Car Accident Severity: Seattle, Washington

(Applied Data Science Capstone – IBM)

This project aims at understanding what factors play a vital role in the severity of car accidents in Seattle, Washington using Data Science Toolkit and predicting them prior to take necessary measures to avoid them using Machine Learning techniques.

Name: Vedant Mane

Project: Car Accident Severity Prediction

Course: Applied Data Science Capstone

Specialization: IBM Data Science Professional Certificate

Contents:

1. Introduction
   1.1. Background
   1.2. Problem
   1.3. Stakeholders
2. Understanding Data
   2.1. Data Source
   2.2. Data Cleaning
   2.3. Feature Selection
3. Methodology
   3.1. Exploratory Data Analysis
   3.2. Model Selection
4. Results
   4.1. Logistic Regression
      4.1.1. Model
      4.1.2. Classification Report
      4.1.3. Confusion Matrix
      4.1.4. Model Evaluation
   4.2. Decision Tree
      4.2.1. Model
      4.2.2. Classification Report
      4.2.3. Confusion Matrix
      4.2.4. Model Evaluation
5. Discussion
   5.1. Average F1 Score
   5.2. Precision
   5.3. Recall
6. Conclusion
7. References

1. Introduction:

1.1. Background:

Seattle, also known as the Emerald city, is the largest city in both the state of Washington and the Pacific Northwest region of North America. It is home to a large tech industry with Microsoft and Amazon headquarters in its metropolitan area. The city has urban population of over 3.4 million as reported by PopulationStat. [1] As reported by curbed.com [2] in 2017, the total number of personal vehicles in Seattle in 2016 is approximately 435,000. The car population has more than doubled in the Seattle area since 2010. This tremendous increase in the number of vehicles has lead to higher number of accidents on the road explained merely by a simple probability. Worldwide, approximately 1.35 million people succumb to death due to road crashes every year, on average a total of 3,700 people lose their lives everyday in the road and an additional 20-50 million people suffer non-fatal injuries, often resulting in long-term disabilities.

1.2. 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 can cost up to $871 billion in a single year. According to the WSDOT data, a car accident occurs every 4 minutes and a person dies every 20 hours due to a car crash in the state of Washington. 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 car accidents can be reduced based on a few factors.

1.3. Stakeholders:

The reduction in the car accidents can be beneficial to several government bodies that work towards improving those road factors and car drivers themselves who may take precautionary measures to reduce the severity of the accidents.

2. Understanding Data:

2.1. Data Source:

The dataset requires pre-processing so that the machine learning model classifies the prediction to a categorical variable. The dataset has a total of 37 features and 194673 observations with variation in number of observations for every feature. The total dataset has high variation in the lengths of almost every column in the dataset due to the missing data values. These values could have been beneficial for our prediction algorithm had the data been present. The Metadata information for the dataset can be found at the link here.

2.2. Data Pre-processing:

The model aims to predict the severity of the car accident, considering that, the variable "SEVERITYCODE" of Severity Code that was in the form of 1 (Property Damage only) and 2 (Injury Collision) were encoded as 0 (Property Damage only) and 1 (Injury Collision). The attribute "INATTENTIONIND" for Driver's Attentiveness at the time of accident was considered 1 were value was 'Y' and 0 where data appeared missing. Similarly, the attribute "UNDERINFL" was encoded as 0 for the values 'N', '0', 'Nan' while 1 for 'Y', '1'. In the case of "SPEEDING", the 'Y' value was considered as 1, while missing values were considered to have the value 0. The Light conditions were at the time of accident were categorized into 4 groups based on the type of light and time of day as Daylight, Dark with Street Lights, Dark without Street Lights & Dark with unknown lighting conditions. Similarly, Weather Conditions were categorized as Clear, Cloudy or Low Visibility, Windy, Pouring for Raining or Snowing or Sleet/Hail/Freezing Rain and Unknown for missing or unknown values. The condition of the roads was classified as Dry, Wet, Blocked or Unknown conditions. The "JUNCTIONTYPE" attribute was classified into 2 types based on whether the accident took place at the intersection or elsewhere. Then, dummy variables were created for all the categorical columns (ROADCOND, LIGHTCOND, WEATHER).

Further we will normalize this data using the StandardScaler function and fit this data for our dataset and transform the same for model building.

2.3. Feature Selection:

Feature	Description
Location (Address Type)	Description of the general location of the Collision.
Weather Condition	A description of the weather conditions during the time of the collision.
Car Speeding	Whether or not speeding was a factor in the collision.
Light Conditions	The light conditions during the collision.
Road Condition	The condition of the road during the collision.

Junction Type	Category of junction at which collision took place
Number of People involved	The total number of people involved in the Collision.
Number of Vehicles involved	The number of vehicles involved in the collision.
Driver's Attentiveness	Whether or not the Driver was attentive.
Driver under influence	Whether or not the Driver was under influence.
Severity Code	A code that corresponds to the severity of the collision: • 1 — Property Damage • 2 — Injury Collision

3. Methodology

3.1. Exploratory Data Analysis

3.2. Model Selection

4. Results

4.1. Logistic Regression
4.1.1. Model
4.1.2. Classification Report
4.1.3. Confusion Matrix
4.1.4. Model Evaluation
4.2. Decision Tree
4.2.1. Model
4.2.2. Classification Report
4.2.3. Confusion Matrix
4.2.4. Model Evaluation

5. Discussion

5.1. Average F1 Score

5.2. Precision

5.3. Recall

6. Conclusion

7. References

[1] "populationstat.com," PopulationStat, October 2020. [Online]. Available: https://populationstat.com/united-states/seattle. [Accessed 03 October 2020].

[2] "seattle.curbed.com," Curbed.com, 10 August 2017. [Online]. Available: https://seattle.curbed.com/2017/8/10/16127958/seattle-population-growth-cars-transit. [Accessed 03 October 2020].