Predicting The Likelihood of Severe Accidents

Dona Ray

April, 2018

1. Introduction

Car accidents are an inevitable problem when it comes to driving. We may get into one, or may be indirectly affected by one, due to long traffic delays an accident miles down may have caused. Whether its a minor or a major accident, there are substantial costs to be borne by both parties involved.

Fatal car accidents alone cost the state of California tax payer more than $4 billion a year, according to the Center for Disease control and Prevention. Apart from medical bills and emotional loss to the victim's family, there are other costs like, loss of productivity, administrative costs, property damage, unrecoverable costs, increase in insurance, taxpayer costs and increase in research for new safety features. In addition, there are police costs relating to the the time spent in attending and reporting accidents by officers. There are also non-economic costs, like road damage and environmental devastation, leading to more taxes to do repairs and cleanup. It also has indirect costs for those not involved, because crashes also mean traffic delays, resulting in loss of productivity at work as well need for extra resources for traffic intervention.

A new study by the World Bank, found that reducing road accident fatalities and injuries could lead to substantial long-term income gains for low to middle income countries. The study finds that countries that do not invest in road safety, could potentially lose about 7% to 24% in GDP growth over a 24 year period. Using data on traffic fatalities and economic indicators for 135 countries, the study concluded that reducing traffic deaths by 10% could lead to an increase in per capita real GDP by 3.6% over two and half decades.

2. Data

Clearly then, increasing road safety measures leading to lower fatalities and reducing the severity of accidents is welfare improving for all. To recommend policy to the authorities, we could first need to understand the key factors that cause serious accidents. To answer this question, we analyze road accident data from UK, with information about injury related accidents that were reported to the police in the year 2016. Our goal is to analyze the data to understand the various factors that cause serious accidents. Specifically, the idea is to look into and understand what drivers are highly correlated with the probability of an occurrence of a fatal or serious accidents.

This is interesting from many perspectives. For the local authorities, it is imperative to know what are the factors that cause accidents, so that they can look into the possibility of implementing policy that will lead the the reduction of such occurrences. For this dataset, the number of accidents at speed limit 30 is uncharacteristically high, almost 85% of the total number of accidents. One policy recommendation may be then to reduce the limit. Again, if weather conditions lead to an statistically significant increase in the probability of accidents, the recommendation would be for authorities to set lower limits during bad weather or install additional road signs. We also need to consider other factors that may be correlated with the occurrence of an serious accident like time of the day or week. Demographic data will give us further insights into the reasons that lead to serious accidents. In particular, we look at age, sex of the driver, and information about the area where the accident occurred, whether rural or urban.

The data is from UK, downloaded from the website "data.gov.uk". The files provide detailed road safety data about personal injury accidents in Great Britain that were reported to the police in the year 2016. There are three data files, Accidents, Casualty and Vehicle and each file has a column with an accident ID. The Casualty and Vehicle data also have a unique vehicle reference number. The zipped data file consists of the following three datasets in CSV format: accidents, vehicles and casualties.

2.1 Accidents

This dataset has 31 columns and 136621 rows. Some features have missing or ‘Nan’ values. In particular, LSOA_of_Accident_Location', 'Location_Northing_OSGR', 'Location_Easting_OSGR', 'Longitude', 'Latitude' have missing values. These features, which have information about location, latitude, longitude or county will not be used in the study and have been dropped. In addition to these features, ‘Speed Limit’ and ‘Time of Accident’ 37 and 2 missing values. Clearly, these are important variables for the analysis. There are several ideas that have been suggested in the literature on how to deal with missing values. We could impute them by a summary statistic like mean or median or do a simple linear regression to predict the missing value. But, that is suboptimal, since not only are we using information already in the data, but adding some noise at the same time. We cannot leave it as it is, as these features will most likely be predictive in the supervised learning algorithm, and creating separate dummy variables for a few observations, 37 and 2 respectively, will not be predictive. Since it’s a small number as a percentage of the dataset, the corresponding 39 rows have been dropped. After deleting the 5 columns and 39 rows, the Accident dataset now has 26 columns and 136582 rows. Some variables like ‘Road Conditions’ or ‘Light Conditions’

have rows (few hundreds to 1) with missing information, indicated by -1. These are categorical variables, and one possibility is to simply create a separate category, and then drop them from the features dataset during modeling to avoid multicollinearity.

2.2 Vehicles

The vehicle dataset has 22 columns and 252,000 rows. There are no missing values indicated by ‘Nan’, but several variables have missing values indicated by ‘-1’. The home area type of the driver (urban, rural or small town) has 47,490 missing values, while “Age Band of the Driver” has 29,418 missing values indicated as ‘-1’. As before, these will treated as a separate category while creating dummy variables during modeling.

2.3 Casualties

The Casualty dataset has 15 columns and 181384 rows and no variables with missing values indicated by ‘Nan’. But several variables like ‘Home Area Type of Casualty’ have missing values indicated by ‘-1’.

