CS 5304 Assignment 3

Tuan-Chun Chen tc674@cornell.edu
Yuekun Wang yw2222@cornell.edu

April 20, 2019

1. Problem Overview

In this assignment, we used PySpark and SQL in DataBricks to handle the Big Data.

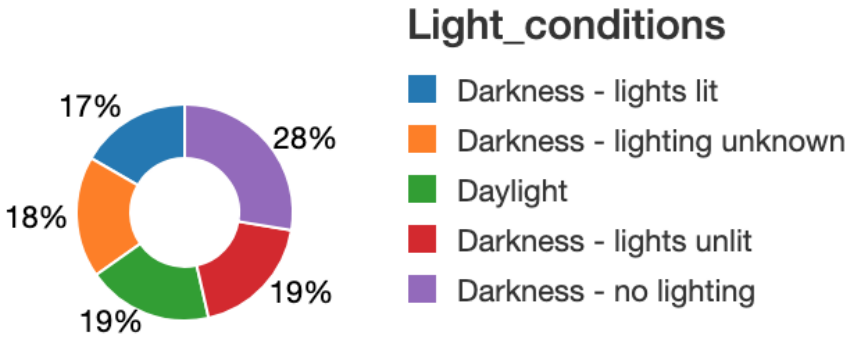
2. Data Description

Two CSV files contain accident information and vehicle information in the UK from 2005 to 2017. We want to analyze the data sets to predict the severeness of accidents.

