Predicting the severity of Car Accident

Kushagra Kumar Rajput

September 5, 2020

1. Introduction

1.1 Background

Today owning a car is no longer a luxury in-fact it is now a necessity of having a car in every household. As the number of cars is increasing day by day on the road the risk of accidents is also increasing. Accidents not only cause human life loss, but financial losses are also incurred. All accidents do not have the same intensity or severity and losses occur vary largely based on the severity of the accident. The severity helps the government and other agencies to predict the number of serious patients that are going to be there. Moreover, with the help of severity government can also plan to work in those areas where frequently more severe accidents are going to occur based on the location of accidents. Furthermore, severity prediction can also help in planning for the weather conditions when accidents are more prone to be severe.

1.2 Problem

Data that might contribute in determining the severity of the accident might include the number of people involved in the event, number or type vehicles involved in the accident, how were the weather conditions at the time of the incident and many more. This project aims to predict the severity of the accident based on these data.

1.3 Interest

The Government and the insurance companies would be very much interested in the prediction of severity of the accidents as it will help in better planning the road safety so that number of severe accidents can be reduced.

2. Data acquisition and cleaning

2.1 Data Sources

There is a large amount of data present on the internet regarding road accidents, but to find the data set to meet the project requirement which to find data set with severity was a bit difficult. Finally, after a lot of google search, I decided to move forward with data on Kaggle.com provided by the Department of Transport UK.

2.2 Data Cleaning

Data used in this project was downloaded from the Kaggle.com which was published by the Department of Transport UK. There were some of the rows which were missing the values, so I

decided to remove those entries from the data as I already a large amount of data even after removing those entries.
Since no data set is perfect as it is and the same is true for this data set a well. Many attributes would not help to learn and hence predict the severity of the accident.

2.3 Feature Selection

After cleaning the data, removing all the null values and the redundant data there were 607216 data samples and 31 features in the data set. The definition of each feature was looked upon and the best features were chosen from the feature set that is going to help in determining the target variable i.e. accident severity.
The list of dropped features is:

- Accident_Index
- Location_Easting_OSGR
- Location_Northing_OSGR
- Longitude
- Latitude
- Date
- Time
- Local_Authority_.Highway
- Local_Authority_.District
- X1st_Road_Number
- X2nd_Road_Number
- Did_Police_Officer_Attend_Scene_of_Accident
- LSOA_of_Accident_Location
- Police_Force
- Day_of_Week
- X1st_Road_Class
- Speed_limit
- Junction_Detail
- Junction_Control
- X2nd_Road_Class
- Special_Conditions_at_Site
- Carriageway_Hazards
- Urban_or_Rural_Area

All these features were not considered in calculating the severity of the accidents as either these were some unique id given to the data or date, time etc.

Following is the list of features that were used to train the model:

- Number_of_Vehicles
- Number_of_Casualties

- Road_Type
- Pedestrian_Crossing.Human_Control
- Pedestrian_Crossing.Physical_Facilities
- Light_Conditions
- Weather_Conditions
- Road_Surface_Conditions

3. Exploratory Data Analysis

3.1 Target Variable “Accident Severity” calculation

In the given data set the Accident severity contains three values 1-Low severity, 2-Medium severity and 3-High severity, or we can say that data is divided into three classes based on the severity attribute. Just by taking an overview of the problem it seems to be a classification problem, where we need to classify data into either of the three classes based on the values of the given feature set.

3.2 Histograms of different features

We have plotted various histograms of the feature set just to gain an insight of what type of data are we dealing with. Fig 1 shows the histogram of different feature sets. Here we can clearly observe which individual features have contributed to more number of road accidents.

