Data Science in the Wild: A3

Yuemin Niu [yn276] Jingxuan Sun [js3422]

Problem Statement

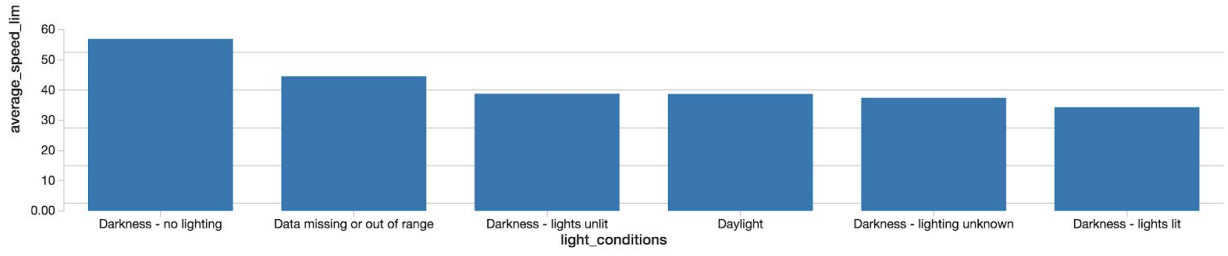
In this assignment, we performed large scale data analysis with Apache Spark. The dataset we are going to analyze is UK Road Safety dataset with one table containing accident information and one with vehicle information. We first used SQL to conduct some key queries. Then we built logistic regression and random forest models to predict the severeness. We figured out that properties of vehicles like age of the vehicle and vehicle type, properties of drivers like the age and gender and properties of sites like road conditions give rise to the occurrence of serious accidents.

The full code that is used for the analysis can be found at [https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2391717547056929/2617863170152724/7728845932782585/latest.html].

1. Data Analysis

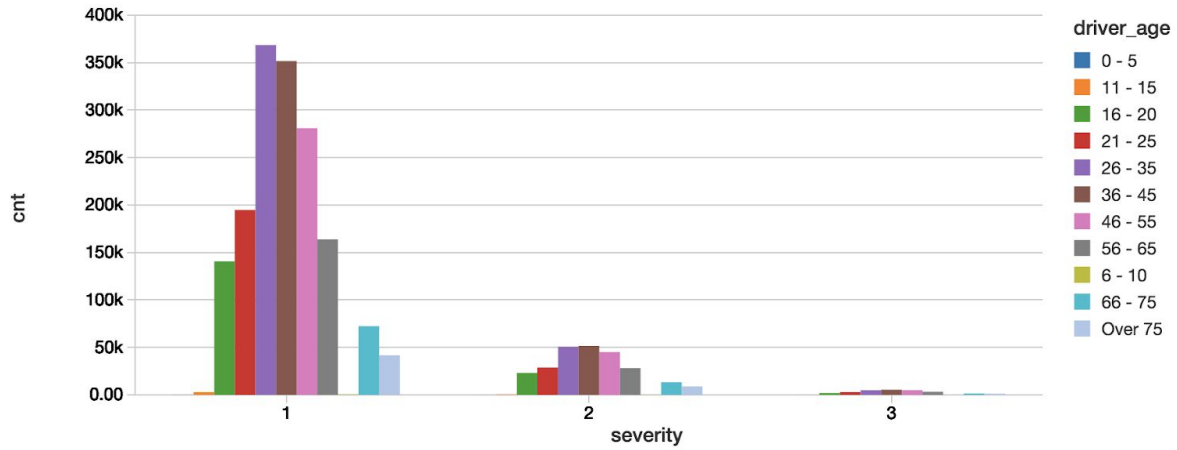
1.1 Speed Limits and Lighting Conditions

We selected average speed limits under different lighting conditions and figured out that sites in darkness and without lighting have the highest speed limit. While the sites with lights unlit or lit have relatively lower speed limits. We believed that speed limit has nothing to do with daylight or darkness and is not related to the road condition. The sites without any lighting usually will be some rural areas and there won't be many cars on the road, thus the speed limit is not strict. While roads with lighting, whether it's unlit or lit, tend to be in an urban area with more cars and complicated traffic conditions.



1.2 Accident Types and Drivers Age

Since the two features are in different tables, we first joined two tables then selected the number of observations grouped by different accident types and age bands. Most of the accidents are slight ones and it can be clearly observed that in each severity group, drivers from 26-55 have a higher tendency of accidents. We also assume that this age group made up most of the drivers.



