Accident Severity Precition

Capston project - IBM Data Science Professional Certificate

Sergio Quesada Dorigné
October 2020

Introduction

Unfortunately, car accidents happen.

Would it be possible to understand/predict the severity of an eventual impact accident for a car driver if ever happening?

Such a predictive model could be interesting for instance:

- *For Insurance companies* to set a final price for products based on customer/drivers characteristics.
- *For government authorities* understand which parameters cause fatal accidents, and to forecast casualties that may occur during a year based on vehicles registered and drivers records and plan measures to reduce those casualties numbers.
- *To generate awareness in individuals* so that they are aware of the need of replacing their car, be even more careful when it rains or when it's a foggy day, etc.

Dataset information, definition and understanding

We will be using the database published on Kaggle.com called "UK Accidents 10 years history with many variables" that collects accidents that took place from 2005-2014 in UK roads. Data are stored in 3 tables: accidents, vehicles and casualties and can be found in the following link:

https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables

- *Accidents file*: main data set contains information about accident severity, weather, location, date, hour, day of week, road type, etc.
- *Vehicles file*: contains information about vehicle type, vehicle model, engine size, driver sex, driver age, car age, etc.
- *Casualties file*: contains information about casualty severity, age, sex social class, casualty type, pedestrian or car passenger, etc.

Dataset information, definition and understanding

Tables have been joined in order to build a larger dataset using the key parameter Accident_Index of the accident that is common to all tables.

Then, most likely variables to be useful a priori for building the model have been conserved.

Table 2. Preselected variables.

Variable Name	Table of origin	Initial data type	Definition
Age_Band_of_Driver	Vehicles	Categorical	The age of the driver is stored in ranges (i.e. 0-10, 10-20, etc.).
Sex_of_Driver	Vehicles	Categorical	Sex of the driver (i.e. male of female).
Age_of_Vehicle	Vehicles	Integer	Age of the vehicle (i.e. 5, 9, 14, etc.)
Light_Conditions	Accidents	Categorical	Was there daylight or not?
Road_Surface_Conditions	Accidents	Categorical	Dry, wet, snow, ice on road?
Urban_or_Rural_Area	Accidents	Categorical	Where did the accident takes place (i.e. city, small city or rural)?
Date	Accidents (from Date variable)	Categorical	Date of the accident.
Casualty class	Casualties	Categorical	Is it the driver, a passenger or a pedestrian?
1st_Point_of_Impact	Vehicles	Categorical	Where did the impact took place (i.e. no impact, front impact, etc.).
Speed limit	Accidents	Integer	Speed limit of the street/highway in which the accident took place (30, 50, etc.).
Casualty Severity	Casualties	Categorical	Severity of the accident based on casualties table information (i.e. fatal, severe, slight).
Day of week	Accidents	Categorical	If it's happening on Monday, Tuesday, etc.
Vehicle Type	Vehicles	Categorical	If it's a car, a bus,a bike, etc.
Accident Severity	Accidents	Categorical	Severity of the accident based on accidents table (i.e. fatal, severe, slight).

Methodology

1. The tables *Accidents*, *Vehicles* and *Casualties* are joined using *Accident_Index* variable common to all tables. Afterwards we can drop *Accident_Index* column. Moreover, multiple variables are removed because not apparently helping to solve the problem. Dataset contains now (4287593, 16)

2. We analyze missing data, note that some variables have it signaled as -1 and are transformed into *NaN* values previously.

```
Accident_Index                0
Vehicle_Type                554
1st_Point_of_Impact        2418
Sex_of_Driver                46
Age_of_Driver            405664
Age_Band_of_Driver       405664
Age_of_Vehicle          1155488
Casualty_Class                0
Casualty_Severity             0
Accident_Severity             0
Date                          0
Day_of_Week                   0
Speed_limit                   0
Light_Conditions              0
Road_Surface_Conditions    4824
Urban_or_Rural_Area           0
dtype: int64
```

As a consequence, we remove the columns *Age_of_Driver*,*Age_Band_of_Driver* and *Age_of_Vehicle*. We also drop the *Accident_Index* column and duplicates that would existe. Afterwards we remove all the rows that would have NaN values. We have the following dataset (4279897, 12).

3. From *Casualty_Class* we keep values “1” that refer to drivers as casualty and from *Vehicle_Type* we select only cars whose value is ‘1’ and we get a dataset containing (2158083, 10).

4.For integrity of data we checked that *Casualty_Severity* and *Accident_Severity* not always have the same values. We will keep only the events that have the same value. We now have a dataset of (2015253, 9).

Methodology

5. We remove events that have no value for *Sex_of_Driver*, getting now dataset of (1918837, 9).

6. We recategorize *Light_Conditions* in a way that we only differentiate between Light (1) and darkness (0) conditions. Dataset remains (1918837, 9).

7. For Road_Surface_Conditions we will recategorize in a way that we only differentiate between dry (1) and wet/ice/snow (2). Dataset remains (1918837, 9).

8. For *Urban_or_Rural_Area* we recategorize so that we only differentiate between big cities (0) and small cities/rural area (1). Dataset remains (1918837, 9).

