Factors affecting severity of injuries in UK car accidents

Andy Birch

26/07/2021

Contents

1. Introduction

This report presents an analysis of the factors affecting the severity of injuries sustained as a result of road traffic accidents in the United Kingdom. The data originates from the UK Department of Transport, but has been sourced from a Kaggle dataset [1]. The full data is very large, so in order to keep the processing times manageable, only accidents from 2014 involving two vehicles have been analysed. This is discussed further in the following section and still provides an ample number of observations to perform this analysis.

The objective is to develop a model that will identify the key factors that affect *casualty_severity* (the target variable), which is the seriousness of injuries sustained in an accident and make predictions when the severity is unknown. The real life application of this analysis could include educating drivers about the main risks, or assisting the emergency services prioritise their response to such accidents.

The *sensitivity* and *specificity* of models will be one way that their predictive power will be assessed, as they measure the proportion of serious or slight injuries correctly predicted respectively. It will be important to score highly on both metrics, so the mean of both scores (*balanced accuracy*) will also be considered.

However, the primary metric used to measure success against this objective will be *precision*, which is the proportion of predictions called as "positive" (a serious injury), that were serious injuries in reality. This measure is key because it factors in prevalence, which is the proportion of observations that have a serious casualty. In this case prevalence is very low, so we will need to achieve very strong levels of both *sensitivity* and *specificity* to avoid false positives swamping the true positives.

A low precision score will indicate that a model has little practical use, as there will be a low probability of a casualty actually being seriously injured when such a severity is predicted.

[1] https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables

2. Data structure

The full dataset available from Kaggle is very large (both in terms of observations and features) and in order for this analysis to be performed on a home computer only a subset will be used. The dataset also needs to be uploaded to GitHub for others to be able to repeat the analysis, which again means that the size of the dataset needs to be reduced significantly. Whilst the size is reduced, there are still 111,460 observations, which is comfortably large enough to conduct a thorough analysis. The following restrictions were used to create the dataset to be used:

- *Year of accident = 2014.* The full dataset includes accidents between 2005 and 2014, but one year is ample data
- *Accidents involving two vehicles.* The analysis considers features of the vehicle in which the casualty was travelling and the other vehicle involved in the accident. Some accidents involved 3 or more vehicles, including these would make the analysis significantly more complex
- *Exclude some vehicle types.* Only accidents solely involving pedal bikes, motorbikes, cars and goods vehicles have been included. A small number of number of accidents involved trams, farmyard vehicles, horses etc, including all these vehicle types would created a large number of features
- *Exclude some features.* The full dataset contains 70 features, many of which are categorical with multiple values, which would create a huge number of features when encoded to be used in models. Features were removed where they were not useful for the intended analysis (location, police force name), were very poorly populated or provided little information as the vast majority of observations had the same value (i.e. is car left hand drive?).

The *casualty_severity* has three levels in the raw data - slight, serious and fatal. In this analysis the latter two levels have been combined for two reasons. Firstly the incidence of fatal casualties is very low at 0.56% which exacerbates the issue of prevalence. Secondly, and most importantly, there are a number of factors that could affect if a seriously injured casualty dies that are not captured in the data, including the response time of emergency services and the quality of care received in hospital. Therefore reducing the problem to a binary classification (was the casualty seriously injured or not) will increase the strength of the model and usefulness of predictions.

The data is initially held in three different tables with the following features:

Accidents

Field	Description
Accident_Index	Unique identifier for each accident
Date	Date of the accident
Time	Time of the accident
Road_Class	Class of road (motorway, “A” road, “B” road etc)
Road_Type	Type of road (dual carriageway, single carriageway etc)
Speed_limit	Legal speed limit for road in mph
Junction_Detail	Type of junction (roundabout, t junction, not a junction etc)
Light_Conditions	Combination of natural daylight and street lighting
Weather_Conditions	Combination of precipitation (rain, snow etc), wind and fog
Road_Surface_Conditions	Was road wet, icy, covered by snow etc

Vehicles

Field	Description
Accident_Index	Unique identifier for each accident
Vehicle_Reference	Unique number for each vehicle involved in the accident
Vehicle_Type	Type of vehicle (pedal bike, motorbike, car, HGV etc)
Vehicle_Manoeuvre	Type of maneouvre vehicle performing (turning, overturning, stopping, none etc)
Skidding_and_Overturning	Did vehicle skid, jackknife and / or overturn?
Vehicle_Leaving_Carriageway	Where did vehicle leave carriageway (nearside, offside, did not leave etc)
Hit_Object_off_Carriageway	Did vehicle hit object when off the carriageway?
First_Point_of_Impact	Was initial impact on front, back, nearside, offside etc?
Sex_of_Driver	Sex of the driver
Age_of_Driver	Age of the driver

Casualties

Field	Description
Accident_Index	Unique identifier for each accident
Vehicle_Reference	Unique number for each vehicle involved in the accident
Casualty_Reference	Unique number for each casualty in the vehicle
Casualty_Class	Was casualty a driver, passenger or pedestrian
Sex_of_Casualty	Sex of the casualty
Age_of_Casualty	Age of the casualty
Casualty_Severity	Was casualty seriously injured? This is the target variable
Car_Passenger	Was car passenger sat in the front or rear?

