Accident analysis and severity prediction

Antonio Lotti

September 19, 2020

Abstract

This study was done in the context of the Coursera course Applied Data Science Capstone (https://www.coursera.org/learn/applied-data-science-capstone) of the IBM Data Science Professional Certificate.
The case study for this project was the prediction of the severity of a car accident.
The present report summarizes the data acquisition and cleaning, the exploratory data analysis, the details of the predictive models developed and their testing and results over a set of independent data. The full notebook can be found on Github in the repository https://github.com/AntonioBL/Coursera_Capstone.

Contents

1 Introduction: the problem

One of the main concerns of our modern cities is road traffic. Daily jobs, as well as shopping, spare time and recreation activities, sports and social activities, cause a large number of people to move every day on our road system, sometimes for a considerable amount of their time.
Road traffic is not only a problem for cities. Sometimes, for example in small centers with poor public transportation service, the use of a motor vehicle becomes a necessity for everyday work and activities. Public road transport, non-motorized vehicles, such as bicycles, and pedestrians are also part of the movements on our roads.

Road accidents have thus become a huge problem for our activities.
They are a problem not only in terms of the risk we run when driving, or walking, or cycling on the roads, but also in terms of time we lose when stuck in queue after a road accident. For the society, road accidents are a cost in terms of human lives, and in terms of money and resources spent in preventing accidents and taking care of injured people.

Road accidents thus represent a great loss from a lot of different points of view for our society.

The ability to estimate the likelihood of a road accident and its possible severity given a certain set of conditions has therefore become a hot topic in the last decades. Predictive models that take into account for example road, time, place and weather conditions to predict the severity of a possible accident can be of advantage for different groups.
Among the main stakeholders we could include governments, police forces, road authorities, who can use those predictive models to identify the most critical conditions and determine the most effective countermeasures to reduce the number or the severity of road accidents, for example patrolling particular areas during critical times.
Even normal drivers, cyclists or pedestrians could benefit from these studies, for example they could be warned about dangerous situations, or in case of an accident these models could predict the possible severity and possible time needed for the traffic to resume its normal course. Drivers could also be warned by driving assistance systems using these models, when the road and weather conditions require a more focused attention because of a higher probability of a severe accident. Along these lines, self-driving systems could implement further security protocols to be adopted at the appearance of such critical circumstances. Insurance companies could also benefit from such predictive models, identifying the level of criticality for the typical road routine of people (e.g. when going to work).

The aim of this brief study is to develop a predictive model of accident severity, considering as input variables a set of risk factors such as, for example, weather and road conditions, time of the

day, or specific road intersection type.

2 Data acquisition and cleaning

The data used in this study were collected by UK police forces and made available to the public, excluding the sensitive variables, through the open data UK government site (https://data.gov.uk/).

The database on UK accidents holds records on road accidents with the current set of definitions and detail of information since 1979. UK police forces use a standard form STAT19 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/230590/stats19.pdf) to report information on road accidents.
These datasets are made available to the public under the Open Government Licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/).

The web address for the download is the following:
https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

The databases provide detailed data about the circumstances of road accidents involving injuries or deaths. In this latter case the data refer to people killed immediately or who died within 30 days since the accident.

For this study, only the most recent years were taken into account. In particular, only the 2017 (`Acc.csv`) and 2018 (`dftRoadSafetyData_Accidents_2018.csv`) 'accident' databases were considered and merged into a single database (the details of 'casualties' and 'vehicles' databases were not taken into account). The 2016 database (`dftRoadSafety_Accidents_2016.csv`) was used for model testing.

In the UK Government open data site an additional table is available with weight corrections for the 'Slight' and 'Serious' labels of the accident severity class; this is because the evaluation of accident severity changed during time. For each accident falling in one of those labels, the table introduces a fractional split between these two labels; for example one 'Serious' accident could be split into 0.9 'Serious' an 0.1 'Slight'.
Such weight corrections were not used in this study.

The main strengths of these databases are the following:

- their data are rigorously validated before distribution, for example by the relevant police forces or local road authorities;
- they are collected with the same methodology across the whole UK;
- they include a lot of details regarding the accidents.

The main weaknesses of these data are due to the fact that they record only accidents involving at least one vehicle in which at least one person was injured, and which were reported to the police. The first point determines the fact that these data do not contain details of damage-only accidents, i.e. with no human casualties. The second point has the consequence that accidents on private roads or car parks are not reported, as well as a considerable proportion of non-fatal injury accidents. This may determine an under-sampling of non-fatal injury accidents.

All the data variables are coded instead of displaying textual strings; for some columns '-1' means an unknown or undefined value. Each accident is identified by a unique accident_index; 31 additional columns, with meaningful names, describe the accident details.

The accident severity column does not contain null values. Some of the other columns contains null (or -1) values. For the analysis, the following columns were dropped:

- 'Accident_Index', since no connection to the casualties or vehicle databases were taken into account;
- 'Location_Easting_OSGR', 'Location_Northing_OSGR', 'Longitude', 'Latitude', 'Police_Force', 'Local_Authority_(District)', 'Local_Authority_(Highway)', '1st_Road_Number', '2nd_Road_Number', 'LSOA_of_Accident_Location', since the geographical position or the department in which the accident happened was not taken into account;
- 'Junction_Control', '2nd_Road_Class', since roughly 40% of the values for these columns were null;
- 'Pedestrian_Crossing-Human_Control', 'Pedestrian_Crossing-Physical_Facilities', because a distinction between accidents involving pedestrians and other accidents was not considered in this study;
- 'Did_Police_Officer_Attend_Scene_of_Accident', 'Number_of_Vehicles', 'Number_of_Casualties', because the aim of the model is to predict the severity of an accident given the particular weather, time and road conditions, not the possible number of vehicles involved and casualties, or the attendance of a police officer.

From the 'Date' column only the month was kept; from the 'Time' column only the hour was kept. For columns 'Special_Conditions_at_Site' and 'Carriageway_Hazards' null values were substituted with the mode, corresponding to 'no special conditions' (value '0') and 'no carriageway hazards' (value '0'), respectively; in case of particular conditions it is highly probable that it would have been reported by the those attending to the accident. A similar reasoning can be done for the missing values of 'Junction_Detail', considering 'Not at junction or within 20 metres' (i.e. '0' value) as the most probable case for those accidents (0.55% of the total accidents).
All other rows containing null values in at least one of the variables taken into account were dropped. The original combined database contained 252617 rows. After this cleaning operation, the database contained 249441 rows; less than 1.5% of the initial data were lost.

In addition to the accident severity column ('Accident_Severity'), which will be used as target column, 13 other columns were kept:

- 'Day_of_Week', 'Hour', 'Month', to identify the time during the year, the week and the day when the accident happened;
- '1st_Road_Class', 'Road_Type', 'Speed_limit', 'Junction_Detail', 'Urban_or_Rural_Area', to identify the type of road involved in the accident, and its characteristics;
- 'Light_Conditions', 'Weather_Conditions' to identify the visibility during the accident;
- 'Road_Surface_Conditions', 'Special_Conditions_at_Site', 'Carriageway_Hazards', to identify the temporary characteristics of the road when the accident happened.

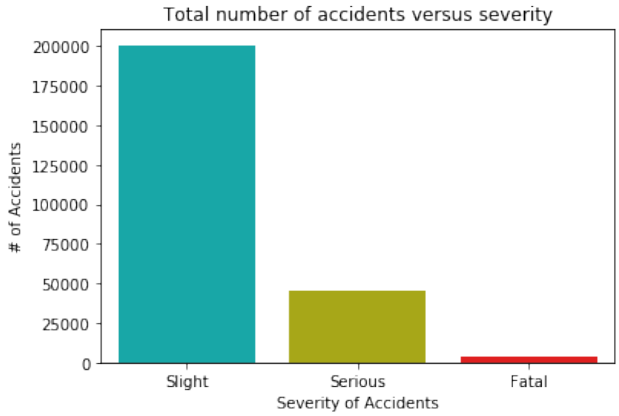
3 Exploratory Data Evaluation

Some preliminary visual investigations were carried on to have a better understanding of the relationships between data and their impact on the accident severity target.

3.1 Accident Severity Distribution

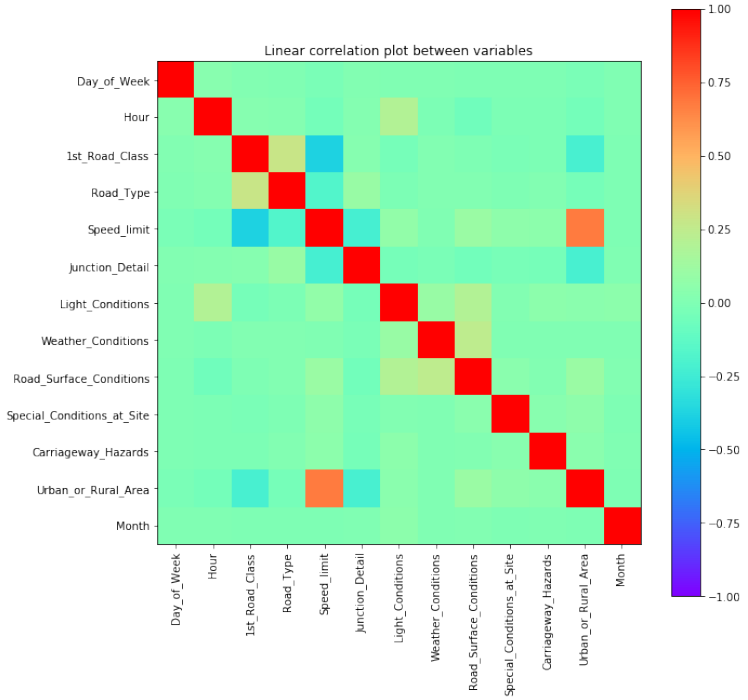
From the overall histogram distribution of accident severity, shown in the following figure, roughly 80.4% of the accidents in the sample were considered slight accidents (i.e. no deaths and no long term hospitalization or serious injuries). Roughly 18.2% were serious injuries and roughly 1.3% were fatal accidents.

The severity distribution was remarkably unbalanced towards slight injury accidents. It was important to take into account this fact when building the prediction model.



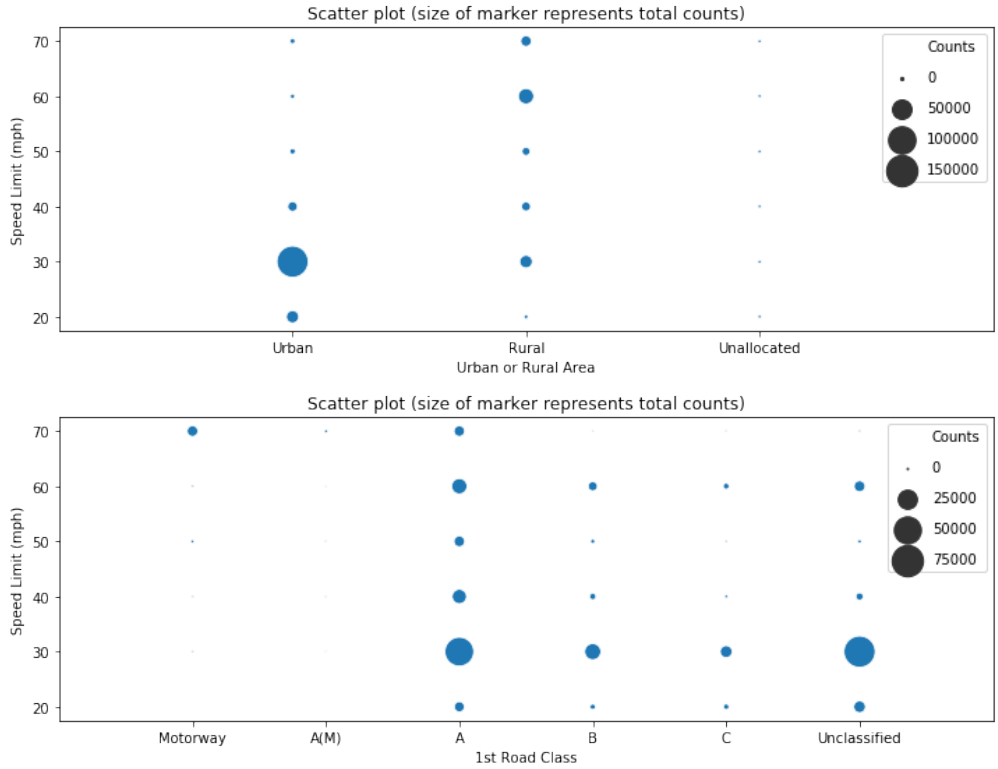
3.2 Correlation Map

The first analysis performed on the data was the creation of a correlation map to verify if it was possible to reduce the number of independent variables and redundant information.



