Predicting the Severity of Collision in Seattle City

Yanbo Lyu

October 11, 2020

1. Introduction:

Seattle is a gorgeous City to live in. However the car collision problem always harass the City.

In most cases, Road condition, lighting condition, weather and speeding are the main causes of occurring accidents that can be prevented by better regulations

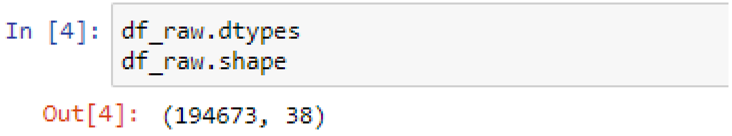
The target audience of the project is local Seattle government, police and citizens, The model and its results are going to provide some advice for them to make insightful decisions for reducing the number of accidents and injuries for the city.

2. Data acquisition and cleaning:

2.1 Data Source

Using the data provided by Seattle DOT on Collisions, I will investigate the connection of severity of car accidents with Road condition, lighting condition, weather and speeding.

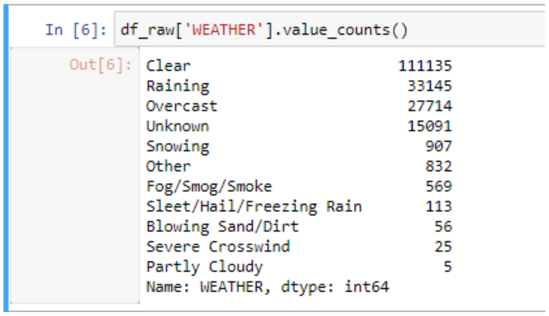
This data provides collisions from 2004 to the present in Seattle. The Data is not clean and an initial analysis of the data shows that there are 194,673 records and 38 fields. This dataset only includes two types of severity – Property Damage (Severity 1) and Injury (Severity 2).



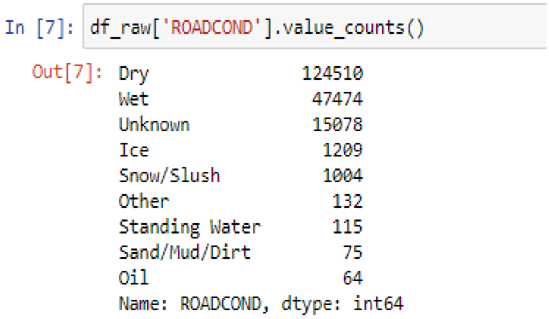
```
In [4]: df_raw.dtypes
        df_raw.shape

Out[4]: (194673, 38)
```

2.2 Data cleaning and Feature

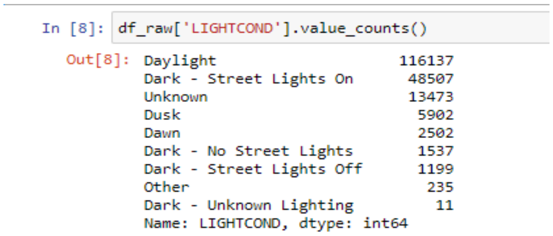
From the data frame we can clearly see that we do not need use all of the data. We want to focus the relation between Road condition, lighting condition, weather and speeding with Severity.



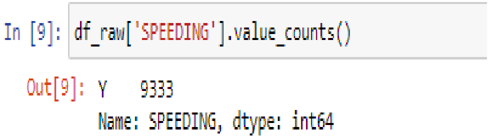
```
In [6]: df_raw['WEATHER'].value_counts()

Out[6]: Clear                       111135
        Raining                      33145
        Overcast                     27714
        Unknown                      15091
        Snowing                        907
        Other                          832
        Fog/Smog/Smoke                 569
        Sleet/Hail/Freezing Rain       113
        Blowing Sand/Dirt               56
        Severe Crosswind                25
        Partly Cloudy                    5
        Name: WEATHER, dtype: int64
```



```
In [7]: df_raw['ROADCOND'].value_counts()

Out[7]: Dry               124510
        Wet                47474
        Unknown            15078
        Ice                 1209
        Snow/Slush          1004
        Other                132
        Standing Water       115
        Sand/Mud/Dirt         75
        Oil                   64
        Name: ROADCOND, dtype: int64
```



```
In [8]: df_raw['LIGHTCOND'].value_counts()

Out[8]: Daylight                    116137
        Dark - Street Lights On      48507
        Unknown                      13473
        Dusk                          5902
        Dawn                          2502
        Dark - No Street Lights       1537
        Dark - Street Lights Off      1199
        Other                          235
        Dark - Unknown Lighting         11
        Name: LIGHTCOND, dtype: int64
```



