Intro (Topic and reasons)

In 2019, U.S. traffic crashes cost employers $72.2 billion in direct crash-related expenses which include medical care, liability, lost productivity and property damage. It means that reducing the number of car crashes is an important task. The proactive approach to decreasing car accidents include preventing potential unsafe road conditions and making people aware of factors increasing the probability of getting into car accident. One way to implement this approach is to create car accident prediction and crash severity models. Knowing crucial factors leading to car crashes and geography of those, authorities might be able to distribute financial and human resources better.

Dataset description

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data was collected from February 2016 to Dec 2020, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

The link to the dataset <https://www.kaggle.com/sobhanmoosavi/us-accidents>

Acknowledgements

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “[A Countrywide Traffic Accident Dataset.](https://arxiv.org/abs/1906.05409)”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "[Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.](https://arxiv.org/abs/1909.09638)" In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Objectives

The main objectives for this research are to create a machine learning model that could predict car accident severity and to find the key factors affecting the accident severity. Along with that we could find top 5 places in the US with the highest number of accidents.

For this dataset the Logistic Regression will be used as a main tool to create a prediction model.

Severity in this project is an indication of the effect the accident has on traffic, rather than the injury severity

Technologies we plan to use:

Pandas, Seaborn, Matplotlib, Numpy, PostgreSQL, Tableau

Project Team and responsibilities:

Deepali Bhardwaj – GitHub

Aleksei Pronin – Machine Learning

Sandip Patel – Database

Team members decided to meet up on a daily basis in Zoom at 6:45p.m. EST.