The vehicle data includes some features that relate to the other car involved in the accident, indicated with the suffix "_B". For example *Age_of_Driver* shows the age of the driver in which the casualty was travelling, whereas *Age_of_Driver_B* is the age of the driver of the other car involved in the accident.

The casualty data is joined to the vehicle data using the *Accident_Index* and *Vehicle_Reference*, the accident data is subsequently joined on the *Accident_Index*. This creates a single dataset with 33 columns and 111,460 rows.

Some features have numerical values that indicate missing information (e.g. "not known"), these are replaced with NAs to correctly identify them as missing and allow corrective actions to be taken.

At this point the data is split into a training set and a test set. The former will be 90% of the observations and will be used to tune and test different models and the latter will only be used for the final evaluation of the chosen model.

3. Exploratory data analysis

Data overview

The following table shows the class, number of distinct values and the presence of missing values for each variable:

	Type	Distinct_values	Missing_values
Accident_Index	character	77,348	0
Vehicle_Reference	integer	2	0
Casualty_Reference	integer	11	0
Casualty_Class	integer	3	0
Sex_of_Casualty	integer	3	1
Age_of_Casualty	integer	102	1,441
Casualty_Severity	factor	2	0
Car_Passenger	integer	3	0
Vehicle_Type	integer	10	0
Vehicle_Manoeuvre	integer	19	1
Skidding_and_Overturning	integer	6	0
Vehicle_Leaving_Carriageway	integer	9	0
Hit_Object_off_Carriageway	integer	12	0
First_Point_of_Impact	integer	6	5
Sex_of_Driver	integer	3	234
Age_of_Driver	integer	98	1,534
Vehicle_Type_B	integer	10	0
Vehicle_Manoeuvre_B	integer	19	2
Skidding_and_Overturning_B	integer	6	0
Vehicle_Leaving_Carriageway_B	integer	10	2
Hit_Object_off_Carriageway_B	integer	13	2
First_Point_of_Impact_B	integer	6	4
Sex_of_Driver_B	integer	3	8,170
Age_of_Driver_B	integer	92	16,165
Date	character	365	0
Time	character	1,416	0
Road_Class	integer	6	0
Road_Type	integer	6	247
Speed_limit	integer	6	0
Junction_Detail	integer	9	0
Light_Conditions	integer	5	0
Weather_Conditions	integer	8	3,168
Road_Surface_Conditions	integer	6	125

Several fields have missing values, this will be corrected before any models are constructed. The target variable *Casualty_Severity* is a factor with 2 levels as discussed previously and the majority of the other variables are integer, which will work well with the packages used in this analysis. There is some further preparation of the data before it can be used for modelling, this is discussed in the next section. The majority of variables contain a small number of unique values that represent values for categorical features.

As noted previously, the target variable is heavily unbalanced in the data:

Casualty_Severity	count
0	90,099
1	10,214

The prevalence of serious injuries is just 10.2%, so any model will need to achieve very high accuracy to provide useful predictions.

Checking for outliers

The majority of features are categorical and the values that they contain all map directly to descriptions provided in the data guide provided via Kaggle. This guide can also be found on the GitHub repo for this analysis. Therefore we know that there are no erroneous values for these features.

We can visualise the distribution of numerical features to check for outliers. Firstly looking at *Age_of_Driver*, we can see 1,682 values below 16 years of age, which is below the legal driving age in the UK. This chart shows the distribution casualties by driver age, limited to drivers aged 25 years and younger:

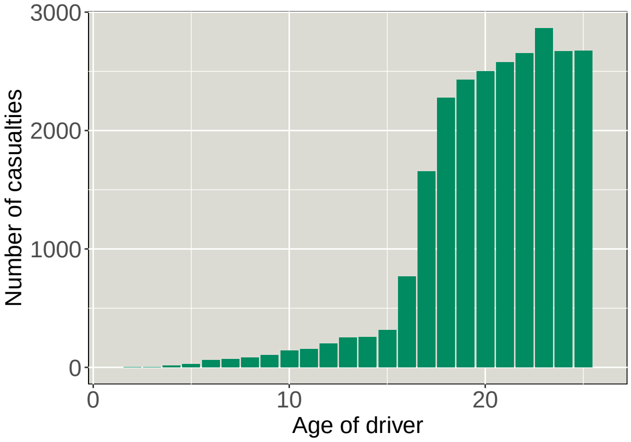

Figure 1: Number of casualties by drivers age

It is possible that accidents involved people driving at an illegal age, but it has been assumed that these values are more likely a result of data errors. Therefore any values below 16 years, for *Age_of_Driver* and *Age_of_Driver_B*, have been set to 16.

The distribution of the number of accidents by week number shows no areas of concern. The box-plot below highlights three outliers with low values, but the bar chart helps explain the reason. The two most extreme outliers are for week number 1 and 53, both of which were only 4 days. And weeks 52 and 53 would be expected to be lower due to the Christmas holidays, leading to fewer people travelling on the roads for business or leisure purposes.

