Classification of severity for new traffic events

Leonardo Vela

Applied Data Science Capstone

IBM Data Science Professional Certificate

September 2020

Business understanding

Every person that needs to move from a place to another one, develop a mental route to follow in order to reach the target place, however there is a lot of possible choices to take, and many considerations to keep in mind. It could seem like trivial choose the shortest route, however there are other issues to consider, such as how long this route can take, road conditions, hour, traffic and not less important, is this a safe route? All these issues are considering whether for people that want to move but also for people in charge of the road's maintenance, emergency teams, police and government entities. Every year the government entities assign a budget for their different issues in charge, including the emergency services. For these reasons is relevant knowing more about the behaviour of accident on the roads, in order to define when could be useful having more emergency teams and equipment. Therefore, we will study the car accident severity data which will provide us information about accidents reported and their relevant variables that could make us understand better what conditions are more likely to trigger an accident, as well as when the events could happen, giving information to government entities about when and how much resources should to assign for emergency services.

Data Understanding

The data provided to make this classification contains all records of traffic collisions of Seattle city since 2004 to present. This dataset is updated weekly and includes all types of collisions, including different vehicles, cyclists and pedestrians.

The dataset contains 194673 records and 38 attributes. This dataset contains several variables present in every accident, however all of them were not suitable for our work, therefore the model construction used a subset.

In a first approach the following attributes were taking in account as contributors to make good estimations:

INCKEY: As ID event.

INCDTTM: The date and hour are very important to know when could happen a new emergency.

INATTENTIONIND: This data is also important to determine in which zones the involved not took care of the situation, perhaps so much distraction on the zone, noises, lights, advertising.

WEATHER: The weather can trigger not suitable conditions on the roads, like poor visual and grip.

ROADCOND: Good conditions make safer the roads.

LIGHTCOND: Good conditions of light always will be better to driving safety.

HITPARKEDCAR: This will allow identify if the zones prone to have many parked cars have high likely to be involved in an accident.

SEVERITYCODE: As our label or data to be estimate.

As above mentioned, the severity code was the target, it has the following domain:

-3. Fatality.

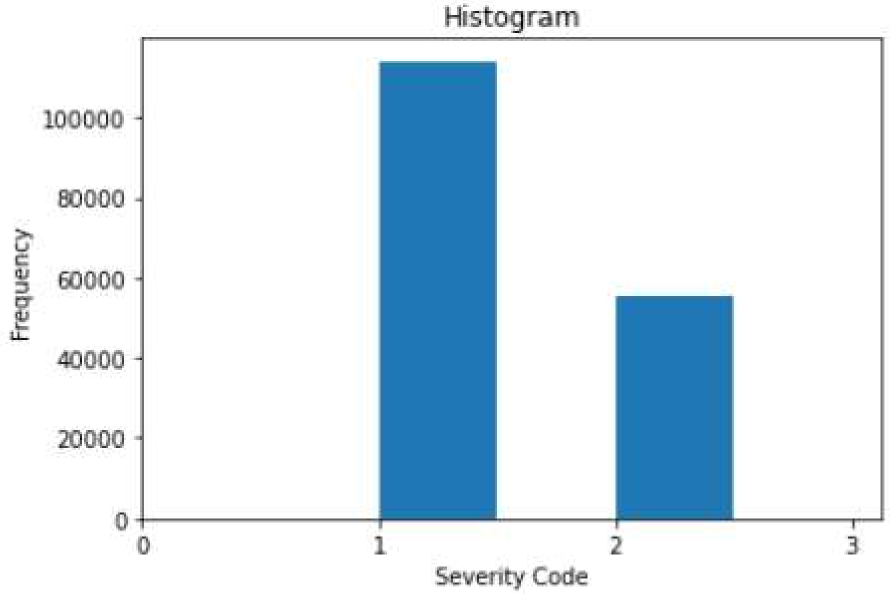
-2b. Serious injury.

-2. Injury.

-1. Prop damage.

-0. Unknown.

However, the image below shows a histogram with just two of those classes. This suggests that the data is unbalanced, therefore this must be balanced to avoid biases in the outcomes.



