ACCIDENT SEVERITY ANALYSIS

UK 2018 Road Safety Data

Data Science Capstone Report

This report outlines the initial data analysis and data science process in aiming to create a predictive machine learning model for Accident Severity on UK roads.

George Galloway
IBM Data Science Capstone Coursera

Introduction

Safety is an incredibly important topic. From a young age we are taught to look both ways before crossing the street. Car manufacturers spend millions per year investigating, testing and improving the safety of their vehicles. Government organizations create safety standards and ensure quality. And as technology continues to improve, Data Scientists enter into the equation. Despite the importance of having cars that are built to deal with collisions, it is just as important to investigate ways to mitigate the occurrence of car crashes. With the abundance of readily available data on the internet, as well as the computer programming systems available, it is up to the imagination of data scientists to invent new insights and solutions to the road safety issue. For Governments, car manufacturers, road maintenance companies, emergency services and even drivers themselves, the availability of insightful analytical (predictive) models can enable smarter and safer decision making. Allowing drivers to be able to respond to sudden changes in situations and to prepare and alert emergency services for swift responses.

The main issue being approached in this Capstone Assignment will be the investigation into a predictive machine learning solution for car collision severity. The main goal is to create a system that can predict the severity of a car crash depending on potential factors such as location, weather and road type. This could lead to navigation systems being able to calculate not only the fastest route home, but the safest.

Data

For this project, I will be using Road Safety data for 2018 provided by the UK Department for Transport. This data is generated through the police STATS19 accident reporting form. This results in well structured data that contains not only the severity of the accident and the number of casualties, but also factors such as weather and road conditions. Furthermore, a latitude and longitude value is recorded for the data, allowing easy visual map integration. The dataset contains in total 122,635 samples and 32 features. Irrelevant features will of course be removed.

I have conducted an initial investigation and visualization which is available to view on my IBM Watson Notebook via the following link. This initial investigation made use of APIs such as Pixiedust for graphic visualization and Folium for location visualization.

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/e66e7764-c753-46c5-81b8-59f00ec6af50/view?access_token=befda76c3ae750ae13637f46e46619ee3ae7b4eb13621c6aeba71244bf92453b

On my GitHub it is also possible to find the data csv file as well as an excel file detailing the data and the meaning of the variables.

https://github.com/GSWGalloway/Coursera_Capstone

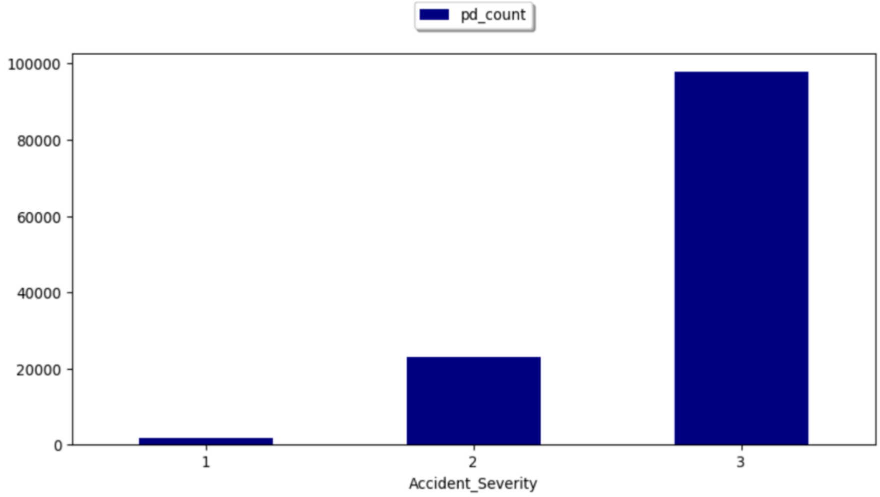
Due to the large volume of well-structured data, and the 32 factors included per data entry, I hope to discover potential correlations from which a predictive model can be generated. I will initially clean the data, remove any null values, and conduct a statistical Pearson correlation analysis. The analysis will of course center around the severity data of the accidents, which is recorded as follows:

