1. Introduction / Business Understanding

1.1 Background

Motor vehicle accidents statistics in the United States are sobering with over 36k fatal crashes and 2.7 million people injured in 2018 alone[1]. Motor vehicle traffic crashes are the leading cause of death for youth (16-20) and young adults (21-24)[2] in the United States. The economic and societal costs are enormous, exceeding 836 billion US dollars in 2010. New methods to help reduce the number of accidents can go a long way in saving lives and the economic costs associated with these crashes.

1.2 Business Problem

Finding ways to reduce accidents has a vast array of interested stakeholders: government, police, first responders, hospitals, insurance companies, parents, drivers, passengers, users of public transportation, etc.

The city of Seattle, Washington, USA has a Department of Transportation (SDOT) which collects and publishes accident data. We will use this data to try to create a model to predict accident severity which in turn can be used by SDOT to help alert stakeholders that there are conditions present which are highly associated with collisions.

SDOT can warn stakeholders via mobile text alerts, Twitter, mobile app alerts, electronic signage on the roads or publishing to news organizations. SDOT can also alter speed limits if necessary, suggest an increased police presence in an area, and warn first responders and hospitals so they can be prepared when conditions favor an accident they will need to be involved in.

2. Data

2.1 Data Source

The data used in this model comes from the Seattle Department of Transportation's Open Data Portal:

https://data.seattle.gov/

Will we be using the collision data which can be found below but was provided by Coursera for use with this project.

https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

Metadata for use in understanding all the attributes can be found below but has been provided by Coursera for use with this project:

https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

[1] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812948
[2] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812951

2.2 Data Summary

The data set is of an adequate size having more than 194k rows (collisions) and 37 attributes for each accident and is recent, with entries from January 2014 to May of 2020. It also contains the value will be trying to predict in SEVERITYCODE

SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown

2.3 Data Cleaning

While the data is large it will need to be cleaned up to be used in our model. Many of the collision records contain missing values, so they will be removed. Some of the attributes are not useful in helping to predict accidents, so we will remove them. The data set also has a bias towards one type of accident (Severity 1), so we will resample to reduce the bias.

2.4 Feature Selection

Correlation analysis will be performed to see which attributes contribute the most to the SEVERITYCODE.