The Economic Value of Predicting Accident Severity by Machine Learning Algorithms: An Analysis of UK Traffic Accidents from 2005-2019

Nikolas Anic (14-606-800), Sophia Bieri (15-727-373), Marco Funk (15-708-704)
Mauro Gübeli (16-704-801), Yannik Haller (12-918-645)
Second Semester Project in Machine Learning for Economic Analysis
University of Zurich

January 8, 2021

Contents

1 Introduction

Traffic accidents are among the most critical issues facing the world as they cause deaths, injuries, and fatalities. Furthermore, they result in economic welfare losses and social costs. According to the World Health Organization (2020), 1.35 million people die each year as a result of road traffic crashes, and between 20 to 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. Consequently, traffic accidents in most countries cause costs in the order of 3% of their gross domestic product. While the demand for vehicles is rising continuously, the number of vehicles on the road and traffic jams increase as well, especially during rush hours.

An accurate model to predict and help identifying the most influential factors for the severity of accidents is highly valuable for governments and insurances. Analyzing the determinant factors of accident severity can help to increase knowledge that can be used to prevent these accidents. In recent years, with the advancements of information technology, machine learning techniques became increasingly helpful predicting the severity of accidents. Krishnaveni and Hemalatha (2011) have conducted an analysis of traffic accident incidences in Hong Kong. The authors used several classification models to predict and detect the severity of injury and causes of accidents. Beshah and Hill (2010) employed Naive Bayes, Decision Tree and K-Nearest Neighbors classifiers to build prediction models to assess the injury severity. The focus of that study was to predict the role of road-related factors for traffic accident severity. Another study of Chen et al. (2016) used Support Vector Machine models to investigate driver injury severity patterns in rollover crashes using two-year crash data collected in New Mexico. AlMamlook et al. (2019) conducted a similar study focusing on using machine learning techniques to predict accident severity in Michigan.

The study at hand focuses on various machine learning techniques using a comprehensive dataset of traffic accidents taken from the Department for Transport of the United Kingdom (2020). In the following section, the dataset and data preprocessing are explained in a first step. Then, in a second step, an explanatory data analysis is presented. Afterwards, in section 3, various machine learning models are applied to predict the severity of traffic accidents. Namely, we apply Logistic Regression, Decision Tree, Artificial Neural Networks, as well as K-Nearest Neighbors classifiers. In addition, section 3 includes a feature importance evaluation, in which the explanatory power of each accident severity predictor under study is assessed by means of the applied Logistic Regression model. Section 4 then concludes and provides policy implication suggestions for insurance companies and governments.

2 Data

2.1 Data Gathering and Preprocessing

In order to understand to what extent and through which channels Machine Learning models can function as a valid basis for predicting trends and conditions influencing traffic accidents, we use data from three distinct datasets taken from the Department for Transport of the United Kingdom (2020).

The datasets cover registered accidents in all legal divisions of the United Kingdom for a time period of 15 years, ranging from 2005 to 2019. The individual sets cover a total of almost 70 feature variables, including information on three major sections. First, information on accident circumstances is provided for each distinct accident. This includes variables on the timing of the accident, location parameters, road categories, speed limits as well as physical conditions of the surrounding environment. Second, information regarding vehicle indicators for each vehicle involved in a listed accident are provided in the form of indications on vehicle types and constitutions, accident manoeuvres, baseline characteristics of the drivers as well as impact characteristics. Lastly, we obtain information on casualty parameters (for each person involved in an accident), such as severity of the health related consequences of the accident, baseline information on the victims, as well as means of transportation. An overview of all feature variables is applicable in the document "STATS19 Variable lookup data guide" on the website of the Department for Transport of the United Kingdom (2020).

Following up, a variable pre-selection process is conducted. We first assess the prevalence of missing or unclear values of each feature by calculating the respective fraction of incomplete data and obtaining a visual representation in form of a histogram. Then, inspecting the features visually, we define a threshold manually for which we assume that inclusion does not substantially restrict the observations on other features. Next, we conduct assessments based on the variation of each feature. This is based on the idea that, if a factor does not greatly vary, its statistical importance on the outcome variable is likely to be limited. As such, we also inspect visually the shapes of the individual distribution and perform an iterative selection process. Building upon the initial selection strategy, we then define which variables are either too difficult to interpret if left in the model or are unlikely to add predictive power given their co-existence with other variables. Examples for such variables constitute too detailed descriptions on road characteristics (such as the meter distance to the nearest roundabout or different levels of pedestrian crosses) or variables which are deemed to express a similar factor based on different characteristics (such as road classes and road number levels).

All steps taken, we obtain 18 remaining variables that we use for our analysis. These include:

1. Accident Severity: A 3-level categorical character string indicating the severity of an accident. Ranging from Slight over Serious to Fatal
2. Light Conditions: A categorical variable covering 7 specific light conditions
3. Weather Conditions: A categorical variable covering 7 specific weather conditions
4. Road Surface Conditions: A categorical variable covering 5 specific surface conditions
5. Road Class: A categorical variable covering 5 distinct road classed, indicating motor-ways, trunk-ways, distributional roads, minor roads and non-classified roads
6. Urban Area: An indicator variable defining urban environment status
7. Hour of day: A continuous variable previously defined as "time of accident". This variable was transformed into hourly intervals to obtain a better interpretability and

improve policy predictions. As such, we defined 24 intervals covering each hour of the day

8. Day of Week: A categorical variable covering the 7 specific days of the week which are indicated as Monday to Sunday
9. Month: A categorical feature indicating in which month the accident occurred
10. Year: A categorical feature indicating in which year the accident occurred
11. Sex of Driver: The gender of the driver causing the accident, given as male or female
12. Age of Driver: A categorical variable covering the age of the driver in bins. Starting with 0-17, 18-20, 21-25 and then in 10 year intervals approaching 75 and ending on 75+
13. Speed Limit: The respective speed limit in mph where the accident occurred
14. Vehicle Type: A categorical variable previously given by multiple sub-categories for cars and motorcycles. These sub-categories were merged into either the major category cars or motorcycles
15. Engine Capacity: The engine capacity of the vehicle causing the accident. Defined individually for cars and motorcycles. Both categories are assigned into a 4-level cluster
16. Junction Detail: Covers details on environment surrounding the accident. Summarised into a 3-level categorical feature with levels junction, open street and roundabout. This is to indicate in which street setting the accident took place
17. Multiple Vehicles involved: A binary feature indicating 1 if more than one vehicle was involved in the accident
18. Age of vehicle: A categorical feature indicating the age of the vehicle at accident date. Binned into 4 levels, ranging from 0-1 to +10 years

Next, we depict in greater detail how these remaining variables are cleaned and transformed from quasi-continuous into categorical features (where this was necessary). We do these transformations by implementing a bin-type structure, with bin widths assessed individually for each feature type. This approach allows us to apply one hot encoding techniques required for several machine learning algorithms (e.g. logistic regression, decision tree, artificial neural network). We use this strategy for a total of 7 variables. The exact transformation of the variables is applicable in Table 1.

Table 1: Transformation of Features to Categorical types

Feature	**Transformation steps**	**Outcome Variable**
Vehicles Involved	Transform the number of vehicles involved into a dummy indicating 1 if accident was not singular	Indicator variable: Multiple vehicles involved
Junction Details	Transform and merge multiple accident surrounding indicators covering traffic conditions into three major categories	Categorical variable indicating accident at (I) Junction, (II) Roundabout, (III) Open Street
Time of Accident	Transform the exact time of the accident into hourly intervals	Categorical variable with 24 hourly intervals
Vehicle Type	Transform and merge major vehicle types (public and private cars, motorcycles) and exclude minor vehicle types (trams, horses, trucks, buses)	Categorical variable indicating status of vehicle as (I) car or (II) motorcycle
Age of Driver	Transform age of driver into pre-defined categories considering also youth driver and pension status	Categorical variable indicating age bins from 0-17, 18-20, 21-25, 10-year intervals up to 75, 75+
Age of Vehicle	Transform age of vehicle into categorical bins considering also new dealership	Categorical variable indicating age bins from 0-1, 2-5, 6-10 and 10+ years
Engine Capacity	Transform and merge engine capacities into 4-bin intervals for both, motorcycles and cars	Categorical variable indicating engine capacity bin. Range for motorcycles: 0-125, 126-350, 351-600, 601-1150 cc. Range for cars: 1151-1999, 2000-2999, 3000-3999, 4000+ cc.

After the multi-step transformation, we merge the variables together and obtain a dataset consisting of roughly 1.5 Million observations portrayed by 18 features and covering a total time period of 15 years for all legal areas of the United Kingdom.

2.2 Research Question

Having constructed a large-scale dataset with the attributes described above, we now focus on the specific research questions aimed to answer. As is apparent, the feature variables allow for a thorough analysis of multiple subjects related to traffic cycles and accident structures. Overall, it is our aim to better understand the underlying mechanisms leading to traffic accidents. However, an especial impact to both public as well as private decision-makers is likely to lie in the field of severity assessments. As such, we focus on the variable *accident severity*, a categorical feature covering three levels of accident impacts, ranging from slight (I) to serious (II) and fatal (III). Setting the focus on this specific outcome can allow for two

main considerations.

On the one hand, it is our aim to assess whether and to what extent state of the art machine learning methods are able to predict accident cycles. As such, we would like to understand which techniques can best mirror the underlying relationship between feature variables and accident severity from a technical standpoint. Especially, this objective can provide further insight into prediction accuracy on traffic as well as accident behaviour, provide an overview of key assumptions the individual models are based on and, accordingly, help define strengths and weaknesses of different methods, in forms of technical feasibility, accuracy and predictability.

Having assessed the ability to portray the underlying relationships from a statistical perspective, it is our aim to define and select features which act as a strong signal on accident prevalence. We want to know which features are likely to influence accident severity, both at a general as well as an impact specific level. Doing so can help define situations in which accidents are expected to be especially prominent given the predominance of these variables. Identifying certain patterns can thus aid decision-makers in two ways. First, new projects can incorporate measures to mitigate the prevalence of these features or even prevent them entirely. Second, already existing infrastructure can be re-evaluated and updated accordingly.

To summarise, we focus in this report on both the overall predictability of traffic accident cycles as well as the prevalence of certain features to serve as signal of accident likelihoods, by assessing different machine learning methods and subsequently performing feature selection using these models. Both should provide further insight into accident forecasting and aid public as well as private organs in making more informed decisions on road safety and accident precaution.

2.3 Exploratory Data Analysis

In this part, we analyse the underlying behaviour of our feature variables to assess their distributional forms, observe their co-existence with other variables and define, through visual inspection, to what degree individual features might have an impact on the classification framework. As such, we will perform a first analysis that is likely to give preliminary results, which are expected to be captured by the subsequent models for prediction and disclosure of severity class conditional feature importance.

In order to perform this type of data analysis, we produce bar plots for each of our feature variables, which are shown in Figures 16-32 in the appendix. These bar plots visualize the severity class conditional densities for each category of the feature under consideration. Phrased differently, the plots depict the share of the total number of accidents observed in a specific severity class, that occurred within the respective feature's categories. Furthermore, to interpret these plots appropriately it is important to be aware of the unequal distribution of the three accident severity classes. Thus, we start by providing an overview of the frequency of accidents occurred in each severity class during the observation period by means

of the histogram shown in Figure 1. As is expected, we observe that the dataset is very unbalanced, as it contains 1'239'548 slight, 221'973 serious and 20'851 fatal accidents.

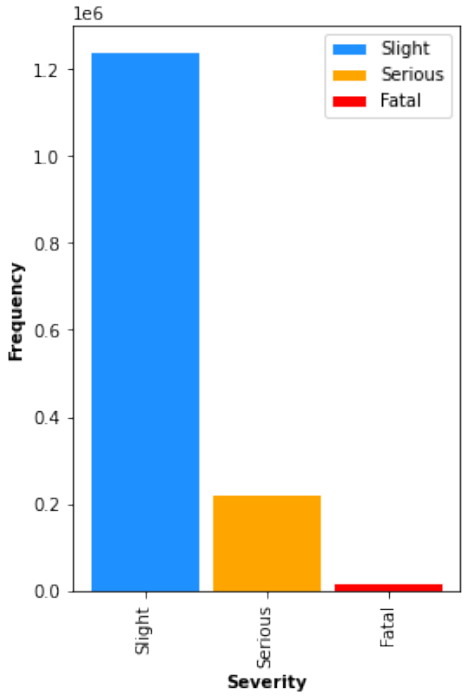

Figure 1: Severity Class Distribution

Given this information, we are now able to perform some exploratory data analysis to depict preliminary patterns in the data.

If we look at Figure 16, the following observations can be made. Overall, most accidents appear to occur in either 30 mph or 60 mph zones. This indicates two things. First, as these are the two most common speed limits in Great Britain (Department for Transport of the United Kingdom, 2019), it is somewhat intuitive that more accidents are observed within areas where these speed limits apply. Second, many accidents appear to occur on cross-country roads. These roads connect individual towns or villages with each other, but are not actually considered as main roads. As such, they are likely to be less developed and, at the same time, still heavily used. With respect to development status (e.g. lights, width of roads, indications, surface), it might be that these roads contract a greater risk of exposure to accidents. Importantly, not only slight, but also more serious injuries appear to peak at this mileage. That might arise from two possibilities. For once, these roads might also be used by non-motorised vehicle types (pedestrians, bikers) which would more quickly suffer more severe injuries following an accident. As such, it might be that many accidents occur in a transgression zone (going from 20 to 30 mph for instance). Secondly, one could assume that speeding is an issue within these zones. As accidents at a speed of 30 mph are less likely to lead to serious injuries when two vehicles collide, the possibility that these accidents often occur with a greater velocity must be considered. The second peak is at 60 mph. Within the UK, such roads are labeled as single carriageways, which are one level below motorways.

