ML Hackathon

Date: 18th November 2018

Predicting the severity of Road Accidents for better emergency services in the vicinity.

Team TGT

Gautami Gupta	**Tejas Kotha**	**Tanmay Jain**
IMT2016069	IMT2016112	IMT2016023

Overview

181,384 accident casualties were recorded on Britain's roads in then year 2016 and out of which 1,792 were fatal. The long-term trend in the numbers killed and injured from road accidents has been declining, especially in the last two decades.

According to the World Health Organisation, more than 1.25 million people die each year as a result of road traffic crashes. The newly adopted 2030 agenda for Sustainable Development has set a target to halve the global number of deaths and injuries from road traffic accidents by 2020. We believe that this target can be only achieved if the rescue teams are able to provide with the required first aids to the victims of these accidents and the necessary road accident preventions by the concerned traffic authorities.^

Introduction

We feel that it is very important that in the areas prone to accidents should have emergency medical consultant who can act immediately. The availability of a medical examiner is not the only determinant of the victim's health but they should also be equipped with the required medical instruments and first aids. This can only be achieved if we are able to predict the type and severity of these accidents. This will also provide the basis for determining and monitoring effective road safety policies to reduce the road accident casualty toll.

We aim to predict the severity (Fatal, Severe or Slight) of accidents with the help of the records of the past 10 years collected by the UK Traffic Police Department.*

*: Department for Transport, Road Accident Statistics Branch. (2015). *Road Accident Data, 2014*. [data collection]. UK Data Service. SN: 7752, http://doi.org/10.5255/UKDA-SN-7752-1
^: https://researchbriefings.parliament.uk/ResearchBriefing/Summary/CBP-7615

Dataset

The data used in this classification problem are records of the traffic accidents that happened in UK between 2005 to 2014 and derived by the UK government. We collected the dataset from Kaggle (https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales).

The size of the dataset is 1.6M x 35. We divide the training and the testing data in the ratio 1:5. The attributes on which we are training our model on are as follows:

- Number_of_Vehicles: number of vehicles involved in the accident or damaged
- Number_of_Casualties: number of people injured/died in the accident

- Day_of_Week: the day of the week already encoded into numbers from 1-7
- Time: timestamp of the accident, in the format of HH:MM
- Year: year in which the accident took place.
- Road_type: type of the road, eg. One way street, Two way street, single carriageway etc
- Speed_limit: speed limit of the road on which the accident took place
- Pedestrian_Crossing-Human_Control: status of a human patrol around the accident spot
- Pedestrian_Crossing-Physical_Facilities: status of the facility for the pedestrians near the accident spot
- Light_conditions: Condition of the light at the time of the accident, eg. if it dark with/without the street light or daylight
- Weather_Conditions: Weather at which the accident took place, eg, snowy, sunny, rainy etc
- Road_Surface_Conditions: Condition of the road, eg dry or wet
- Urban_or_Rural_Area: It is in already encoded form, 1 for urban and 0 for rural.
- Accident_Severity: This is the column we will be predicting, 1 for fatal accidents, 2 for serious accidents and 3 for slight accidents.

Preliminary Findings

As a start, we explored the data to figure out some of the peculiarities of this problem. And our preliminary findings then shaped the approaches we decided to take.

- The data was highly biased. The number of fatal accidents were 1.3% of the total number accidents, whereas slight accidents were contributing upto 85% in the total accidents taking place.
- There were many attributes with object data type which had only 5-6 unique values:
    - Road_Type(Single Carriageway, Dual Carriageway, One way street, roundabout, slip road)
    - Pedestrian_Crossing-Human_Control(None within 50 metres, Control by other authorised person, Control by school crossing patrol)
    - Pedestrian_Crossing-Physical_Facilities(Zebra Crossing, Pedestrian phase at traffic signal junction, No physical crossing within 50 metres, Central refuge, non junction pedestrian crossing, footbridge or subway)
    - Light_conditions(Daylight: Street lights Present, Darkness: Street lights Present and lit, Darkness: Street lights Present but unlit, Darkness: No street lightening, Darkness: Street lightening unknown)
    - Weather_Conditions(Raining without high winds, Raining with high winds, Fine without high winds, Fine with high winds, Snowing without high winds, Snowing with high winds, fog or mist)

    - Road_Surface_Conditions(Wet/Damp, Dry, Frost/Ice, Snow, Flood)

- We used feature_importance to find the importance of each attribute and how much it contributes in the prediction of Accident severity. We dropped few attributes as they don't seem to contribute to problem statement in hand like the name of the nearest highway authority, if the police attended the scene, English Coordinates etc. The attributes which we dropped were :

