Seattle City Traffic Accident Severity Analysis

Applied Data Science Capstone

By: Sarah Fallon

Date: 28 October, 2020

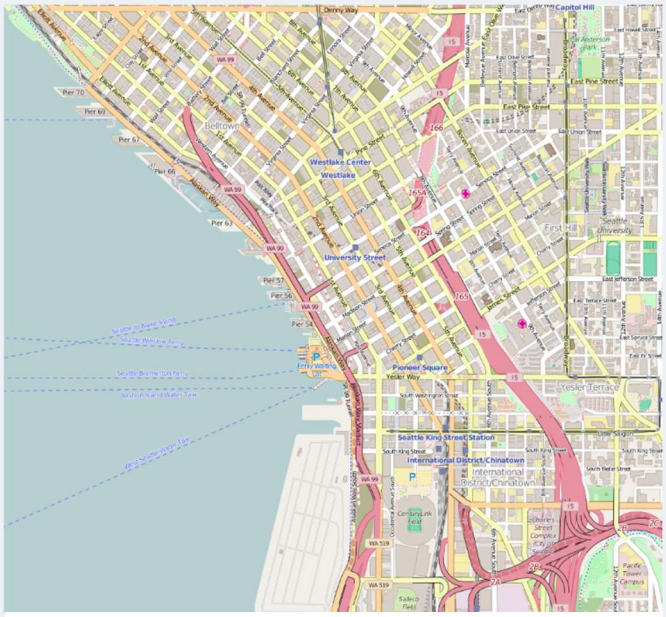
TABLE OF CONTENTS

1. INTRODUCTION

a. Business Problem

Ever since 2017 when *Seattle Times* reported that the city drivers were ranked the fifth worst in the nation[1], the Seattle City Police Department (SCPD) has been under constant pressure from the city government and its citizens to reduce the car accidents and associated damage, increase likelihood of safe driving, reduce damages and injuries, and save lives.

They have now hired our data scientist as part of their team to see if a model can be created to help predict the severity of car accidents in order to better utilize their resources and more efficiently and effectively deploy personnel to each accident reported. The initial analysis they are interested in would be a prediction based on weather and road conditions in order to help better deploy resources, warn the public for high severity days, and ultimately save lives.



[2] This Photo by Unknown Author is licensed under CC BY-SA

b. Stakeholders

The SCPD have developed a task team to research the issue and determine the extent to which a model would be useful. They are the main stakeholders in the analysis in order to help improve road conditions, warn drivers on potentially high accident severity days, and deploy appropriate resources in the event an accident is reported.

The SCPD is answering a call by the city government, mandating improvement of the city traffic accident ratings over the next 5 years. Funding has been tied to the ratings numbers as the city government is looking to ensure traffic improves.

---

[1] Clarridge, Christine. (2014). 'Washington state drivers 5th worst in nation — and trending in the wrong direction, new study says', *The Seattle Times,* 14 December. Available at: https://www.seattletimes.com/seattle-news/transportation/washington-state-drivers-5th-worst-in-nation-and-trending-in-the-wrong-direction-new-study-says/
(Accessed: 28 October, 2020).

[2] https://en.wikipedia.org/wiki/Pioneer_Square,_Seattle
(Accessed: 28 October, 2020)

Ultimately, the citizens of the city benefit from the final analysis by way of announcements to help understand when additional caution is needed and knowing that the SCPD is equipped with the necessary information to respond to accidents appropriately.

c. Data Description

SCPD has provided their hired data scientist with all of the traffic collision data from 2014 to October 2020. It includes 37 attributes and is labeled by severity of the accident. Within the attributes is the following information for each accident:

- Collision address
- Date and time
- Type of intersection where the accident occurred
- Severity of the collision (unknown, property damage, injury, serious injury or fatal)
- Number of people or objects involved
- Driver state (distraction, under the influence, speeding)
- Weather
- Road condition
- Light condition

As weather and road conditions will help the police to send out public warnings or announcements to help caution drivers on days where more severe accidents are likely, the initial focus of the models created will be on these attributes.

2. Data Understanding

a. Detailed Data Description

Upon review of the data, it was noted that there were 195,046 accidents within the dataset and they spanned across all of the severity levels (0-unknown, 1-property damage, 2-injury, 2b-serious injury, and 3-fatality). However, the data was unbalanced such that the majority of accidents were property damage:

Severity level	Number of accidents in dataset
Unknown	2
Property Damage	133,818
Injury	57,829
Serious Injury	3,058
Fatality	339

It was also noted that the majority of accidents were occurring on clear weather days (114,788) and those with dry road conditions (128,573), which may not be as expected by the SCPD. However, we continued with the preparation of the data in order to determine if the weather and road conditions may still influence the severity of the accidents occurring if not the frequency of occurrence.

b. Data Preparation

To prepare the data for analysis, the following steps were taken:

1. Delete inconclusive (unknown or other) accident data within our dependent variable (severity level) or either of our independent variables (weather and road condition) as these would not be helpful in determining the impact on accidents.

