Coursera Capstone Project report: GB car accident analysis 2018

Devansh Gadodia

Abstract

Streets accidents are responsible for remarkable psychological and economic consequences on society in the UK and, more in general, the world. Recent cuts to the budget require the UK Department of Transport to choose a field to invest in order to decrease traffic accidents. The development of a machine learning model based on accidents data provided by the Stats19 form can help in solving this problem, serving also as a useful tool for predicting the risk of a fatal accident given a series of informations normally in posses- sion of insurance companies.
All the classification models employed didn't lead to a satisfactory accuracy level and require further re- finement and data collection; however some possible counter-measurements to apply in the short term were found, such as the increment of traffic stops (especially for motorbikes with more than 125 cc) and measures aimed to increase visibility and better regulate traffic flow in correspondence of areas with high pedestrian density and road junctions.

Keywords: Car, accident, Great, Britain, UK, machine, learning

Contents

* Corresponding author

1. Business Understanding

Street accidents are part of the major public health problems in the world, affecting both economically developed and developing country. The association with high rates of mortality and morbidity is strong and it translates into a huge economic burden for society. If we have a look at middle- and high-income countries, the striking difference is that, although the number of vehicles involved in accidents is glob- ally the same, low-income countries represent 80% of the total deaths, while the latter only the 7%. Of course, strict regulation of driving rules and the process of accurate street maintenance aimed to mitigate the number and the gravity of accidents didn't start immediately, but as the number of vehicles per per- son increased and the society became very interconnected. Not only middle-income countries have more limited resources to put into safer infrastructures, but their society is continuously changing at a a speed that most of high-income countries reached only in the past. However, beside a safer way to design streets and other structures, the *human factor* seems to be the most prominent risk factor: the World Health Or- ganization indeed stated that "Strong policies and enforcement, smart road design, and powerful public awareness campaigns can save millions of lives over the coming decades."(1).
Despite being one of the safer country in Europe regarding the number of deaths of car accidents over the population, the UK is interested in further decreasing road casualties. Unfortunately, the EU budget for road safety measures has been cut significantly in recent years(2), so this big resources limitation requires to understand the right field of intervention that will lead to the biggest reduction in street causalities.
Given this limitation, the initial question that the local authorities and the UK Department of Transport would ask is "in which area should we invest more to decrease the economic and social burden caused by streets accidents?".
Accordingly to the form used to report a street accident (Stats19), its severity can as- sume three distinct and hierarchical values: "slight", "severe" and "fatal". It is obvious that a fatal accident has more profound an negative consequences than one or more slight accidents. The machine learning model that the stakeholders asked to build should therefore distinguish the three categories (with a gra- dient of severity) on the basis of a series of features belonging to different domains, so that it would be possible to unravel their relative contribution to the three classes and take appropriate countermeasures. In addition to people working directly in the field of streets safety, this model could be useful also for insurance companies, hospitals and regional administrators: knowing if fatal or severe accidents happens more frequently in rural areas (and if this feature is very "important" for correctly predicting fatal or severe accidents) translates into a less centralized system of street-rescue and in an increment of hospital services located far from big cities; vehicle- and life-insurance policies should also be revised accordingly to the new risk factors individuated.

2. Data

I collected three different databases storing informations and details of car accidents occurred in UK in 2018. The data was collected using the STATS19 accident reporting form and carried different informations with some redundancy:

- `dftRoadSafetyData_Vehicles_2018`: data corresponding to technical info of every vehicle involved in the accident (such as "age", "type of vehicle", "engine capacity" . . .), but also of the driver (like band of age and IMD decile)
- `dftRoadSafetyData_Casualties_2018`: data corresponding to individual informations of the casu- alties the casualties such as'' sex", "severity", "age", "role at the moment of the accident" (driver, passenger or pedestrian) . . .
- `dftRoadSafetyData_Accidents2018`: data corresponding to general informations regarding the site of the accident, such as "road conditions", "light conditions" and "weather conditions".

The 2019 dataset contained records not older than June, so, to avoid incomplete data or bias due to the absence of variables linked to a given season, I decided to use the 2018 dataset instead. Another implicit

Feature	Dataset	Class	Reason for inclusion in the model
Vehicle type	Data Vehicles 2018	Technical factor	Different vehicles can show different accident risk
Skidding and Overturning	Data Vehicles 2018	Dynamic factor	Can be a proxy of bad road conditions or weather conditions
First point of impact	Data Vehicles 2018	Dynamic factor	Frontal impacts can be different in the severity if compared to back impacts (reachable speed is different) Engine
Capacity		Technical factor	For each vehicle category, is a proxy of the engine size and therefore of the vehicle weight
Age band of driver	Data Vehicles 2018	Human factor	Young and old age driver could show higher accident risk Age of
Vehicle	Data Vehicles 2018	Technical factor	Can highly affect accident occurrence and gravity
Sex of driver	Data Vehicles 2018	Human factor	Data needs to be checked for this important factor Casualty class
	Data Casualties 2018	Dynamic factor	Allows distinction in accidents with/without pedestrians Speed limit
	Data Accidents 2018	Dynamic factor	Proportional to the road-associated risk
Junction detail	Data Accidents 2018	Technical factor	Can highly affect accident severity
Light Conditions	Data Accidents 2018	Dynamic factor	Affecting visibility and therefore accident risk; is a mixed feature because depends on natural and artificial light Weather
conditions	Data Accidents 2018	Natural factors	Can highly affect accident risk
Urban or rural area	Data Accidents 2018	Dynamic factor	Influencing vehicle density and other factors

Table 1: Initial selection of features for different databases

assumption of my choice was that statistics of car accidents sampled every year should be very similar to the ones sampled in the same decade. For this reason, the reliability of the model should also be checked in the long term.

