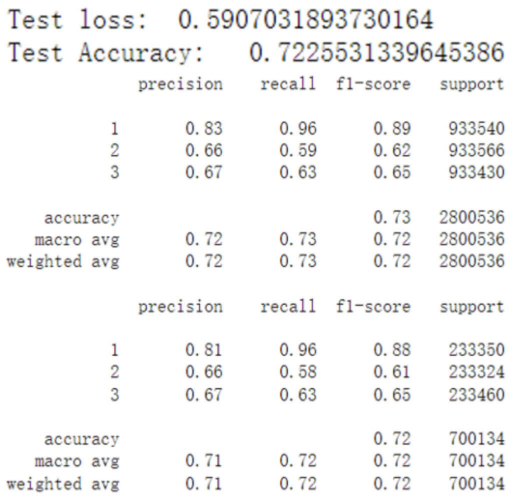
Secondly, we increased the hidden layer from 2 to 4 and changed epochs to 10. We found the accuracy with 72.26% is higher than the original model's accuracy.



```
ann_model3 = tf.keras.Sequential()
ann_model3.add(Dense(512, kernel_initializer='normal', activation='relu', input_shape=(27,)))
ann_model3.add(Dense(64, kernel_initializer='normal', activation='relu'))
ann_model3.add(Dense(64, kernel_initializer='normal', activation='relu'))
ann_model3.add(Dense(64, kernel_initializer='normal', activation='relu'))
ann_model3.add(Dense(num_classes, kernel_initializer='normal', activation='softmax'))
ann_model3.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
history3 = ann_model3.fit(x_train_ann, y_train_ann, validation_data=(x_test_ann,y_test_ann), epochs=30, batch_size=200, verbose=1)
```

Code 7: Construction of ANN after Optimization (hidden layer from 2 to 4 and changed epochs to 10)



```
Test loss:  0.5907031893730164
Test Accuracy:   0.7225531339645386
              precision    recall  f1-score   support

           1       0.83      0.96      0.89    933540
           2       0.66      0.59      0.62    933566
           3       0.67      0.63      0.65    933430

    accuracy                           0.73   2800536
   macro avg       0.72      0.73      0.72   2800536
weighted avg       0.72      0.73      0.72   2800536
              precision    recall  f1-score   support

           1       0.81      0.96      0.88    233350
           2       0.66      0.58      0.61    233324
           3       0.67      0.63      0.65    233460

    accuracy                           0.72    700134
   macro avg       0.71      0.72      0.72    700134
weighted avg       0.71      0.72      0.72    700134
```

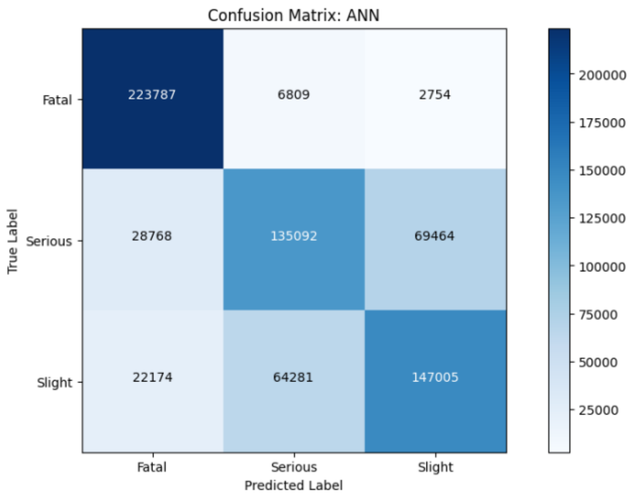


Figure25: Classification Report of ANN after Optimization

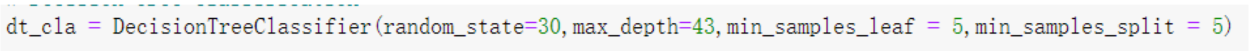
Figure26: Confusion Matrix of ANN after Optimization

