Capstone Project Car accident severity

By: Leandro Recova

August 29th, 2020

1 – Introduction and Business problem.

- The goal of this capstone project is to provide a machine learning model that can predict the severity of car accident based on a dataset provided by the course.
- Based on the information of previous accidents, weather condition, traffic jam, the model could provide the likelihood of an accident for a driver that might be driving along the road.
- This kind of warning would be very important so the driver can make decisions while driving in such conditions.
- This model could be integrated with a GPS software application that runs in the user application based on the trajectory followed by the driver.

2 - Dataset Description

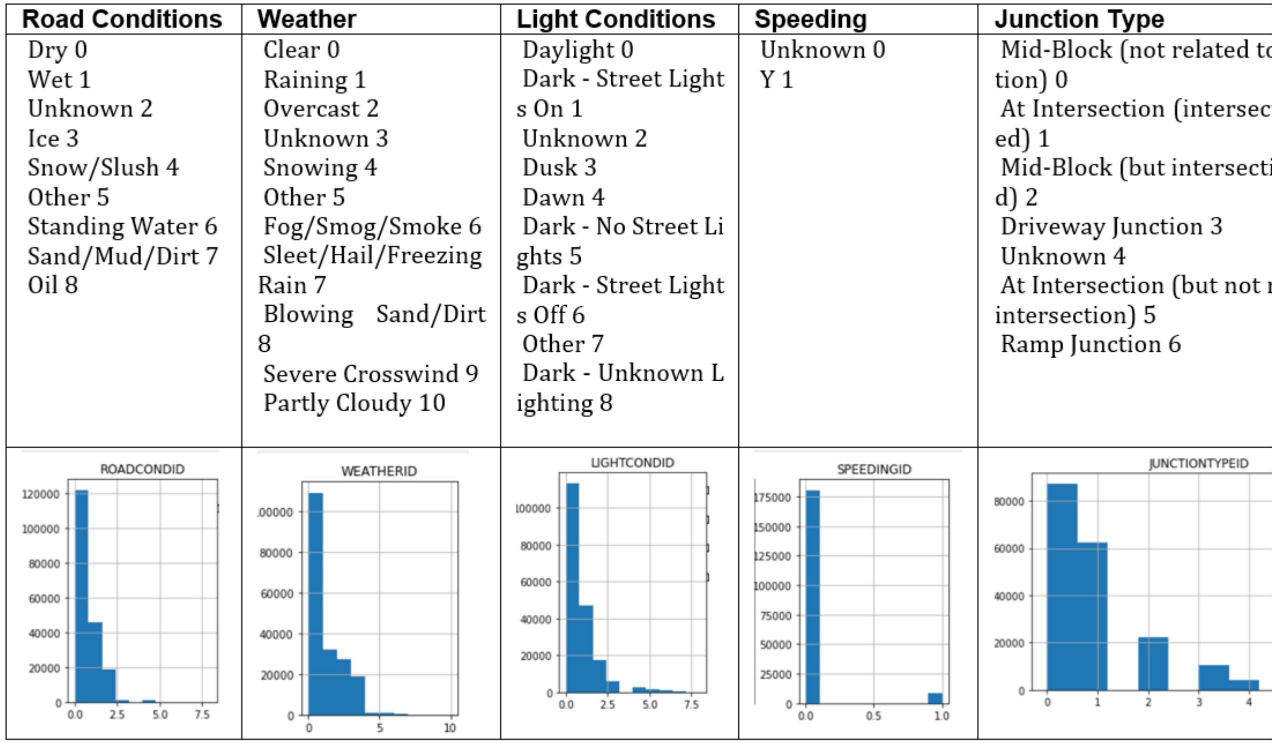
- The dataset that will be used during this capstone project was provided in the week 1 of the course.
- It has a list of 38 fields described as listed in the table below.
- By analyzing the data, the dataset will have to go through a pre-processing problem since some columns still have blanks, inconsistency of data description (E.g: Y, N, 0, 1 values in the same column).
- Based on the fields presented in this dataset, we will have to go through a pre-processing of the fields, make sure we select the features that will have a impact in the model, and remove those columns that will not be necessarily for this project.
- We will use a logistic regression model to predict the severity of car accident based on this dataset.

3 - Methodology

- As mentioned in the previous section, the data was analyzed and we selected few features for this model. They are listed below:
- **Location:** Fields X and Y representing the longitude and latitude, respectively. We also renamed these fields during the data processing.
- **Road Condition**: Field Road Cond. We created an additional column called ROADCONDID to have a number associated with one of each type of conditions listed in the dataset.
- **Weather**: This is an important factor and we also categorized this field by associating a number with each of the conditions listed.
- **Light conditions**: Like the previous item, we also created a column called WeatherID to represent the particular category.
- **Speeding**: This field had multiple blanks and also a "Y" when speeding was flagged. We added fill it out the blanks with 0 and the "Y" field with 1 in a new column called SpeedingID.
- **Junction Type:** We categorized each of the junction types by creating a new column called JunctionTypeID.