The three database have important informations that, in the light of the questions I need to ask, could be defined as *complementary* between each other. In order to build a good model and to address the prob- lems described in the section1, I need to include different kind of parameters that may deeply affect the probability that a serious accident will take place. In particular, I have identified four main feature classes:

1. Natural factors: they can deeply decrease road safety, independently of the driver's skills, and are not always predictable. Even if largely anticipated, the human intervention aimed to counteract them can be very limited, and is therefore restricted to road maintenance and safer traffic rules (e.g. lower speed limits)

2. Human factor: Despite ideal road and weather conditions, distractions, progressive deterioration of driving skills due to ageing and inexperience can be determinant in causing a serious accident

3. Technical factor: all the factors linked to the proper working of the means of transport ("vehicle age", "vehicle type" . . .) their characteristics ("engine capacity", "propulsion code" etc.) or the road configuration ("urban/rural location", "road surface condition" etc.).

4. Dynamic factor: all the factors related to the car accident dynamic (slippery of the car/motorcycle, points of impact, role of the person involved . . .).

As stated previously, the principal aim of the project is to individuate, if possible, a general strategy for decreasing the occurrence of severe or fatal car accidents in the UK (which, instead, is increasing year on year(3)). Considering the huge social and economic burden deriving from it, the priority is given to the fatal and severe ones. The nature of the factors influencing the probability that a severe or fatal car accident manifests is heterogeneous: weather can't be controlled, and the ageing process can't be stopped, but a lower speed limit and more restriction in terms of driving licence renewal can make the difference. Then, the next step is to select all the features that are directly and indirectly related to all the possible field of intervention. Ideally, the selected features (showed in the Table1) would be equally representative of the four different domains previously described.

3. Methodology

"Data casualties" and "Data vehicles" databases have more records than "Data accidents"; this is due to the fact that they store informations at the level of single casualty or vehicle, respectively. In order to merge these two databases with the "Car accidents" data, I needed to condense their informations to the level of single accident (more technically, grouping their features by `Accident_index`). The first step was then to select from the three databases only the features needed for our model; to properly match the specific row during the merging process of "Data casualties" and "Data vehicles", the features `Accident_Index` and `Vehicle_Reference` were also kept.

After having imported as dataframes the three different dataset with the chosen features listed in the table 1, the presence of incomplete data was solved by deleting all the rows with a missing value in at least one column. Once collected the data, this was visually explored to better understand if and how different features could be transformed to clearly describe some inner pattern of car accidents or just simplify the interpretation by the investigators/stakeholders.

3.1. Data elimination after exploratory analysis

Many features were not included in the very first stage of the model because exploratory analysis re- vealed that there was an equal repartition among the three severity categories. Surprisingly, the exploratory analysis revealed that:

- The street accidents frequency among different days of the week was almost identical
- The street accidents frequency among single or double carriageway was very similar
- More than 95% of car accidents occurred in absence of control of pedestrian crossing
- The street accidents frequency among the three different severity categories were the same in both sunny and rainy days (which formed more than 90% of the total cases). The feature "Weather condi- tions" was therefore not considered for building the model.
- The street accidents frequency among different road conditions was almost identical
- More than 96% of the street accidents took place in streets devoid of special conditions (such as roadworks, defective road signs . . .)
- More than 97% of the street accidents took place in streets devoid of carriageway hazards (such as previous accidents, animals on road . . .)

3.2. Data simplification and merging

Feature merging or combinations can reveal hidden patterns lying inside certain categories or can create new informations. For instance, the feature `Casualty_Class` of the "Casualty database" had initially three different values corresponding to the role of the person involved in the accident:

- 1: Driver
- 2: Passenger
- 3: Pedestrian

After having applied the One Hot Encoding on it, a binary code for each casualty was obtained, carrying as information the presence or absence of drivers, passengers or pedestrians. Counting all the records for each class within each single accident and filtering for class number 3, allowed the creation of the new feature "Pedestrian involvement", which refers to the number of pedestrian involved in a given accident. Regarding the other features, the following modifications were applied:

- *Vehicle Type*: "Car" and "rented-car/taxi" categories were merged into a general "Car" category. I then grouped the other categories into "Pedal cycle", "Motorcycle up to 125" cc, "Motorcycle up to 500 or more". The other categories were "Unknown" and ''Other". As for all the other features listed below, "other" and "unknown data" were excluded from the analysis.
- *Age Band of Driver*: Four different categories were created, "Age to small for having driving license" (0-15 years old), "Young Driver" (16-35), "Intermediate Driver" (36-65) and "Old Driver" (older than 65). The latter one is the one I was more interested in because naturally linked to the loss of reflexes and attention.