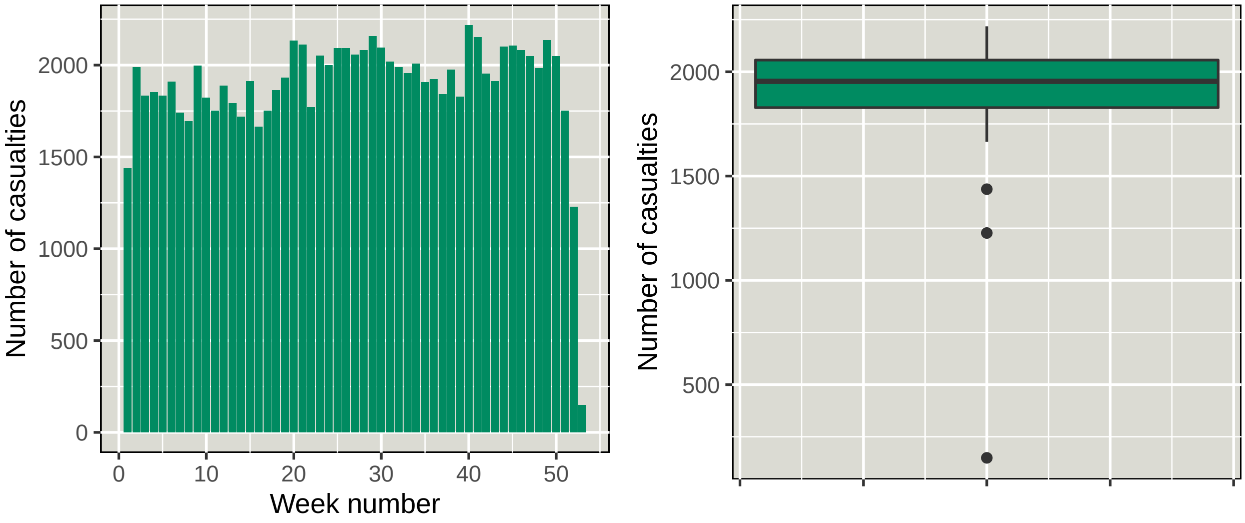

Figure 2: Number of casualties by week in the year

A quick check of the distribution of accidents by time of day also shows a logical pattern without any outliers. There are far fewer accidents in the early hours of the morning when the majority of the population are at home, but the number increases rapidly as people start their morning commute, peaking in the 8 o'clock hour. The number of accidents then quickly declines and then increases again, reaching a peak in the 17 o'clock hour with the evening rush hour. The number of accidents steadily declines hour by hour into the night.

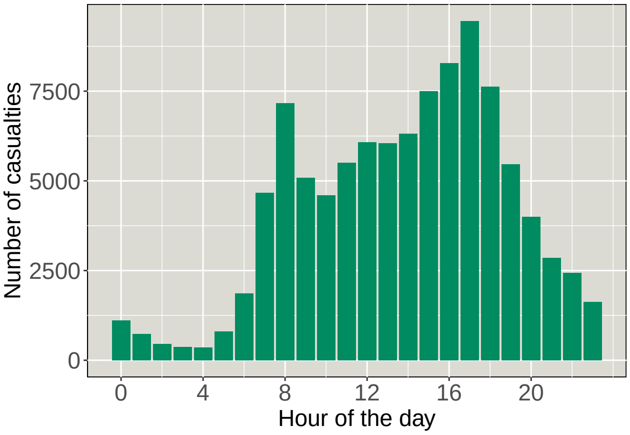

Figure 3: Number of casualties by hour in the day

Predictive power of features

The perceived wisdom is that speed, driver inexperience and sex of drivers are all factors in the severity of accidents, we can see how will this is supported by our data.

Starting with speed, we are provided with the legal speed limit for the road on which the accident occurred, although unfortunately we cannot know that actual speed of the vehicles. The chart below shows the proportion of casualties that were classified as serious, split by speed limit. It shows a clear upward trend as the speed limit increases, until 70 miles per hour when it suddenly declines. This could be due to the type of road and the safety features they include, a central reservation for example that prevents cars crossing into oncoming traffic. If the combination of road speed and type is important, then it may be used by the models used later in this analysis.

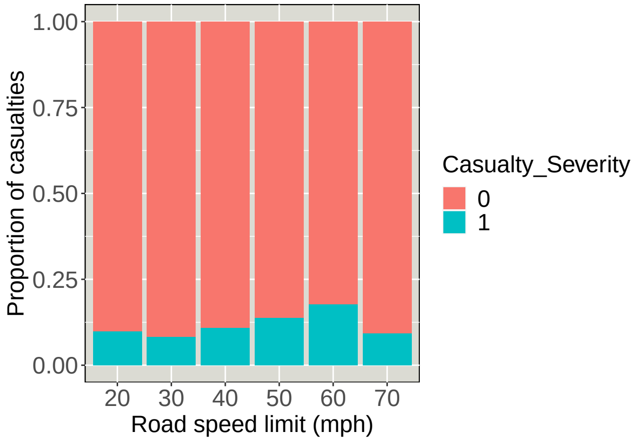

Figure 4: Proportion of casualties by severity and speed limit

Applying the same analysis to the age of the driver (of vehicle in which the casualty was travelling) we can see that the rate of serious casualties does initially decline as the age of the driver increases. However it then consistently increases from about 40 years and up, which could be as a result of older people being more susceptible to serious injuries from a crash than younger people. Again, such a combination of features could be used in the models developed later.

