Coursera - IBM Data Science

Applied Data Science Capstone

A project report on

Car Accident Severity Analysis

By

Ayush Anand Jha

October 1, 2020

1. 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/. The intention of this document is to describe what is the purpose of the project and what analysis is performed in order to make this project to generate useful results.

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.

1.1. 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 These insights, could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accidents can occur as well as saving both, time and money, In addition, 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.

1.2. Target Audience

The target audience for this project is the daily travelers and drivers of the city. Also, 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, and 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.

2. Data Understanding

2.1. 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 features/attributes can be used in the implementation of the model. The label of the dataset is severity, which describes the fatality of an accident.

To perform this analysis, we need the following data:

1. Accident location
2. Road conditions
3. Weather condition
4. Junction
5. Car speeding
6. Number of people involved
7. Light conditions
8. Number of vehicles involved

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”

2.2. Features

The data set contains 38 columns:

1. SEVERITYCODE: A code that corresponds to the severity of the collision.
2. X: The longitude
3. Y: The latitude
4. OBJECTID: ESRI unique identifier
5. INCKEY: A unique key for the incident
6. COLDETKEY: Secondary key for the incident
7. REPORTNO: Report number
8. STATUS: Match or unmatch
9. ADDRTYPE: Collision address type: Alley, Block, Intersection
10. INTKEY: Key that corresponds to the intersection associated with a collision
11. LOCATION: Description of the general location of the collision.
12. EXCEPTRSNCODE: Enough or not enough information
13. EXCEPTRSNDESC: Enough or not enough information
14. SEVERITYCODE: Duplicated column
15. SEVERITYDESC: A detailed description of the severity of the collision

16. COLLISIONTYPE: Collision type
17. PERSONCOUNT: The total number of people involved in the collision
18. PEDCOUNT: The number of pedestrians involved in the collision. This is entered by the state.
19. PEDCYLCOUNT: The number of bicycles involved in the collision. This is entered by the state.
20. VEHCOUNT: The number of vehicles involved in the collision. This is entered by the state.
21. INCDATE: The date of the incident.
22. INCDTTM: The date and time of the incident.
23. JUNCTIONTYPE: Category of junction at which collision took place
24. SDOT_COLCODE: A code given to the collision by SDOT.
25. SDOT_COLDESC: A description of the collision corresponding to the collision code.
26. INATTENTIONIND: Whether or not collision was due to inattention. (Y/N)
27. UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol.
28. WEATHER: A description of the weather conditions during the time of the collision.
29. ROADCOND: The condition of the road during the collision.
30. LIGHTCOND: The condition of the road during the collision.
31. PEDROWNOTGRNT: Whether or not the pedestrian right of way was not granted. (Y/N)
32. SDOTCOLNUM: A number given to the collision by SDOT.
33. SPEEDING: Whether or not speeding was a factor in the collision. (Y/N)
34. ST_COLCODE: A code provided by the state that describes the collision.
35. ST_COLDESC: A description that corresponds to the state's coding designation.
36. SEGLANEKEY: A key for the lane segment in which the collision occurred.
37. CROSSWALKKEY: A key for the crosswalk at which the collision occurred.
38. HITPARKEDCAR: Whether or not the collision involved hitting a parked car. (Y/N)

3. Exploratory Data Analysis

3.1. 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:

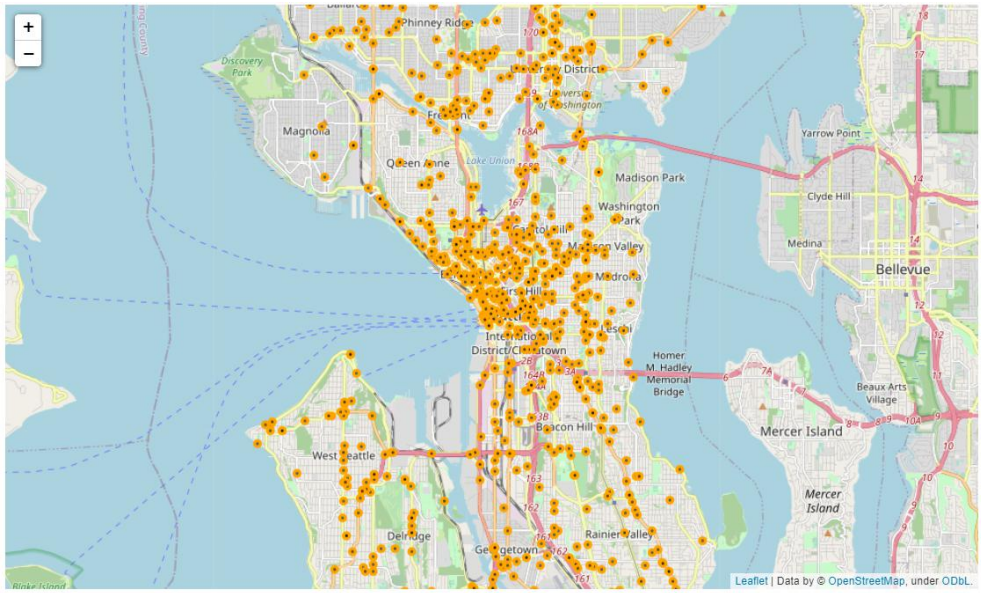
1. Accident location: Latitude("Y" column - float), Longitude("X" column - float)
2. Road coditions: "ROADCOND" column - text
3. Weather condition: "WEATHER" column - text
4. Junction: "JUNCTIONTYPE" column - text
5. Car speeding: "SPEEDING" column - boolean
6. Number of people involved: "PERSONCOUNT" 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

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

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

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).

We can use the Folium library to see the collision distribution on the map.



Further, 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.

```
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
```

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

Moving further we'll create to additional features for the dataset called "dayofweek" and "hourofday" from "INCDATE" and "INCDTTM" respectively. This will lead us to seperate the dependent and independent variables, and balancing the unbalanced dataset.

3.2. Model Building

It is very essential to build a model that can be used to further development of the project and end up giving useful results in order to predict the severity.

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

We'll start with this importing necessary libraries and splitting the dataset into train and test dataset.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_rus, y_rus, test_size=0.2, random_state=4)
```

And then we implement our models accordingly.

3.3. Model Evaluation

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. And the table below depicts the result.

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

3.4 Model Conclusion

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.

4. 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 of this documented report. This project is also going to help individual in determining the best model among chosen ones.