Bertel, David

bertelmendoza@gmail.com

Abstract

This report summarizes the analysis of a data set containing collisions in the city of Seattle. This report presents a visualization of the data and a model to predict the severity of the collisions depending on certain conditions.

predicting collisions severty in the city of seattle

Coursera capstone project

Table of Contents

[BUSINESS UNDERSTANDING 2](#_Toc55219214)

[Background 2](#_Toc55219215)

[Problem 2](#_Toc55219216)

[Interest 2](#_Toc55219217)

[DATA UNDERSTANDING 2](#_Toc55219218)

[Veracity 2](#_Toc55219219)

[DATA PREPARATION 4](#_Toc55219220)

[INITIAL VISUALIZATION 5](#_Toc55219221)

[MODELING 7](#_Toc55219222)

[EVALUATION 8](#_Toc55219223)

[DEPLOYMENT 8](#_Toc55219224)

# BUSINESS UNDERSTANDING

## Background

The city of Seattle has been gathering traffic information about collisions and their severities. It has by now a dataset with these accidents and some features associated to the accidents. These features may be used to identify correlation or causation depending on the features in the dataset, which we are going to investigate in this report.

## Problem

The worst outcome of an accident is casualties. Even if there are no casualties, normally when an accident occurs the people involved in it have to wait for the police to clarify the facts before being able to remove the cars. This process causes delays in the traffic.

## Interest

All people using a car are subject of traffic delays and therefore would benefit if there is an improve in the traffic i.e. if it would be possible to predict delays and best routes depending on certain features. On the other hand, companies developing autonomous cars can benefit from the outcomes of this study since it would help the to identify high risk situation or areas. This information about risk would be very valuable for insurance companies.

Last but not least, being able to predict the severity of accidents on certain areas may be useful for the city of Seattle to deploy emergency services or even a proper traffic infrastructure.

# 

# DATA UNDERSTANDING

## Veracity

Before starting any analysis of the data, the first step is to check the veracity of the data. In this case I am using the link provided in Cousera which contain the following dataset:

|  |  |
| --- | --- |
| *Data Set Basics* | |
| Title | Collisions—All Years |
| Abstract | All collisions provided by SPD and recorded by Traffic Records. |
| Description | This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present. |
| Update frequency | Weekly |
| Keywords | SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle |
| *Contact information* | |
| Contact organization | SDOT Traffic Management Division, Traffic Records Group |
| Contact person | SDOT GIS Analyst |
| Contact Email | DOT\_IT\_GIS@seattle.gov |

The data set includes a description of all the attributes in the data set. Link to the description of the data set: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

An initial look at the data set is always important. To check the shape and format used in the attributes:

A picture containing table

Description automatically generated

A picture containing graphical user interface

Description automatically generatedTable

Description automatically generated

# 

# DATA PREPARATION

Before starting with any kind of predictions it is important to know the dataset and the format of the information contained in it. Although it is not recommended by some data scientists, because it could bring biases to the analysis, a preliminary visualization of the data helps to have a first idea of the problem. Some important features that are going to be used are described in text. Using proper transformation, I converted them into a number

*Table

Description automatically generated*

Table

Description automatically generated

Attributes

After analyzing the information in the dataset, I concluded that reducing the prediction of the severity of an accident can be reduced to some particular conditions: WEATHER, ROADCOND and LIGTHTCOND. These features represent the weather, road condition and light condition respectively.

# INITIAL VISUALIZATION

Although it is, by some, not a recommended practice because it may introduce biases, an initial visualization of the data is useful to understand the problem we are trying to solve. There some cases where just the visualization is needed to find a solution to the problem.

In the first place may be useful to identify where are these collisions actually happening. With the use the library *folium* I displayed the collisions in the city of Seattle. As expected, these collisions happen mostly in the city center. This information will help us to draw some conclusions at the end of the report.

Map

Description automatically generated

In the map I displayed a representative number of incidents (500). In the data set there is some more detailed information about the location of these collisions. From the data we can see that only collisions type 1 and 2 are registered (1 = “property damage” and 2 = “injury”) and that most of the collisions do not occur in intersections but in address type Block.

Graphical user interface, text, application, email

Description automatically generated

From the initial visualization another important finding is the fact the number of incidents has being decreasing in the last years. There a particular decrease in 2020 probably due to COVID that has force many people to work from home. Being Seattle a city home of Microsoft and Amazon, it is probable that a lot of the jobs related with these two companies can be done from home.

Chart, line chart

Description automatically generated

# 

# MODELING

To avoid some attributes having more importance than others in postprocessing I used the *sklearn* to normalize the data:

Graphical user interface, text

Description automatically generated

After normalizing the data the next step is to define training and test data. I have taken 80% of the data set for training and the rest for testing the model:

A picture containing text

Description automatically generated

For training the model I found that the optimal value of k is 4. Giving a good accuracy in the results.

Graphical user interface, text, application, email

Description automatically generated

# 

# EVALUATION

To analyze the data set I used the K- Nearest neighbor method. By reducing the prediction to 3 features I reached an accuracy of the model of around 70%. This information would surely help all the stakeholders interested in the prediction. As mentioned before, 20% of the data set information was used to test the model.

# DEPLOYMENT

The city of Seattle has been gathering information about accidents and their severities. It has by now a dataset with these accidents and some features associated to the accidents. These features may be used to identify correlation or causation depending on the features in the dataset, which we are going to investigate in this report.

The worst outcome of an accident is casualties. Even if there are no casualties, normally when an accident occurs the people involved in it have to wait for the police to clarify the facts before being able to remove the cars. This process causes delays in the traffic.