IBM Data Science Professional Certificate Powered By coursera

coursera

Applied Data Science Capstone

Capstone Project - Car Accident Severity

OMID KARAMI

Applied Data Science Capstone

September 2020

Report

Contents

Abstract

One of the important indicators for measuring the health index of developed countries is the rate of accidents and road deaths. According to the World Health Organization, road traffic injuries caused an estimated 1.35 million deaths worldwide in the year 2016[1]*. In this report I will explain the Supervised Machine Learning techniques to predict the road accident severity through the data for 221525 accidents from 2004 until September 2020 in the Seattle City. It's expected based on this modelling, we can help the organizations that are involved in road accidents and traffic such as emergency services, traffic centre, insurance and public transport companies, as well as the municipality and also passenger.*

1. Introduction

The goal of this project is the prediction of road accident severity. This study utilized a dataset consisting of 221525 recorded accidents and 40 attributes in Seattle City in Washington State by Seattle Department of Transportation (SDOT). Analyzed and build balanced machine learning models were developed by applying prediction methods. One of the important indicators for measuring the efficiency of service provision in road network systems of each country is the number of road accidents. The main purpose of the study aims analysis of injury accident and fatal accident to predict the accident severity. Its measurement comprehensively considers statistical relationship among variables such as average speed on road section, average traffic volume per day, period of time, weather conditions, physical characteristics of accident area, and causes of accident.

1.1 Business Problem Definition

In this study, the following issue have been tried to be answered accurately, both in terms of studying the number of injury and fatal accident, weather condition as well as a complete study of the information related to Seattle transportation network systems.

- Predict the road accident severity in the Seattle transportation system.

1.2 Target the Right Audiences

Specifically, insurance and public transport companies, as well as the municipality and all passengers who traveling on intercity routes are among those who will be interested in the results of this research. Numerous reasons can contribute to road accidents. Rough road conditions, bad weather condition such as wet ground Poor lighting can impair visibility, confined or congested traffic routes can increase the likelihood of collisions. Consequently, municipalities or road construction and public transport companies can review the results of this study to provide appropriate solutions to prevent road accidents. Also, passengers can avoid unnecessary travel by checking the weather conditions and road traffic before their trip.

[1] https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

2. Related Research

Due to the rapid spread of car accident and its impact on fatal indicator in each country, a large number of studies have been conducted in this regard. In this section, I have reviewed two related articles and outlined the methods they have presented.

1. Impacts of accident severity factors and loss values of crashes on expressways in Thailand[2]
2. Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models[3]

The first paper focuses on finding factors that affect the accident severity. The speed on a road section influences the severity of crashes. (Vatanavongs et al., 2014).
The second paper presents a comparison between two modeling techniques, Bayesian network and Regression models, by employing them in accident severity analysis. Three severity indicators, that is, number of fatalities, number of injuries and property damage, are investigated with the two methods, and the major contribution factors and their effects are identified. (Fang et al., 2013).

3. Data

3.1 Data Description

In order to achieve a successful project, it is important to use a manageable dataset to analyze it. For this purpose, all data must be available and also to achieve the desired result, with high accuracy and efficiency, its size should be large enough, and of course all pre-processing conditions, including data cleaning and normalization, should be considered. This dataset includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. (Timeframe: 2004 to Present). The dataset is available in Comma Separated Value (CSV) format. For further information you can visit the official website of Seattle Department of Transportation[4] through the following link. (Click Here)

The dataset consists of 39 independent variables and 221,525 rows. The dependent variable or target variable, “SEVERITYCODE”, contains value corresponds to the severity of the collision. It takes the values 0, 1, 2, 2b or 3. The description of “SEVERITYCODE” are provided in the “Attribute Information” metadata which available in the official website of Seattle Department of Transportation[5] and you can see as follows:

[2] Vatanavongs Ratanavaraha, Sonnarong Suangka, Impacts of accident severity factors and loss values of crashes on expressways in Thailand, IATSS Research, Volume 37, Issue 2, 2014, Pages 130-136, ISSN 0386-1112, https://doi.org/10.1016/j.iatssr.2013.07.001.

[3] Fang Zong, Hongguo Xu, Huiyong Zhang, "Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models", Mathematical Problems in Engineering, vol. 2013, Article ID 475194, 9 pages, 2013. https://doi.org/10.1155/2013/475194.

