Predicting Accident Severity

1 Introduction

Around the world, early warning systems are in place that work to predict and notify the population of an impending disaster, ahead of time. The problem defined here is to apply the same concept to predicting the probability of an Accident and it's severity. The aim is to predict severity of an accident using weather and local conditions, in addition to some personal/specific features.

This model is targeted towards use by Road and Safety Authorities as well the common user. The envisioned application is for the common user to receive prior warning about which roads to avoid or whether particular roads on the calculated route have a higher chance of accidents and warrants further attention and precautions. Road and Safety authorities could also this model to identify accident hotspots and prevalent problems.

2 Data acquisition and cleaning

2.1 Data Sources :

There are numerous datasets available on Accident Data, collected and hosted on government websites. The particular dataset used for this model is UK Road Safety : Traffic Accidents and Vehicles sourced from Kaggle.

The dataset comprises of two files :

1. AccidentInformation.csv: every line in the file represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns. Date range: 2005-2017
2. Vehicle_Information.csv: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016

The two above-mentioned datasets can be linked through the unique traffic accident identifier (Accident_Index column).

2.2 Data Cleaning

Once the two datasets are merged using Accident_Index column as a key, we begin to clean the data to prepare for modelling. There were several issues with the data.

Firstly, all rows have null/empty column values were dropped. Furthermore, all rows having column values out of range (-1) as defined in the dataset were dropped from the dataframe.

Secondly, plotting the frequency distribution of Accident Severity categories, it is evident that the dataset has imbalanced classes. The ratio of Slight:Serious:Fatal Accidents is 85:13:2.While this is to be expected, given that Slight accidents are much more common than Serious or Fatal accidents,the imbalance needs to addressed to ensure the model is not biased.

Another issue is that the current dataset consists of over 2 million rows.This is a large amount of data to process, requiring computing power that i do not have easy access to.Over 1.3 million rows of data are of class Slight Accident severity.Therefore, i am downsampling the majority class (‘Slight accident Severity’) so as to reach maximum dataset size of 600,000.

In addition to this, the class “Fatal” Accident severity is severely under-represented in the dataset with ~2% (22300 rows) belonging to this class. Preliminary runs of the model showed that the model performance on predicting accidents of Fatal Severity was below par, due to the lack of enough data available to train the model. Initial runs produced results with prediction accuracy less than 10%. This was a major contributor to bringing down the average model performance. Therefore, we are combining the 'Serious' and 'Fatal' categories of Accident Severity into a composite "Serious or Fatal" category.

2.3 Feature Selection

Next, we choose the attributes that will be relevant in helping us predict Accident Severity and Probability of the Accident. Particularly, we choose information and attributes that could contribute to the occurrence of an accident and it’s severity and exclude attributes that are dependent on the accident’s occurrence.

I have selected attributes that describe the conditions prevalent during the occurrence of the accident, namely, Weather, Light, local environment attributes in addition to personal data like Vehicular and Driver data fields.

With an iterative process of Visualization and modelling, the following features were finally chosen:

1. Light Conditions
2. Road Surface Conditions
3. Weather Conditions
4. Road Type
5. Sex of Driver
6. Day of the Week
7. Speed Limit
8. Driver Age Band
9. Urban or Rural Area
10. Junction Detail
11. Vehicle Type

2.4 Feature Optimization :

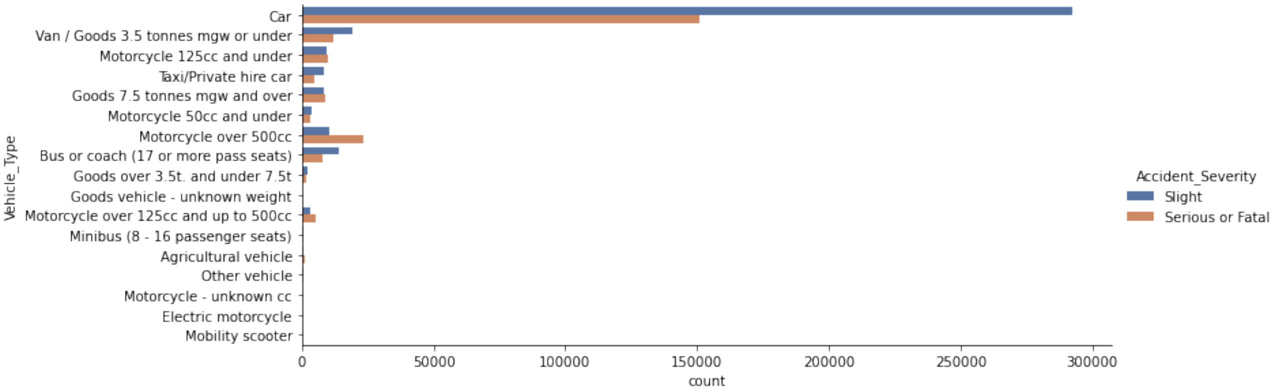


Figure 1: Vehicle Type Categories

Some features selected for the model, have categories with a degree of overlap. While these features are highly specific, the result of having too many categories leads to each category having relatively less data to train each category. Another step i have taken to improve performance is to redefine these categories into a more general overall feature.
For example, when we look at the different member categories of the attribute ‘Vehicle Type’, it is evident from the figure that there are multiple categories each for two wheelers, cars, and buses.
I have combined these sub-categories into one general category whose definition included each of the sub-categories.

