Analysis on Car Accident Severity

AYUSH ANAND JHA

Introduction

- The project is based on the dataset provided in the course of Applied Data Science Capstone of Road Accident occurred in Seattle
  - generated from https://data.seattle.gov/
- This project is to take a glimpse at different accidents that occurred in Seattle City.
- The objective of the project is to design a system that can be used to avoid or tackle any future occurrences of road accidents that can be caused due to several reasons based on past data.

Problem

- The idea is to provide a pre-determined possibility of occurrence severe accident prior to the movement of the vehicles in traffic that can help several travelers who are keen to drive on a particular lane or highway.

- Analysing a significant range of factors, including weather conditions, special events, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed

- this knowledge of a severe accident situation can be warned to driver so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.

Who Are The Target Audience?

- The daily travelers and drivers of the city
- those who possess an interest in machine learning can use this project for research purposes
- Anyone who is keen to try roads of the city would've the pre-determined possibility of encountering an accident
  - It can help them to avoid that path or drive to save time as well as life
- Government should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year.
- Others interested could be private companies investing in technologies aiming to improve road safety.

Data Acquisition

- The data that will be used in the project is the one that is provided with the course 'Data-Collisons.csv' consist of 38 features as column and around 1.9 million records of accidents in rows.

- These information can be obtained from Seattle Department of Transportation (SDOT). SDOT has an open data platform which can be found in "https://data.seattle.gov/". In this platform, they update their information about collisions weekly. We can find all information we need in this dataset.

  - The attribute information details can be found in "https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf"

Data Preparation

- The problem is predicting the severity code by using the independent variables. Hence, it is a classification problem. The "severity" depends on the following data:

1.Accident location: Latitude("Y" column - float), Longitude("X" column - float)

2.Road coditions: "ROADCOND" column - text

3.Weather condition: "WEATHER" column - text

4.Junction: "JUNCTIONTYPE" column - text

5.Car speeding: "SPEEDING" column - boolean

6.Number of people involved: "PERSONCOUNT" column - integer

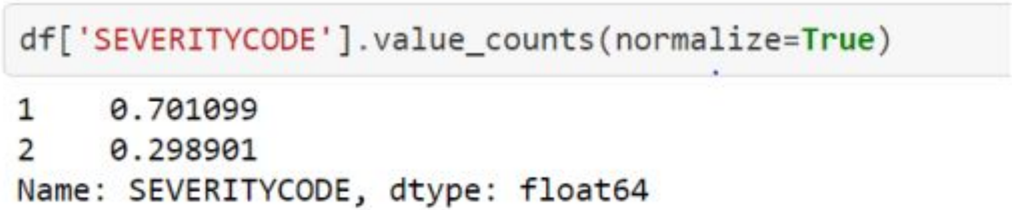
7.Light conditions: "LIGHTCOND" column - text

8.Number of vehicles involved in: "VEHCOUNT" column - integer

9.The date time when the accident occurs: "INCDATE", "INCDTTM" columns - text

Data Preparation - Severity

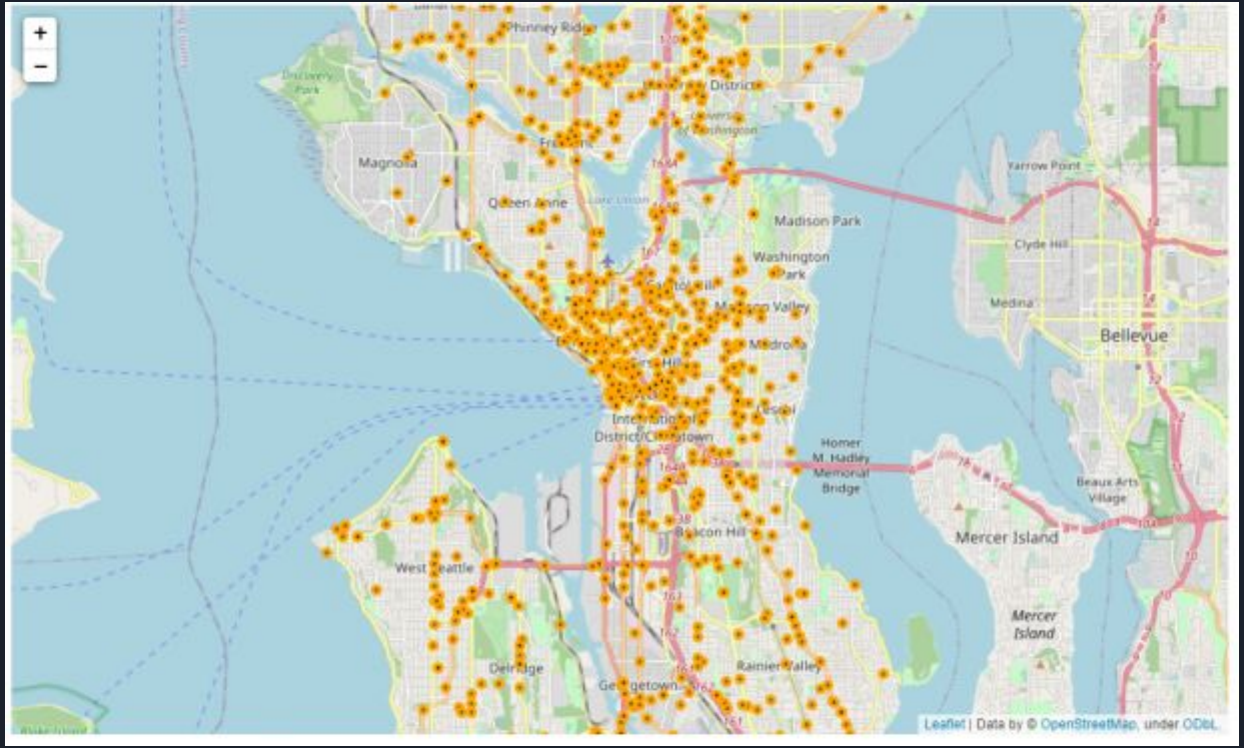
We can see that the dataset contains only 2 severities: "1" (prop damage) and "2" (injury). It will limit the prediction because the classification can not perform with the label which doesn't exist in dataset such as "3" (fatality), "2b" (serious injury) and "0" (unknown).



```
df['SEVERITYCODE'].value_counts(normalize=True)
```

```
1    0.701099
2    0.298901
Name: SEVERITYCODE, dtype: float64
```

