Predicting Vehicle Accident Severity

Stuart Watt

16/04/2020

Contents

1 Introduction

Vehicle accidents are sadly a common source of physical trauma in everyday life. Further motivated the financial burden of these events to individuals and greater society, reducing their frequency and severity has been consistently placed at a high priority both in the public and private sectors. As a consequence, great volumes of data about this domain has been collected over several decades. A well known example is the UK government road accident and safety datasets, which are published every year since 1979. Here, details about road accidents known to UK police that involve personal injury are recorded, albeit with those occurring on private roads or car parks not included.

For recent years, details about accidents are recorded in a set of 3 consistent relational tables with every accident specified by an identification index. The first contains information about the accident itself, such as location, time of incident, number of vehicles and road conditions. In the second table information about involved vehicles for each accident is recorded, such as types, engine displacements and points of impact. The third table contains casualty information, such as the age, sex, and injuries sustained for each individual and event. Of interest is a classification of injury severity, with “Slight”, “Severe” and “Fatal” levels. From this, an *accident severity* is defined. Slight accident severity is defined as a personal injury accident where all injuries are minor, such as whiplash, bruising and strains. Meanwhile, fatal accidents are when one or more individuals are killed though their sustained injury, either immediately or within 30 days of the accident. Lastly, severe accidents occur when none injured persons succumb to their injuries but at least one suffered an injury of greater severity such as fractured bones, deep cuts and unconsciousness. It is noted that there is a reported selection bias in collection of the data, as many minor accidents are known to not be reported to the police.

In this project an attempt predicting accident severity using several other features is made. Information of all accidents recorded from 01/01/2016 to 31/12/2018 from the aforementioned data source was downloaded and imported. After features were selected, exploratory data analysis was performed along with further feature engineering and selection before a final dataset was obtained. Then data from 2016 and 2017 was used in conjunction with various algorithms to train a variety of machine learning models. Finally, performance was tested and analysed using 2018 data.

2 Data Import & Cleaning

Data from variables of interest was loaded and categories recoded according to their descriptions in the data handbook. It is worth noting that some factors described in the handbook did not occur in the dataset.

The following features were initially extracted: accident severity (the class feature), number of vehicles, number of casualties, date, day of week, time of day, road type, speed limit, lighting conditions, weather conditions, road surface conditions, area type and finally the average age of casualty. All variables were taken from the accidents table apart from the average age of casualty. This was obtained from the casualty table through performing a grouped average by accident index and then joining with the existing dataset. Categories were recoded from numeric values with reference to the variable lookup document provided from the data source.

3 Exploratory Data Analysis & Feature Selection

The purpose of the exploratory data analysis was to investigate any dataset issues and to see if any insight could be gained about the dataset. Particular attention was placed on relationships between the class feature with the others.

3.1 Overview

The following table provides an overview of the initial dataset.

Table 1: Number of rows for each accident severity category.

Accident_Severity	n
Slight	316772
Serious	67424
Fatal	5042

First is to note that there are (thankfully) more observations with the lesser severity, with a small portion of the dataset containing fatal accidents. Below is a list of categories of categorical variables.

```r
accidents %>% select_if(is.factor) %>% map(~levels(.))
```

```
## $Accident_Severity
## [1] "Slight"  "Serious" "Fatal"
##
## $Day_of_Week
## [1] "Tuesday"   "Wednesday" "Thursday"  "Friday"    "Saturday"  "Sunday"
## [7] "Monday"
##
## $Road_Type
## [1] "Single carriageway" "Roundabout"         "Dual carriageway"
## [4] "One way street"     "Slip road"          "Unknown"
##
## $Light_Conditions
## [1] "Darkness - lights unlit"     "Darkness - lights lit"
## [3] "Daylight"                    "Darkness - no lighting"
## [5] "Darkness - lighting unknown"
##
## $Weather_Conditions
## [1] "Fine no high winds"    "Raining no high winds" "Fog or mist"
## [4] "Other"                 "Unknown"               "Fine + high winds"
## [7] "Raining + high winds"  "Snowing no high winds" "Snowing + high winds"
##
## $Road_Surface_Conditions
## [1] "Dry"                  "Wet or damp"          "Flood over 3cm. deep"
## [4] "Frost or ice"         "Snow"
##
## $Urban_or_Rural_Area
## [1] "Urban"       "Rural"       "Unallocated"
```

3.2 Missing values

Both road type and weather conditions have an "unknown" category. This is an explicit option in the vehicle accident reporting form. These were converted to NA values. There are also "other" and "unallocated" levels in the weather conditions and area type variables, respectively. I decided not to convert these to NA values at this stage, since these provide information about what the respective value isn't.

Also, in the lighting conditions variable the "darkness-lighting unknown" category seems redundant. Information already is contained through other variables about whether an accident occurred at night or not. For this reason, this category was replaced with NA, too.

A count of missing values for each column was performed as below.

Table 2: A count of missing values for each feature containing at least one.

Feature	Missing values
Weather_Conditions	12102
Light_Conditions	7534
Avg_Age_of_Casualty	7114
Road_Type	6344
Road_Surface_Conditions	3929
Speed_limit	37
Time	18
Urban_or_Rural_Area	1

Proportions of missing values by class category is shown below.

Table 3: Missing values broken down by class category.

Accident_Severity	Total_Observations	Total_Missing	Percent_Missing
Slight	316772	25557	8.067948
Serious	67424	3681	5.459480
Fatal	5042	229	4.541849

Observations containing missing values make up a small fraction of the dataset (even more so with less frequent class variable categories). Given the large quantity of data, I decided to remove them. Also, categories with no instances are removed.

3.3 Variation & Composition

The distribution of each feature alongside the composition of the class variable is investigated. Also the dataset is adjusted incrementally.

3.3.1 Number of Vehicles

The distribution of number of vehicles was investigated.

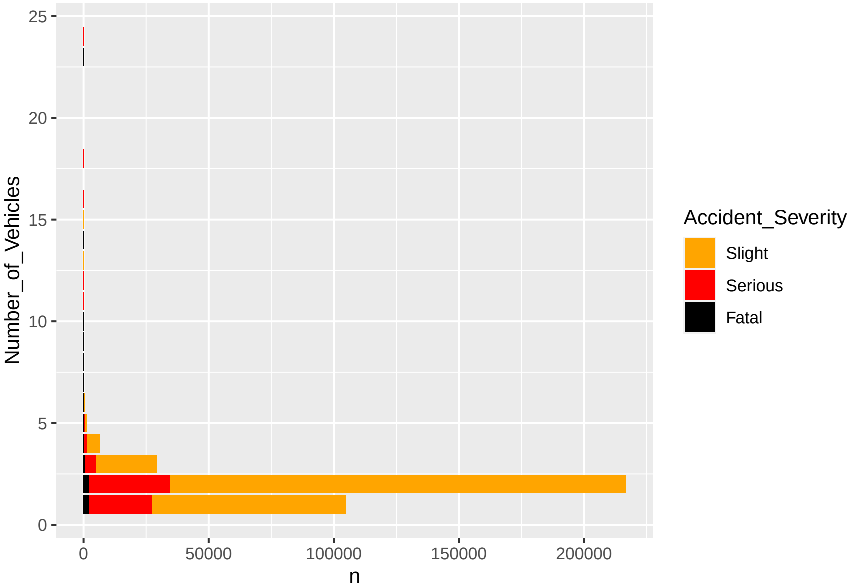


Figure 1: Number of observations for differing number of vehicles.