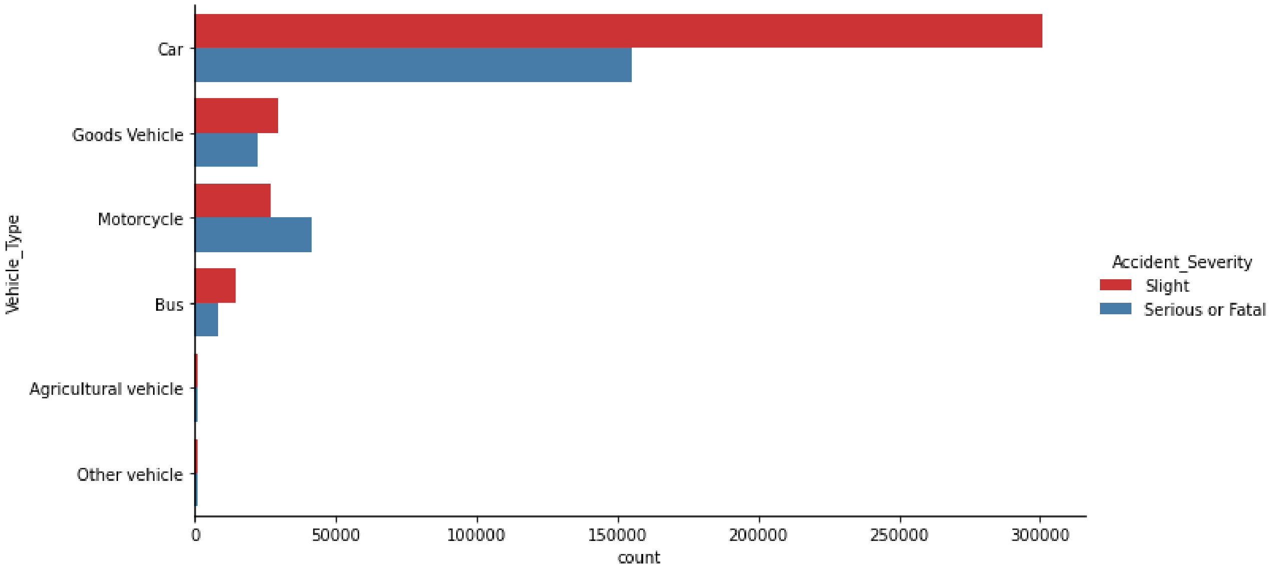


Figure 2: Vehice Type Post-Optimization

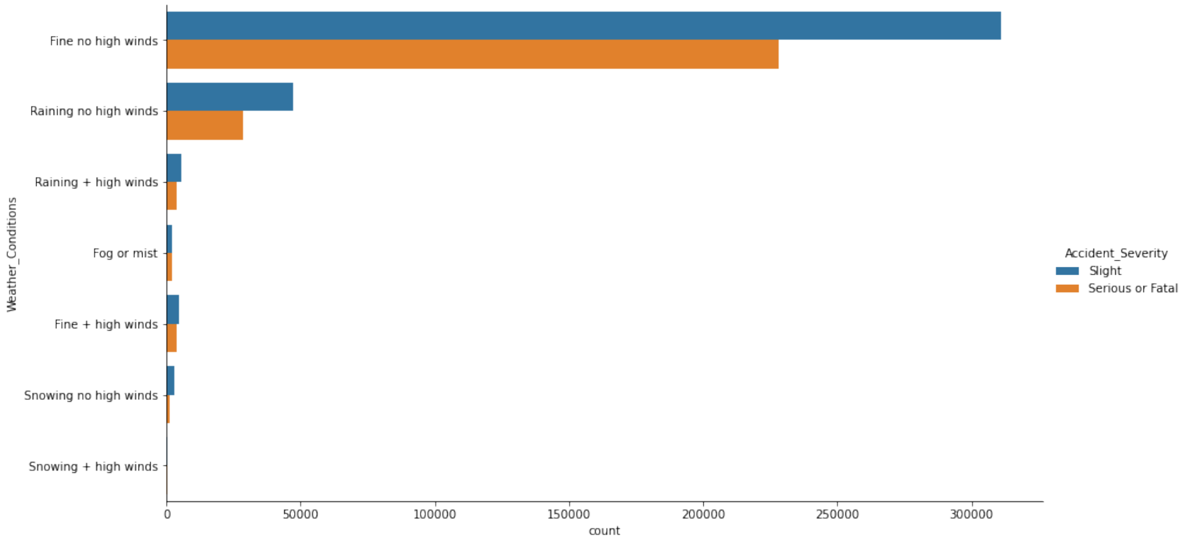


Figure 3: Weather Conditions

Another example is the Weather_Conditions attribute, which are split to numerous categories with some categories barely having any data. We are combining these similar categories together to reduce the number of different categories and increase the data available in each category.