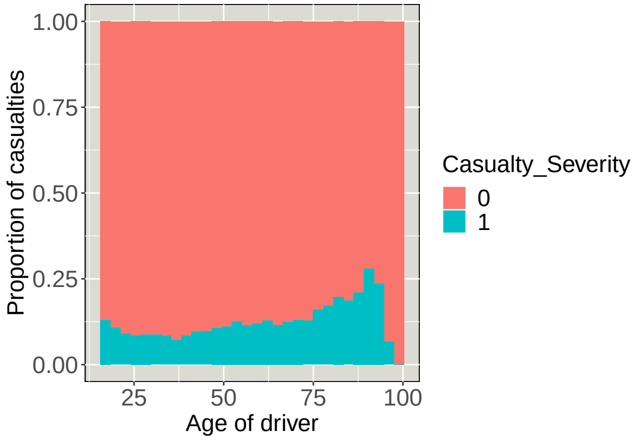

Figure 5: Proportion of casualties split by severity and age of driver

Turning now to the sex of the driver where we observe a far larger proportion of casualties where the driver was male resulted in serious injuries at 12.1%, compared to 6.37% for female drivers. Such a difference in proportions suggests that it could be of practical significance and a chi squared test confirms that the result is statistically significant:

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 812, df = 1, p-value <2e-16
```

This brief analysis suggests that there are several features that could provide effective predictive power for potential models, which is encouraging, and we could continue investigating all of the features in this manner. Indeed it seems reasonable to expect that many of the other features could also be factors in casualty severity, including road conditions, vehicle type, light conditions and so on. Such analysis would be quite time consuming however and would not show if they will provide sufficient accuracy to be of any practical use. The only way to test this is to go ahead and start building models and analysing their outputs.

4. Data preparation

Before models can be built there are a few more steps of data preparation that need to be undertaken.

Firstly many machine learning models will not run with missing values, therefore the missing data has been imputed with the help of the MICE package.

The next task is to create new features to capture the cyclical nature of the *Time* feature. This is necessary to ensure models understand that, for example, 23:55 is closer to 00:05 than it is to 12:00. To achieve this time is represented by a pair of coordinates in a circle, where the x-axis is the sin of Time and the y-axis is the cosine of *Time*. This chart shows the distribution of casualties by time of day, note that the lighter shade in the chart indicates fewer observations which is consistent with the earlier chart:

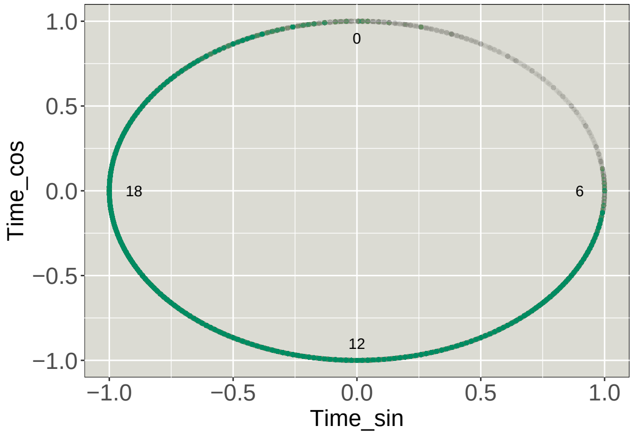

Figure 6: Cyclical distribution of casualties by hour in the day

A new feature *day of week* has also been added, as this could capture different driver behaviours or impairment (e.g. alcohol). This too is a cyclical variable and has been encoded in the same manner as *Time*:

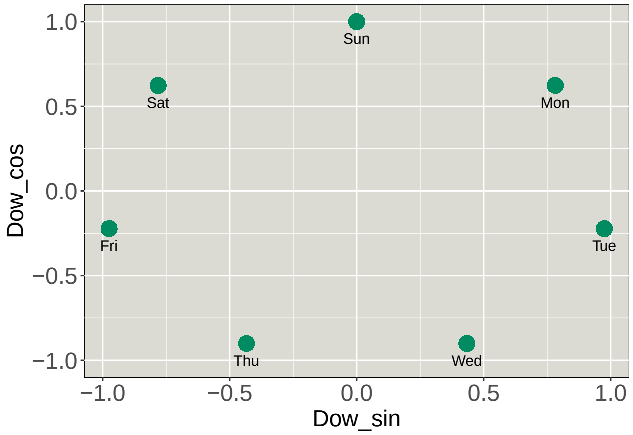

Figure 7: Day of week as a cyclical variable

The penultimate step of data preparation is to apply one hot encoding for the categorical features. These features are provided as numerical values, but they do not have any logical order to them. Supplying figures like this to machine learning algorithms would likely result in very poor models as the algorithm would assume such a relationship exists between the figures.

Categorical features with a small number possible values are encoded by directly creating new columns.

For categorical features with a large number of values, the caret package has been used create dummy variables and then fill out values of one or zero as appropriate. The option *fullrank* is always set to true to avoid perfect multicollinearity between features. As a simple example, if we have a feature that can only take values of A, B or C, then dummy variables only need to be created for A and B. If they are both zero for an observation, then by definition we know that the value was originally C. Including a dummy variable for C as well would introduce perfect multicollinearity between these variables.

The addition of the dummy variables increases the number of columns significantly to 108.

The final step is to split out 10% of observations from the training set to use as a validation set . The training set will be used to train and build models, these can then be tested against the validation set to allow a decision to be made about the final model to be used. Only then will the test set be used against the final model selected to provide the final assessment of the chosen model.

5. Model selection

i. Basic model - prior probabilities

A good starting point for the assessment of competing models is a basic model that guesses whether a casualty received serious injuries or not. This will provide a good baseline against which more complex models should be able to achieve more impressive results.