1 – Fatal

2 – Serious

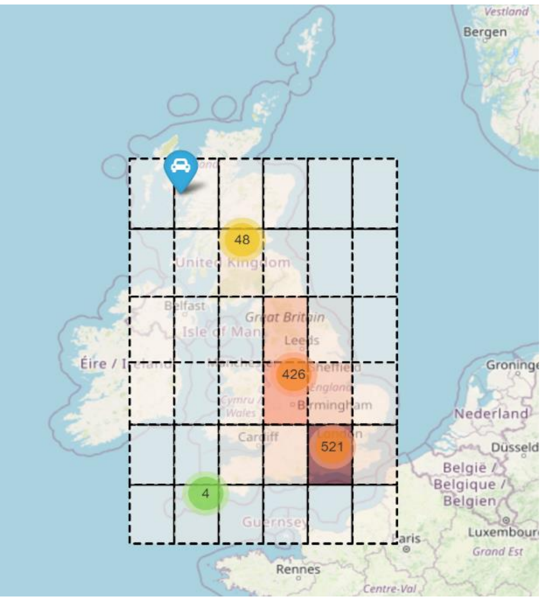
3 – Slight

Visualizing the severity data clearly shows that the vast majority of accidents fall under the category ‘slight’. As can be seen in the graph below.

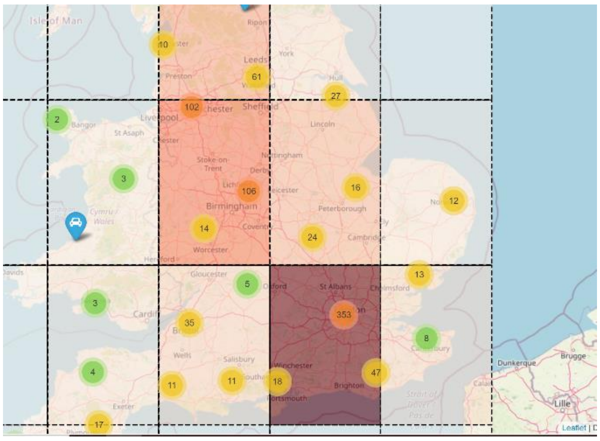


Methodology

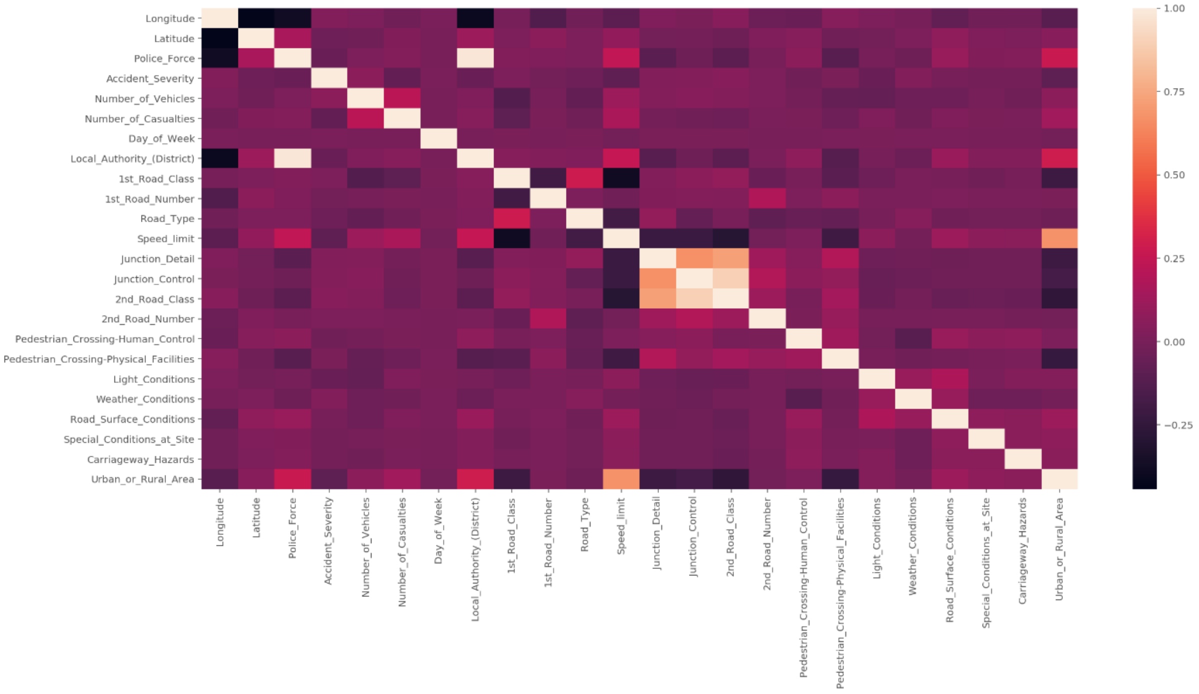
Initially, a thorough Explorative Data Analysis was conducted. In order to ensure this was not only useful for myself but also for others, I ensured to make this very graphic and include a number of clear visualizations of the data. I decided a useful tool, not only for myself but for any other involved party, would be a map visualization on which clearly areas of higher risk can be identified. In order to achieve this, I made use of the Folium API. To ensure IBM Watson could easily handle the processing, I created a random sample of 1000 data points. Using markers from the location coordinates in the data csv file I could not only create an interactive map, but also portray onto this a heatmap, identifying the worst areas for accidents. The resulting map clearly indicated which areas were higher risk and by clicking on a marker the exact details of the accident can be read.