The correlation map (in the above figure) showed a certain degree of positive correlation between

speed limit and 'urban or rural area' variables, and negative correlation between speed limit and '1st road class'.
The correlation of these variables was visualized by mean of scatter plots, shown in the following figures, where the dimension of the marker is proportional to the number of counts.



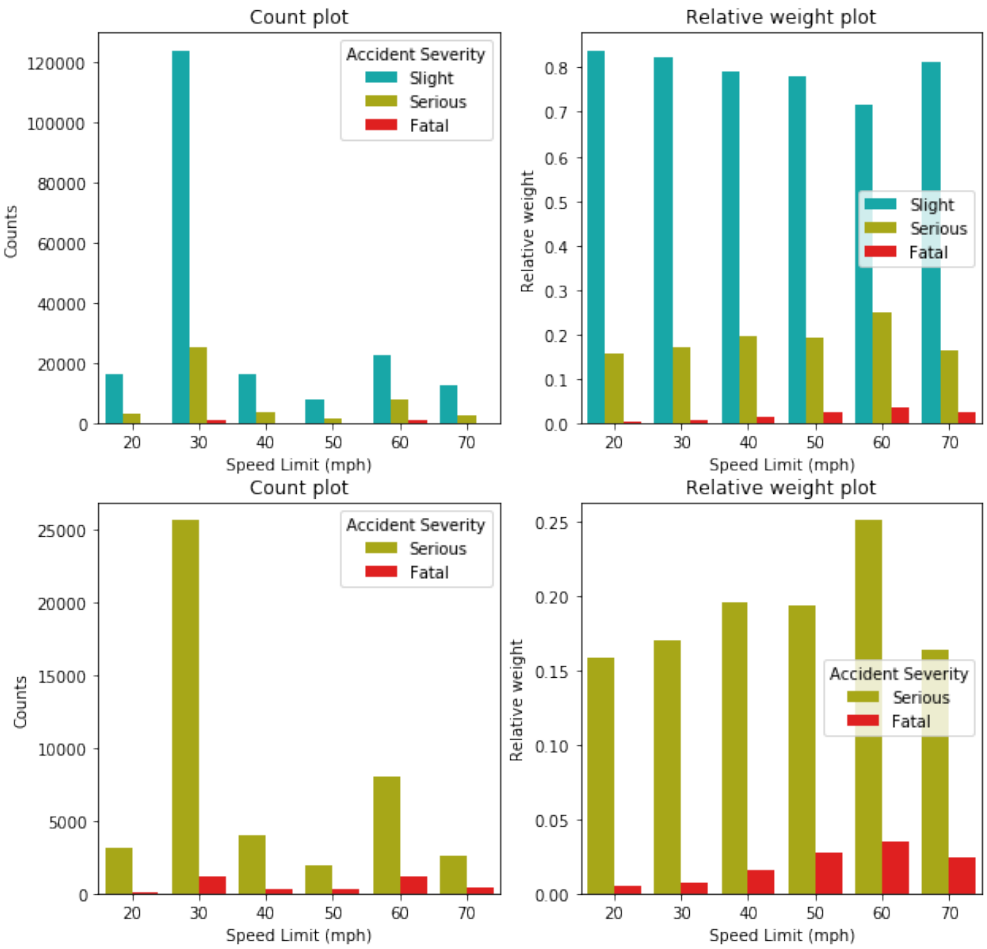
The relation between 'Urban or Rural Area' and speed limit was more evident, since most of the low velocity roads (30 mph speed limit) were in urban areas, while most of the high velocity roads (60-70 mph speed limit) were in rural areas.

Motorways are high speed roads, with speed limit in general 70 mph. A, B and C road speed limit spanned over the full range. Most of the unclassified roads were low velocity roads (30 mph).

Because of the partial correlation between the data and in order to simplify the model, during this study it was decided to drop '1st_Road_Class' and 'Urban_or_Rural_Area' columns when training the models.

3.3 Speed Limit

The Accident Severity distribution as a function of road speed limit was plotted as a histogram count distribution, as well as a histogram distribution in which the bar heights were normalized for each speed limit value to the total number of counts of accidents at that value. This kind of graphic could give a better visualization of factors influencing the severity of the accidents.
Both graphs are shown in the following figure. The bottom row shows only the serious and fatal accidents.

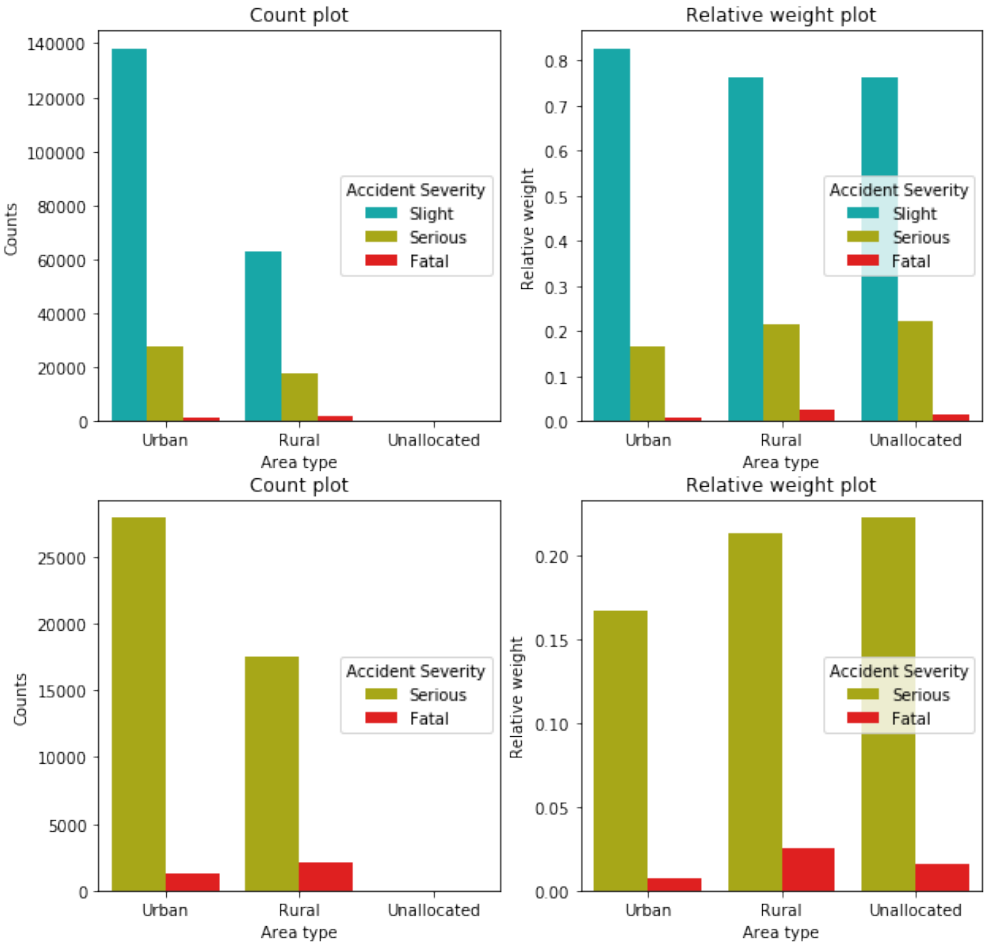


Most of the accidents in the database happened at low velocities (30 mph). A secondary peak appeared at 60 mph.

However, when considering the relative weight inside the speed limit variable, for low speed limit the percentage of serious and fatal accidents was lower than in the case of large speed limit. In particular, the relative weight of fatal accidents reached its maximum for speed limit 60 mph. Indeed, the number of fatal accidents at 30 mph limit was roughly the same as 60 mph limit, but the total number of accidents at 30 mph was much larger than in the 60 mph case.

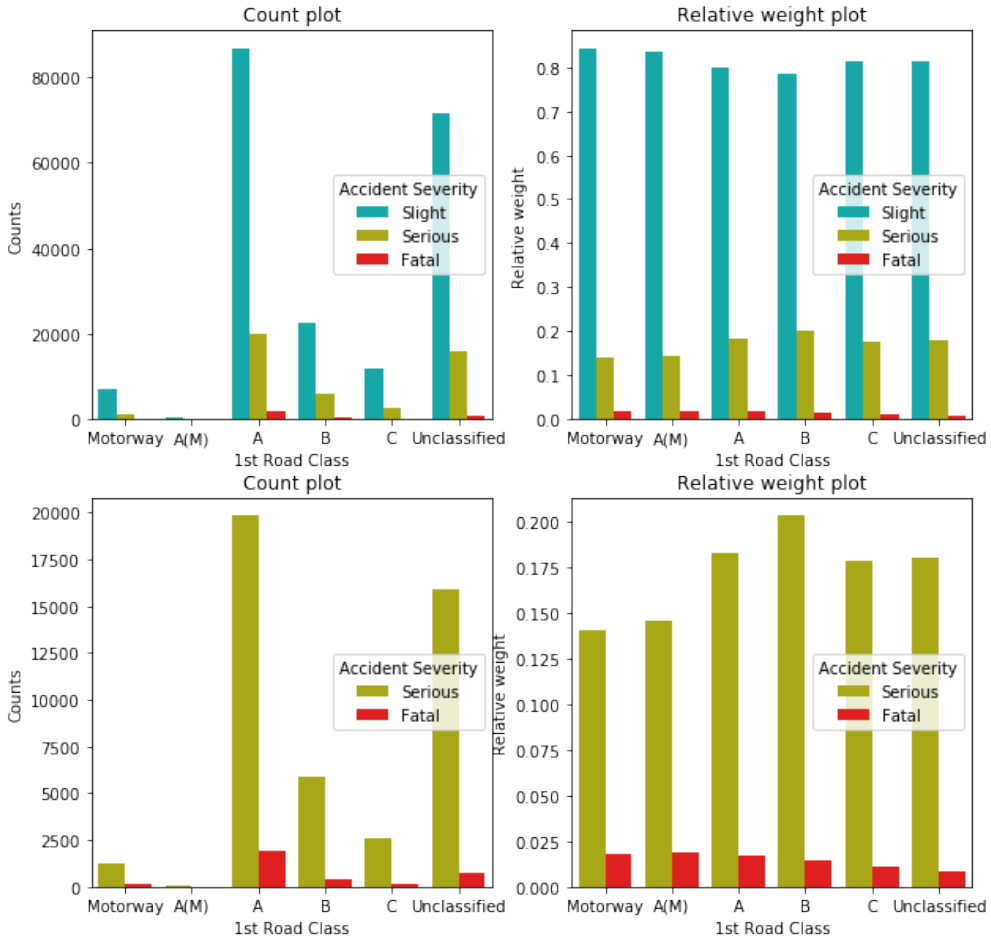
The histograms of counts and relative weights were used to visually evaluate all the different variables of the database.

3.4 Urban or Rural Area



Most of the accidents happened in urban areas. However, when considering relative weights, the relative percentage of serious and fatal accidents in rural areas was higher. As found with the correlation analysis, this is probably caused by the fact that high velocity roads were more frequent in rural areas.
This variable was dropped in the model fitting.