    - Police_Force
    - Junction_Control
    - Junction_Detail
    - Did_Police_Officer_Attend_Scene_of_Accident
    - LSOA_of_Accident_Location
    - Carriageway_Hazards
    - Special_Conditions_at_Site
    - Location_Easting_OSGR
    - Location_Easting_OSGR
    - Local_Authority_(District)
    - Local_Authority_(Highway)

Visualization

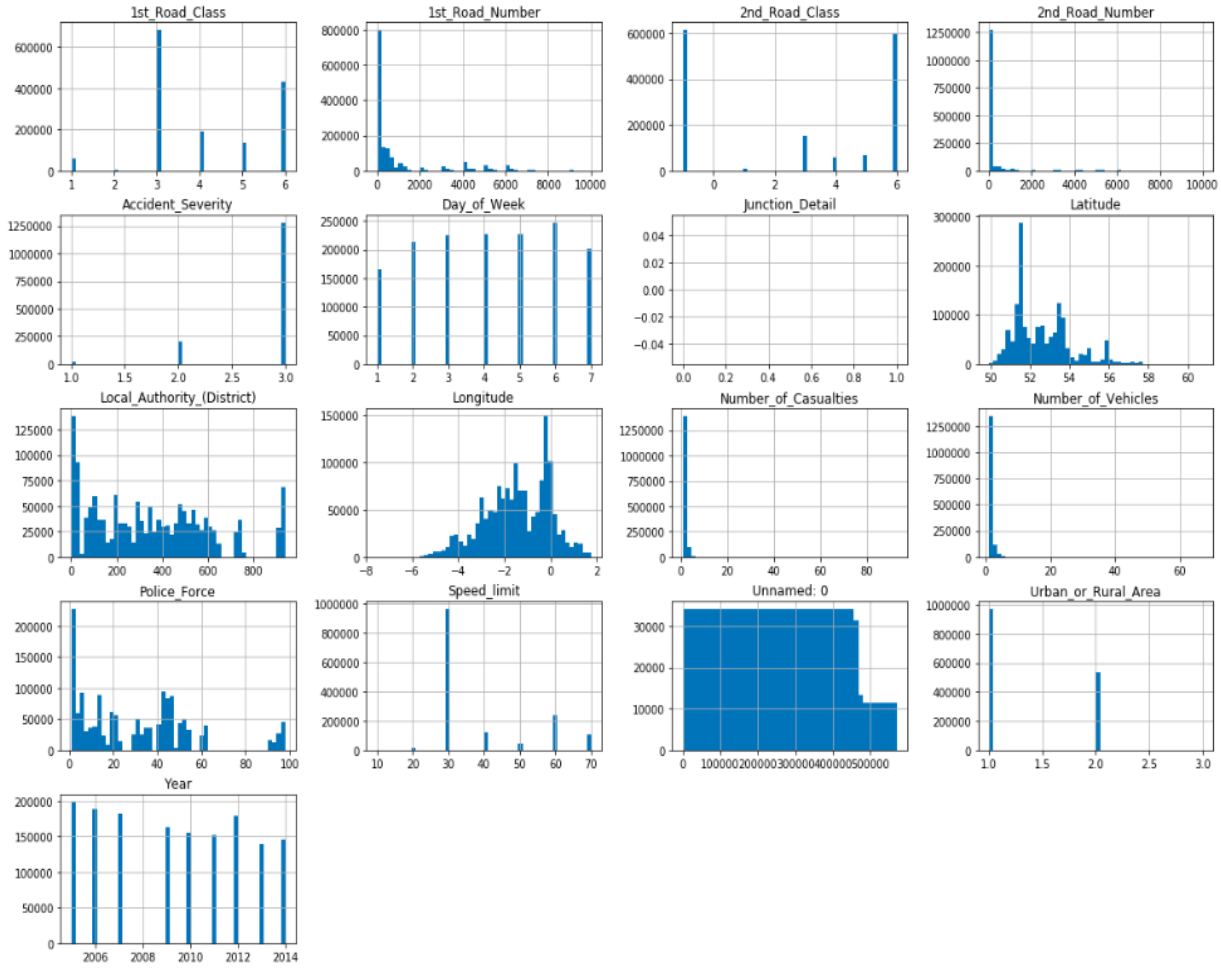
- We observed a decreasing trend in the number of casualties from 2005 to 2014.

	Year	Number_of_Casualties
0	2005	271017
1	2006	258404
2	2007	247780
3	2009	222146
4	2010	208648
5	2011	203950
6	2012	241954
7	2013	183670
8	2014	194477

- We observed that most of the accidents take place in the weekends.