Such roads are also often constituted as motorways, but also lack more often infrastructural maintenance. As with the 30 mph zones, one could consider that a combination of poor constitution of these roads, large traffic volumes as well as speeding might be responsible for the degree of slight, serious and lethal injuries apparent in this setting. Furthermore, we observe that the severity conditional densities of more severe accidents tend to be larger in areas with a higher speed limit. This is also expected, as accidents are much more likely to lead to serious injuries when the causing vehicle comes along with a higher velocity.

Looking at Figure 17 reveals that, overall, most accidents appear to occur during rush hour times in the mornings and afternoons. This is intuitive as an increased traffic volume increases the risk of accident exposure. Further, it can be seen that most more serious accidents, both proportionally as well as in real numbers, occur during late afternoon rush hours. Potentially, this is the time when most traffic occurs and individuals are more likely to make mistakes as they are exhausted from work and don't pay the attention required in each situation. Another factor might be impatience. This could occur when people just want to get home or also have to catch up to some plans made in the evening hours. As such, it is likely that stress or time pressure plays a central role in the given accident distribution. Moreover, we observe that fatal accidents show a comparatively high density during the late evening and night time hours. This indicates that bad light conditions and tiredness of the driver could also be important factors that increase the probability of lethal accidents. Another aspect could be excessive speeding during the night time which lead to such types of accidents, as people may expect a comparably moderate police presence during these times.

Also, by inspecting Figure 18 we can observe that slight and serious accidents occur quite evenly distributed throughout the week, with reaching the peak on Friday. Interestingly, we observe an opposite trend in slight and fatal accidents on weekends, as relatively fewer slight accidents are reported, while fatal accidents appear to be comparatively more likely. When considering potential factors driving the relative increase in fatal accidents on weekends, we claim the following two major forces. First, many motorcyclists and sports car owners tend to go for a ride on weekends in order to follow their passion. The routes for such tours are known to lead over winding passes and other dangerous roads. Moreover, they often involve speeding, risky overtaking maneuvers and rather coltish driving in general - all of which are factors that are likely to increase the risk of fatal accidents. Second, since most events involving excessive consumption of alcohol or other debilitating substances take place on weekends, we claim that it is more likely that people drive under the influence of mind impairing substances, which again would increase the probability of rather severe accidents. The decrease in slight accidents, on the other hand, might be based on the following two points. First, traffic is simply lower throughout weekends when people are not using the vehicle for work purposes. Second, traffic might be associated with less stress during weekend periods and people might be less exhausted when using a vehicle. As such, rush hour commuting is less prone. Together, these two observations can support the claims made above and help explain the pattern observed. That is, overall most accidents appear in settings with a large traffic volume, more non-motorised participants at more stressful times and less-developed infrastructures.

In Figures 19 and 20 can be seen that for both, month and year, only decent signals are apparent. Overall, slight accidents appear to partly decline over the years, whereas the rate of serious incidents appears to remain stable. One reasoning for this pattern could be that less severe accidents might just be better controllable and thus, be better preventable through spreading of awareness, whereas more serious accidents appear to rely on factors which are harder to control individually. Further can be observed, that the relative density of fatal accidents appears to decrease more drastically over the years than for slight accidents. This might be due to the development of new technologies (e.g. better airbags or automatic breaking systems) which better protect the occupants of a vehicle or even prevent lethal accidents from occurring in the first place. For months, we don't see any particular pattern, as accidents from all severity classes appear to be rather uniformly distributed. The slight dip observed in February is probably due to the simple fact that this month just has fewer days than all the others and therefore the prior probability of accidents is lower. Overall, however, it seems likely that both variables are not very appropriate predictors.

Further, in Figure 21 can be seen that approximately 70 percent of all accidents occurred when multiple vehicles were involved. However, there is only a great discrepancy for slight accidents. For more serious accidents, there appears to be no strong indication that they occur significantly more frequent when multiple vehicles are involved. This might be since more serious accidents might often occur in settings where drivers do not pay attention to their surrounding environment while being in a higher-speed setting. Cases of such might be driving at times with relatively low commuting volume (such as in the night on a less developed road). As such, if attention span is assumed to be an effective predictor of more serious accidents and if attention is rather given when commuting is intense, then it is likely that such accidents occur also quite frequently in a more remote setting.

Coming to road classes (whose class conditional distribution can be seen in Figure 22), most accidents appear to occur in either trunk or distribution roads[1]. Trunk roads are the most heavily used form of roads in the UK. They are similar to second-class highways, connecting most major towns and cities. As opposed to motorways, they do not only connect the largest economic areas within the UK, but go a level deeper. Here, normal speed limits of 60-70 mph are given. However, often these roads also lack infrastructural maintenance and are not as developed as motorways. This could give intuition as to why motorways experience such a small accident prevalence compared to trunk roads. This would suit the argument made above that technical maintenance of roads might impact accident occurrence in multiple ways. On the other hand, we see that distributional roads also experience a more frequent distribution. These are often large minor roads, indicated at 30 mph, which serve as transition roads, leading through less populated areas within towns or between villages. As such, they are also more frequently used by non-motorised vehicles, which would explain why more serious accidents appear to be, proportionally, more prevalent within these roads. Inspecting Figure 23 reveals that most slight accidents (and therefore, as this is the predominant class, also the overall number of accidents) appear to occur in urban areas. This is

[1]We omit unclassified roads, as we are unable to clearly depict their occurrence, but assume that these are proportionally distributed to less developed roads - as such no motorways

somewhat intuitive, as traffic is usually much heavier and more dense in cities than in rural areas. However, we also observe that the density of lethal accidents shows a reverse pattern. This might be because, due to the generally lower speed limits (i.e. 30mph in most urban areas vs. 60mph in most rural areas) and heavier traffic apparent in urban areas, speeding is less possible there than in rural areas. As such, it is much less likely that the accident causing vehicle came by with a very high velocity. Consequently, serious accidents are less likely to occur in urban areas. This is to be expected, as we already saw a similar image when comparing the densities between the speed limits 30mph (common in urban areas) and 60mph (common in rural areas).

Figure 24 depicts that most accidents appear to occur at junctions or on the open street, with junctions being responsible for somewhat more slight injuries and open street being responsible for, in absolute and relative terms, more serious accidents. While the latter is quite intuitive, it is still a little surprising that more slight accidents occur within the setting surrounding junctions. One explanation might be that drivers are more focused on open street than on junctions, where speed levels are lower and directions from which vehicles are approaching increase. Thus, they are likely to incorporate cases of rear-end or side-way collisions. Also here, one could argue that junctions are potentially less maintained compared to open streets and lack clearness, imposing risks for dead angles or seeing other participants too late. Also, as junctions are likely to occur in a low-speed environment, participants might not pay the amount of attention they would on open street.

Further, when looking at Figure 25 we observe that most accidents are caused by cars. However, this extreme predominance of accidents caused by cars is mainly due to the fact that there are just more cars on the roads than motorcycles. On the other hand, when we compare the relative share of slight and more severe accidents between cars and motorcycles, we observe a clear pattern. Motorcyclists are (in relative terms) much more likely to cause serious and fatal accidents. This is to be expected because, in contrast to cars, the architecture of the vehicle alone makes it less possible to protect the driver or occupants from serious injury in the event of an accident. Hence, the vehicle type of the driver who caused the accident seems to be a quite expressive factor for distinguishing between slight and more serious accidents.

In Figure 26 the severity class conditional distribution per engine capacity category of vehicles is visualized. At first glance, it is conspicuous that most accidents are caused by vehicles with a displacement of 1151 to 1999 cubic centimeters (cc). This is to be expected, as this is the most common class for cars - the predominant vehicle type in our dataset. Interestingly, we can observe that for both, motorcycles and cars, more serious accidents appear to be relatively more likely as engine capacity increases. The reason for this might be that, as increasing displacement usually leads to more power and torque, vehicles with a higher engine capacity are better able to accelerate very fast. This ability is, especially for individuals belonging to the target group of such highly motorised vehicles, very tempting and therefore often leads to uncontrolled speeding. As a consequence, it is more likely that accidents involving highly motorised vehicles are caused with a high velocity, which dramatically increases the probability of severely injured casualties. Thus can be concluded, that

engine capacity category could be considered as a potentially quite important predictor for accident severity status.

Coming to light conditions (see Figure 27), accidents are mostly prevalent for daylight conditions. This is intuitive given the fact that most accidents occur during rush hour times and decline substantially at non-rush hour times. Interestingly, we can observe that, although not as many accidents occur in nighttime conditions, they appear to lead to a greater proportion of more serious accidents. Interestingly, also, is the fact that in settings where darkness is apparent but lights are not lit, the density of serious and, especially, lethal accidents are perceptibly higher compared to slight accidents. As such, it is likely that lack of visibility and too late realisation of a potential collision serves as an indicator to the severity of an accident. Thus, one could argue that, although more situations for potential accidents occur during more visible settings, they are likely to be preventable due to a better visibility and earlier reaction options.

Looking at the densities for weather and road surface conditions shown in Figures 28 and 29, respectively, one can see a quite similar image. Most accidents occur in settings where rain or dry weather is apparent. This is intuitive, as the UK is a country in which these are the two most common forms of weather. Interestingly, we do not see any form of increase of accidents under foggy or icy conditions. That is, either these events occur so seldomly that individuals are not exposed to them enough such that a statistical trend is visible, or they are not able to create a "shock moment" (comparable to the spike in accidents in snow rich countries at the time the first snow arrives). Overall, one can say that most accidents occur in ordinary weather settings, and weather and road conditions appear to be somewhat more strongly correlated with each other.

Considering Figure 30, it appears as if males were responsible of both, more slight and more serious accidents, although the discrepancy for slight accidents is less extreme. This might be based on two reasons. For once, males might be more frequent drivers and more likely to own a car. As such, they are more frequently represented in traffic and have a higher exposure to accidents. Further, males might also be more likely to be less attentive in traffic, less risk-averse drivers or more confident in their driving skills. As such, they might underestimate certain risks or overestimate their ability to forecast and handle situations, which increases the likelihood of being involved in accidents.

Also, when looking at Figure 31 we can observe that there appears to be a strong discrepancy of accident prevalence for new and older vehicles. However, this is only apparent for recently purchased cars vs. cars older than 2 years. The reasoning for this pattern might be twofold. First, car owners of new cars might be more attentive to prevent accidents, as their utility of the car is still large and their appreciation of the good is still strong. On the other hand, it might also be the case that new cars are simply not as common as used cars. That is, their relative occurrence in traffic is proportionally lower compared to cars with an increased age. A third reason might also be that new cars have more professional accident preventing systems, such as line recognition or distance measures or driver observation technologies, which could lead to a better support of the driver through predictability and a non-declining

attention span.

Lastly, when looking at Figure 32 we can observe that a lot of accidents are caused by people with an age between 26 and 45. This is intuitive when considering that, especially in the UK, these are the most frequent drivers. Especially, these drivers use the vehicle for work purposes and are thus more likely to encounter psychologically stressful situations, such as in the context of rush hours. However, as the age ranges of bins for agents below the age of 26 are relatively smaller (i.e. a range of 3 or 5 years vs. 10 years for age 26+), the magnitude of the shown bars are not directly comparable to each other. If we, hypothetically, would construct a bin for 18-25 by stacking the bars of the categories 18-20 and 21-25 on top of each other, we would observe that, overall, most accidents are caused by these drivers. This could result from a combination of lack of experience, over-confidence and, in some instances, consumption of mind-altering substances while driving. Furthermore, we can observe that, compared to the category of drivers between the age of 66 to 75, the density of fatal accidents appears to increase significantly for individuals above the age of 75, while the densities for slight and serious accidents remains on a similar level. This is intuitive as casualties older than 75 are simply more vulnerable and therefore are more likely to succumb to their injuries after an accident. However, taking all together, we can still observe that driving appears to be especially vulnerable to accidents in situations where traffic is high and is also due to psychological traits as well as road infrastructure environments.

All things considered, we can see that accidents are based on a combination of both, individually-psychological as well as infrastructural, environment-based factors. To this extent, we observe that features such as gender or age of the driver, road surface conditions, speed limits, time of occurrence as well as road classes appear to be important predictors of both, accident occurrence as well as severity level distinctions.

2.3.1 Multicollinearity Issues

A common problem we will face in regression models and during the model selection process is that of multicollinearity. That is that some features are highly correlated or that the variation in some features is explainable by a linear combination of some of the others. In said cases, feature selection might be potentially biased since, in the case of logistic or tree-based classifiers, permutation-based mechanisms might misinterpret the drop in prediction power of individual features if they are likely to be easily expressed by other features. If this is the case, then permutation importance (and mean-decrease-in-impurity importance) decreases for more collinear variables, making them appear less important. Furthermore, when high multicollinearity is present among the features, the retrieved parameters of a model may be estimated with suboptimally high variance.

An initial, although not conclusive, method to assess whether multicollinearity is present within the data is by inspecting correlation plots. However, creating such a plot requires either directly continuous features or categorical variables that are somewhat reasonably rankable (e.g. accident severity) and can therefore be encoded into a numerical format. Since, in the dataset we are using, this approach is not applicable to several features (e.g.

