Analysis and Forecast of Road Accident

1. Introduction

1.1 Background

With the rapidly development of economy, the number of vehicles and travelers continue to grow. Road traffic safety becomes a critical factor that influences the lives and society property heavily. It would bring a huge negative impact on the development of society if the we ignore this topic.

1.2 Problem

Road Accident is usually determined by multiple factors. Weather condition, road type, vehicle type and many other factors can contribute the results. This project aim is to predict the possibility of the condition or condition group that will cause the road accident.

2. Dara acquisition and clearing

2.1 Data Source

The data for this project is found in Kaggle, published by SDOT Traffic Management Division, Traffic Records Group. The link is here. The dataset contains a detailed csv file Data-Collisions.cvs. Each row represents a unique accident, and each column represent a possible factor that would impact the accidence. There are total 37 attributes in the file.

2.2 Data Cleaning

After checking the data, I found there are several problems with datasets.

Firstly, several features have missing values: ADDRTYPE, LOCATION, UNCTIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, SPEEDING and ST_COLCODE. There are blank cells under these features. Because they contain qualitative information, I use UNKNOWN to replace blank cells

Secondly, the format of several features is not consistent or is not appropriate. For INCDATE, because the time is the same for all samples as 00:00:00+00, I remove all the detailed time, only keep date. For INCDTTM, some samples have detailed hours and minutes, but some are not. The time is worth to consider as an important factor that will affect the road safety, and the samples that are missing time is not a big volume. Therefore, I add 00:00 to the samples which miss detailed time. UNDERINFL is another feature has format problem. It use 0,1,Y and N. To unify the data, I convert Y to 1 and N to 0.

2.3 Feature selection

After cleaning and observing the data, I found a lot of repeat information and choose to abandon. For example, EXCEPTRSNCODE and EXCEPTRSNDESC. They indicates that the samples don't have clear location information, but UNKNOWN in LOCATION can help. Same logic, INCDATE is also repeated.

I also abandon X,Y, INCKEY, COLDETKEY, SEGLANEKEY, CROSSWALKEY, SDOTCOLNUM and HITPARKEDCAR because they only contain some useless code or ID, and don't have too much help for our analysis.

After discarding unnecessary features, total 27 features are selected.