Accuracy of ANN after Optimization: 0.7226

F1-Score of ANN after Optimization: 0.88

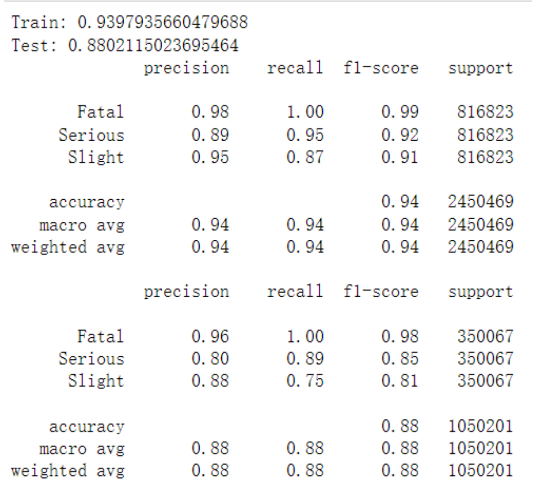
- Decision Tree Optimization

In this model, we change the Max-Depth, which can also be described as the length of the longest path from the tree root to a leaf. we increased the maximum depth of the decision tree and use the min_samples_leaf and min_samples_split functions to prune the decision tree to avoid overfitting. As expected, the accuracy of the decision tree optimized arrived 88.02%.



```
dt_cla = DecisionTreeClassifier(random_state=30,max_depth=43,min_samples_leaf = 5,min_samples_split = 5)
```

Code 8: Construction of Decision Tree after Optimization (increased the maximum depth, use the min_samples_leaf and min_samples_split)



```
Train: 0.9397935660479688
Test: 0.8802115023695464
              precision    recall  f1-score   support

       Fatal       0.98      1.00      0.99    816823
     Serious       0.89      0.95      0.92    816823
      Slight       0.95      0.87      0.91    816823

    accuracy                           0.94   2450469
   macro avg       0.94      0.94      0.94   2450469
weighted avg       0.94      0.94      0.94   2450469
              precision    recall  f1-score   support

       Fatal       0.96      1.00      0.98    350067
     Serious       0.80      0.89      0.85    350067
      Slight       0.88      0.75      0.81    350067

    accuracy                           0.88   1050201
   macro avg       0.88      0.88      0.88   1050201
weighted avg       0.88      0.88      0.88   1050201
```

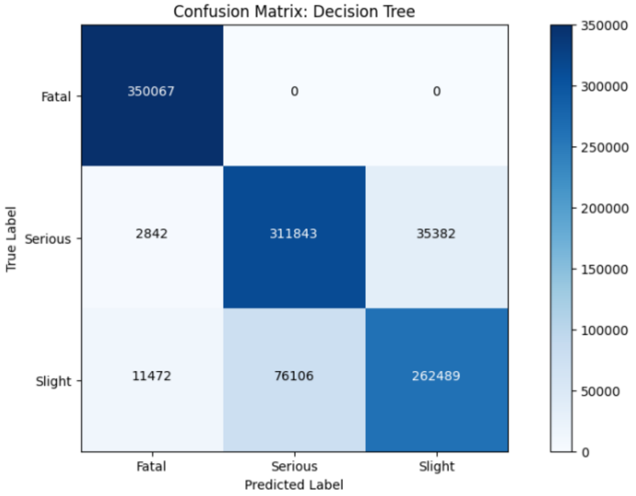


Figure27: Classification Report of Decision Tree after Optimization

Figure 28: Confusion Matrix of Decision Tree after Optimization

Accuracy of Decision Tree after Optimization: 0.8802

F1-Score of Decision Tree after Optimization: 0.98

6. Model Evaluation & Comparison

6.1. Model Evaluation

After preparing all the datasets and performing the required preprocessing and data modeling steps, the predictive model was evaluated, and the prediction results were recorded. The results of the machine learning classification analysis used to predict casualties in UK road crashes are summarized in the table below.

Model	Training Data	Accuracy	F1-Score
LR (Before resampling)	Imbalanced	81.92%	0
LR (After resampling)	Balanced	50.10%	0.58
XGBOOST	Balanced	60.06%	0.67
Decision Tree	Balanced	69.21%	0.78
Decision Tree (Max-Depth)	Balanced	88.02%	0.98
ANN-Relu	Balanced	65.57%	0.76
ANN-Swish	Balanced	63.44%	0.73
ANN-Sigmoid	Balanced	60.24%	0.68
ANN-Relu (with 4 hidden layers & 30 epochs	Balanced	72.26%	0.88

