Predicting the Severity of Road Accident

Kamaleshwar

August 31, 2020

1. Introduction

1.1 Background

Road traffic accidents are just one example of a type of accident which can happen in adverse weather conditions. Whether it happened because of a faulty piece of equipment at the roadside, another road user, or being knocked over by a car whilst crossing the road, there may be a chance to make a claim with a personal injury solicitor. Accidents vary from claim to claim, but the contributing factors are often very similar, if not the same.

Factors known for causing accidents during bad weather are:

- Visibility
- Other motorists
- Faulty vehicle or equipment
- Poor road conditions
- Snow & Ice (ungritted roads)

Therefore, it is advantageous to accurately predict whether an accident can happen and what will be the severity of it. For example, this information can be used to avoid accidents by choosing alternatives for the travel.

1.2 Problem

As a nation on wheels, driving around is part of the life. Imagine You are driving to another city for work or to visit friends. It is rainy and windy and on the way, you come across the terrible traffic jam on the other side of the highway. Long lines of cars are barely moving. As you keep driving police car start appearing from afar shutting down the highway. Its an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to happening.

Now wouldn't be great if there is something in place that could warn you given the weather and road conditions about the possibility of you getting into a car accident

and how severe it would be, so that you would drive more carefully or even you change your travel if you are able to.

1.3 Interest

Government organizations, Individuals, Road safety Organizations

2. Data acquisition and cleaning

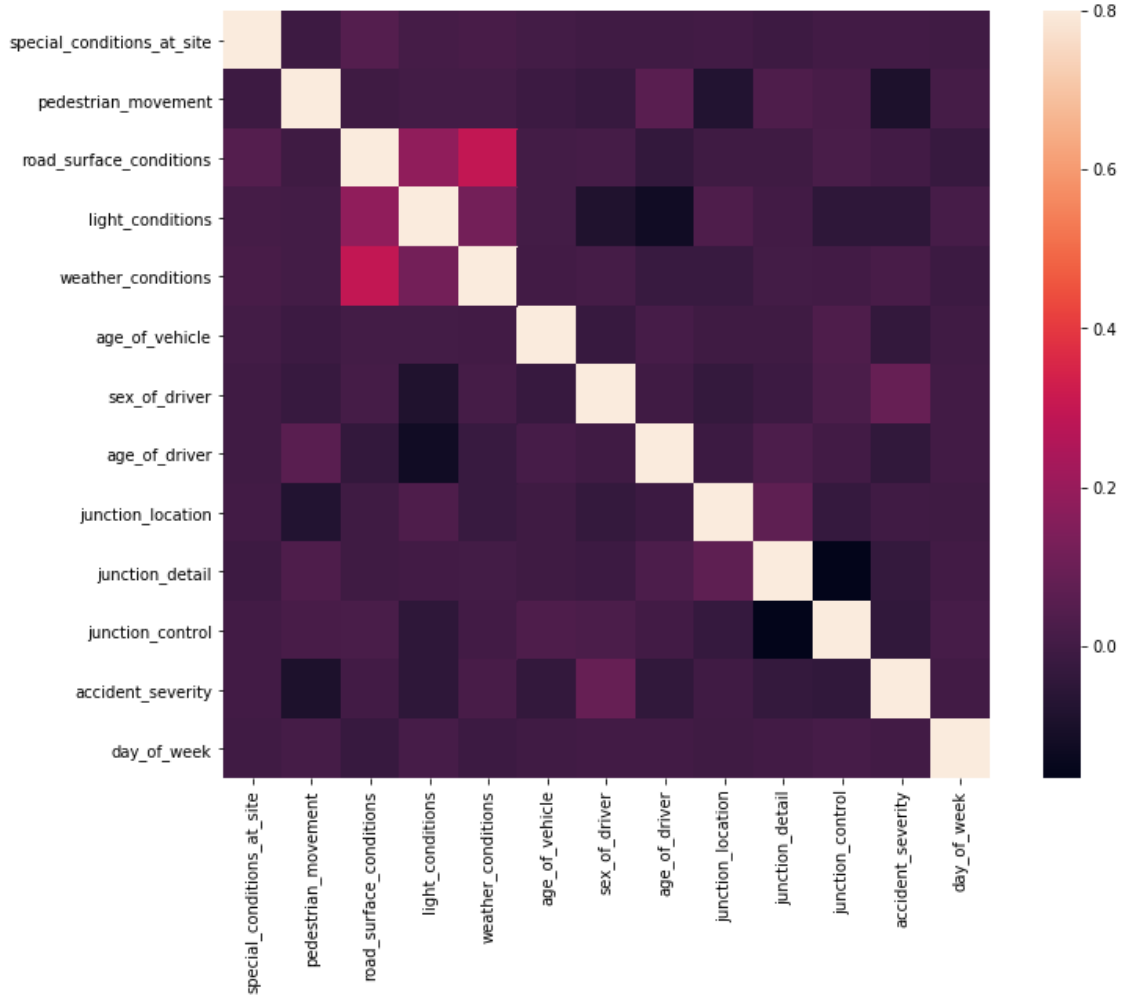
2.1 Data sources

Data has been obtained from Kaggle.com.