9. For *1st_Point_of_Impact* we will only consider situations in which there is indeed an impact therefore we remove events having value '0'. Dataset is now (1859852, 9).

10. Concerning *Speed_limit* we differentiate between accidents with speed limit equal or below 30 (0) and above 30 (1). Dataset is now (1859852, 9).

11. Concerning the *Date* we extract the Month, but we don't really see an impact of it. Dataset is now (1859852, 9).

12. We finally differentiate *Day_of_Week* between weekend (Friday, Saturday and Sunday = 0) and the rest of the week (Monday, Tuesday, Wednesday and Thursday = 1). Dataset is now (1859852, 9).

13. We binarize Accident_Severity and we will have 'Fatal' = 1 and 'Severe/Slight' = 0.

We save the obtained dataset into *'carsForModeling.csv'*.

Methodology

Table 3. Correlation of variables with *Accident_Severity*.

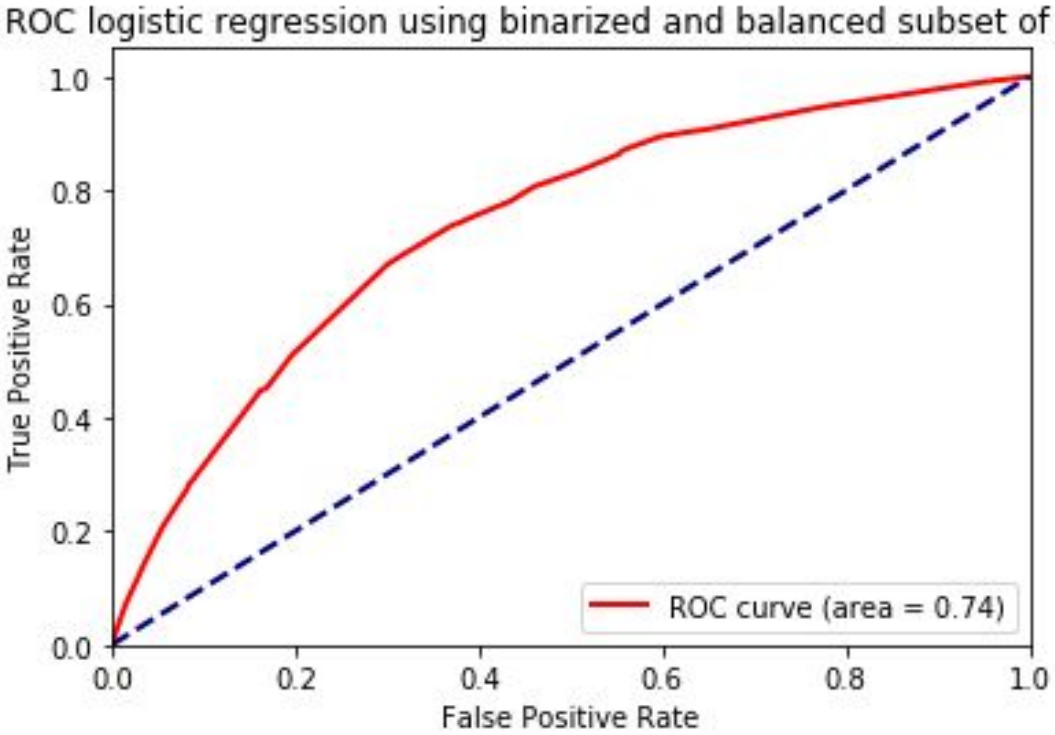
	Sex_of_Driver	Light_Conditions	Road_Surface_Conditions	Urban_or_Rural_Area	Speed_limit	Day_of_Week	Accident_Severity
Sex_of_Driver	1.000000	0.082586	0.003862	-0.006513	-0.015494	0.043430	0.042699
Light_Conditions	0.082586	1.000000	-0.185675	0.018758	0.000178	0.038032	0.045380
Road_Surface_Conditions	0.003862	-0.185675	1.000000	0.103964	0.100384	0.015135	0.004397
Urban_or_Rural_Area	-0.006513	0.018758	0.103964	1.000000	0.639037	-0.004701	-0.065206
Speed_limit	-0.015494	0.000178	0.100384	0.639037	1.000000	0.001049	-0.053038
Day_of_Week	0.043430	0.038032	0.015135	-0.004701	0.001049	1.000000	0.021133
Accident_Severity	0.042699	0.045380	0.004397	-0.065206	-0.053038	0.021133	1.000000