Data preparation

The data preparation includes deal with no data, assign correct formats, transform categorical variables into numerical variables and extract data from the raw data.

Following this lineament, no data was found in the attributes, WEATHER, ROADCOND, LIGHTCOND, X, Y, ADDRTYPE, including data values like unknown, 0, or NaN, which were dropped

Format correctly some attributes also was required, is the case of UNDERINFL which is an attribute that indicates if the accident was influenced by alcohol. This attribute contains several values that indicates the same such as 0 and N to indicates no alcohol influence and 1 and Y to indicate alcohol influence, therefore a formatting was applied getting just 1 or 0 values. Furthermore, in order to get a standard dataframe, other attributes were formatted with the same values, such as INATTENTIONIND, PEDROWNOTGRNT, SPEEDING and HITPARKEDCAR.

Then, categorical variables like WEATHER, ROADCOND, LIGHTCOND and ADDRTYPE were converted into numerical variables by means of dummy variables.

Finally, in order to estimate if an accident occurs in the time range with highest traffic, the hour was extracted from date making a new attribute called high_traffic_hours.

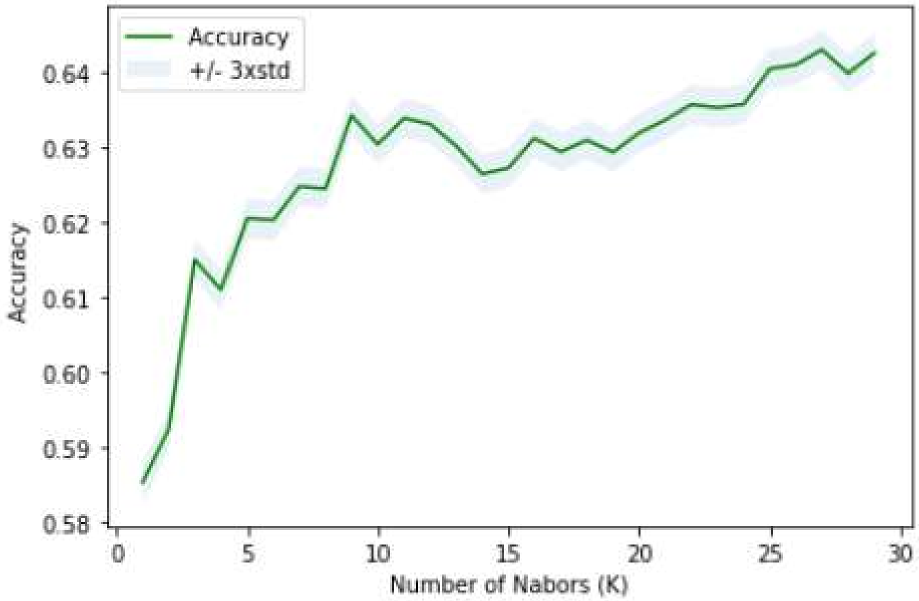
Once the data was correctly prepared just left one step more, balance the data. As was mentioned previously, the data has more than double in one severitycode than in the other, this can create biased outcomes, therefore an under-sampling technique was applied getting 55707 rows per each classification.

Modelling

There are several machine learning algorithms that help to estimate data, but what algorithm we need to use depend on what kind of outcome we expect. As we need obtain a code classifier that indicates if a new accident can belong of one the categories, the models that help to classify like K-nearest, decision tree and support vector machine were applied, then compared their accuracy's metric and select one of them as our main estimator.

In order to consider their accuracy's metric, it was required split the dataframe in train and test data, this allows to evaluate the models using the test data with 30% of the whole dataframe to compare the original class and the predicted class.

