PREDICTING THE SEVERITY OF AUTO ACCIDENTS

Henry Chikwendu

August 23, 2020

1. INTRODUCTION

1.1 Background

Vehicular travel is an essential to the global economy as workers commute to work, families go on vacations trips, the United States Postal Service delivers mail via road networks across the country in automobiles, even food supply chains connecting from farm lands to your local supermarket require vehicular travel to sustain demand and the list is endless, indeed vehicular travel is essential for humanity in the 21st Century.

However one of the biggest challenges to this means of transport is safety. Annually approximately 1.35 million people die in auto accidents with an average of 3,700 deaths per day worldwide while more than 38,000 people die from auto crashes on US highways annually. Auto crashes are the leading cause of death in the United States amongst people aged 1-54.

Therefore it is paramount that Safety on these roads cannot be over emphasized which can be done by understanding the factors that contribute to these crashes, proper safety measure can be put in place.

1.2 Problem

We have a dataset which we will analyze and derive insights and also try to predict how severe road accidents could be by taking into consideration some contributing factors in other to assist the likes of Health-care workers and emergency services to help to target key areas prone to severe road accidents take action more quickly potentially saving lives.

2.DATA ACQUISITION AND CLEANING

2.1 Data Sources

The dataset was provided by seattle.gov: Link for road Accidents from 2004 to 2020 within the city of Seattle, Washington.

2.2 Data Description

The data consists of 194, 673 samples and 37 features which consists of attributes like weather conditions as at when some accidents happened, the number of persons involved in the accident, the Longitude and Latitude of the collision which helps locate the exact location of these collisions. Below are the key features used in building our model to predict the severity of accidents in the city of Seattle.

SL No.	Feature	Description	Reason for Selecting
1	ADDRTYPE	Collision at Alley, Block, Intersection	Gives the likelihood of collision at these places
2	PERSONCOUNT	Number of people involved in the collision	Gives an indication of severity
3	PEDCOUNT	Number of pedestrians involved in the accident	Gives an indication of severity
4	PEDCYLCOUNT	Number of cyclists involved in the accident	Gives an indication of severity
5	VEHCOUNT	Number of vehicles involved in the accident	Gives an indication of severity
6	INATTENTIONIND	Whether the person was not paying attention	Not paying attention can result in accident
7	UNDERINFL	Whether the person was driving under influence	DUI can cause accidents
8	WEATHER	Weather conditions	Bad weather can cause accidents
9	ROADCOND	Road conditions	Wet roads can cause skidding
10	LIGHTCOND	Light conditions	Light conditions affect visibility
11	PEDROWNOTGRNT	Pedestrian right of way was granted or not	
12	SPEEDING	Whether speeding or not	Speeding causes accidents
13	COLLISIONTYPE	Collision Type	Type of collision gives severity of accident
14	HITPARKEDCAR	Whether or not the collision involved hitting a parked car.	Hitting a parked car causes property damage
15	Year	Year of accident	Did one year have a lot of accidents
16	Month	Month of Accident	Does month affect number of accidents
17	Day	Day of accident	Day of month
18	Hour	Time of accident	Are accidents caused majorly at night
19	Weekday	What day of the week accident happened	Are accidents caused more on certain days of the week

