Severity of car accidents

Andrés Felipe Vargas Quintero

23/29/2020

Content

1. INTRODUCTION

Traffic accidents, little instants in which the life can change, and in some occasions, end; year by year thousands of lives are lost due this terrible incidents which can be caused for the most insignificant reasons, as animals on the way or a distraction while driving.

There is not a unique reason for the traffic accidents, the causes are too diverse and there are so many factors intervening at the same time that every case needs to be evaluated in detail in order to find all the hidden answers, answers which can be founded on the driver, the speed, the weather, etc.

On this project we'll find the correlation between the specific factors and the severity of the accidents, understanding severity as the amount of human lives affected on the incident, for this we'll need a dataset in which all the different conditions present on the moment of the incident are signed.

2. DATA

For this project we will use the dataset *"Accidents de trànsit amb morts o ferits greus a Catalunya"* wich reports all the traffic incidets reported on Catalonya on the period 2010-2020 with the details about weather conditions, road conditions and number of casualties, the data set can be found on the link "https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb".

This dataset has all the information about the traffic accident occurred in Catalonya between 2010 and 2020, this way it is possible to evaluate if there is interference of climatic, conditional or time specific situations that can affect on the total amount of victims that the accident leaves.

Some of the column of the dataset have the same information but expressed on different ways, as the case of the columns referred to the presence of haze on the moment of the accident, or the columns referred to the illumination on the way, also there are columns which information is not relevant to determine the severity of the accident as the columns referred to the exact city or town in which the accident occurred.

3. METHODOLOGY

The methodology used on this project will be based on a multiple linear regression model, the main target that is the total number of victims on the accident (F_VICTIMS on the dataset), this column has the total amount of victims affected on the accident, counting deaths and injuries.

On the categorical columns, the missing data is signed as “sense especificar”, so this expression will be kept as missing data (np.nan) at the beginning of the analysis in order to avoid misunderstandings on the moment of analyzing each one of the variables.

The columns will be analyzed using scatter plots, this way we’ll be able to see which are the truly relevant variables on the moment of predict the severity of the car accidents, for this we’ll use the seaborn library, plotting regplott() for numerical columns, and catplot() for categorical variables.

Using ANOVA analysis will be determined if the categorical variables are directly correlated to the target or not, the unit columns, in which is signed the amount of different kind of vehicles involved on the accident will be analyzed using a multiple linear regression model.

With the visual analysis we could notice that variables with a better accuracy on ANOVA are actually the variables which less affect the target, the reason for this is because there is no significant different between the different values of the variable and their relationship with the target.

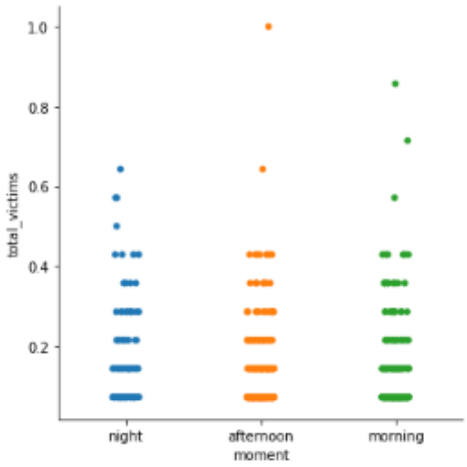

Fig. 1

On the image (fig.1) we can see the plot of the variable “moment”, this variable has a good performance of ANOVA with an f value of 5.65 and p value 0.003, but checking on the scatterplot it is evident that the essence of the variable is the same with respect to the target, having only some noise on the high amounts of the target.

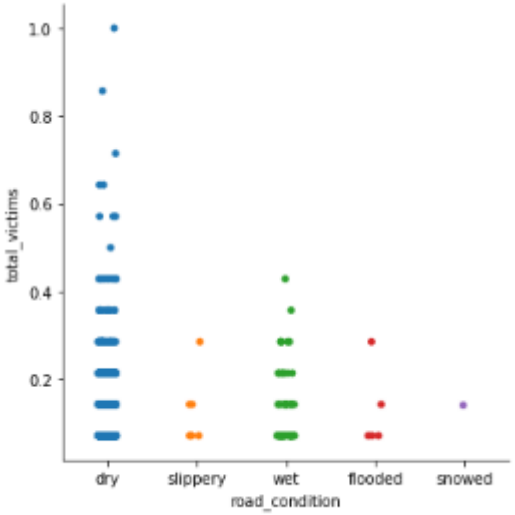

Fig. 2

On the other hand is the variable “road conditions” (fig. 2), tis variable has not so good ANOVA performance, the f value is 0.021 and the p value 0.99, but it is noticeable that the “dry” value is statistically more affective to the target that the “flooded” or “snowed” values.