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

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

- 0: Unknown
- 1: Property/vehicular damage
- 2: Minor injury
- 2b: Serious injury
- 3: Fatality

3.2 Data Cleaning

High quality data enables strategic and business systems to have a better insight of problem solving. Data cleaning is one of the most important steps in pre-processing stage. Eliminating unwanted, duplicate and irrelevant observations allows even simple algorithms to have significant insights into the data and achieve better results. Detect the outlier data and remove it, as well as handling missing data, help to producing quality data and model's performance.

Missing target variable: A simple look at the target variable shows that a large amount of data is not useful for building model. There are many accidents with SEVERITYCODE =0 which means "Unknown" severity. The goal of this model is to predict the car accident severity, consequently the data that is unknown should be deleted.
Redundant and useless data: There are some columns containing useless or redundant data. These columns can be removed from the data frame.
Convert categorical variable to numeric: As I mention before, the target variable "SEVERITYCODE" is a categorical variable. Convert the categorical code [0, 1, 2, 2b, 3] to numeric code [0, 1, 2, 3, 4] in order. This means, relabeling '2b' as '3' and '3' as '4'.

3.3 Feature Selection

The feature selection process in many cases enhances the performance of the machine learning model. This process involves reducing the number of input variables to reduce computational costs in developing a predictive model. In addition, the selection of features includes the selection of the most important and relevant features. In this study, the following features are key features in predicting the severity of accidents, which will be fully discussed in the Methodology section.

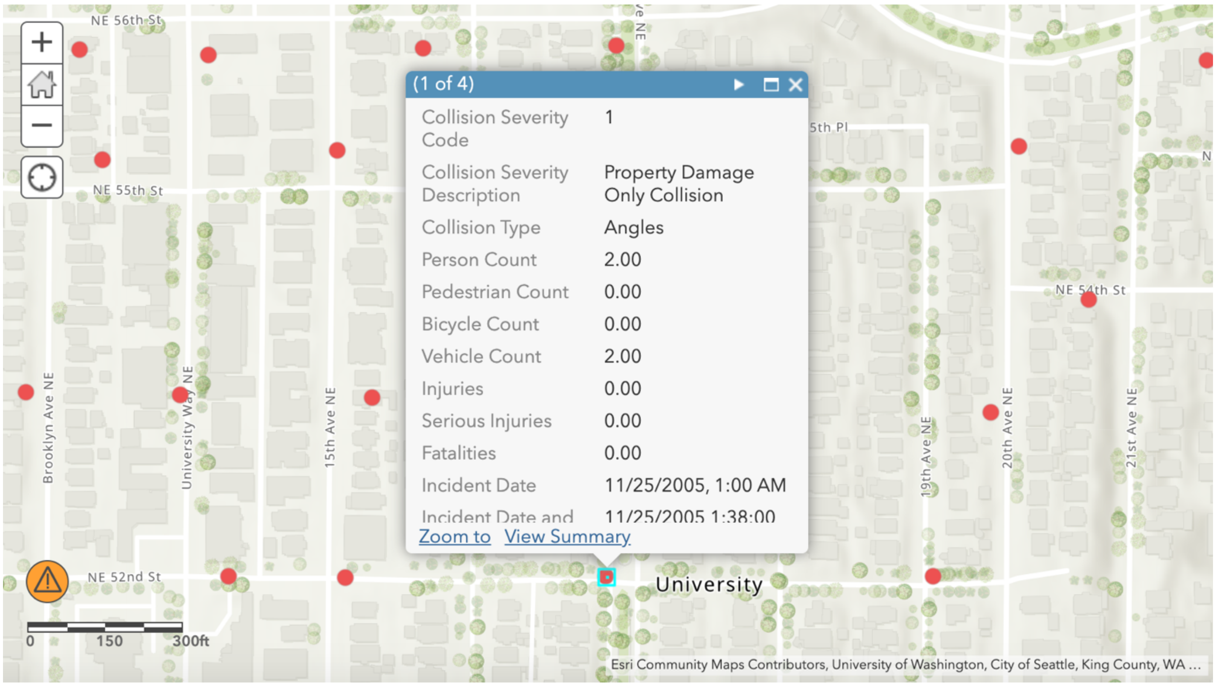
- **WEATHER**: *A description of the weather conditions during the time of the collision.*
- **ROADCOND**: *The condition of the road during the collision.*
- **LIGHTCOND**: *The light conditions during the collision.*

Eventually, the Target or Dependent variable will be **'SEVERITYCODE'**. It's used to measure the accident severity. The following features play an important role in measuring the severity of accidents. **'WEATHER'**, **'ROADCOND'** and **'LIGHTCOND'**.

3.4 Data Exploration

Regarding to have a good insight of data exploration, I obtained information respecting the collision from the ArcGIS Online[6] (Fig.1) that enables to connect people, locations, and data using interactive maps. This map and ocean of another map are available in ArcGIS Online website.