3.5 First Road Class

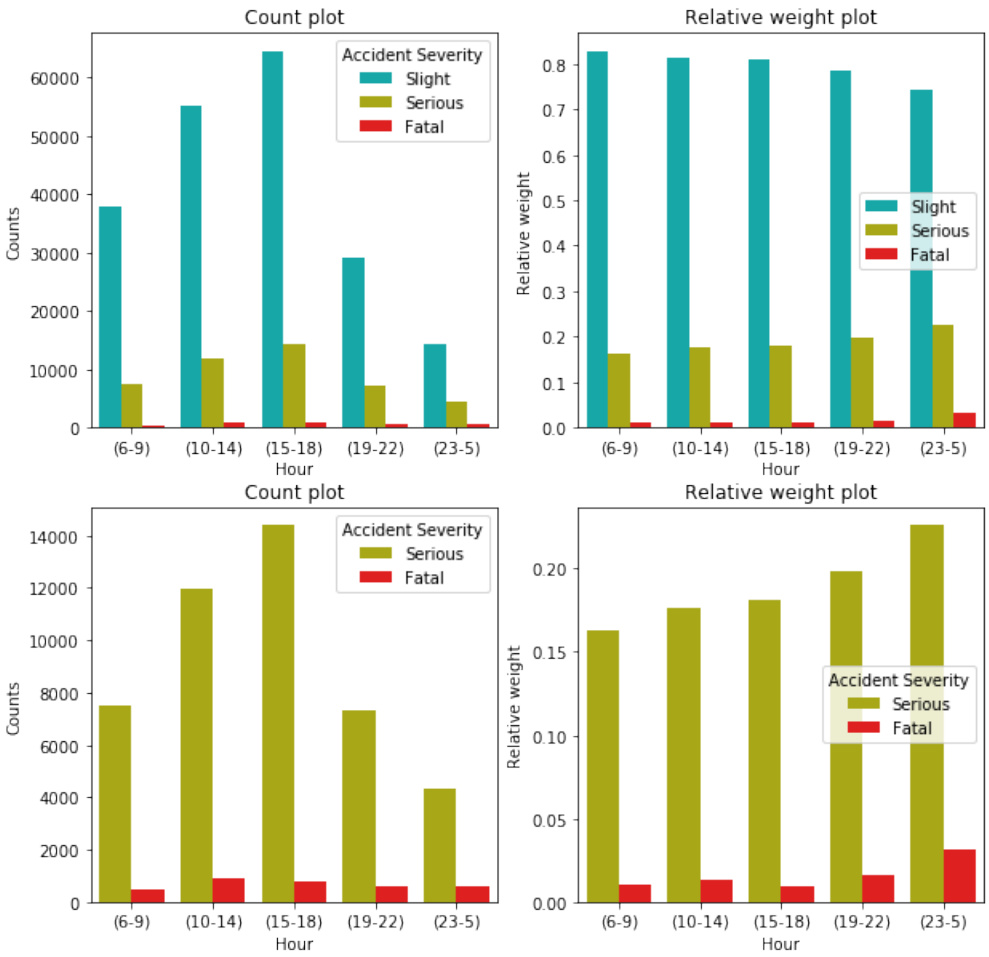


The classes of the roads with the majority of accidents were: A and Unclassified
When considering the relative weight of accidents, serious accidents had a maximum for B class roads. Fatal accident relative weight slightly decreased when passing from Motorways and A classes to B and C classes. This is probably due to the partial correlation of road class and speed limit.
This variable was also dropped for the model fitting phase.

3.6 Hour

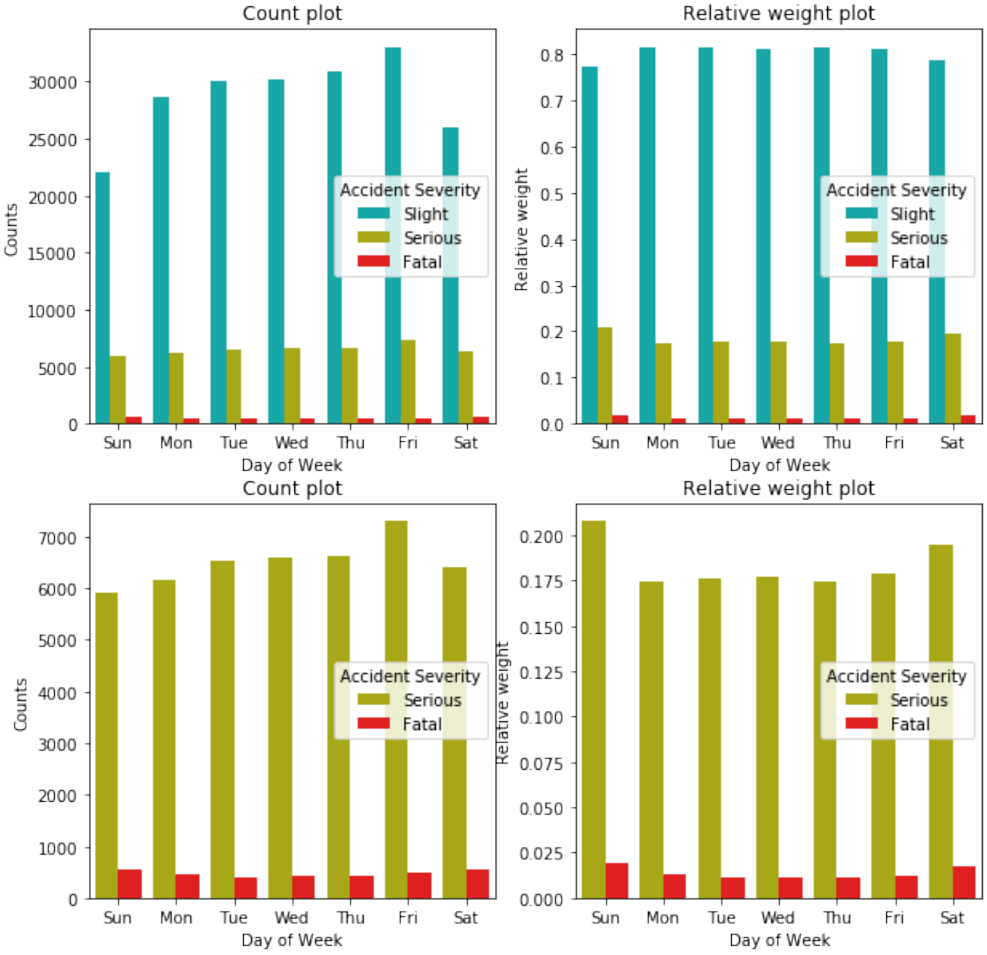
When looking at the distribution with respect to hour of the day, there were two peaks corresponding to rush hours, between 7 and 9 and between 15 and 18. The max accident count happened at 18, during evening rush hour.
The plot of relative weight of accidents, i.e. number of accidents normalized to the total number of accidents happened at that hour, showed that during morning rush hour the relative weight of slight accidents was larger, while at night (from 23 to 5) the relative weight of fatal accidents was larger, with a peak around 4.

The hour variable was grouped in a series of time periods: Morning Rush (6-9), Day (10-14), Evening Rush (15-18), Evening (19-22), Night (23-5).



These time period still exhibited the features observed in the 'hour' variable, namely the peaks at rush hours and the higher relative weight (shown in the graphs on the right in the above figure) of serious and fatal accidents at night.

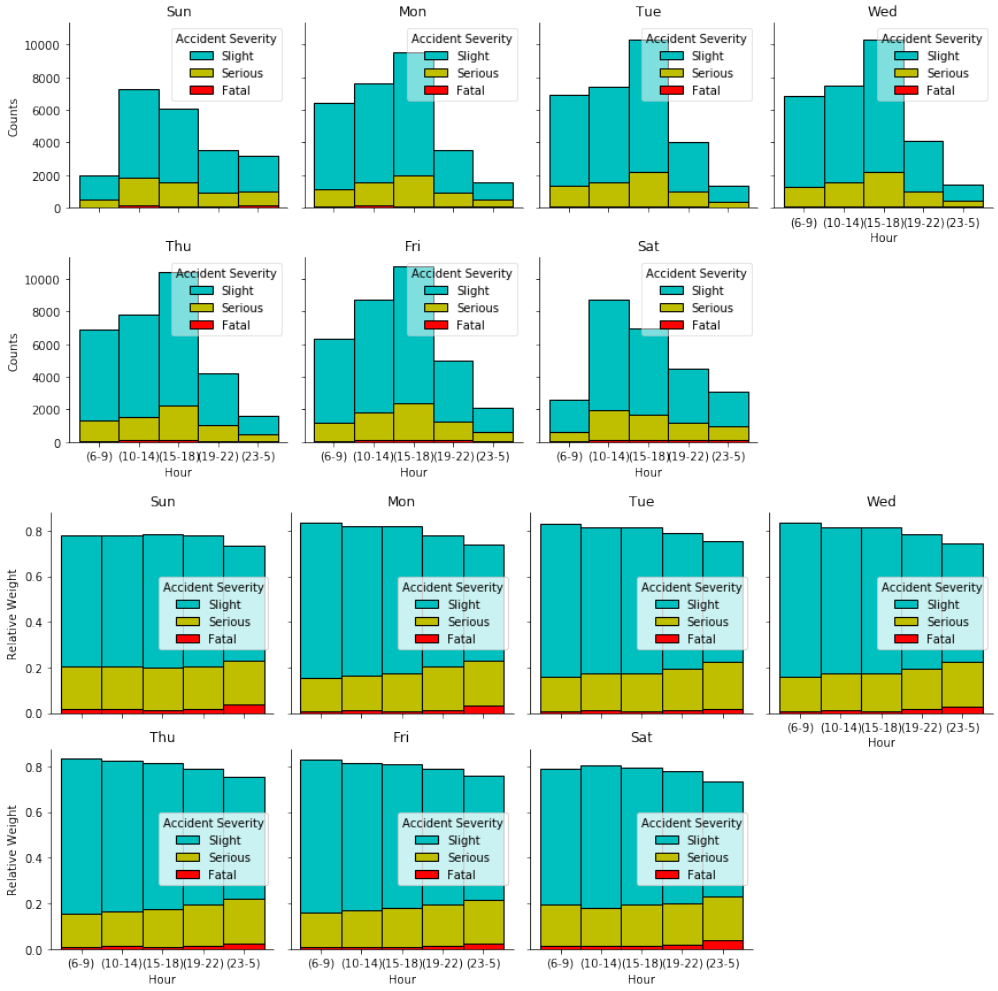
3.7 Day of the Week



During Saturdays and Sundays the frequency of accidents was slightly lower. There was a peak on accident counts on Fridays.

However, there was no clear correlation between relative weight of accident severity and the day of the week, except for a small increase in the relative weight of serious and fatal accident in the weekend.

The combined contribution of day of the week and time period was analyzed, as shown in the following figure.

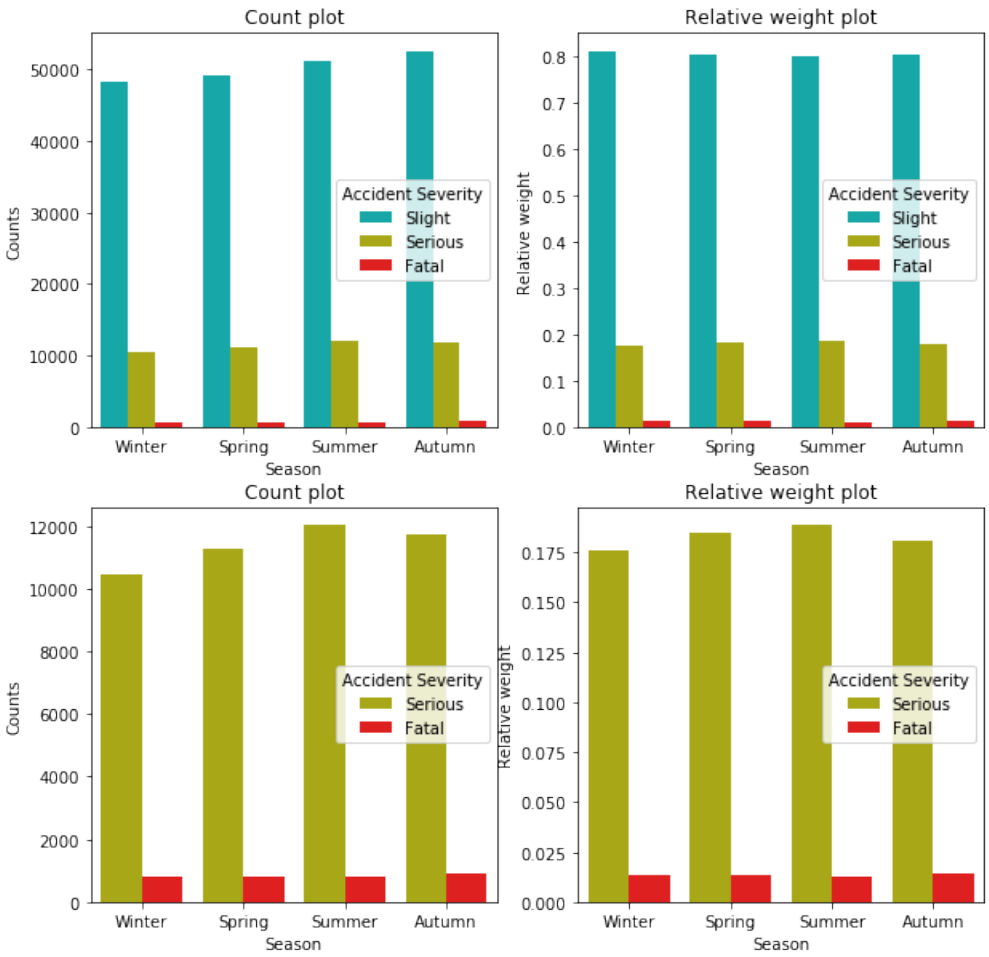


The days in the weekend (Saturday and Sunday) featured a different count distribution than the other days; this was particularly evident for rush hours. The relative weights were similar for all the days.
The variable 'Day of the Week' was therefore reduced to a variable 'Weekend' indicating if the day belonged to the weekend or if it was a working day.