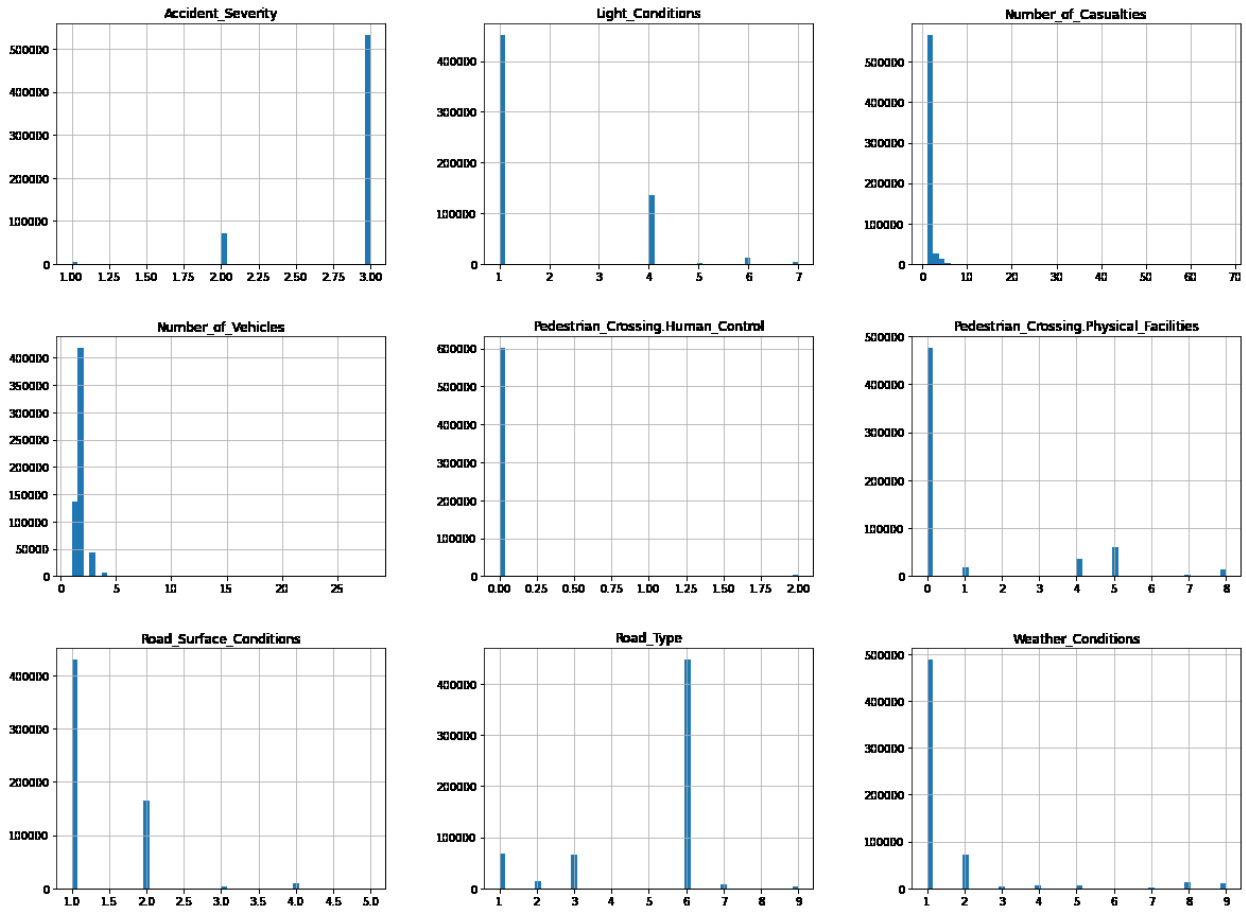

fig1. histogram of feature set

3.3 Scatter Matrix

First of all we divide the attributes into two attribute groups as it makes it easy for us to visualize the plot.

attributes1=["Number_of_Vehicles","Number_of_Casualties", "Road_Type", "Pedestrian_Crossing.Human_Control","Accident_Severity"]

attributes2=["Pedestrian_Crossing.Physical_Facilities", "Light_Conditions", "Weather_Conditions", "Road_Surface_Conditions","Accident_Severity"]

The scatter matrix gives a good idea whether the features are related to each other are not.

Fig 2 gives a scatter matrix for the first group of attributes. Whereas, the Fig 3 gives a scatter matrix for the second group of attributes.