In order to have a better understanding of the dataset, before starting building the model, we first explore some of the main features of the data set. Visualization helps us to have a better insight of the data and also be able to build a more accurate model. In this section, through histogram, tried to get a better perspective on the data. In Fig. 2 plotting out the accident severity, as well as the number of people involved in the collision in the collision in Fig.3 and the frequency of the number of total injuries and fatalities in the collision are shown in Fig.4 and Fig.5.



Collision Severity Code	1
Collision Severity Description	Property Damage Only Collision
Collision Type	Angles
Person Count	2.00
Pedestrian Count	0.00
Bicycle Count	0.00
Vehicle Count	2.00
Injuries	0.00
Serious Injuries	0.00
Fatalities	0.00
Incident Date	11/25/2005, 1:00 AM
Incident Date and	11/25/2005 1:38:00

Esri Community Maps Contributors, University of Washington, City of Seattle, King County, WA ...

Figure 1 - ArcGIS Online

[6] https://www.arcgis.com/home/index.html

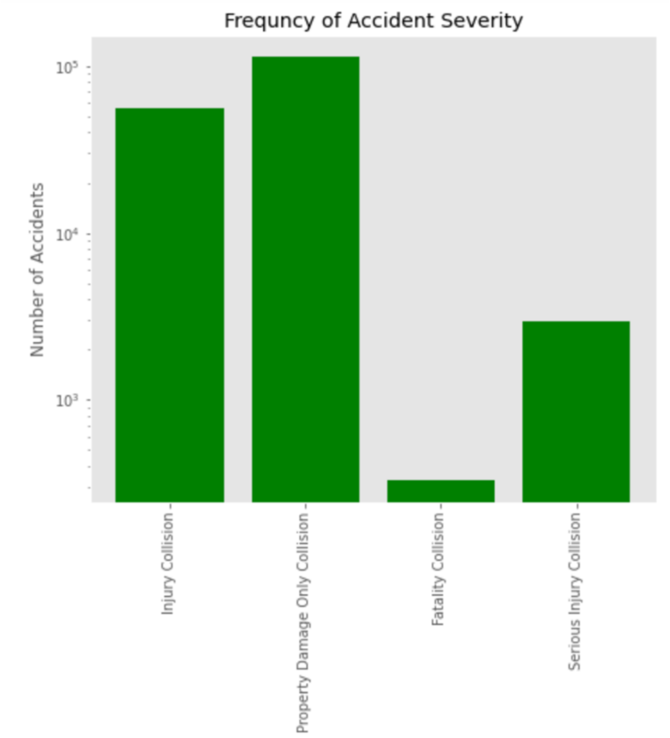


Figure 2 - Frequency of Accident Severity

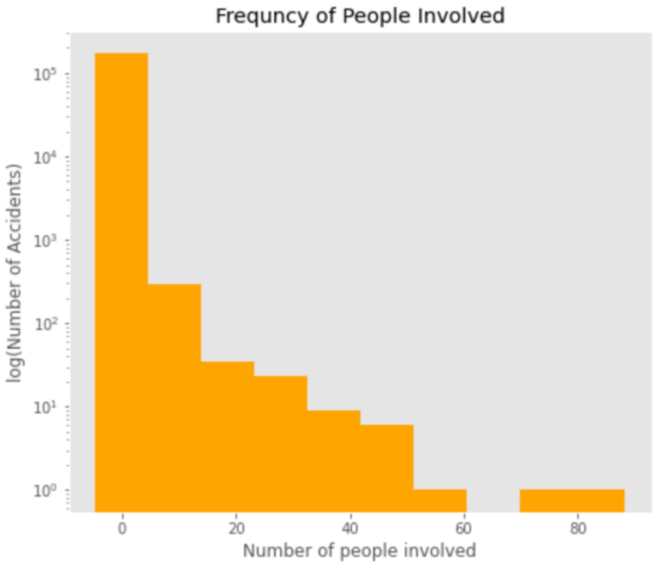


Figure 3 - Frequency of People Involved in the Collision

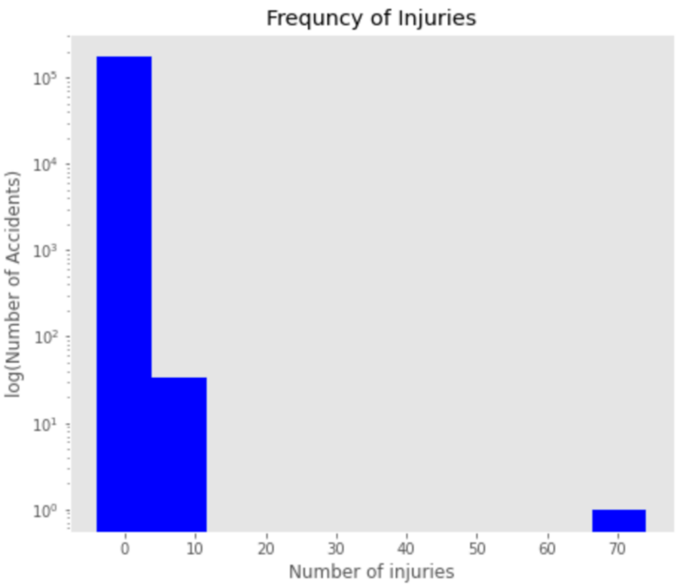


Figure 4 - Frequency of Injuries

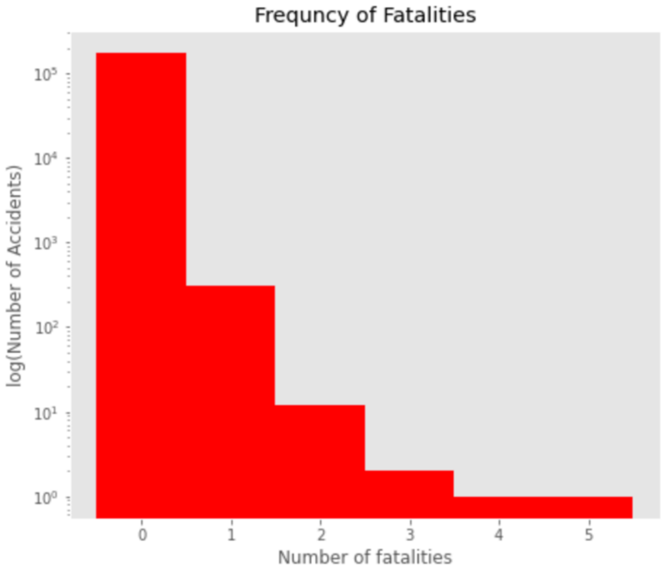


Figure 5 - Frequency of Fatalities

4. Methodology