Road Class, Weather Conditions, Junction Detail), we decide to argue from an intuitive perspective instead. In the following, we list all combinations of feature variables in our dataset for which we suspect that they are strongly correlated, and therefore could cause multicollinearity to some extent.

1. Weather Conditions and Road Surface Conditions
2. Hour of Day and Light Conditions
3. Road Type and Speed Limit
4. Road Type and Junction Detail
5. Vehicle Type and Engine Capacity
6. Urban Area and Speed Limit

From these considerations, one could suspect that indeed some multicollinearity might be present among the features in this dataset. However, we discuss our approach to counteracting this problem in the following.

There are several techniques to tackle the issue of multicollinearity. However, their feasibility and suitability depends strongly on data availability as well as the purpose of the analysis. First, for each pair (or group) of collinear variables, we could in principle just drop one of the features which are highly correlated with each other, as the variation in one of them already captures most of the joint explanatory power to predict the dependent variable. However, this approach is not desirable in the present analysis, as one of our aims is to rank our features according to their importance in predicting a specific accident severity class. As such, we prefer not to drop any of the variables in our dataset. We therefore chase another approach to counteract the high estimation variance caused by potential multicollinearity - namely, using a very large sample size relative to the number of features. As increasing the number of observations in a dataset relative to the number of features helps to decrease the variance of the estimated model parameters, we claim that, in the present case, the sample size is probably large enough (i.e. 1'482'372 observation vs. 120 features after creating indicator variables for each category of our explanatory variables) that we get sufficiently precise estimates and therefore multicollinearity issues are a negligible problem.

3 Methodology & Results

3.1 Model Prediction

Having assessed the relations between our features' categories and accident severity classes graphically in the previous section, we now go a step further and attempt to predict the underlying relationships by means of multiple models.

Highly specific prediction and targeting is becoming more prevalent in an increasing number of fields, including the field of insurance. This section uses the above named variables to

predict accident severity, allowing insurance companies to adjust their policies and prices according to these features of individuals, vehicle features and area specific features.

3.1.1 Logistic Regression

The first model we use to assess prediction accuracy is a logistic regression model. As a standard approach, we transform the feature variables into a binary format and apply a l2 regularisation. This allows us to further account for issues of multicollinearity while simultaneously mitigating potential concerns of overfitting.

Regarding the accuracy score, we see that the model performed quite well. With the data, we retrieve an accuracy of 0.8365, which is quite large considering other classification algorithms. This implies that, of all predicted classifications, we were able to predict nearly 85 of 100 correctly. This includes True Negatives as well as True Positives. As such, we are able to capture a sufficient part of the underlying relationship concerning the feature variables.

Another interesting implication is to observe the classification accuracy through a precision-recall matrix. Recall that a classification algorithm predicts the probability for each observation to belong to each of the given classes and then chooses the respective class as predicted value for which said probability is largest. Comparing these predictions to the actual values for each class, we can retrieve a matrix consisting of four values for each class. These are True Positives, which is the number of occurrences that a predicted value is positive given the actual value is positive, True Negatives, which is the number of occurrences that a predicted value is negative given the actual value is negative, False Positive, which is the number of occurrences a predicted value is positive given the actual value is negative, and False Negatives, which is the number of occurrences a predicted value is negative given its actual value is positive. As such, it is our aim to obtain many of the former two while limiting the number of times we obtain the latter two.

Given these values, we can form precision and recall calculations. We need to form a precision-recall matrix in our case, since the dataset with which we are operating is imbalanced. As such, most of the observations we obtain are within classification level 1, indicating slight accidents. Thus, the actual True Positive Rate (TPR) vs. False Positive Rate (FPR) comparison does not hold a reliable verdict, and purely accuracy oriented metrics might not be of much value.

As such, we are also unable to simply draw a commonly used tool, called the Receiver Operating Characteristic (ROC) curve. This is simply because the ROC curve plots TPR vs. FPR. Because TPR only depends on positives, ROC curves do not measure the effects of negatives. As such, if you use the Area under the Curve (AUC) directly from an ROC plot, you do not place more emphasis on one class over the other, so it does not reflect the minority class well.

To combat this problem, we are using the definitions obtained above. Remember that the definition of recall is the number of True Positives divided by the number of True Positives plus the number of False Negatives. Precision is defined as the number of True Positives

divided by the number of True Positives plus the number of False Positives. While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant. Because precision is directly influenced by class imbalance, the precision-recall curves are better to highlight differences between models for imbalanced datasets.

Further, we assign a F-beta score. The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.

	precision	recall	f1-score	support
Fatal	0.00	0.00	0.00	16632
Serious	0.51	0.03	0.05	177389
Slight	0.84	1.00	0.91	991877
accuracy			0.84	1185898
macro avg	0.45	0.34	0.32	1185898
weighted avg	0.78	0.84	0.77	1185898

Figure 2: Classification Report Logistic Regression

When we look at the confusion matrix and the subsequent classification report, we can observe that the results are somewhat different from the pure accuracy score.

First look at the confusion matrix. Interestingly, the model did never predict any accident to be of class fatal. This is intuitive to some extent, given the relatively low frequency such accidents occurred within the data. Overall, we can observe that only 1.5 percent of all observations included a fatal accident. As such, the model might simply have lacked a sufficient number observations to accurately define which feature combination distinctively causes a situation in which fatal level accidents are more likely than non-fatal ones. Supporting this claim, one could argue that, throughout all observations, there appears to be no setting in which an individual is more likely to suffer from fatal consequences when involved in an accident than from non-fatal ones. This is intuitive to some extent, given that individuals predominantly would mitigate their risk of exposure to these settings by circumventing such environments and that these environments would have already been reconstructed to mitigate risks of fatal accidents. Thus, it is likely that each feature combination leading to more serious accidents is as likely to lead to only slight accidents.

For us, this implies that we are unable to actually say much in particular on how to predict fatal accidents given the logistic regression algorithm. Further, this gives the first major implication on why a purely accuracy-oriented score is likely to lead to false conclusions on overall prediction accuracy. This bears the consequence of a more selective reasoning. In order to assess whether this failure to assign fatal status to accident severity, one could bear several considerations. On the one hand, one could claim that the model is not entirely sufficient given that it is unable to predict each kind of accident severity. If one was to only look at prevention for fatal accidents, then clearly we could not work with the given model.

On the other hand, one could argue that, as the proportion of fatal to overall accidents is barely 1.5 percent, and since the data catches a large number of accidents throughout the entire UK over a multiple-year horizon, a threshold of only 20'851 fatal accidents is already at an acceptable level regarding traffic fatality. Given that it is likely that earlier studies and prevention campaigns aimed specifically at reducing the number of most serious accidents, one could argue that this number is the result of the effectiveness of said programs. As such, it might be more useful to attempt predicting only serious or even slight accident patterns, given their large prevalence and, absolutely speaking, large costs they may cause, both socially as well as financially.

Looking at serious accidents, we find mixed results. On the one hand, we can observe that precision is slightly above 50 percent. This indicates that the model assigned in less than half of the cases a correct positive verdict. Looking at it from a prevention-only perspective, this is not necessarily a bad, but rather an inefficient, result. Given that more serious accident predictions would likely cause decision-makers to increase prevention efforts into these situations, one could argue that too much is better than too little. However, it would become serious if financial resources are constrained and, as such, an efficient allocation of funds is necessary. On the other hand, we find that the recall measure is practically zero. This implies that the model assigned many times a slight accident status to situations in which a serious accident occurred. In the case of accident prevention, this is not optimal. Comparing to the situation described before, we would now obtain many feature combinations in which accidents would be predicted as not serious enough. As such, a suboptimally low allocation of funds could be sent to situations in which these combinations are prevalent, potentially underestimating the danger they pose. By being unable to actually depict the number of relevant observations accurately, we thus should be even more careful in allocating too much weight to the accuracy measure. Especially, we now can support the claim made above that many situations which can be labeled as serious might also be labeled as slight and that, consequently, the exact allocation of severity status depends likely on many more parameters than the ones we obtained in our set.

Now, looking at all scores, we obtain an unweighted (macro-oriented) as well as a weighted average of f1-scores. Depending on the policy setting we would like to observe, we can choose these values accordingly. If we want to focus on the total financial and social consequences of accidents and assign each accident severity an approximately equal severe psychological outcome, one could argue that a weighted average of all scores would be sufficient to estimate the efficiency of our model. As such, being able to predict less serious situations with substantial accuracy might lead to a desired reduction in overall accident rates and could subsequently decline total costs resulting of all accidents decline. This would be similar to a prevention strategy in which it is the aim to reduce only minor costs but at a large scale. Then, we could use the weighted f1-score of 0.77, which is roughly 8 percentage points lower than the initial accuracy metric obtained. However, if it is the aim to primarily prevent more serious accidents, and if it is thought that the psychological consequences of different severity levels are heterogeneous, then using a model which is likely to underestimate accident severity of slight and serious character might not be a good idea. Also, if the financial consequences of serious accidents are disproportionally larger, then a potential reduction in

less severe accidents might not outweigh the remaining costs caused by serious accidents. As such, one would be better off by using the macro average and obtain a score of 0.34, which is substantially lower than the two previous scores.

However, a third option would be to only regard results within accident severity categories of 1 or 2 and disregarding lethal accidents. This is based on the claim made above in which lethal accidents are either at such a low level already or that they constitute characteristics which are not distinctively measurable and thus cannot clearly be predicted. In this case, one could simply take the non-weighted average of the first two categories, given that one would want to take into consideration the very low recall value given for serious accidents. As such, we would obtain a f1-score of 0.48. Depending on the extent to which one would like to weight this False Negatives for serious accidents, one could individually assign weights to both scores. This would depend on the argument made above, in which one must trade-off the consequences, both financially as well as socially, of prevention efforts into any category.

3.1.2 Decision Trees

In this subsection, we develop a Decision Tree (DT) to predict accident severity. We use the preprocessed data, as described in section 2. The only difference is, that we have to transform the output variable into an integer format in order to calculate the tree. The idea is, that we build an optimal tree, where optimal means, that we optimize four tuning-parameters: maximum depth, minimum number of samples per split, minimum number of samples per leaf and the maximum number of features. The criterion defining optimality here is the accuracy. It is further validated with a ten-fold cross-validation approach.

The analysis (which can be found in a notebook as reference) suggests that the optimal tree has the following values for the tuning parameter: maximum depth = 5, minimum number of samples per split = 0.0001 (stated as a fraction of the entire data), minimum number of samples per leaf = 0.0001 (stated as a fraction of the entire data) and the maximum number of features = 119.

We then use the optimal tree to get the predictions. The model achieves an accuracy score of 0.8373, which is almost the same accuracy score as the logistic regression model.

```
array([[246543,   1201,      0],
       [ 43051,   1436,      0],
       [  3983,    261,      0]])
```

Figure 3: Confusion Matrix Decision Tree

	precision	recall	f1-score	support
1	0.84	1.00	0.91	247744
2	0.50	0.03	0.06	44487
3	0.00	0.00	0.00	4244
accuracy			0.84	296475
macro avg	0.45	0.34	0.32	296475
weighted avg	0.78	0.84	0.77	296475

Figure 4: Classification Report Decision Tree

In the above classification report, number 1 represents slight accidents, 2 represents serious accidents and 3 represents fatal accidents. Interestingly, the DT model never predicts a fatal accident, as in the logistic regression model. It is even more astonishing, that precision, recall as well as the f1-score are mostly identical to the logistic regression model. Of course, it is very likely that the numbers do not match perfectly, as we only observe the decimal numbers. However, the reasoning behind the observed figures would be very similar to the one from the logistic regression. Hence, we refer to the argumentation given in the beforehand section on the logistic regression model.

3.1.3 Artificial Neural Network

In this subsection, an Artificial Neural Network (hereinafter referred to as ANN) classification prediction model is built to predict accident severity.

Before actually building the ANN, some additional preparation steps must be undertaken on the preprocessed dataset. For ANNs, as well as other prediction models, it is beneficial to have all categorical variables as binary dummies, which is the first step to be carried out. The second small step, is to encode the dependent variable, since we are dealing with a multi-class classification task.

Once these two steps are completed, we split the dataset into a training and test set, using a test size of 0.2.

The ANN is built using the Keras library as an interface to Tensorflow. The model is of the sequential class and contains one input layer, three hidden layers and one output layer. For the number of neurons, we choose to have six neurons in each layer, since additional neurons do not seem to lead to an improvement in model performance. In the output layer there are three neurons, since the dependent variable has three possible outcomes. As an activation function, we set the Rectified Linear Unit (ReLU) function in the hidden layers and the Softmax function in the output layer, which is appropriate for multi-class classification. The optimizer is set to be the Adam optimizer, since it is shown to be most efficient. The loss function employed is the categorical cross entropy and the metric used is accuracy. The results of this ANN model are represented in Figure 5.

```
Confusion Matrix

[[      0    1640   19211]
 [      0    7834  214139]
 [      0    6349 1233199]]

Accuracy: 0.84


Classification Report
```

	precision	recall	f1-score	support
Fatal	0.00	0.00	0.00	20851
Serious	0.50	0.04	0.07	221973
Slight	0.84	0.99	0.91	1239548
accuracy			0.84	1482372
macro avg	0.45	0.34	0.33	1482372
weighted avg	0.78	0.84	0.77	1482372

Figure 5: Results of Multi-Class ANN with Unbalanced Dataset

What happens is that, while the accuracy is relatively high (0.84), since the sample is extremely unbalanced with respect to the outcome variable ('Slight': 1'239'548, 'Serious': 221'973, 'Fatal': 20'851), the class 'Fatal' is never predicted and the class 'Serious' is only predicted very seldomly. What we do get, however, is an extremely high recall of 0.99 for the class 'Slight'.

```
Confusion Matrix

[[ 12936   3919   3996]
 [ 67577  68875  85521]
 [215914 246464 777170]]

Accuracy: 0.58


Classification Report
```

	precision	recall	f1-score	support
Fatal	0.04	0.62	0.08	20851
Serious	0.22	0.31	0.25	221973
Slight	0.90	0.63	0.74	1239548
accuracy			0.58	1482372
macro avg	0.39	0.52	0.36	1482372
weighted avg	0.78	0.58	0.66	1482372

Figure 6: Results of Multi-Class ANN with Balanced Dataset

To solve for this issue, we create a balanced dataset, by randomly picking 20'000 observations of each class. Training the same ANN on the balanced dataset, we get the results shown in Figure 6.