3.8 Month

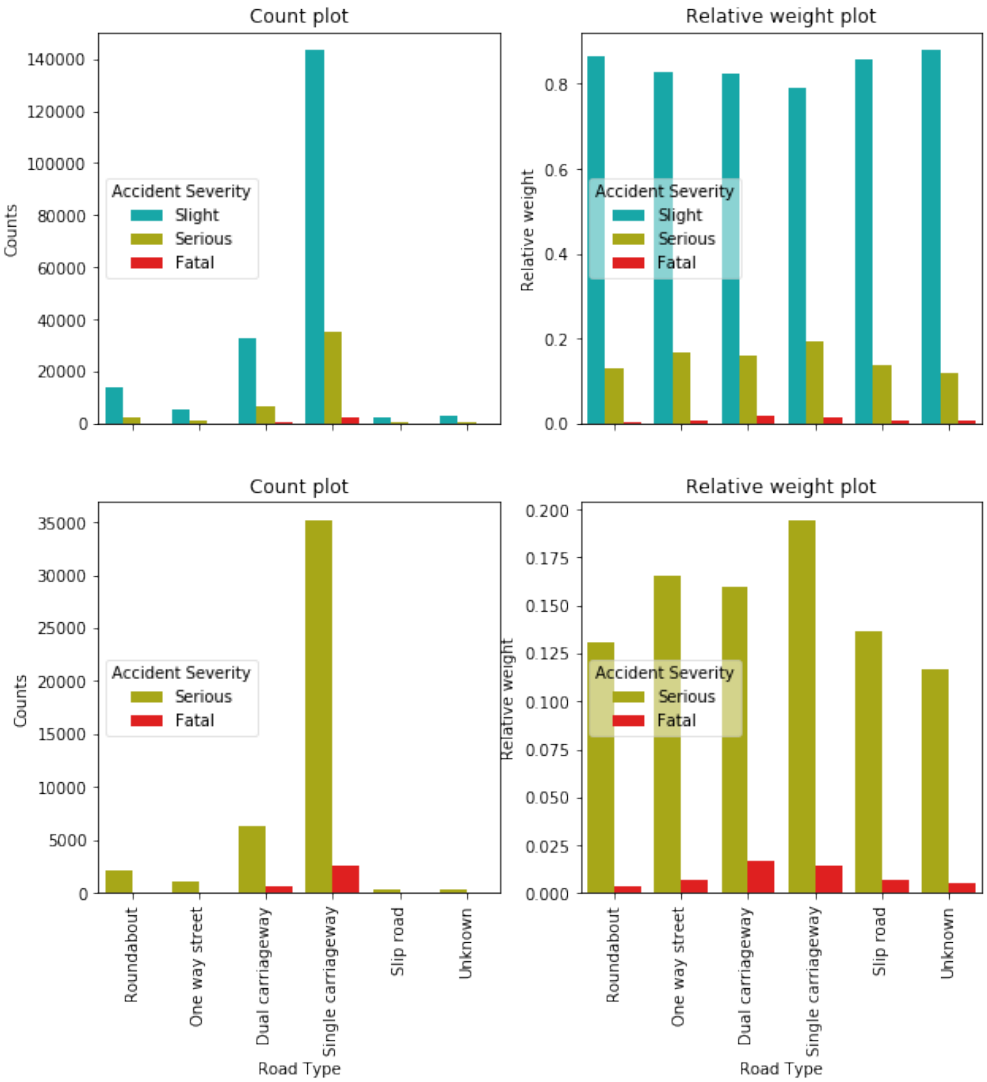
There was no clear indication of the influence of months on accident severity. The count distribution showed three small maxima at January, June and November.
However, the relative weight distribution was basically the same for all months.

The months were grouped by meteorological season, as shown in the following graphs.



Both count plot and relative weight plot seemed similar among the different seasons.

3.9 Road Type

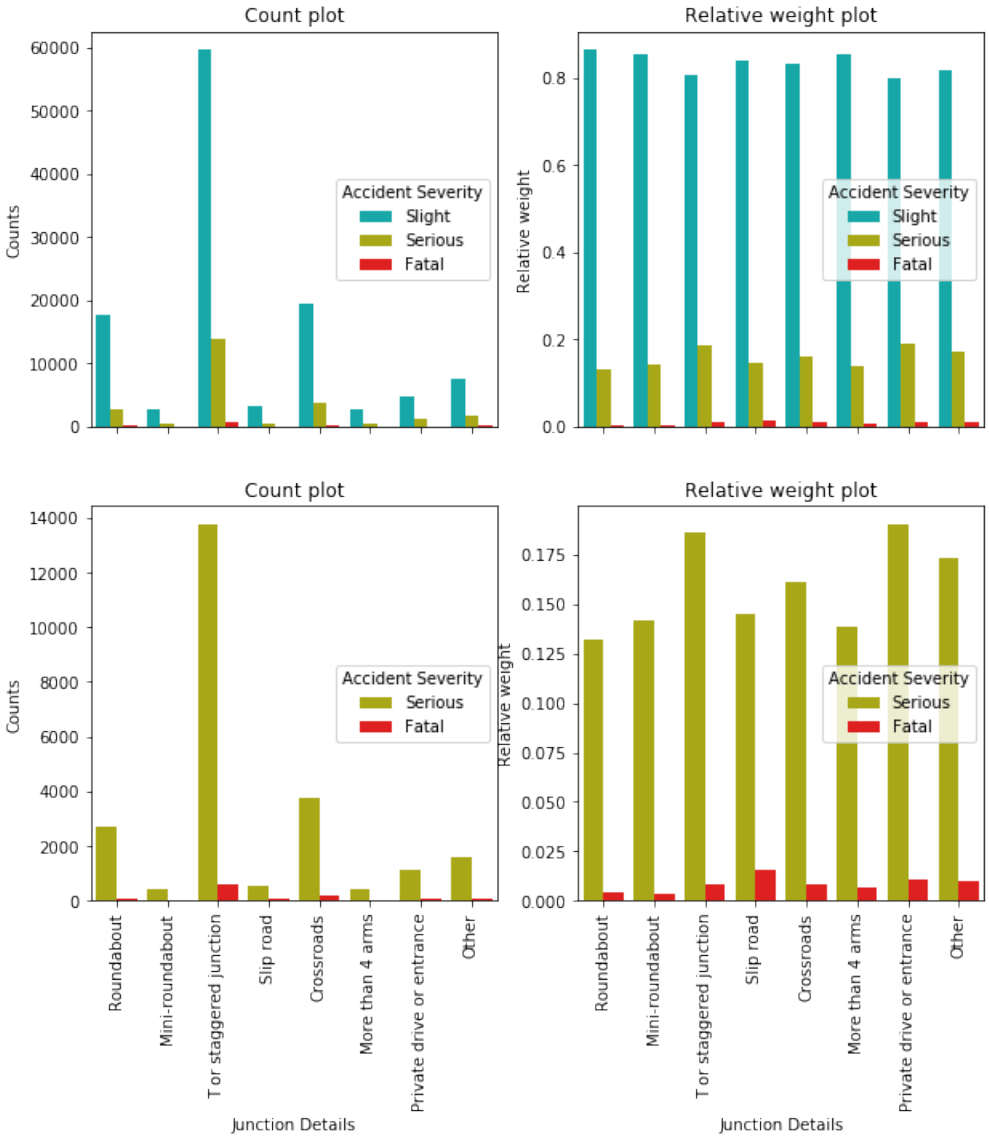


The majority of accidents happened on single carriageway roads, followed by dual carriageway roads and roundabouts.

When considering the relative weight (right part of the above figure), dual carriageway roads had the highest relative value of fatal accidents. This is probably related to the fact that these roads have higher speed limits.

Roundabouts showed the smallest relative weight of fatal accidents for the road type category.

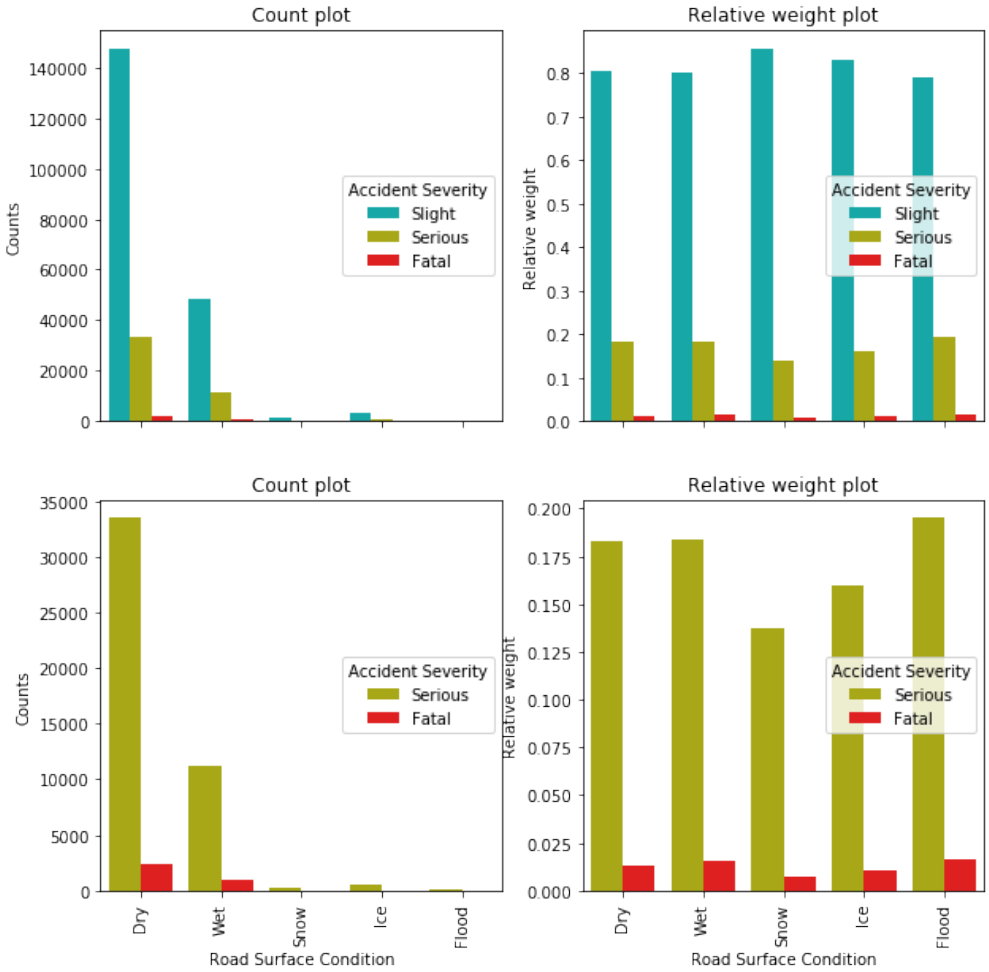
3.10 Junction Details



For what concerns the accident count, the majority of the accidents that happened in correspondence of road junctions happened at T or staggered junctions, followed by normal crossroads and roundabouts.

When considering the relative weight of accident severity (right part of the above figure), the junctions with higher relative weight of serious accidents were 'T or staggered junctions' and 'private drives'. The maximum of the relative weight of fatal accidents was for 'slip road junctions', followed by 'private drives', 'T or staggered junctions' and 'normal crossroads'.