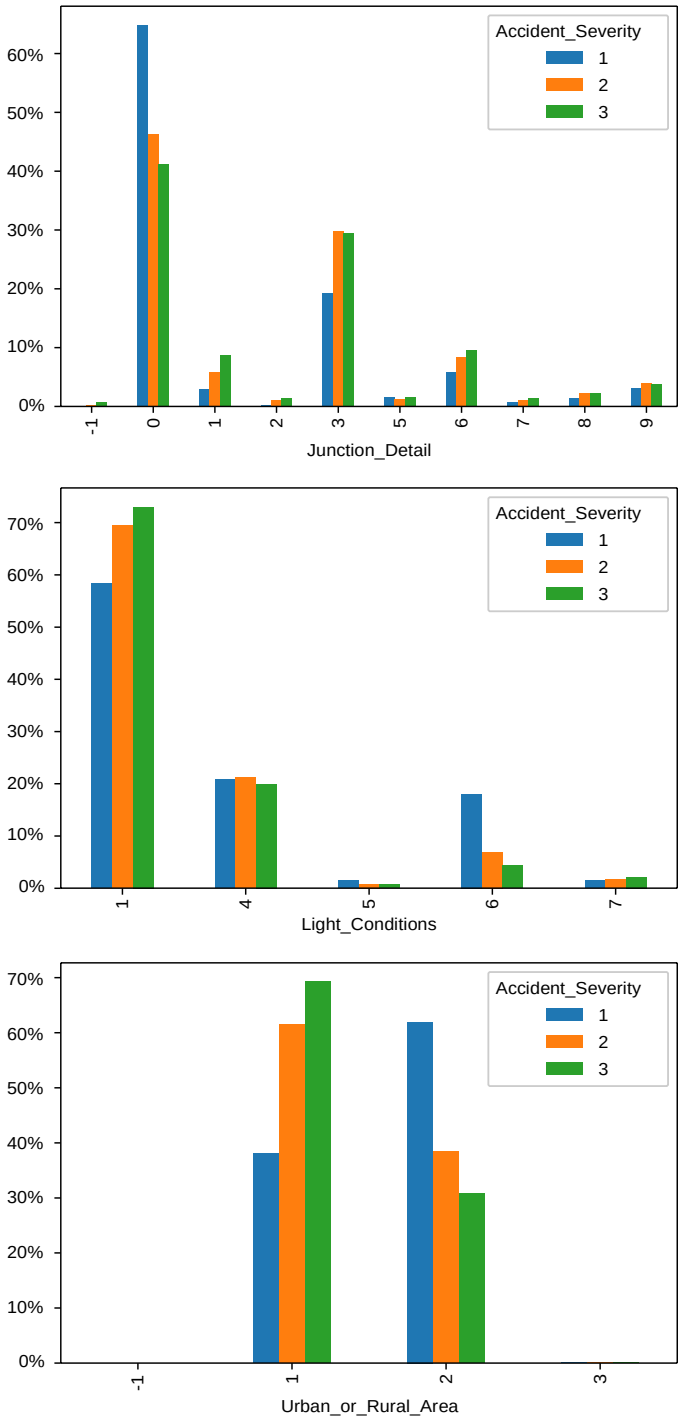

Figure 1: Bars displaying the percentage of accidents distributed along different classes of the features of interest and color coded accordingly to class severity (blue = fatal, orange = severe and green = slight.

- *Age of Vehicle*: the vehicle was considered as "new" if not older than 10 years, "middle" if not older than 20 years and "old" if older than 20 years. Old vehicles are more susceptible to damage or