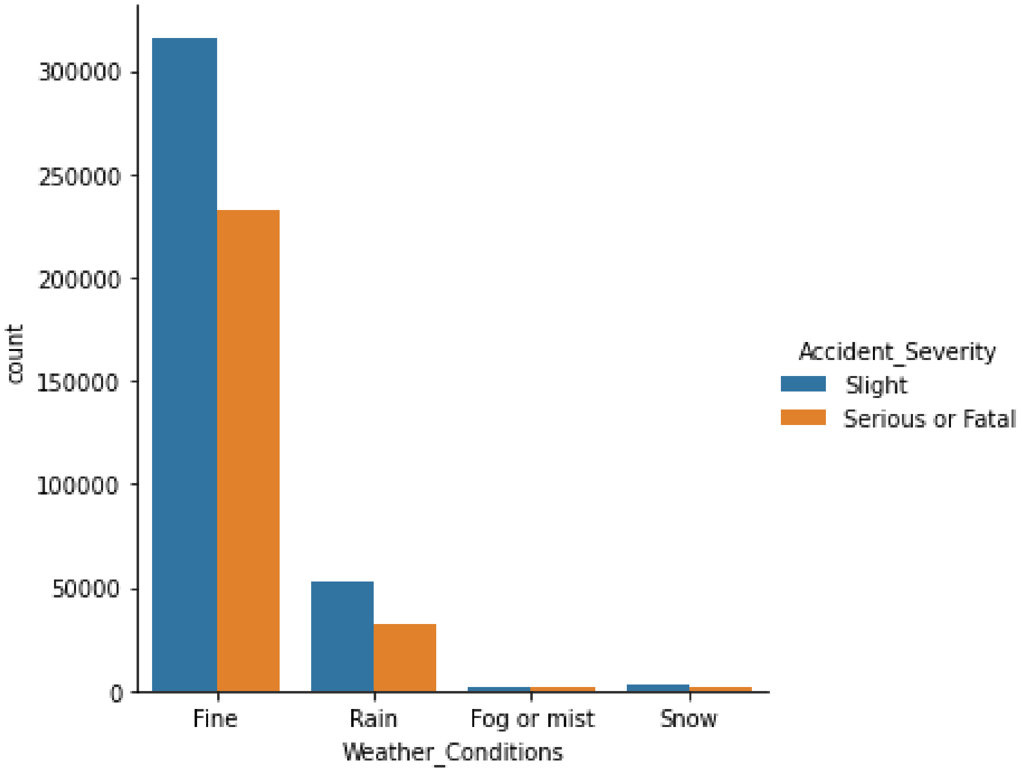


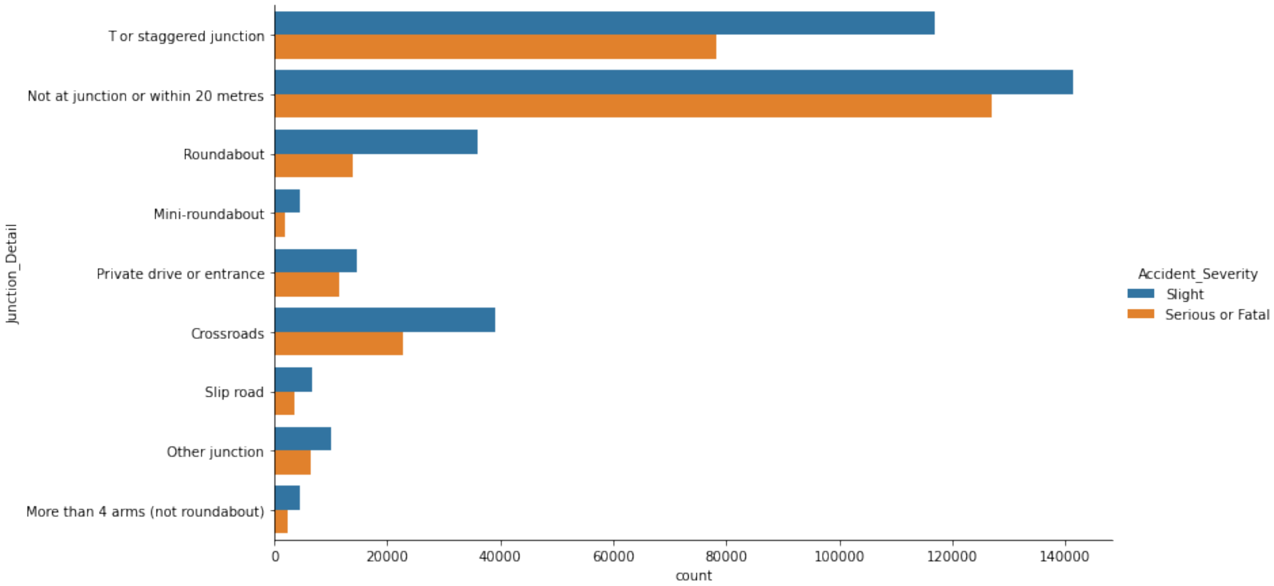
Figure 4: Weather Conditions after optimization

A similar process is carried out on all the other selected features, to optimize them towards better model performance.

3 Data Analysis

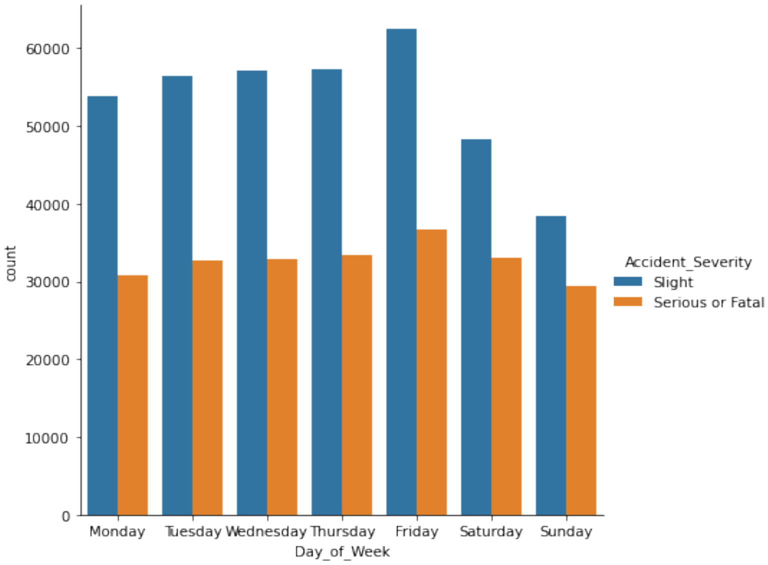
In this section, i have used the Seaborn package to plot the selected features against the target variable, which is Accident Severity.