Primarily Captures Road Accidents in UK between 1979 and 2015 and has 70 features/columns and about 250K rows.

2.2 Feature selection

The features that are used to predict the occurrence of Accident are found by plotting a heat map to determine the correlation between the variables.



As seen by the graph, there no linear relationships present, besides between the added features of weather condition, road surface, and light condition. This makes sense, as weather-, road-, and light conditions are dependent on each other. When it is raining, one can presume that the road

condition atthe same time is also wet. Absence of other linear relationships can be explained by the fact that almost everything is a categorical variable. Even the weather related conditions barely achieve 0.4 on the Pearson correlation as they are nominal features as well. Hence there is no justification and indication touse predictive models based on linearity

3. Exploratory Data Analysis

3.1 Calculation of target variable

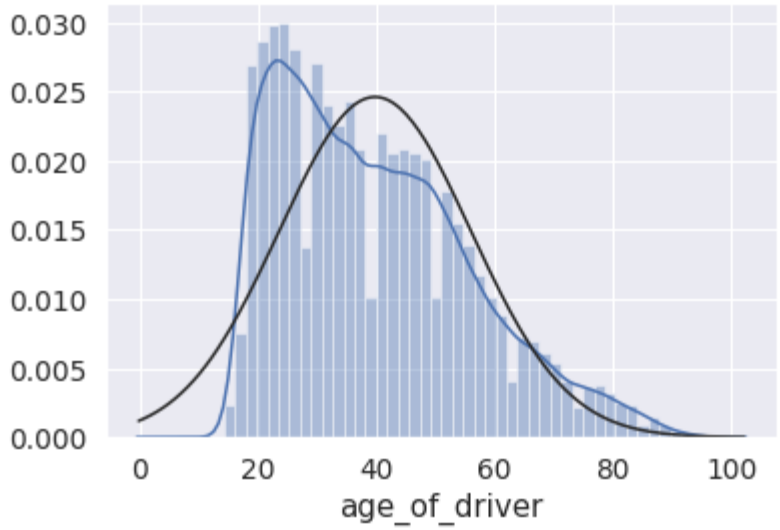
The following features are used to predict the severity of an accident.

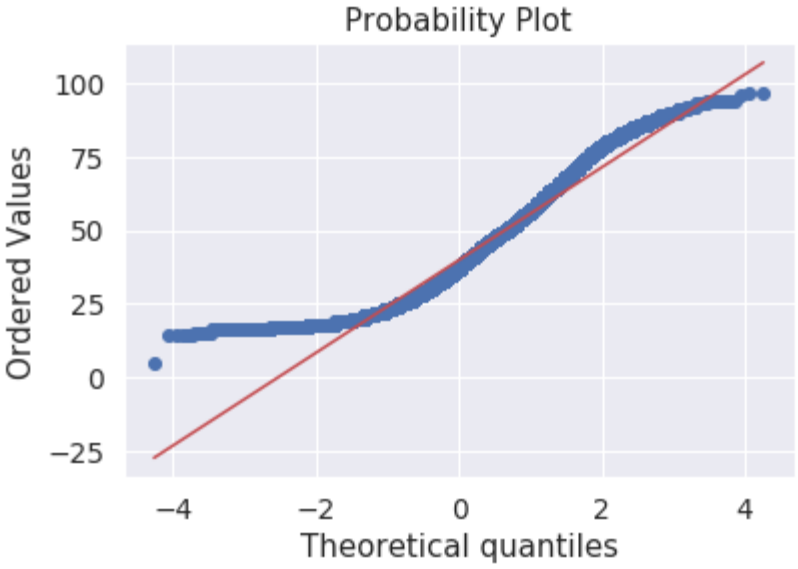
```
Index(['special_conditions_at_site', 'pedestrian_movement',
       'road_surface_conditions', 'light_conditions', 'weather_conditions',
       'age_of_vehicle', 'sex_of_driver', 'age_of_driver', 'junction_location',
       'junction_detail', 'junction_control', 'day_of_week'],
      dtype='object')
```

All the classification algorithms are used to predict the Severity of accident. And the one with best evaluation score will be selected.

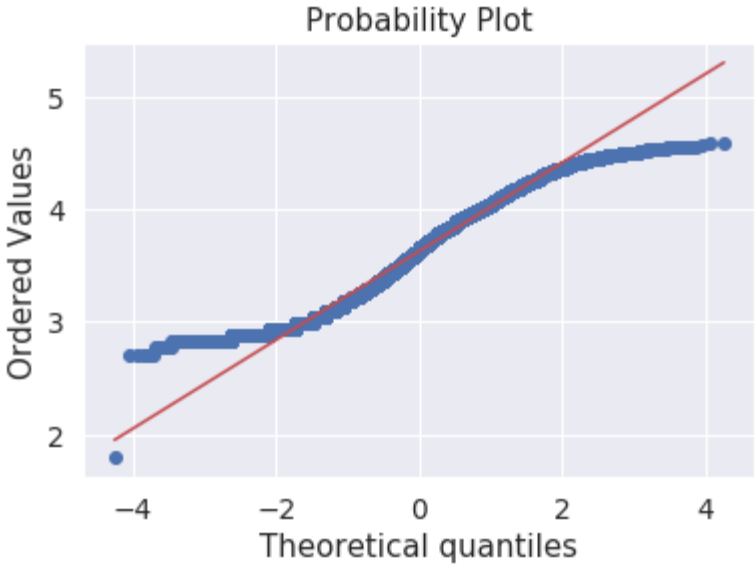
3.2 Normalizing the Non Categorical variables.

