Applied Data Science Capstone - Car accident severity

Peer-graded Assignment

Table of contents

Introduction - Business Problem

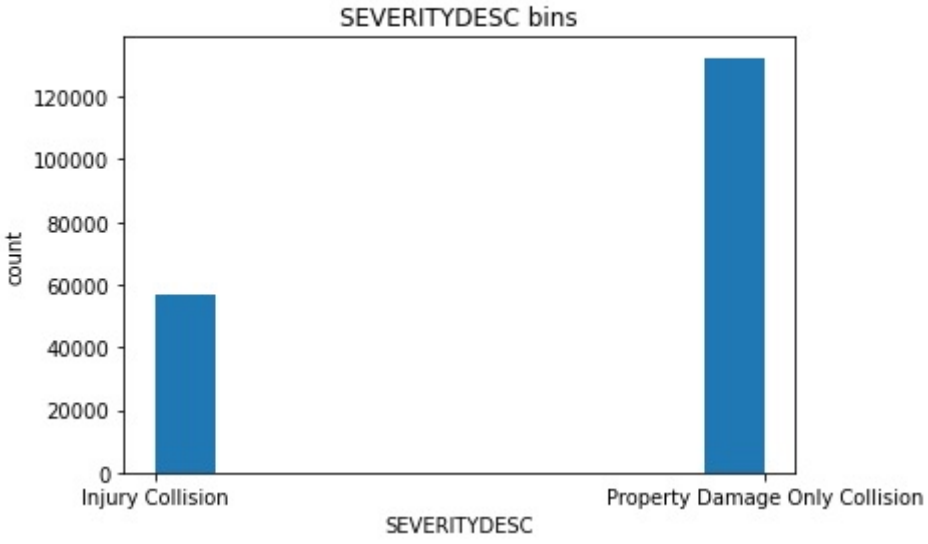
The objective is to predic the probability of severity of an accident (injury or property damage only) given the weather and road conditions. Prediction will be done with the Logistic Regression.

Methodology

For this project I will mainly concentrate on the Road and Weather Conditions to estimate severity of the accident.

Analysis Defining X 'Weather - Clear', 'Weather - Raining', 'Road Condition - Dry', 'Road Condition - Wet' , y 'SEVERITYCODE' for dataset. Normalizing the dataset Train/Test dataset Modeling (Logistic Regression with Scikit-learn)

```
Text(0.5, 1.0, 'SEVERITYDESC bins')
```



```
Property Damage Only Collision    132451
Injury Collision                   57092
Name: SEVERITYDESC, dtype: int64
```

```
df['WEATHER'].value_counts()
```

```
Clear                       111116
Raining                      33141
Overcast                     27702
Unknown                      15080
Snowing                        907
Other                          830
Fog/Smog/Smoke                 569
Sleet/Hail/Freezing Rain       113
Blowing Sand/Dirt               55
Severe Crosswind                25
Partly Cloudy                    5
Name: WEATHER, dtype: int64
```

```
df['ROADCOND'].value_counts()
```

```
Dry               124432
Wet                47450
Unknown            15068
Ice                 1206
Snow/Slush          1002
Other                132
Standing Water       115
Sand/Mud/Dirt         74
Oil                   64
Name: ROADCOND, dtype: int64
```

•••

	Weather - Clear	Weather - Raining	Road Condition - Dry	Road Condition - Wet	SEVERITYCODE
count	189543.000000	189543.000000	189543.000000	189543.000000	189543.000000
mean	0.586231	0.174847	0.656484	0.250339	0.301209
std	0.492509	0.379837	0.474883	0.433209	0.458784
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	1.000000	0.000000	0.000000
75%	1.000000	0.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

•••

```
Train set: (132680, 4) (132680,)
Test set: (56863, 4) (56863,)
```

Modeling Logistic Regression

```
array([[0.68, 0.32],
       [0.68, 0.32],
       [0.68, 0.32],
       ...,
       [0.68, 0.32],
       [0.66, 0.34],
       [0.68, 0.32]])
```

First column is the probability of class 1, P(Y=1|X), and second column is probability of class 0, P(Y=0|X)

jaccard index

```
0.6945465416879165
```

Classification Report

```
              precision    recall  f1-score   support

           0       0.69      1.00      0.82     39494
           1       0.00      0.00      0.00     17369

   micro avg       0.69      0.69      0.69     56863
   macro avg       0.35      0.50      0.41     56863
weighted avg       0.48      0.69      0.57     56863
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/sklearn/metrics/classification.py:1143: UndefinedMetricWarning: Precision and F-score
are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
```

Log loss

```
0.5990610314273476
```

Conclusion

The objective was to predic the probability of severity of an accident (injury or property damage only) given the weather and road conditions using Logistic Regression. Accidents with property damage are more likely, also clear weather conditions and dry road conditions have highest number of accidents.