2.4 Outliers

There are 25 data points with number of casualties more than 12, and 13 data values where number of vehicles involved is 10 or more. Clearly, these are outliers, but a very small percentage of the overall dataset. Moreover, while some supervised learning algorithms like Logistic Regression are sensitive to outliers, there are several that are robust to outliers, like Decision Trees and Support Vector Machines.

2.5 Data Merge

Each dataset has a ID variable, that uniquely identifies the accident. This is unique in the Accident dataset, but may have non-unique values both in the Vehicle and the Casualty datasets respectively. This is because a particular accident may involve one or more vehicles and casualties. To proceed with our analysis, we would then need to merge the three datasets. The first merge involves the Vehicle and the Casualty data sets. Its an outer join, on two indices, namely, Accident ID and the Vehicle Reference column. The second merge involves this merged dataset with the Accident data. This is also an outer join, but now on the index Accident ID. The final merged dataset has 252,000 rows.

3. Data Exploration

The accident data has information on road conditions, speed limits, time and date as well as information on weather conditions at the time of the accident. We look at some bar charts below to understand some of the key as well as interesting features. The y-axis shows the frequency or the total count of accidents and the x-axis plots the categorical variable. For example, Figure 1 has the different speed limit categories on the x-axis, and the y-axis shows the total number of injury-related accidents that were reported to the authorities . It can clearly be seen that the mode is at 30, which is about 62% of all accidents, followed by speed limit 60, at about 13.5%. A significant number of accidents occur at T-sections and crossroads, although the latter is almost half of the number of T-sections. In addition to T-sections roundabouts are also susceptible to accidents.

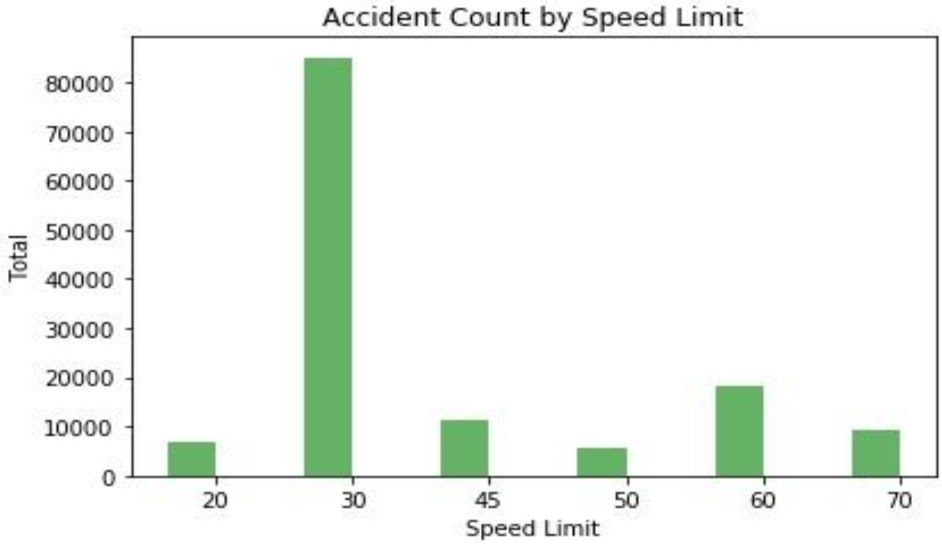

Figure 1

It is possible that certain weather patterns may make roads unsafe for driving. Figure 3 below shows that 99,181 or 72% of all injury-related accidents occured in fine weather with no winds, rain, or snow. It seems that drivers do drive more carefully during bad weather or possibly there are fewer drivers on the road when is rain, snow, fog or a combination of these patterns. Another interesting question is if more accidents occur during the day or at night when the lighting may or may not be poor. From figure 4 we can conclude that, most accidents occur when daylight is good. Clearly, poor lighting is

not a factor. In this dataset, 112,210 or 82.2% of overall injury-related accidents that were reported to the authorities occurred during daylight.

