**Introduction**

Road traffic accidents are common safety problem around the world. This automotive accidents result in over result in over 30,000 fatalities in the United States annually. Road traffic accidents are estimated to cost the US economy approximately $810bn per. Identifying the factors which influence accident severity is therefore of paramount importance.

In an effort to reduce the frequency of car collisions in our community, this project will leverage existing accident data to predict the different accidents' severity given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.I will use different supervised machine learning algorithms and select the machine learning model that gives the highest prediction accuracy.

The study of building this kind of model will be of significance to a lot of stakeholders and beneficiaries: (1) town/city planners, who may be able to use the model to inform their road planning and traffic calming strategies; (2) emergency service responders, who may be able to use the model to predict the severity of an accident based on information that’s provided at the time the accident is reported in order to optimally allocate resources across the city, and (3) traffic police officers.

**Data**

A comprehensive dataset of over 190,000 observations collected occurring between 2004–2019 in the Seattle city area was obtained from the Seattle Open Data Portal. The dataset has almost 40 columns describing the details of each accident including the weather conditions, collision type, date/time of accident and location. To accurately build a model to prevent future accidents and/or reduce their severity, we will use the following attributes — ADDRTYPE, WEATHER, ROADCOND, VEHCOUNT, PERSONCOUNT.

**Methodology**

exploratory data analysis

inferential statistical testing

machine learnings - why

**Methodology**

I used Jupyter Notebooks to conduct that analysis and imported all the necessary Python libraries like Pandas, Numpy, Matplotlib, and Seaborn. The data was mostly categorical so I stuck to graphical representation to see correlation between various variables.

I started by importing the csv file and to prepare the data, I dropped the columns we do not need from the dataset, i.e., columns that do not have values or where the values are unknown. Even though this is an important factor, I dropped Speeding entirely because it is missing over 180,000 values and this can hamper the results.

Then, I began choosing columns to use from the dataframe that I created. The columns that I chose were SEVERITYCODE, which assigns a crash a value of 1, which means no injury, and 2, indicating injury, COLLISIONTYPE, which describes the type of crash, WEATHER, which describes the weather at the time of crash, ROADCOND, which describes the condition of the road at the time of crash, LIGHTCOND, which describes the light conditions at the time of crash, INATTENTIONIND, which describes whether the driver was distracted, and UNDERINFL, which describes whether the driver was under the influence.

Lastly, I visualized the data in the form of bar graphs. I filtered out the columns I wanted from the provided .csv and then called value\_counts to graph the mostly categorical data.

**Conclusion**

Most crashes happened in clear, dry, and bright conditions. Most days are clear, dry, and bright, so it’s no surprise that most car crashes occur under these conditions. I also found out that crashes with a distracted driver or an impaired driver are statistically more likely to result in injury, which is also not a surprise. The results of the data indicate to city officials that they should ask drivers to be more alert in ideal conditions