3.1.1 Junction Detail



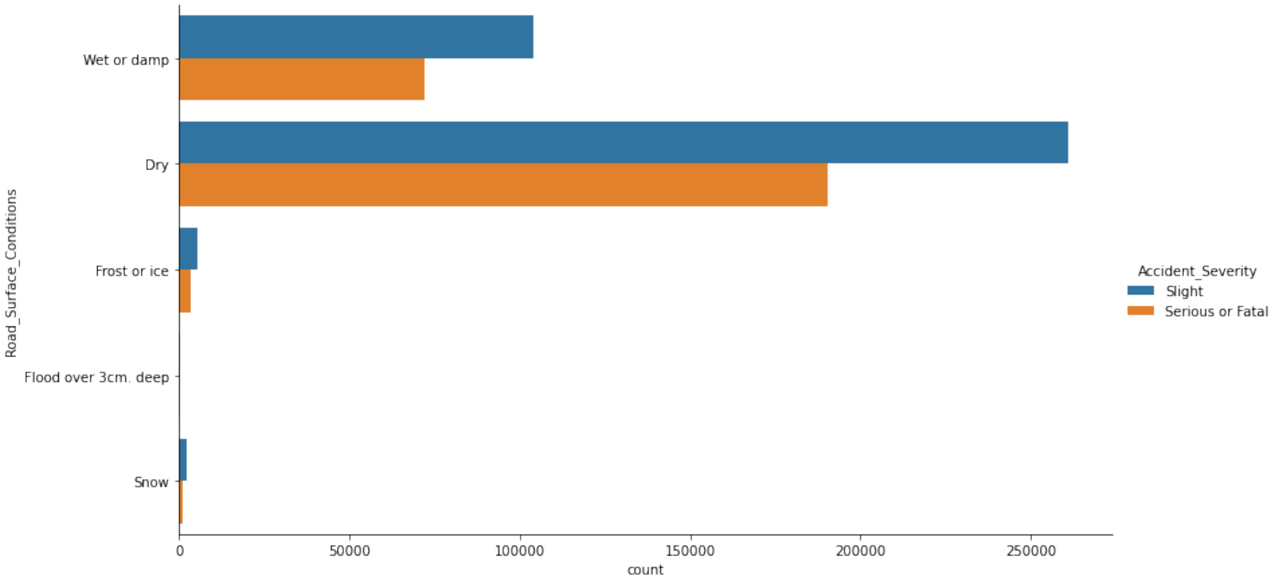
Tracking the trends across Accident Severities and Junction categories, the majority of accidents have occurred at locations that further than 20 metres from the nearest junction.

3.1.2 Day of Week



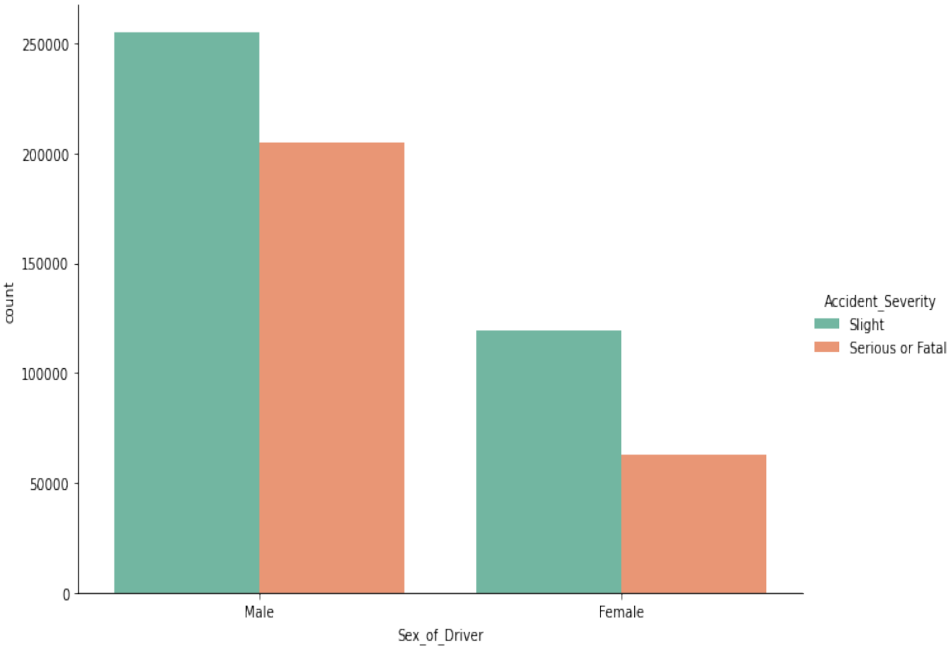
The distribution of accidents over the days of the week illustrates that a higher number of accidents occur on the weekdays compared to the number of accidents occurring on Saturdays and Sundays,i.e, the weekend.