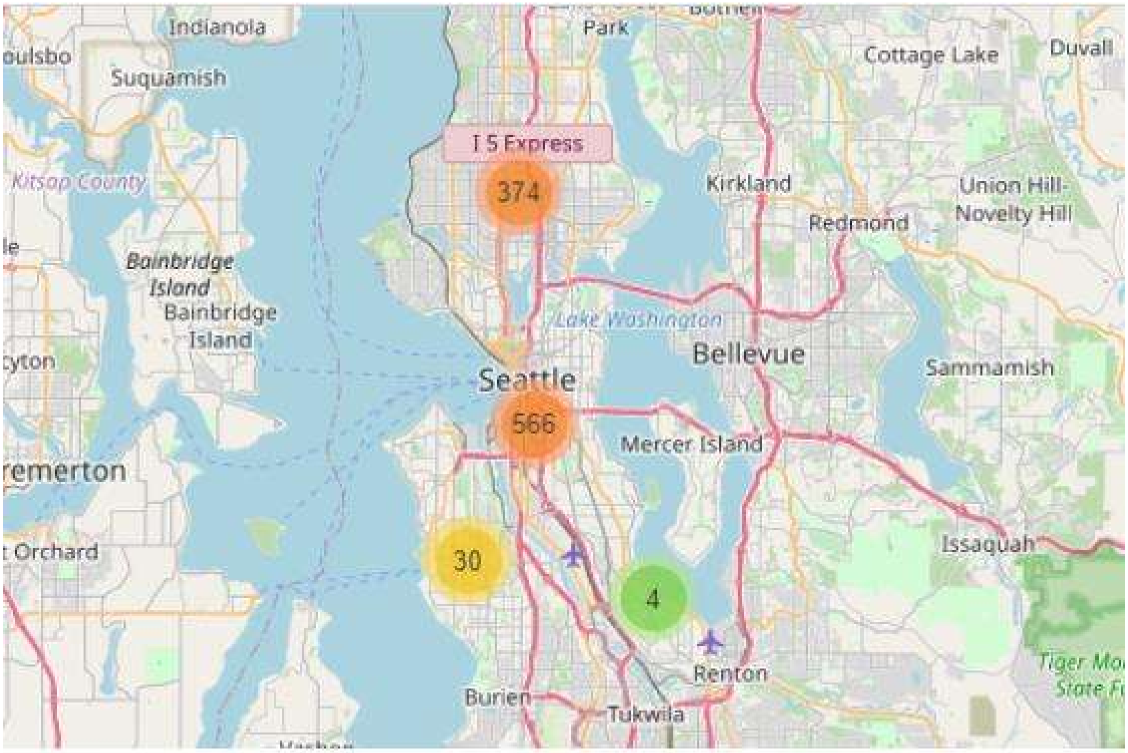
The first model used was k-nearest and to determine what k number was more suitable, an iterative process was executed testing k in a range from 1 to 30. The best accuracy was 0.6429 and it was shown with k = 27.



The second model applied was decision tree, using the entropy criterion an accuracy of 0.6413 was found. Finally, the support vector machine was applied with the kernel radial basis function (rbf), getting and accuracy of 0.6474 which was the highest accuracy found among the three models.

Results and discussion

The domain of severity code is composed by 5 codes, however the dataset just present 2 of theses codes, showing a no balanced dataset, with a ratio of 2:1, in other words every 2 cases of prop damage there is one of injury. Also, by means a sample, a map shows the downtown as the zone with more accidents.



In order to compare among different models the accuracy to predict the severity for a new event, three models were applied, k-nearest, decision tree and support vector machine, there was not a big difference among their results, their accuracy was 0.64. It is reasonable, if we consider that the correlation among variables is too weak, the highest correlation respect to the target was 0.215321 corresponding to the pedestrian involved in the accident, and the lowest value was -0.000375 and correspond to dawn weather condition. Finally, the emergency service can expect be ready to attend an emergency that involves injury at the downtown and the surrounding areas to green lake, so it could help to stablish new emergency points to treat injuries and check the routes to hospitals and medical centres.

Conclusions

Thanks to computer tools is possible to processing thousands of entries in a few minutes, this allows people to be more creative and analytic. However, this kind of computer tools work well if the data is correctly prepared, and this task can take much time because this implies understand very well the problem to solve and what mean the data to do it.

Furthermore, this is an iterative process, perhaps make the process once again or more times if is required can help to understand or highlight insights in data that can improve the models.

As the data juts have 2 codes regarding to the 5 classificatory codes, it is not possible to predict a code different of these codes.