It is evident that the vast majority of accidents involve two vehicles or less, while several accidents have greater than 10 vehicles involved. To visualise the proportion of of serious and fatal accidents with the number of vehicles, the bars are scaled evenly in the following figure:

There does not seem to be much relationship for number of vehicles less than four, which constitutes the majority of the dataset.

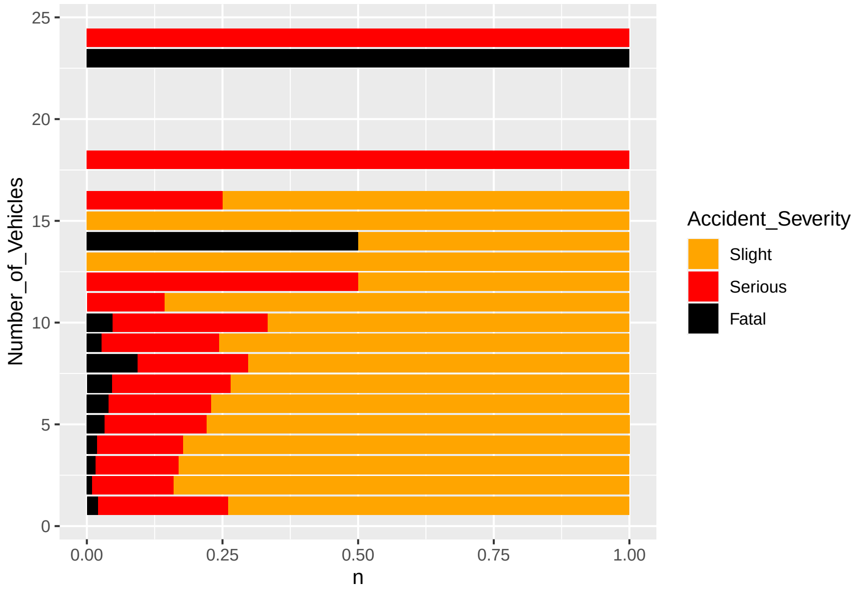


Figure 2: Severity proportions for differing number of vehicles.

3.3.2 Number of Casualties

A similar investigation to the distribution of number of casualties is performed as with the number of vehicles:

We note that the minimum number of casualties observed is 1, which also happens to be most frequent. No accidents record 0 casualties because these incidents are not defined as being an accident according to the data source definition. The distribution decays quickly, with a few outlier observations with greater than 20 casualties.

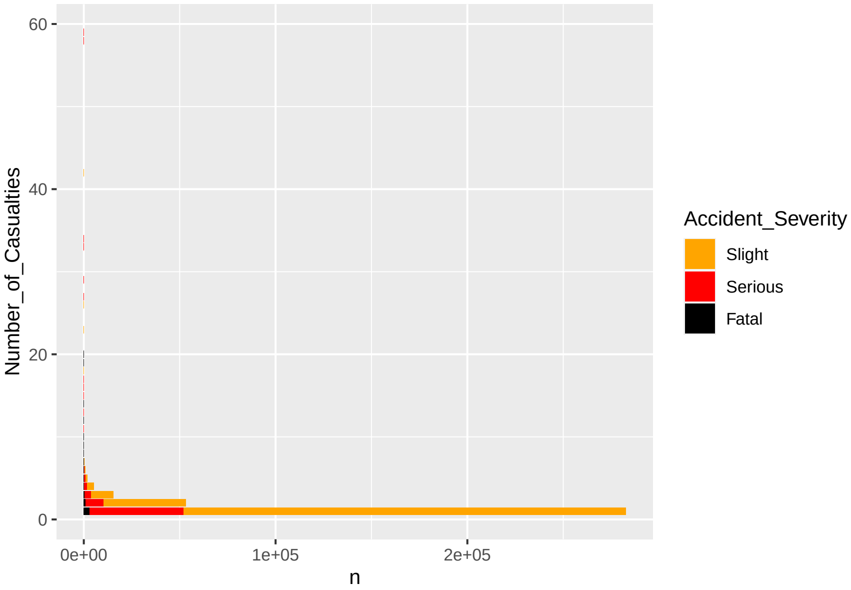


Figure 3: Number of observations for differing number of casualties.

We similarly look at the proportions of each bar:

There seems to be an relationship between the number of casualties and the accident severity. This makes sense, since the larger the number of people injured the greater the probability that one injury will be severe or lead to a fatality. Unfortunately no information about the number of passengers of vehicles is given in the dataset.

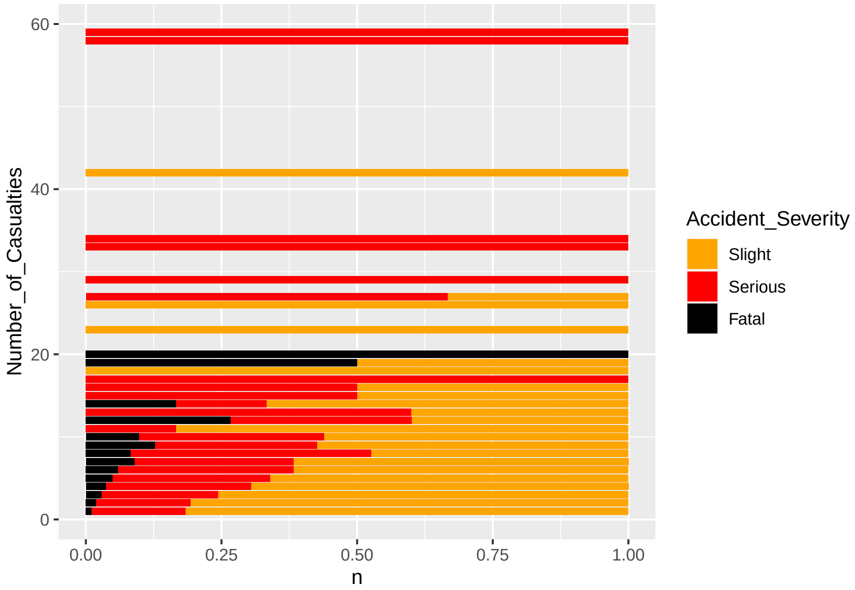
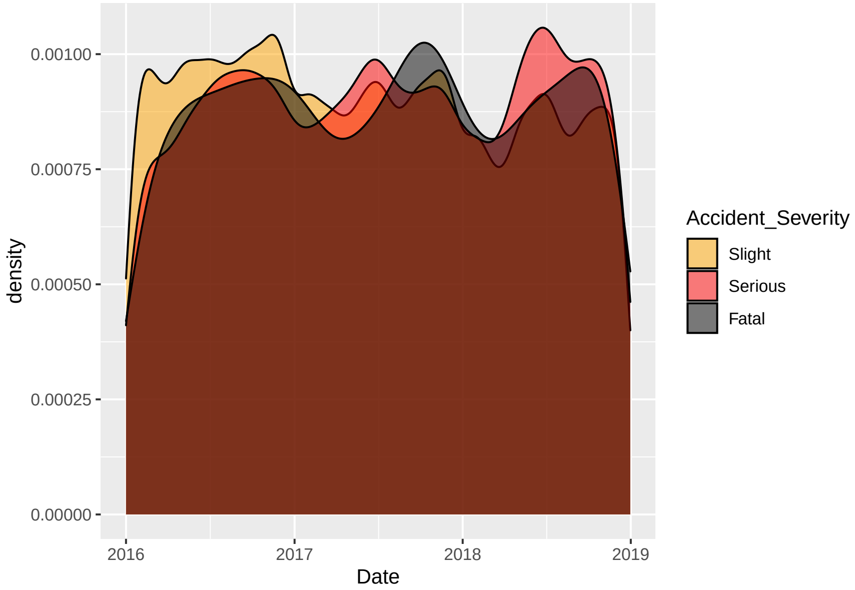


Figure 4: Severity proportions for differing number of casualties.

