**IBM Data Science - Report Block 9 (Applied Data Science Capstone)**

**Introduction where you discuss the business problem and who would be interested in this project**

I​ developed a model that predicts the severity of an accident given the kind of location and the weather, road and light conditions. The model will indicate the severity of a car accident. This is of interrest for people at 911 in the USA. Based on the outcome of the model they can decide whether to send only police of also ambulance and/or fire department.

When the people of 911 are called they ask a lot of questions. But not all of these questions can be answered by the caller. One of the main questions will be if there are any injuries. If the caller doesn’t know the model can predict the chance of injuries based on a couple of variables that are easy to determine. These questions are based on the data I use and is described in the next section.

**Data where you describe the data that will be used to solve the problem and the source of the data**

G​iven the data set with collisions in Seattle I use the severity (SEVERITYCODE) as variable to predict, the dependent variable. And I use road conditions (ROADCOND), weather conditions (WEATHER), light conditions (LIGHTCOND) and the kind of location (ADDRTYPE) as the independent variables. These four have the greatest impact on the severity of car accidents.

S​EVERITYCODE: 1 (property damage only), 2 (injury also).

R​OADCON: Dry, Wet, Unknown, Ice, Snow/Slush, Other, Standing Water, Sand/Mud/Dirt, Oil.

W​EATHER: Clear, Raining, Overcast, Unknown, Snowing, Other, Fog/Smog/Smoke,

Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy.

L​IGHTCOND: Daylight, Dark - Street Lights On, Unknown, Dusk, Dawn, Dark - No Street Lights, Dark - Street Lights Off, Other, Dark - Unknown Lighting.

ADDRTYPE: Block, Intersection, Alley.

I don't use inattention (INATTENTIONIND), under influence (UNDERINFL) and speeding (SPEEDING), because these three are influenced by the driver.

The source of the data: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

**Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why**

As a start I inspected all the data in the data set and determined the usability for the case. SEVERITYCODE is of course the main variable. It’s not an indicator of the chance on an accident but it gives the severity. Given the fact there has been an accident. This is the dependent variable.

SEVERITYCODE has two different values: 1 and 2. 136485 records have SEVERITYCODE 1 and 58188 records have SEVERITYCODE 2. Total number of 194673 records with a valid value. This equals the total number of records so no missing values.

I did this check also for the other four, independent, variables. All of them had missing values. I changed the values in these records to “Unknown”. We still don’t know the correct value but at least they have a value. I also checked if the number of valid values equaled the total number of records. And this was the case in all four situations.

The values of SEVERITYCODE are 1 and 2. A number. This is OK. The other four values had a description (see the previous question). I changed all these to integer values.

The machine learning technique we’re dealing with is classification. This is a supervised technique. And we’re dealing with categorical variables, as stated above. Therefor the classification algorithm I’ve used is Logistic Regression.

**Results section where you discuss the results**

I split the data in train and test data and evaluated the used model by calculating different scores.

|  |  |
| --- | --- |
| Method | Score |
| Jaccard | 0.70 |
| Log\_loss | 0.59 |
| F1 | 0.58 |

And the confusion matrix is as follows.

A screenshot of a cell phone

Description automatically generated

When we look at the confusion matrix, we see that the majority of the predictions is SEVERITYCODE 1. We have just 41 cases of the 14447 where the SEVERITYCODE 2 is predicted right. This is a very low number.

A recommendation could be to add some more independent variables to raise this number. I tried this and added VEHCOUNT. This raised the results as follows.

|  |  |
| --- | --- |
| Method | Score |
| Jaccard | 0.71 |
| Log\_loss | 0.59 |
| F1 | 0.62 |

And the new confusion matrix as follows.

A screenshot of a cell phone

Description automatically generated

Now we have 889 cases where the SEVERITYCODE 2 is predicted right**.** This is a substantial raise compared to the situation we don’t use the variable VEHCOUNT.

**Discussion section where you discuss any observations you noted and any recommendations you can make based on the results**

With just a couple of independent variables the model doesn’t score well. I started with just WEATHER, LIGHTCOND and ROADCOND and SEVERITYCODE 2 was not predicted. Not at all. Then I added ADDRTYPE and the model improved. And by adding VEHCOUNT the model improved further. But still, a lot of SEVERITYCODE 2 is not predicted by the model.

A suggestion I would make is to invest the further addition of independent variables and to see whether the model improves. I suppose that LOCATION, for instance, improves the score. I didn’t use this variable because there are too many different addresses and it’s easy to use two similar, but not identical, descriptions for the same address. Too difficult for the purpose I developed the model for.

I added JUNCTIONTYPE and the model improved further. I didn’t add the results in this report, but it supports me in the idea that adding more independent variables will improve the model.

On one hand I wanted to keep the model simple to use. With a few independent variables. On the other hand, I wanted a model with a high accuracy. With a lot of independent variables. After all I finished in between. More independent variables but not the accuracy I hoped for.

**Conclusion section where you conclude the report**

This was not an easy assignment. But I learned a lot. It gave me more confidence that I can develop a model on my own and create a report or presentation to present it to stakeholders. On the other hand, there is still a lot to learn. But I’m anxious to do so.