3 - Methodology

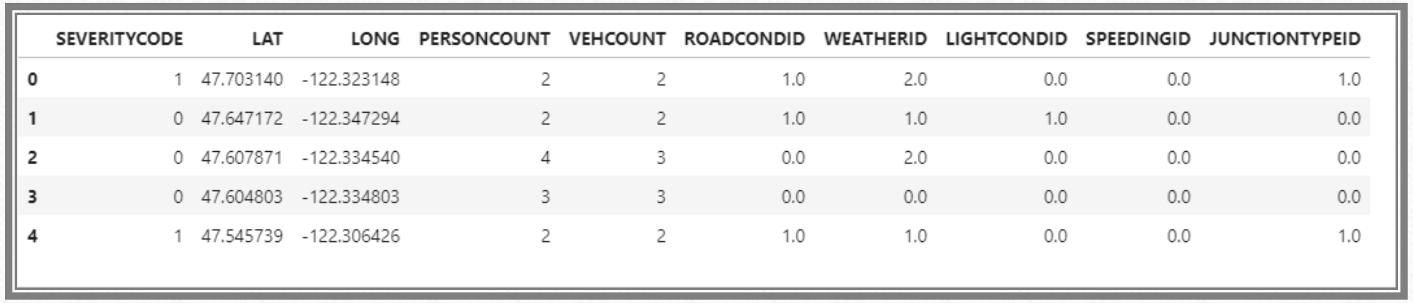
- Table 1: Features used in the LR Regression model and its histograms.



Road Conditions	Weather	Light Conditions	Speeding	Junction Type
Dry 0 Wet 1 Unknown 2 Ice 3 Snow/Slush 4 Other 5 Standing Water 6 Sand/Mud/Dirt 7 Oil 8	Clear 0 Raining 1 Overcast 2 Unknown 3 Snowing 4 Other 5 Fog/Smog/Smoke 6 Sleet/Hail/Freezing Rain 7 Blowing Sand/Dirt 8 Severe Crosswind 9 Partly Cloudy 10	Daylight 0 Dark - Street Lights On 1 Unknown 2 Dusk 3 Dawn 4 Dark - No Street Lights 5 Dark - Street Lights Off 6 Other 7 Dark - Unknown Lighting 8	Unknown 0 Y 1	Mid-Block (not related t tion) 0 At Intersection (intersec ed) 1 Mid-Block (but intersecti d) 2 Driveway Junction 3 Unknown 4 At Intersection (but not intersection) 5 Ramp Junction 6

3 - Methodology

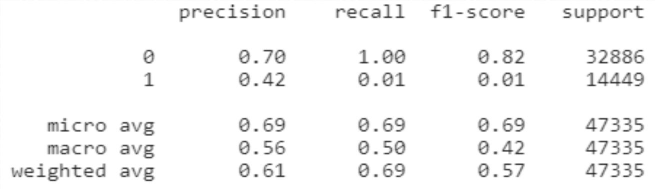
- In most of the accidents, it is clear the road conditions dry, weather clear, light condition day light, not speeding, and junction types mid-block are the ones with most of the accidents.
- Once the categories have been defined, we created a panda dataframe called **finalDF** with the fields required for the logistic regression algorithm. Below is a snapshot of the first rows of the input data.



	SEVERITYCODE	LAT	LONG	PERSONCOUNT	VEHCOUNT	ROADCONDID	WEATHERID	LIGHTCONDID	SPEEDINGID	JUNCTIONTYPEID
0	1	47.703140	-122.323148	2	2	1.0	2.0	0.0	0.0	1.0
1	0	47.647172	-122.347294	2	2	1.0	1.0	1.0	0.0	0.0
2	0	47.607871	-122.334540	4	3	0.0	2.0	0.0	0.0	0.0
3	0	47.604803	-122.334803	3	3	0.0	0.0	0.0	0.0	0.0
4	1	47.545739	-122.306426	2	2	1.0	1.0	0.0	0.0	1.0