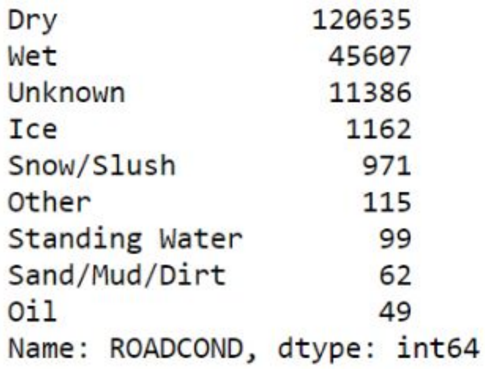
Map For Collision Distribution



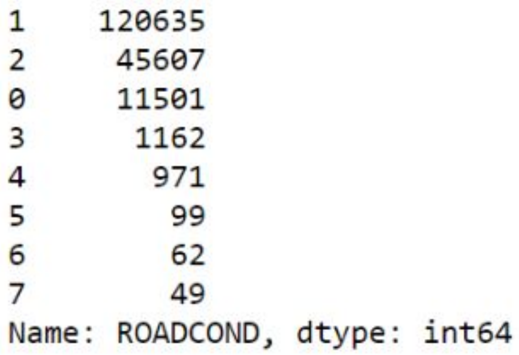
Data Preparation

❖ We've to drop the missing value of the longitude and latitude in order to data preparation. Also later we’ve to encode the categorical data, one of the example is given below.

❖ We’ll perform same operation for other attribute such as WEATHER, JUNCTIONTYPE, and LIGHTCOND.



```
Dry               120635
Wet                45607
Unknown            11386
Ice                 1162
Snow/Slush           971
Other                115
Standing Water        99
Sand/Mud/Dirt         62
Oil                   49
Name: ROADCOND, dtype: int64
```



```
1    120635
2     45607
0     11501
3      1162
4       971
5        99
6        62
7        49
Name: ROADCOND, dtype: int64
```

Model Building and Evaluation

❖ In this analysis we are going to used following models as we find our categorical:

1. K Nearest Neighbor (KNN)
2. Decision Tree
3. Logistic Regression

❖ Evaluation is important as it shows the clear picture of how much efficient the models were after being trained and tested. In this project F1-Score and Jaccard Score are used as evaluation metrics.

Evaluation Table

Metrics / Models	KNN	Decision Tree	Logistic Regression
F1-Score	0.64	0.66	0.62
Jaccard Score	0.47	0.52	0.47

Results

- From the above evaluation of different classification models, we can observe that F1-score of the different models didn't varied much, yet, Logistic Regression model was significantly better choice for the project (score = 0.62).
- However, according to Jaccard Score both KNN and Logistic Regression equally suits the requirement with score of 0.47
- It can be concluded that three models chosen for the development and evaluation are being studied and verified altogether.

Conclusion

- The exploratory analyses of the extracted dataset and the models that were built in order to develop a proper system that can predict car severity for the intended target audience mentioned in previous section.
- This project is also going to help individual in determining the best model among chosen ones.