Accident Severity Prediction

Siddharth Gangwar

September-2020

Section 1: Introduction

Background:

Seattle, also known as the Emerald city, is Washington State's largest city, with home to a large tech industry with Microsoft and Amazon headquartered in its metropolitan area. As of 2020, it has a total metro area population of 3.4 million (www.macrotrends.net). The total number of personal vehicles in Seattle in the year 2016 hit a new high of nearly 444,000 vehicles. In one South Lake Union census tract, the car population has more than doubled since 2010.

Objective:

The purpose of this project is to predict the severity of an accident by training an efficient machine learning model with the help of existing accidents data from 2004-20. This project is majorly focused on predicting rarer classes accurately such as Serious and Fatal.

Problem:

The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. The project aims to predict how severity of accidents can be reduced based on a few factors.

Stakeholders:

The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

Section 2: Dataset

The dataset used for this project is based on car accidents which have taken place within the city of Seattle from the year *2004* to *2020*. This data is regarding the *severity of each car accidents* along with the time and conditions under which each accident occurred. The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of *1* (Property Damage Only) and *2* (Physical Injury). Following that, *0* was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity.

Feature Selection:

Feature Variables	Description
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)

Section 3: Methodology

Exploratory Analysis:

Considering that the feature set and the target variable are categorical variables with the likes of weather, road condition and light condition being above level 2 categorical variables whose values are limited and usually based on a particular finite group whose correlation might depict a different image then what it actually is. Generally, considering the effect of these variables in car accidents are important hence these variables were selected. A few pictorial depictions of the dataset were made in order to better understand the data.

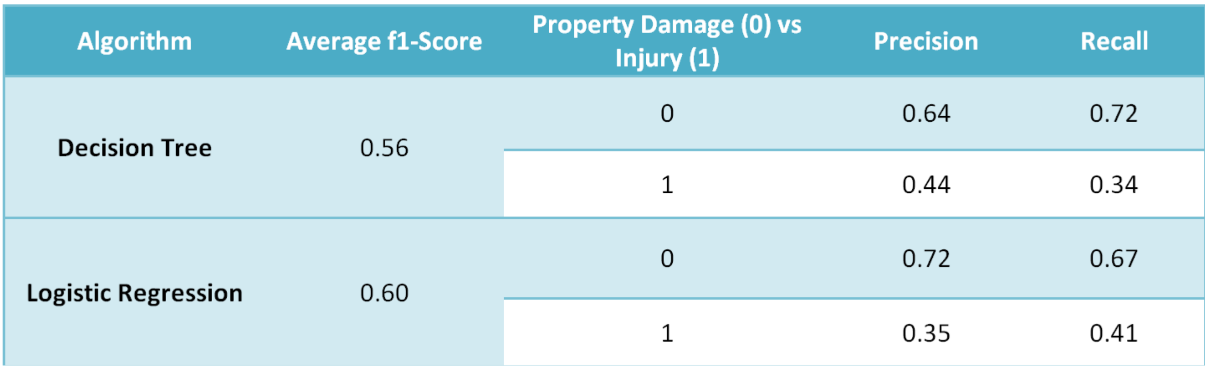
Machine Learning Models:

- **Logistic Regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable

- **Decision Tree Analysis:** The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Results:

The results of each of the three models had variations among them, one worked very well at predicting the positives accurately while the other predicted the negatives better.



Algorithm	Average f1-Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic Regression	0.60	0	0.72	0.67
		1	0.35	0.41

Recommendations:

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.