Sandhya Srinivasa (48018117)
Emily Arcidiacono (47608461)
Chase Vaught (47950218)
Daniel Laureano (47485260)
EMIS 3309
5/10/2021

Analysis of Road Accidents in UK From 1979-2015

Executive Summary:

To begin the project, we analyzed a dataset that contains information about car accidents in the United Kingdom. Furthermore, the dataset's response feature is accident severity which is a categorical value as there are three possible outcomes: 1(fatal), 2 (serious), and 3 (slightly severe). We use different statistical methods to analyze the relationship between certain features and accident severity, in order to understand the significance of these features. Using our findings from the statistical learning methods, we were able to decipher the significance of each chosen feature.

For our statistical models, we chose K nearest neighbor (KNN), LDA, and Random Forest. The first statistical learning method we used was KNN. When we set K to a value of 3, we got our misclassification error to equal 19.86. For the second technique, we applied the classification method of LDA. In our results, we found that the misclassification error is 15.15%. The third method we used was random forest. The number of variables chosen to randomly sample as candidates at each split is four, which we received by square rooting the number of features. The misclassification error rate for Random Forest was 16.31%. KNN was removed as a candidate because it had the largest misclassification error rate. This was most likely due to the large size of the dataset. Although LDA had the lowest misclassification error, we chose random forest due to its capability to handle large data sets and to identify the relative significance of variables.

In conclusion, our findings are listed below regarding how different variables affect the servility of the accident. When the age of the car is used as a feature, the prediction performance of accident severity is stronger. In other words, there is a strong relationship between the severity of the accidents and the age of the car. After analyzing the gender of the driver, we found that there is not a strong relationship between that and the severity of the accident. Certain weather conditions might increase the rate of accidents, but they do not have a large impact on the severity of the accident. We found that point of impact has a strong correlation on accident severity. Lastly, we noticed that right or left hand driving does not affect accident severity, as it was determined as the least important feature.

Problem Description:

The audience for our dataset are individuals who are choosing what type of vehicle to buy and are considering the safety of the vehicle. From the dataset, we are trying to understand how the type of car and other features involved in the accident affects the severity of the

accident. By letting our response column be the severity of the accident, we will be able to predict the severity of the accident depending on the age of the car, car type, and other factors. Some of the more specific questions we will be answering are listed below.
How does the age of the car affect the severity of the accident?
Is there a relationship between the sex of the driver and the severity of the accident?
How does weather conditions affect accident severity?
Does point of impact affect accident severity?
How does right hand/left hand driving affect accident severity?

Data Cleaning:

We analyzed a dataset from Kaggle that contains information about car accidents in the United Kingdom between the years of 1979 and 2015. Furthermore, we wanted to understand how certain features involved in the accident affects the severity of the accident. The main reason for choosing this dataset was the number of observations the set provided. The dimensions of the original dataset were made up of 70 columns and 285,331 rows. Having this many entries will vastly improve the accuracy of our models. However, we decided to look at only 15 of the features from the set as many of the other features contained missing values. First, we eliminated about 40 features based on if the feature contained a large amount of missing values. Next, we chose the final 15 features depending on whether we thought the feature would have a significant effect on the response column, accident severity, as there were other columns that were more about the casualties and the place of the accident. For example, some of the features we got rid of were the journey_purpose, day_of_week, and the longitude and latitude of the accident. Our original data set contained two types of NA values. There were 10 features that contained NA values. We did not include these features in our dataset because they contained many missing values. The value "-1" is used for NULL or out of range values for the rest of the features. We wrote R code to eliminate all the entries that contained "-1" values. Through data cleaning we eliminated 102,373 entries or 35.8% of our original data that had to be removed because of missing or out of range values. The resulting dataset is 15 columns with 182,958 rows and no missing values. The dataset mostly contains categorical values which did not have any measurement error. However for our numerical features represented by age, there is measurement error since the ages are not exact. The ages were represented with whole numbers meaning they did not account for how old the person was in terms of months and days after their birthday.