The purpose of this study is to predict the severity of accidents using the supervised learning algorithms. To achieve an accurate prediction model, the data is performed by pre-processing steps. Before building the model, we split dataset in to training and testing subsets. In this approach, the parameter test_size was set to 0.3, meaning that 70% of the balanced data were used for training the model and 30% of the data were reserved for testing. Regarding the target or dependent variable, we can use to prediction model to predict the accident severity by two different categories in Machine Learning Models. Regression and Classification models.

To build the model, three different algorithms are used in this project including supervised machine learning techniques. These algorithms are listed below;

- Decision Tree
- K-Nearest Neighbor (KNN)
- Logistic Regression

Jupyter Notebook are used to implement these algorithms, as well as important Python libraries such as Numpy, Pandas, Scikit-Learn (python machine learning library). After implementing machine learning algorithms and building models, I have evaluated these three models using evaluation methods and you will see the results in the following sections.

5. Results and Evaluations

According to the table below, you can see the accuracy of each of the algorithms used. In this evaluation, we have used Jaccard Similarity Score, F1 Score and R2 Score.

Model	F1 Score	Jaccard Score	R2 Score
Decision Tree	0.67	0.52	-0.33
K-Nearest Neighbor (KNN)	0.67	0.46	-0.39
Logistic Regression	0.76	0.61	0.03