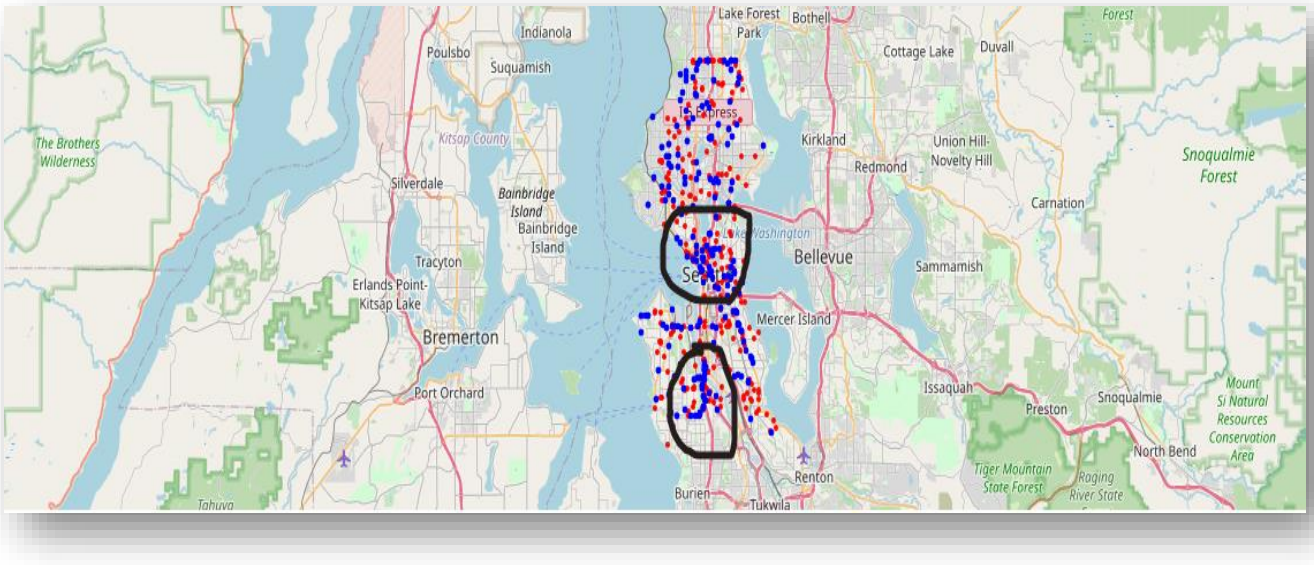
3. METHODOLOGY

3.1 Data Analysis

ADDRTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDTTM	INATTENTIONIND	UNDERINFL	WEATHER
Intersection	2	0	0	2	3/27/2013 2:54:00 PM	NaN	N	Overcast
Block	2	0	0	2	12/20/2006 6:55:00 PM	NaN	0	Raining
Block	4	0	0	3	11/18/2004 10:20:00 AM	NaN	0	Overcast
Block	3	0	0	3	3/29/2013 9:26:00 AM	NaN	N	Clear
Intersection	2	0	0	2	1/28/2004 8:04:00 AM	NaN	0	Raining

I used python **folium** library to visualize geographic details Taking some key features like speeding and driving under the influence of alcohol we can see some key areas that have higher rates of collision compared to other locations across the city from 2013 to 2020

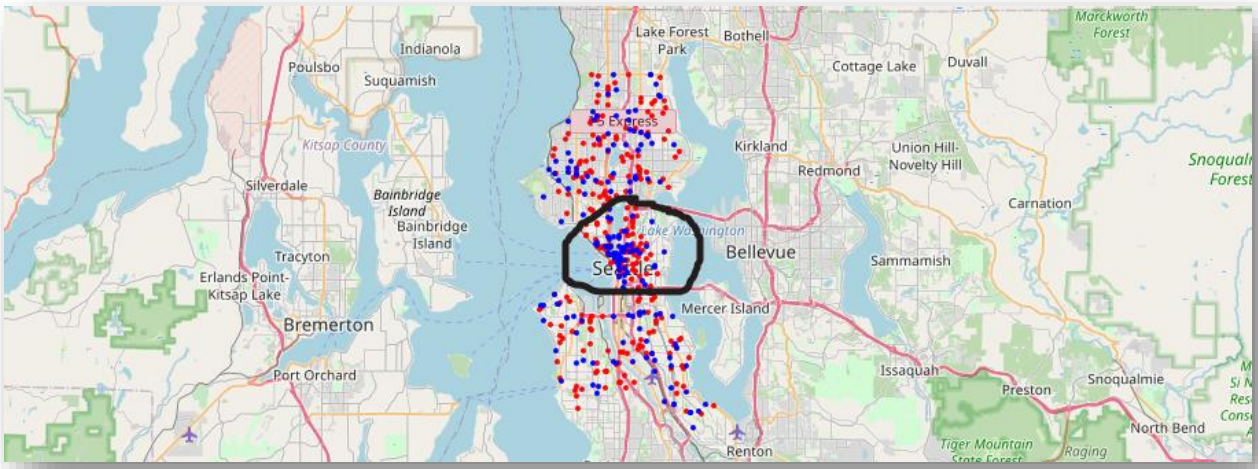
1. Speeding



Accidents caused by Speeding (2013–2020)

The image above shows circled spots on the map which indicates areas highly prone to accidents it would be crucial for stakeholders like the city mayor and emergency services to investigate these areas to reduce accidents rate.

