Data Description:

The data used in this project is a public dataset shared by Sobhan Moosavi, a research scientist at Lyft. The whole dataset contains traffic accident records from Feb. 2016 to Jun. 2020, covering 49 states of the United States. These traffic data were collected by two APIs that broadcast traffic events captured by a variety of entities: US departments of transportation, law enforcement agencies, traffic sensors etc. A subset of the dataset that contains only accidents that happened in Chicago was used for this project. There were 42 features in the original dataset, but after careful feature engineering and feature selection, only 15 features were kept for further analyses. The data can be categorized into three types:

1. Geospatial data:

The data includes the GPS coordinates of the start and end point of the event. Moosavi and the fellow researchers also converted the GPS data into augmented addresses consisting of street number, street name, city, country etc. using Reverse Geo-Coding. This specific street information is extremely important to this project as it is used to create a new feature summarizing the type of street where the event happened.

1. Environmental data:

The environmental data is mostly weather information that can provide context for the accidents, including temperature, humidity, pressure, precipitation etc. The weather information is recorded within 15 minutes from the start time of the accident.

1. Points of interest:

POI are locations annotated on a map as amenities, traffic signals, crossings etc. These POI data is coded in binary form and should give us a sense of external traffic complication around the accident, that might augment or alleviate the traffic impact.