3.1.3 Road Surface Conditions



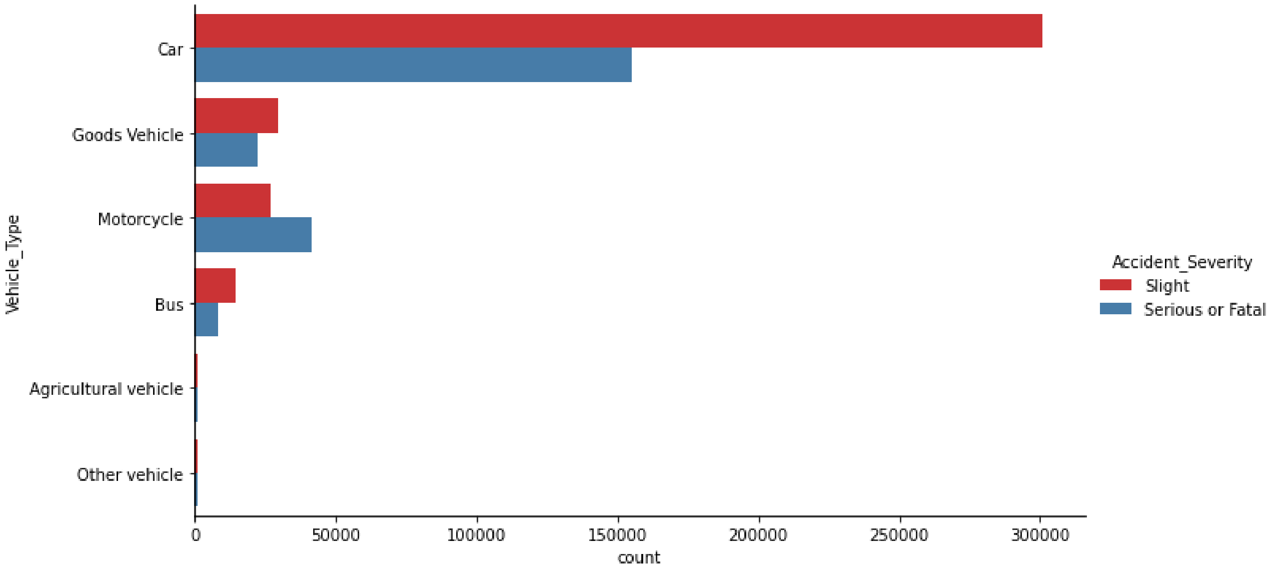
The distribution of accidents across Road surface conditions illustrate that the number of accidents that occur in Dry conditions are more numerous than any other category.This probably speaks out to the human error factor. However, the number of accidents that occur when the road is wet or damp is not a negligible number either. The wet and damp roads contribute to lower friction and makes driving harder in these conditions.

3.1.4 Sex of Driver



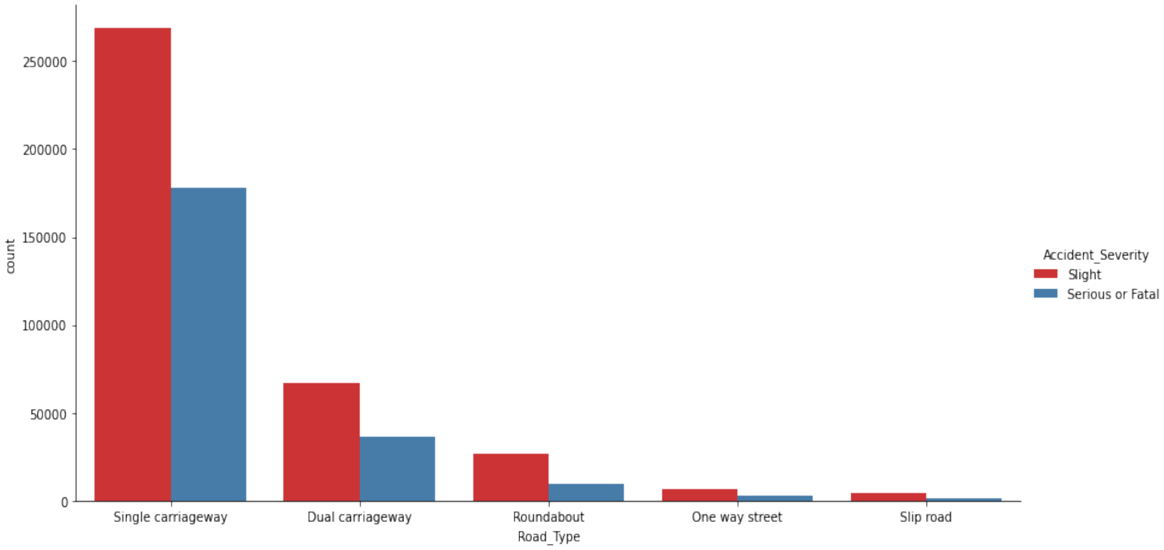
Accident severity distribution against the sex of the driver illustrates that Men are more likely to have collisions compared to women across both Slight and Serious/Fatal categories.