Predictions are made for the validation set using the prevalence observed in the training set as the probabilities for the sample function. These predictions are then compared to the actual casualty severity and the results shown in a confusion matrix:

	Reference 0	Reference 1
Prediction 0	8,053	915
Prediction 1	957	107

We can see that the model is good at predicting when a casualty is not seriously injured, but fails to correctly predict those that are injured seriously. The following table shows the key results that will be used to compare all the models in the analysis, the metrics are the following:

- Overall: The proportion of predictions that are correct
- Sensitivity: The proportion of serious casualties that are correctly predicted as such
- Specificity: The proportion of minor injuries that are correctly predicted as such
- Balanced: The mean average of the sensitivity and specificity metrics
- Pos_pred: The positive predictive value or precision, which is the proportion of predictions for serious injuries that were serious casualties in reality

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101

As might be expected with guessing, the results are not very impressive. Whilst the specificity is strong at 89.4%, the ability of this simple approach to correctly predict the serious casualties is very low at just 10.5% (the sensitivity measure). It therefore achieves a poor balanced accuracy score and a precision value of 10.1% that will not be of practical use.

ii. Random Forest

The first true machine learning algorithm used in this analysis is the random forest, it has been selected as one of its outputs is the relative importance of each feature. This can then be used for feature selection, i.e. using a smaller number of features with the best predictive power.

When the random forest is initially run, we find that prevalence is causing a significant issue, the model predicts a very small number of serious casualties:

	Reference 0	Reference 1
Prediction 0	8,980	963
Prediction 1	30	59

This allows the model to achieve a high level of overall accuracy because the specificity is so high. The sensitivity is extremely low however as the model fails to predict injuries that were serious in reality. It is interesting to note that the precision is high in this case, but the overall number of positive predictions is so small that the result is actually very poor. This highlights the benefit of considering several metrics when rating model outcomes.

Overall	Sensitivity	Specificity	Balanced	Pos_pred
0.901	0.058	0.997	0.527	0.663

To address the issue of prevalence, the training set is adjusted to be balanced. To achieve this, the number of observations for casualties with non serious injuries is reduced to the number of serious casualties, by taking a random sample. Running the random forest again we can see that the accuracy has improved significantly. There is a strong improvement in the balanced accuracy score, and the precision is considerably better than the simple guess used previously.

	Reference 0	Reference 1
Prediction 0	5,942	228
Prediction 1	3,068	794

Overall	Sensitivity	Specificity	Balanced	Pos_pred
0.671	0.777	0.659	0.718	0.206

It may be possible to further improve this model however, by tuning the *mtry* parameter, which defines the number of features that are randomly selected to be used each time a tree is split. The default value is the square root of the number of features (rounded down), which in this case is ten. The *train* function from the the *caret* package has been used to test a range of *mtry* values within 5 above and below the default value using cross validation. Five fold cross validation with 90% of the records is used for all parameter tuning.

The optimal value of *mtry* is found to be 9 and the model using this value marginally improves accuracy over the default model:

	Reference 0	Reference 1
Prediction 0	6,019	242
Prediction 1	2,991	780

Overall	Sensitivity	Specificity	Balanced	Pos_pred
0.678	0.763	0.668	0.716	0.207

A useful output of the random forest model is the relative importance of each feature when the model was built. This can be used to reduce the features in the dataset to a smaller number of the most useful variables, which has the twin advantages of improving computational speed and, more importantly, reducing the likelihood of overfitting models to the training data.

