A COLLISION SEVERITY PREDICTION MODEL

José Carlos Del Solar Rueda

2020 - October - 03

1. Introduction

Nowadays one of the main causes of death in the road are car accidents, these are caused due to known conditions like weather, state of the road, etc.
The purpose of this Capstone Project is to build a data science model that would help Us whether a car accident would cause property damage only or be an injury collision in the worst case scenario.

This project is designed to help the local authorities take better decisions about car accident prevention and also for the public in general that could use this information as a tool to make better choices in their daily life.

2. Data

2.1. Data Source

To build this model I used the shared data of traffic collisions from the city of Seattle. Based on that data I analyzed the registered traffic collisions variables like: Weather condition, Road condition, and light conditions during the collision.
Data was extracted from: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

At the beginning the dataset had 194,673 rows with 37 different attributes. For the purpose of this exercise null rows and non relevant columns had to be cleaned up. The data included all sorts of car collisions from 2004 to the present.

2.2. Data Cleaning

Here is a brief explanation of each of the variables used for the model:

- SEVERITYCODE: A code that corresponds to the severity of the collision.
- WEATHER: A description of the weather conditions during the time of the collision.

- ROADCOND: The condition of the road during the collision.
- LIGHTCOND: The light conditions during the collision.

3. Methodology

To solve this case I used the logistic regressión classification algorithm, let me explain why. In logistic regression, we have one or more independent variables such as Weather, Light condition and Road Condition which I grouped and transformed into numerical values. And also, to predict the outcome we have a categorical class such as Severity of the Collision, which we call the dependent variable.

3.1. Preprocessed data

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight

So given this dataset, I tried to predict the severity of a car accident, using the explained independent variables. But first I had to analyze the composition of the raw data. Then we preprocess the data frame and select only the key features. In this process I had to clean the unuseful rows or not statistical significant data, like “Unknows” and “Others”. Then I had to make sure to use only numerical data for the independent variables. Drop the null rows, and finally We are ready to execute the logistic regression algorithm to evaluate the results.

```
In [5]: df['SEVERITYDESC'].value_counts()
Out[5]: Property Damage Only Collision    136485
        Injury Collision                   58188
        Name: SEVERITYDESC, dtype: int64

In [6]: df['SEVERITYCODE'].value_counts()
Out[6]: 1    136485
        2     58188
        Name: SEVERITYCODE, dtype: int64
```

Out[9]:

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	0	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	0	Raining	Wet	Daylight

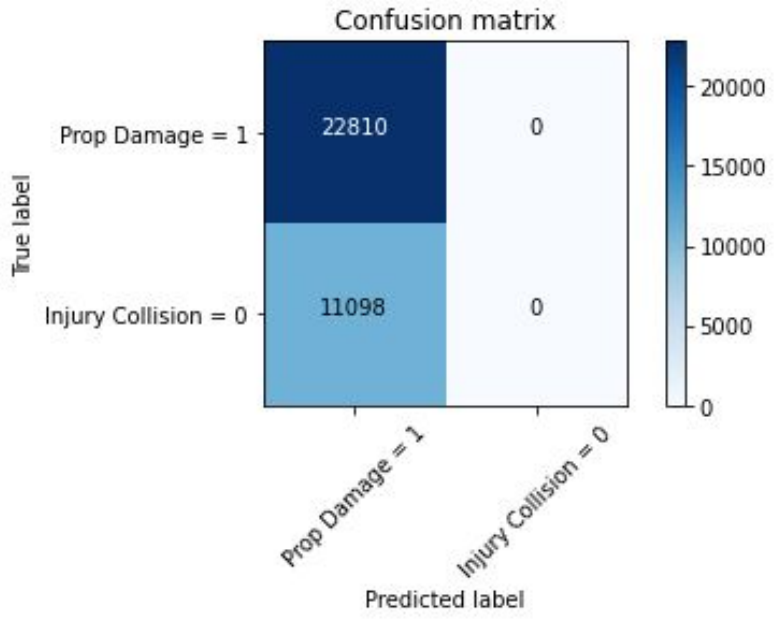
3.2. Processed data

```
In [35]: churn_df.head(5)
```

Out[35]:

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	0	2.0	1.0	0.0
1	1	1.0	1.0	1.0
2	1	2.0	0.0	0.0
3	1	0.0	0.0	0.0
4	0	1.0	1.0	0.0

```
Confusion matrix, without normalization
[[22810     0]
 [11098     0]]
```



4. Discussion

As you can see in the previous graphic, 22,810 collisions out of a test sample of 33,908 are correctly predicted as “1” Property Damage only. That means that means that 68.92% of the time a person that drives through the roads of Seattle with bad weather, poor road conditions and or absence of good light conditions could have a car accident with property damage.

On the other hand 11,098 samples of collision were misclassified as damage property only when they should have been an injury collision type. This means that 32.72% of the time a collision might result in injury consequences for the people involved.

With jaccard's index for **accuracy evaluation** we can see there is an accuracy prediction of **67.27%**

```
In [47]: from sklearn.metrics import jaccard_similarity_score
         jaccard_similarity_score(y_test, yhat)

Out[47]: 0.6727026070543825
```

5. Conclusion

The purpose of this project was to build a data science model that would help us predict the possibility of getting into a car collision and the severity of it. This would help local authorities to take better decisions about car accident prevention and also for the public in general that could use this information as a tool to make better choices in their daily life. We did this first by understanding that a logistic regression model would be the best classification algorithm due to the variables found in the dataset. Then I had to identify the key independent variables: Weather conditions; Light conditions; Road conditions.

After cleaning and processing the data into the algorithm we can see that from one out of three collisions in the city of Seattle could end up in a serious injury for the persons involved. Whatsmore, using Jaccard's evaluation index we can also identify an accuracy of 67.27% for the prediction model. Finally, we can conclude that if we could have data of the cars not involved in a collision every time they go out to the road, we could take the prediction model to a new level. We could even try to predict if they would or would not get into a car accident. The possibility of having that kind of data would be wonderful, and thus with reason we can say that having good data is the new gold.