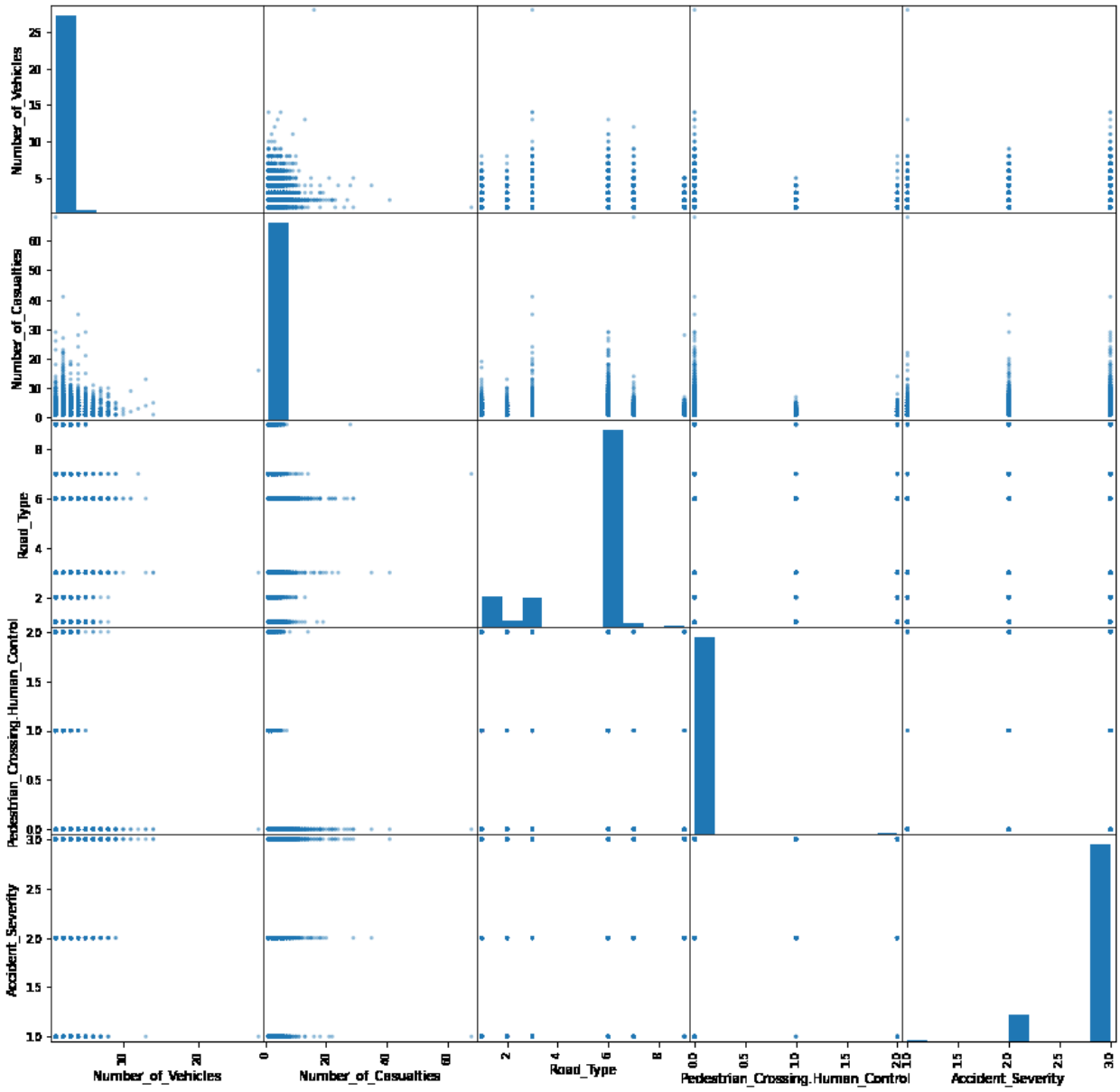

fig2. scatter matrix for attribute group 1

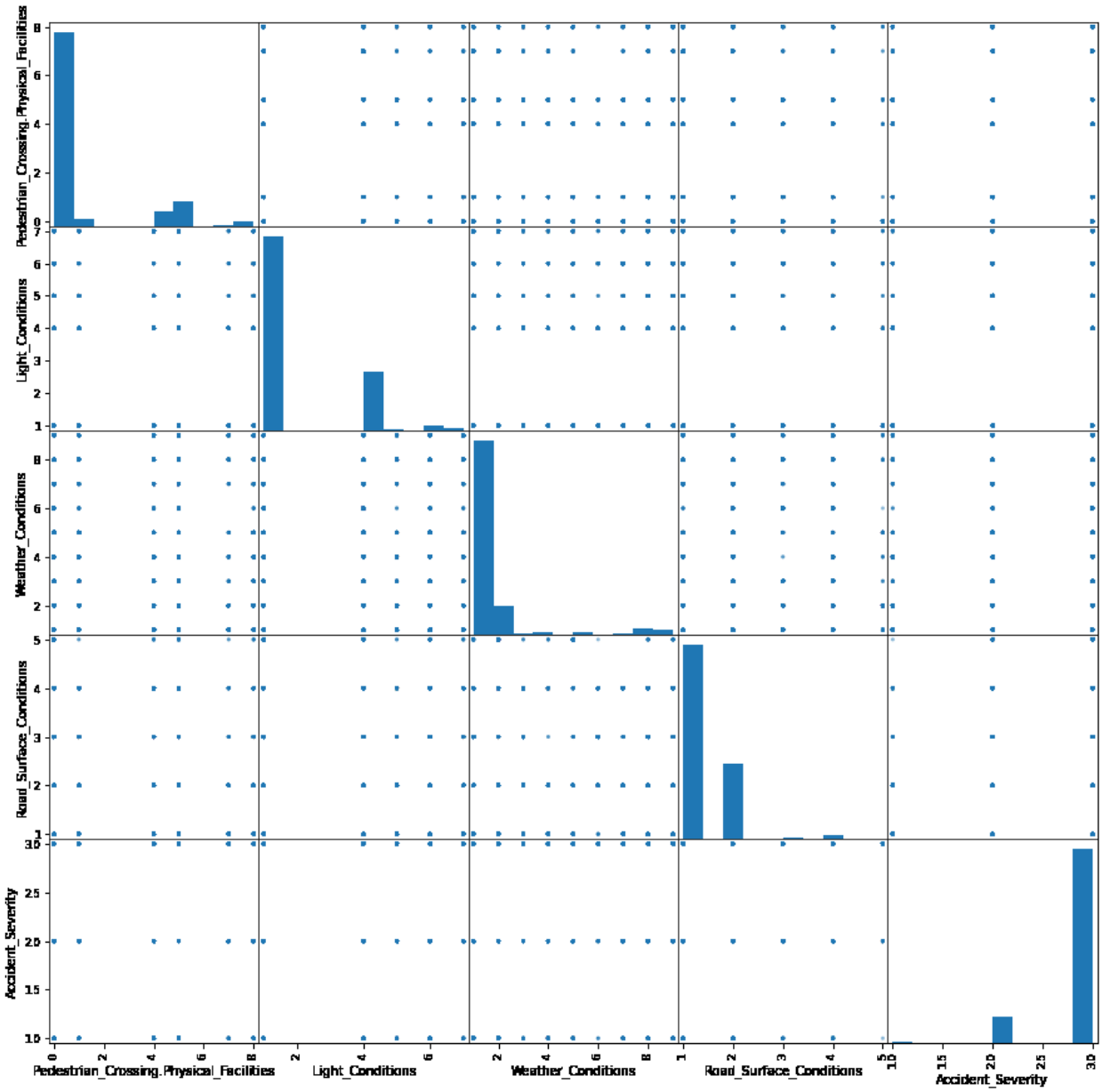

fig3. scatter matrix for attribute group 2

3.4 Correlation between features

Here we have used the correlation function of the pandas library to get a numeric value for the correlation between the features in the data. The below table gives a correlation value between Accident Severity and other features.

Feature	Corr Value
Accident_Severity	1.000000
Number_of_Vehicles	0.094896
Weather_Conditions	0.027842
Road_Surface_Conditions	0.018471
Pedestrian_Crossing.Human_Control	0.003217
Pedestrian_Crossing.Physical_Facilities	-0.014971
Road_Type	-0.035094
Light_Conditions	-0.048739
Number_of_Casualties	-0.056895

4. Classification models

The application of classification models is very straightforward. We divided the samples into three classes. The number of samples in each class were about the same. We chose logarithmic loss as the metric here because the results would probably be presented with probabilities and logarithmic loss puts more emphasis on the probabilities than other metrics. Logistic regression, SVM, KNN and Decision Tree were tuned and built. Among the individual models, the SVM model performed the best (~40.5% Log Loss), and the Decision tree performed similarly as the SVM model, though the differences between models were small.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.766098	0.819745	0.217224
Decision Tree	0.768568	0.819072	0.392933
SVM	0.768568	0.819072	0.405461
Logistic Regression	0.768568	0.819072	0.398769

5. Conclusions

In this study, we analyzed the relationship between Accident Severity and other road conditions. We identified road conditions, weather conditions and number of vehicles involved in accidents among the most important features that affect accident severity. We built classification models to predict the accident severity. These models can be very useful in helping to determine the severity of the accidents and help the local authorities to be ready for more severe conditions in advance.