3.11 Road Surface Conditions

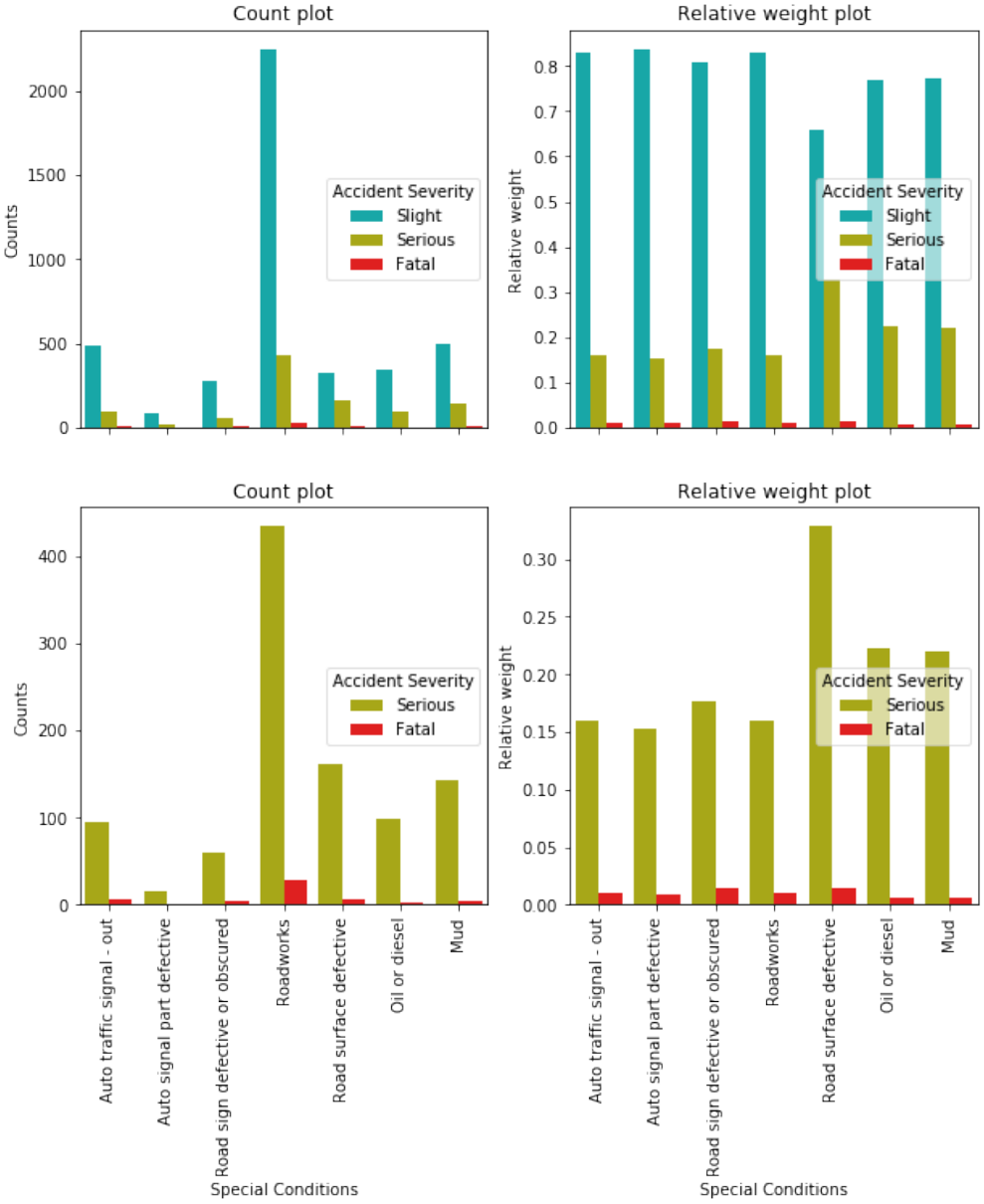


For what concerns road conditions, most of the accidents in the database happened on dry roads, followed by wet roads.

The relative weight of serious accidents showed a maximum for flood, followed by wet and dry roads. The relative weight of fatal accidents was higher for wet and flood road surfaces.

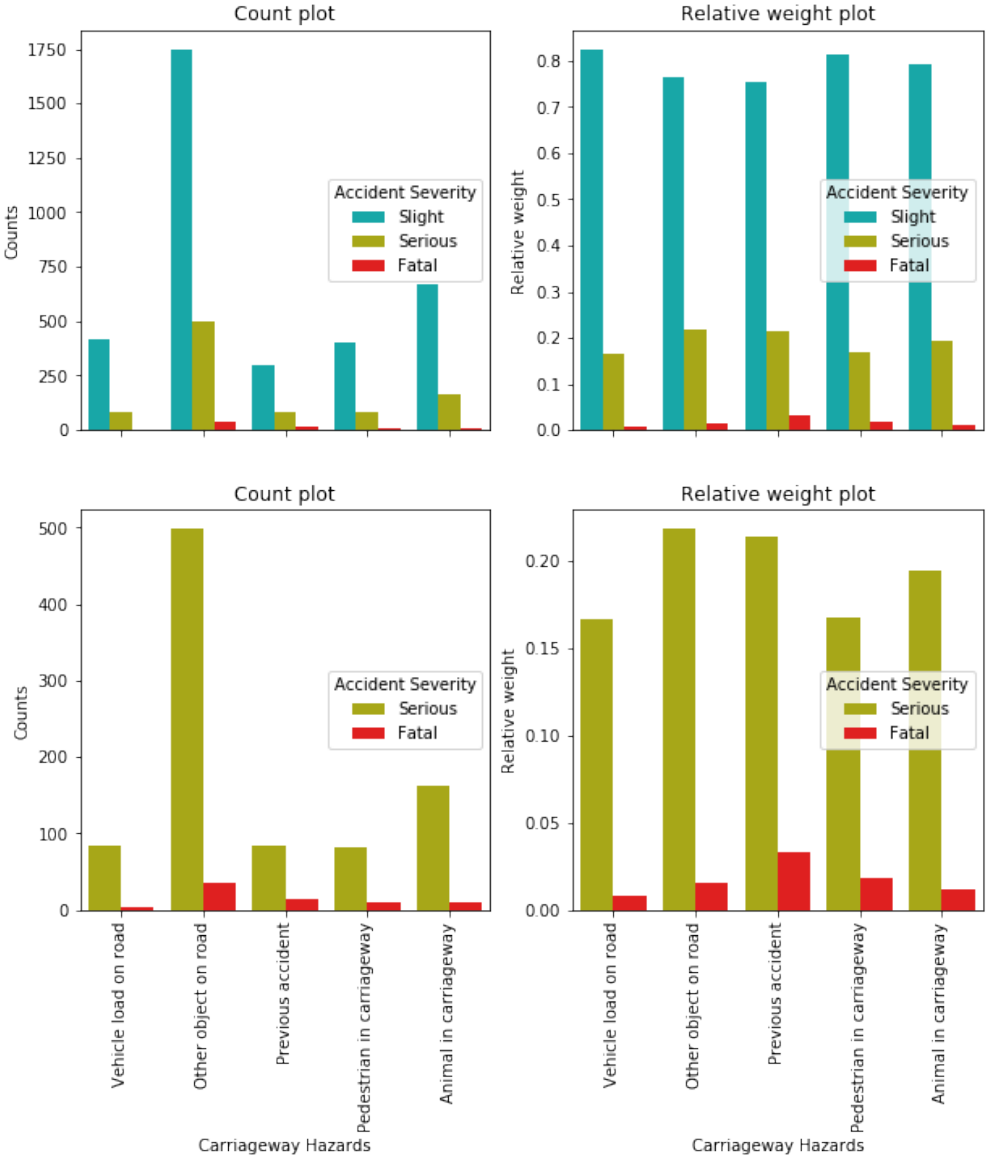
Snow and ice surface conditions showed the lowest relative weight of serious and fatal accidents. Probably, prevention measures, modern winter equipment, together with a more conscious attitude of people during these extreme surface conditions, contributed to this fact. However, the number of statistical samples for these two cases was small, and the presence of only few samples could bias the analysis.

3.12 Special Conditions



Most of accidents happened without special road conditions. The majority of accidents that happened with special road conditions were in correspondence of roadworks.
The weight plots showed that the maximum relative weight of serious accidents was for defective road surfaces, followed by oil and mud. For fatal accidents, the maximum relative weight corresponded to defective road signs and defective road surface.

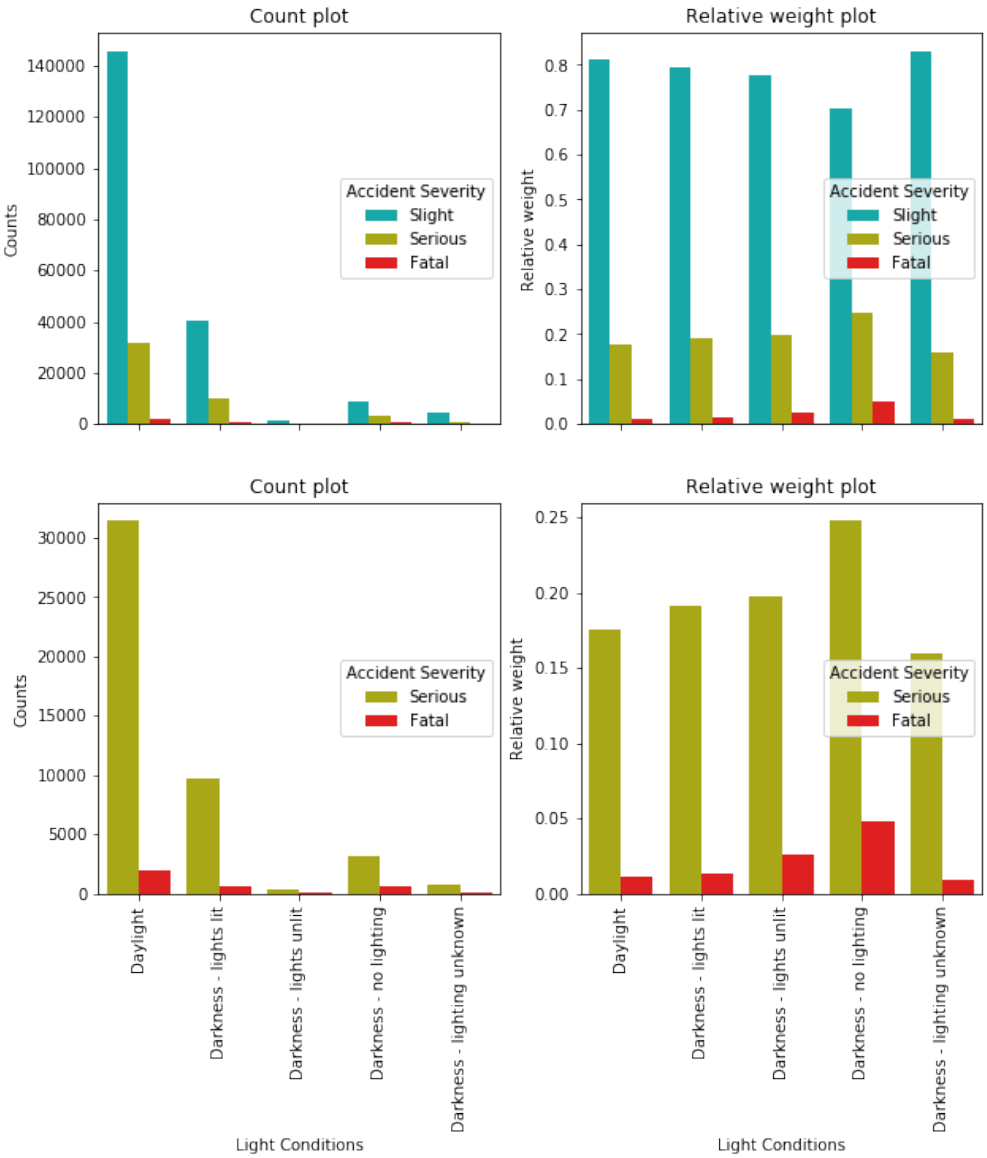
3.13 Carriageway Hazards



In case of hazards in the carriageway, the majority of accidents happened for objects or animals in the carriageway (note that most of accidents did not have hazards in the carriageway).
The relative weight plots showed that the presence of a previous accident had a higher relative weight of serious and fatal severity with respect to the other cases of carriageway hazards. High relative serious severity happened also in the case of general objects in the carriageway.