This visualization clearly indicated that London is a high risk area, followed by the Birmingham and Leeds counties.



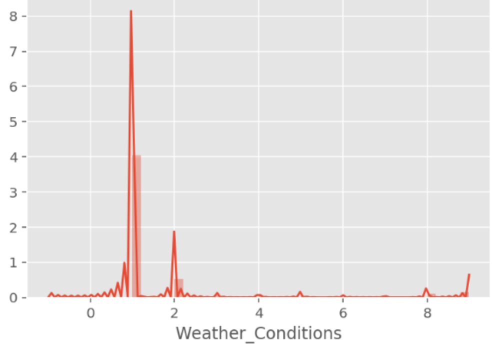
In order to be able to create a **reliable** predictive model, it is very useful to have positively correlating data. Sadly however, as the below correlation diagram reveals, the only noteworthy correlation is that between “Urban or Rural Area” and “Speed Limit”. Sadly this image doesn’t reveal a correlation between weather conditions and accident severity, as I had initially been hoping.

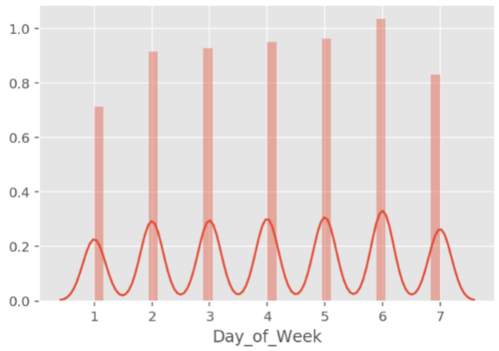


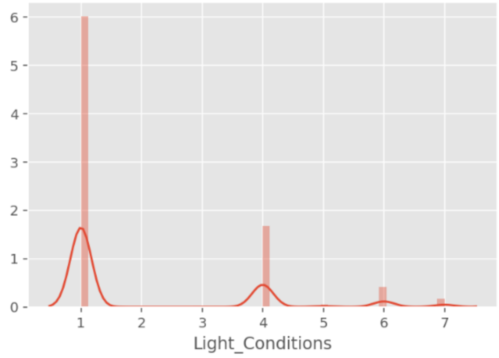
However, I continued with 4 data features which I hoped would aid in generating a reliable predictive model. In doing so I decided to focus on creating a predictive model based around weather conditions, rather than location based. I believe my Folium map visualization has already generated a valuable system for identifying locations of high risk, and I now wish to continue focusing on the conditions experienced by the driver. This could potentially result in a model not only useful for the UK, but also easily implementable elsewhere.

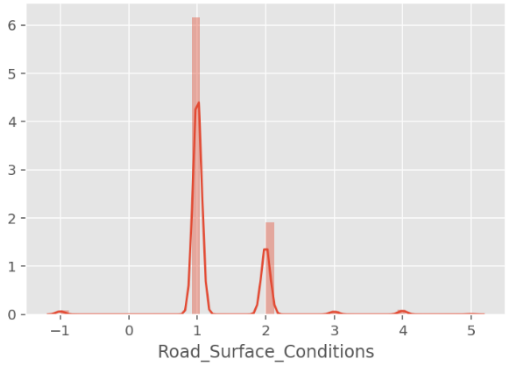
Results

As more data analysis was conducted, it became clear that with the UK data it would not be possible to generate a reliable model. I have been able to identify areas of increased risk of traffic accidents, however so far it does not seem feasible to form a predictive model due to the lack of correlations. There may be a higher risk of a crash in bad weather, but due to the larger number of vehicles on the road in good weather, it is impossible to define features that play a role in accident risk.