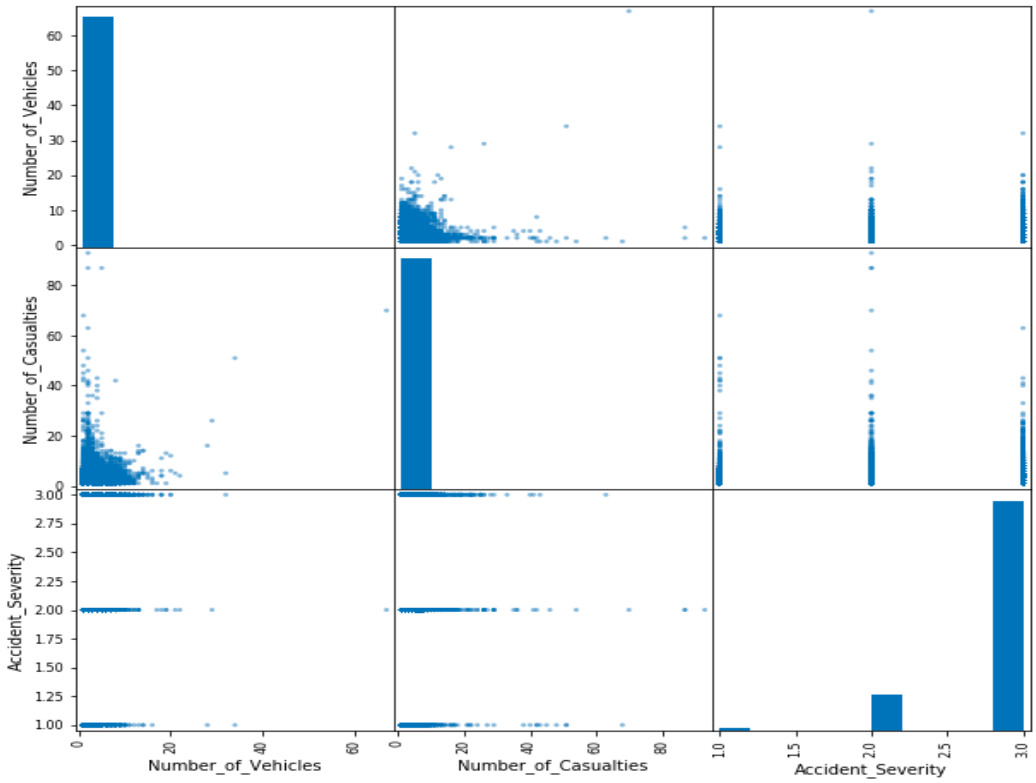
	Day_of_Week	Number_of_Casualties
		sum
0	6	331934
1	5	299044
2	4	297756
3	3	294476
4	7	285261
5	2	284043
6	1	239532

- Data distribution over the whole dataset



- Severity of the accidents on the entire UK map based on the latitude and the longitude given:

- Scatter Matrix of Accident Severity, Number of Vehicles and Number of Casualities:



Preprocessing

- Dropping the columns in our dataset which aren't useful to our model using data.drop()
- Dropping all the rows with the null values in each columns using data.dropna()
- In the Time column, using minutes would not make any effect. Therefore, we have dropped the minutes using slicing and are using just the hours part of the column.
- For the columns which have 5-6 unique values for the whole dataset, we have used label encoding as to convert the string columns into integers. We have XGBoost model which does not support object data types, therefore label encoding helped us converting it into integers.
- Since, Accident_Severity has only 3 possible values, therefore we have used one hot encoding and made 3 new classes Fatal, Severe and Slight as per the need of out model training.

Model

In the end of preprocessing we were left with 13 attributes with which we trained our model. As the data was tabular and didn't need any vectorizers/ NLP for that hence we used traditional tree models for the problem.

Just for experimenting we tried the Naive Bayes models (Gaussian, Multinomial, Complement and Bernoulli) along with Logistic Regression, SGDClassifier, AdaBoost and RandomForestClassifier and combined many of these models using VotingClassifier, Bagging and Stacking methods in various combinations which resulted in 67% accuracy at best.

As we have already mentioned that our data was biased over Slightly severe accidents, we used imblearn library to use undersampling / oversampling classifiers like BalancedRandomForest, RUSBoostClassifier etc. and out of which the best results were given by EasyEnsembingClassifier with the base estimator model being XGBoost whose parameters were derived from GridSearch and n estimators being 12.

Training

We ran our model over the dataset with a train: test of 1:4. We used cross validation with cv = 5.

Conclusion

We were able to predict the severity of these accidents with the accuracy of 70.3% over our test data which consisted of 25% of the entire dataset (375, 455 entries).