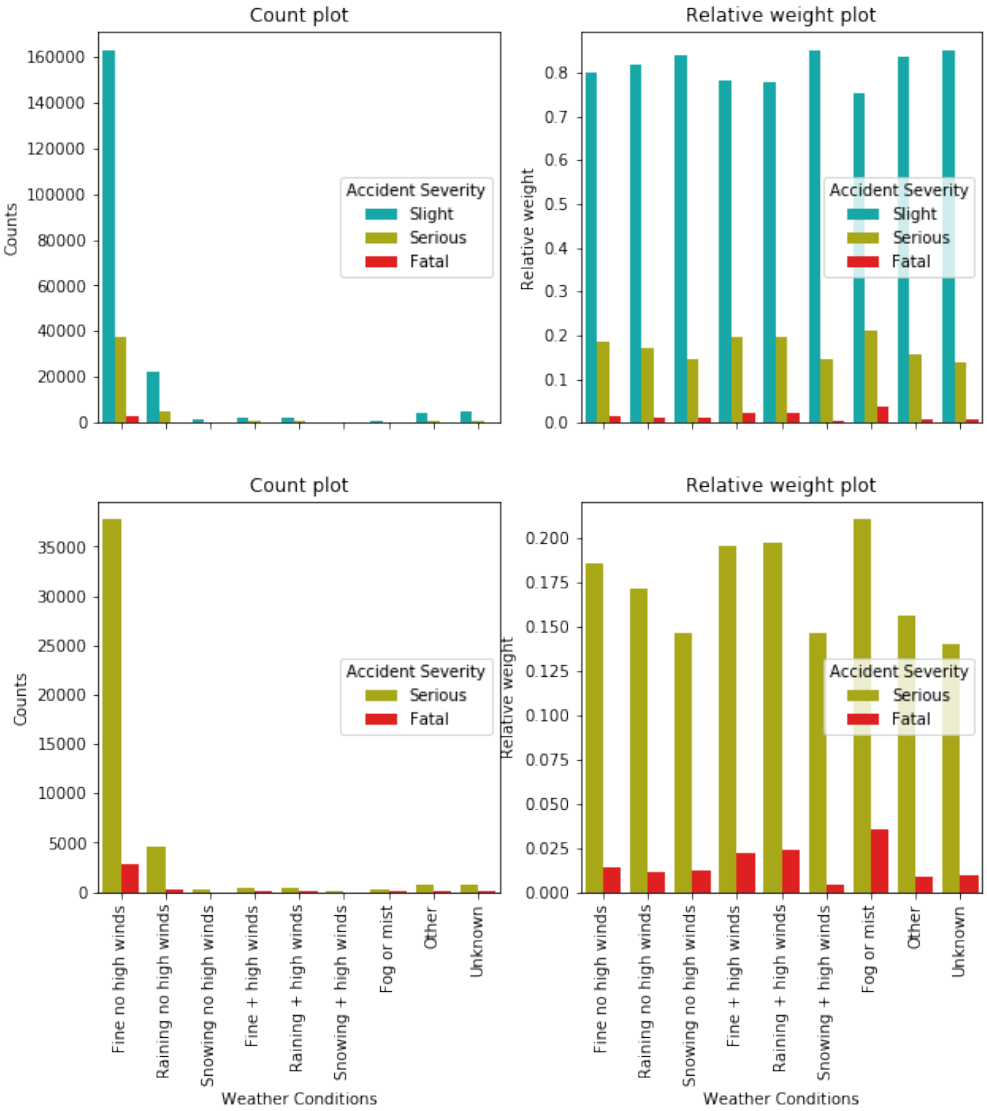
The cases of hazards in the carriageway had a relatively small number of statistical samples different from the case of 'No carriageway hazards', so statistical considerations on them could be biased.
This relatively small statistical sample led to the exclusion of this column in the model training phase.

3.14 Light Conditions



The majority of accidents happened during daylight.
The relative weight plot showed that accidents happening in presence of 'darkness' conditions had higher serious and fatal relative weights. In particular, the situation with the lowest visibility, i.e. 'darkness without lighting', had the highest relative weight for both serious and fatal severity. This confirms the previous observation of a slight higher severity risk for accidents that happened at night.

3.15 Weather Conditions



The majority of accidents happened for fine weather, the second most frequent case was rain.
The relative weight plot showed that the highest risk was for fog and high wind conditions ('fine with high winds' and 'rain with high winds').
The fact that the highest relative risk was in the case of fog highlights the danger of low visibility conditions.

4 Models

After the identification of the independent variables (risk factors), the target variable (i.e. 'Accident_Severity') and a preliminary data analysis and visualization, a series of predictive models were trained on the data.

The problem was a classification problem, since the target variable was a set of labels for accident severity: 'Slight' (3), 'Serious' (2), 'Fatal' (1). The original data already coded the different vari-

ables as numerical variables. Among the variables chosen for the model fitting phase, the only parameter with a real numerical ordering was the speed limit (and possibly time period and season), while for most of the other variables the numbers were simply different labels.
All these considerations led to the choice of classification models for the prediction.

For the classification predictor, the following base estimators were considered and trained:

- K-Nearest Neighbors,
- Decision Tree,

and the following ensemble methods:

- Bootstrap Aggregation (Bagging) Classifier,
- Adaptive Boosting (AdaBoost) Classifier,
- Random Forest Classifier.

Since the data are highly unbalanced, as described in Section 3, a simple `accuracy` score could not be used as the evaluation metric used for model scoring.
Indeed, since the majority class ('Slight' severity) covers roughly 80% of the cases, a simple model which for every input value returns only that majority class label would score 0.8 out of 1 with the `accuracy` metrics, but all the other classes (i.e. the "interesting" cases) would be completely misclassified.

The metrics chosen for model evaluation was the `balanced accuracy score`. It is defined as the average of recall (i.e. true positives divided by the number of real positives) obtained on each class.
Since the number of accident severity classes was 3, the classification model assigning all data to the majority class would score only 0.33 with such metrics. The `balanced accuracy score` ranges from 0 to 1, with 1 being the best score (i.e. all samples correctly classified).

As explained during the exploratory data evaluation, '1st_Road_Class' and 'Urban_or_Rural_Area' variables were dropped because of their correlation with 'Speed_limit', and in order to further simplify the model training, 'Carriageway_Hazards' was also dropped, because the number of serious or fatal accidents with hazards different from 'None' was very small compared to the total of cases.

The features over which the models were trained were ten: 'Road_Type', 'Speed_limit', 'Junction_Detail', 'Light_Conditions', 'Weather_Conditions', 'Road_Surface_Conditions', 'Special_Conditions_at_Site', 'Time_Period', 'Weekend', 'Season'.

4.1 K Nearest Neighbors (KNN)

The feature columns were scaled with a `StandardScaler` for the KNN Classifier fit, which uses a distance-based metrics.
The number of neighbors (K) parameter was tuned with cross validation (4 folds) on the full set of training data. The scoring metrics was the `balanced accuracy score`.

The average values of the scores from the 4 different folds of the cross validation had a peak at K=2, i.e. two neighbors. The average score value was quite low for this classifier, around 0.35.

The model was trained with these parameters: `n_neighbors=2`.

4.2 Decision Tree

For the Decision Tree and the ensemble classification models, particular attention was paid to the unbalancing of data classes.
In particular, for the Decision Tree the option `class_weight='balanced'` was used. This options weights the contribution of the samples by assigning to each sample a weight inversely proportional to its class frequency (i.e. total count of samples in that class).
This situation is roughly equivalent to the *oversampling* of the data belonging to the least populated categories, since the weighted numbers of the samples belonging to different classes becomes the same (for randomly extracted data), without dropping data from the most populated classes.

The number of branches, i.e. `max_depth` parameter, was optimized with cross validation (5 folds) on the full set of training data. The scoring metrics was the `balanced accuracy score`.

The average values of the scores from the 5 different folds of the cross validation had a peak at number of branches around 6-9, with corresponding average score around 0.46.

The model was trained with these parameters:
`criterion='entropy', max_depth=7, class_weight='balanced'`.

4.3 Bootstrap Aggregating (Bagging) Classifier

A Bagging classifier is an ensemble predictor that fits a series of base classifiers each on random subsets of the original data and then aggregates their individual predictions in a final prediction.
For the Bagging Classifier case, the underlying base estimator was declared as a Decision Tree with `class_weight='balanced'`, to overcome the problem of strong data unbalancing; this is roughly equivalent to an oversampling of each of the random subsets considered during model training.
All input data features were considered in the fit.

The number of estimators parameter `n_estimators`, i.e. the number of base estimators in the ensemble, was tuned with cross validation (4 folds) on the training data. The values of mean balanced accuracy was not changing much with `n_estimators` in the range 100-400, with balanced accuracy score around 0.406-0.407.

The model was trained with these parameters:
`base_estimator = DecisionTreeClassifier(class_weight = 'balanced'), max_features = X.shape[1], n_estimators = 300`.

4.4 Adaptive Boosting (AdaBoost) Classifier

An AdaBoost classifier is an ensemble predictor using an ordered series of a base classifier, that are trained on the original data, in which subsequent classifiers are trained on the dataset with adjusted weights to increase the impact of the cases misclassified by the previous classifiers.
Analogously to what was done for the Bagging Classifier, in order to overcome the unbalancing of the dataset, the base model for the AdaBoost classifier was chosen as a Decision Tree with `class_weight='balanced'`.

The number of estimators parameter `n_estimators` was tuned with cross validation (4 folds) on the training data. The values of mean balanced accuracy was not changing much with `n_estimators` in the range 40-80, with balanced accuracy score around 0.403-0.404.

The model was trained with these parameters:
`base_estimator = DecisionTreeClassifier(class_weight = 'balanced'), n_estimators = 70`.

4.5 Random Forest Classifier

A Random Forest Classifier is an ensemble predictor which creates and fits a number of decision tree classifiers on sub-samples of the data and averages (or takes the mode of) their classifications. In this study the Random Forest was used together with `class_weight='balanced'`, so that each sample was weighted inversely proportional to its class frequency. Similarly to the Decision Tree case, this is roughly equivalent to oversampling the classes with low counts.
The parameter `max_features` was left at its default value, i.e. square root of the number of features.

The number of estimators `n_estimators` and maximum depth of the tree classifiers `max_depth` parameters were tuned with a grid cross validation (4 folds) on the training data.

The best parameters found with the grid search were around 6-7 for `max_depth` and 50-70 for `n_estimators`, with balanced accuracy around 0.46-0.47.

The model was trained with these parameters:
`class_weight = 'balanced', max_depth = 7, n_estimators = 70`.

4.6 Undersampling Models from imblearn

In addition to the previous models, which mainly used an artificial oversampling of the data, two ensemble models based on *undersampling* of the most populated classes were used:

- Balanced Bagging Classifier;
- Balanced Random Forest Classifier.

These models were imported from library `imblearn.ensemble`.

4.6.1 Balanced Bagging Classifier

The Balanced Bagging Classifier from library `imblearn` is similar to the `sklearn` implementation, but it uses an additional step to balance the training set at fit time with `RandomUnderSampler`.

The number of estimators parameter `n_estimators` was tuned with cross validation (4 folds) on the training data. The values of mean balanced accuracy was not changing much with `n_estimators` in the range 150-400, with balanced accuracy score around 0.43.

The model was trained with these parameters:
`max_features = X.shape[1], n_estimators = 200`.

4.6.2 Balanced Random Forest

Similarly to the Balanced Bagging, the Balanced Random Forest from `imblearn` package balances the class distribution for each bootstrap with a random undersampler.

The number of estimators `n_estimators` and maximum depth of the tree classifiers `max_depth` parameters were tuned with a grid cross validation (4 folds) on the training data.

The best parameters found with the grid search were around 7-9 for `max_depth` and around 40-80 for `n_estimators`, with balanced accuracy around 0.46-0.47.

The model was trained with these parameters:
`max_depth = 8, n_estimators = 70`.

5 Test the Model

As explained in the data acquisition and preparation section 2, the accident database from 2016 was used to test the trained models.

In order to run the predictive models on the 2016 data, these data were prepared similarly to what was done to the training dataset, namely:

- Replace -1 values with `NaN`;
- Drop unused columns;
- Extract month value;
- Extract hour value;
- Fill `NaN` with 0 for columns 'Special_Conditions_at_Site', 'Carriageway_Hazards', 'Junction_Detail';
- Drop rows containing `NaNs`;
- Transform hour and month to time period and season;
- Transform day of the week to weekend indicator;
- Extract the features considered for these models;
- Scale with `StandardScaler` the feature set for the KNN Classifier.

After data cleaning, 135815 samples (roughly 99.4% of 2016 raw data) were present in the 2016 processed dataset.