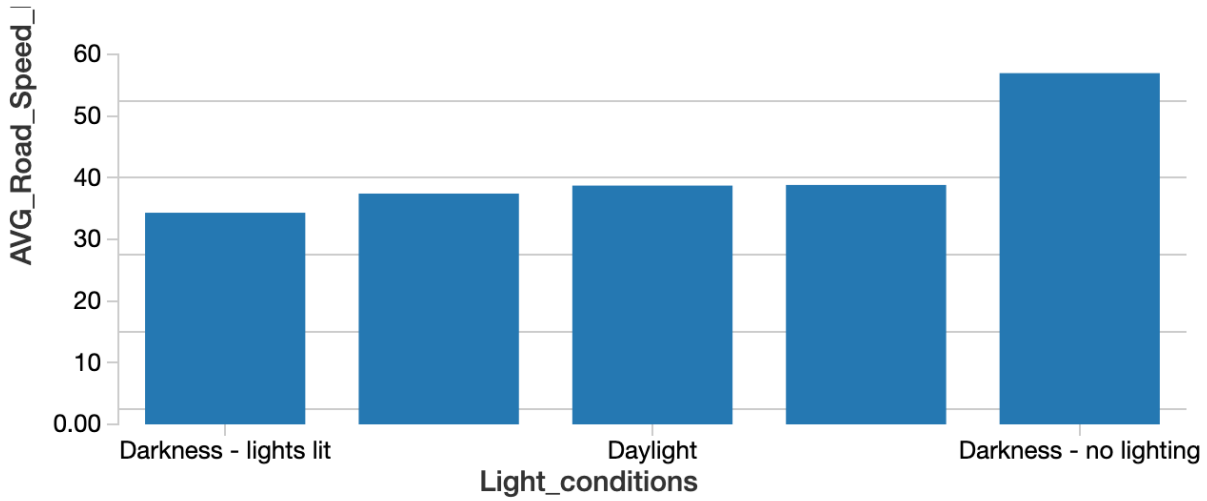
3. Data Analysis

3.1. Average Speed Limit Under Different Lighting Conditions

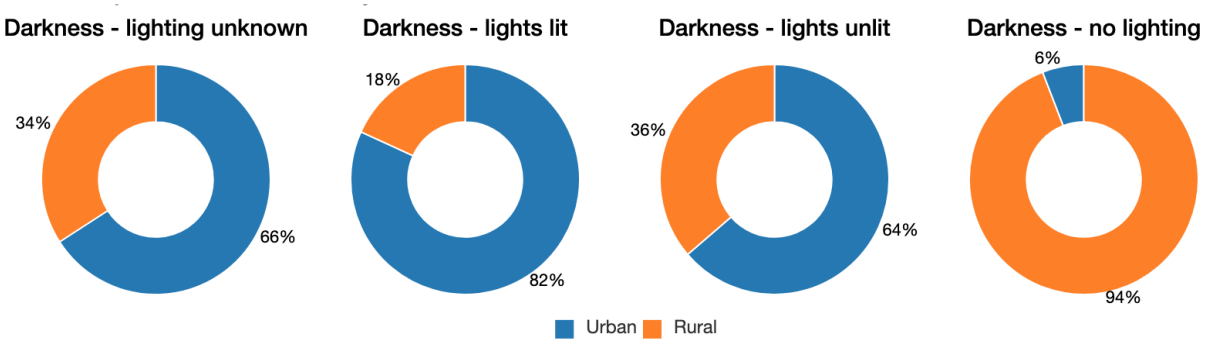
The pie chart shows a general breakdown of different light conditions. It is quite obvious that way more accidents happen during night time.



The following histogram shows the average speed limit at accident sites under different lighting conditions:



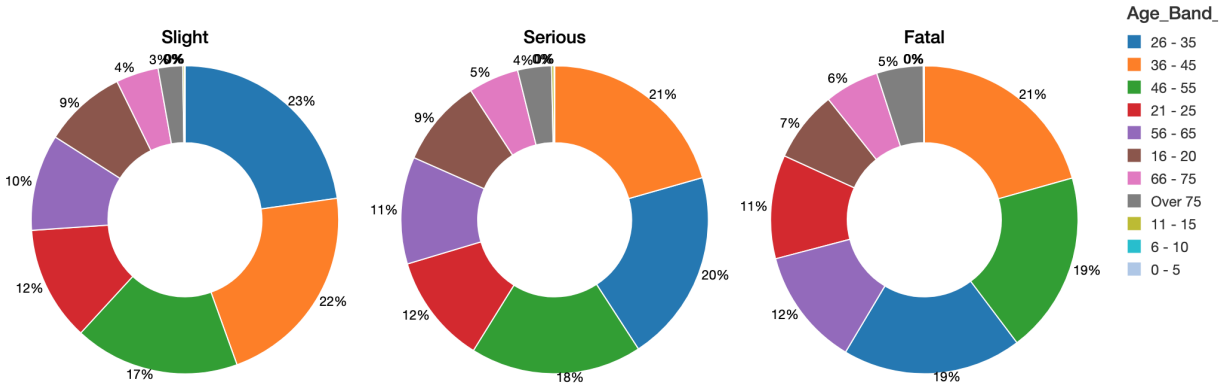
Since speed limit would not change according to the lighting conditions, in other words, speed limit should be the same despite of daylights or darkness, we should look at this problem under two main conditions. First of all, the average we get for the daylight should be the average of speed limit at all accident sites during daytime. Then, we want to take a look at the condition under darkness. We found a significant differences between the average speed limits between 'lights lit' and 'no lighting'. Based on common sense, we know that there should be more lights in the urban areas and fewer lights in rural areas and highways. It also make sense since speed limits in urban areas should be lower than those in rural areas and highways. To further confirm our hypothesis, we ran another SQL query to group by different light conditions in urban or rural areas.



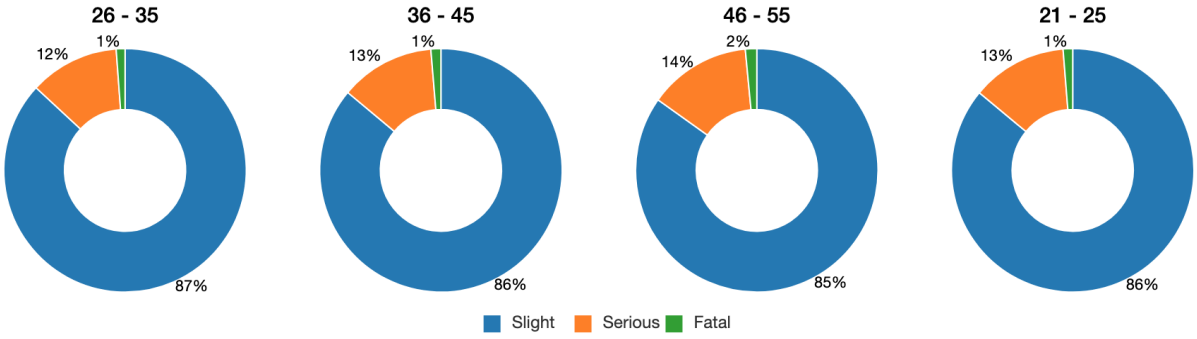
We observed that most 'light lit' are in urban areas while most 'no lighting' are in rural areas which justified our explanations above.

