Applied Data Science Capstone Project

Data Science Report on Car Accident Severity

Introduction : Business Problem

This section contains a description of the problem and a discussion of the background.
With the world population well over the seven billion mark and high vehicular traffic, road accidents are very common. Often, there is a loss of property and even life. It would be great to understand the most common causes so that accidents can be prevented from happening.
An algorithm has to be designed to understand the dynamics of car accident data. For reducing car collisions frequency in a community, the algorithm must predict the severity of a possible accident taking care of the following conditions:

- current weather
- road condition
- visibility condition

The analysis will impart a thorough understanding of the factors and the correlations between them. Studying these relations shall have multiple applications. Devices connected by internet of things can collect information related to the conditions and they may be integrated them into an app.
On any particular day, the conditions can be used to alert drivers about bad conditions. By the virtue of this, drivers may practice an added carefulness during driving or the police may enforce more safety protocols.
The model described will be beneficial in reducing road accidents and the loss incurred because of that.

Data Understanding

This section contains a description of the data and how it will be used to solve the problem.
The dataset of size about 1,90,000 found in the file `Data-Collisions.csv` includes all types of collisions at Seattle from 2004 to present that are updated weekly. The data contains comma separated values having 37 attributes with the following highlighted attributes:

1. Text `ADDRTYPE` of length 12 describing collision address types.
    - Alley
    - Block
    - Intersection
2. Text `SEVERITYCODE` of length 100 that codes the severity of collision.
    - 3
    - 2b
    - 2
    - 1
    - 0
3. Text `WEATHER` of length 300 describing weather conditions during the collision.

4. Text `ROADCOND` of length 300 denoting road conditions during the collision.
5. Text `LIGHTCOND` of length 300 with details of light conditions during the collision.

The data in the csv file is not fit for analysis as there are multiple columns where the features do not have numerical type. Hence, label encoding is required for conversion to the desired data type.
We shall be using the following attributes for our analysis:

Attribute	Data Type
SEVERITYCODE	int64
WEATHER	category
ROADCOND	category
LIGHTCOND	category
WEATHER_CAT	int8
ROADCOND_CAT	int8
LIGHTCOND_CAT	int8

Methodology

The data analysis begins with normalizing the dataset and splitting it in a 70% training and 30% testing. Next, the modeling and predictions are operated on the models as below:

1. k-Nearest Neighbours
2. Decision Tree
3. Logistic Regression

Finally, the evaluation metrics are being used for accuracy testing of our models for:

1. Jaccard Index
2. F-1 Score
3. Logloss

Conclusion

Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).