Table2 Model Prediction Accuracy

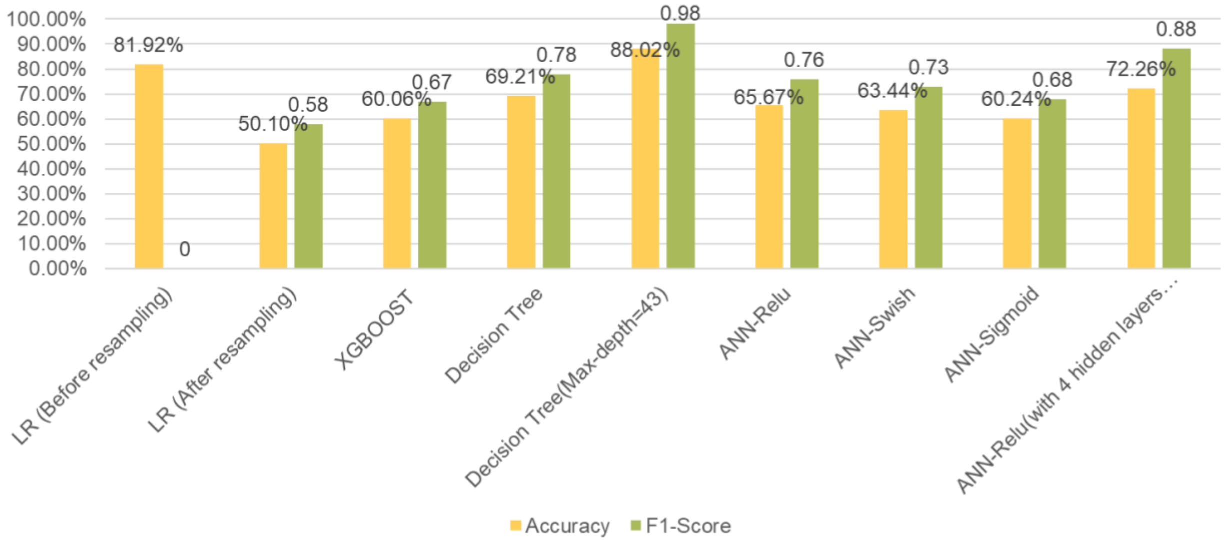


Figure29: Bar chart of Model Prediction Accuracy

6.2. Comparison with other models by various researchers

Comparing the optimal models obtained by various researchers, we found that some researchers did not take into account the non-equilibrium of the data, so they all obtained high accuracy rates. However, machine learning models are prone to bias the majority class in dealing with non-equilibrium, and are obsessed with high test set accuracy, thus neglecting the recall rate, which leads to the reduction of the F1 value. This can lead to models that appear to be highly accurate but are simply unusable.

Experiments comparing classification predictions using a balanced dataset were observed. The results were generally around 80%, while we ended up with an accuracy of 88%, which is one of the more excellent classification prediction results.

Researcher	Optimal model	Accuracy (%)	Training Data
This project	Decision Tree	88.02	Balanced
Nour, M. K. [3]	XGBoost	74.83	Balanced
Arnav Saini [5]	SVM	87.73	Imbalanced
Yun [6]	Random Forest	85.29	Balanced
Siew Lee Koh [7]	Random Forest with only numerical predictors	83.47	Imbalanced

Table3 Comparison of the optimal models of each researcher

In conclusion, from the Table 2 Model Prediction Accuracy, we can clearly see that the classification model with the highest accuracy rate is the decision tree model (optimize Max-Depth) with an accuracy rate of 88.02%, which can be attributed to the nature of the modelling task and the data used. Most attributes have category values where decision trees-based methods are reported to outperform regression-based methods.

Secondly, by comparing the results of other researchers as well as our own initial LR models using unbalanced data sets. We can observe data imbalance can lead to enrichment of classification predictions in the majority class- slight accidents. Because of enrichment, it is not enough to evaluate model only regarding accuracy but combining recall and f1-score. Thus, resampling is a necessary step.

7. Conclusion & Improvement

7.1. Conclusion

In recent years, the traffic safety situation in the whole world has been relatively severe. Road safety problems have become a common social problem, which has a major impact on people's lives and safety, and will also cause property losses, which play a role in social and economic development. How to quickly judge the severity of traffic accidents and how to scientifically prevent serious traffic occurrences has become the focus of traffic safety research.