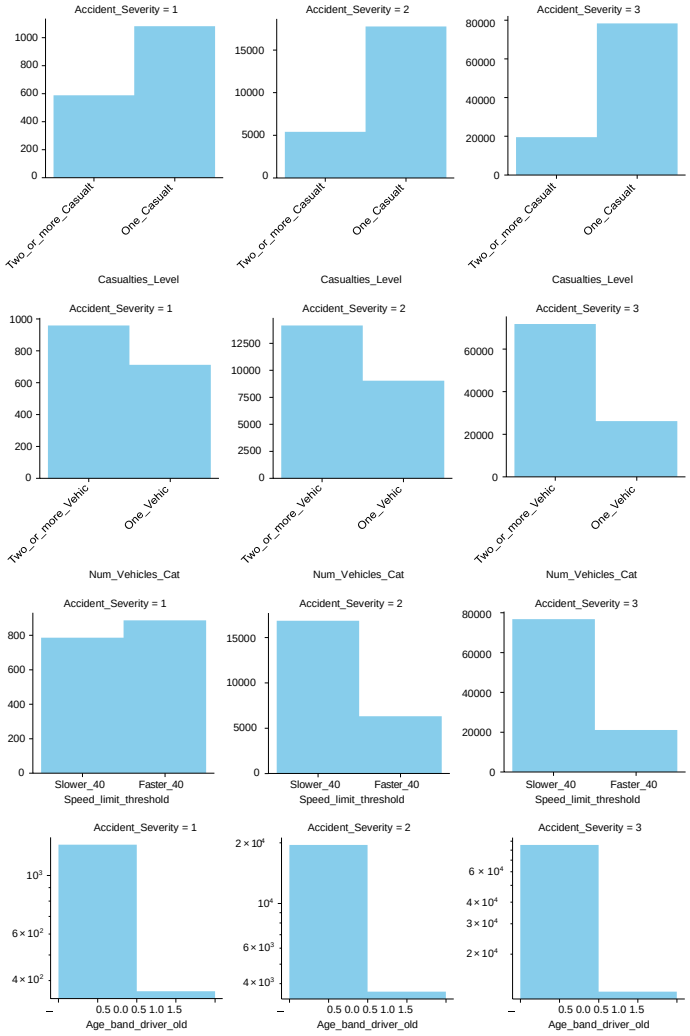

Figure 2: Counts distributions after the reduction of the number of levels within “Casualty Number”, “Vehicle number”, “Speed limit” and “Age-band of the driver”. As we can observe, when compared to the other two classes, fatal events show a higher percentage of two casualties and vehicles involved in the same accident, more accidents occurring at lower speed limits and less old drivers.

dysfunctions.

- *Engine Capacity*: The engine capacity was divided in three categories: “normal” (up to 1200 cc), “high” (up to 2000 cc) and very high (more than 2000 cc).
- *Skidding and Overturning*: Almost 70% of the vehicles involved in car accidents had no skidding or overturn reported; so it was decided to join the “Skidding” and “Skidding and overturning” cate- gories into a binary one: Skidding. Overturning informations was lost because poorly representative (21% of skidding vehicles in fatal accidents *vs* 7% of jackknifed).
- *First Point of Impact*: The first impact location of the crash varied consistently among severity cate- gories, with most of the slight accidents occurring while hitting the back of the vehicle. It was decided to preserve only the information regarding impacts at the front and back of the vehicles.

Accident Severity	Counts before balancing	Percentage before balancing	Counts after balancing	Percentage after balancing
Slight	97,799	79.7%	13,000	42%
Severe	23,165	19.9%	9,000	29%
Fatal	1,671	1.4%	9,000	29%

Table 2: Number of street accidents for different categories before and after the balancing.

The One Hot Encoding solved the problem of eliminating missing data in this process since missing data is simply classified as 0. In all the cases, with the exception of "Engine capacity" and "Age of vehicle", the amount of missing data was very small so the consequences of classifying it as 0 was minimal.
Regarding the Data accidents dataset, the following modifications were done:

- *Number of Vehicles*: the more the severity of the car accidents, the more frequent was the involvement of a single vehicle. To simplify the data and preserve this information, two categories were created, one for accidents involving one vehicle and the other for accidents involving two or more vehicles. The new feature is called `Num_Vehicles_Cat`.
- *Number of Casualties*: the more the severity of the car accidents, the less frequent was the involvement of two or more casualties. All the values were then spread across two categories: "one casualty" and "more than one casualty". The new feature is called `Casualties_Level`.
- *Speed Limit*: accidents occurring in streets with speed limits lower than 40 mph are mostly non fatal. The speed limit values were then condensed into two categories: "lower than 40 mph" and "faster than 40 mph" The new feature is called `Speed_limit_threshold`.
- *Junction Detail*: fatal accidents are more frequent when there is no junction within 20 meters (Figure 1, category 0); non-fatal ones when there is a T-junction (Figure1, category 3). All the other junction types were grouped into a general category (`Junct det other`) so that after the One Hot Encoding process they can be easily and quickly deleted.
- *Light Conditions*: fatal accidents are more frequent when is dark and there is no public lighting system (Figure1category 6). All the other categories were grouped into a general one so that after the One Hot Encoding process they can be easily and quickly deleted.
- *Area*: accident severity is inversely and directly proportional to, respectively, urban (Figure1category 1) and rural area (Figure1category 2). All the other categories were not considered for the analysis.

An example of the modifications applied to the dataset is reported in Figure2.
Next, the sum of the values for all the previously selected features was calculated after grouping them by `Accident_Index`. As expected, after this process, the number of rows was the same for all the three databases and the merging using as index `Accident_Index` was done. The features we were more inter- ested in were then selected without including all the complementary classes deriving from the One Hot Encoding: this could in fact lead to a *multicollinearity problem*.

To inspect for a possible multicollinearity problem, a correlation matrix for all the features was created. As we can observe from the Figure3, the highest correlation was found along the matrix diagonal and there were no highly correlated features; so we proceeded with the feature selection

3.3. Data balancing and features selection

The dataset was divided into the *Target value* `Y` and the *Feature table* `X`.

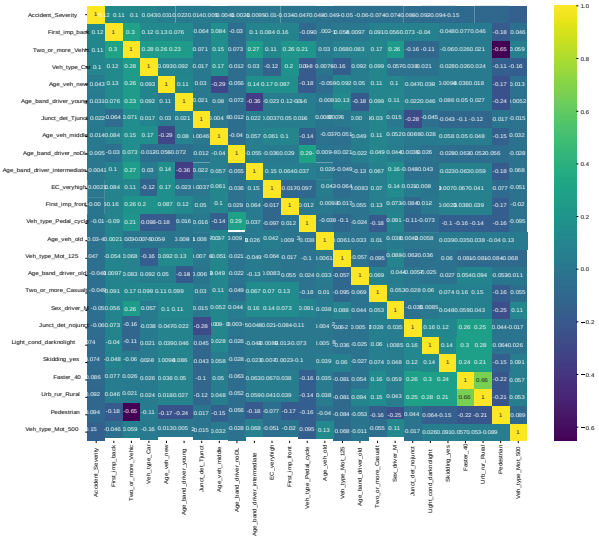

Figure 3: Correlation matrix of all the chosen features. No collinearity problem detected.