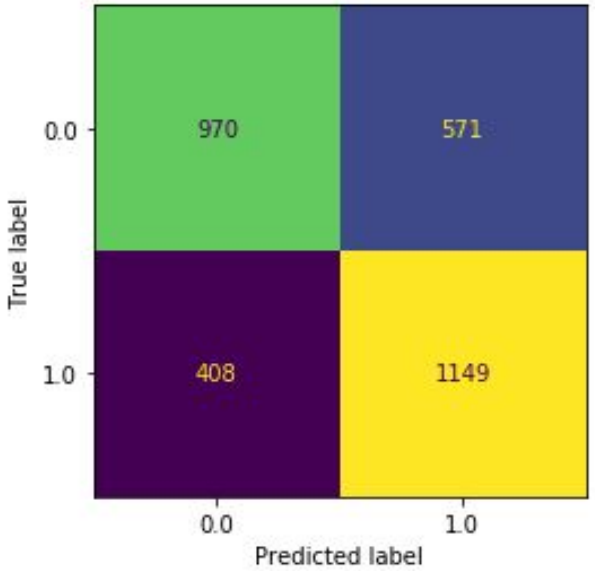
Following Machine Learning techniques to address this binary classification problem and present its metrics:

- Binary classification using Logistic regression
- Decision tree

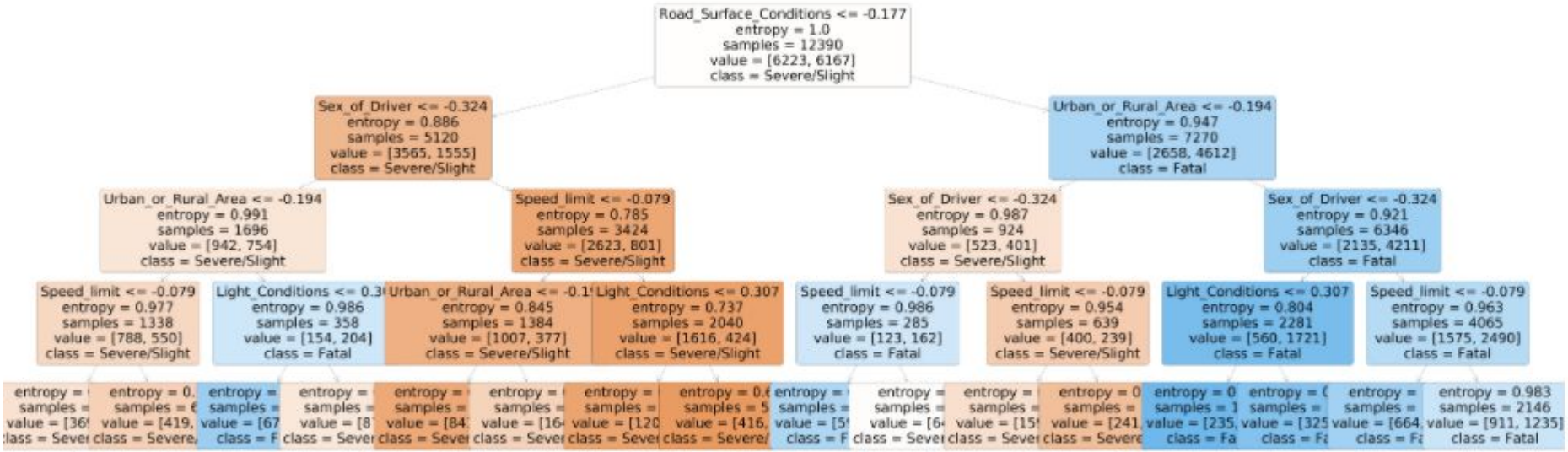
NOTE: dataset is unbalanced, meaning we have much more events that are Accident_Severity = ‘Severe/Slight’ that Accident_Severity = ‘Fatal’, this last one is a limitant. Therefore since now in our dataset we have 7744 ‘Fatal’ events we will select a random subset of data of 7500 events having Accident_Severity = ‘Severe/Slight’ and Accident_Severity = ‘Fatal’.

Binary Logistic Regression





Decision tree



Results

Model	Logistic regression	Decision tree
F1 score	0.672	0.670
Accuracy	0.670	0.670
ROC AUC	0.731	

Discussion

General observations:

- setting a few parameters (accidents for cars, driver as casualty and accidents having an impact), and using some easy to access variables allow to assess with acceptable degree of certainty when a car driver is about to start his journey, if ever having an accident it would be ‘fatal’ or ‘severe/slight’.

- even if those models have been developed using UK data, where traffic happens from the left side, it is very likely that the important variables remain the same for countries where traffic happens from the right side, and therefore those models could be also applicable.

To improve the models

- Have more events containing the age of the driver, so that we don’t have to drop *‘Age_of_Driver’* or *‘Age_Band_of_Driver’*.

- Have a classification of the car concerning its size (i.e. small, medium or big car).

- We have an unbalanced dataset, meaning we have more events Accident_Severity = ‘Severe/Slight’ rather that Accident_Severity = ‘fatal’. Therefore, the metrics of the models are (‘unfortunately’) likely to be improved if ever having more ‘fatal’ events with all required information.

Conclusions

Authorities, car companies and assurance companies could work together to reduce fatality of the accidents by acting on three major axes:

- improve visibility of the roads either because of the weather (fog, rain, storm, snow, etc.) or because it’s night.
- improve the roads surface or car wheels surface and/or design so that when the road is wet or has ice or snow, this doesn’t influence a fatal accident.
- improve roads in small cities/villages and rural roads.

Accident Severity Precition

Capston project - IBM Data Science Professional Certificate

Sergio Quesada Dorigné
October 2020