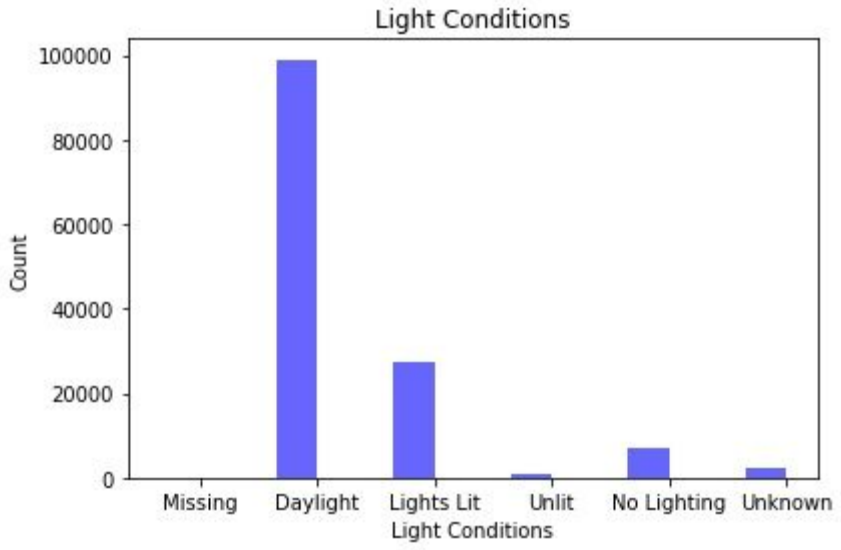

Figure 2

We next analyze time or the frequency or count of accidents by day of the week. We expect more traffic on Fridays, which would lead to more accidents. The data confirms this, but surprisingly by very small margins. While the peak is on Friday, followed by Thursday, with minimum on Sundays, accident count by days of the week is an approximate uniform distribution. We would also like to understand if there is a discernable pattern for number of accidents over the months of the year. As expected, like the days of the week, the number of accidents is approximately uniformly distributed over the months of the year. November and January seem to be slightly higher, but not by much. Since days and month do not differentiate, we next see if time of the day plays an important role. The bar chart depicted in Figure 3 shows the mode is between 5:00 pm and 6:00 pm in the evening[1]. This is followed by 4:00 pm to 5:00 pm and then surprisingly a close third place between 3:00 pm and 4:00 pm and in the morning between 8:00 am and 9:00. 3:00 pm. This is possibly because people seem to be more careful when driving to work as compared to on their way home from work.

[1] The frequency at 5:00 pm in Figure 5 is the total number at that hour, or between 5:00 pm and 5:59 pm.

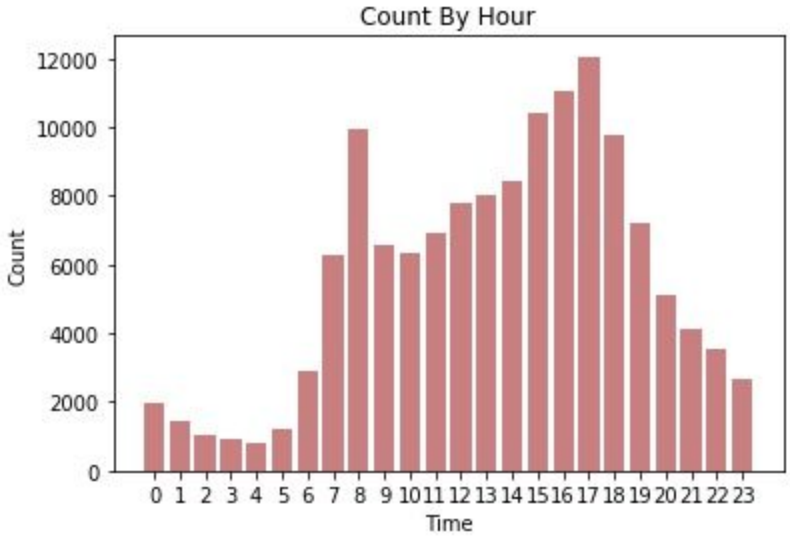

Figure 3

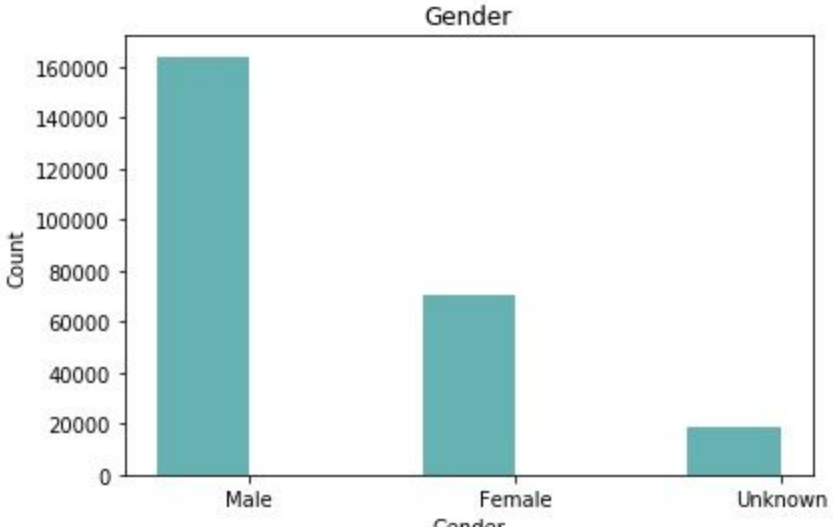

Figure 4

In the bar chart depicted in teal, we will try to understand if demographics play a role in influencing the probability of being involved in an accident. The bar chart on gender in Figure 4, shows that males are twice as likely to be involved in a car accident compared to females. Note that in this dataset, each data point relates to an occurence of an injury-related accident that was reported. The data does not have information about all

the cars that were on the road, or about all drivers. It would be erroneous to infer that males make worse drivers, since we do not know the proportion of male and female drivers in the same time period.

A frequency distribution of the age distribution shows that the maximum number of accidents occur in the age group [26,35], followed by [16,25] and [36,45] respectively. Again, this is unsurprising, yet interesting. More people in the age bands, [26, 35] and [36, 45] are probably drivers. We cannot answer this with certainty, unless if analyze the overall distribution of drivers by age. But, assuming, the age band [16, 25] will have fewer drivers, the number of accidents is disproportionately high. Insurance companies in this scenario seem to be justified in setting higher premiums to this group.