Before starting the evaluation of features selection, I've balanced the dataset. The proportions of street accidents between the different categories were indeed very different, with the minority class (Fatal ac- cident") accounting for the 1.4% of the total cases and the intermediate class ("Severe accident") for the 18.9%. This is an extreme situation that requires an accurate balancing of the dataset: a model built on such imbalanced categories can be indeed affected by the so called *accuracy paradox*, that is the situation where the high accuracy of the model reflects just the underlying class distribution. An accuracy of the 79% in our case, could also be obtained with a model predicting all the cases only as pertaining to the most representative class ("Slight accidents").
To address this problem, I decided to use a double approach: 1.Oversample the

minority class up to 10% of the majority one $(9,000)$

2.Downsample the other two classes to the level of the minority one, except for the majority (set to $13,000$)

The oversampling was achieved by using the SMOTE ("Synthetic Minority Oversampling TEchnique") technique, which consists in selecting randomly an example of the minority class, finding five (or more) neighbours and creating new samples along the line connecting the reference sample and one of the five neighbours.
To downsample the other two classes, I took advantage of the RandomUnderSampler method. Both SMOTE and RandomUnderSampler techniques are part of the `imbalanced-learn API`.
At this point, the dataset was split into two different datasets: *Training data* (corresponding to the 80% of the whole dataset) and the *Test data*; each one having a target value and a feature table.
Before running the models on them, the Feature tables of training and test data were preprocessed in order to get the same range for all the features and avoiding the consequent weight imbalance due to different range scale of the inherent measurement or characteristic.
To obtain a balanced dataset reflecting both in training and test dataset the same proportion of target classes of the original dataset, the option `stratify=Y` (with `Y = target`) was used.

3.4. Modeling

Before building the model, I evaluated the number of features to include in it: a big amount of features requires more time for the model to be built; on the other hand, the additional information provided could

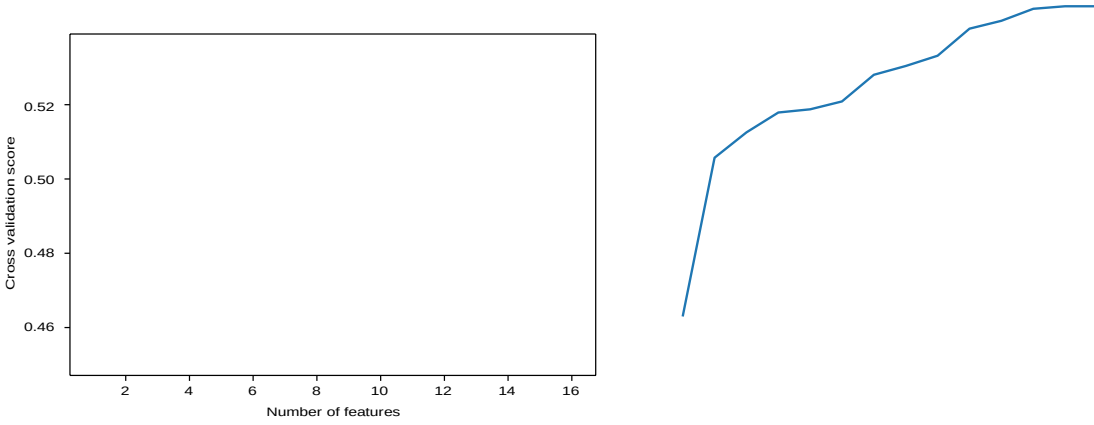

Figure 4: Model accuracy with increasing subset of features after RFECV evaluation. Highest accuracy reached at 13 features.

help in increasing prediction accuracy.
To discard the features not improving the model, I applied a combination of two techniques:

1.Recursive feature elimination

2.Cross validation

The first one evaluates the accuracy of the model for all the features and then exclude one of them in an iterative process in order to reach a given number of features which best contribute to model accuracy. Since I don't have any clue regarding the right number of feature to preserve (that RFE needs), I went for a combination of it with the cross validation technique. In fact, cross validation allowed me to run the model multiple times on different combinations of the dataset, collecting in the end the best scoring collection of available features.

I've run the module `sklearn.feature_selection` with an initial set of 16 features and a Logistic Re- gression model as estimator. In particular, I've chosen as solver `saga`[1], I've set the maximum iteration number to 4, 000, the `multi_class` option to `multinomial` and the `class_weight` to `balanced`). In this way, the model was "informed" about the multi-class nature of the target value and adjusted the weights ac- cordingly to the frequency of the predictive classes (which is not identical).
RFECV selected 15 out of 16 features previously selected as important for the model accuracy (see Figure 4), so I dropped from the model `Age_veh_old`'. The result can change if run on a different selection of the dataset; in other instances RFE suggested to remove `Two_or_more_Vehic` too.