The next step was to normalize the only features that were not categorical: age of the driver and age of the car. Normalization involves taking the logarithm of the given features. This is done to because high values for certain variables computationally skew results more in favour of that variable, than their actual contribution. In this case, age of the driver for example has values ranging from 18-88. When the majority of other categorical variables are binary or limited within 1-8 categories.





In this case, age of the driver and age of the vehicle were the only variables with a high numerical variance, and therefore logarithms were taken of both variables. Furthermore, taking the logarithm of both the age of the driver and age of the vehicle improved the fit by altering the scale, and making the variables more "normally" distributed.



After taking the log, one can notice that the values range from approximately 2.5 to 4.5. This increases the performance of machine learning algorithms, as the numerical values do not have disproportionate amounts of computing value compared to all the other categorical variables.

4. Predictive Modeling

4.1 Prediction without weather based features

I decided to predict the severity without including the weather based features.
Below is the evaluation score for different predictive models.

Machine Learning algorithm scores without weather related conditions

	Model	Score
2	Logistic Regression	92.47
3	Random Forest	92.20
0	Support Vector Machines	92.06
6	Stochastic Gradient Decent	92.06
7	Linear SVC	92.06
5	Perceptron	92.00
1	KNN	90.40
4	Naive Bayes	90.00
8	Decision Tree	86.87

4.2 Prediction with weather based features.

Machine Learning algorithm scores with weather related conditions

	Model	Score
3	Random Forest	92.59
2	Logistic Regression	92.47
0	Support Vector Machines	92.12
5	Perceptron	92.06
7	Linear SVC	92.06
6	Stochastic Gradient Decent	91.19
1	KNN	90.20
4	Naive Bayes	87.94
8	Decision Tree	86.16

5. Conclusions

The results indicated that adding weather-related features to a machine learning algorithm in predicting severity of an accident did not substantially change the accuracy of models. The results indicate a high accuracy. As this is multilabel classification, the accuracy measure in this case computes the amount of labels predicted that exactly match the corresponding set of labels.

However, we have to take into account the accuracy paradox as sometimes it may be desirable to select a model with a lower accuracy because it has a greater predictive power on the problem. In our dataset there is a large class imbalance as most accidents are classified as mild(class 3) as shown in the graph below.

```
Confusion matrix:
 [[   2    2   22]
 [   1   53  274]
 [   3   73 4070]]
Classification report:
              precision    recall  f1-score   support

           1       0.33      0.08      0.12        26
           2       0.41      0.16      0.23       328
           3       0.93      0.98      0.96      4146

   micro avg       0.92      0.92      0.92      4500
   macro avg       0.56      0.41      0.44      4500
weighted avg       0.89      0.92      0.90      4500
```

A model can predict the value of the majority class for all predictions and yield a high accuracy although almost all of the predictions would concern the majority class; hence, yielding a very high accuracy. Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. Precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. Moreover, another metric is F1 score which is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score and returns an compromise between precision and recall.

A clean and unambiguous way to present the prediction results of a classifier is to use a use a confusion matrix. On below is for without weather conditions and one below is with weather conditions included.

```
Confusion matrix:
 [[   2    1   23]
 [   1   56  271]
 [   1   71 4074]]
Classification report:
              precision    recall  f1-score   support

           1       0.50      0.08      0.13        26
           2       0.44      0.17      0.25       328
           3       0.93      0.98      0.96      4146

   micro avg       0.92      0.92      0.92      4500
   macro avg       0.62      0.41      0.45      4500
weighted avg       0.89      0.92      0.90      4500
```

The precision, recall, and F1 score are also at high levels of 0.89, 0.92, and 0.9 respectively - meaning that the classification is successful and the accuracy of the model is more or less 90\% when investigated on multiple metrics.

When taking into consideration the weather condition, the lighting condition, and road surface conditions the accuracy of machine learning models are as follows:

sns.heatmap(cm,annot=True,fmt="d") plt.show()

The results indicated that adding weather-related features to a machine learning algorithm in predicting severity of an accident did not change the accuracy of the model. When adding three features of light condition, weather condition, and the condition of the road surface, the measures of recall, precision, and f1-score remained unchanged.

When looking at the overall performance of all of the algorithms, there was an increase in accuracy between the data with weather conditions when compared to data without weather related conditions. Namely, random forest algorithm increased performance by 0.59%. The previous top performer when no weather related conditions were introduced, Logistic Regression, sustained the same level of accuracy. Hence, it was concluded to further scrutinize the recall, precision, and f1-score of random forest algorithm to see whether there was an actual change in prediction power.