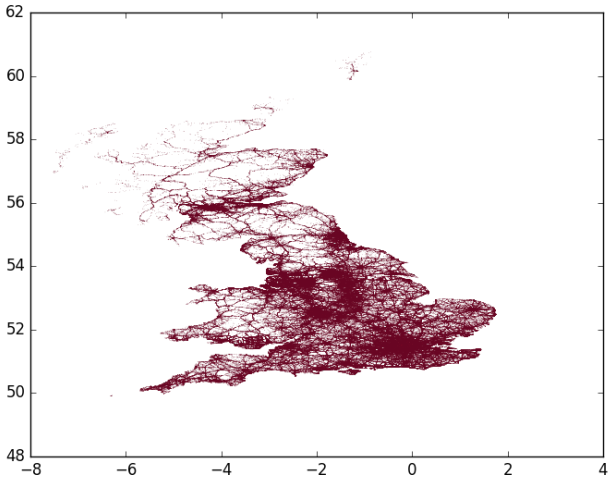
1.3 Serious Accident Ratio and Car Maker

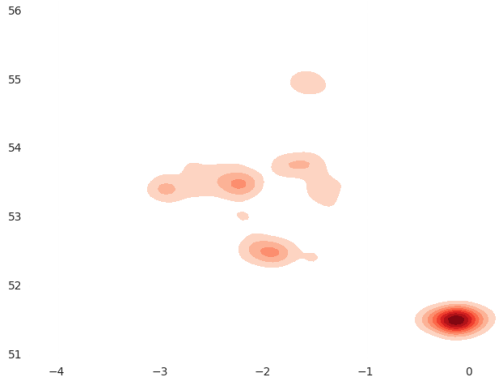
We computed serious accident ratio by dividing the number of serious accidents with total number. We discovered that the manufacturer with the highest serious ratio is Triumph and the one with least ratio is Smart. Here we list manufacturers with top serious accident ratio, and this table shows that there's some correlation between serious ratio and Japanese manufacturers. Details about the ratio and manufacturers and the full list are in codes.

make	serious_ratio_by_manufacturer
TRIUMPH	**0.327633050982899**
KAWASAKI	**0.318240251392658**
YAMAHA	**0.266207258155125**
APRILIA	**0.256673511293634**
SUZUKI	**0.229121241316363**
HONDA	**0.174745333264388**
PIAGGIO	**0.171674623233174**
SUBARU	**0.152138350081795**

1.4 2D Heat Map

We first drew a scatter plot with latitude, longitude and accident ratio. We discovered that the scatter plot showed a nice UK shape. The darker part indicates relatively higher ratio of accidents, which in this case are large cities like London, Nottingham and Birmingham.



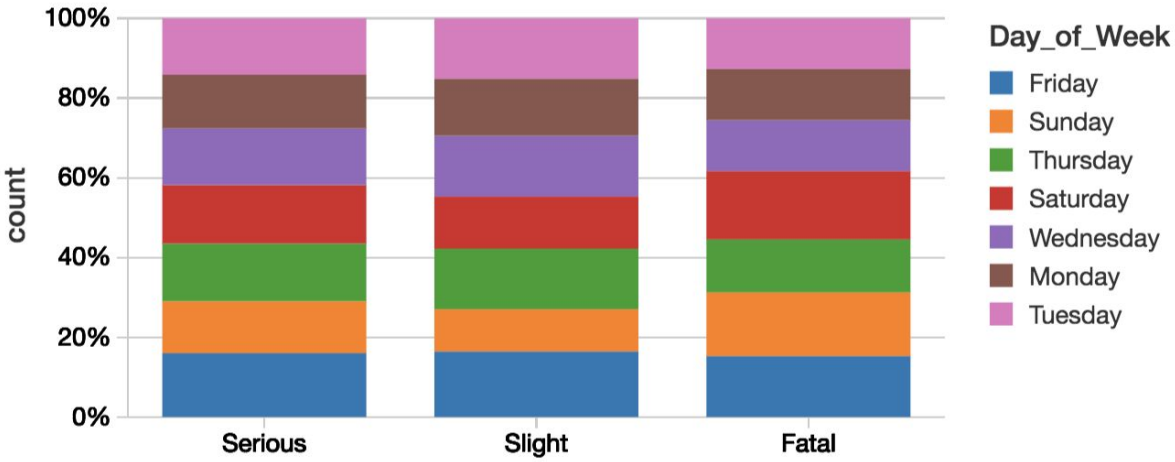


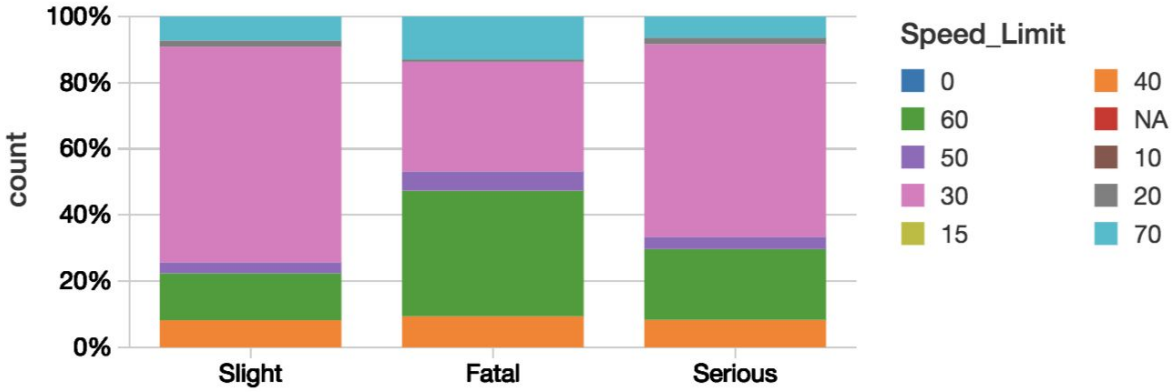
Then we used seaborn library to draw a heat map on the accident ratio. The result is similar to what we have in the scatter plot, that is larger cities like London are more likely to have higher accident ratios.