The principal aim of this model is to predict the severity of UK street accidents by using informations collected through the STATS19 form. The target value is categorical and has three different hierarchical levels: "slight", "severe" and "fatal" accident. Since the correct labelling for each accident is present, I've chosen a supervised model. Among all the possible choices, I've selected the following classification models:

- Logistic regression
- KNN
- Decision tree learning
- Gradient Boosting Classifier

[1]Indicated for datasets in which the number of rows is very large compared to the number of features.

4. Results

For each model I have calculated training accuracy and three different accuracy evaluations of the test dataset (Jaccard score, f1-index and log-loss). A huge drop in accuracy observed after modelling test data is considered as a clue of overfitting. To have additioinal insights regarding the accuracy, a confusion matrix was built for every model.
Before running a model, I needed to identify the right hyper-parameters for the algorithm. Among all the possible options available. A module that comes in handy for this task is `GridSearchCV`, which finds the combination of a series of parameters given by the user that leads to the highest accuracy.

4.1. Logistic Regression

The optimization involved the *solver* and the *strength of regularization* (`C`). I went for a very broad range for C (from 0.01 to 100) and selected only solvers capable of handling multiple target classes (`newton-cg`, `sag` and `saga`).

Then I run the Logistic Regression model with the optimized hyperparameters(C = 10 and solver: sag[2]).

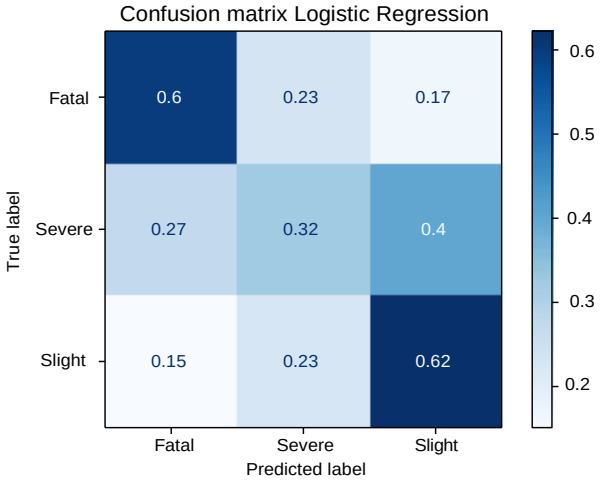

Figure 5: Confusion matrix for the Logistic Regression model.

As we can observe from it, “Fatal” and “Slight” accidents were classified in a sufficiently accurate way; severe accidents instead were almost equally misclassified as pertaining to the other two classes.

4.2. K-nearest neighbors (KNN)

In this case, the only hyper-parameter to optimize is the number of neighbours (k). To do so, I’ve run observed (k where k changed from 1 to 40 and that gave as result the k for which the highest accuracy was 13). a for loop

Also in this case, the two extreme classes were classified quite correctly, while most of “Severe condition” accidents were considered as slight or fatal.

4.3. Decision tree learning

The GridSearchCV was used for optimizing the criterion of splitting (`entropy` or `gini`), the maximal depth of the tree (the maximal expansion of nodes, range: 2, 5 or 10), the minimal sample leaf (or the minimal number of samples required for each node, range: 2, 5, 10 or 20), the minimal samples for split (range: 2, 10 or 20), the maximum number of leaf nodes (range: 5, 10, 20).

[2]This solver is indeed indicated while dealing with a dataset characterized by a large number of rows and a small number of columns/features.

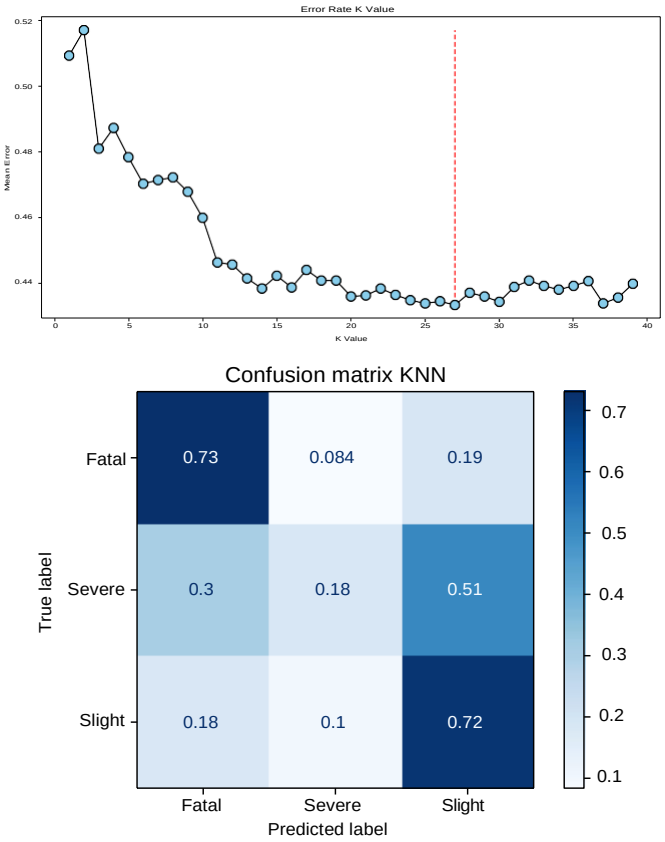

Figure 6: Error rate for increasing k-values and Confusion matrix for the KNN model.