Making the scatterplot analysis we could also notice that the numerical variables do not have too much linearity and the majority of accidents with units different to regular family cars, do not have too much severity, or, at least; there are not many of this vehicles on the accident.

Also the numerical values are analyzed on a visual way, in this case making a regplot from the seaborn library, with this analysis it was determined that not all the numerical values can be evaluated using a linear regression model.

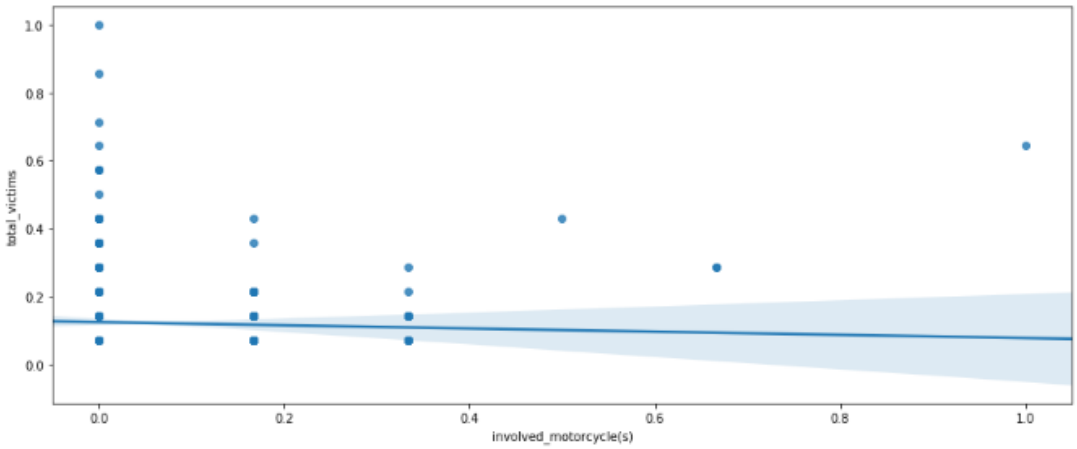

Fig. 3

This regression plot (Fig. 3) displays the variable “involved_motorcycle(s)”, on this case there is a linear model displaying noise on the high values of the variable x.

After the analysis we determined which of the different predictors present on the dataset are actually a relevant factor to determine how many people will be affected on a traffic accident; to evaluate the model will be used the Mean Squared Error (MSE) metric.

4. Results

- Weather conditions and road status, the most relevant factors at the moment to predict the severity of the accident.
- Amount and type of involved units not too relevant.
- Hour in which the accident occurs does not affect the severity of the accident.

5. Discussion

At the beginning the main idea about the dataset was that the most relevant part of the data consisted on the amount and type of involved units, weather and time conditions were considered also as relevant but not as much as the units, and the status of the road was not really important on the first approaches.

After developing the model and making the respective evaluation, was determined that the most relevant variables are the weather and the conditions of the road, variables as “haze” that provide a huge difference between the values affect in a significant way the efficiency of the developed model.

Units involved do not affect too much the result of the model, the amount of victims on each accident is just significantly affected by the presence of family or personal cars, which are involved on the major part of the accidents with victims.

Time conditions are not relevant on the dataset, the data proves that it is not different to have an accident on the morning or on the night, and the type of day is not either a useful predictor on the moment to develop a predictor model.

6. Conclusions

- The most relevant variables in order to predict the severity of a traffic accident are the ones related to the weather and the status of the road.
- The amount and type of units involved on the accident are not soo correlated as they seem to be.
- Good climatic conditions present more severe traffic accidents.
- There is no big difference on the accidents occured on work days and the ones occured on weekends.