Any feature with an importance of less than 10% of the most important feature are removed from the dataset at this stage. We now have 40 features:

Importance	Feature
1	Age_of_Casualty
2	Age_of_Driver
3	Age_of_Driver_B
4	Road_Class
5	Speed_limit
6	Time_sin
7	Time_cos
8	Dow_sin
9	Dow_cos
10	Daylight
11	Street_Light
12	Raining
13	Dual_c
14	Single_c
15	Road_Wet
16	Junc_Crossroads
17	Junc_Roundabout
18	Junc_T
19	Car
20	Motorbike
21	Car_B
22	Held_up
23	Right_turn
24	Stopping
25	Right_turn_B
26	Skidded
27	Offside
28	Nearside
29	Hit_object
30	Impact_Back
31	Impact_Front
32	Impact_Offside
33	Impact_Back_B
34	Impact_Front_B
35	Impact_Nearside_B
36	Impact_Offside_B
37	Male_driver
38	Male_driver_B
39	Driver
40	Male_cas

The reduced set of features generally includes some information from each of the original categorical variables, before they were replaced with dummy variables for one hot encoding. Therefore the reduction of features has retained only the most important values from the categorical variables, rather than excluding any in totality. For example there are eight variables representing the type of junction, but only two are found to be among the most useful features.

The ten most important features (actually the top 12 as time and date appear twice) are:

- Age of the casualty
- Age of each driver

- Type of road and speed limit
- Time of day and day of week
- Light conditions and presence of rain

The random forest algorithm is now trained again with the reduced set of features and continues to perform far better than just guessing:

	Reference 0	Reference 1
Prediction 0	5,951	291
Prediction 1	3,059	731

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193

It is worth highlighting at this point just how accurate the model will need to be to overcome the issue of prevalence and provide useful predictions. The balanced accuracy measure of 68.8% is a good result, but the precision is only 19.3%, which is not going to be very useful. If a model achieved 90% scores for both sensitivity and specificity when the prevalence is 10%, then the precision would be exactly 50%. This is because the true positives would be 9% of the observations (90% of the 10%) and the false negatives would also be 9% (10% of the 90%) and hence the prediction is right half of the time.

iii. K Nearest Neighbour (KNN)

The k nearest neighbour (KNN) is another popular algorithm for classification problems, let's see if it can beat the tree based model. Because knn is a distance based algorithm, the scale of features is very important because features that have a larger range of values will separate observations with a greater distance than features with a small range of values. In this analysis there are numerous binary variables, but also several that have much higher values, including the speed limit and the ages of drivers and casualties. Failing to adjust for this issue will generate very poor results:

	Reference 0	Reference 1
Prediction 0	5,087	411
Prediction 1	3,923	611

Overall	Sensitivity	Specificity	Balanced	Pos_pred
0.568	0.598	0.565	0.581	0.135

Once the data has been scaled the knn algorithm can be trained to identify the optimal number of neighbours for the model. The best knn model now produces much better results than the initial knn model, however it is inferior to the random forest model. Although the balanced accuracy is very similar, it has a much weaker specificity which means a lot of false positives, which then overwhelm the true positives and result in a poor precision.

	Reference 0	Reference 1
Prediction 0	5,103	192
Prediction 1	3,907	830

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193
KNN	0.591	0.812	0.566	0.689	0.175

iv. Logistic regression

The next model used in this analysis is logistic regression, which is well suited to binary classification problems. There are no parameters to tune, so the model can be run without the need to use the train function. The model performs relatively strongly, it has slightly stronger balanced accuracy than the random forest, although its precision is marginally lower:

	Reference 0	Reference 1
Prediction 0	5,921	257
Prediction 1	3,089	765

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193
KNN	0.591	0.812	0.566	0.689	0.175
Logistic regression	0.666	0.749	0.657	0.703	0.198

v. Linear discriminant analysis (LDA)

