Capstone Project - Accident predictor

Applied Data Science Capstone IBM/Coursera

Contents

Introduction – Business Problem

In this project we will try to determine or predict the severity of the traffic accident. Specifically, this solution or analysis will be targeted to the following stakeholders.

A. Mobile Map based applications

Users will be alerted of accident severity in the travel route based on various indicators

B. Vehicle insurance providers

This analysis may be useful for insurance providers to develop quotes based on statistics of accident severity and various indicators

C. Department of Motor Vehicles and other government bodies

This analysis can be used as input to post appropriate alert signage on roads. It can also be used to improve driving conditions and post appropriate speed limit / warnings.

Data

Based on the definition of our problem, factors that will influence our decission are:

- driving under the influence of alcohol or other substances
- speed of the vehicle
- weather / light conditions
- road conditions

For this project we will be using the sample data provided as part of this project

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Metadata is described here:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Methodology

In this project we will direct our efforts on detecting factors affecting the accidents, particularly those with high severity.

In first step we have collected the required **data: severitycode, incident date/time and various other attributes or factors related to the accident.

Of the various attributes availabled to us - these were identified for further analysis:

- Incident Date/Time
- Attention indicator
- Driving under influence indicator
- Weather Road condition
- Light condition
- Speeding indicator

Second step in our analysis will be to cleanse the data and filter out rows/columns with null values. More than 90% of the accidents did not have speeding / attention data and hence eliminated. Next we need to eliminate accidents where DUI was a factor. Weather, road and light condition attributes were converted to category object in order to facilitate analysis.

In third step we will focus on the filter attributes and split the data into test and training sets. We will apply the following **machine learning** algorithms on the training set:

K-means Clustering
Decision Tree
Logistic Regression

In the final step we will evaluate the machine learning models on the test set and evaluate the accuracy for the model with **jaccard index** and **f1 score**.

Results and Discussion

Our assumptions going into the analysis was bad weather, road / light conditions may lead to more severe accidents. We also thought certain months may have more accidents compared to the others. Some months may have high leisure travel - for ex: summer months and this could lead to more accidents. Also, some months may have higher count of accidents due to incline weather. Quick analysis for accident counts by severity proved that these counts were more or less even in the same ballpark range.

The original dataset has 195K observations. Data wrangling and cleaning process left us with 167K observations for analysis. We split the original data 70/30 ratio for training/test analysis. Our results did not improve with 80/20 ratio.

Our analysis evaluated various machine learning models such as **K-means Clustering**, **Decision Tree**, **Logistic Regression** to determine if accident severity can be predicted based on the factors such as **month**, **road conditions**, **weather conditions** and **light conditions**.

Jaccard index score was 67% - our test set and predicted result test matched to a reasonable degree. F1 score or **accuracy rate** of prediction was **54%**. Prediction accuracy using this analysis is not reasonable and may not be acceptable to the stakeholders.

Conclusion

Purpose of this project was to identify if there was certain factors which lead to more severe accidents. We used machine learning models to see if there is a pattern to predict the accident severity.

Our results were inconclusive and there were no clear indicators on whether certain factors lead to more severe accidents than others. We will need to revisit the original dataset and research other attributes/parameters to see if these results could be improved upon. We might also need to pull other relevant datasets for more detailed analysis.