IBM Data Science Certificate Program
Applied Data Science Capstone
Car Accident Severity Analysis

Zhao Zhang

Date: 10/13/2020

1. Introduction

1.1 Background

Motor vehicle crashes in the United States cost $242 billion in economic damage and 32,999 deaths in 2010. Furthermore, there were 3.9 million injuries and 24 million damaged vehicles (Ref1). Property damage and injury-related traffic accidents are a socially and economically disruptive force that often sinks into the back of our minds. Factors such as medical expense, productivity loss, and administrative costs amount to approximately $871 billion damage in societal value (Ref1). This astonishing figure is approximately twice the damage done by Hurricane Katrina, the most costly hurricane in U.S history. However, unlike Hurricane Katrina, road accidents are more easily managed through human effort. For instance, the degree of human error and traffic environment, which increase the likelihood of accident occurrence and the magnitude of damage, can be optimized through education and increased investment in infrastructure.

1.2 Understanding the Problem

This project examines the traffic collision data in Seattle, Washington from 2004 to the present. Through exploratory data analysis and machine learning techniques, it aims to provide some insight in aspects such as traffic management, road infrastructure, and importance of traffic safety. Specifically, the project focuses on two crucial aspects of accidents: human error and environment. Factors such as inattention, rule-breaking actions (e.g., exceeding speed limit) and substance abuse constitute the subjective reason of an accident. On the other hand, natural environment (e.g., weather) and infrastructure (e.g., road and light conditions) form objective reasons for an accident. Gaining perspective on these issues can help the relevant stakeholders develop scientific solutions and minimize the harm done on the road.

1.3 Stakeholders

The main stakeholders of this report are governmental agencies in charge of city planning/civil engineering, transportation, traffic control, and road maintenance. Predictive machine learning models can help identify weak spots that are more susceptible to severe accidents based on real-time data. In turn, short-term tactics such as more stringent traffic regulations and lower speed limits can be placed on high-risk areas. Furthermore, in the long term, maintenance plans and infrastructure investments can be implemented in a more clinical, targeted, and effective approach. In addition to public organizations, individual drivers can also benefit from this report from an educational perspective. There are only benefits from further understanding the harm of irresponsible driving and the unnecessary risks of driving in harsh conditions.

2. Data

2.1 Data Source

The dataset contains weekly-updated collisions data in Seattle, Washington, USA from 2004 to Present (October 2020). The employed open-source government dataset is collected and published by Seattle Department of Transport (SDOT). The original dataset can be found here.The metadata information is released by the SDOT Traffic Management Division. Metadata information can be found here.

The raw dataset contains 194673 entries and 38 columns, representing 194673 recorded collisions and a wide range of attributes of the accidents. See **Appendix 1.** for the basic descriptive statistics and data information of the raw dataset.

2.2 Feature Selection

The selected feature set of variables are shown in Table 1.

Column Name	Description
ROADCOND	The road condition at the time of the accident – e.g., wet, dry
LIGHTCOND	The light condition at the time of the accident – e.g., dark, daylight
WEATHER	The weather at the time of the accident – e.g., clean, rain
UNDERINFL	Whether the driver is under influence – Yes or No
SPEEDING	Whether the driver was speeding – Yes or No
INATTENTIONIND	Whether the driver was paying attention – Yes or No
ADDRTYPE	The address type of the accident – e.g., alley, mid-block
JUNCTIONTYPE	The junction type at the accident location -e.g., mid-block
COLLISIONTYPE	The type of collision that took place – e.g., head-on, rear-ended
PEDCOUNT	The number of pedestrians in the accident -e.g., 3
PEDCYLCOUNT	The number of cycling pedestrians in the accident -e.g., 1
VEHCOUNT	The number of vehicles in the accident -e.g., 2
SEVERITYCODE	The type of damage done in the accident – i.e., property damage, injury

Table 1. Feature Selection and Description

The selected features can be categorized into surrounding environment variables (including physical location of the accident and natural environment) and driver discretion. Other attributes in the raw dataset are neglected for a variety of reasons. For instance, SDOTCOLNUM, a number assigned to the incident, is disregarded because it is irrelevant to the objectives of the project. The employed features provide universal factors in a car accident. Moreover, almost all of the information can be obtained through first-hand report, hence offering value to first responders and emergency service agencies. Therefore, the project aims to provide appliable insight appliable to other areas, even though the

study is solely done on the city of Seattle.

2.3 Data Cleaning

The main data cleaning issues of this project include the presence of categorical variables, missing variables, and data imbalance. The process is primarily conducted through python and its Pandas package. Some variables are also encoded and standardized to suit machine learning techniques. Please check the python notebook attached in my GitHub for the specifics as well as the coding.

2.3.1 Missing Variables

First, data entries with missing values in the target variable column, SEVERITYCODE, are dropped. Second, we also drop entries with missing values in the feature variables (i.e., Road Condition, Weather, Light Condition, Under Influence, Speeding, Inattention, Address Type, Collision Type, and Junction Type). The rationale is the small volume of missing data contained won't make a meaningful impact for this report. The data volume after dropping missing values sit at 188344 entries.

2.3.2 Feature Engineering and Pre-processing

Based on our feature selection, we disregard the columns in the original dataset that are not in Table 1. However, problems such as data format, consistency, and categorical values require remain. Therefore, the feature variables require further processing. For instance, we convert object Yes/No values to integer 1s and 0s, which better fit machine learning algorithms. In terms of consistency, I address those that have inconsistent data formats (e.g., the UNDERINFL column contains both 1s/0s and Y/N values) to a uniform data type (in this case, integer 1s and 0s). Lastly, for columns with categorical variables, such as WEATHER and ROADCOND, different methods are implemented for the exploratory analysis part and machine learning modelling part. I leave the variable in its original string format for exploratory analysis, which contains mostly data visualization. For machine learning models, these variables are encoded by creating dummy variables for each different value. I also convert SEVERITYCODE values, 1(property damage) and 2(injury), into 0(property damage) and 1(injury). As a result, the final machine learning dataset is expanded into 44 columns. It is also notable that data entries with 'unknown' or 'other' categorical values are treated as null values and are therefore dropped. At this stage, there are 167310 entries in the dataset.

Another key issue in machine learning models is data imbalance. There are about 3 times as many property damage collisions in the dataset than injury-related collisions. If this issue is left unaddressed, machine learning techniques are therefore likely to be biased (for instance, if we project all the entries to be property damage related, we can still achieve an accuracy of 75%. Therefore, we down-size the final dataset so that there are an equal amount of property damage collisions and injury-related collisions for our target variable, SEVERITYCODE. Lastly, we standardize all variables using the scikit-learn package to fit machine learning