Next to be tested is a Bayesian classifier, linear discriminant analysis or LDA. Again, this algorithm does not require the use of the train function to tune any parameters. The results are marginally weaker than the logistic regression model, due to a slight decline in precision.

	Reference 0	Reference 1
Prediction 0	5,864	251
Prediction 1	3,146	771

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193
KNN	0.591	0.812	0.566	0.689	0.175
Logistic regression	0.666	0.749	0.657	0.703	0.198
LDA	0.661	0.754	0.651	0.703	0.197

vi. Support Vector Machine

This final individual model in this analysis is the support vector machine (SVM). The first step is to decide whether to run the model for a linear solution, or use one of the available kernels that allow for non-linear decision boundary. The table below shows the results when different kernels are used with their default parameters. It shows that the radial kernel performs significantly better than the others both in terms of the balanced accuracy score and precision.

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Linear	0.645	0.753	0.632	0.693	0.189
Radial	0.650	0.766	0.637	0.701	0.193
Polynomial	0.643	0.730	0.634	0.682	0.184
Sigmoid	0.658	0.680	0.656	0.668	0.183

The radial kernel is therefore chosen for the SVM model, which can now be tuned to find the optimal combination of values for the cost (which defines the penalty for a mis-classification) and gamma (which defines the shape of the decision boundary). This chart shows the accuracy for each combination within the ranges tested, with the optimal parameters being 0.03 for sigma and 0.5 for the cost:

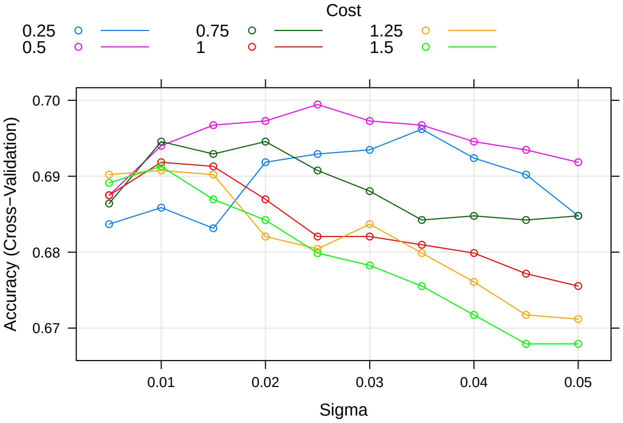

Figure 8: Model accuracy with different model parameters

The SVM model achieves a balanced accuracy and precision which rates it as the fourth best method in this analysis. It is stronger than the knn model, but falls short of the results from the other three methods:

	Reference 0	Reference 1
Prediction 0	5,647	235
Prediction 1	3,363	787

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193
KNN	0.591	0.812	0.566	0.689	0.175
Logistic regression	0.666	0.749	0.657	0.703	0.198
LDA	0.661	0.754	0.651	0.703	0.197
SVM	0.641	0.770	0.627	0.698	0.190

vii. Ensembles

The use of an ensemble can be a powerful approach as the results from the individual models can complement each other and produce a more stable model. Two different ensembles have been tested - the three strongest models (random forest, logistic regression and LDA) and then all five models (aside from the basic guess). A prediction is made based on the majority of "votes" for a positive or negative prediction from the individual models. So if for example the random forest and LDA both predict a serious casualty for an observation, then the three model ensemble would also predict a serious casualty.

The three model ensemble achieves a strong balanced accuracy score, but its precision score is behind some of the individual models:

	Reference 0	Reference 1
Prediction 0	5,886	252
Prediction 1	3,124	770

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193
KNN	0.591	0.812	0.566	0.689	0.175
Logistic regression	0.666	0.749	0.657	0.703	0.198
LDA	0.661	0.754	0.651	0.703	0.197
SVM	0.641	0.770	0.627	0.698	0.190
Ensemble 3	0.663	0.753	0.653	0.703	0.198

The five model ensemble is stronger than the three model version, achieving the highest balanced accuracy of any model, although again its precision score is lower than some of the individual models.

	Reference 0	Reference 1
Prediction 0	5,746	231
Prediction 1	3,264	791

Method	Overall	Sensitivity	Specificity	Balanced	Pos_pred
Basic guess	0.813	0.105	0.894	0.499	0.101
Random Forest	0.666	0.715	0.660	0.688	0.193
KNN	0.591	0.812	0.566	0.689	0.175
Logistic regression	0.666	0.749	0.657	0.703	0.198
LDA	0.661	0.754	0.651	0.703	0.197
SVM	0.641	0.770	0.627	0.698	0.190
Ensemble 3	0.663	0.753	0.653	0.703	0.198
Ensemble 5	0.652	0.774	0.638	0.706	0.195

The following charts show the distribution of casualty severity split by the number of models that “voted” for a positive prediction and whether they were seriously injured in reality. It is clear that the precision improves as more models vote to predict a serious casualty, but the figure even when all five models believe it was a serious casualty is low.

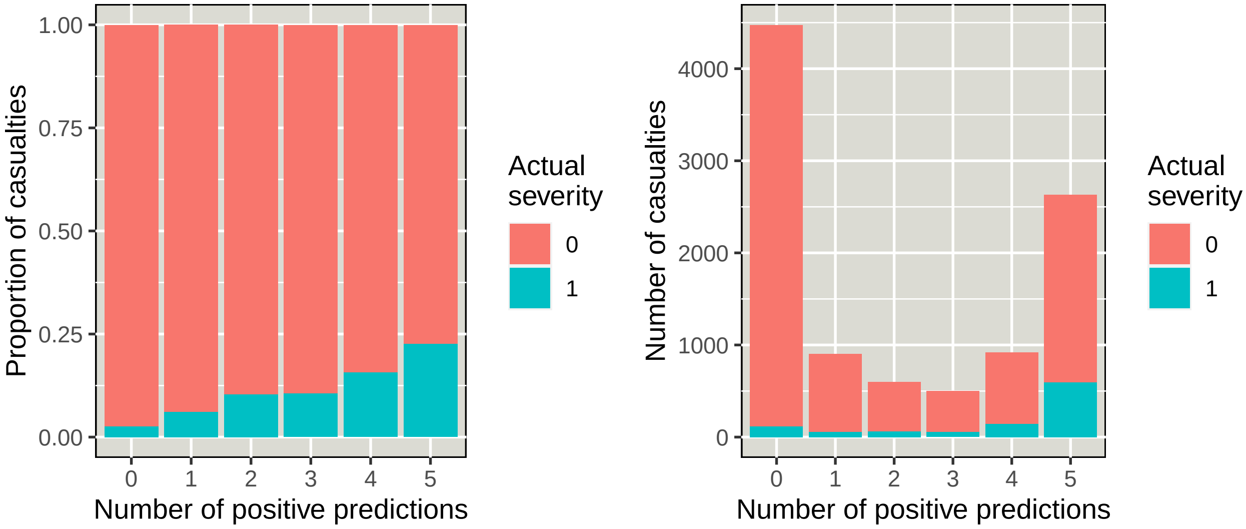

Figure 9: Distribution of casualties by number of positive "votes" and actual severity

viii. Model selection