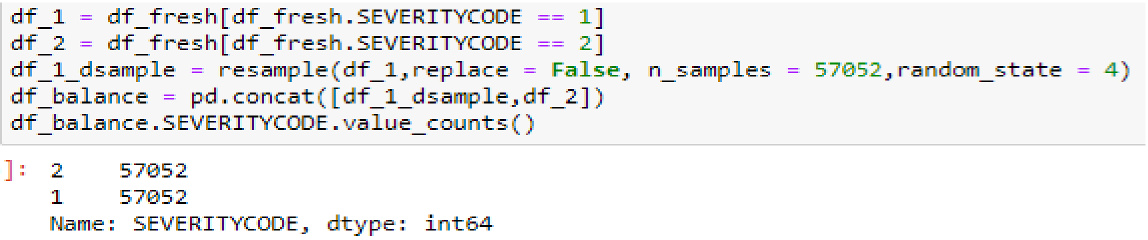
```
In [9]: df_raw['SPEEDING'].value_counts()

Out[9]: Y    9333
        Name: SPEEDING, dtype: int64
```

We can see the data lack of speeding info, so we decide only analyze the relation weather, light condition and road condition with severity. Also, we drop the data which lack of severity info. After that, we get 189,337 total data set including 132,285 severity 1 and 57,052 severity 2 data set.

3. Exploratory Data Analysis

Our target severity is imbalance, with 132,285 severity 1 and only 57,052 severity 2.

Severity 1 size almost three times than severity 2. So we decide use de-sampling method down size severity 1 to match severity 2.



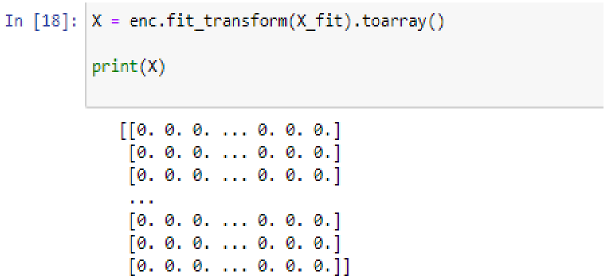
```
df_1 = df_fresh[df_fresh.SEVERITYCODE == 1]
df_2 = df_fresh[df_fresh.SEVERITYCODE == 2]
df_1_dsample = resample(df_1,replace = False, n_samples = 57052,random_state = 4)
df_balance = pd.concat([df_1_dsample,df_2])
df_balance.SEVERITYCODE.value_counts()

]: 2    57052
   1    57052
   Name: SEVERITYCODE, dtype: int64
```

After we down size, we get 57052 severity 1 and severity 2 data size for modeling.

4. Predictive Modeling

4.1 Transform and normalize the data

Before we apply the model, first we transform the string type data to numerical data for analyze. Here we use one hot encoding method.