3.3.3 Time and Date

There are several ways to view the temporal aspect of this dataset. One option is to look at the distribution as a whole.



These distributions do not seem to reveal any clear pattern. Distributions by time and day of week however show some insights:

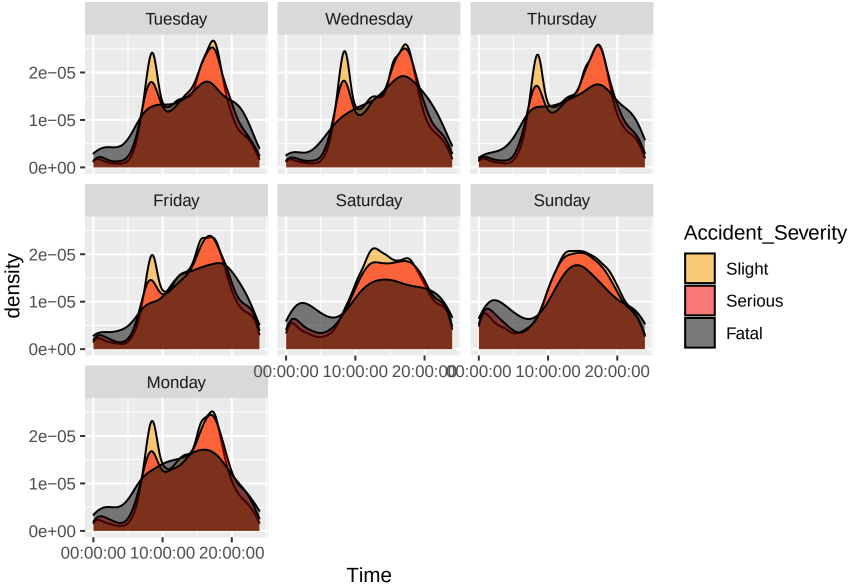


Figure 5: accident severity distributions by time and day of week.

It is evident that the distribution of greater severity accidents is more time invariant. We see that fatal accidents occur with greater relative frequencies in hours close to midnight. However, this may be because of lighting conditions.

We see that overall the slight and serious density curves are similar. Note that peaks occur from Monday to Friday at around 9am and 5pm: this is likely due to greater traffic from people commuting to and from work. In the weekend there is a greater likelihood of accidents early in the morning and late at night. The similarity of weekday and weekend daily distributions motivates me to replace the day of week column with a boolean “Weekend” column.

The problem with encoding the time of day as a simple numerical number is that it does not convey information about cyclical nature. If one were to encode time as the number of seconds past midnight, then 23:59:59 would be maximally different from 00:00:00, even though they are practically speaking very similar. One way to represent this cyclical nature would be through sinusoidal functions:

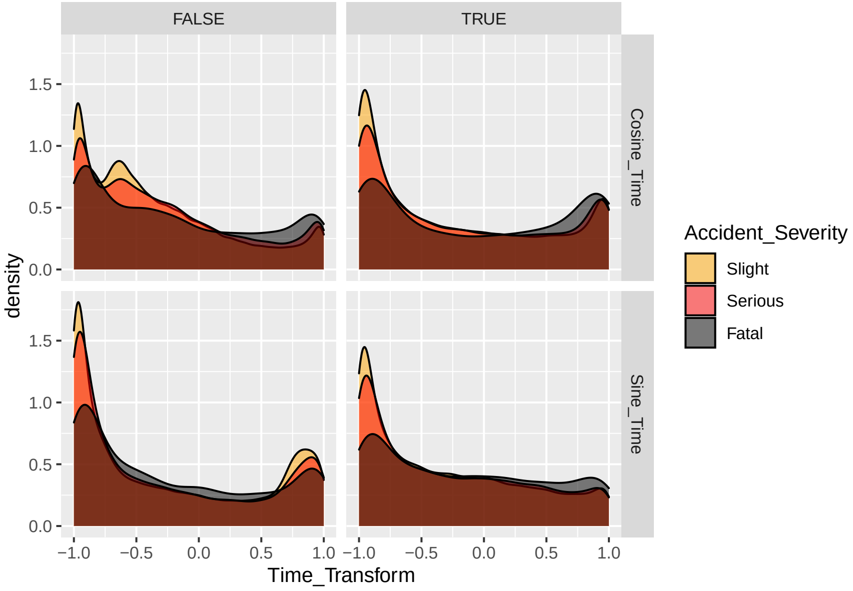


Figure 6: Accident severity distributions of sinusoidal transforms of time. Distributions are further broken down by weekday (left) or weekend (right).

Here we see some separation of densities. These two transformations are included as engineered features. The question remains how much of the observed separation is due to lighting conditions, rather than the time itself.

3.3.4 Road Type

Figure 7: Distributions and proportions of accident severity by road type.

There seems to be some impact here, but it is hard to know whether this provides any information other than what speed limit might.

3.3.5 Speed Limit

When looking at the distribution broken down by accident type, an interesting relationship is found.

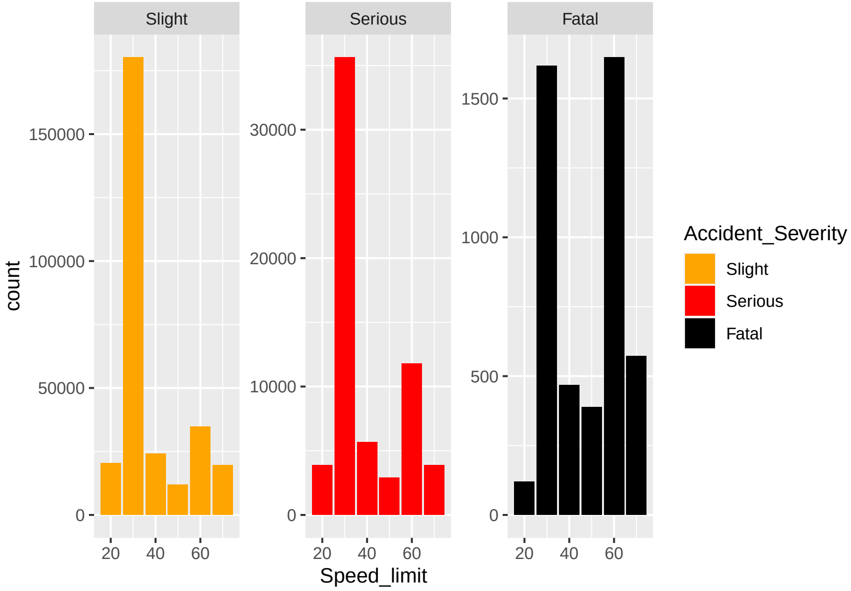


Figure 8: Accident severity distributions against speed limit.

We see that the relative frequency of more severe accidents increases with the speed limit. This of course makes intuitive sense as a greater mean kinetic energy of traffic is implied. It is also evident that 30mph is the modal speed.

3.3.6 Light Conditions

Most accidents happen in daylight, where there is most likely more traffic. Further investigation reveals a higher proportion of fatal accidents with darkness and no lighting or lights unlit.