Our dataset contained 13 categorical features and two numerical features. When analyzing for categorical outliers there were 3 features that contained outliers in their data. The feature was_vehicle_left_hand_drive contained 1,469 outliers that represented "no". The next feature to have outliers was junction_location. It contained 923 outliers that belonged to category 7 that represents "Entering from slip road". The feature road_type contained 1,477 outliers that belonged to category 9 which represents "unknown". The numerical features in our dataset are age_of_driver and age_of_vehicle. The outliers of these numerical features were found by finding the ages that exceeded (Q3 + 1.5*IQR). For the age_of_driver there were 704 outliers that exceeded that calculated age of 87. For the age_of_vehichle there were 4858 outliers that exceeded the age of 17 years.

Data Exploration:

When examining our data, we found some interesting relationships between features. To begin with, we analyzed the age of drivers involved in accidents alongside the severity level, as seen in *figure 1*. Severity is ranked from 1, 2 or 3. 1 being fatal, 2 being serious, and 3 being slightly severe. No matter the severity of the accident, all three levels had a median around 40 years of age. Fatal accidents had a slight increase in its median age when compared to serious accidents. The interquartile range for fatal accidents was the largest, with slightly severe having the smallest. Fatal accidents had the oldest minimum age, while serious accidents had the youngest minimum age. Fatal accidents also had the oldest maximum age while slightly severe had the youngest maximum age. Lastly, it is interesting to point out that only serious and slightly severe had visible outliers in the data.

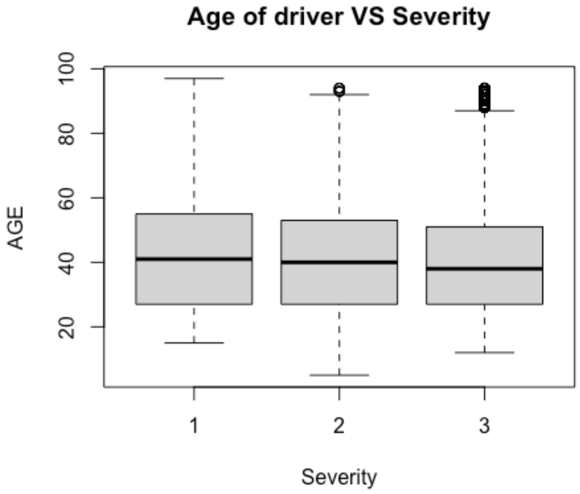


Figure 1

The next two features we analyzed were the age of vehicles involved in accidents and the severity of those accidents. When looking at *figure 2*, you can notice how all three severity levels (fatal, serious, and slightly severe) all have the same median value around 8 years. In other words, the most common number of accidents occured after cars aged around eight. All three severities also had the same interquartile range. All three severities had the same minimum age, around 1 year, and slightly severe had the lowest maximum age at 20 years. Fatal accidents had the lowest number of outliers while serious accidents had a very wide spread of outliers ranging from ~23-100+ years old. It is interesting to point out that the majority of vehicle accidents occurred in the younger years of the car's life. This contradicts the common misconception that older cars are less safe.

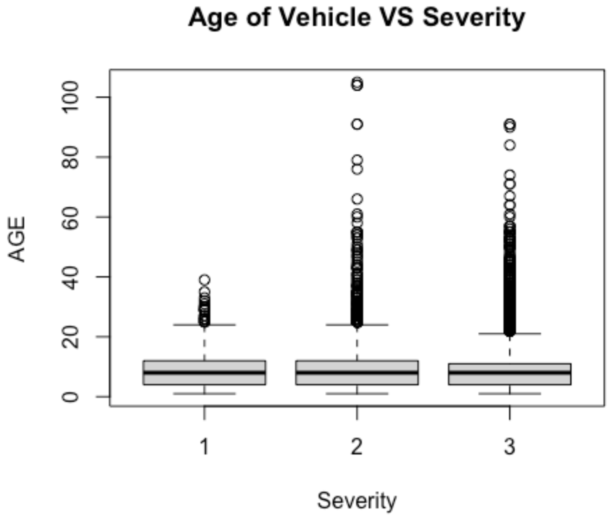


Figure 2

Lastly, we analyzed the relationship between the speed limits on the road at the time of the accidents along with accident severity. Similar to the United States, the United Kingdom uses miles per hour as the unit for their speed limits. As seen in *figure 3*, fatal accidents had a median speed limit of 60 MPH, while serious and slightly severe accidents had a median of 30 MPH. The first quarter took up the entire IQR for fatal and the third quarter took up the entire IQR for both serious and slightly severe. Fatal, serious, and slightly severe accidents had the same maximum values at 70 MPH, most likely the top speed limit in the area. The accident with the minimum speed was a slightly severe accident.