The above graphs indicate from top left to bottom right;

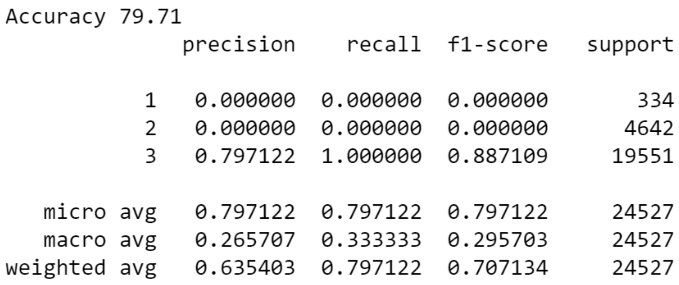
The most accidents take place in good weather,

In 2018 the most accidents took place on Thursday,

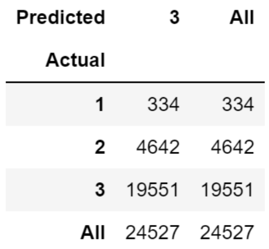
The majority of accidents took place in favorable light conditions, followed by mist/fog,

The majority of accidents took place on dry road surface conditions.

I continued with these four parameters in mind to build a predictive model. I initially split the data into training data and test data and then conducted a number of statistical tests. I made use of 3 predictive models in particular; Random Forest, Logistic Regression and Decision Tree. Ultimately after tuning the Decision Tree Model I received the following results:



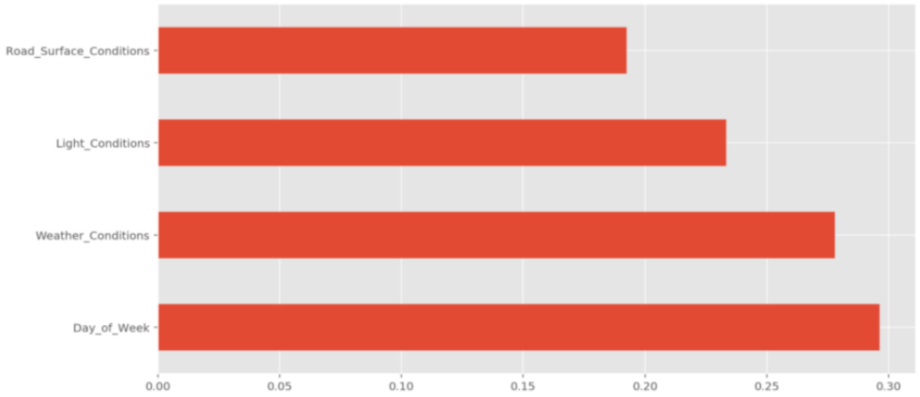
```
Accuracy 79.71
              precision    recall  f1-score   support

           1   0.000000  0.000000  0.000000       334
           2   0.000000  0.000000  0.000000      4642
           3   0.797122  1.000000  0.887109     19551

   micro avg   0.797122  0.797122  0.797122     24527
   macro avg   0.265707  0.333333  0.295703     24527
weighted avg   0.635403  0.797122  0.707134     24527
```



Predicted	3	All
Actual		
1	334	334
2	4642	4642
3	19551	19551
All	24527	24527

As can be seen, an accuracy of 79.71% was achieved.

I also ranked the features by importance in this model, resulting in the following graph.



Discussion

With the data I used, it was not possible to generate a highly reliable model. There was a clear lack of correlations from the beginning and I wouldn't recommend this model for real world use. I believe if I was to expand the data set to several years and match it to data features on the vehicle and driver I could generate a better model. There is also the issue of how to deal with the fact that there are more cars on the road in favorable weather conditions, which of course then results in a higher risk of an accident but shadows the effect of the less appealing and potentially more dangerous conditions.

Conclusion

The Folium visualization of the UK accidents and 'hot-spots' is a great tool and very easy to understand and use. The predictive model created in this project however is not of the highest quality and I would not recommend to be used by law enforcement, navigation systems or drivers in general. This system needs to be expanded with more data and more data features especially, in order to hopefully discover clearer correlations and generate more accurate models.

For someone with absolutely no coding experience but a keen interest in technology, this has been a very challenging and educational experience for me. I initially feared code and programming but now I understand the power of it and the incredible range of potential. I will have to continue practicing to fine tune certain skills but I am quite pleased with the result of this project.