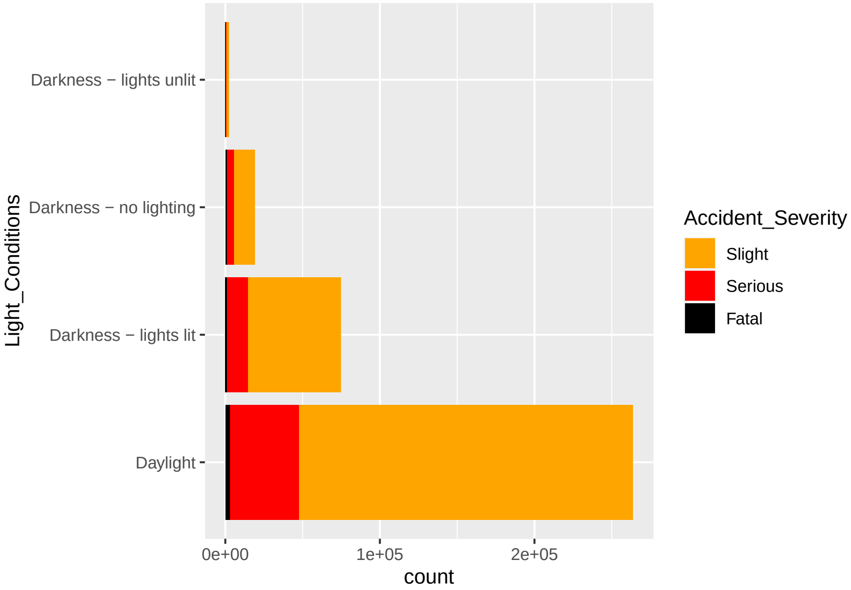


Figure 9: Light condition and accident severity distribution.

A new boolean feature is introduced recording whether daylight or external lighting was present at the event. While the vast majority of cases occur when lighting is on, fatal and serious accidents occur with higher relative frequency when there is no lighting. However, only a small portion of the dataset represents instances with no lighting.

This variable is somewhat redundant for instances during daytime, but perhaps it will aid classification in night-time.

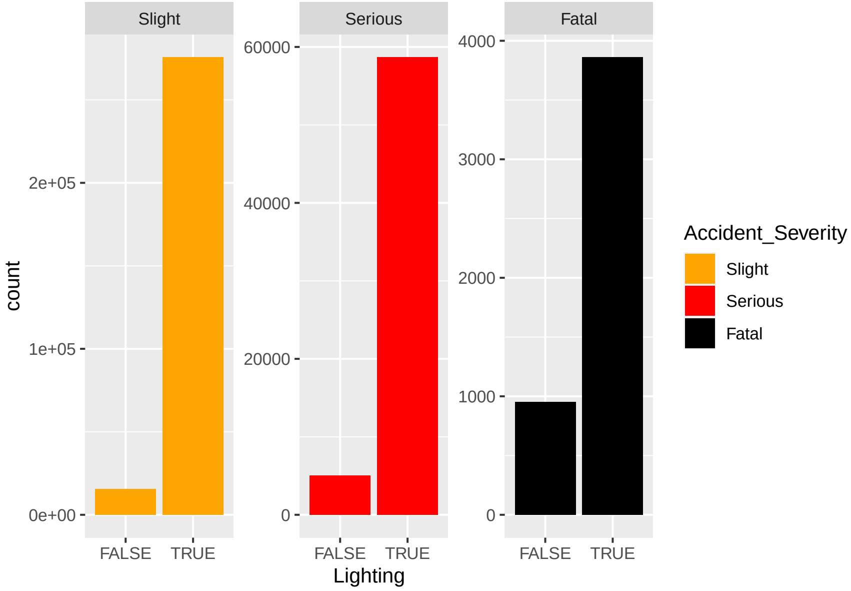


Figure 10: Accident severity distributions against presence of external lighting.

3.3.7 Weather Conditions

The vast majority of accidents occur with fine weather and no high winds:

Proportional analysis of the accident severity by weather conditions does not reveal any insights with confidence, as so few data is collected for observations in some categories. To tackle this, two new boolean columns are made: one for precipitation/fog and the other for high winds:

Surprisingly there does not seem to be much difference. Both these variables are highly unbalanced, but I argue that the high winds variable is too unbalanced for any practicality. As such, this was removed from the dataset along with the unknown precipitation or fog rows.

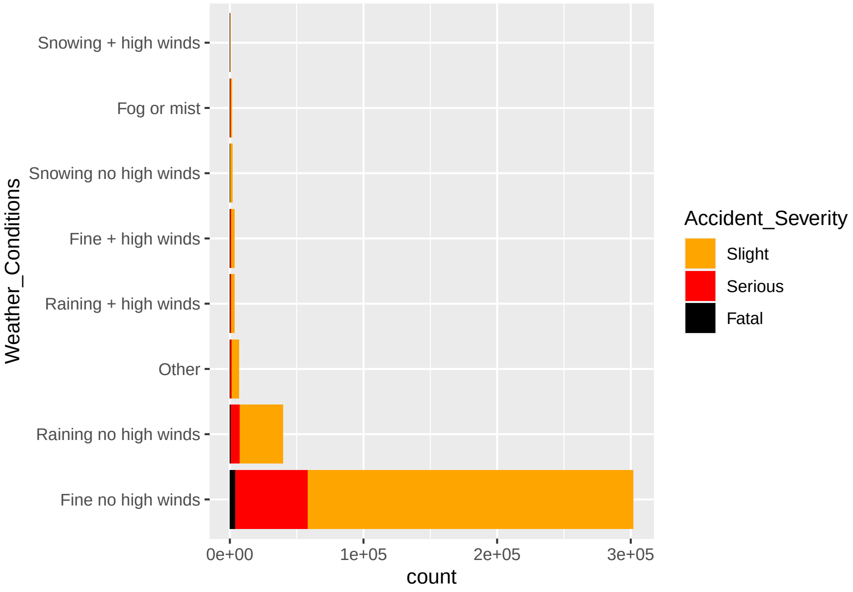


Figure 11: Weather conditions and accident severity distribution.

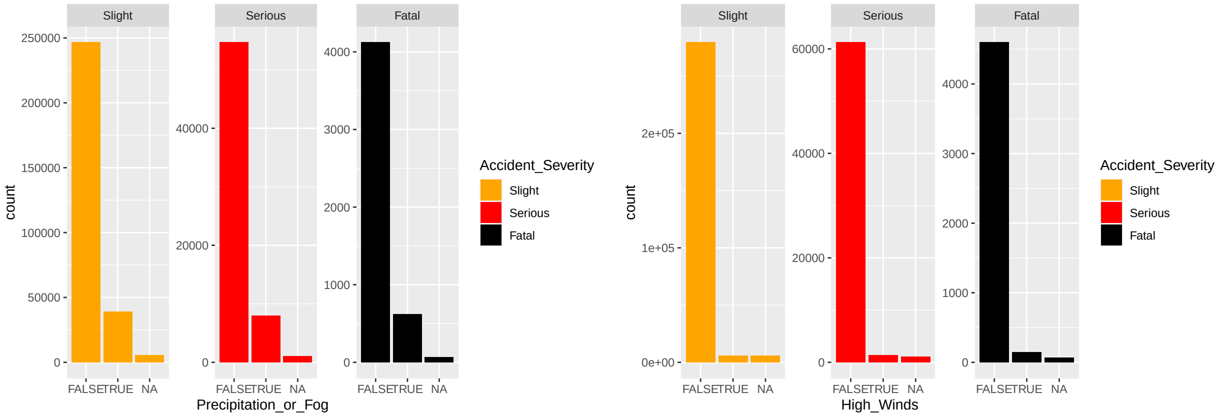


Figure 12: Accident severity distributions against presence of precipitation or fog and high winds.