3.2. Accident Type vs Driver Age

The majority of accidents are composed with age groups of 26-35, 36-45, and 46-55 since most drivers falls into these age groups.



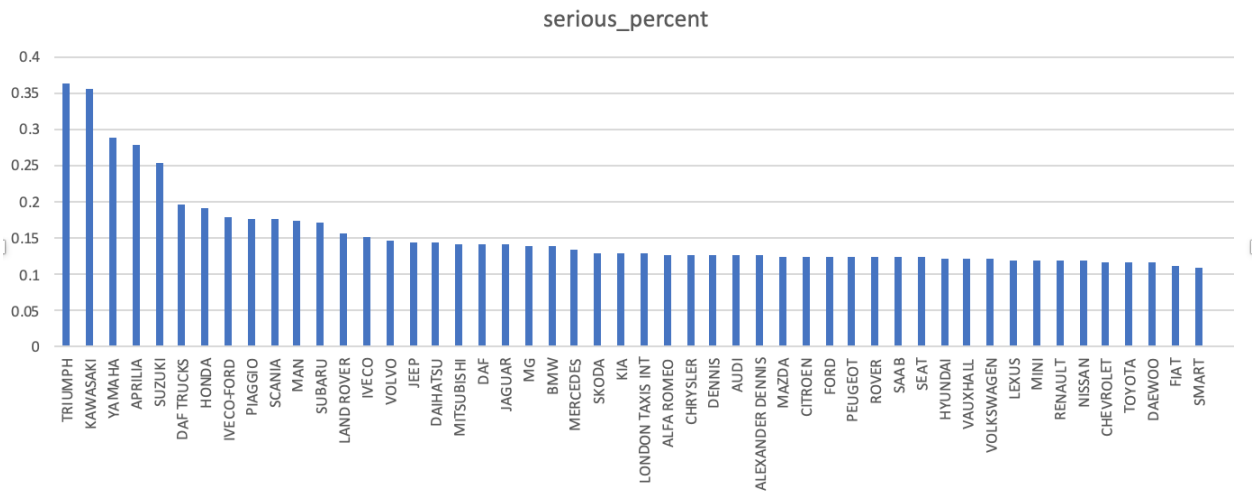
Among these groups, drivers in 46-55 have a slightly higher rate of involving in a serious/fatal accident.



For a detailed table of the number of accidents according to accident type and driver age, please refer to the code.

3.3. Serious accident ratios from top 50 most common manufactures

In this question, we wanted to find the serious accident ratio from the top 50 most common manufactures. To calculate serious accident ratio, we count the number of serious and fatal accidents and divided by the total number of accidents taken placed by one manufacture. We get the following graph ordered by how often the model of cars appears in accident data set.