4 – Results and Discussion

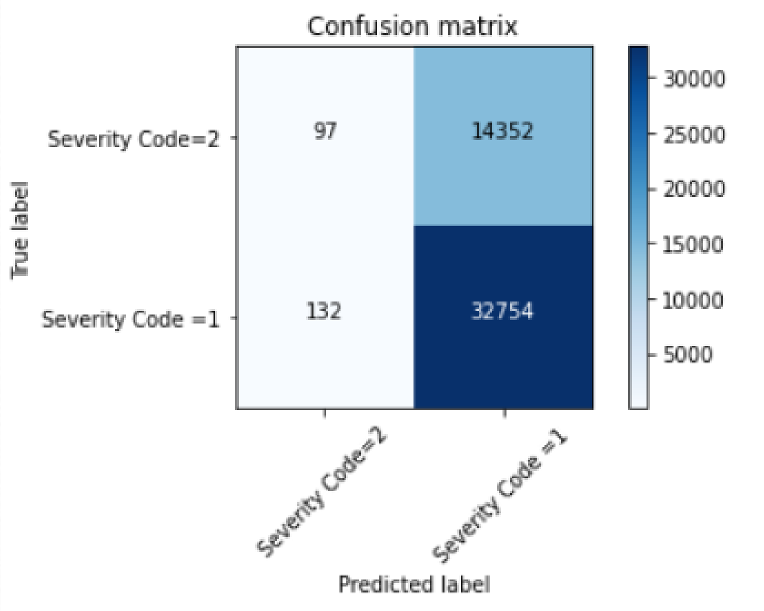
- We obtained a 70% precision of accuracy in estimating the severity code 1 and 42% for severity code 2.
- The model did not estimate well the severity code 2.
- One of the reasons could be the fact that there might be some human decisions in classifying the codes based on all factors used in this model.
- We also tried to play with different values of the regularization factor, but the results seem to be within the same values obtained above.

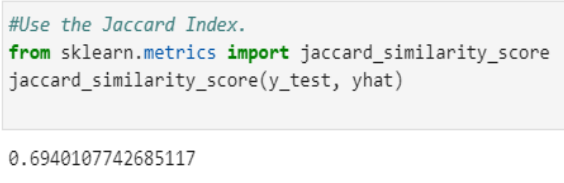


```
              precision    recall  f1-score   support

           0       0.70      1.00      0.82     32886
           1       0.42      0.01      0.01     14449

   micro avg       0.69      0.69      0.69     47335
   macro avg       0.56      0.50      0.42     47335
weighted avg       0.61      0.69      0.57     47335
```



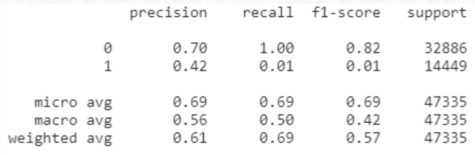


```
#Use the Jaccard Index.
from sklearn.metrics import jaccard_similarity_score
jaccard_similarity_score(y_test, yhat)
```

```
0.6940107742685117
```

4 – Results and Discussion

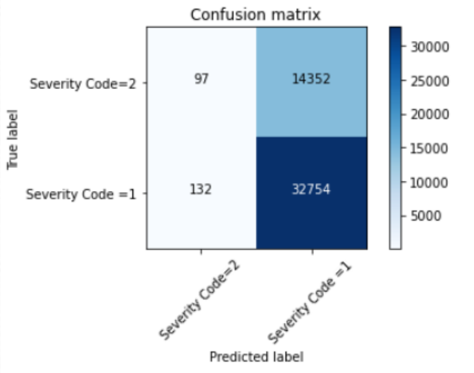
- We obtained a 70% precision of accuracy in estimating the severity code 1 and 42% for severity code 2.
- The model did not estimate well the severity code 2.
- One of the reasons could be the fact that there might be some human decisions in classifying the codes based on all factors used in this model.
- We also tried to play with different values of the regularization factor, but the results seem to be within the same values obtained above.

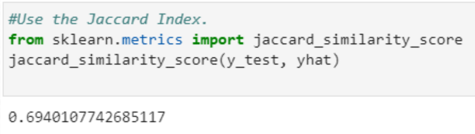


```
              precision    recall  f1-score   support

           0       0.70      1.00      0.82     32886
           1       0.42      0.01      0.01     14449

   micro avg       0.69      0.69      0.69     47335
   macro avg       0.56      0.50      0.42     47335
weighted avg       0.61      0.69      0.57     47335
```





```
#Use the Jaccard Index.
from sklearn.metrics import jaccard_similarity_score
jaccard_similarity_score(y_test, yhat)
```

```
0.6940107742685117
```

5 – Conclusions

- There was a significant time spent during the data preparation, choice of the features, categorizing the features that could have a more impact in the outcome prediction, and the choice of the method to evaluate the results.

- The logistic regression method was utilized since we had to estimate between two severity codes. We used the liblinear solver and obtained a jaccard index of 0.69 and a precision of 70% of severity code 1 and 42% for severity code 2. The estimation for severity code 2 was not the result expected, but, after analyzing the dataset, it seems there might be a human decision factor to decide when to choose these codes.

- The final recommendation left for the dataset owner would be to reevaluate how the codes 1 and are being assigned and if any human factor is used to decide between the severity of these codes. A evaluation of this method against future decisions would help to better classify these codes and provide a better prediction of accidents for insurance companies and drivers that might be driving in a high risk road.