3.3.8 Road Surface Conditions

This variable is clearly highly unbalanced. The low amount of instances with the 3 least popular categories raises doubts about any observed patterns regarding proportional severities. It seems logical to simplify this classification.

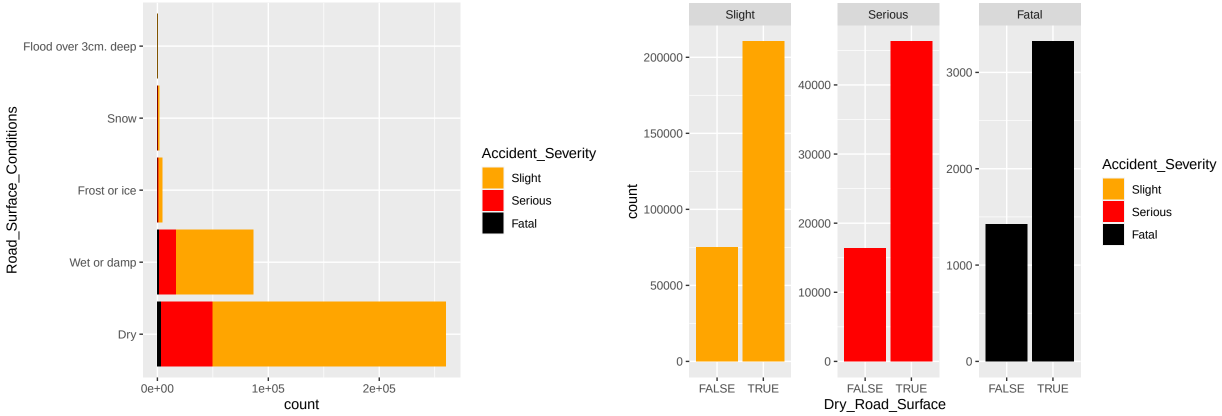


Figure 13: Distribution of road surface conditions and accident severity.

Interestingly, there seems to be little relationship. However, fatal accidents appear to be relatively more frequent when the road surface is not dry. Note that this does not comment on whether accidents occur with greater relative frequency in non-dry road conditions, just that the proportion of accident severities do not seem to change much.

3.3.9 Area Type

The distribution of area type is as below. We see that more severe accidents occur with greater relative frequencies in Rural areas, although more accidents occur in urban areas overall. Perhaps the distribution of speed limits differ in area types.

Table 4: Area type with accident severity distribution.

Urban_or_Rural_Area	Accident_Severity	n
Urban	Slight	192498
Urban	Serious	37311
Urban	Fatal	1739
Rural	Slight	92995
Rural	Serious	25354
Rural	Fatal	3006
Unallocated	Slight	44
Unallocated	Serious	16

To simplify categorisation, rural and unallocated area types are merged in a new binary column.

3.3.10 Average Age of Casualty

There is an interesting relationship between the average age of casualty and the distributions when grouped by severity. Accidents of higher severity occur with greater relative frequency with those older in age. This may be because of increased frailty of aged populations, or perhaps a non-uniform distribution of means of transport with age.

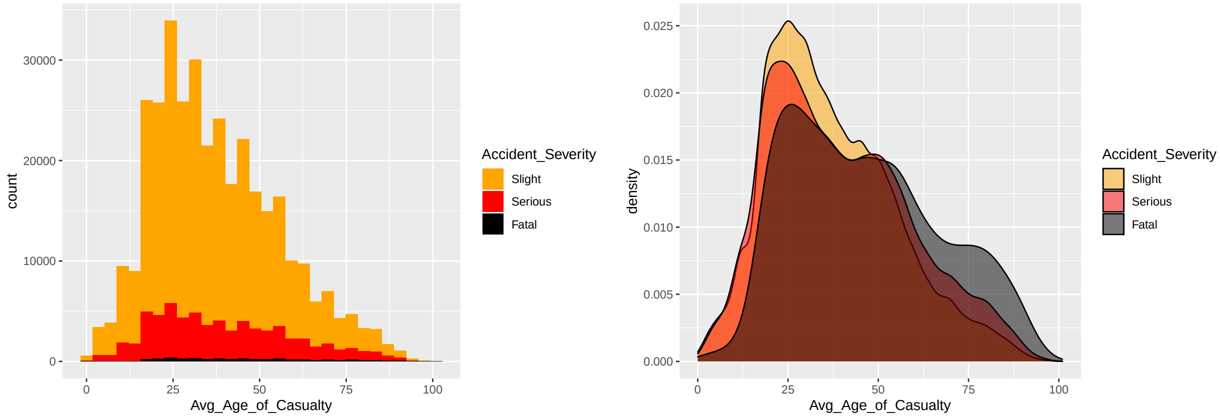


Figure 14: Distribution of accident severity with average age of casualty.

3.4 Covariation

Here I will look at the interaction of two or more non-class variables at once with the aim of gaining understanding about some of the previous insights.

3.4.1 Number of Vehicles and Casualties

We see that the most frequently occurring combination is one casualty with two vehicles involved. The distribution appears quite smooth. Note the logarithmic colour scale. Observations with few vehicles and many casualties perhaps involve transport vehicles such as buses.

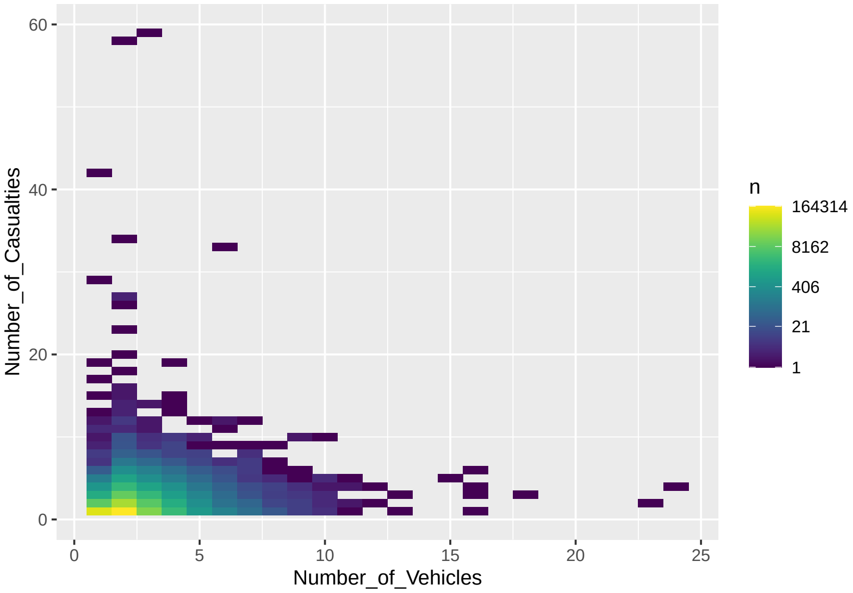


Figure 15: Two dimensional histogram of number of casualties and vehicles.