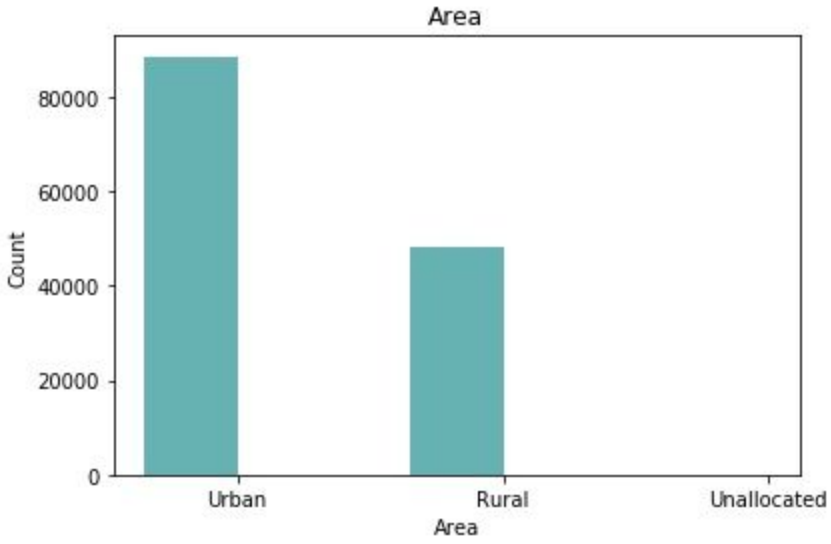

Figure 5

The final chart, Figure 5, on demographics compares the number of accidents in urban and rural areas. The count in urban regions is almost twice as more as in rural areas. Given that urban areas are more densely populated, the numbers on this bar chart is not surprising.

3.1 Comparison by Slight Vs Severe or Fatal Accident

While the above plots are informative, we can gain further insights by comparing the frequencies in two categories “Slight” and “Severe or Fatal”. We would also like to understand the distribution of the counts in each category of the target variable.