We see that the accuracy drops, but is nevertheless better than a random choice. What the table additionally shows is that the precision, recall and f1-score of the classes 'Fatal' and 'Serious' increased, leading to an increase in prediction performance for these two classes at the expense of the 'Slight' class. In this respect, further work could be done by combining the baseline probabilities of having an accident of a certain severity class with some prediction model.

Next, we may believe that for insurance purposes, differentiating between 'Serious' and 'Fatal' may not be as relevant. Therefore, we create a second ANN, namely a binary ANN, where the classes 'Serious' and 'Fatal' are merged. By turning the task into a binary classification task, a few adjustments must be undertaken to the ANN. The first adjustment concerns the activation function in the final output layer. This is now no longer a Softmax function, but a Sigmoid function (i.e. the Softmax's counterpart for binary classification problems). Secondly, the categorical crossentropy loss function is changed to the binary crossentropy loss function. With these adjustments, we get the results using the binary ANN with the unbalanced dataset in Figure 7.

```
Confusion Matrix

[[1227848   11700]
 [ 227688   15136]]

Accuracy: 0.84


Classification Report

                   precision    recall  f1-score   support

           Slight       0.84      0.99      0.91   1239548
Serious and Fatal       0.56      0.06      0.11    242824

         accuracy                           0.84   1482372
        macro avg       0.70      0.53      0.51   1482372
     weighted avg       0.80      0.84      0.78   1482372
```

Figure 7: Results of Binary ANN with Unbalanced Dataset

As in the multi-class ANN with the unbalanced dataset, this binary ANN mostly provides the prediction 'Slight', which is to be expected, although it does also provide predictions for the second class. We receive a relatively high accuracy and high precision, recall and f1-score for the class 'Slight'.

Finally, this binary ANN is also trained on a balanced dataset, with 240'000 observations of each class. The results are displayed in Figure 8.

```
Confusion Matrix

[[803472 436076]
 [ 90977 151847]]

Accuracy: 0.64


Classification Report

                   precision    recall  f1-score   support

           Slight       0.90      0.65      0.75   1239548
Serious and Fatal       0.26      0.63      0.37    242824

         accuracy                           0.64   1482372
        macro avg       0.58      0.64      0.56   1482372
     weighted avg       0.79      0.64      0.69   1482372
```

Figure 8: Results of Binary ANN with Balanced Dataset

We see that the binary classification ANN outperforms the multi-class ANN in terms of accuracy (0.64) and further measures (precision, recall and f1-score). From this model, we can conclude that it outperforms random choice by quite a bit.

3.1.4 K Nearest Neighbors

In this subsection, we present a k-nearest neighbors algorithm (hereinafter referred to as KNN) in order to make predictions on the severity of accidents.

The KNN model is built using the KNeighborsClassifier function provided by the sklearn package. Important parameters for the classification algorithm are the number of neighbors to use for queries, the weighting function of the neighbors used for prediction, the algorithm used to compute the nearest neighbors and the distance metric used for the classification model. In this model, we are using a $k = 5$ KNN model. For the optimal choice of algorithm to compute the nearest neighbor, we will use the 'auto' option. This option will attempt to choose the most appropriate algorithm between BallTree, KDTree and a brute-force search. Furthermore, the classifier uses a uniform weight function, where all points in each neighborhood are weighted equally. To calculate the distance, the standard euclidean distance function is used.

Unfortunately, the KNN classification algorithm is highly computationally intensive and a classification on the whole dataset, consisting of 1'482'372 datapoints, cannot be done within a reasonable amount of time. Therefore, a random subsample including 40% of all datapoints is used for the calculation. After that, the dataset is further split into a training and test set, using a test size of 0.2. The metric to assess the classifier is accuracy. The results of a $k = 5$ KNN model are represented in Figure 9.

```
Confusion Matrix

[[   15   200  1405]
 [   85  1468 16370]
 [  159  3628 95260]]

Classification Report

              precision    recall  f1-score   support

       Fatal       0.06      0.01      0.02      1620
     Serious       0.28      0.08      0.13     17923
      Slight       0.84      0.96      0.90     99047

    accuracy                           0.82    118590
   macro avg       0.39      0.35      0.35    118590
weighted avg       0.75      0.82      0.77    118590
```

Figure 9: Results of a $k = 5$ KNN with Unbalanced Dataset

The accuracy is relatively high (0.82) and similar compared to the results of the other classification methods presented above. However, the KNN algorithm is able to predict the class 'Fatal' despite a highly unbalanced sample (around 84% of all accidents are 'Slight' accidents) with respect to the outcome variable. Furthermore, we see that for the class 'Slight' the recall is high (0.96), but much lower for the classes 'Serious' and 'Fatal' (0.08 and 0.01, respectively).

To solve the issue of an unbalanced sample, we create again a balanced dataset, by randomly picking 20'000 observations of each class. We train the same KNN model using 80% of that balanced dataset, but test it on the unbalanced dataset. We get the following results in Figure 10.

```
Confusion Matrix

[[  2684   1062    536]
 [ 33525  29466  21981]
 [144593 165855 181417]]

Classification Report

              precision    recall  f1-score   support

       Fatal       0.01      0.63      0.03      4282
     Serious       0.15      0.35      0.21     84972
      Slight       0.89      0.37      0.52    491865

    accuracy                           0.37    581119
   macro avg       0.35      0.45      0.25    581119
weighted avg       0.78      0.37      0.47    581119
```

Figure 10: Results of a $k = 5$ KNN with Balanced Dataset

Similar to the ANN model, we see that the accuracy drops. However, it is slightly better than a random choice. What Figure 10 additionally shows is that the f1-score of the classes

'Fatal' and 'Serious' increased, leading to an increase in prediction performance for these two classes at the expense of the 'Slight' class.

Furthermore, we can do the same analysis as in the ANN section and merge the classes 'Serious' and 'Fatal'. The results of this binary KNN with an unbalanced dataset are shown in Figure 11.

```
Confusion Matrix

[[ 1768 17775]
 [ 3787 95260]]

Classification Report

                precision    recall  f1-score   support

Serious & Fatal       0.32      0.09      0.14     19543
         Slight       0.84      0.96      0.90     99047

       accuracy                           0.82    118590
      macro avg       0.58      0.53      0.52    118590
   weighted avg       0.76      0.82      0.77    118590
```

Figure 11: Results of a binary $k = 5$ KNN with Unbalanced Dataset

Given the unbalanced dataset, this binary KNN predicts mostly 'Slight' accidents, which is to be expected. The accuracy, precision and recall scores for the 'Slight' class are comparable to the multi-class KNN. However, these scores are better for the merged 'Serious' category.

Again, we can do the same analysis for a balanced dataset consisting of 30'000 observation for each accident category. The results for the binary KNN trained on a balanced dataset are shown in Figure 12.

```
Confusion Matrix

[[ 39946  33067]
 [198076 273860]]

Classification Report

                precision    recall  f1-score   support

Serious & Fatal       0.17      0.55      0.26     73013
         Slight       0.89      0.58      0.70    471936

       accuracy                           0.58    544949
      macro avg       0.53      0.56      0.48    544949
   weighted avg       0.80      0.58      0.64    544949
```

Figure 12: Results of a binary $k = 5$ KNN with Balanced Dataset

Unsurprisingly, we see that again the binary classification KNN outperforms the multi-class KNN model in terms of accuracy (0.56) and further measures (precision, recall and f1-score). However, it still lacks the precision of a unbalanced model.

3.2 Feature Selection with a Logistic Regression Model

As we have seen, the overall prediction accuracy of our model depends on several individual considerations. However, we are also interested in depicting which features are more likely to assign prediction for a particular severity class. This information could aid decision-makers in deciding which feature combinations they should particularly pay attention to when designing or evaluating the risk a situation could pose on traffic incidents. As such, we will follow the following strategy to assess feature importance.

We use the magnitude of the coefficients resulting from the estimated logistic regression model to assess feature importance. In this case, the higher a coefficient, the more important the corresponding feature is on assigning prediction status. Remember that we assigned a binary feature to each category of our explanatory variables. Thus, the number of features increases dramatically. In order to obtain the relevance of each feature, we simply rank the features according to their size to see which were more prominent in assigning accident severity status.

As we can see, we obtain 120 features in a binary format. We now want to assess which features are indicators for categories of which baseline predictor variable. Doing so requires us to identify which binary feature belongs to which parameter we introduced. Importantly, we can only do this for each of the three severity classes separately. As such, it is important to understand which parameter is a strong predictor for each class separately and, in a next step, combine the information obtained from each class. Let's first obtain the results for the coefficients for the class of slight accidents.

3.2.1 Feature Selection on Slight Accidents

Figure 13 shows the features ranked according to the explained variance each feature contributes to the model. In this case the features are plotted against their relative importance, that is the percent importance of the most important feature (i.e. the feature with the largest coefficient in absolute terms).

As we can observe, the most important predictor variables appear to be of Vehicle Type, Engine Capacity, Junction Detail, Speed Limit, Road Class, Urban Area environment as well as Time of Accident. We first focus on Vehicle Type. As we can see, driving a car is an important predictor of a slight accident. As we will outline in the section on lethal accidents, this does not necessarily imply that cars are more likely to cause less severe accidents. Rather, one may hypothesize that due to its historical distribution, the model assigned a large predictive ability to this factor based on its disproportional prevalence of slight accidents compared to more serious accidents. A similar verdict can be made on the road category "Motorways", which is the road class with the highest speed limit. Looking at Engine Capacity and Speed Limit, it is intuitive to assume that agents causing an accident either with a low capacity instrument or in a low speed environment are less likely to cause a substantial form of damage, mainly based on the situation's physical restraints. This can be supported by the roundabout indication, which poses a traffic situation that is predominant in a low-speed environment. Moreover, we can see that more extreme road surface

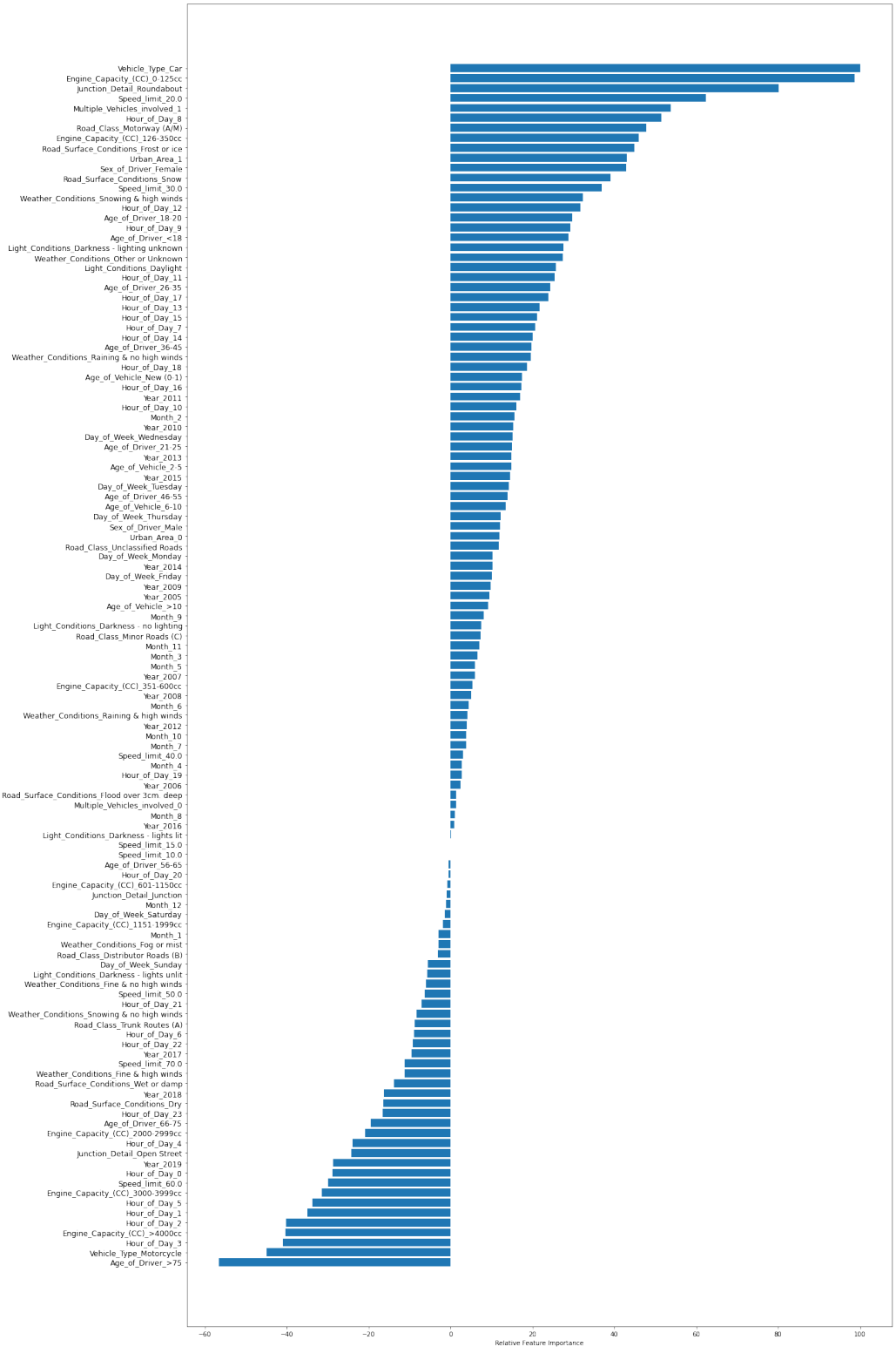

Figure 13: Feature Importance for Slight Accidents

conditions, such as frost or snow, are also an indication of slight accidents. This may also be intuitive, as these are situations which, in the United Kingdom, are rather rare events. As such, traffic agents are less experienced within this terrain and thus more likely to act incorrectly. On the other hand, one could argue that, especially because these situations act as a surprising factor, agents behave even more carefully, which would mitigate the potential for greater severity. Lastly, one can observe that rush hour morning traffic is more likely to lead to slight accidents. This may be based on the notion that traffic amount is largest around this time. Further, as people are in a stressful situation, they are more likely to make mistakes. As such, we can follow up on the argumentation that a combination of psychological stress and traffic prevalence is likely to be a combination that increases the likelihood of accidents. Interestingly, as we can see below, afternoon rush hour is a predictor of more severe accident status, which might allow the interpretation that awareness, rather than traffic prevalence and amount, is a more important predictor of severity status, since agents are not exhausted from work in the morning hours.