The confusion matrix of the Decision Tree model revealed a better classification of the “Severe” class at the expense of a worse accuracy for the the two classes when compared to the previous models.

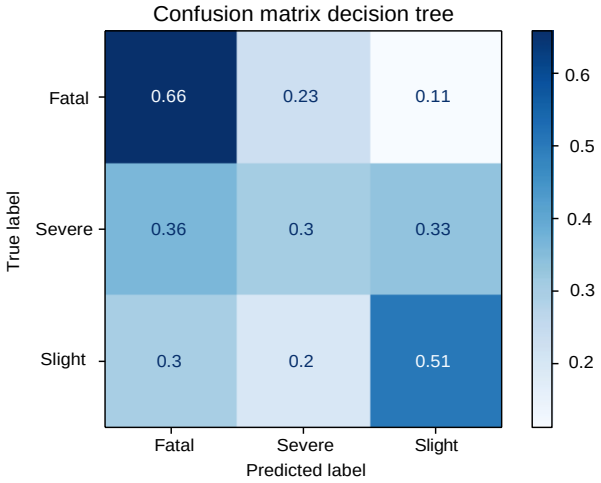

Figure 7: Confusion matrix for the Decision Tree Learning model.

Model	Training accuracy	Jaccard coefficient	f1 score	Log Loss
Logistic Regression	0.53	0.36	0.52	0.98
KNN	0.57	0.38	0.53	0.98
Decision Tree	0.51	0.34	0.50	0.98
Gradient Boosting Classifier	0.57	0.34	0.50	0.98

Table 3: Number of street accidents for different categories before and after the balancing.

4.4. Gradient Boosting Classifier

I've also tried an ensemble method, which consists of several weak models (learners) combined se- quentially so that they will for a strong learner, improving prediction accuracy. I've optimized the hyper- parameters ("learning rate") by using a for loop giving the accuracy score of training- and test-datasets and selecting a value of 0.25. The number of chosen features for the model was slightly more than the half of the total.

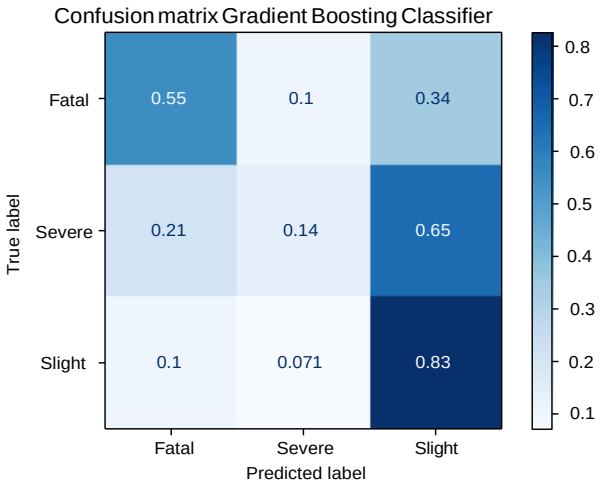

Figure 8: Confusion matrix for the Gradient Boosting Classifier.

The confusion matrix revealed a strong tendency toward a misclassification of all accidents as "slight accidents".

Log-loss is a measure of the performance of a classification model, which takes account the probability that is assigned to the target.

The list of the selected features with their *relative* importance is reported in the following table

5. Discussion

5.1. Accuracy evaluation

To have a clear idea regarding the performance of all the models employed, I' created a table with all the scores measured in different ways as well as the accuracy on trained data.

From this, it emerges that the average accuracy for every model was low, with Logistic Regression and KNN showing the higher score in terms of f1-score and LogLoss. In most of the cases, the model was able to correctly predict fatal and slight accidents; severe accidents were instead predicted as slight accidents in most of the cases (Logistic Regression, Decision Tree and Gradient Boosting Classifier), while KNN had the

worse “severe accident” prediction in general (with only 26% of the cases correctly predicted (see Figure 6). The correction for over-fitting was acceptable, since the drop of the accuracy level obtained for training- and test-data are not so different. Another important consideration is the low level of accuracy obtained with the training data, suggesting further data refinement and/or collection before deploying the model.

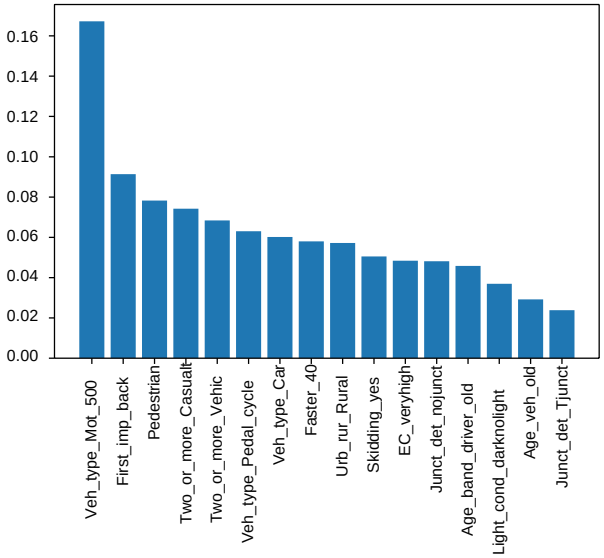

Figure 9: Feature importance.