3.1.5 Vehicle Type



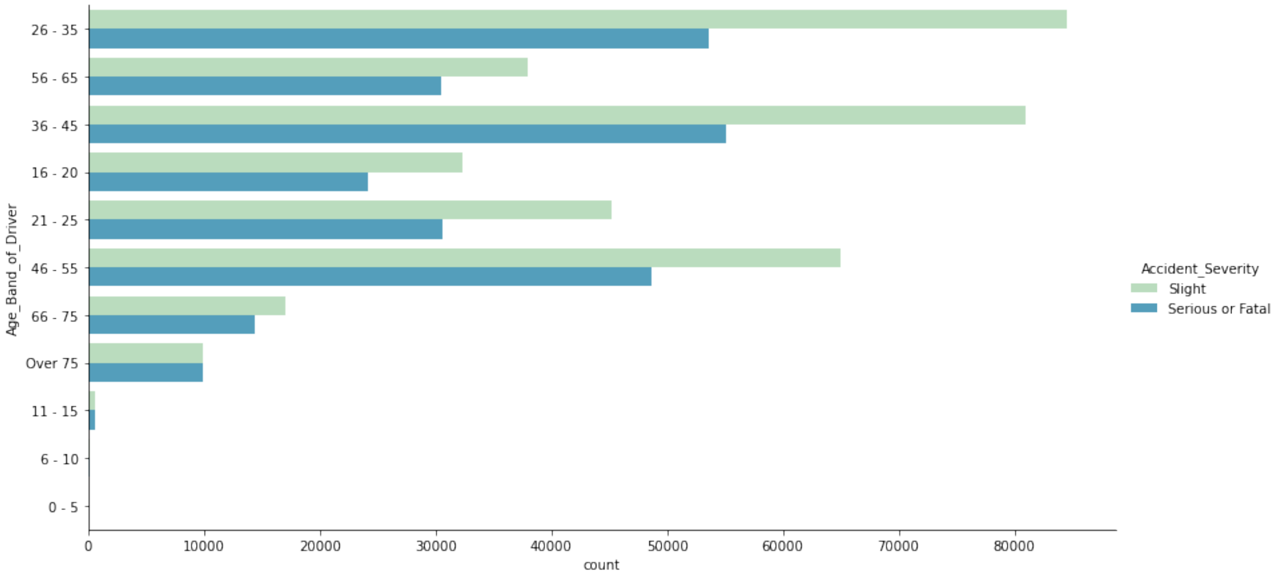
The interesting observation that can be made from the plot of Vehicle type v/s accident Severity shows that motorcycle accidents tend to be more serious or fatal compared to slight accidents.

3.1.6 Road Type



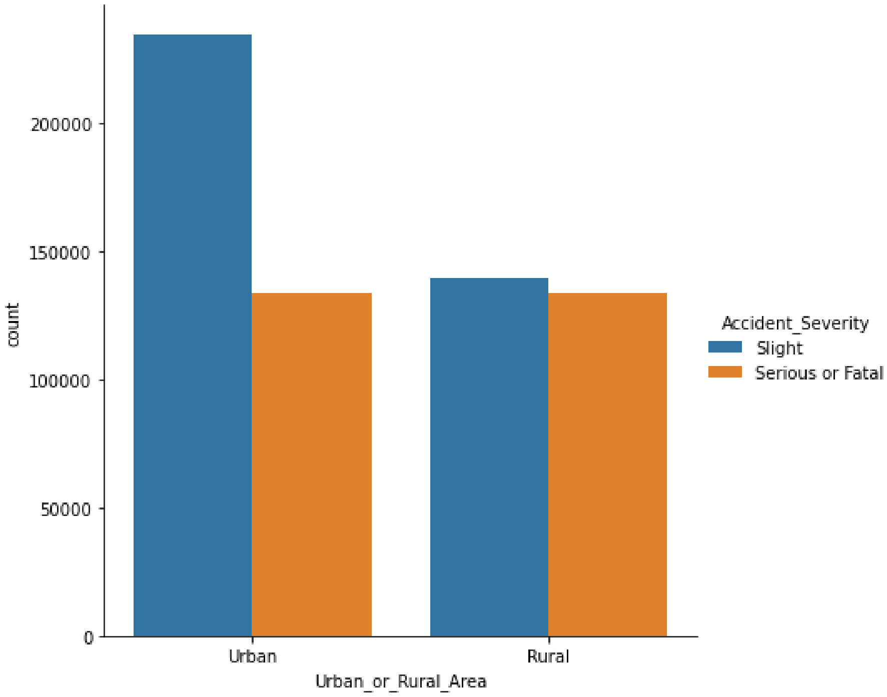
From the graph, we glean that most accidents occur on a Single carriageway where there is no separation between opposing flows of traffic.

3.1.7 Age Band of Driver



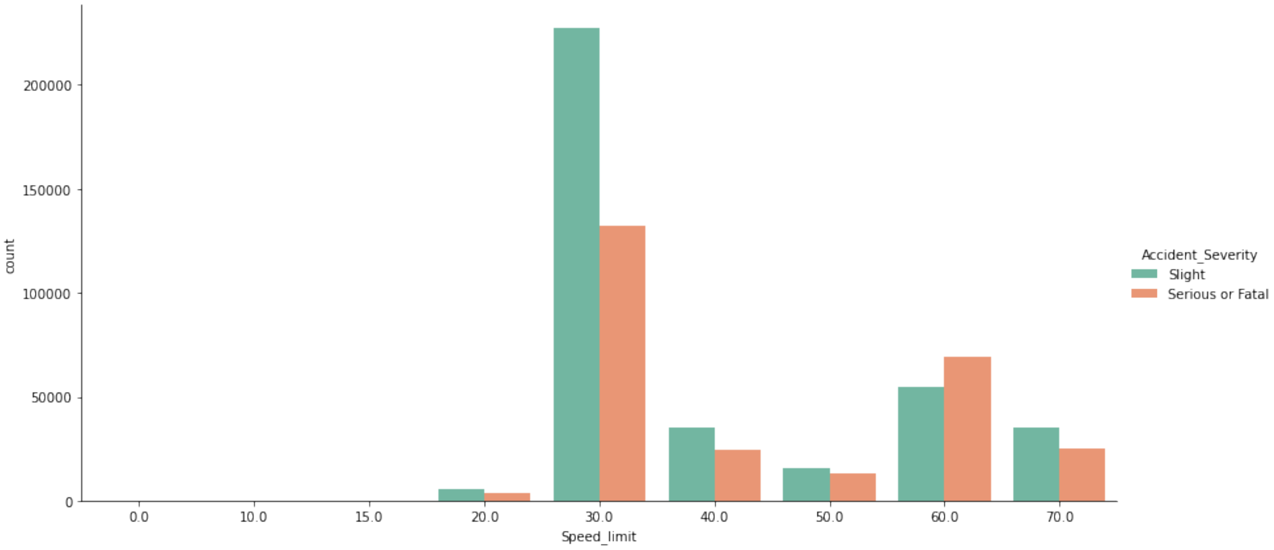
Plotting Accident Severity v/s DriverAge Band shows that drivers aged 26-35 have the maximum number of collisions. Suprisingly, drivers in the age bands of 36-45 are just as likely to have accidents.

