**Method**

Predicting the severity of road traffic accidents is a multiclass classification problem , that is we are seeking to classify Severity on a rating scale of 1 – 4 to indicate the duration of disruption likely to occur as the result of an incident.

Modelling is performed using Python (version 3.9.13) in Jupyter Notebooks (version 6.4.12)

To model Severity we have used a subset of features of the dataset containing both numerical and categorical variables. We reduced the number of features in our dataset removing those features that would not be useful to the model as well as those features that were a proxy or categorisation of another feature that already existed within the dataset.

Predictor variables included Start\_Lat, Start\_Lng, Distance, Temperature, Wind\_Chill, Humidity, Pressure, Visibility, Wind\_Speed, Precipitation, Amenity, Bump, Crossing, Give\_Way, Junction, No\_Exit, Railway, Roundabout, Station, Stop, Traffic\_Calming, Traffic\_Signal, road\_type, weekday, month, and year.

Predicting Severity is a multiclass classification problem therefore we have evaluated the performance of K-Nearest Neighbours, Random Forest and Gradient Boosting which are algorithms capable of performing multiclass classification.

It was necessary to normalise numeric features in the dataset prior to dividing the dataset into a testing and training set. We reserved 30% of the data for testing and the remaining 70% was utilised for training of each model. The training and test sets each included 25,754 and 11,034 observations respectively.

SMOTETomek from the Imbalanced-learn package (version 0.9.1) was used to address class imbalance in the training data. This approach was chosen to address the significant disparity in the size of majority and minority classes. The SMOTETomek package combines the SMOTE oversampling technique to generate synthetic examples of the minority class and then applies the Tomek Links technique to under sample the same class to get better separation between class clusters.

The resulting training set included 79,754 rows of data with data distributed similarly distributed across each class of Severity.

Table X. Observations by Severity in Resampled Training Data

Graphical user interface, application

Description automatically generated

**K-Nearest Neighbours**

K-Nearest Neighbours is a supervised machine learning algorithm capable of performing multiclass classification. K-Nearest Neighbours works on a principle assuming every data point falling in near to each other is belongs to the same class. It classifies each new data point based on its similarity to it’s nearest neighbours forming clusters of data points belonging to the same class. The K-Nearest Neighbours classifier can accept both numerical and categorical predictor variables and therefore was considered to be appropriate for the task of classifying Severity and provide a point of comparison for the performance of Random Forests and Gradient Boosting.

K-Nearest Neighbours was performed using the KneighborsClassifier from the scikit-learn package (version 1.1.1) . RandomizedSearchCV from the Scikit-learn package was used to perform 5-fold cross validation to identify the optimal value of K (nearest neighbours). Given the size of the dataset and the computational demand associated with cross-validation of each different combination of hyperparameters a random search was used for model tuning due to its advantages in processing time when compared with a grid search.

The Weighted F1 score was used for model selection.

The training set was used to train the K-Nearest Neighbours classifier. We then assessed the accuracy of our model using our holdout test set and evaluate the accuracy of the model in classifying Severity.

K-Nearest Neighbours performed best where K = 2. The cross-validated K-Nearest Neighbours classifier resulted in a classification accuracy of 84.21% and weighted F1-score of 84.74%.

**Random Forest**

Random Forests are capable of producing highly accurate classifiers whilst also reducing the risk of overfitting making it suitable for the task of predicting Severity. Random Forest was performed using the Scikit-learn package in Python. RandomizedSearchCV was used to perform 2-fold cross validation.

It was identified that a model using 1500 estimators, a maximum tree depth of 60, a minimum of 5 samples per split and 1 sample per leaf resulted in the most accurate Random Forest classifier.

The resulting model correctly classified Severity 87.61% of the time and reported a weighted F1-score of 87.95%.

**Gradient Boosting**

Gradient Boosting uses an ensemble of decision trees to predict Severity. The primary difference between Gradient Boosting and Random Forests lies in how the decision trees are created and aggregated. In Gradient Boosting, decision trees are built additively and sequentially where each new tree tries to predict the error left over by the previous model whereas Random Forests train each tree independently and at random.

Gradient Boosting was performed using the XGBoost package (version 1.5.0) in Python. RandomizedSearchCV was used to perform 2-fold cross validation.

It was identified that a model using learning rate of 0.1, maximum depth of 100, minimum child weight of 2, gamma of 0.0 and column sample by tree of 0.5 resulted in the most accurate Gradient Boosting classifier.

The cross-validated Gradient Boosting classifier resulted in a classification accuracy of 88.50% and weighted F1-score of 88.22%.

**Results and Discussion**

Disruption to traffic flow results in real costs to communities and economies. Disruption caused by traffic accidents can lead to real costs to individuals and families, for example when a parent may be late to collect a child from care, or loss of income where an individual is late to work. Businesses also incur additional costs caused from supply chain and workforce disruptions.

It is for this reason that understanding factors that influence severity of traffic flow disruption has substantial benefit to responsible stakeholders including planners, emergency services, traffic management agencies and policy makers.

In evaluating the model, we must consider the accuracy of the model and its performance in correctly classifying the severity of a traffic incident. The result of each model is summarised in Table X with Gradient Boosting (88.50%) being the most accurate classifier and K-Nearest Neighbours being the least accurate (84.21%).

Table X. Summary of Classifier Performance

Graphical user interface, text

Description automatically generated

Gini importance was used to determine the features of most importance to the Gradient Boosting classifier as the best performing model. The results are shown in Figure X and identify the month the incident occurred, location of the incident, length of traffic congestion, select weather conditions and day of week were most important to the Random Forest model in predicting Severity.

Figure X. Feature importance to the Gradient Boosting model

Chart

Description automatically generated

Interestingly, points of interest and traffic features played very little role in determining severity. Despite this we should not discount the relationship between traffic network design and incident severity noting the dataset does not contain information as to whether these features have contributed to an incident. Further, their lack of contributions to incidents may in fact demonstrate that certain traffic calming measures contribute to a reduced likelihood of incidents occurring.

Modelling enables us to identify opportunities to have the greatest impact to reduce incident severity and deploy resources efficiently.

In practise, predictive modelling performed with a Random Forest classifier can assist agencies and policymakers to plan the allocation of resources to incident response based on the time of year, select weather conditions and proximity to locations of high severity incidents. Whilst there may be a tendency to locate response teams near locations that have a high frequency of incidents, the use of predictive modelling enables agencies to better deploy resources to different locations based on time of week and year and forecast weather conditions with a view to respond quickly to reduce high severity traffic incidents specifically.

It's important that we also consider the practical cost implications to agencies in taking action or failing to take action based on the predictions of the model. There is a real associated cost with reducing the severity of traffic incidents both directly in the deployment of resources to respond to incidents, and indirectly with the trickle-down impact of disruptions to communities and businesses.

A strategy of elimination of traffic incidents is unrealistic and therefore effectively mitigating the occurrence of incidents through road network design and maintenance programs can be balanced with resourcing and location of incident response teams to have the greatest impact on reducing economic cost associated with traffic flow disruption.