Looking at all the variables, we can assume that the model created a pre-selection process which is intuitive when considering natural traffic behaviour and characteristics.

3.2.2 Feature Selection on Serious Accidents

As it is applicable, we can observe that roughly similar feature variables are responsible for class assignment likelihoods as above, but, importantly, these variables are often of an opposite sign. As such, we can observe that also Vehicle Type, Engine Capacity, Speed Limits, Hour of Day as well as Weather Conditions are prominent feature variables. Interestingly, we can also observe a positive time trend on more serious accidents, implying that, overall, more serious accidents tended to increase throughout the observational period (which is, as stated, the opposite of the trend for slight and fatal accidents). If we neglect the time trends, then we can observe that an indication of motorcycles, Speed Limits, Rush Hour times as well as accident that occurred without involvement of multiple participants appear to have the strongest variation. In this case, the coefficients for all three variables are positive. This is also intuitive, and, more importantly, these observations complement the reasoning given when considering slight accident situations. Especially, the 20 mph zone and indication of motorcycles appears to be a strong predictor for more serious accidents, supporting the fact that non-car participants are exposed to a greater risk in being more seriously injured. Also, following the statement given in the data analysis part, one could argue that many more serious accidents occur when people are in situations in which they do not necessarily have to focus (by driving e.g. alone a non-maintained road after work without great traffic) or are in situations that inflict great stress on them (e.g. in the case of rush hour stress).

Overall, we can observe that mostly the hour of the day, speed limits, involvement of low capacity vehicles, extreme weather conditions or junction details within lower-speed environments (such as within towns) appear to be larger signals on accident classification when considering the two most prominent accident classes.

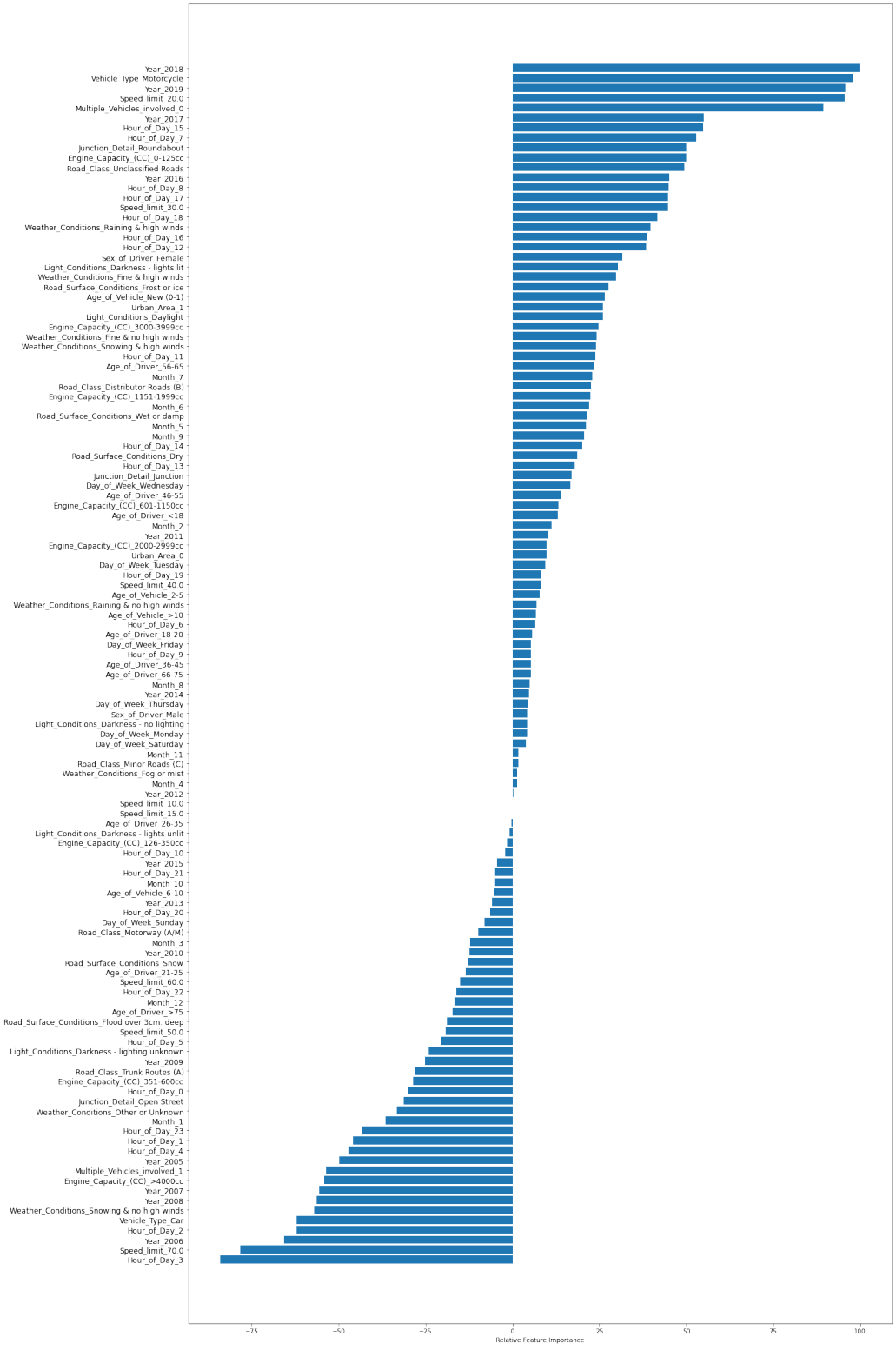

Figure 14: Feature Importance for Serious Accidents

3.2.3 Feature Selection on Lethal Accidents

Although our logistic regression model is unable to predict any observation to be classified as lethal, it might nevertheless be interesting to observe which features are, comparably speaking, more prominent in assigning this level. This is based on the notion that, despite stronger feature importance for the remaining two classes, one can still gain some understanding on which characteristics could signal lethal accidents but also happen to occur in situations which signal less serious levels. As such, they can be regarded as complementary.

In our case, we can observe that only relatively few binary feature variables obtain a substantially large importance. Interestingly, we can observe that the dispersion of feature variables is roughly mirrored compared to the case of slight accidents. As such, we can see that binary features with a relative importance of larger than 40 percent arise from 9 original features (i.e. Engine Capacity, Junction Detail, Age of Driver, Sex of Driver, Speed Limit, Hour of Day, Road Surface Conditions as well as Urban Area indicators). If we dig deeper into these features, we can observe that Engine Capacity of 0-125 cc, Junction Details as well as a Speed Limit of 20 mph are the strongest, and also all negative. As such, they can be interpreted as coefficients which indicate that these attributes cause the model to decline the likelihood to put an observation into accident level lethal. This is intuitive, as a low capacity observation might indicate small motorcycles involved in an accident, a participant less likely to be the causing factor of a lethal accident. Further, we see that small speed limits or urban area locations, as well as an environment including a roundabout, cause the model to decline the likelihood of an observation belonging to the lethal class. This is also intuitive, given that in these areas, the aforementioned participants are more likely to be prevalent and an agent's velocity is comparatively small. Another interesting observation is that females apparently cause less lethal accidents compared to their male counterparts and are thus also a declining factor for this class. Somewhat surprisingly, we observe that car appears to be a predictor which reduces the likelihood of lethal classification. This might be twofold. For one, compared to motorcycles, cars are more secure, can withstand a greater damage and protect a traffic agent to a greater extent. Further, prevalence of a negative sign for this category does not necessarily imply that it is less likely to lead to fewer lethal accidents, but rather that it is as, or even more, likely to lead to a different status based on the historical distribution of accident severity. As such, the model is likely to assign a negative sign in this class due to the disproportional prevalence of the variable in different classes compared to the class of interest. Ultimately, we can observe that larger engine capacity, higher speed limits as well as night time are important predictors of lethal accidents. All these factors are intuitive. On the one hand, we were able to show that situations including declining speed limits are an important predictor against lethal status. Accordingly, one should expect that the ability to follow a higher speed should serve as the opposite indication. In addition, one should argue that night time hours act as a factor for driver's awareness as well as road environment, both which, as previously argued, in case of a deteriorating status, are likely to be important predictors of severity status.

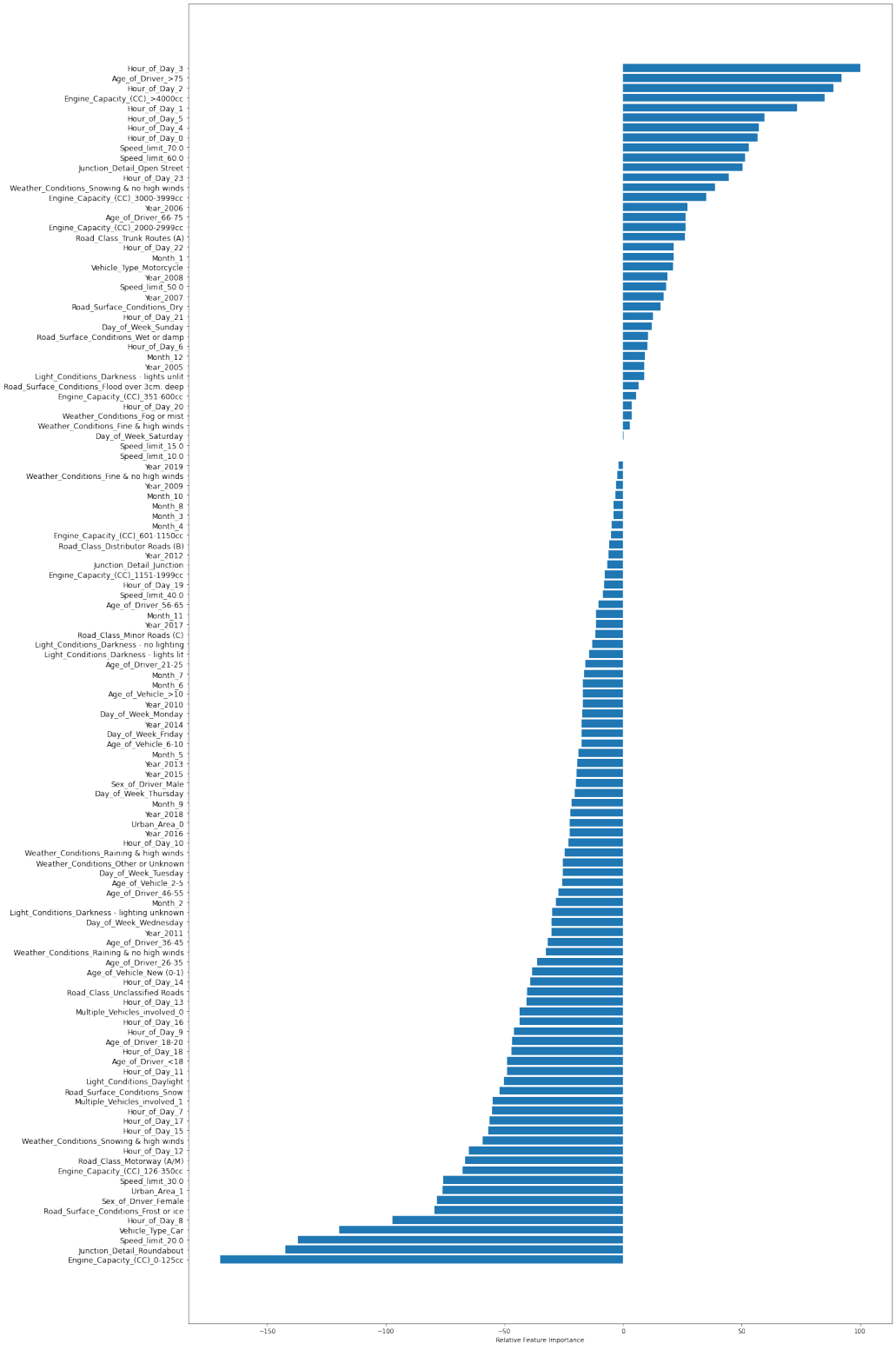

Figure 15: Feature Importance for Lethal Accidents

3.3 Feature Importance and Signals

Overall, one can observe that many of the originally selected variables appear to have an impact in assigning accident status, both on a general as well as a severity-specific level, although the latter appears to be more dominant. As such, it is apparent that, from this model, decision-makers should rather focus on factors which are prominent for one specific accident type, instead of attempting to mitigate accidents by improving feature conditions which are applicable for all levels. Especially in combination with the classification report described above, one must compare the overall prediction accuracy of the model with more nuanced levels and then create a combined metric in which one assesses both prediction levels as well as feature importance. In the case of a logistic regression framework, it might be most useful to focus on slight and serious accidents, while incorporating less feature types for serious levels, as the model appeared to be less precise and induced a lower recall metric.