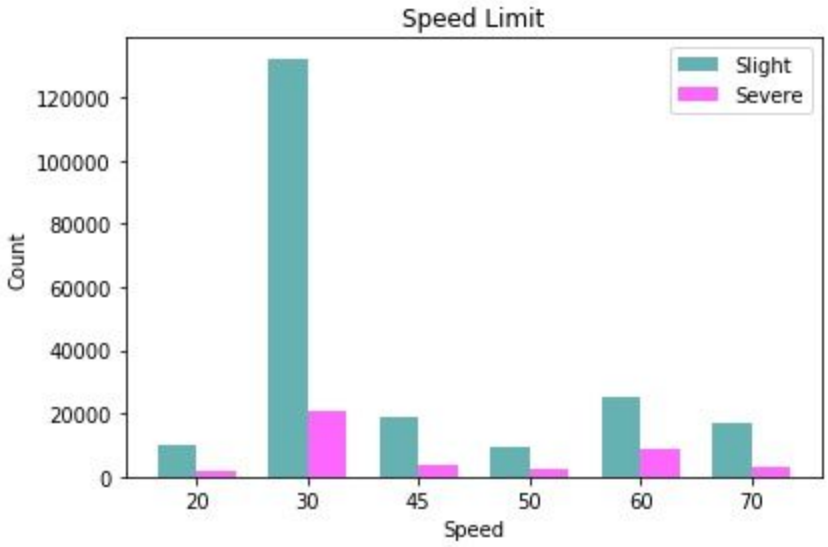

Figure 6

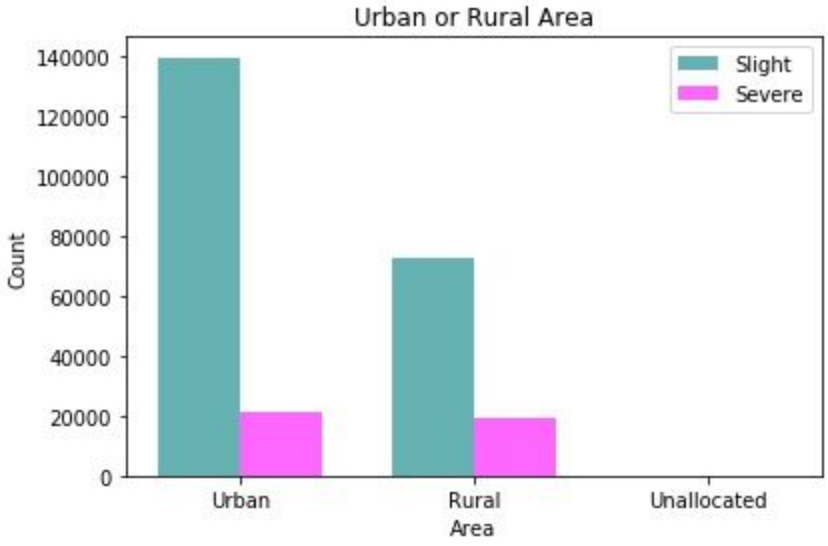

Figure 7

In figure 6, number of accidents are shown by speed limits. But now, we can compare the total number of “Slight” and “Severe” accidents in each of the speed categories. The teal color shows the count of “Slight” accidents while the magenta color shows the count of “Severe” accidents. Note that while absolute number will be less, we will gain insights by looking at relative proportions between categories. As previously noted, the maximum number of accidents occur at speed limit 30. We can see that the proportion of “Severe” accidents is higher for each of the subsequent categories. The proportion is maximum at speed limit 60, the number of “Severe” accidents being a little over one-third of the total number. Higher speed limits do make accidents dangerous, but we note an interesting fact that the proportion is less than one-fourth when the speed limit is 70.

For the different junction categories, the ratio does not change much. For each group, T-section or crossroads, or slip road or something else, the ratio is close to one-fifth. We also study the proportion of slight and severe accidents by gender. While it is higher for males, it is only marginally so. In other words, its about one-fourth for males, and close to one-fifth for females. And finally, the last bar chart, Figure 7, shows the counts of “Slight” and “Severe” accidents by urban and rural areas. While the number of overall accidents is double in urban areas as compared to rural areas, the number of “Severe” accidents is almost equal in both areas! The ratio in rural areas is approximately one-third, while in urban areas its a little over one-sixth.

3.2 Inferential Statistics

To develop an effective model for prediction, we would like the feature variables to be independent or uncorrelated with each other. In addition, good predictors will be highly correlated with the target or endogenous variable y. The dependent variable y takes the value 1 if the accident is categorized as severe, and zero otherwise.

Prior to fitting our data to our model to make predictions, we will try to understand the relationship between the variables and the target, as well as between the features themselves. For a classifier like the Logistic Regression, we would like the features to be independent. The intuition is that if there is collinearity between two features, the classifier is not learning any new information from both features. In the extreme case, if two features are perfectly correlated, there is no additional information by including the second feature. Note that, independence is a stronger condition than correlation, since correlation refers to a linear relationship. Two variables may be uncorrelated in a linear

sense, but have a non-linear relationship. So, when we want to make decision about the feature space, clearly, one condition is that they need to be independent. Secondly, we should include features that are highly correlated with the dependent or endogenous variable, the target variable we would like to classify. In our example on the accident data, the problem is to classify an accident as slight or severe. Or more precisely, to understand the factors that may cause one.

3.3 Hypothesis Testing

We first analyze two demographic variables gender and type of area, urban or rural. We would like to understand if they are correlated with the target variable. If correlation is present, then these variables will be good predictors in the classification model.

1. **Target and Gender**: We will test the null hypothesis that the severity of accident and the sex of the driver are uncorrelated or independent. These are both categorical variables, which implies we will use the test of difference of proportions. The sample size is large and *np > 5*[2]*, n(1-p) >5*, so Central Limit Theorem can be applied. Using the formula for the z-statistic under the null hypothesis that both proportions are equal, the value of the z-statistic is: $z = 30.93.$ Because $30.93 > 2.58$, we reject the null at 1% level of significance.

2. **Target and Area Type:** We also test the null hypothesis that area type and target are independent. As before, because *np, n(1-p) > 5 ,* Central Limit Theorem can be applied. The z-statistic can easily be computed and is: $z = -49.08.$ This a two-tailed test, so we reject the null again at 1% level of significance, because $-49.08 < 2.58.$

   From the test results above, both features, sex of the driver and area type of the location of the accident, urban or rural, is correlated with the target[3]. The results imply that male drivers are more likely to be involved in a serious accident. And occurrences of serious accidents is higher[4] in rural areas.

3. **Target and Speed Limit:** We would also like to analyze the relationship between speed limit and the severity of accidents. Speed limits have 7 categories, 20, 30,

[2] Sample size is denoted by n and p is the proportion or the number of males in the sample with target = 1. Note that np > 5 has to be satisfied for all four categories for a two variable test.
[3] The target variable is the likelihood of a serious accident as opposed to a slight one.
[4] Both the hypothesis can be statistically tested by non-parametric methods like Permutation tests.

45, 50, 60 and 70. To test the independence of two categorical variables, where at least one of them has more than two categories, (in this example, speed limit), we will employ the chi-square test of independence. The assumptions are the same as in the previous two examples, each category must have at least 5 data points. The test statistic can easily be computed and is: $\chi 2 = 3246.26$. Comparing this to a $\chi 2$ distribution with 6 degrees of freedom, we reject the null at 1% level of significance, because 3246.26 > 16.81.

The above tests indicate that gender, area type and speed limit will be good predictors in a supervised classification problem where we classify our target variable as a severe or a slight accident.