The random forest was the best individual model and its precision outperformed even the ensemble models. However, the five model ensemble achieved the highest balanced accuracy and the use of an ensemble has the added advantage of being more robust and less prone to overfitting as it several models have to agree in order for a prediction to be made.

Therefore the five model ensemble had been selected as the final model and will now be tested against the test_set.

6. Final model results

The same steps that were necessary for the training set are repeated for the test set - imputing missing figures, correcting drivers age when under 16 and one hot encoding for the categorical variables.

Each of the models developed above are then used to make predictions for test set and a majority vote is used to decide a final prediction. The results are quite close to the scores achieved for the validation set, which shows that we avoided over or underfitting the model:

	Reference 0	Reference 1
Prediction 0	6,380	304
Prediction 1	3,632	831

Overall	Sensitivity	Specificity	Balanced	Pos_pred
0.647	0.732	0.637	0.685	0.186

The charts below show the observations in the test split by the number of votes for a positive prediction (a seriously injured casualty) and the actual casualty severity.

As we saw previously, the proportion does rise as more positive predictions are made, but still only 22% when all 5 call it serious. Equally 2.64% of casualties where all models called it zero were actually seriously injured. The former is the more challenging issue as it means the precision is poor.

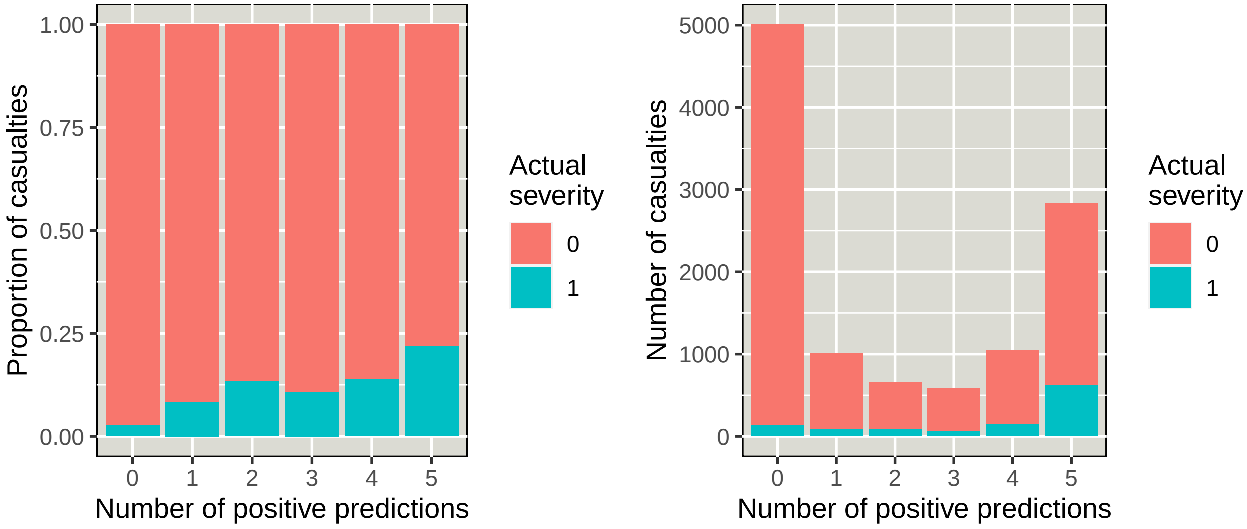

Figure 10: Distribution of casualties by number of positive "votes" and actual severity

7. Conclusions and next steps

This analysis has identified the key factors that affected the severity of injuries sustained by casualties of road traffic accidents in the UK in 2014. Predictions have been made against an unseen data set of data and achieved a balanced accuracy of 68.5%, which is a reasonable result. The low prevalence of serious injuries however, means that this is not sufficient to be of practical use. Indeed only 18.6% of the predictions for serious injuries were actually serious (the precision score).

The final model used was an ensemble of three different machine learning algorithms and we observed that many seriously injured casualties were predicted to be non serious by every model. Similarly many casualties with non serious injuries were predicted to be serious by every model.

If it were possible to access additional data, the following may be helpful to improve the model predictions:

- *Actual speed of impact* – we already know the speed limit for the road on which the accident occurred, but it would be more insightful to know the speed at which the collision occurred. It could be the case that the vehicle(s) were travelling in excess of the legal speed limit. Equally it be true that the vehicle(s) braked sufficiently to avoid a high speed impact.
- *Impairment of drivers* – some accidents involve drivers who are distracted by a mobile phone or are under the influence of alcohol or drugs. This could be a factor in the severity of injuries, if the actions of these driver(s) are not as effective as those who are not impaired.
- *Road worthiness of vehicles* – it would be useful to understand whether unroadworthy vehicles result in more serious injuries during a collision. An unsafe car could brake or steer badly, leading to a more serious collision and a higher likelihood of serious injuries. It could also be the case that an unroadworthy car is less structurally sound and fails to provide adequate protection to those inside it.
- *Use of seat belts* - it is reasonable to believe that being involved in an accident when not wearing a seat belt would lead to more severe injuries, it would be useful to include this in our modelling.
- *Underlying health conditions of casualties* – some casualties may have underlying health issues (e.g. be more susceptible to a heart attack) or be taking medication (e.g. blood thinners) that could increase the likelihood of sustaining serious injuries in a collision.

It may also be insightful to extend the analysis with the current data features by considering additional machine learning algorithms (e.g boosting techniques, including AdaBoost), alternative time periods or different types of accident (e.g. single vehicle accidents).