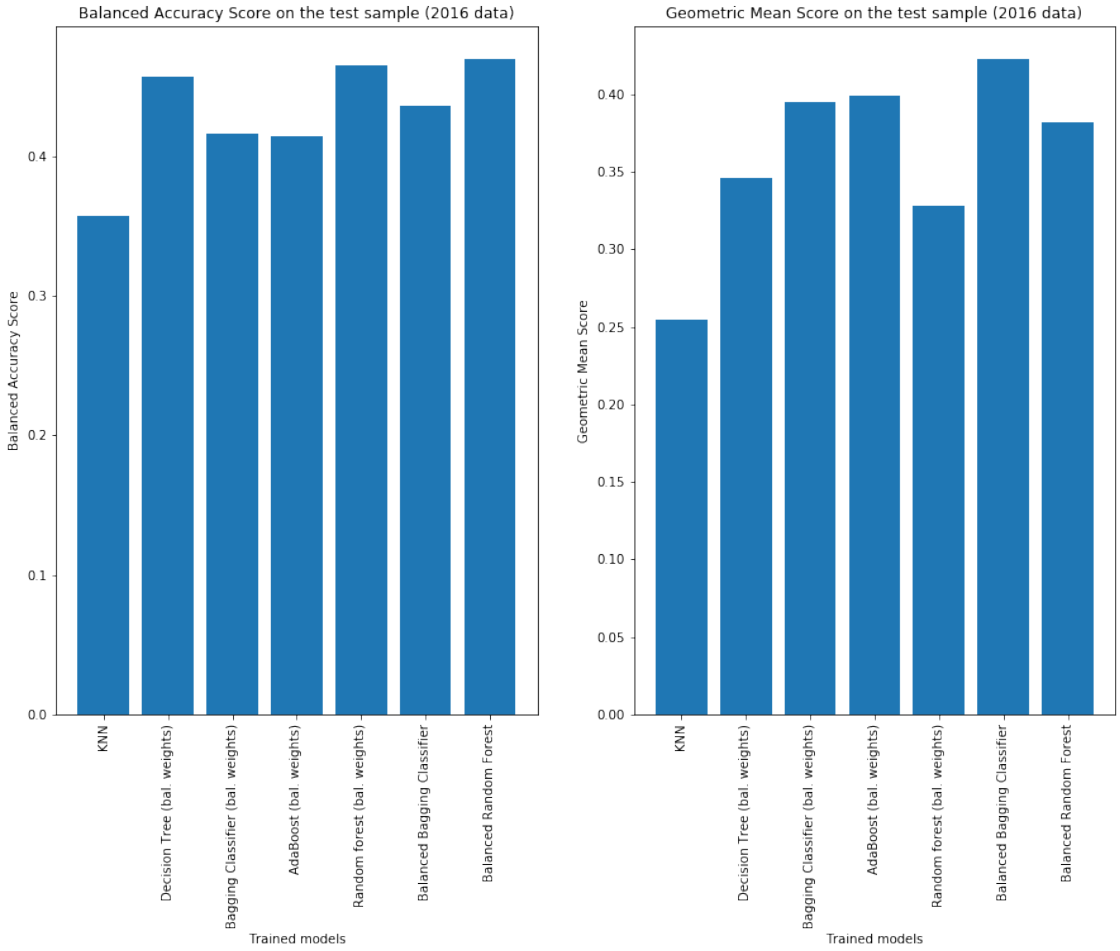
6 Results and Discussion

The trained models were used on the 2016 dataset.
In order to verify the ability of accident severity prediction, two metrics were considered:

- Balanced Accuracy Score from `sklearn.metrics`
- Geometric Mean Score from `imblearn.metrics`

As already explained in the model training section, the Balanced Accuracy Score is the average of recall scores per target class.
The Geometric Mean (G-Mean) score is the Nth-root of the product of class recalls, with N the number of classes.
Both of these metrics score from 0, the worst value, to 1, the best value. The Balanced Accuracy scores 1/N (with N the number of classes) for a model which predicts only one single class output; the Geometric Mean scores 0 if at least one of the class is unrecognized (i.e. 0 predicted counts) by the model (for example, the dummy predictor which predicts only the majority class).

The following figure summarizes the scores for the different models trained in this study, applied to the 2016 dataset.



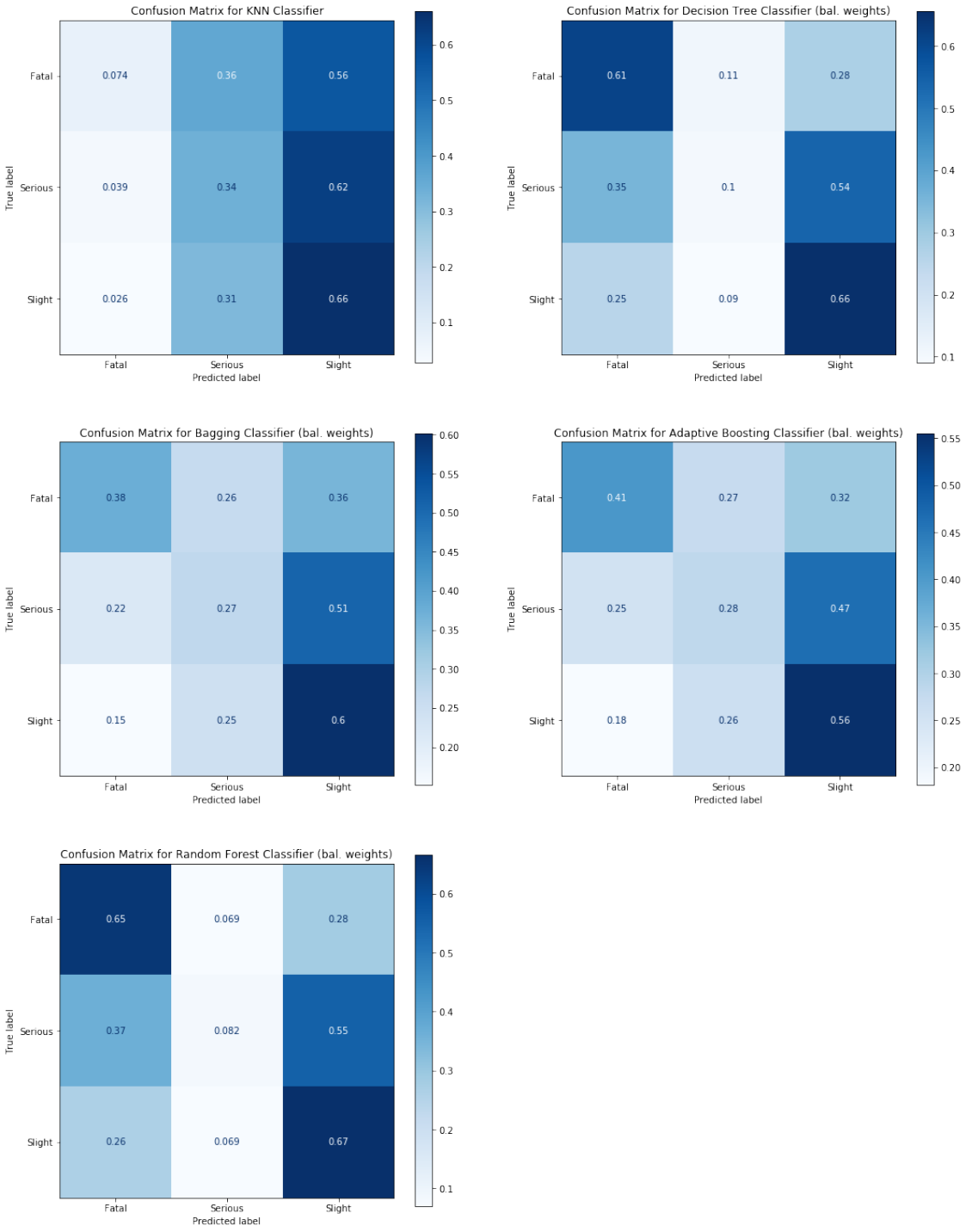
The KNN model scored 0.36 with Balanced Accuracy, only slightly better than a majority class predictor. It scored 0.26 with G-Mean.

For what concerns the Balanced Accuracy, all the other models scored slightly more than 0.4, with Decision Tree and Random Forest models (Balanced and with Balanced Weights) scoring between 0.46 and 0.47. The Bagging and AdaBoost scored around 0.42, while the Balanced Bagging scored roughly 0.44. The best score was obtained by the Balanced Random Forest Classifier with 0.47.

With the G-Mean, the highest score was obtained by the Balanced Bagging Classifier with 0.42, followed by the Bagging and AdaBoost Classifiers around 0.40. The Balanced Random Forest scored 0.38, the Decision Tree 0.35 and the Random Forest with Balanced weights 0.33.

All the scores were not extremely high. The models which performed best, according to these metrics, were the Balanced Bagging and the Balanced Random Forest Classifiers, i.e. the two *ensemble* models with data *undersampling*.

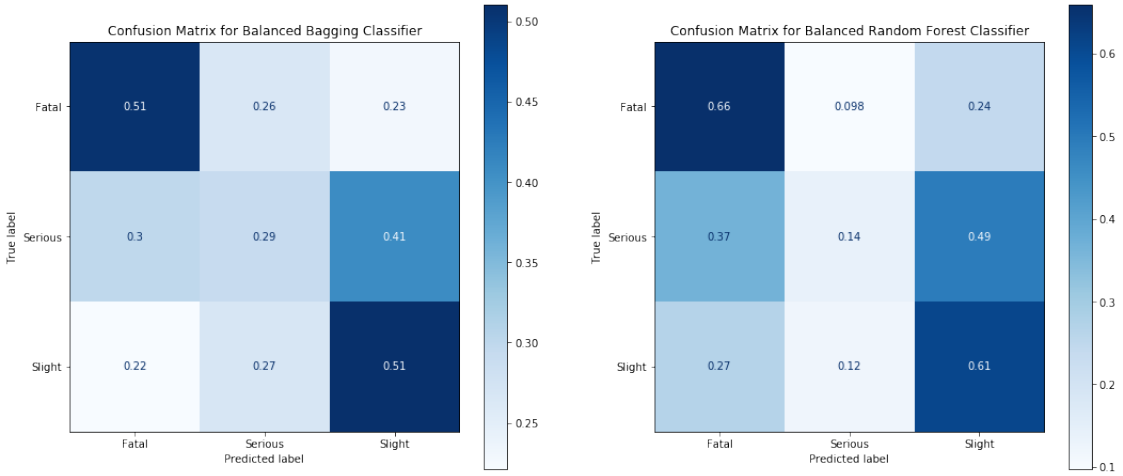
The confusion matrix for the models on the 2016 data are shown in the following figure, where each cell was normalized with respect to the total number of actual true cases for that label.



As seen in the confusion matrix, the behavior of KNN Classifier was skewed towards a majority class predictor: roughly 60% of the samples of each class were predicted as the most frequent label ('Slight').
Bagging and AdaBoost Classifiers performed better in the prediction of the different classes. However, actual 'Serious' class was still mostly predicted as 'Slight' (50% of cases), and correctly predicted only 27-28% of times; true 'Fatal' were correctly predicted only roughly 40% of times.
The Decision Tree and Random Forest with balanced weights performed better for both 'Slight'

and 'Fatal' classes, reaching relative accuracy (recall) around 61-67% for both classes. However, actual 'Serious' cases were still missclassified roughly 50% of times as 'Slight' and 40% as 'Fatal'.

The following figure refers to the ensemble model with undersampling.



The Balanced Bagging was the one with best performances (except for the poorly labeling KNN) on the 'Serious' class, with correct predictions for roughly 30% of actual true cases; it reached a recall around 50% for both 'Fatal' and 'Slight' cases.
The Balanced Random Forest performed similarly to the Decision Tree and Random Forest with balanced weights, with recall of 66 and 61% for 'Fatal' and 'Slight', respectively. It still wrongly predicted 49% of true 'Serious' cases as 'Slight', and the recall for 'Serious' was around 14%.

