Predicting the severity of an accident

Villa Fernando

October 2, 2020

1. Introduction

One of a common issue in the modern society is the car accident, product of the traffic in cities. Weather conditions, special events, traffic jams, the day of week and other factors could be explained this problem and with that is possible to predict the severity of the accidents can be performed.

This problem is relevant for the governments, because they can act faster if know the impact of the accident and with that reduce the mortal accidents. For that reason, is necessary to explain and discover more knowledge about this topic.

2. Data

For explain and resolve the prediction about severity of an accident I use the set "Data-Collisions". This set contains features that I describe in the Introduction. I considered other features like:

Data Users:

- Severity Code
- PERSONCOUNT
- VEHCOUNT
- Bicycles: is a new variable about if there are bicycles in the collision (values 1,0)

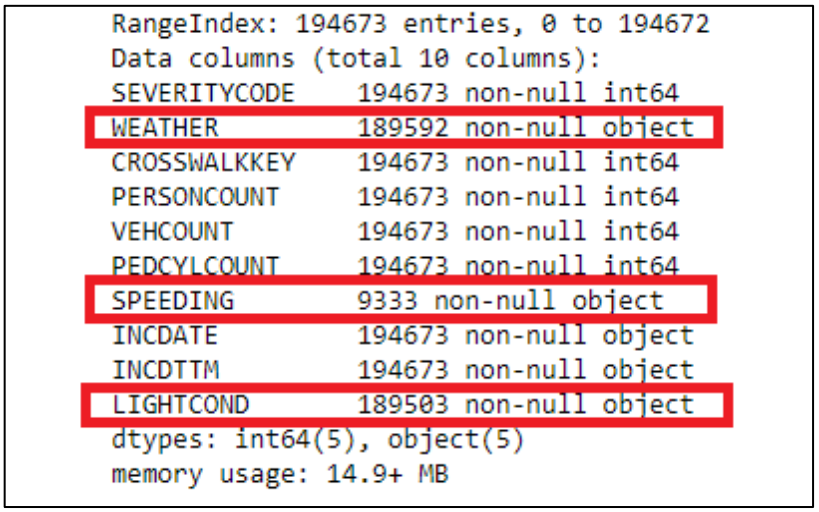
Other Features:

- TYPEWEATHER: is a new category and i use 3 class. Good weather(0)-- Clear,PartyCloudy , Bad wather(1)-- all of bad weeather,and others(2)-- Other and Unknown.
- SPEEDING
- INCDATE
- INCDTTM
- WEEKDAYTYPE : Values 0-- about 0 to 3 dayof week,1--- 4 day of week,2--- 5,6 day of week.
- LIGHTTYPE: is a new category and i use 3 class. Dark(0)- is a dusk,all types of dark without artifficial lights,Clear(1)- dawn and dark with street lights, Other types and unknown(2)
- CROSS: If is the collision in a cross walk or not (1 yes, 0 no).This feature obtain if the value of CROSSWALK KEY=0 then COS=0 , otherwise CROSS=1.

3. Methodology

3.1 Cleaning Data

Understanding the data and only chose the feature to mention in the DATA section, I discovered some features that have some missing value.



For that reason, is needing to replace the missing values in the features in red and create new variables.

a) Weather:

- Replace missing values with the most common value, in this case “Clear”
- Create the variable “TYPEWEATHER” that have values in 3 different categories: 0 (“Clear” and “Party Cloudy” weather), 1 (“Bad Weathers), 2 (“Other” and “Unknown” Weather)

b) Speeding:

- Replace missing values with the opposite unique value. The only values that is not null is "yes" for that reason I assume that values "null" are "not".
- Replace values yes to 1 and not to 0.

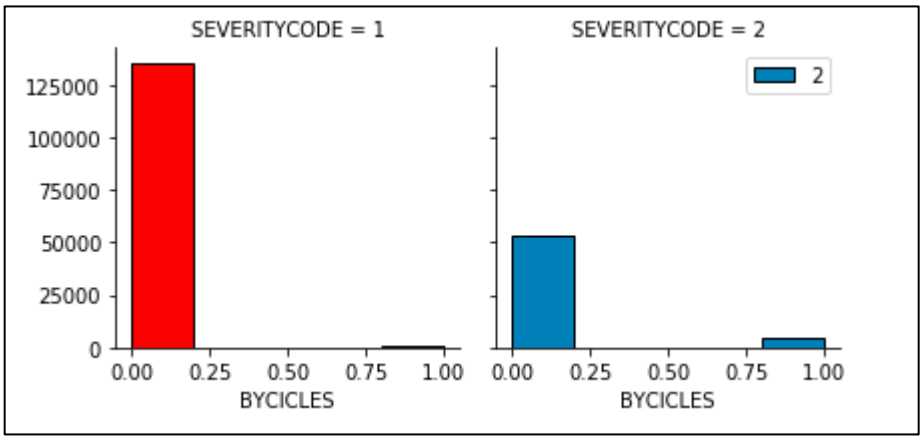
c) Lightcond:

- Replace missing values with the most common value, in this case “Daylight”
- Create the variable “LIGHTTYPE” that have values in 3 different categories: 0 (Dark condition), 1 (with Light condition), 2 (“Other” and “Unknown”)

3.2 New Features and Analysis

a) Bycicles:

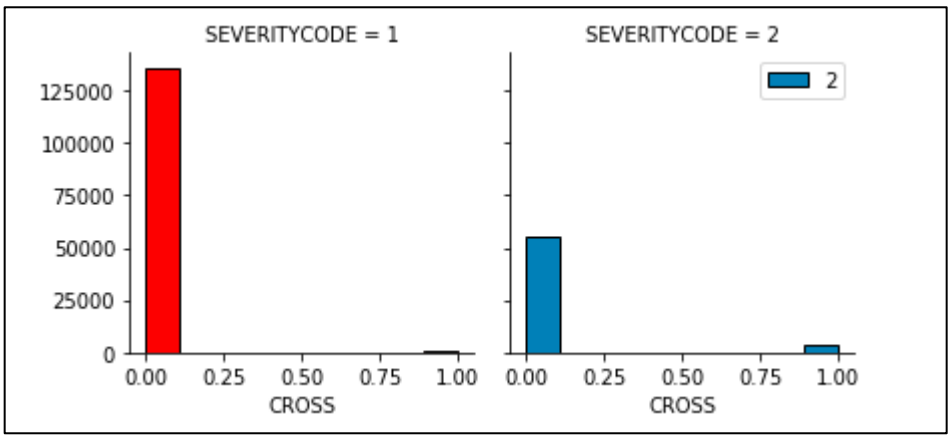
- Create the variable “BYCICLES” that have values 2 different values, if value of “PEDCYLCOUNT” is more than 0 then BYCICLES is 1, else is 0.



Is more probably that in the accident don’t find a person in Bycicles. But if there are is more probably that have a several accident.

b) Cross:

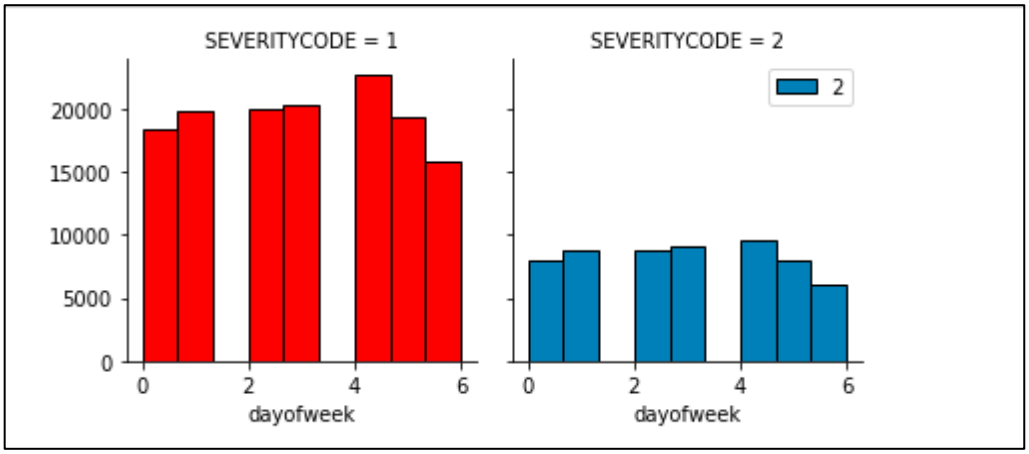
- Create the variable “CROSS” that have values 2 different values, if value of “CROSSWALKKEY” is more than 0 then CROSS is 1, else is 0.



Is more probably that in the accident does not in a cross way. But if there are is more probably that have a several accidents.

c) Weekdaytype

- Create the feature “dayofweek” about the feature “INCDATE” that have values of 1 to 7.



- About the distribution, first 3 days of week we have a similar tendency in two types of severitycode.
- In the 4 day the value increase and 5 and 6 day dicrease
- For that reason I need to créate the variable “WEEKDAYTYPE” that have 3 categories: **0**-- about 0 to 3 dayof week,**1**--- 4 day of week, **2**--- 5,6 day of week.

3.3 Correlation:

The correlation analys is a good method to discover what feature have more affinity with the tarjet (severitycode). I have these results:

SEVERITYCODE	1.000000
BYCICLES	0.214702
PEDCYLCOUNT	0.214218
CROSS	0.182314
CROSSWALKKEY	0.175093
PERSONCOUNT	0.130949
SPEEDING	0.038938
dayofweek	-0.015246
WEEKDAYTYPE	-0.017153
VEHCOUNT	-0.054686
TYPEWEATHER	-0.104996
LIGHTTYPE	-0.119548

The features in green are the variables that I use in the predictive model, but is not needed CROSSWALKKEY because the variable CROSS is better and create from CROSSWALKKEY.

4. Predictive Modeling

There are two types of models, regression and classification, but in this problem, I prefer use a Classification Models and Logistic Regression because the target is binary.
Before training the models, I use a 20% to the data to test the value, and 80% to train the model.

The different models that I train are these:

- KNN NEIGHBOARD: with 6 clusters
- DECISION TREE
- SVC
- LOGISTICS REGRESSION

4.1 Performances of different models

Using the method of F1-Score and Jaccard (Accuracy) I have these results:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.73	0.68	NA
Decision Tree	0.74	0.66	NA
SVM	0.74	0.67	NA
LogisticRegression	0.74	0.67	0.56

About the results, I choose the KNN algorithm because have a better F1-score even if the Jaccard result is a little less than the others. F1- Score is a better to evaluate than the Jaccard method.

5. Conclusions

In summary, it was found that the KNN predictive model is more suitable for predicting the severity of an accident. Additionally, the model considers only 5 variables, so it makes prediction more efficient (since too much data should not be collected) and would allow the authorities to act more quickly.

6. Future directions

In the case study, an accuracy of 73% was obtained, which for a predictive model is a good indicator, however there are still 27% of precision to be explained, which would merit the evaluation of new variables or the use of more sophisticated models such as networks. neuronal. This would be a better version for future work.