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

Questions we hope to answer:

* Is there a correlation between severity and other available features (like temp, visibility, day or night etc.)?
* What are the top 10 States and top 10 Cities with most accidents?
* Is there a trend over time (day, month, year)?
* What time do most of the accidents occur?
* Identify Red , YELLOW and GREEN  Accident zone in US
* Monthwise Accident analysis from 2016 to 2019 ?
* In which year had more accidents in US?
* How much weather responsible for accidents?
* On which location we have to take care to avoid accident during our next journey between two cities of US?

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.

Data Exploration phase

At this phase the data was analyzed for null values, data types and number of unique values.

It was found that data set had multiple data types that later were converted to suitable ones for machine learning model. Some features had a lot of missing values up to 69%, so they were dropped and for the rest of the features we decided to drop rows. Initial visualizations were done before transforming the data and after transformations. You can see them all in the “EDA.ipynb” file located in the repository and also in Google Slides by clicking the provided link.

The major findings are: dataset had a lot of missing values; features had different data types; Los Angeles is the city with the biggest number of accidents; the most number of all accidents per city is less than 2500; the most severe accidents make only 7,5% of all accidents; there are 2 peaks with bigger number of accidents during the day - morning peak 8-9a.m. and afternoon peak 4-6p.m.; on the weekend the number of accidents decreases almost by half; December is the month with the biggest number of accidents; more than 85% of accidents happen on the right side of the road; most of the accidents happen during fair weather.

Machine Learning Phase

One of the main questions was about correlation between severity of car accidents and other features. The Logistic Regression model was chosen for this project because it is a classification algorithm used to find the probability of event success and event failure.

Logistic Regression has both advantages and disadvantages:

Advantages:

* It is easier to implement, interpret, and very efficient to train
* It makes no assumptions about distributions of classes in feature space
* It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative)
* It is very fast at classifying unknown records
* Good accuracy for many simple data sets and it performs well when the dataset is linearly separable
* Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets

Disadvantages:

* The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables
* If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting
* It constructs linear boundaries
* It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set
* Logistic Regression needs independent variables to be linearly related to the log odds (log(p/(1-p))

Data preprocessing and feature engineering and selection:

Since our future machine learning model is based on predicting severity, we decided to regroup severity from four categories into two by labeling the most severe one with 1 and the rest of categories with 0. The regrouped data was imbalanced and the dataset was big enough so we undersampled the 0 value and oversampled 1.

Additional visualizations were made to show correlations with the regrouped data. Based on plots, a part of the road features, like 'Amenity','Bump','Give\_Way','No\_Exit', 'Railway','Roundabout','Traffic\_Calming' were dropped as they had no effect on the number of accidents. Also, features with only one category were dropped as they have no value for our analysis. There was inconsistency in Weather and Wind features, so we cleaned them and regrouped using .loc function and regular expressions. The next step was to create new weather features and drop the original one. 69% of data in 'Number' and 29% of 'Wind\_Chill(F)' was missing, so it was decided to drop these columns. Although 'Precipitation(in)' also had a lot of missing values, we decided to leave this column as rain may change road conditions and potentially affect the number of accidents. In the rest of the features the number of missing values was insignificant, so we dropped missing values.