3.4.2 Speed Limit & Area Type

The most probable speed limit for accidents is 30mph and 60mph for urban and non-urban environments, respectively. This holds true even when broken down by accident severity. Perhaps these speed limits are the most common across UK roads when taking traffic into account. Non-urban areas have a greater distribution of accidents by speed limit.

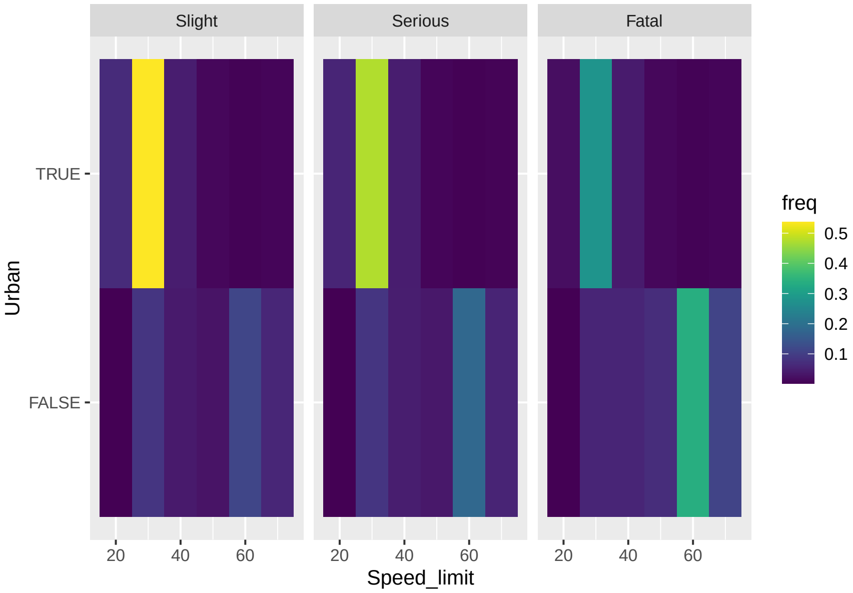


Figure 16: Accident Severity Distribution by Area Type and Speed Limit.

3.5 Final Feature Selection

Correctly identifying features to use is a critical component of any machine learning task. Too few can result in oversimplification, while too many leads to the so called "curse of dimensionality" phenomenon, where (among other problems) the feature space becomes too sparse for the amount of training data available. Here I justify the chosen features for classification purposes.

Firstly, it is obvious not to use the accident index as a predictor variable.

In the temporal analysis during EDA, interesting relationships were found between accident severity and both the time of day and day of week variables. These were refined into two three new features: two containing sine and cosine transforms of the time of day variable to represent the cyclical aspect of time of day, and one of boolean type signifying if the accident occurred on the weekend or not. Meanwhile, no other interesting relationships were found. It is possible that seasonality is related to the target variable through various means, but these are likely represented through other variables, for example those describing weather and road conditions.

The lighting conditions were used to construct a variable signifying whether the environment was lit or not, either naturally or artificially. This simplifies what the essential component relevant for this classification purposes. As such, I remove the original light conditions feature in place with this.

The weather conditions variable contained many categories and was impractically unbalanced, this is replaced with a single boolean feature recording whether there was any fog or precipitation present at the accident.

A few features contained not only a categorisation schema that was not ideal for the purposes of this research, but also a high degree of imbalance with minority categories containing few instances. The weather conditions variable was replaced with a single boolean feature recording whether there was any fog or precipitation present at the accident. Also, the road surface conditions variable was replaced with one signifying dry road surface conditions and the variable signifying area type was made into a boolean variable.

Altogether, the final dataset may compared to the original as follows:

Table 5: Summary of original and final dataset.

Dataset	Rows	Percent_Slight	Percent_Serious	Percent_Fatal
Original	389238	81.38260	17.32205	1.295351
Final	352963	80.89715	17.75852	1.344334

4 Modelling

4.1 Strategy

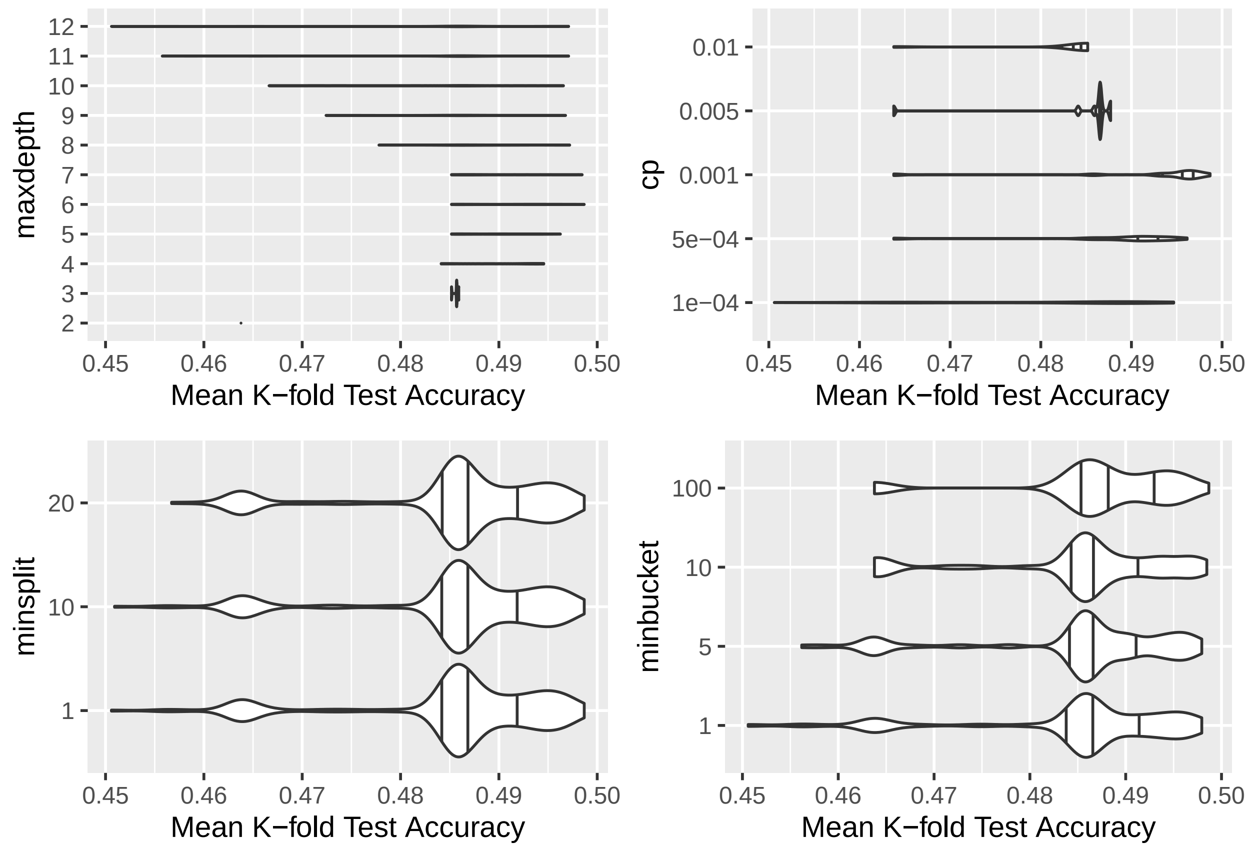
The dataset is split into training and testing sets by year, with observations in 2016-2017 as the training set and 2018 as the testing set. Furthermore, the class distribution in the training set is balanced through under-sampling. Specifically, samples of observations with non-minority classes are taken without replacement until each accident severity category contains equal amount of instances. The total number of observations in the training set becomes 9483, with each class category containing 3161 each.

