CAPSTONE PROJECT

CAR ACCIDENTS SEVERITY PREDICTION

BY: CHENG YAT YEUNG

Index

1. Introduction
2. Data Description
3. Data Preparation
4. Data Analysis and Visualization
5. Conclusion

1. Introduction

Background Description

Traffic collision, also called car accident, is one of the most common accidents in modern society. It often results in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved. In 2013, 54 million people worldwide sustained injuries from traffic collisions, and also caused 1.4 million deaths. Therefore, it raises a serious concern for many citizens around the world.

Our motivation is to predict the accident severity of any road, which will play a crucial factor for traffic control authorities to take proactive precautionary measures.

Problem Description

Car accidents can take place due to serval reasons, for example, vehicle design, driver negligence, road environment, traffic condition etc. Therefore, it is important to predict the severity of the car accidents under different conditions. And the purpose of this project is to predict the severity of an accident, the cause for these accidents and suggest how to reduce the number of accidents by training an efficient machine learning model with the help of existing accidents data.

2. Data Description

The dataset used in this project contains data about car accidents provided by SPD and recorded by Traffic Records. The dataset provides several attributes such as the weather during the time of accident (WEATHER), road condition (ROADCOND), visibility of the area (LIGHTCOND) and type of road junction (JUNCTIONTYPE). The label for the data set is 'SEVERITYCODE' (Target variable), which describes the fatality of an accident, it can be measured and predicted the severity of an accident based on a scale of 0-5. 0: Little to no Probability (Clear Conditions) 1: Very Low Probability – (Chance or Property Damage) 2: Low Probability – (Chance of Injury) 3: Mild Probability – (Chance of Serious Injury) 4: High Probability – (Chance of Fatality)

3. Data Preparation

Main objective of this step is to get the pre-selected variable for machine learning. It includes the steps Exploratory Data Analysis, dealing with missing values, dropping features and converting the data types.

These data mainly come from two sources MapQuest and Bing. We are trying to understand the severity cases provided by each source. MapQuest reported less accidents with severity level 4 which cannot be seen in the plot itself, whereas Bing reported almost the same number of level 4 accidents as level 2.

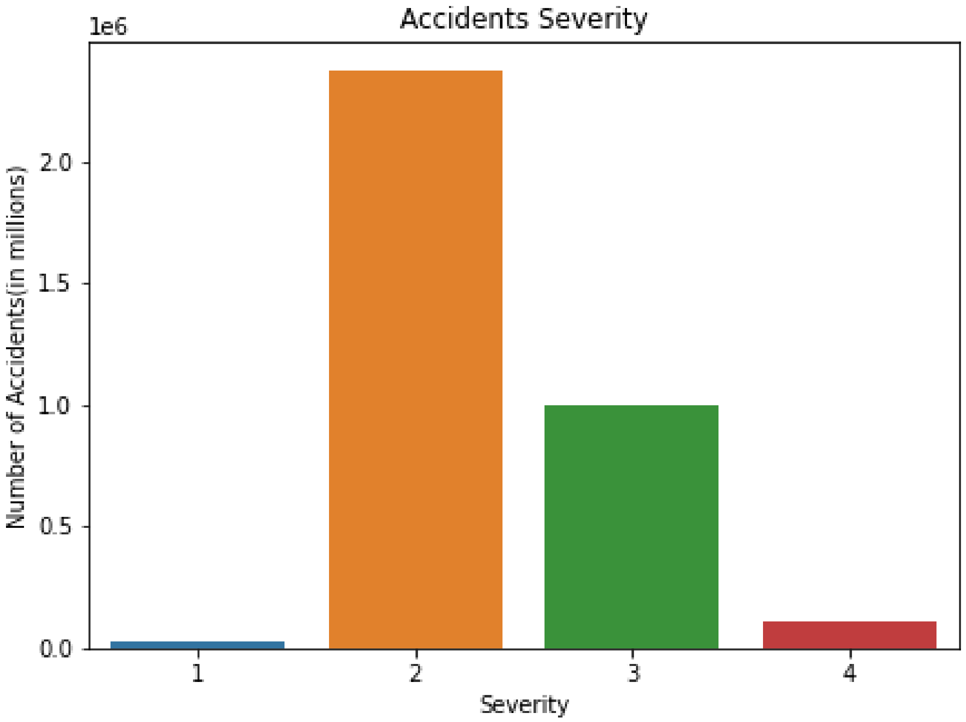
Converting Start Time, End Time and Weather Timestamp to the real date time columns. Also, simplifying the Wind Direction and Weather Condition features to avoid complications.
Also splitting up the Start Time feature into Day, Month, Year, Weekday, Hour and Minute. All the POI and POD features data types are modified to Integer ditypes for our convenience.

Dropping the few rows as they are less in count compared to the total sample values. City, Zip code, Airport Code, Sunrise Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight and etc...