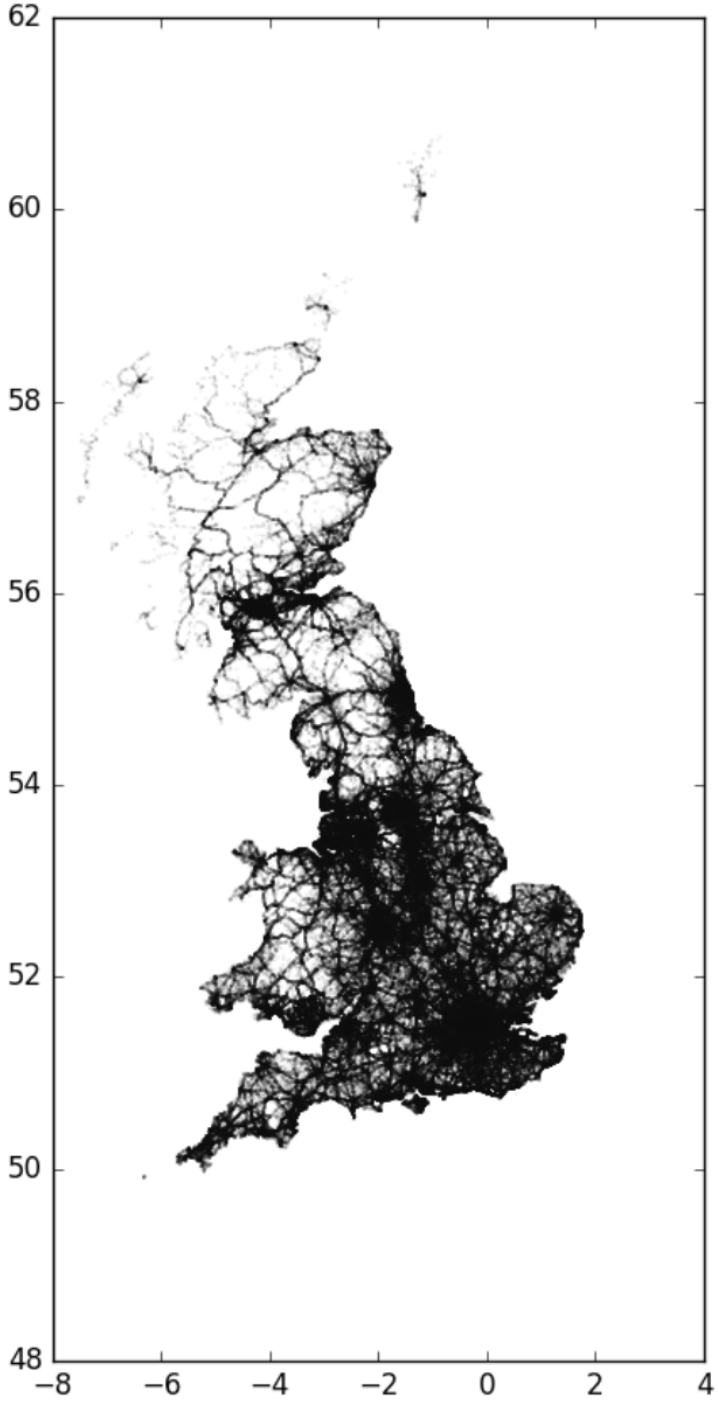


From the histogram, we observed that the highest accident ratios in the popular manufactures appear in Triumph(0.363), Kawasaki(0.356), Yamaha(0.288), Aprilia(0.278), and Suzuki(0.254), which means these models are more likely to cause serious accidents comparing to other models.

For a detailed table of serious accident ratios for all 50 manufactures, please refer to the code.

3.4. Heatmap

The below heat map is generated according to the geographic information of each accident.


62
60
58
56
54
52
50
48
−8
−6
−4
−2
0
2
4

4. Severeness Prediction

4.1. Feature Selection and Preprocessing

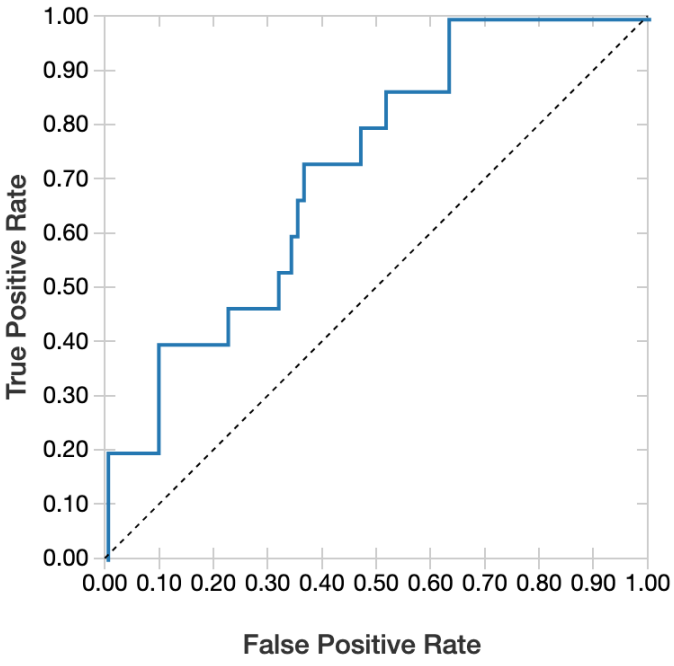
We seleceted 'Road_Surface_Conditions', 'Age_Band_of_Driver','Vehicle_Manoeuvre', 'Vehicle_Type', 'Sex_of_Driver' as our categorical features, and 'Number_of_Casualties', 'Number_of_Vehicles', 'Speed_limit', 'Age_of_Vehicle' as our numerical features since we proposed these features should be most relevant to predict the severeness of the accident.

