IBM Data Science professional certificate

Capstone project, Sylvian Munier, august 2020

Introduction

The aim of this project is to try to predict the severity of an accident with respect to different variables found in the statistics report of the city of Seattle.

The purpose is not an absolute model, but to give indications where improvement can be made.

The audience is the security officer in charge of the road security.

As we are looking at a model of the reality, the first step is to find what values can make a good predictor and delete all the rows containing no tangible information.

Data – pre-analysis

In a first approach, I tried to look upon the impact of the weather alone. I realized then that the majority of accident were happening in dry condition. Then I tried to find if, with regard of the number of rainy days, I can equilibrate the occurrence, but I found that half of the year, there is rain.

So, I could not find a way to show the increase of the number of accident and their severity due to rain or snow.

I checked if I can find a pattern concerning the number of pedestrians impacted, but again, nothing really conclusive.

But I found some interesting things (see the ipython notebook for more information named pre-analysis on my github : https://github.com/SylvianMunier/IBM-Data-Science-capstone-project).

But first, take a look on the number of cases with regard to the road condition:

Dry	124510
Wet	47474
Unknown	15078
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

We see that around 2/3 of the accident happens on dry condition.

Note that, I did not keep all the variables to make my model, because we can't predict, for example, the oil spilt on the road.

When I tried to find if more pedestrians are impacted or if the severity of the accident increase with the weather, I found this:

```
There is 47.44 of severity 2 accident in dry conditions for 100 severity 1 accident
There is 49.63 of severity 2 accident in wet conditions for 100 severity 1 accident
There is 29.17 of severity 2 accident in ice conditions for 100 severity 1 accident
There is 19.95 of severity 2 accident in snow conditions for 100 severity 1 accident
```

We see that the severity of accident decreases with snowy or icy conditions, that's probably because people tends to be more cautious on those conditions and also that they drive more slowly.

In the same vein, the difference might not be significative, but we see a slight increase in the severity when the road is wet.

It is possible that people tend to drive as they were on dry condition, and then, hit those unfortunate souls with more speed, hence increasing the number of serious injuries. But it is only a supposition, more careful analysis has to be conducted.

We find the same results with the number of pedestrians involved.

```
Proportion of pedestrian in accident when the road is dry :0.04
Proportion of pedestrian in accident when the road is wet :0.05
Proportion of pedestrian in accident when the road is ice :0.02
Proportion of pedestrian in accident when the road is snow :0.03
```

Here the difference between dry and wet is more significative, as the distance of braking is increased, the proportion of insufficient anticipation increases.

But when there is snow or ice, the car travels with less velocity and then there is more time to react.

This tries to confirm the suspicion made earlier, but it is still not sufficient to be sure.

What can be of interest for the authorities is that:

	index	JUNCTIONTYPE
1	Mid-Block (not related to intersection)	89800
2	At Intersection (intersection related)	62810
3	Mid-Block (but intersection related)	22790
4	Driveway Junction	10671
5	At Intersection (but not related to intersection)	2098
6	Ramp Junction	166

You see were there is the bigger number of accidents, and so you can add some actions to diminish those numbers.

One action can be more sensibilisation of the driver concerning driving on wet conditions, or how to act at an intersection.

Maybe analyse the comportment of drivers in those areas and find suitable actions to run.

Data

The data set is a statistical report of the city of Seattle, it is composed of 39 columns with around 200k rows.

All the columns on the dataset cannot be of any use. Like the geographical location of the accident, as we are not incorporating that in our model (though it can be of interest, I reckon that). The main reason is that the column has to have less that 10 different variables to be used properly. If I use the geography to model, you have to take into account a density of accident decreasing from the impact zone, and it can become really mathematically heavy, and I thought that was not the point here.

Finally, I choose the K-nearest neighbours as algorithm to use, and the simpler the model is, the better.

So I keep the severity of the accident as dependent variable (that is “y”. Column SEVERITY) that I will predict using the road condition (ROADCOND), the light condition (LIGHTCOND) and the type of junction (JUNCTIONTYPE).

For every column, there was empty lines, so I just dropped them.

I also keep only those five variables for the road conditions:

	index	ROADCOND
1	Dry	119728
2	Wet	45675
3	Ice	1107
4	Snow/Slush	840
5	Standing Water	104

As we can’t predict if there is oil on the road, for example.

Methodology

For more details on the model creation, see “model creation” in my github:
https://github.com/SylvianMunier/IBM-Data-Science-capstone-project

First, I had in mind to make a regression to model the increase of number and severity. But I soon realized that it would not be possible, as the data were discreate. So, I imagine this problem like a churn prediction for a telecom company with 2 outputs: severity = 1 or 2, so I opted for K-nearest

neighbours (KNN). Then I choose type of data containing few variables, like the 3 I chose, that is: junction type, road and light conditions. And for the same reason I dropped the location (too many lines, so you have to keep only the first rows, and doing that, I would have skewed my repartition).