3.1.8 Type of Area



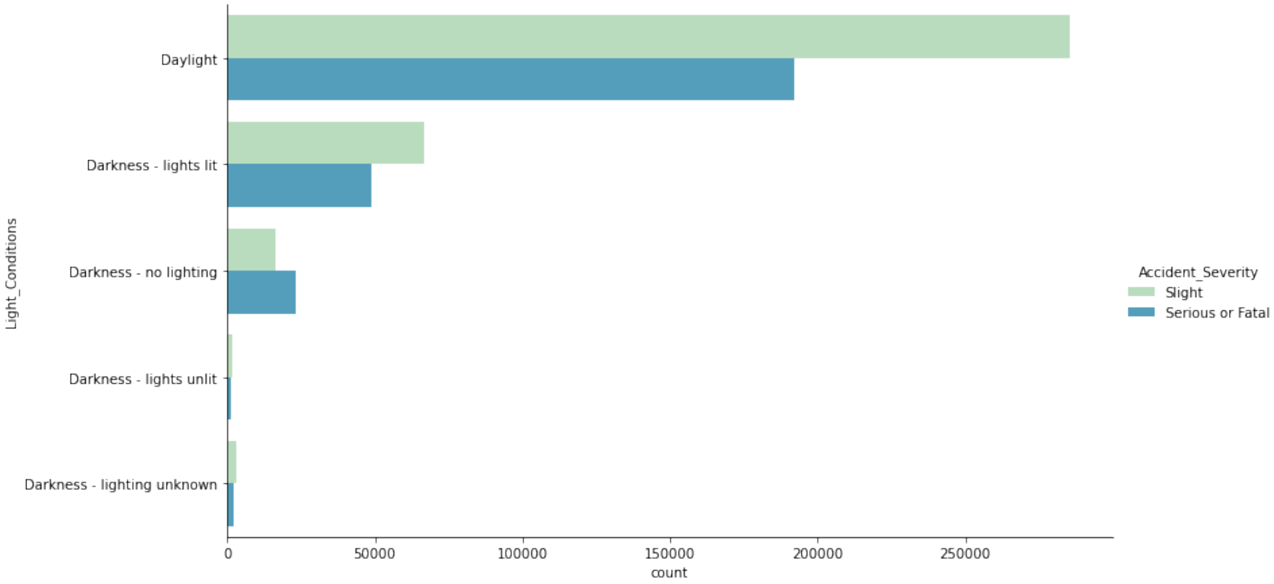
The majority of Slight accidents occur in Urban areas.This is probably due to the larger vehicular concentration that exists in Urban areas.When it comes to serious or fatal accidents, both rural and urban areas have similar numbers.

3.1.9 Speed Limit



Speed limits can be used a proxy/measure of speed that a vehicle can achieve on that road. While it is not as useful as actual speed of vehicle leading to the collision, it informs as to the most frequent speed at which accidents occur. Though there are the possibilities of outliers, the general recorded speed of vehicle in most cases will be around the mentioned speed limit. At the speed limit of 60, the number of serious or fatal accidents exceed the number of accidents of slight severity,

3.1.10 Light Conditions



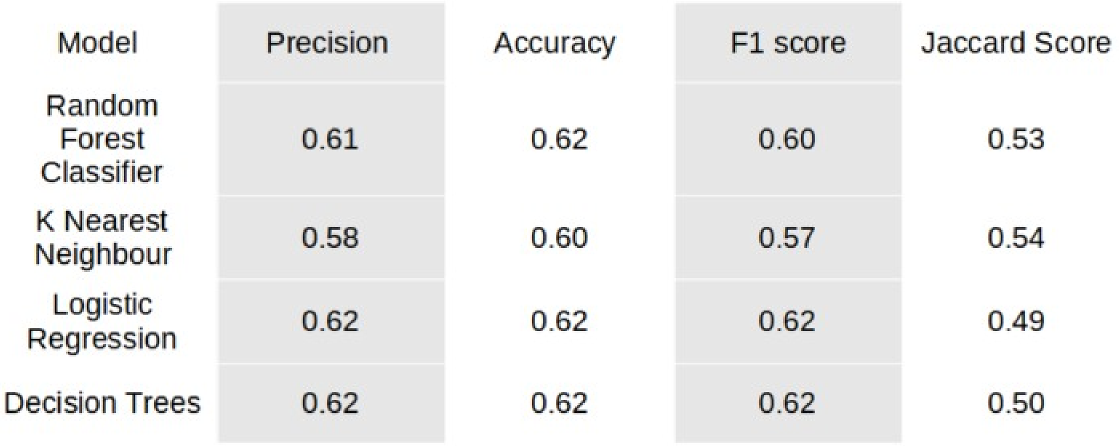
The majority of recorded accidents occur at daytime suggesting that visibility plays a far lesser role in accidents that previously thought. An interesting observations is that when it is dark with no external lighting, the number of Serious of fatal accidents exceed the number of slight accidents.