Afterword, this dataset is used along with with cross validation methods to train various models from different classification algorithms. An attempt to select optimal hyperparameters is done through a controlled search of the hyperparameter space. Accuracy is chosen as the performance metric to be optimised. This is chosen as it is seen to be the most appropriate given the balanced nature of the under-sampled training set.

4.2 Algorithms

4.2.1 Decision Tree

Firstly, a simple decision tree algorithm from the "rpart" package was chosen. The Gini index was used as the splitting criteria and the following parameters was adjusted: 1. *"maxdepth"*: the maximum depth of the decision tree, 2. *"cp"*: a parameter determining the minimum improvement in performance for node splitting, 3. *"minsplit"*: the minimum number of cases in a node which can result in splitting, 4. *"minbucket"*: the minimum number of cases in a leaf node. These were tested in a grid search fashion using 10-fold cross validation using the same splits for each hyperparameter set. Results are visualised below.



As shown above a variability of about 5% mean accuracy was exhibited. The most accurate configuration tested is as follows:

```
## Tune result:
## Op. pars: maxdepth=6; cp=0.001; minsplit=10; minbucket=100
## acc.test.mean=0.4986799
```

These parameters were then trained for the whole (balanced through under-sampling) dataset. The resulting model is shown in the figure below.

We see that speed limit is present at the root of the tree along with other nodes, providing evidence that it is an important variable for classification.

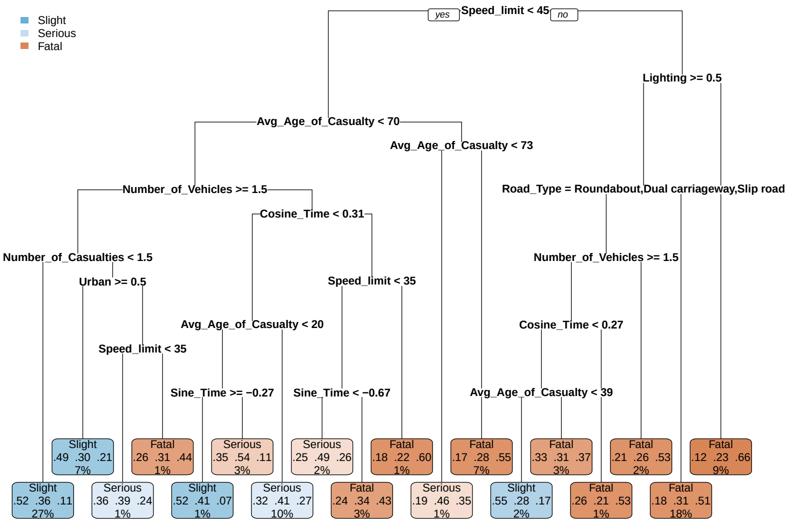


Figure 17: Trained decision tree model diagram.

4.2.2 Random Forest

An ensemble of decision trees were then used through the random forest algorithm implemented in the "randomForest" package. The number of individual learners was set at 300. Also a search was made for the optimal value of the "mtry" parameter which controls the number of features that are randomly sampled at each node. The results are as follows.

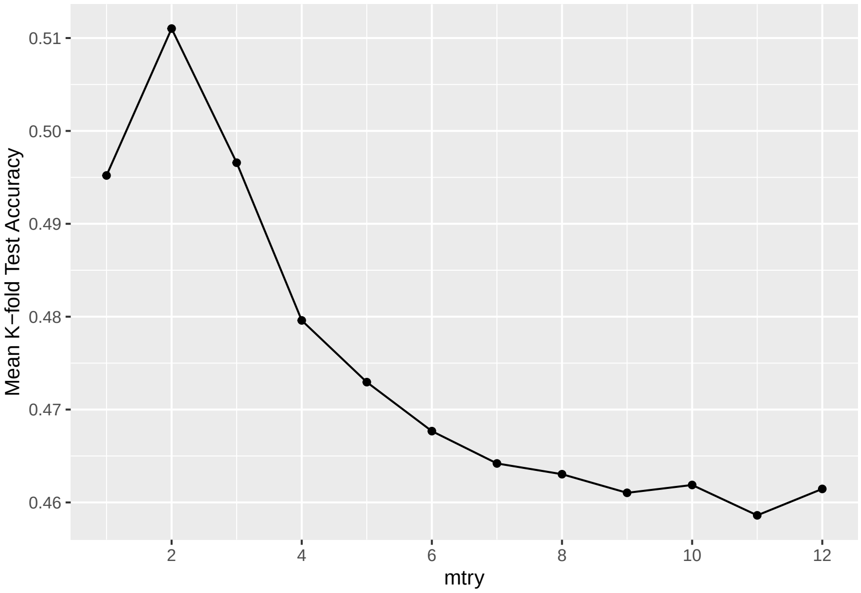


Figure 18: Mean k-fold cross validation test accuracies against varying mtry parameter values.

Similar mean accuracies are exhibited as that obtained with the single decision tree. A general negative trend is seen with increasing the "mtry" parameter past 2, suggesting that greater variety of trees yields better performance.

A model with the optimal parameters found is trained on the training set. The following plot is used to check if the number of trees is adequate. Since the error rates stabilise with increasing trees it appears that there are enough learners.

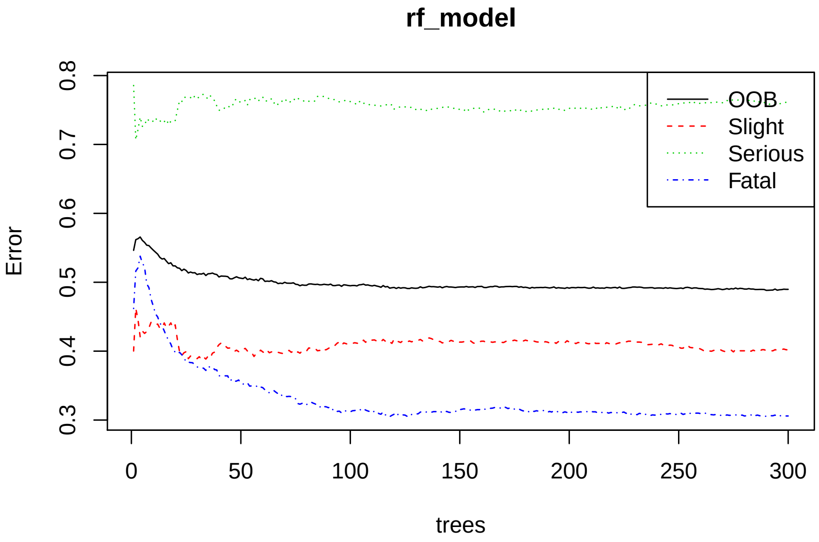


Figure 19: Out of bag mean error along with error rates by category types.

We note the poor performance with correctly classifying serious accident severity instances.

The random forest algorithm also yields a measure of feature importance. The results from the training are as follows:

Table 6: Feature importances as reported by the random forest algorithm.

Feature	MeanDecreaseGini
Avg_Age_of_Casualty	520.60290
Cosine_Time	451.47283
Sine_Time	429.62070
Speed_limit	270.71545
Number_of_Vehicles	187.68175
Number_of_Casualties	154.22971
Urban	130.17448
Road_Type	127.75760
Lighting	76.18210
Weekend	68.93950
Dry_Road_Surface	65.05104
Precipitation_or_Fog	50.83336

This ranking provides interesting insight when considered with the decision tree model diagram and EDA investigations.

