Introduction to Data Mining

DATS 6103 – Spring 2020

PROJECT PROPOSAL

US Reported Accidents (Visualization and Summary)

Submitted by:

Daniel Frey

Gregg Legarda

Siwei Wang

Submitted to:

Dr. Amir Jafari (PhD.)

Problem

We have a huge dataset of 3 million observations for car accident reports in the US. What does it mean and how can we make sense of this data? We cannot interpret anything just by looking at this huge dataset. We want to find out the relationships and patterns of this data to help us understand the underlying meaning.

Solution

We will be using data mining techniques such as preprocessing, cleaning, data analysis and modeling. For preprocessing, we will: remove unnecessary columns and take random sampling of data; clean the dataset by formatting and removing observations with missing data; perform data analysis such as EDA (mean, sd, median, min, max, quartiles); and use modeling techniques such as regression, decision trees, or random forest as needed. We will also use data visualization such as histograms, scatter plots and charts to help us understand the meaning behind this data.

Dataset

We will use a dataset obtained from Kaggle titled: “A Countrywide Traffic Accident Dataset (2106 - 2019)”. According to the description, “This is a countrywide car accident dataset, which covers 49 states of the United States. The accident data are collected from February 2016 to December 2019, using several data providers, including two APIs that provide streaming traffic incident 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. Currently, there are about 3.0 million accident records in this dataset.”

Algorithm

It is still to be determined. We will likely use classification or clustering algorithms.

Softwares

We will use Python as our programming language. For our data management, we will use Pandas, Numpy, and Matplotlib. For our visualization we will use Python-compatible software such as Leaflet, Geopandas and Plotly. For our application, we will likely use PyQt5 and Plotly Dash.

Performance

Performance will be determined from the accuracy of our model. We will determine the model used soon.

Schedule

Group Discussion	Completed
Research possible ideas	Completed
Pick datasets & topic	Completed
Research resources, determine software & packages needed	Completed
Submit group proposal	April 3
Pre-process data (clean up and formatting)	April 4
Perform data analysis	April 5
Build front end GUI	April 10
Integrate results to GUI	April 10
Build geographical maps	April 10
Integrate maps to GUI	April 15
Plot in data points to geomaps	April 15
Implement chosen modeling technique	April 15
Integrate techniques to GUI	April 20
Add extra features	April 20