4 Predictive Modelling

There are two types of learning, unsupervised and supervised. Since this is about predicting the target variable, i.e., Accident Severity, this is a supervised learning model, namely classification. Classification is a type of machine learning where a model is trained to predict a class label for given input data.

4.1 Classification models

The different classification models used for this problem were Logistic Regression, K-Nearest Neighbours, Decision trees and Random Forest classifier. The K Nearest Neighbour was run with number of neighbours set to 4.Since this algorithm is time consuming, only a few runs of the model with different K values was possible. However,the dataset produced best results at K = 4.The other models were chosen for their memory efficiency and their speed,given the large dataset. I identified precision, recall,jaccard and F1 scores as the metrics by which to judge model performance.



Model	Precision	Accuracy	F1 score	Jaccard Score
Random Forest Classifier	0.61	0.62	0.60	0.53
K Nearest Neighbour	0.58	0.60	0.57	0.54
Logistic Regression	0.62	0.62	0.62	0.49
Decision Trees	0.62	0.62	0.62	0.50

Table 1: Model Performance

From the table that has recorded the chosen evaluation metrics for each model, it is evident that all of the model have similar performance with minimal differences. It is clear that the KNN model has the worst performance out of all the models, however it is difficult to declare a model that has a clear performance advantage over the others.

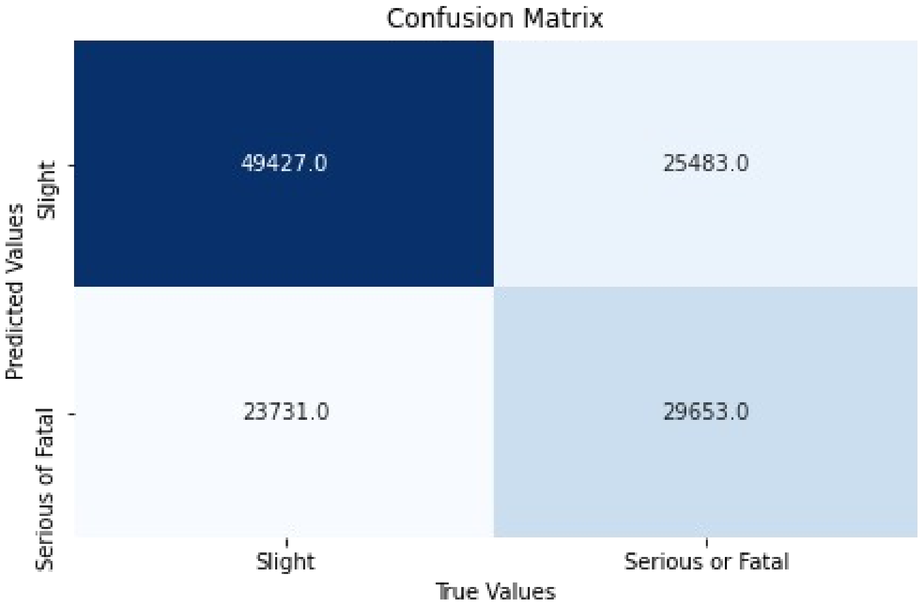


Figure 5: Confusion Matrix of Decision Tree Model

Choosing the Decision Tree model and viewing it’s confusion matrix as displayed above gives us a better idea of the model’s performance. The model is best able to predict Slight accidents correctly, indicating a good True Positive score. The model is able to distinguish True Negatives but with a lower degree of accuracy. It also struggles in distinguishing False Negatives and False Positives. This is where the performance is lost and is an area that needs to be addressed in future work.

5 Results and Discussion

In this study, I have analysed the relationships between local weather data, environment data and vehicular and driver data and it’s impact on accident severity. I have built a model to predict the Accident Severity along with it’s probability. These models can useful in helping an end user assess the chances of an accident and it’s severity before setting out on a journey. This can help inform their choices to postpone their travel, or, to use alternate routes or mediums of transport.

Some models could benefit from additional processing power, for example, the Decision Tree model could have the maximum depth increased and K Nearest Neighbours model might profit from the ability to user higher values of K.

The models reflect the data which collected at a macro level and is therefore lacks the specific contextual data to improve it’s prediction accuracy. Ideally, to improve the model's performance, we should narrow our field of study to a smaller region in the UK, a city preferably and work on predicting accident severity for that smaller region and include locally relevant attributes. These models can be considered as an initial foray into the problem of predicting accident severity and definitely show that there is promise in this endeavour and warrants deeper analysis and implementation.

6 Conclusions and Future Work

I was able to achieve ~16% improvement from the initial model in the classification problem, and ~62% accuracy. However, there is still scope for significant improvement of the model in this study. Additional data on average traffic flow at locations of the accidents would contribute to an increase in accuracy of predicted probabilities.

Models in this study mainly focused on individual well-defined features. Using the model with location specific data and trends from Google Maps/Waze will help in improving model performance and also personalizing the predictions. It could also suggest alternate routes avoiding accident prone areas or alternate forms of transport. With the additional data and context that these applications could provide, it would also be possible to identify accident hotspots particular to a specific weather/light condition. For example, a particular road sees increased number of accidents when it is raining. The transport and road safety organization could then look at the data and suggest measures to reduce the risk of accidents, like imposing a lower speed limit when it is raining. These suggested applications for the model are obviously much more difficult to execute and realize, but if done, could work to reduce accidents and increase safety on dangerous roads.