It is important to state that we only partially controlled for multicollinearity issues through correlation but did not execute some forms of feature dependence measures. Also, we did not work on any imputation strategies such as given by tree bases models. In this case, we solely looked at variable importance from a static perspective and disregarded, to some extent, the step-wise feature selection process. As such, we did not fully account for the issues mentioned above. Thus, although the features selected by the logistic regression model to have the largest impact on level assignment appear to be intuitive, there still might be some statistical caveats which cannot be fully controlled for by using this type of model.

4 Conclusion

In conclusion, our paper provides the following policy implication suggestions for insurance companies. Firstly, an increase in the premium for highly motorized vehicles. This could be in the form of a (vehicle type dependent) gradual increase using an exponential weighting factor in order to penalize very highly motorized vehicles more. Secondly, we suggest an increase in the premium for drivers between the ages of 18 and 35. This could also be done in combination with years of driver's experience, which would require further research. Furthermore, insurance companies may want to increase their premium for nighttime driving. This could for instance be implemented on a subscription basis by which drivers pay a premium to be ensured on each night they drive. Another possibility would be to vary the cost coverage depending on when an accident incidence occurred. Additionally, we suggest geographic specific premia, whereby premia could depend on zipcode or other geographically bounded areas (i.e. bounded by some kilometer radius), which could be determined using a geographically weighted regression or cluster analysis. Which radius or area bound should be chosen as well as the specific premia for those areas is a further research opportunity in this respect.

Economic implications of this would be a reduction in overall accidents, in particular more severe accidents as well as creating an incentive for the use of public transport, especially if driving (through higher insurance costs) becomes more expensive.

With regards to broader policy implications for the UK governments, we propose the following. We suggest an overall better maintenance of trunk routes in order to tackle accidents in general. Further, we propose a government law and registry in which trunk routes must be assessed for their condition and taken care of if needed every few years. The optimal number of years between restorations should be assessed by an expert. In addition, we suggest better lighting of trunk routes, especially at night and in bad weather conditions. Perhaps this lighting could be adjustable to the surroundings and environmental conditions (i.e. natural light and visibility). We also recommend public educational programs and campaigns to enhance the drivers' awareness of the effect of tiredness and impaired driving on accident severity. Such programs could include visual presentations as well as data presentations and could possibly already be started in the last year of High School and then be continued - perhaps even made mandatory. To go even further, a repetition course, which assesses driving behavior could be made mandatory every few years, in which different tests are administered, collecting data on drivers' traits and then using this data for further research. Furthermore, governments could perceive "safety while driving" as a public good and subsidize companies doing research on technologies in the interest of driving safety (such as new technologies in the field of automation), which would prevent accidents in the first place. Moreover, we advise the government to promote public transport during known rush hour times, for instance by closing certain lanes off to non public-transport vehicles. This would firstly allow for an increased number in public transportation vehicles (such as busses) and secondly drivers would, anticipating more traffic, think twice before deciding to commute by car. Regarding this point, we recommend that governments should foster public transport in general as it would simultaneously reduce the number of accidents and be in line with the governments' climate goals. Moreover, it may be beneficial to create forward-looking conditions in which travel in general, whether by public transportation or by personal vehicle, is less necessary. This would include flexible working conditions (such as home office possibilities) and improved delivery services (for shopping). Especially in the current times of the Covid-19 pandemic, such considerations are becoming more and more importance. Lastly, the government could incorporate traffic cycle predictions or customized and flexible traffic route generators into the governmental website or provide such applications via mobile apps. This could for instance be undertaken in partnership with a GPS provider like Google, which already has a plethora of data for such use.

Economic implications of this would as well be a reduction in overall accidents, which entails in lower ex-post monetary costs (costs for insurance companies, hospital costs, etc.) as well as a reduction in deaths. This would increase welfare and general health. Especially in times of a worldwide pandemic, where hospital capacity is extremely limited, avoiding unnecessary accidents is critical.

In terms of the prediction and prediction models it is crucial to get a somewhat even prediction accuracy across all classes, which is tackled in this paper by using balanced datasets for training. In this respect further work could be done, for instance by employing a weighting scheme. Such prediction models could particularly be utilized for setting highly specific insurance premia, which are anyway likely to become more common in the age of Big Data. This could be done for both car as well as health insurances. Moreover, further work could

be done in predicting not just the severity, but also the frequency of accidents in specific regions. This would allow the government to take prescient actions, such as increasing the hospital capacities and other performance driving factors for hospitals (e.g. staff and equipment) in such areas.

Although our results rely on a rather large dataset, we recognize that there are nevertheless certain limitations. One of them being the fact that the dataset only covers the legal area of the UK and therefore it is possible that our findings lack of external validity. One possible further approach could be to conduct a meta-analysis, aggregating multiple studies from different countries to test for replicability. Furthermore, the used dataset focuses solely on cars and motorcycles, while other forms of transportation are dropped. These may be of particular interest and may provide further insights on accident severity and traffic accidents in general.

In conclusion, our paper provides various policy implications for both, insurance companies as well as governments, and dives into the economic implications within the topic of traffic and accidents. We acknowledge that a lot of further work is to be done and highlight some research opportunities. Research in this area is particularly imperative due to its significant impact and is becoming even more important in the age of Big Data. Fortunately, Big Data is also one of the most important factors that facilitates such analyses.

References

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, (pp. 272-276). IEEE.

Beshah, T., & Hill, S. (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. *AAAI Spring Symposium: Artificial Intelligence for Development* (Vol. 24, pp. 1173-1181).

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention, 90*, 128-139.

Department for Transport of the United Kingdom (2020). *Road Safety Data.* Retrieved 23. December 2020 from https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Department for Transport of the United Kingdom (2019). *Road Lengths in Great Britain 2019.* Retrieved 7. January 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/860685/road-lengths-in-great-britain-2019.pdf

Krishnaveni, S., & Hemalatha, M. (2011). A perspective analysis of traffic accident using data mining techniques. *International Journal of Computer Applications, 23*(7), 40-48.

World Health Organization. (2020). *Road traffic injuries.* Retrieved 13. January 2021 from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Appendix

