IBM Data Science Professional Certificate – Capstone Project

By:

Mohamed Elnaghy

Table of Contents

[Introduction 2](#_Toc52326435)

[Data 2](#_Toc52326436)

[Methodology 3](#_Toc52326437)

[Data Preparation 3](#_Toc52326438)

[Data Wrangling 3](#_Toc52326439)

[Exploratory Data Analysis 3](#_Toc52326440)

[Model Development and Evaluation 4](#_Toc52326441)

[Results and Discussion 4](#_Toc52326442)

[Conclusion 5](#_Toc52326443)

**Table of Figures**

[Figure 1 Leaflet Map for the Accident I will be Using 3](#_Toc52326456)

[Figure 2 Results 5](#_Toc52326457)

[Table 1 The Two Categories of Attributes 4](#_Toc52326465)

# Introduction

Road accidents are a severe concern for most nations worldwide because accidents can cause severe injuries and fatalities and substantial economic losses. According to the World Health Organization's Global Status Report, approximately 1.25 million deaths happened per year are because of road accident injuries, and most fatality rates were in lower-income countries. Not only road accidents caused deaths and injuries, but also it affected the economic and social conditions of in terms of health care costs of injuries and disabilities. The World Health Organization (WHO) estimated the economic costs derived from road accidents reached 518 billion USD per year in high income countries and 65 billion USD per year in medium and low-income countries. Hence, no one can deny the importance of studying the accidents, and analyzing the data to mitigate and reduce the traffic deaths. In this project, CRISP-DM (Cross Industry Process for Data Mining) methodology will be used to predict any road's accident severity by training an efficient machine learning model with the help of existing accidents. This will play a crucial factor for traffic control authorities to take proactive precautionary measures.

# Data

For this project, I'll be using US accidents dataset acquired from Kaggle.com. the dataset covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. There are about 3.5 million accident records in this dataset, and more than 40 attributes. The accidents severity is a number between 1 (least impact on traffic) to 4 (significant impact on traffic). Some of the attributes are start Latitude, start longitude, end latitude, and end longitude. However, the end latitude and longitude are more than 70% missing, so these two attributes will be neglected. Full weather condition of each accident is listed in the dataset (humidity, Temperature, wind, precipitate, pressure), so I’ll be using these data to analyze and train my model to predict the severity of the accident based on the weather condition. Besides, I’ll be using different correlation to discover what is the most weather condition that contributed to the accident.

# Methodology

## Data Preparation

### Data Wrangling

At first, I used the whole dataset that I have, which has more that 3.5 million rows. However, I had too many kernel problems, where the kernel dies in the middle of a regression analysis. As a result, I had to limit the data to only 10000 rows. The dataset has 49 columns, and after understanding the dataset, I found out that I won’t be using some of the attributes, so I dropped the unwanted columns. Then, I replaced the NaN values with the mean for some columns and deleted the rows that contain the NaN values.

### Exploratory Data Analysis

I wanted to picture the accidents location for the limited dataset on map, so I used folium to create a leaflet map (Figure 1). I used the correlation function to get the correlation between the attributes. Stop, Crossing, Junction, and Traffic\_Signal were the highest values. However, all the correlation coefficients were very low (between 0.0001 to 0.2). That's why I used Pearson's correlation as well. Similarly, it gave me low values, so I'll be going with the abovementioned attributes. As a result, I divided the attributes into two categories; the first category was the numerical attributes, and the second category was the Boolean attributes (Table 1).

Map

Description automatically generated

Figure Leaflet Map for the Accident I will be Using

Table The Two Categories of Attributes

|  |  |
| --- | --- |
| Numerical Attributes | Boolean Attributes |
| Wind Speed | Stop |
| Visibility | Crossing |
| Humidity | Junction |
| Precipitation | Traffic Signal |

## Model Development and Evaluation

I used Supervised Learning to develop my model, as I am using fully labeled data, which means that each example in the training dataset is tagged with the answer the algorithm should come up with on its own.

I used the train test split and used 10% of the data for testing. Then, I used different classification techniques (K Nearest Neighbor, Decision Tree, Logistic Regression, Support Vector Machine). For evaluation, I used different evaluation techniques (jaccard similarity score, f1 score).

# Results and Discussion

From the results shown in figure 2, the Boolean Attributes has higher accuracy than the Numerical Attributes in all the classification except K Nearest Neighbor (f1 score). The Decision Tree for the Boolean Attributes is the highest accuracy in both jaccard similarity score (63.6%) and f 1 score (57.7%). However, I think we need more data attributes with higher correlation coefficient to be able to get more accurate model.

Table

Description automatically generated

Figure Results

# Conclusion

In Conclusion, I get an access to the dataset from Kaggle.com, which consists of 3.5 million rows and 49 columns of information about car accidents in the US from 2016 to 2020. After, cleaning the data and prepare it for model execution, I divided the desired attributes into Boolean and Numerical Attributes. Then, I used different classification technique, and with evaluation, I was able to get the best model with the highest accuracy, which was the Decision Tree for the Boolean Attributes.