4.2.3 Naive Bayes

The naive bayes algorithm as implemented in the "e1071" package was used to complement the tree based algorithms used previously. No hyperparameters are available to tune for this algorithm. The same folds for k-fold cross validation as before was used to test the algorithm, obtaining the following results:

```
## Resample Result
## Task: training
## Learner: classif.naiveBayes
## Aggr perf: acc.test.mean=0.4508047
## Runtime: 1.54699
```

Once again, a model was trained on the full training set.

4.2.4 Support Vector Machine

Finally, classification by partitioning feature space was attempted using the SVM algorithm as implemented in the "e1071" package. The dataset was scaled as by default in the function implementation. Both linear and quadratic versions were tested using a polynomial kernel with degree 1 and 2, respectively. Also, the cost parameter was adjusted. Unfortunately, due to the computational expense of training the algorithm k-fold cross validation was not feasible. Stratified holdout using 10% training and 90% testing on the training dataset was used instead, with the same split being used for each parameter set. Despite the lower portion of training data in this scenario, it seems likely that adequate data was available to select parameters adequately given the number of instances of the dataset.

The results are shown in the plot below.

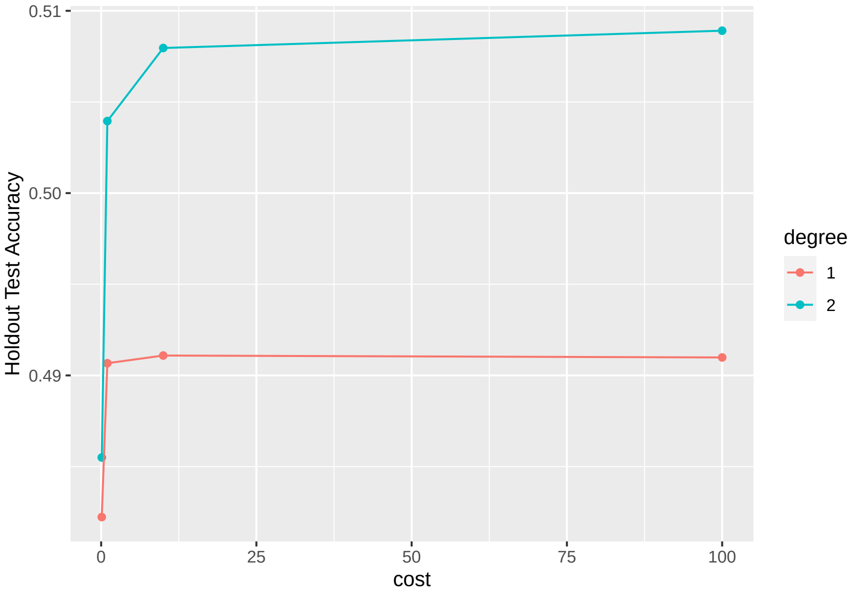


Figure 20: Accuracies of myperparameter search for support vector machine algorithm on training set using a hold out method.

We see that polynomial of degree 2 gave consistently greater accuracy than with using a linear kernel and that increasing the cost parameter resulted in increasing accuracy. The optimal hyperparameter set found was used to train a model with the full training set.

A visualisation of the feature space partitioning is shown below using both the training model and data. The two chosen variables, average age of casualty and cosine of time, were previously ranked as most important by the random forest classifier. The space here is not well separated, highlighting classification challenges with this dataset.

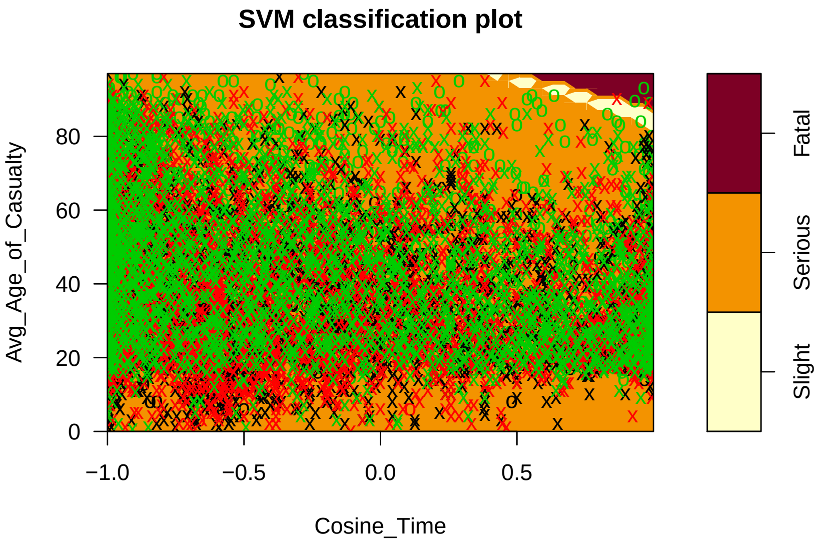


Figure 21: A sample visualisation of feature space partitioning with the model trained using all training data. Note apparent lack of separation.

5 Results

The trained models were tested on the testing data consisting of accidents occurring in 2018. Results are shown in the table below.

Table 7: Various performance metrics of models tested against testing data.

Algorithm	Accuracy	Balanced_Accuracy	Kappa	Precision_Slight	Precision_Severe	Precision_Fatal
Decision Tree	0.5054313	0.4955064	0.0808381	0.5751214	0.2081150	0.7032828
Random Forest	0.5278616	0.5035439	0.0901106	0.5987544	0.2287967	0.6830808
Naive Bayes	0.6371121	0.4488677	0.0786592	0.7701160	0.1099467	0.4665404
SVM	0.5257379	0.5026847	0.0878911	0.5945032	0.2355205	0.6780303

Because the testing set is unbalanced, accuracy is not the best evaluation metric. The percentage of 2018 accidents recorded as slight severity is 79.0798344%, so the models fail to outperform a simplistic model that predicts slight severity regardless of input. More appropriate metrics are balanced accuracy and kappa values. Unfortunately these show that the overall performance of the models are poor, with kappa values showing only slight agreement between predicted and true labels.

Precisions for class categories show the source of poor performance. We see that classification of severe accidents is particularly poor compared with slight and severe. From the exploratory data analysis, similarities between slight and severe accidents were found in the time and average age of casualty features, which were of high importance for the tree based algorithms.

The most distinct result is with the Naive Bayes algorithm. This predicted slight accidents with good precision, resulting in a higher accuracy. However, performance for prediction of severe accidents was very poor. The decision tree model predicted fatal accidents with the highest accuracy although all other performance metrics are worse than with support vector machine and random forest, which have very similar performance.

Because these trade-offs are present, the most successful algorithm for this classification problem is heavily dependent on the purposes at which classification is being used for. However I argue that the random forest algorithm provided the most balanced classifier, since it is observed to score best for balanced accuracy and kappa statistic along with displaying relatively good performance for all other metrics.

6 Conclusion

This project has shown the difficulties with prediction of vehicle accident severity. Although variables that are related to accident severity have been identified, the implemented algorithms struggled to distinguish between slight and severe but non-fatal accidents. Fatal accidents were predicted with fairly good success, while severe accidents were predicted very poorly. An alternative classification problem could be developed to distinguish between fatal and non-fatal accidents. This would likely result in greater success.