Proposal for Capstone Project

Title: Predicting the severity of a motor vehicle collision

Problem: The safety of drivers, passengers, pedestrians and property is an ever-growing concern when it comes to the operation of vehicles on the road. With 6,452,000 motor vehicle crashes and 37,133 crash-related deaths in 2017 alone, significant costs in terms of life, money and property are incurred as a result of collisions each year. To make things more complicated, the number of vehicles in the US keeps growing every year, creating a need to adopt greater safety measures on the road. Insurance companies play a significant role in minimizing and handling automobile collisions, and they do so through services provided by charging insurance premiums. These insurance rates could be more accurately predicted based on factors such as location, weather, road conditions, light conditions, population, collision date/time, whether the driver was DUI etc., most of which are dependent on the location. We propose to use collision data from the city of Seattle to implement a machine learning model to predict the severity of collisions using the above variables and others.

Audience: Some of the biggest insurers in the US such as Statefarm, Esurance and Allstate have stated location as one of their top criteria for determining rates. These and many other insurance companies could use this model to determine more accurate rates for their customers. Moreover, they could also provide useful information to their customers such as looking out for red flags when buying expensive cars in more accident prone locations which would likely increase their insurance premiums. Another audience for this model could be rideshare companies as well as Google Maps and Apple Maps. Also, traffic control departments could collect this data firsthand and make it available to the aforementioned companies.

Data: The data will be acquired from the City of Seattle Open Data Portal and consists of 212,760 instances of vehicle collisions in Seattle while the timeframe of the data spans from 2004-2018. A number of collision event features such as date of collision, location of collision, type of collision (sideswipe, rear-end, parked car etc.), number of people involved, number of fatalities, junction type (intersection, driveway junction, mid-block etc.), weather conditions, road conditions, light conditions, collision description etc. are described which will be used to predict the level of severity of a collision. In total, 40 variables are included in the dataset but this number will be reduced to include only the ones crucial for building the model. As location would play

an important role in the modeling process, addresses for each instance will be extracted using Python's *reverse_geocoder* library and added to the dataset.

Modeling: A multi-label classification approach will be implemented in order to predict the severity of a collision. There exist 5 categories of severity - fatality collision, serious injury collision, unknown, injury collision and property damage collision, in decreasing order of severity. Due to property damage covering a majority of the collisions (62.21%), it will be interesting to create a model that accurately predicts each level. Injury, unknown, serious injury and fatality collisions make up 26.46%, 9.77%, 1.39% and 0.15% respectively. Different classification models will be tested and evaluated to determine the one which provides the best results based on the chosen metric.

Challenges: Since the data consists of instances only within the city of Seattle, it will be difficult to generalize the model results to a larger area such as the US. Although, similar modeling techniques could be deployed to country-wide data or data from other states.

Deliverables:

- Jupyter notebooks (code):
    - Data acquisition and cleaning
    - Exploratory data analysis (EDA)
    - Modeling and prediction
- Project Report