2. The weather types and road condition codes were mapped to numerical values for calculation purposes:

Weather	Number Assigned
Clear	0
Partly cloudy	1
Overcast	2
Raining	3
Fog/smog/smoke	4
Blowing sand/dirt	5
Sleet/hail/freezing rain	6
Snowing	7
Severe crosswind	8

Road Condition	Number Assigned
Dry	0
Wet	1
Sand/mud/dirt	2
Standing water	3
Snow/slush	4
Oil	5
Ice	6

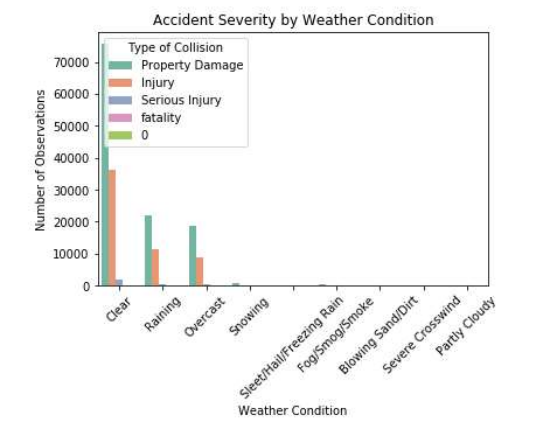
3. The severity codes were altered to create all numeric values for calculation purposes (1 – property damage, 2 – injury, 3 – serious injury, and 4 – fatality).

4. Finally, the data was balanced using downsampling and upsampling methods to avoid highly skewed data towards lower severity accidents (10,000 severity 1, 3,007 severity 2, 3,000 severity 3, and 1,000 severity 4).

3. Methodology of Analysis

a. Initial Data Analysis

Confirming our understanding of the initial data observations, a bar graph was created to show the distribution of accident severity across the various weather conditions.



Initially, the correlation between the weather, road condition, and severity of the accidents was reviewed. It was noted that the weather and road condition were positively correlated but that the severity of the accidents did not contain an initial linear correlation with either of these attributes. Upon reviewing the p-value and Pearson coefficient, this understanding was confirmed.

Next, the K-nearest neighbor and Decision Tree modeling methods were chosen as potential models that could better predict the severity of accidents based on these attributes.

b. Machine Learning Models

The first model created was a K-nearest neighbor model using both independent variables of weather and road condition. It was created with a train/test split of 80/20 and the optimal k-value was determined to be 5.

The second model created for comparison, was a Decision Tree. Again, both independent variables were used in this model and it was created with a train/test split of 70/30.

4. Results

a. Model Accuracy Review

Through review of the two models, a comparison was completed as to the accuracy of each to determine which would be more accurate in determining the severity of accidents. Both were very close in their accuracy; however, KNN was a slightly better with the following accuracy percentages noted:

Machine Learning Model	Accuracy
K-Nearest Neighbor	58.9%
Decision Tree	58.5%

5. Conclusion

Based on the numbers identified above, the conclusion is that that particular weather and road conditions only have somewhat of an impact on the occurrence of accidents and the severity of any property damage or injuries that may result. The best model to help the SCPD with the prediction would be the KNN model identified.

In addition, based on analysis of the data, the Police Department may be interested to learn that the majority of accidents already occur on clear weather days and those with dry road conditions. This suggests that drivers may already be applying caution or avoiding the roads on bad weather/road condition days and/or are letting down their guard on clear weather days. This may be useful in the type of public messaging they would like to pursue to reduce accidents.

6. Further recommendations

Further analysis could be undertaken with this data if the Police Department would like to investigate additional aspects for deployment of resources to accidents beyond just notification of public for precautionary measures. Additional models could look at some of the information generally available within an initial accident report call, such as: number of people/objects involved, intersection type, crosswalk involved, lane in which the car was when the accident occurred, day/time of accident, and the state of the driver (speeding, under the influence, distracted). Although this analysis would not help in making specific warnings to the public regarding a specific accident threat any given day, it may help them to better deploy resources to each accident with the ultimate intent of assisting injuries and reducing fatalities by providing timely and proper care.

Thank you for reading this report.