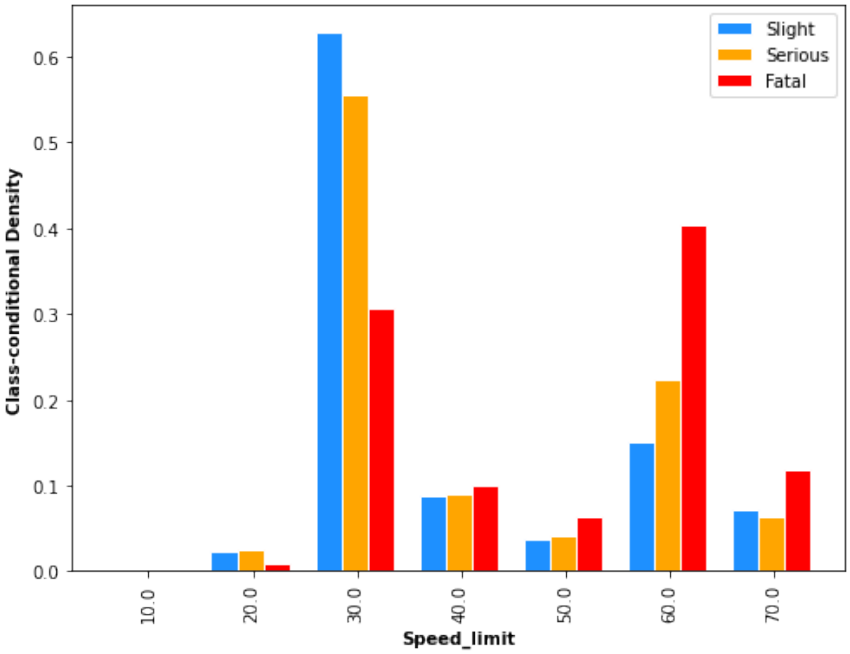

Figure 16: Speed Limit Distribution

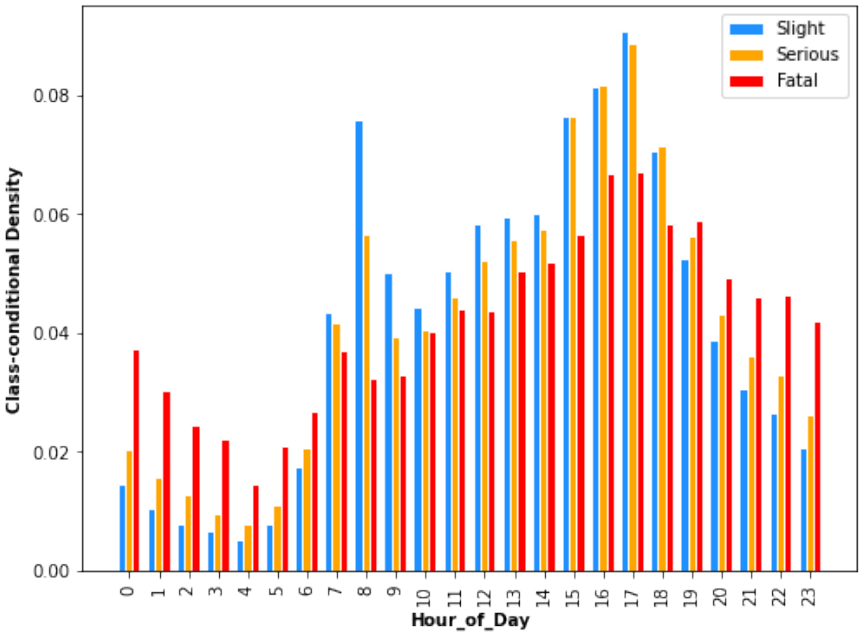

Figure 17: Daytime Distribution

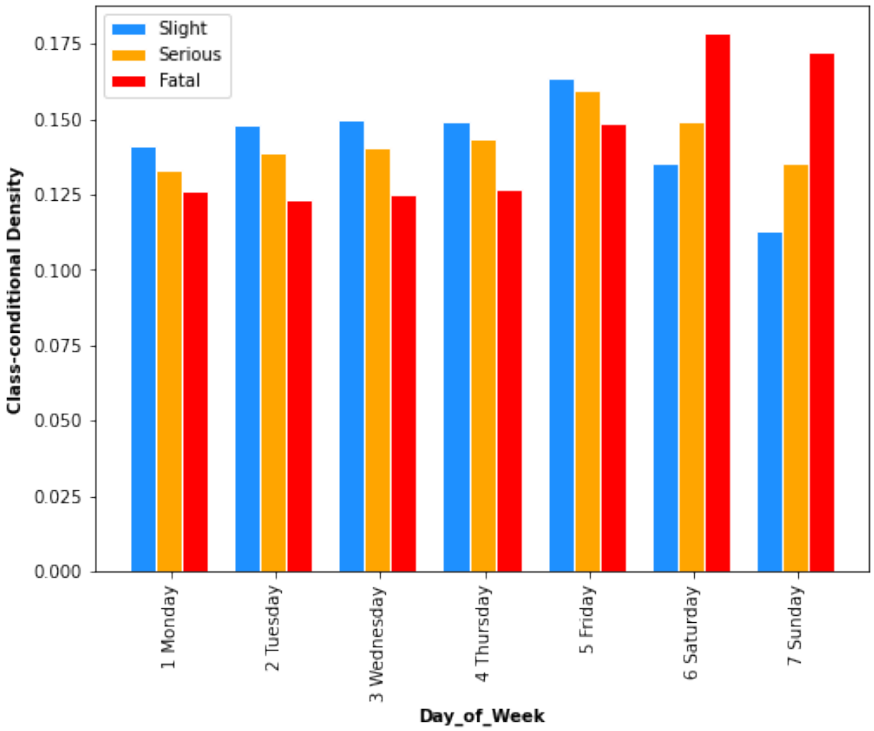

Figure 18: Weekday Distribution

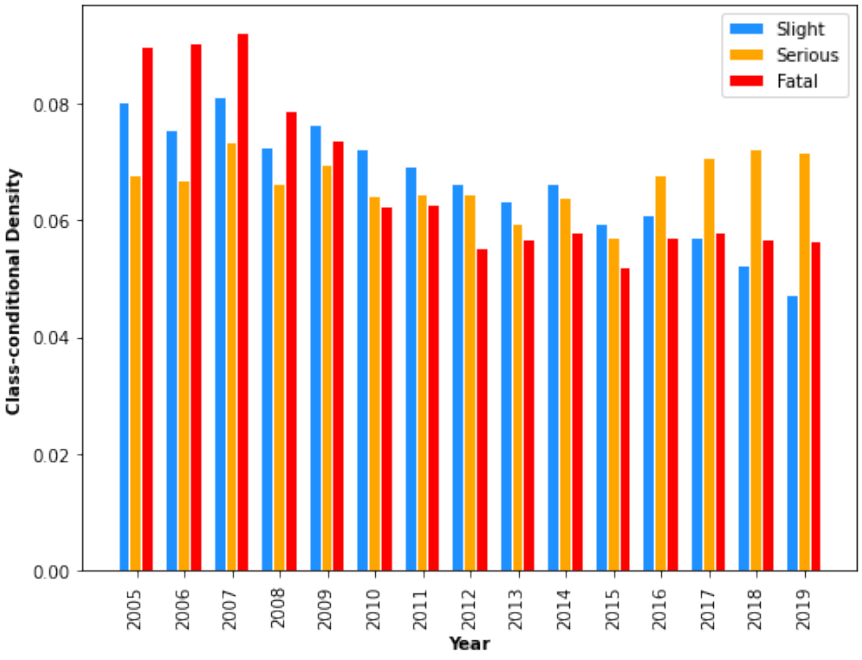

Figure 19: Year Distribution

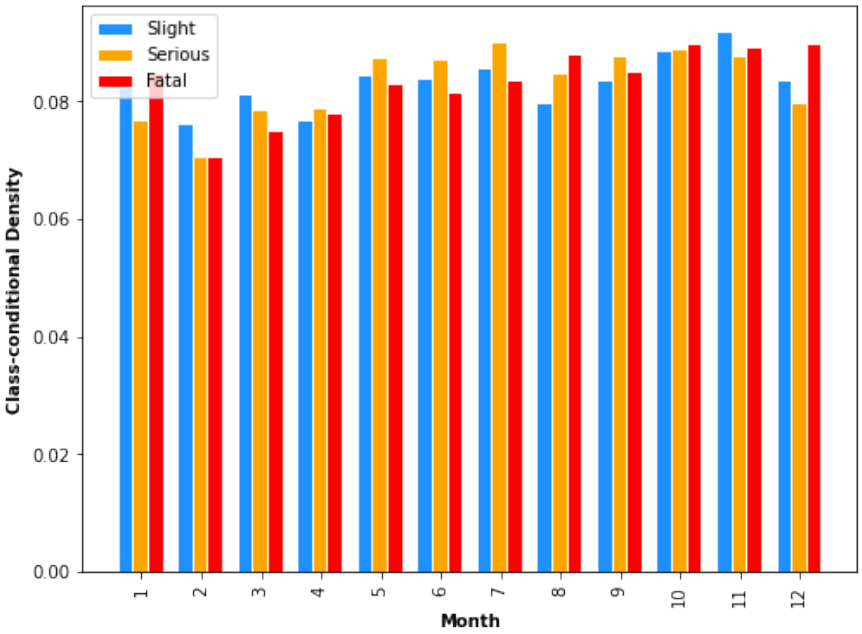

Figure 20: Month Distribution

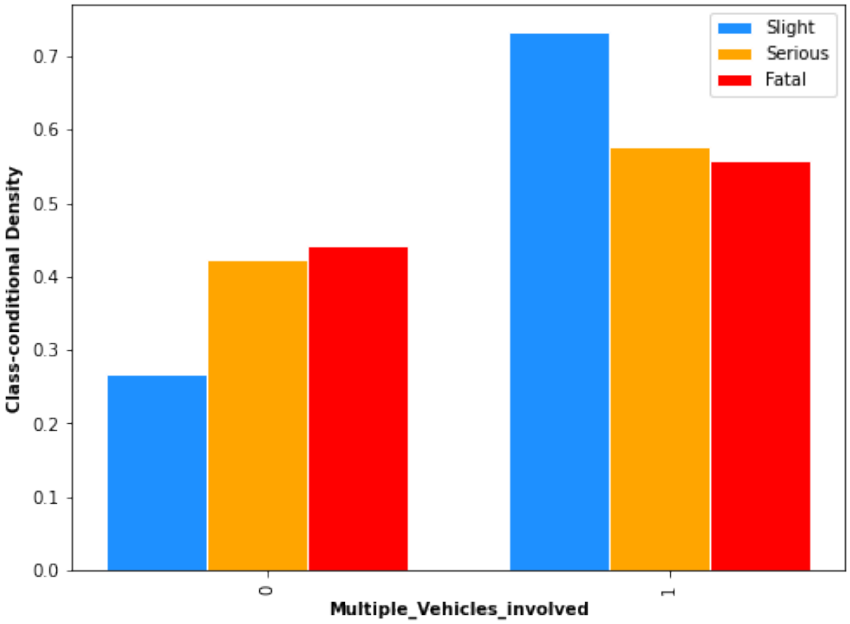

Figure 21: Involved Vehicles Distribution

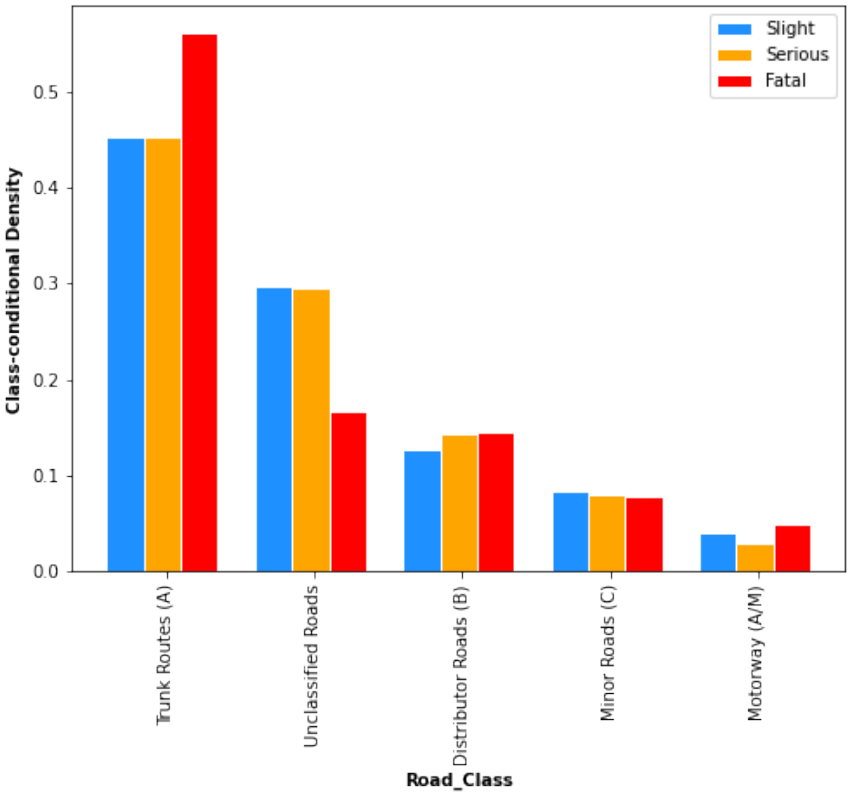

Figure 22: Road Classes Distribution

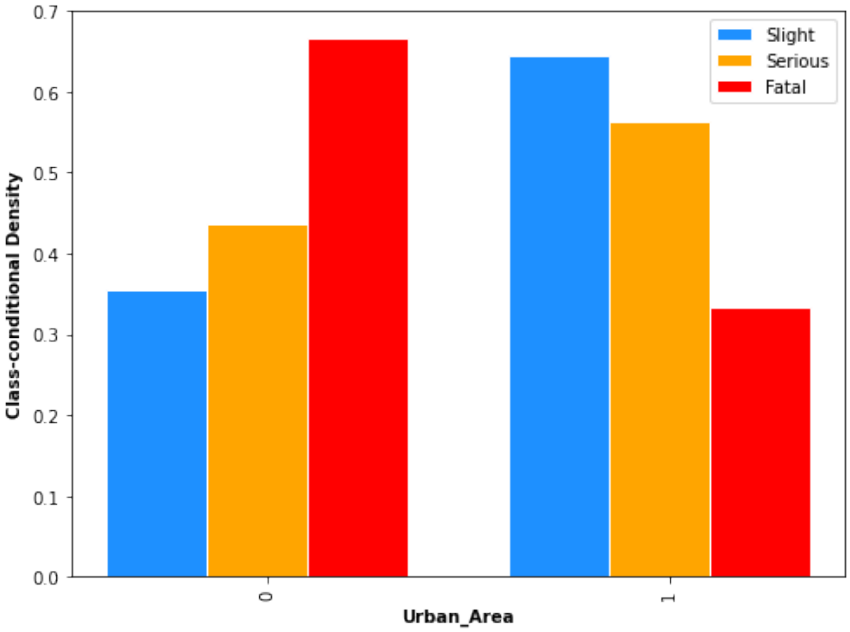

Figure 23: Urban-Rural Distribution

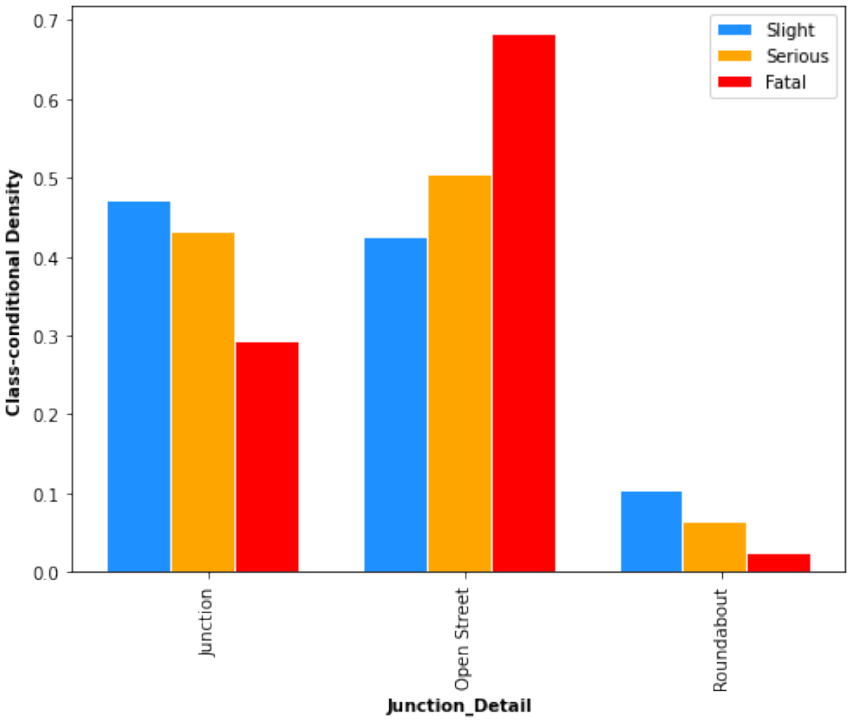

Figure 24: Junction Detail Distribution