4. Modeling

Our goal in this paper is to understand and identify the factors that cause a serious accident. The data includes the target variable, this is a supervised classification problem. We then define our target or dependent variable y as follows: if the accident is categorized as serious (serious injury or fatality), y takes the value 1, and if the it is a slight accident, y is zero. The feature space X is an n x k matrix, where k is the number of features or variables, and n is the number of data points. The independent variables are the categorical variables like weather, road type, junction type, road surface conditions, speed limits, as well as demographic variables like sex age of the driver. It also includes non-categorical discrete variables like time of the accident and age of the driver.

4.1 Cross-Validation

For a simple model with very few features, the model may fail to capture important patterns in the data. When this happens, the model is said to be under-fitting or biased. On the other hand, as a model gets more complex, it may pick up patterns in the data that do not generalize to the overall population. In this scenario, we say that the model is overfitting the data, because it is finding interesting but chance occurrences in the data that do not generalize. This is also referred to as a model with high variance. While we clearly do not want our model to be biased[5], models with high variance[6] are also

[5] Models that do not include relevant or predictive features are biased.
[6] Models that include irrelevant or features that are not good predictors of general patterns in data have high variance.

undesirable. One reason why overfitting may occur or go undetected is if training and testing is done on the same dataset. To avoid such scenarios, we split our data into a training and testing or hold out data set. Testing the model on a hold out data that the model has not been trained on, ensures that the model does not get learned on occasional erroneous labels.

We randomly split the data into a 80% training set and a 20% testing set. We then use a k-fold cross-validation technique[7] on the training set. The k-fold cross-validation then splits the training data into k folds, and uses (k-1) folds in each iteration for training the model. At the end of the cross-validation process, each data point is used only once for testing but k-1 for training. We can then evaluate the performance of the classifier by taking an average of the accuracy score of all the models of all k iterations.

4.2 Naive Bayes Gaussian

We first train a Naive Bayes Gaussian classifier to the data as a benchmark model. In spite of its simplicity, and its assumption of conditional independence[8], this model often performs well in supervised learning problems. The target variable is skewed with about 25% of y values with value 1, or classified as a serious accident. The goal of this paper is to identify the factors that cause or predict the likelihood of a serious accident. If the data is unbalanced, then the optimizing algorithm may identify most of the non-serious accidents. This could lead to a very high R-square or accuracy score. In fact, training the classifier on the unbalanced data, gives a high accuracy above 80%, but correctly classifying less than 10% of the serious accident. To avoid this, we balance the data, either by doing random sampling with replacement prior to training, or by incorporating it in the classifier by passing an argument classweight = 'balanced'[9]. For the Naive Bayes algorithm, its not possible to pass a classweight ='balanced' in the classifier while training the model, so we need to do this prior to training the model. We use random sampling with replacement to rebalance our data so that both target classes are now weighted equally.

[7] In our model we use k=5

[8] Naive Bayes assumes conditional independence in the feature variables. Nevertheless, this is a weaker assumption than the full independence assumption implicit in Logistic or Linear Regression.

[9] While the classweight='balanced can be easily incorporated in the Logistic Regression and Decision Tree classifiers as an argument, for the Naive Bayes and the KMeans classifiers, the data had to be resampled prior to training.

Naive Bayes	Classification Report			
Target	**Precision**	**Recall**	**F1-score**	**Support**
y=0	0.87	0.70	0.78	42409
y=1	0.23	0.46	0.30	8078
Avg/Total	0.77	0.66	0.70	50487

Table 1

From Table 1, the Naive Bayes classifier has a precision of 87% of the 'slight' accidents and about 24% for 'severe' accidents'. Precision is the ratio or percentage of correct predictions, of all predictions in that category. This implies the model is correct 87% of the times it predicts y=0, but only about 24% of the times it predicts y=1. For our problem, we are particularly interested in the 'severe' accidents or when y=1. The recall values indicate the ratio or percentage of correct prediction of all instances of that class label. The Naive Bayes estimator then correctly classifies 70% of the 'slight' accidents and about 46% of the 'severe' accidents. This prediction is on the 50,487 sample data points of the test data, or the hold-out test data. The classifier is doing well in identifying 'slight' accidents, and while correctly labeling close to half of the severe accidents. The overall accuracy is about 66.23.

4.3 K-Nearest Neighbors (k-NN)

K-nearest neighbors or k-NN is a non-parametric[10] method of estimation for both regression and classification formulation. For our classification model, the output is a class membership, classified by a majority vote of its k closed neighbors. For a regression model with k-NN, the output is the average or median value of its k nearest neighbors. We can also assign weights to the neighbors, where closer neighbors may have a higher weight than 'distant' neighbors. Performance of the k-NN classifier is affected when the class distribution is skewed, as in our dataset. The instances of 'severe' accidents is heavily dominated by the occurrences of 'slight' accidents. As such, the data is skewed and the more frequent class will tend to dominate the out of a

[10] As opposed to Naive Bayes and Logistic Regression, we make no assumption on the parametric distribution of the errors for the K-nearest neighbors algorithm.

new data point. This is because they tend to be more common among the k nearest neighbors due to their large numbers. To overcome this problem, we resample the data similar to the case of the Gaussian Naive Bayes. Prior to training the classifier, we resample the data with replacement so that both classes are now represented uniformly in the dataset. Since distances are measured in Euclidean distance between pairs of samples, this will be influenced by the measurement unit. The results will be dominated by attributes with a relatively large range. In our example, the age attribute will influence the results of the classifier. To avoid this, we scale the data, so that each attribute now is Gaussian with mean zero and unit variance.