We first used OneHotEncoderEstimator to encode categorical data and combined numerical data into the feature column. Then we generated label from accident severeness. We combined the serious and fatal accidents together and make the model binary.

After splitting the dataset into training and testing data, we observed the data imbalance of having way too many slight accidents that resulted in a prediction of all accidents being slight accidents. In order to deal with this problem, we randomly selected a number (184,812 in this case) of slight accidents to match with the number of serious/fatal accidents.

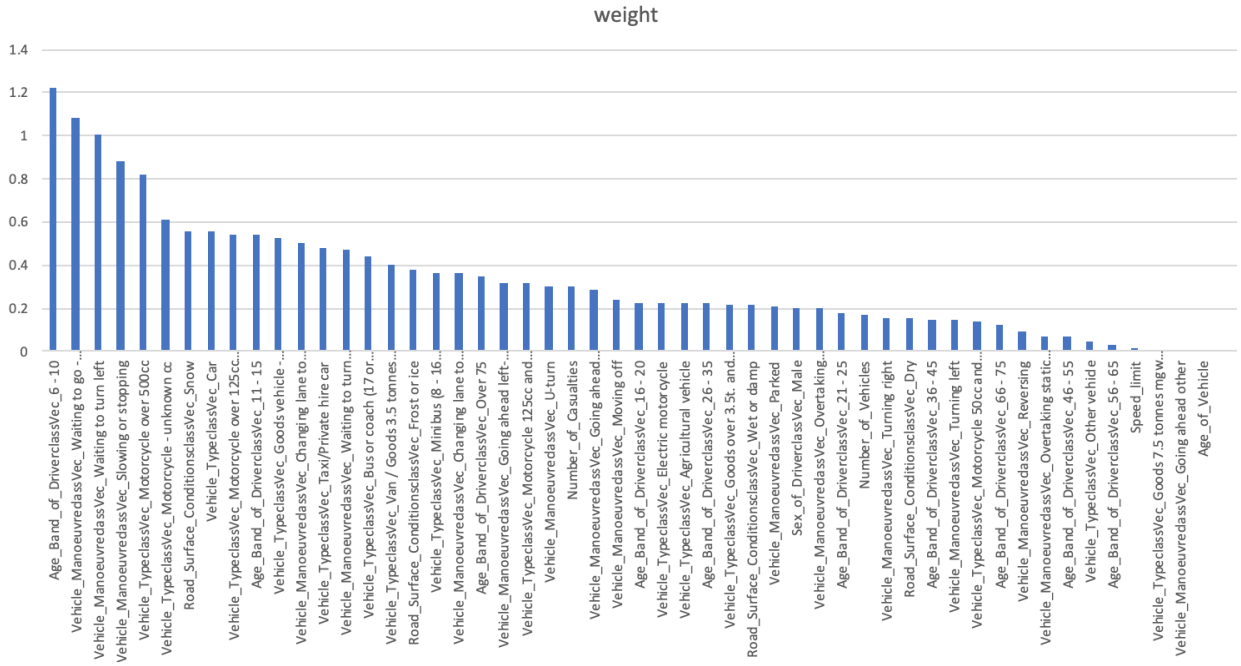
4.2. Logistic Regression

We get the ROC curve as following with AUC equals to 0.678:



If we take a look at the accuracy rate for each class, we found that the model is 58.44% accurate for serious/fatal accidents and 65.83% accurate for slight accidents.

The feature importance for the logistic regression model is shown below:



We can tell that young-aged drivers, vehicles waiting to go/turn left/slowing or stopping, Motorcycles are highly likely to involve in a serious/fatal accident. Moreover, the roads covered with snow are also likely to cause serious/fatal accidents.