Traffic Collision Data Exploration Report

Introduction about the project

This project is about the traffic collision， which causes serious impact on person’s and city traffic network on safety, also will increase the government capital & resource.

By studied and analysis the data of collision can help to get the trend of change on timing line, relevant properties and people, which can display information straightforward.

And This result can assist the government when they are planning the traffic network construction, arrange resource on traffic coordination in daily.

For public, people can also understand the potential risk distribution on collision occurrence in location and timing, so that they can have an alert in mind to pay more attention and avoid tragedy

Data source

The data for this project is from Kaggle(https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data)

- The data is based on Los Angeles America
- The data duration is from 2010 Jan to 2019 Jun
- The data structure is as below: (more detail and definition about the data, pls refer above link of Kaggle)
  - DR Number
  - Date Reported, Date Occurred, Time Occurred
  - Area ID, Area Name
  - Reporting District
  - Crime Code, Crime Code Description
  - MO Codes
  - Victim Age, Victim Sex, Victim Descent
  - Premise Code, Premise Description
  - Address, Cross Street, Location, Zip Codes
  - Census Tracts
  - Precinct Boundaries
  - LA Specific Plans
  - Council Districts
  - Neighborhood Councils (Certified)

Methodology

Firstly, do the data cleansing on below items:

- Drop unnecessary columns
- Derive additional columns from original ones for convenience further classification
- Handle NaN and NULL value in the records

Secondly, draw plot to get the trend of collision change on

- Time line
- Victim
- Location

Third, to pick up the high-risk location (accidents > 10 over 9 years) and low-risk location (accidents <= 10 over 9 years), and do logistic regression base on the venues nearby with these location point

And get the model for prediction

Result

From the time line the highest risks of collision occurrence are

- Each Oct of year
- Each Fri of week
- During 15:00 - 18:00 of each day, and at weekend the risk at late night and mid-night are same risk as that duration

From the victim information, we can aware that

- Male is higher risk than female
- Young people is higher risk than older people

The model finale can predict around 64% accuracy with the verification data

Observations

Base on the result, as LA is densely populated city and caused the accident increased to high level at rush hours after work in working days, and as LA is a nightlife and tourist city, so the accident occurrence in relatively high level at weekend days is no supervise

For another hand, from victim information, perhaps male and young people are more aggressive than others (female and older people) and these people are easier to violate the traffic rules and have quarrel no matter as driver or passenger which will increase the higher chance on accident

Discussion

In this project, so far only apply logistic regression on prediction, can try below items in future:

- To classification for the venues by names, so that can get more clearly feature value
- To do the train/test data in folds, so that can decrease the possibility of overfit
- To try other algorithms like decision tree, SVM and see whether higher accuracy can get