From Table 2, the precision for y=0 is 0.86 and 0.19 for y=1. While doing well in the 'slight' category, it is of all the data point it classifies as 'severe', it is only accurate 19% of the cases[11]. The recall column shows the classifier labels 69% of y=0 and 39% of y=1 labels correctly. The overall accuracy is 0.64 while the average F-score[12] is 0.68.

k-NN	**Classification Report**			
Target	**Precision**	**Recall**	**F1-score**	**Support**
y=0	0.86	0.69	0.76	42409
y=1	0.19	0.39	0.26	8078
Avg/Total	0.75	0.64	0.68	50487

Table 2

4.4 Logistic Regression

Logistic Regression is a supervised learning method for binary classification. The output of the logistic regression model is interpreted as the log-odds ratio of belonging to a particular class. In other words, it estimates the probability of class membership over a categorical class. In this case too, with unbalanced data, the classifier may identify most of the 'slight' accidents, and very few of the 'severe' accidents, giving a very high accuracy score. To weigh both classes equally, we incorporate the argument classweight='balanced', in the classifier. Similar to k-NN and naive bayes, logistic regression can also overfit the data, leading to high variance. To detect this, we perform

[11] It is correct in 3174 out of 16,435 it classifies as y=1.
[12] We set beta=0.5, so the F-score is the geometric mean of the precision and recall scores.

k-fold cross validation using GridSearch and F-score as our score to test the performance on the test data. To control overfitting, we need to add a regularization parameter in the classifier. We do an exhaustive search over a range of values for the inverse of the regularization parameter C from 0.001 to 1000[13].

From Table 3, precision for y=1 is 0.88 and 0.22 for y=1. The recall for the 'severe' category is higher at 0.54, compared to the other classifiers. This is good, as the model is classifying more than 50% of the 'severe' category correctly. It is not doing as well in the 'slight' or y=0 class, correctly classifying 63% of all 'slight' accidents, compared to more than 70% for the four other models. Overall the classifier is not doing as well since the accuracy is lower at 0.61.

Logistic Regression	**Classification Report**			
Target	**Precision**	**Recall**	**F1-score**	**Support**
y=0	0.88	0.63	0.73	42409
y=1	0.22	0.54	0.31	8078
Avg/Total	0.77	0.61	0.66	50487

Table 3

4.5 Decision Tree

Decision tree is a supervised learning technique which can be used in both classification and regression problems. In a classification problem, final leaves represent class labels. Leaves are assigned the class label by majority. In a regression tree, the leaves are assigned the mean or median of the target value of all samples in the end leaf.

The Decision tree algorithm will have similar problems like the logistic regression, with unbalanced data. So again, to weigh both classes equally, we use the classweight='balanced' argument in the classifier. Unlike k_NN, since best splits are performed using gini or entropy criteria, data does not need to be scaled. Decision trees

[13] Note that C in this case is the inverse of the regularization, so high value of C indicate weak regularization and vice-versa.

are popular in supervised learning, but also prone to overfitting. This can be overcome, by pruning the tree, by restricting the number of leaves or end nodes, or by specifying the minimum number of samples in each leaf. As before, we perform, a k-fold cross validation using GridSearch and F-score as our score to test the performance on the hold-out test data. We do an exhaustive search over a range of values for the following hyperparameters: minimum samples at each leaf'[14], maximum number of leaves, and finally using both the gini and the entropy criterion to find the optimal tree.

From Table 4, we can see that the recall column shows that the decision tree is correctly classifying 75% and 39% of 'slight' and 'severe' categories respectively. This gives an overall accuracy of 0.70, which is better than the other three classifiers.

Decision Tree	**Classification Report**			
Target	**Precision**	**Recall**	**F1-score**	**Support**
y=0	0.87	0.75	0.81	42409
y=1	0.23	0.39	0.29	8078
Avg/Total	0.77	0.70	0.72	50487

Table 4

4.6 Random Forest

Random Forests are an ensemble learning method for classification and regression. They are a way of averaging a multitude of deep decision trees, and in the process reducing variance. The idea is to compute k number of trees[15] using a random sample of the data for each tree. Furthermore, for each split in each tree, only a random sample of the feature space is available. This is because, we want each tree to be as uncorrelated as possible. So in addition to the randomness of the data, feature variables are also randomly selected at each split. The final result is then computed by averaging over all the trees.

[14] This restricts the node from splitting further if number of samples is less than the given threshold.
[15] Number of trees is a hyperparameter and needs to be predetermined.

Random forests are an ensemble of many decision trees. As such, while we do not need to scale the data, we do have to balance the classes. In addition to tuning the hyperparameter k=number of trees, we also tune for minimum number of samples in each leaf, maximum number of nodes and the criteria to find the best split: Gini impurity or entropy for information gain. To find the optimal forest, we do an exhaustive search over several hyperparameter values using GridSearch and F-score as the scoring function.