5.2. Outcome interpretation

I’ve also calculated the feature importance by using an Extra Trees Classifier (10 trees selected, see Figure9). From this analysis it emerges that the presence of a motorbike with more than 125 cc is strongly associated with fatal accidents and is the more important features among all considered. Other (relatively) important features are:

- First impact back
- Presence of pedestrians
- More than one casualty involved
- More than one vehicle involved
- Involvement of pedal cycles
- Involvement of pedal cycles
- Speed limit higher than 40 mph
- Rural area of the accident

Exploring the decisional tree, we see that independently of the number of casualties and pedestrian involvement, the association with a fatal accident is strong where the first impact of a vehicle is not on the back and the accident takes place in streets with junctions and high speed limit. The accidents severity was more serious if the amount of pedestrian involved in accidents located in street with low speed limits was consistent. The severity get worse also if the accident involves powerful motorbikes in a rural areas with

low speed limits.
Leaves with high entropy still are retained in low speed limits streets where there are not so many junctions and only one vehicle is involved. The same situation occurs in high speed limit streets with only one casualty and no pedestrian involved.

6. Conclusion

All the classification models employed predicted quite correctly the more extreme severity classes for street accidents in UK ("fatal" and "slight"), obtaining very poor performance regarding the intermediate severity class ("severe").
It emerges that powerful motorbikes are associated with severe-to-fatal accidents even in rural areas where the speed limits is below 40 mph; in a higher speed limit context, the first point of impact located on the sides or the front of the vehicle is strongly associated with accidents fatality and the presence of pedestrian and street junctions leads to a more severe accident-class prediction.
The poor accuracy of all the models investigated do not allow model deployment and require its refine- ment with new data and transformations of old data to better discriminate the "severe accident" category. Nonetheless, insights in order to ameliorate in the short term the incidence of fatal accidents (which obvi- ously have the highest priority) can be deducted:

1. An increment of traffic stops with special attention for drivers of mid-high powerful motorbikes in both rural and urban areas. The fact that fatal accidents occurred even in streets with low speed limits, does not take into account the fact that the driver could have been driving above the speed limit.

2. Since the presence of pedestrians is proportionally associated with an increment of accident severity prediction and fatal accidents augments in presence of street junctions, the Department of Transports should better regulate traffic flow in correspondence of high pedestrian density areas and road junc- tions.

Regarding the model refinement, several factors explaining the possible discrimination between differ- ent severity levels should be individuated in a context with only one vehicle or one casualty is involved and there are no street junctions within 20 meters.
Another way to obtain a suitable tool for individuating the field of intervention could come from the use of two distinct target classes: "fatal" and "not fatal" accidents. Indeed, fatal accidents have the priority in the safety agenda of UK Department of Transport and the funds destined to the project could not be enough to intervene in such a huge program. This simpler model could also be useful for the definition of a life-insurance policy.

References

[1] WHO, Global status report on road safety 2018.
[2] Eu funds for road safety multiannual financial framework 2014- multiannual financial framework 2014-2020 saving lives saving lives on eu roads until 2020[online] (2019).
[3]A. Dhani, D. Robineau.Reported road casualties in great britain: provisional estimates year ending june 2019[online] (2018).

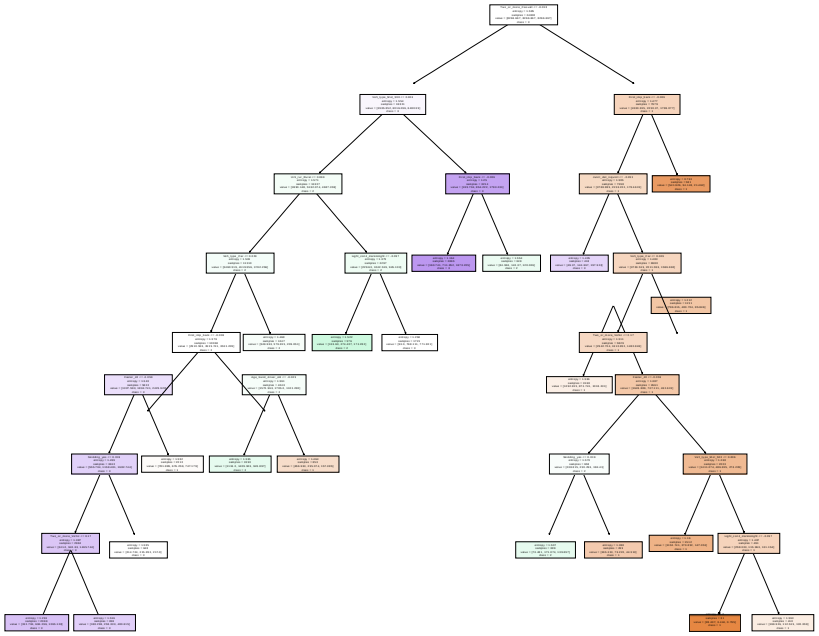

Figure 10: Diagram of Decision Tree output.