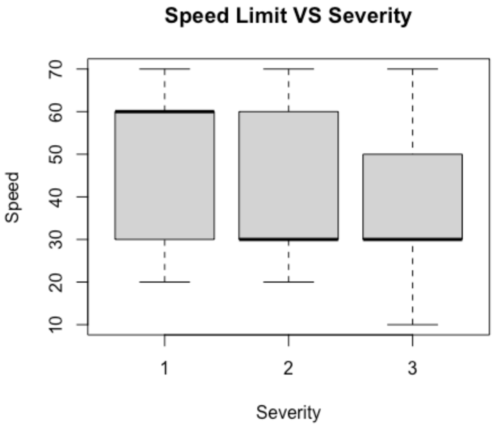
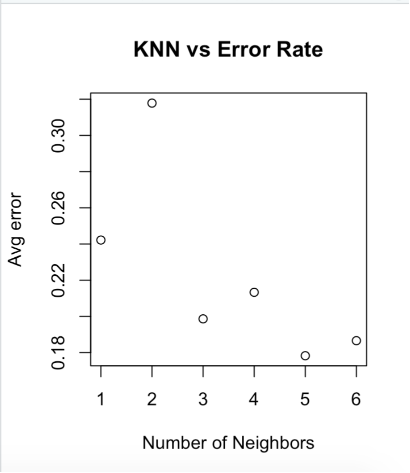


Figure 3

When examining the statistics for individual features, we found a few interesting details. To begin with, the average vehicle type involved in the accidents was a car. The other types included motorcycles, vans, buses, taxis etc. The severity of accidents had an average between serious and slightly severe. The most common weather condition was in between fine and moderate rain. The average condition of the roads was in between dry and damp. And lastly, we found it interesting that the point of impact was hitting the back of the vehicle. Analyzing these features contributed to our findings and revealed key trends in the data. We used these findings to help us decide on which features to examine more closely.

Statistical Learning Methods:

The first statistical learning method we used was K nearest neighbor, or KNN. This method was easy to implement and did not take long to calculate. Fortunately, KNN is a great method to use when dealing with missing values, which in our case we had. Lastly, since our data was mainly categorical, we knew that KNN would be a good method to use to examine the data. An issue we had using this method was that KNN does not reveal a prediction for important features which was one of the questions we tried to answer. This issue hindered us from analyzing the data thoroughly in terms of features. Lastly, when we set K to a value of 3, we got our misclassification error to equal 19.86%. After we found this value, we tried different values for K and found that the error rate was lowest when K had a value of 5, as seen on the graph displayed here. When K equals 3, the misclassification error is equal to 17.83%.