This paper uses the British traffic accident data for analysis. Since the data consists of multiple files, this paper first integrates the data and selects variables that may be required for modeling from multiple aspects and performs missing value processing for some missing cases in the data. The data were then subjected to descriptive statistics and exploratory analyzes which proved to be a very interesting and challenging project. An exploratory analysis of the data provided insights into the field of road accidents occurring in the UK. Then use the correlation coefficient matrix and random forest to screen the variables, and finally select the features with a cumulative contribution rate of the top 70% for subsequent modeling. By analyzing the LR model, we found that the data was seriously unbalanced, so we used random sampling to balance the data set. Then use logistic regression, XGBoost, neural network and decision tree models to model the balanced data set, and then get the two optimal models, and then perform model optimization on these two models, that is, parameter adjustment, and finally get the best performance model. Finally, it is analyzed and compared with the research results of various researchers.

Through the analysis, it is found that whether there are data such as the number of victims, the area, and the age of the vehicle will help the traffic law enforcement department to quickly determine the severity of the accident; at the same time, the traffic law enforcement department should strengthen the education and management of young and old drivers Measure the specifications of motor vehicles on the road, pay attention to the form and direction of vehicles, etc.

7.2. Limitation

Models are often difficult to customize, meaning it is difficult to change their parameters or adapt them to new datasets. Our data is downloaded from the official website of the UK.

Models are often expensive to train, which means building an accurate model can take a lot of time. It took us a lot of time to preprocess the data and then run the model to get the results.

Models cannot explain unstructured data. We delete several data with text explanation.

With the major technological advancements in the field of self-driving cars in recent years, fully autonomous vehicles have become ubiquitous in the world. And as the amount of telematics data collected and stored increases, the safety and impact of fully autonomous vehicles on road issues will be a very important topic and a topic of great interest to analysts. Accident predictions for fully autonomous vehicles will also be rebuilt.

7.3. Improvement

Another factor stress that models developed were only basic models that will be good starting points for further studies. Improvements can be made by incorporating new explanatory variables that are related to the influence of human characteristics and behavior on accident cause.

It is hoped that this accident prediction model, available to any user in the future, can assist them in providing safety precautions and future work on safety issues, especially relevant agencies, and future road users.

By capturing the underlying risk distribution that determines the probability of future crashes everywhere, without any historical data, we can find safer routes and help city planners design safer roads and even predict future crashes.

Reference

[1] World Health Organization. (n.d.). Global status report on road safety 2018. World Health Organization. Retrieved October 20, 2022, from https://www.who.int/publications/i/item/9789241565684

[2] World Health Organization. (n.d.). Road traffic injuries. World Health Organization. Retrieved October 20, 2022, from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[3] Nour, M. K., Naseer, A., Alkazemi, B., & Jamil, M. A. (2020). Road traffic accidents injury data analytics. International Journal of Advanced Computer Science and Applications, 11(12).

[4] Feng, M., Zheng, J., Ren, J., & Liu, Y. (2020, February). Towards big data analytics and mining for uk traffic accident analysis, visualization & prediction. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing (pp. 225-229).

[5] Saini, A., Gauba, N., Chawla, H., & Ali, J. (2021). Road Accidents Analysis Using Comparative Study & Application of Machine Learning Algorithms. Journal: WSEAS TRANSACTIONS ON COMPUTER RESEARCH, 78-86.

[6] Yun,T.X.(2021). Learning Based on Road Traffic Accident Severity Analysis and Prediction. Retrieved November 9, 2022, from https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202202&filename=1022406379.nh

[7] Koh, S.L., Katannya k., Roger T.K.E., Meraldo A., & Yeok K.C. (2019), Analysis and Prediction of Traffic Accidents in London, Retrieved November 9, 2020, from https://github.com/katannyak/UK-Traffic-Accident-Analysis-and-Visualization/blob/master/docs/Final_report_Team47.pdf

[8] Sarkar, A., & Sarkar, S. (2020). Comparative assessment between statistical and soft computing methods for accident severity classification. Journal of The Institution of Engineers (India): Series A, 101(1), 27-40.

[9] Asare, I. O., & Mensah, A. C. (2020). Crash severity modelling using ordinal logistic regression approach. International Journal of Injury Control and Safety Promotion, 27(4), 412-419.

[10] Kononen, D. W., Flannagan, C. A., & Wang, S. C. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Analysis & Prevention, 43(1), 112-122.

[11] Road Safety Data. (n.d.). Retrieved October 20, 2022, from https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[12] R. Foundation, “https://www.racfoundation.org/motoring-faqs/mobility#a26.”