The KNN algorithm need to have exclusively numerical values, so I replaced all the strings with a corresponding numerical value (1 to 5 or 6).

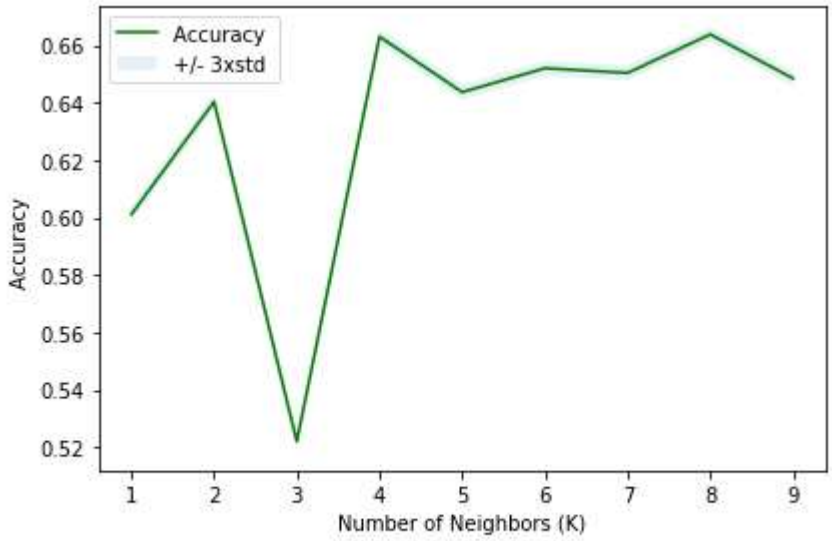
	index	ROADCOND
1	Dry	119728
2	Wet	45675
3	Ice	1107
4	Snow/Slush	840
5	Standing Water	104

	index	LIGHTCOND
1	Daylight	111583
2	Dark - Street Lights On	46481
3	Dusk	5623
4	Dawn	2391
5	Dark - No Street Lights	1376

	index	JUNCTIONTYPE
1	Mid-Block (not related to intersection)	74984
2	At Intersection (intersection related)	59343
3	Mid-Block (but intersection related)	21198
4	Driveway Junction	9982
5	At Intersection (but not related to intersection)	1794
6	Ramp Junction	153

The data were allocated 80% for training and 20% for test.

Then, for the model, I tried to automate the process to find the best K for the KNN algorithm, in this case, K=8. As seen in this graph (notice that, I had an envelope representing ± 3 standard variation, and notice also that the envelope is really small, so the standard variation is quite small) :




```
The best accuracy was with 0.663820130781404 with k= 8
```

After what my model can predict the severity with a score of $R^2 = 0.664$, which is a good result considering the simplicity of the model (that is 66% of the variation are explained by my model).

The code can be modified to add a scan() function to let the user add his own variables for prediction.

Conclusion & discussion

The pre-analysis gives:

```
There is 47.44 of severity 2 accident in dry conditions for 100 severity 1 accident
There is 49.63 of severity 2 accident in wet conditions for 100 severity 1 accident
There is 29.17 of severity 2 accident in ice conditions for 100 severity 1 accident
There is 19.95 of severity 2 accident in snow conditions for 100 severity 1 accident

Proportion of pedestrian in accident when the road is dry :0.04
Proportion of pedestrian in accident when the road is wet :0.05
Proportion of pedestrian in accident when the road is ice :0.02
Proportion of pedestrian in accident when the road is snow :0.03
```

We see that the severity of accident and the number of pedestrians involved, decreases with snowy or icy conditions, that's probably because people tend to be more cautious on those conditions and also that they drive more slowly.

In wet conditions, it is possible that people tend to drive as they were on dry condition, and then, hit those unfortunate souls with more speed, hence increasing the number of serious injuries.

In wet conditions the distance of braking is increased, the proportion of insufficient anticipation increases too.

But when there is snow or ice, the car travel with less velocity and then there is more time to react.

Those observations need to be analysed more thoroughly though.

What can be of interest for the authorities is that:

	index	JUNCTIONTYPE
1	Mid-Block (not related to intersection)	89800
2	At Intersection (intersection related)	62810
3	Mid-Block (but intersection related)	22790
4	Driveway Junction	10671
5	At Intersection (but not related to intersection)	2098
6	Ramp Junction	166

You see where there is the bigger number of accidents, and so you can add some actions to diminish those numbers.

One action can be more sensibilisation of the driver concerning driving on wet conditions, or how to act at an intersection.

Maybe analyse the comportment of drivers in those areas and find suitable actions to run.

The model can predict the severity with an $R^2 = 0.664$, which is a good result considering the simplicity of the model.

The code can be modified to add a scan() function to let the user add his own variables for prediction.