Table 1 - Summary of the Performance Metrics of the Machine Learning Model

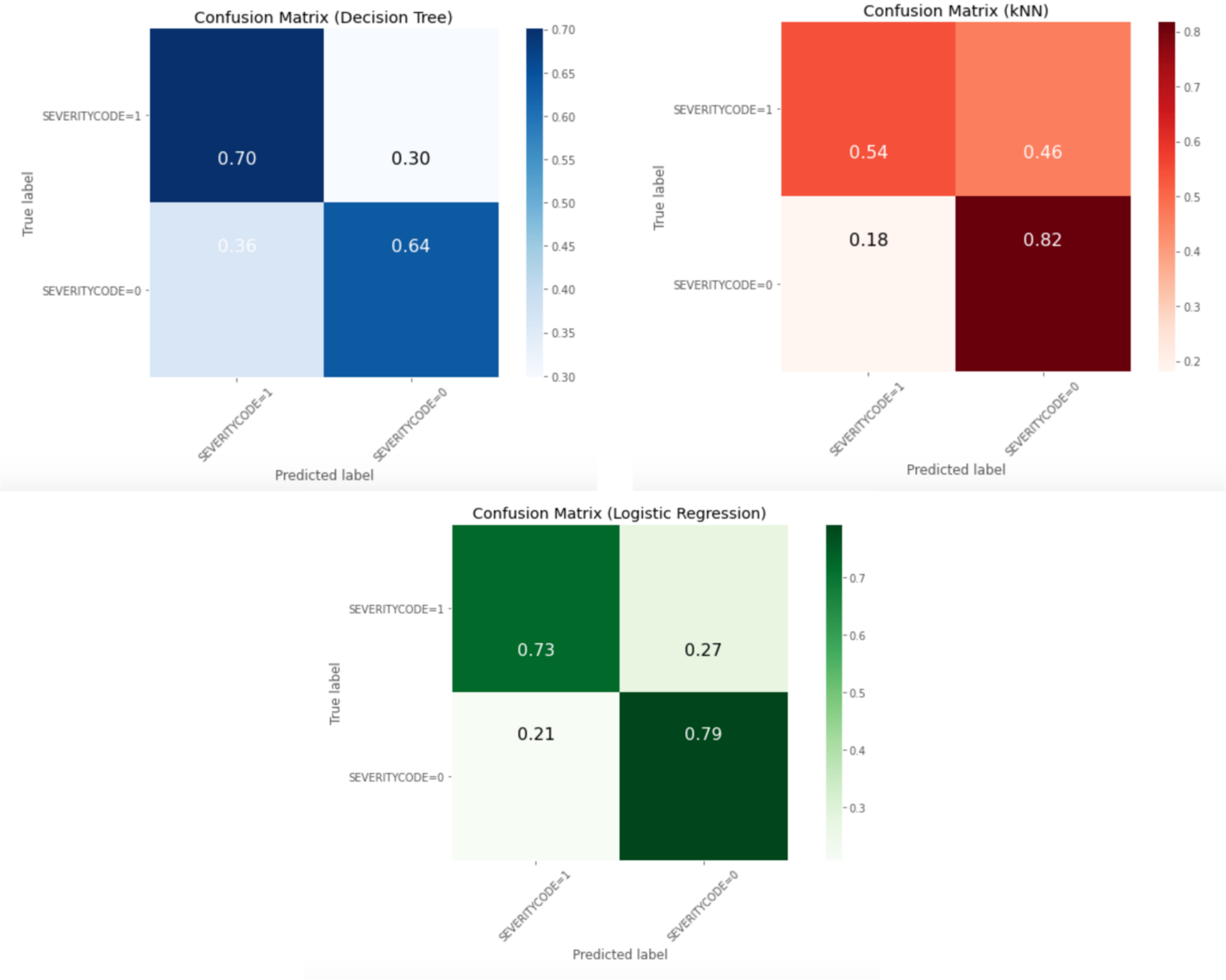


Figure 6 - Confusion Matrices of the Machine Learning Model

6. Discussion

Regarding Table 1, the F1 score and Jaccard Similarity Index for each model is given. The LR model have the highest F1 scores. The F1 score of the kNN and DT model is equal with 0.67. having an F1 score of 0.67. The Jaccard Similarity scores for the three models are 0.52, 0.46 and 0.61, respectively. One key advantage of the LR model over the DT and kNN models is simplicity of implementing and as you see, the high accuracy in both F1 score and Jaccard Similarity score.

7. Conclusion

The purpose of this study is to predict a model for the severity of road accidents. Such a model can be adapted with any road traffic network anywhere in the world. The traffic data set in this study was from the Seattle Department of Transportation. As you can see in the evaluation section, the accuracy of this model is very good. It has 82% accuracy. Therefore, based on this modeling, we can help the organizations that are involved in road accidents and traffic such as emergency services, traffic center, insurance and public transport companies, as well as the municipality and also passenger.

References

Vatanavongs Ratanavaraha, Sonnarong Suangka, Impacts of accident severity factors and loss values of crashes on expressways in Thailand, IATSS Research, Volume 37, Issue 2, 2014, Pages 130-136, ISSN 0386-1112, https://doi.org/10.1016/j.iatssr.2013.07.001.

Fang Zong, Hongguo Xu, Huiyong Zhang, "Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models", Mathematical Problems in Engineering, vol. 2013, Article ID 475194, 9 pages, 2013. https://doi.org/10.1155/2013/475194.

https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

https://www.arcgis.com/home/index.html

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

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