6.1 Balanced Models for 2 target classes

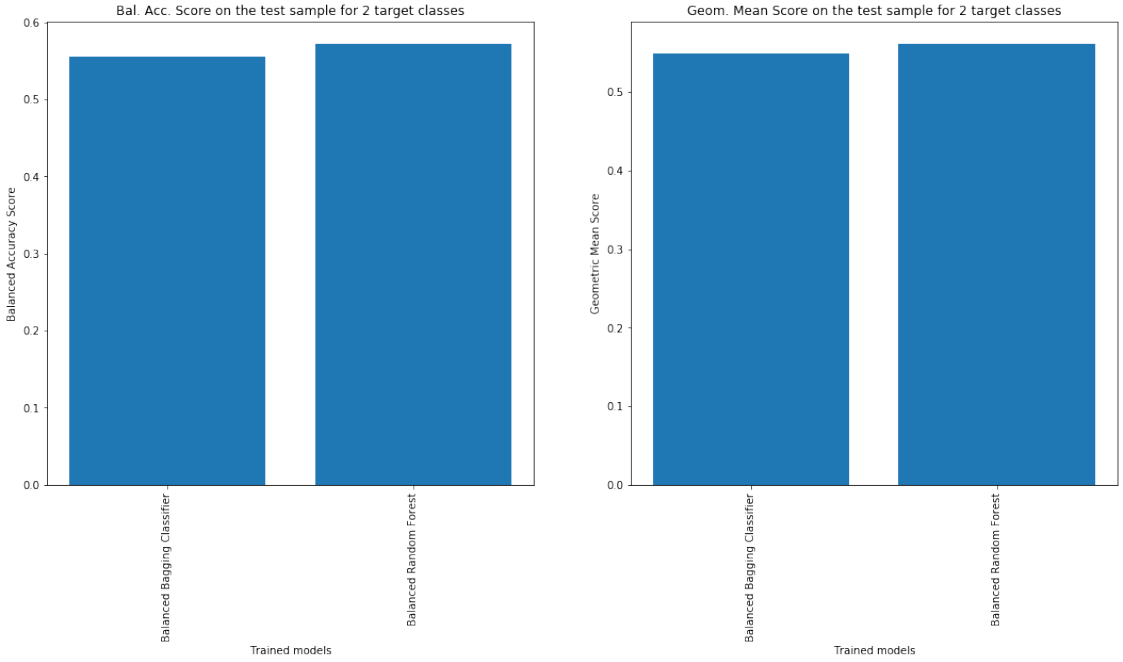
Given the performances of these models, as a tentative solution to improve the classification, the problem was further simplified by merging together the 'Fatal' and 'Serious' target classes, since these cases could both result in high cost for the society.

Only the Balanced Bagging and the Balanced Random Forest Classifiers were trained on the simplified dataset, since these were the best performing models in the previous analysis.

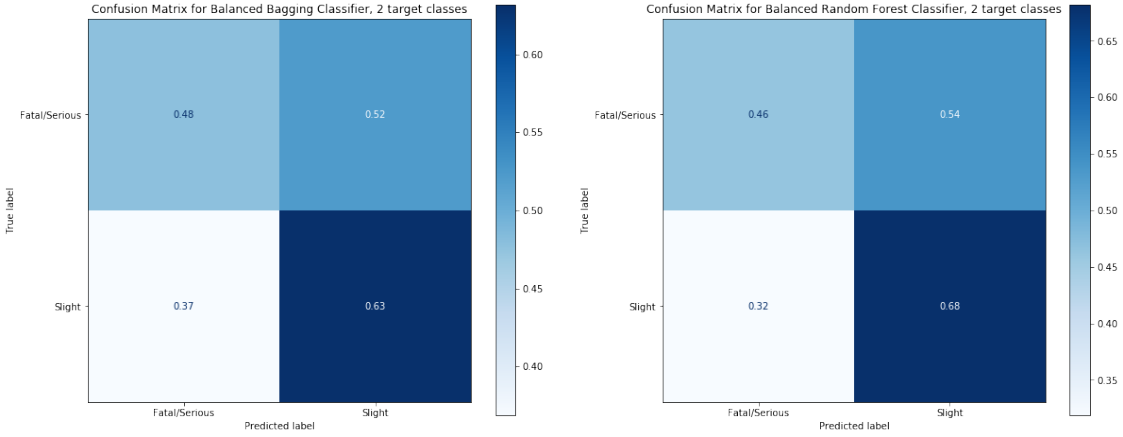
For the Balanced Bagging, the number of estimators parameter `n_estimators` was tuned for the Balanced Bagging with cross validation (4 folds) on the new training data. The values of mean balanced accuracy was not changing much with `n_estimators` in the range 50-400, with balanced accuracy score around 0.54. The model was trained with these parameters:
`max_features = X.shape[1], n_estimators = 100`.

For the Balanced Random Forest, the number of estimators `n_estimators` and maximum depth of the tree classifiers `max_depth` parameters were tuned with a grid cross validation (4 folds) on the training data. The best parameters were around 7-9 for `max_depth` and around 30-80 for `n_estimators`, with balanced accuracy around 0.57. The model was trained with these parameters:
`max_depth = 9, n_estimators = 30`.

The following figure shows the Balanced Accuracy and Geometric Mean scores for these two models, tested against the 2016 data (where 'Fatal' and 'Serious' classes were combined).



Both models performed similarly, with a Balanced Accuracy around 0.56-0.57 and G-Mean around 0.55-0.56. The Balanced Random Forest was performing slightly better than the Balanced Bagging. The new confusion matrices are shown in the following figure.



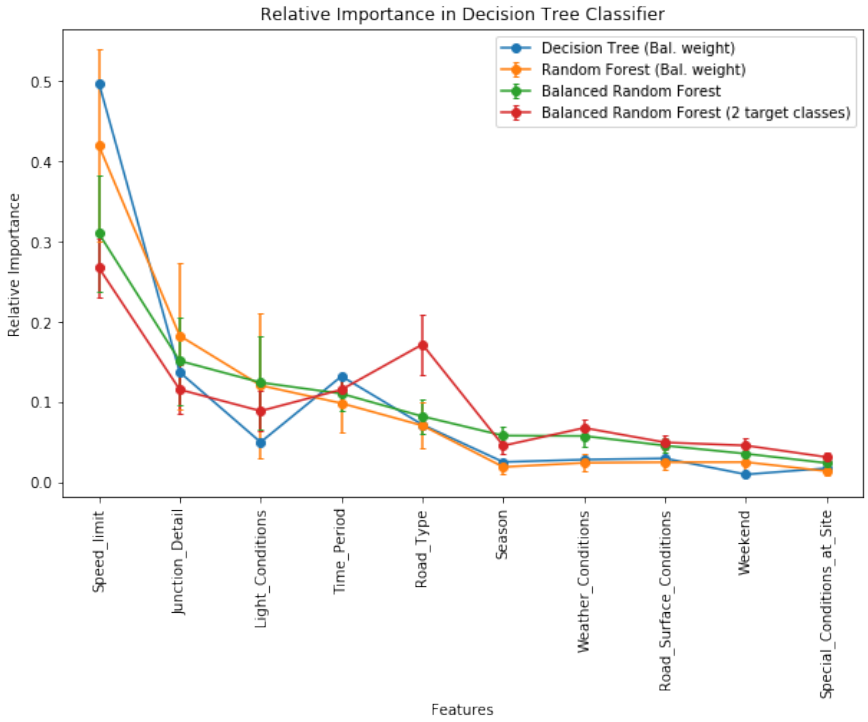
As confirmed by the scores, the outputs of the models were similar.
Roughly 63-68% of true 'Slight' accidents were correctly predicted, but only 46-48% of true cases belonging to the combined class 'Fatal or Serious' accidents were correctly identified.

6.2 Feature importance

As an additional step, the use of tree classifiers enabled the possibility of ranking the input feature in order of importance in the classification process.
In general, this ranking enables the selection of the most important features for the discrimination between the cases, since it ranks the contribution in reducing the impurity of branches.
For the present problem, such an analysis could help in identifying the features for which a most evident distinction was present between slight, serious and fatal accidents, and somehow deter-

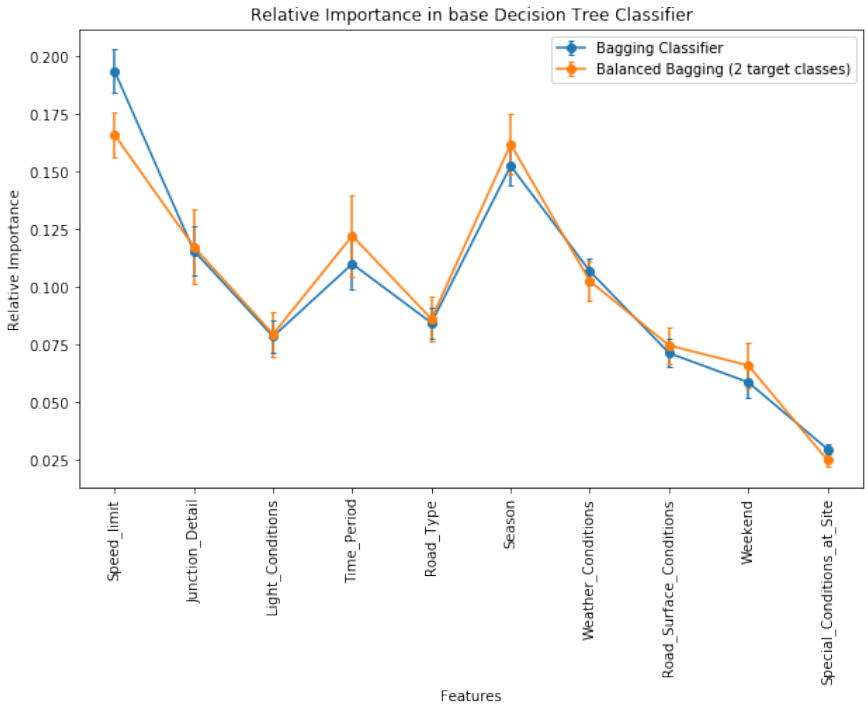
mine the feature which had a relatively "most dangerous" or "most safe" situation.

The following figure shows the relative importance of the features for the Decision Tree and Random Forest models.



The most important feature was speed limit. This confirms the initial observation (Section 3) about the relative weight of serious and fatal accidents increasing with increasing speed limit, and thus probably the speed at which the accidents happened.
The next important features were junction detail, light conditions, time period and road type. This confirms the fact that visibility (light conditions, time period) was an important factor in the severity of the accidents.
It also points out that the type of junction was a discriminant factor for severity prediction. This means that some kind of junctions could be relatively more dangerous than others. For example, T or staggered junctions and slip roads had high relative weights of serious and fatal accidents, respectively. Roundabouts were the junctions with the lowest relative weight of both serious and fatal accidents.
The Random Forest trained on the 2 classes dataset showed "road type" as the second most important feature; even in this case, roundabouts had the lowest relative weight of serious and fatal accidents.
Special conditions at site, weekend label, road surface conditions, weather conditions and season had lower relative impacts in the classification.

The following figure shows the average relative importance of the feature for the base Decision Trees used in the Bagging Classifiers.



Both Bagging Classifiers had similar feature importance rankings. The most important was speed, like in the previous graph. However, season scored second with these algorithms; from the relative weight graph in Section 3 the correlation between season and accident severity was not clearly evident.
Time period, junction detail and weather conditions followed in the importance ranking.
Special conditions at site, weekend label, road surface conditions were still at the bottom of the ranking order.

7 Conclusions and Perspectives

In conclusion, this study analyzed UK accidents data from 2017 an 2018, and trained a series of models to predict the severity of accidents. It concentrated on the ability of the model to correctly predict critical accidents, i.e. fatal and serious accidents.

It showed that the best performances, among the models tested during the study, was reached by undersampling ensemble models, namely Balanced Bagging and Balanced Random Forest. However, no model could achieve recall values larger than 50% for all the classes at the same time, even when aggregating the serious and fatal labels into one single label.

It showed the relative importance of features retrieved by the trained models: speed limit, visibility conditions and junction types had the most relative weight in accident severity classification.
In particular, higher speed limits had a relatively larger chance of high severity in accidents.
Bad visibility was shown to be an important factor increasing the probability of high severity.
The evaluations on junction types suggested roundabouts as the relatively (to the total number of accidents happening in correspondence to them) most safe type of junctions, while T or staggered junctions and slip roads were the relatively most dangerous ones.

Possible future development for this study could be the generalization of the classification for

datasets divided into the different kind of road users involved in the accidents, for example drivers, bikers, cyclists, pedestrians etc, included in the additional tables available from the UK open data site.
Identifying the most dangerous situations for each road user category could help Governments and police forces in targeting focused prevention measures. Moreover, it could trigger industries working in the field of safety to work on improved security devices for the identified particularly dangerous conditions for those road users.

Other possible future developments could be the use of different classification models, such as Gradient Boosting, or the inclusions of features which were dropped in this study.

Another possible development could be the analysis of the evolution of risk factors during the years, to determine the efficacy of prevention measures and vehicle safety improvements.

It must be noted that on top of the *relative weight* of factors in the severity of the accidents, *absolute counts* (i.e. count histograms) are very important, since the total cost for the society derives from a combination of both relative weight of severity, and happening frequency.