ORIE5741 Project Proposal

Muyuan Liu(ml2699) Duanduan Zhu(dz223)

October 3rd 2021

1 Problem and Objective

While accidents cause traumas and tearful moments, they also inevitably cause traffic distress which can lead to potential economic loss. From minor congestion to highway closure, one effective way to deal with such issues is to predict and prevent them. Our team aims to predict car accident severity based on the contributing factors. Consequently, we hope to be able to answer the questions such as "will thunderstorms cause severe accidents?", "What kind of roads are the most dangerous?", etc... A solution would be especially beneficial for traffic safety authorities because they can prevent accidents from happening by eliminating the risk factors. For instance, they can fix the road conditions that are deemed dangerous by the model. Moreover, they can also prevent accidents from making a prolonged impact by predicting accident severity and preparing emergency reaction plans accordingly, in order to achieve faster resolution and limit economic loss.

2 Data Description

Our analysis will use the "US Accidents" dataset. It contains 1.5 million accident records in 49 states of the US from February 2016 to December 2020. Accident records consist of only traffic features rather than driver and vehicle information. More specifically, the 47 features in the dataset cover 5 aspects including the weather, accidents sites, accident time, general locations, and the affected traffic. Our model intends to predict accidents severity using the other 46 explanatory variables. And the outcome "severity" is originally defined as the traffic delays caused by the accidents and takes ordinal values of 1 to 4 where 4 is the most severe. The rest 46 features contain various data types, missing values, and have unique encoding rules and meanings. Cleaning and understanding these features will be the primary focus of our next stage.

3 Why is the dataset useful?

The dataset contains sufficient observations and dimensions, which allow us to conduct in-depth exploratory data analysis from multiple perspectives and detect patterns and abnormalities. We plan to summarize the accident severity and distribution by locations, weather, on holidays, or during certain time periods such as the outbreak of the COVID-19 pandemic. Next, we intend to identify key factors by feature selection and engineering. Last, we will construct regression and other predictive models and train and test them on this large enough dataset, ensuring prediction performance and comprehensive evaluation.

4 Reference

Dataset and Descriptions: `https://smoosavi.org/datasets/us_accidents`