4. Data Analysis and Visualization

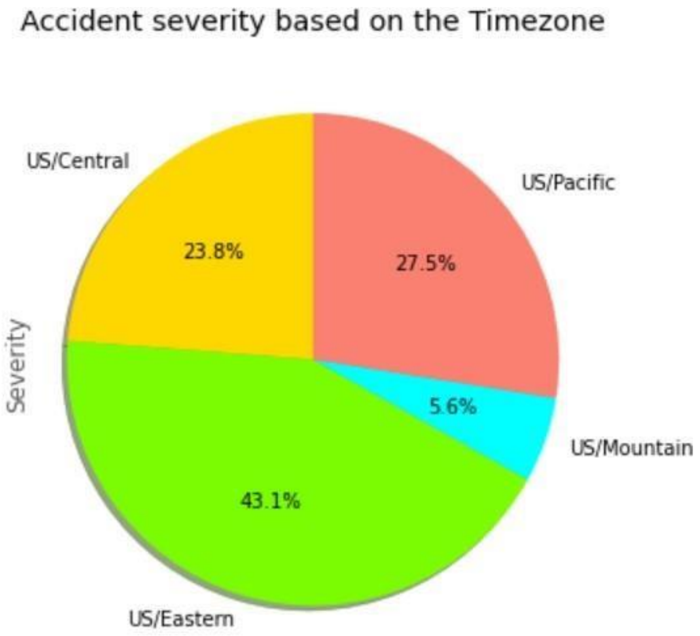
We are performing the analysis of the available dataset using different features. Initially we are just identifying the how many accidents were happened on each severity which is the dependent variable and the values we have to predict using classification algorithms.

Figure 1: Severity Distribution



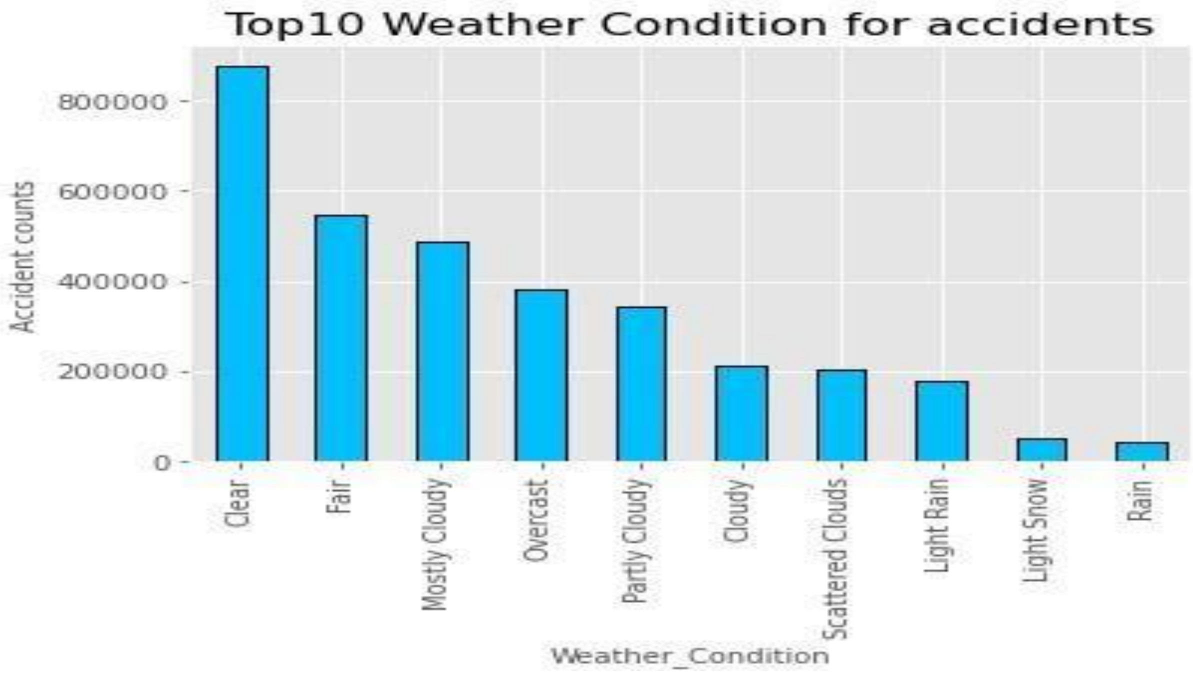
Most accidents happened during the daytime, especially AM peak and PM peak. When it comes to night, accidents were far less but more likely to be serious.

Figure 2: Severity based on Time zone



Most of the accidents occur Clear and Fair conditions.

Figure 3: Highest Weather condition for Accidents



Based on the above visualizations of US Accidents dataset after Exploratory Data Analysis highly impacted Features with values are:

- Weekdays other than weekends
- US Eastern Areas
- Right side of the Roads
- Daytime
- Clear and Fair cloud condition
- Wind is Calm and Western Side

5. Conclusion

This section summarizes the findings on Accident Severity Prediction obtained from both EDA and the Data Analysis section of this report. Above analysis are clearly providing the counterintuitive answers to the questions. Our Analysis mainly demonstrate the top categories of the Weather conditions, Time Features, POI features shows that Clear, Calm, Traffic signal, Eastern areas and Sunlight are the ones with high accident occurring features even though the weather conditions were bad, Heavy Storm and visibility also poor.