2. Severeness Prediction

2.1 Feature Selection

The basic idea of feature selection is to filter those features with relatively large variation among three accident types. For example, the total number of serious and fatal accidents increases during weekends. Also, higher speed limits are more likely to be associated with serious and fatal accidents. More details can be found in codes.





Based on this selection rule, we included 15 features in our model. Among them, 12 are categorical and 3 are numerical. The selected features are below:

Day of Week	Number of Casualties
Junction Control	Number of Vehicles
Light Conditions	Speed Limit
Urban or Rural Area	
Age Band of Driver	
Sex of Driver	
Vehicle Type	
Road Type	
Special Conditions at Site	
Weather Conditions	
Age of Vehicle	
Vehicle Manoeuvre	

2.2 The Model

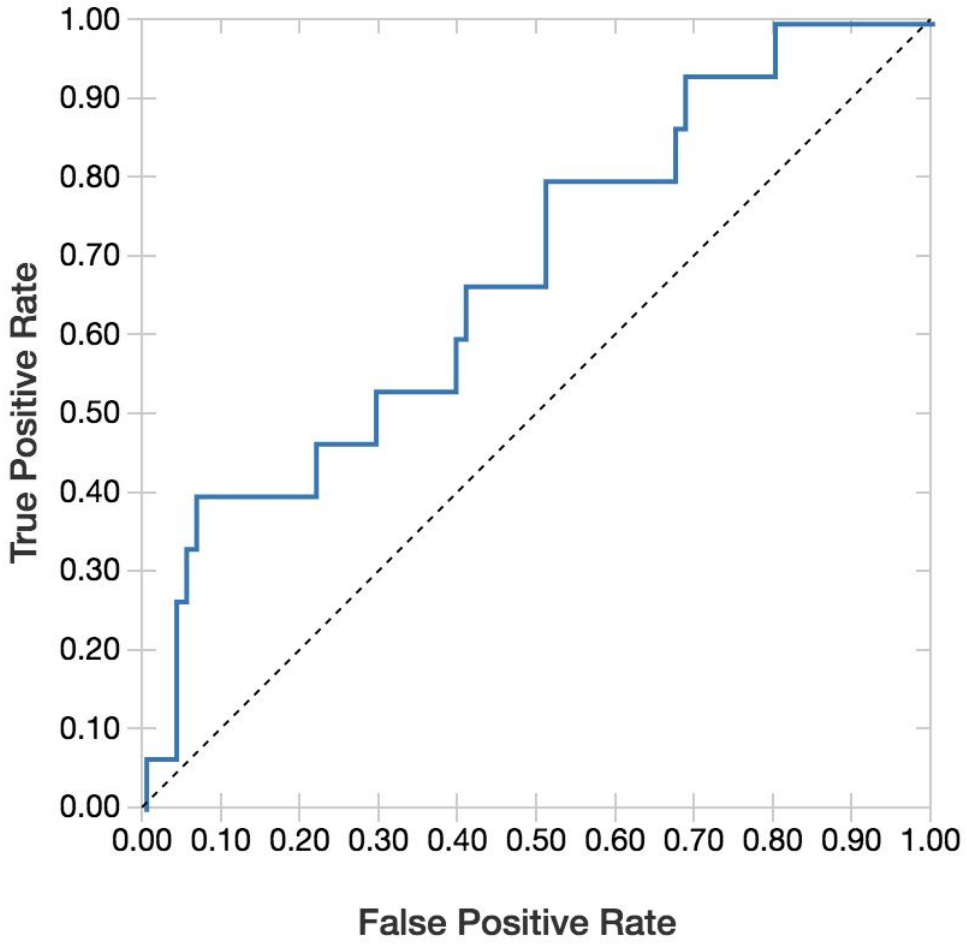
Since the target variable we are going to predict is severeness, which is categorical, we tend to select random forest and logistic regression models. We notice that the severeness has three

categories and the distribution is not balanced, thus we combine the serious and fatal classes and transformed the target variable into a binary one.

The model performance is summarized as following.

Model	**AUC**
Logistic Regression	0.688
Random Forest	0.649

The logistic regression gives us better performance. The ROC curve is below.



2.3 Feature Importance

We then extracted weighted feature coefficients as feature importance. The most important feature is Age of Vehicle, then follows the Gender of Driver, Road Type, Age of Driver and Vehicle Manoeuvre. The results make sense to us. For example, the older the vehicle is, the more likely it is involved in a serious accident. Also, more serious accidents happen on slip roads and dual carriageway. Besides, when the manoeuvre is changing lanes, the chance of a serious accident goes up. One interesting finding is that when weather conditions are fine, the likelihood of an accident is high. The possible reason might be that the safety consciousness of drivers goes down on good weather and results in more serious injury on the road. The full list of feature importances can be found in the code.