IBM Applied Data Science Capstone Report

1. Introduction

1.1 Background

Car accidents are a very significant cause of serious injury and death in the United States. According to the National Safety Council, 38,800 people lost their lives in car crashes in 2019, and about 4.4 million people were in accidents severe enough to require that they seek medical attention for their injuries. Some of the individuals will suffer from permanent disability. This results in significant distress, decreased quality of life, and financial burden for many individuals every year. It also has greater societal impacts in the form of lost productivity and subsequent impacts on the economy, as well as heavier burdens on the medical system.

There are a great number of factors that may contribute to whether an accident will be severe or minor. These might include, but are not limited to, road conditions, light conditions, the weather, driver inattention, driver intoxication, speeding, and the area in which people are driving. This project seeks to determine if there are any reliable indicators that particular conditions significantly elevate the risk of getting into a more serious car accident – one that involves bodily injury rather than just property damage. This project will focus on one area, Seattle, and attempt to determine if there are significant risk factors for car accident related injuries that may be addressed so as to improve outcomes.

1.2 Interest

There are numerous parties who may have an interest in this type of assessment. From a public health standpoint, government bodies could use the information gathered from this study to determine the best changes that can be made to infrastructure in order to reduce fatalities and serious injuries (such as increasing lighting in certain areas, dealing more aggressively with poor road conditions, such as heavy snow, or modifying how streets or intersections are constructed), or to determine the best driver behaviours to target with policing or public education efforts (such as distracted driving, intoxicated driving, or speeding).

Private companies may have an interest in the assessment as well, as the information gathered here may allow for the development of consumer products aimed at improving driver safety. The manufacturers of GPS units in vehicles, or tech companies that provide driving and GPS related products (such as Google Maps), could use this type of information to provide added value to their customers. This could look like expanding the functionality of GPS units to include current weather, road conditions, lighting conditions, or other information that would output an accident risk assessment to the user, and provide the user with guidance on how they should proceed according to the current risk level.

2. Data Acquisition and Cleaning

2.1 Data Sources

I opted to use the provided data set for the course, as I am very new to data science and coding, and wished to keep things relatively simple so that I could focus on the rather large number of new skills I need to practice. As such, the data was acquired from the IBM Applied Data Science Capstone course content. Additional information on the data set was acquired from the Seattle

GeoData website (https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0).

The dataset originates from Seattle Department of Transportation, and represents data from 2004 to the present. It provides 194,673 observations, with most observations having 38 attributes. The attributes are diverse, and include things such as the weather, road conditions, lighting conditions, whether or not a driver was speeding, the number of people involved, the seriousness of injuries, details of how the accident occurred, where the accident occurred, and much more.

2.2 Data Cleaning and Feature Selection

My goal for this project was to focus on attributes that are most easily observable and/or modifiable prior to an accident occurring. As such, I decided to use weather, road conditions, light conditions, speeding, inattention, being under an influence, and address type (intersection, block, or alley). Other attributes were omitted because:

1) They were too similar to the outcome being predicted (accident severity), such as injuries, serious injuries, fatalities, and collision type.
2) They contained information that could only be determined after the accident occurred, and would be unlikely to be helpful in the goal of reducing accidents through public health or technological intervention.
3) They were simply not useful for my intended approach (e.g. GPS coordinates, ESRI identifiers, etc.)

I considered including dates and times, however the time is missing in many of the timestamp observations. I may decide to include the day of the week as a variable, once I have been able to spend more time assessing the data.

For attributes with a binary classification, i.e. driver inattention and speeding, null values were assumed to be “no” (values were either null or “Y”), and the data was updated accordingly. Null values and values indicated as “unknown” for all other attributes resulted in the observation being removed, as there was no reasonable means of filling in the data with a mean, or another similar measure. This is primarily due to the fact that all the attributes being assessed contain categorical values. At this time, the number of observations being utilized has been reduced to 169,578, with 8 attributes (including the target attribute) associated with each observation.

The target variable, accident severity, is limited to either a “1” or “2” in the data set, despite the fact that the accompanying documentation indicates 5 potential severity codes. This is therefore a binary outcome. As such, solving the desired problem requires a predictive classification model, which necessitates a machine learning approach. Many of the machine learning approaches covered within the certification courses may be applied, including K-Nearest Neighbours, Decision Trees, Logistic Regression, and Support Vector Machine. I intend to assess the accuracy of each prior to determining which would be best suited to the task of predicting accident severity. Measures of accuracy such as Jaccard Index, F1-score, and Log Loss will be utilized for this purpose.