US Accidents Analysis Project
By: Raneem Alhumaidan

Abstract

The project has two goals, first, it will strive for understanding the cause and effect rules of the accidents to analyze and generate insights on the traffic accidents and key factors affecting the accidents, also perform data visualization, to describe this data more vividly. Second, build several machine learning models: Accurately predict accident severity and predict street side of an accident in the United State, in order to help and understand to identify the patterns of how these serious accidents happen.

Design

As we all know car accidents are one of the biggest issues we have in this era, they can cause heavy injuries and deaths. Even more, it increases traffic jams dramatically thus creating more time and money loss. Therefore, data science has been considered the most direct and reliable way to attack a problem, tracing it to the root and predicting what and when the next consequences will take place. This project with a good dataset will follow the same direction and try to solve a specific real-world problem, data analysis can be an efficient method to extract useful information in order to figure out the cause and effect rules of the accidents, which will decrease number of accidents , that will help to avoid heavy losses and improve people's life experiences.

Data

The dataset was obtained from Kaggle, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2020, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. The dataset has about 1.5 million accident records and 47 features that contains information about the severity, start and end time of the accident, also had a more detailed about accident coordinate location, the length of the road extent affected by the accident, weather condition, visibility, humidity, wind direction, speed, shows the period of day (i.e. day or night) and the location (State, County, city, street name and number).

Algorithms

- ***Feature Engineering:***

  1- Create a feature that contains the time duration to clear an accident.[Time_Duration(mi)]

  2- Create a feature to check if the accident happened in the weekend or not.[Is_weekend]

- ***Models:***

  I built two models, one to predict accident severity and the other to predict accident street side, in both I have used Logistic regression, decision tree, and random forest classifiers before settling on the random forest as the model with the strongest performance. Also I have tried Random Forest with fitting important features only, and compared the results with the full features.

- ***Model Evaluation and Selection:***

  I took sample of my data because the CPU limitation, I chose California data which has more than 28 thousand records to work with it. The challenge is the data was implanced , to solve the problem I used class wight to balance it. Split it into 70% for training and 30% for testing. All the F1 scores of predictions for both severity and street side are reported it the table below:

Model 1: Predict Accident Severity

Full Features	Important Features
F1- Score : 0.61	F1- Score : 0.63

Model 2: Predict Accident Street Side

Full Features	Important Features
F1- Score : 0.67	F1- Score : 0.72

Tools

- Numpy and Pandas for data manipulation.
- Scikit-learn for modeling.
- Matplotlib, Seaborn, plotly, Bokeh, and Folium for visualizations.