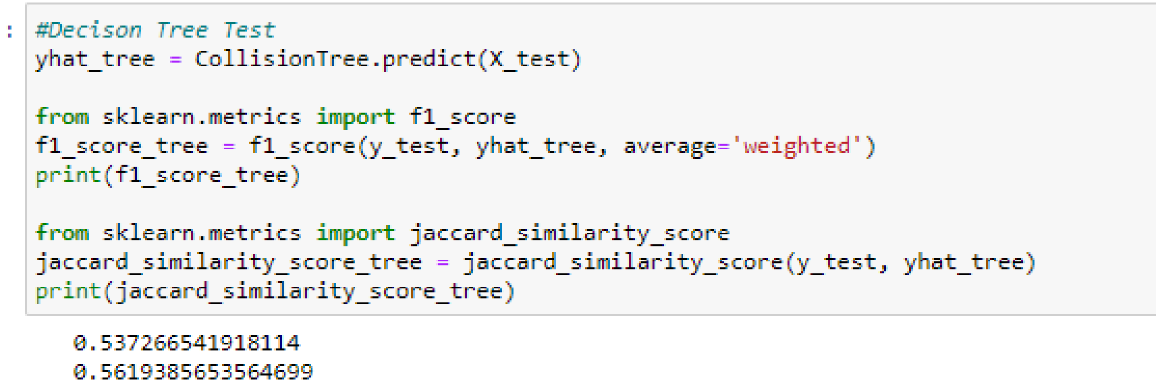


```
In [18]: X = enc.fit_transform(X_fit).toarray()

         print(X)
```

```
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

4.2 Decision Tree Modeling

Now data is fully ready, clearly we know the problem type is classification. So we first apply decision tree model to make prediction.



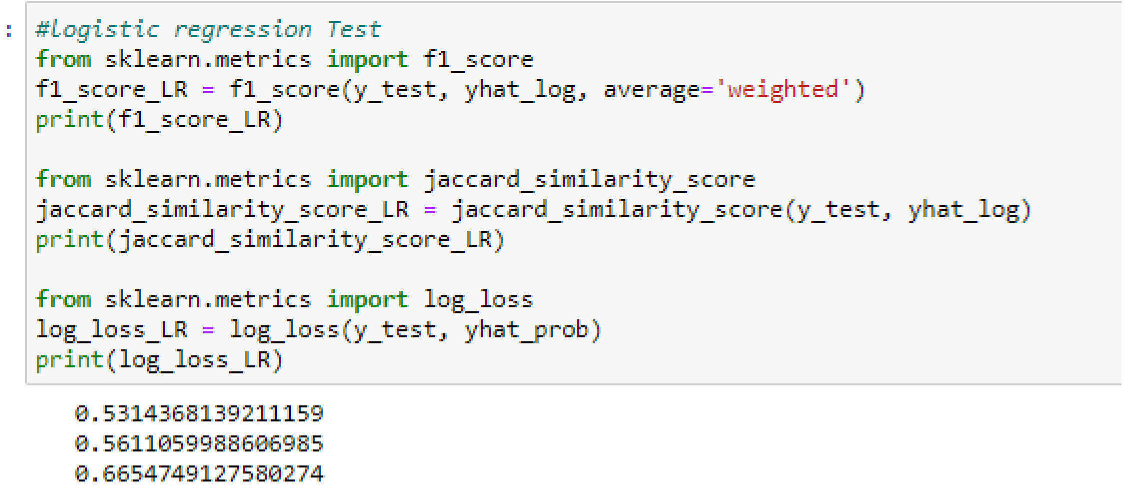
```
: #Decison Tree Test
  yhat_tree = CollisionTree.predict(X_test)

  from sklearn.metrics import f1_score
  f1_score_tree = f1_score(y_test, yhat_tree, average='weighted')
  print(f1_score_tree)

  from sklearn.metrics import jaccard_similarity_score
  jaccard_similarity_score_tree = jaccard_similarity_score(y_test, yhat_tree)
  print(jaccard_similarity_score_tree)
```

```
0.537266541918114
0.5619385653564699
```

Here we use f1 score and jaccard score to check the performance of the model, we can see that the three parameter do have some connection with severity.

4.3 Logistic Modeling

Next we apply logistic regression model.



```
: #logistic regression Test
  from sklearn.metrics import f1_score
  f1_score_LR = f1_score(y_test, yhat_log, average='weighted')
  print(f1_score_LR)

  from sklearn.metrics import jaccard_similarity_score
  jaccard_similarity_score_LR = jaccard_similarity_score(y_test, yhat_log)
  print(jaccard_similarity_score_LR)

  from sklearn.metrics import log_loss
  log_loss_LR = log_loss(y_test, yhat_prob)
  print(log_loss_LR)
```

```
0.5314368139211159
0.5611059988606985
0.6654749127580274
```

We get similar f1 score and jaccard score here with decision model.

5. Conclusions

Base on above analysis, we can conclude that the three variable: weather, light condition and road condition have some impact on severity (property damage and injury) We suggest citizen in Seattle be alert when they driving in the city, try to go out during better weather.

And the Seattle government should improve the traffic condition by improving the lighting condition and road condition in the city to make the city nice and safer.