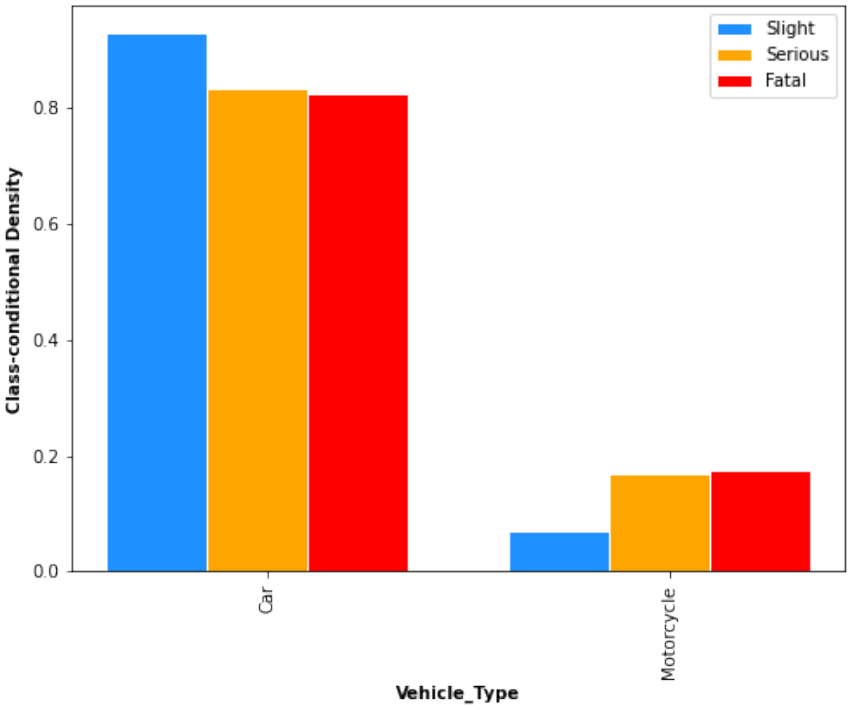

Figure 25: Vehicle Type Distribution

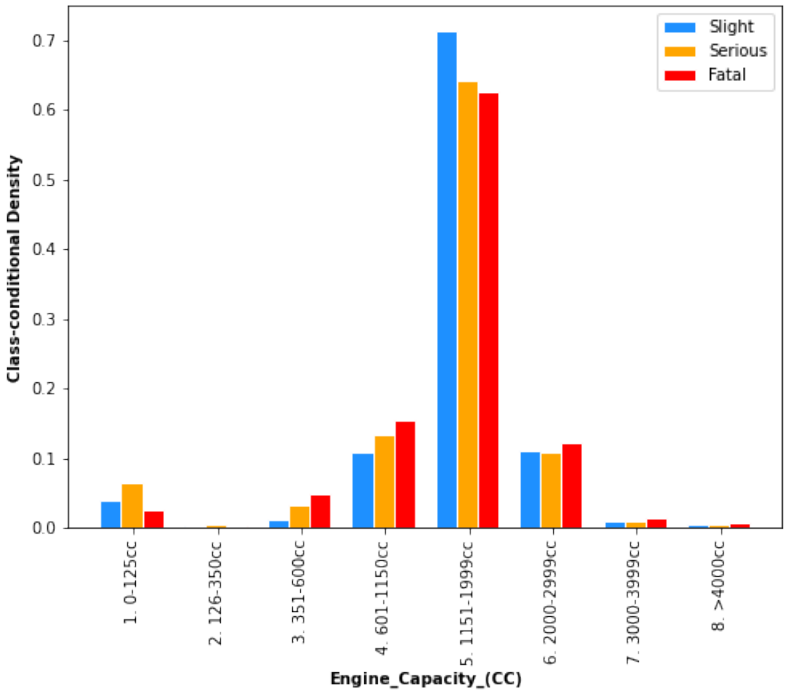

Figure 26: Engine Capacity Distribution

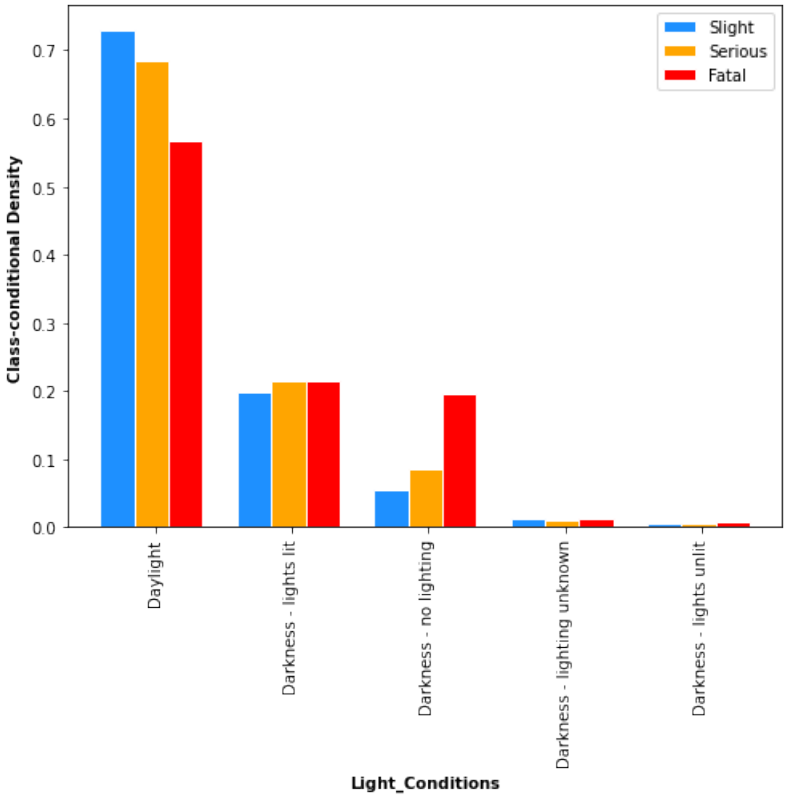

Figure 27: Light Condition Distribution

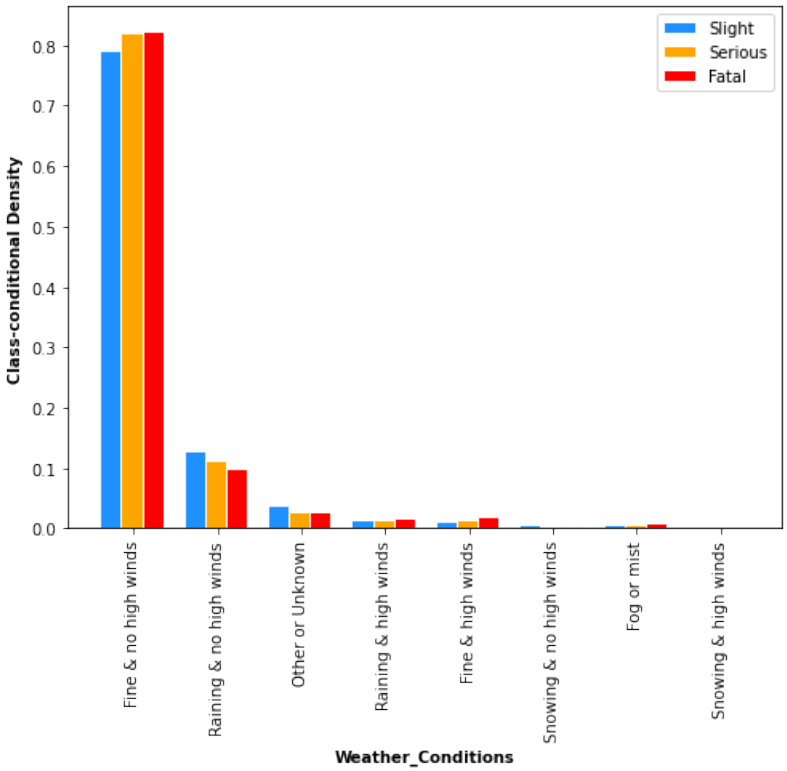

Figure 28: Weather Condition Distribution

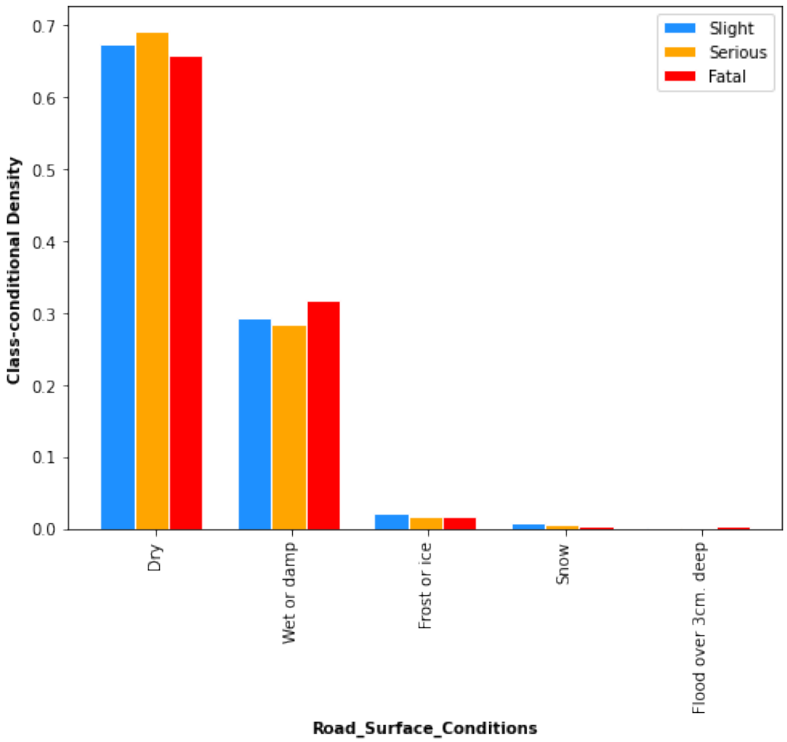

Figure 29: Road Surface Condition Distribution

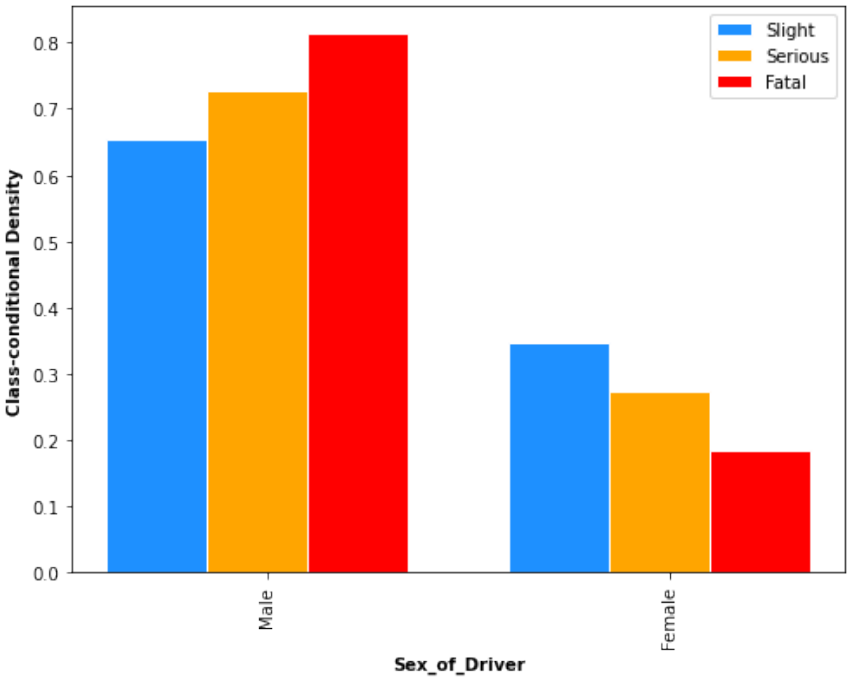

Figure 30: Gender of Driver Distribution

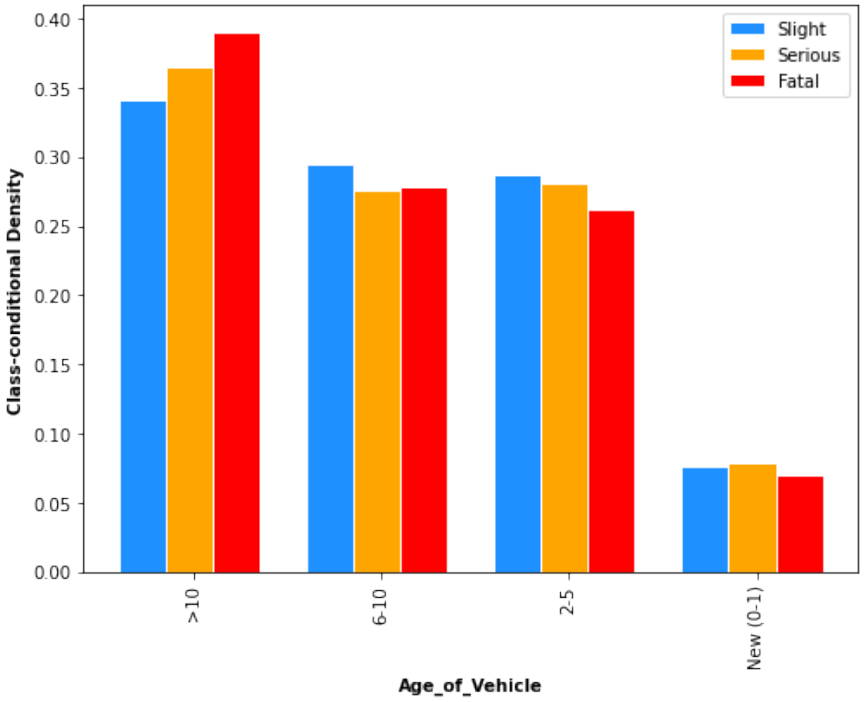

Figure 31: Age of Vehicle Distribution

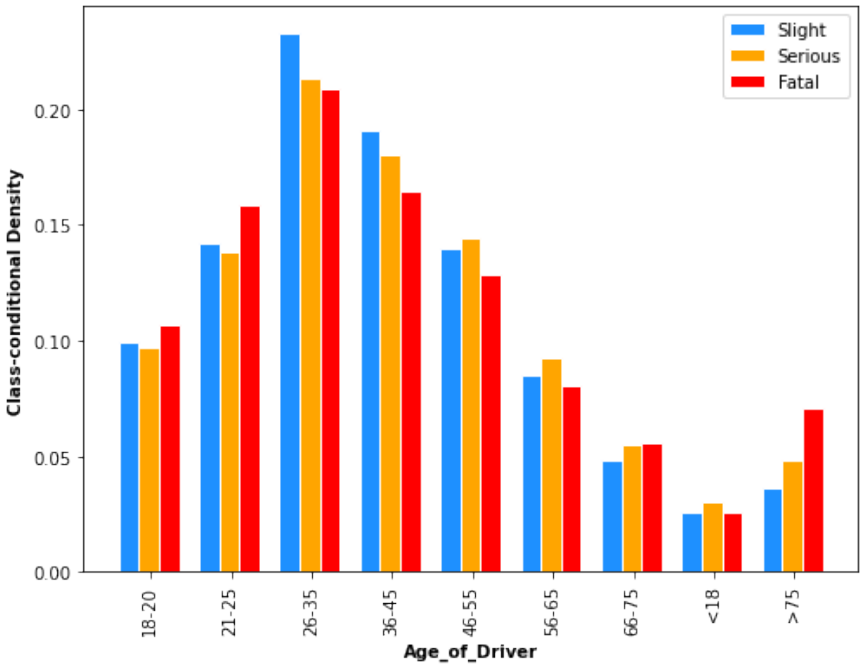

Figure 32: Age of Driver Distribution