Table 5 shows the Classification report of the Random Forest. The recall column shows that it is correctly classifying 72% and about 45% of the 'slight' and 'severe' cases correctly. This gives an accuracy score of 0.67, which is surprisingly lower than the decision tree.

Random Forest	**Classification Report**			
Target	**Precision**	**Recall**	**F1-score**	**Support**
y=0	0.87	0.72	0.79	42409
y=1	0.24	0.45	0.31	8078
Avg/Total	0.77	0.68	0.71	50487

Table 5

We can compare the different models by their accuracy score, F-score and the ROC-AUC score, listed in Table 6 below. Using Grid Search with k-fold cross-validation has improved the accuracy for both k-NN and Decision Tree classifiers, from 0.59 to 0.64 and from 0.64 to 0.70 respectively. The accuracy for the unregularized and regularized models are almost same, 0.6127 for the unregularized, and 0.6129 for the optimized regularized model. Grid search has reduced accuracy for the Random Forest from 0.70 to 0.68. This is possible because, we are using the F-score as our scoring metric in the Grid Search.

Classifier	Unoptimized			Optimized		
	Accuracy	F-score	ROC score	Accuracy	F-score	ROC score
Naive Bayes	-	-	-	0.66		0.58
k-NN	0.59	0.21	0.52	0.64	0.22	0.54
Logistic Regression	0.61	0.25	0.58	0.61	0.25	0.58
Decision Tree	0.64	0.22	0.56	0.70	0.25	0.57
Random Forest	0.70	0.21	0.54	0.68	0.26	0.59

Table 6

4.7 Feature Importance

We next would like to understand which features are more predictive. Table 7 ranks the top 7 features by the Gini Importance. The most predictive feature is feature is age of the driver, with a Gini Importance of 0.42. The second most predictive feature is time, or more precisely, hour of accident with a Gini Importance of 0.23. The third, is speed limit at 60, but the coefficient is low at 0.065. Weather, road conditions, Speed limit 30 and single lane roads also make the top 7 important features but with a low Gini Coefficient.

Feature	Gini Importance
Age	0.4225
Hour	0.2335
Speed Limit 60	0.0656
Dry Road	0.0503
Good Weather	0.0361
Speed Limit 30	0.0268
Single Road	0.0258

Table 7

5. Policy Recommendation

From our analysis above, age of the driver and the time or hour of the day are important predictors of severe accidents. In fact, age is the most important predictor, and so, if insurance premiums are not staggered, then clearly low-accident prone age-groups will be subsidizing high-accident prone age-groups. Since most accidents occur at 5pm, authorities should look into incentives for work from home as well as different work timings across corporations, so traffic is more uniformly distributed across different times of the day. Putting up billboard signs, requesting drivers to be alert and cautious as gentle reminders, are will also help reinforce improved and safer driving practices, thereby reducing the incidence of accidents.

Certain road types are also more susceptible to road accidents. In particular, T-sections and cross-roads need to be given special attention by the authorities. Placing additional stop signs or traffic lights may help alleviate the problem. Further research needs to be done on speed limits, especially the 30 and 60 limits. One possibility is to lower speed limits from 30 to maybe 25, and then analyze if this significantly reduces the incidents of accidents. Weather and light conditions play little on no role. This seems counter-intuitive, but one possibility is that, bad weather and poor lighting may lead to lighter traffic.

6. Future Research

As insightful as this study has been, there are several things that can be looked into for future research. As is always true in any work in Data Science, we can improve our model predictions in two ways: one is the availability of better data. For example, the variable, purpose of journey, has 66% missing observations, and therefore could not be incorporated in our analysis. The mode for the total number of accidents is at 5 pm, which is rush hour back from work traffic . It is very likely then, this variable would have been a predictive feature in our analysis. Information about traffic controls, whether its controlled by traffic light, stop sign or an authorized personnel or its uncontrolled has 22% missing data. Again, this is a key feature for predicting the likelihood of an accident. We also have about 12% missing observations for age, another important feature.

In addition to more informative or cleaner data, better modeling techniques can also help in building models with superior predictive performance. While this study delves into a few supervised learning methodologies, there are several more like neural networks, support vector machines as well as other ensemble models like bagging, AdaBoost or stacking ensemble models. It will be useful to see if model performance can be improved by enhanced data as well as advanced supervised modeling methodologies.

It may also be insightful to find the factors that cause all accidents. For this analysis, we would need all traffic information within some time window, for different road conditions and at different times of the day. Furthermore, what are the costs associated with auto incidents? As noted, there are monetary costs and non-monetary costs. It would be interesting to be able to quantify these costs to improve our understanding of the long-term effects of reducing accidents on the overall welfare. This analysis has been done on data from the year 2016. The government of UK website has data on previous years, so we could also do time-series analysis on some of the key features, to see if trends on the target variable as well as the exogenous variables have been changing over time.