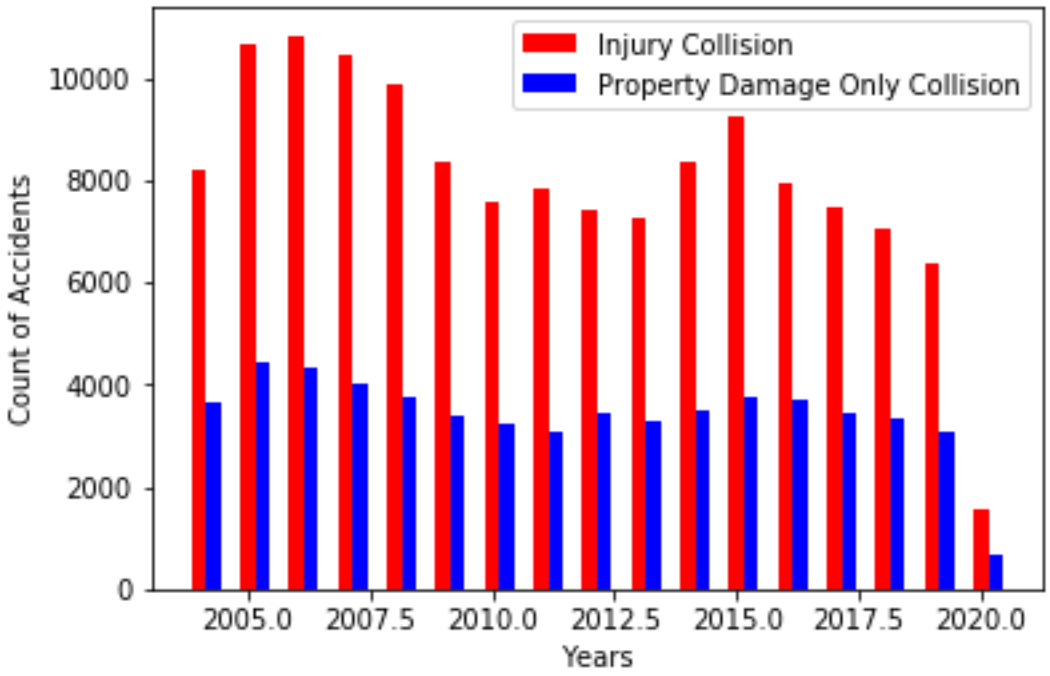
2. Under the Influence



Accidents caused Under the influence (2013–2020)

The above map shows, the points where accidents are caused due to DUI, I recommend the city of Seattle to introduce increase awareness of the dangers of driving while impaired by putting up road signs to educate drivers around those areas in the circled marks.

Looking at the distribution of accidents through the years according to the severity such as injury collision and property damage only and we see that over 10,000 people were involved in injury collisions from 2015 to 2017 and 2020 records lower than 2000 causalities as at when this data was recorded.



3.2 MODELLING

Using the following classifiers to get the prediction whether given certain attributes (features = X), the severity of the accident (labels = y)

1. Logistic Regression
2. Support Vector Machine
3. XGBoost Classifier
4. Random Forest Classifier

RESULTS

The Results of the classification are as follows:

CLASSIFIER	F1-SCORE	ACCURACY
Logistic Regression	0.75	0.70
Support Vector Machine	0.69	0.75
XGBoost Classifier	0.73	0.76
Random Forest Classifier	0.72	0.74

Conclusion:

The data-set has been used to classify the severity of the accidents based on certain select features.

The exploratory data analysis shows density of accidents based on geography based on Speeding, Driving Under Influence, In-attention and Hitting Parked Cars.

From a machine learning standpoint. The most important features were: Collision Type, Person Count, Vehicle Count and Address Type. The Gradient Boost algorithm performed the best.