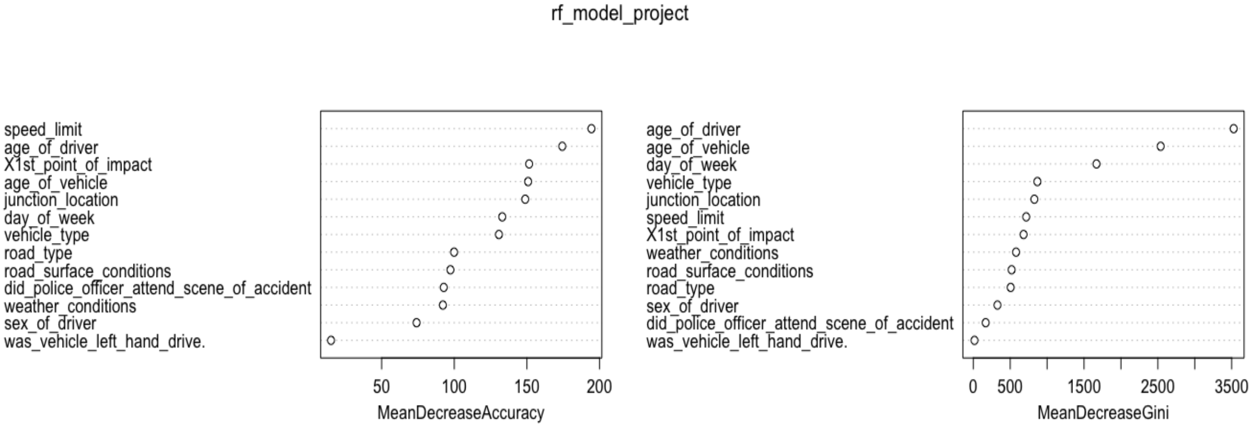


For the second technique, we applied the classification method LDA. We used this method because we assumed that all classes have similar variance, also our dataset's response feature is categorical and has multiple classes. However, we also considered the cons of LDA, such as that it assumes that every class has the same variance, and it may not perform well if this assumption is not true. The dataset's response feature was the accident severity, which is a categorical value because with three possible outcomes for accident severity: 1, 2, 3. We predicted these values based on the type of car as well as the number of previous accidents the car has had. In our results, we found that the misclassification error is 15.15%, however this is deceiving because LDA is predicting most of the samples from class 3 of accident severity. This might be due to skewness in the dataset (for example, we have a lot of class 3 and only a few values from classes 1 and 2). This being the case, we would have to re-process the data in order to balance the data set.

The third and last technique we applied was Random Forest. While applying Random Forest, the biggest issue we had was how long it took to process and the fact that sometimes it never processed. In order to mitigate this issue we used a smaller amount of the data, around 100000 observations, to run Random Forest on R. However, Random Forest is able to handle missing values and large datasets as well as show the relative importance for the features. The number of variables chosen to be randomly sampled as candidates at each split is 4, which was derived from square rooting the number of features. The misclassification error rate we found was 16.31%.

To summarize the error rates, KNN was 19.86%, LDA was 15.15% and Random Forest was 16.31%. First, we ruled out KNN due to the fact that it had the largest misclassification error rate. The reason for the large misclassification error rate could possibly be because our data set was so large. The entire training data is processed for every classification and it also assumes equal importance to all features when deciding the predicted values, and we realized there was an imbalance in our classification data. From LDA and Random Forest, LDA had a lower misclassification rate but we chose Random Forest because of how it's able to handle large data sets and to identify the relative significance of variables, which was an important part of our objective.

In Random Forest, Mean Decrease Accuracy is used to understand relative significance. The Mean Decrease Accuracy is a rough estimate of the loss in prediction performance when that particular feature or variable is omitted from the training set. The way mean decrease accuracy is calculated is from the ratio of correct classification to total records. The Mean Decrease Accuracy allows us to decide how significant/important a feature is in comparison to others. For example, speed limit plays a bigger role in accident severity than the road surface conditions. The Mean Decrease Gini is another form of deciding how significant or important a feature is in comparison to others, but we only used Mean Decrease Accuracy to decide in our analysis. The figure below shows the relative significance of each feature. On the graph, the more you move right, the more significant that feature is.



Conclusion:

We analyzed a dataset from Kaggle that contains information about car accidents in the United Kingdom between the years of 1979 and 2015. Through analyzing outliers and data cleaning we were able to generate a new dataset containing features we thought would have a significant impact on the response column, accident severity. The features with the most significant effect on accident severity was speed limit followed by the age of the driver. The least significant features that impacted accident severity was whether the driver was driving with their left hand and the sex of driver. This did not match our prediction of that 1st point of impact would have the largest effect while the day of the week would have little to no impact. This was found through the Random Forest Mean Decrease Accuracy. Mean Decrease Accuracy is used to understand relative significance of each feature. We applied the statistical models: KNN, LDA, Random Forest in order to model to find which model produced the lowest misclassification rate. LDA provided the lowest rate of 15.15%. The dataset contained many instances of class 3 (slightly injured) compared to the other classes. We recommend using the same amount of observations from each class of accident severity to train the data in hope for better predictions and lower misclassification rates. We also recommend that vehicle brands test their vehicles at different speed limits and age groups of drivers as they are some of the most significant features that impact the severity of an accident.

References:

Dataset: https://www.kaggle.com/akshay4/road-accidents-incidence
Understanding How to Find Importance in Random Forests:
https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/