Serious Traffic Injuries in Seattle
A Data Analysis
9/8/2020

Colin McLean

Table of Contents

Introduction

Background:

Seattle is the largest city in the Pacific Northwest of the United States. It has a population of around 750,000 people. Being a big city, it has a lot of cars, estimated at nearly 500,000 cars in the city! Because of this, there are naturally a lot of accidents which occur. Seattle averages about 200 severe injuries per year due to vehicle accidents.

Problem:

An analysis of data behind these accidents may help identify the contributing factors that lead to severe injuries. This project's goal is to identify factors which predict serious injuries in vehicle accidents in Seattle.

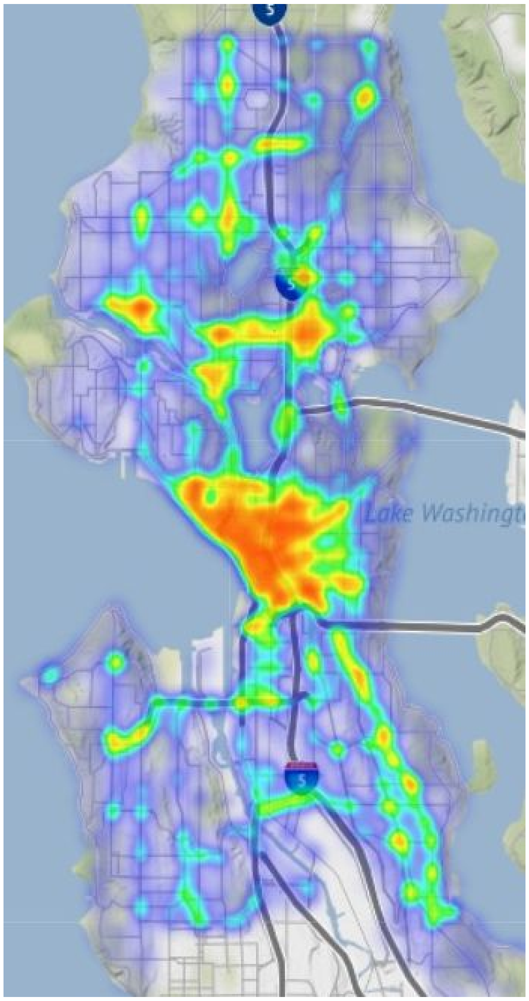
Interest:

Seattle has a program called "Vision Zero" where their stated goal is to end traffic deaths and serious injuries on city streets by 2030. I think that this project could help them accomplish that goal and it would fit perfectly with what their aims are.

Methodology

Data Source:

The dataset I will be using for this project will be a set provided by Coursera which comes from the Traffic Records of the Seattle Police Department. It records various data about collisions which occurred in the city since 2004 https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf



Heat Map of Seattle traffic accidents

Data Preprocessing:

The dataset has a large amount of information, including 39 columns of data. A lot of this data was redundant or excessive so I had to simplify it.

‘INTKEY’, ‘EXCEPTRSNCODE’,’EXCEPTRSNDESC’, ‘EXCEPTRSNDESC’’ and ‘SDOTCOLNUM’ columns all had significant amounts of missing data so they were dropped.

‘LOCATION’, ‘ST_COLDESC’, ‘CROSSWALKKEY’, ‘SEGLANEKEY’, ‘PEDROWNOTGRNT’ I determined were not helpful variables for the scope of my project so I removed them as well.

‘STATUS’, ‘REPORTNO’, ‘COLDETKEY’, ‘INCKEY’, and ‘OBJECTID’ were all columns for identifying the row and specific accident. These were for their records and are not relevant to the data model I’m using.

‘SEVERITYDESC’, ‘JUNCTIONTYPE’, ‘PERSONCOUNT’, ’PEDCOUNT’, ’PEDCYCLOUNT’, ‘SEVERITYCODE.1’, ‘HITPARKEDCAR’, ‘SDOT_COLDESC’, and ‘SDOT_COLCODE’ were all redundant columns that I dropped. For example PEDCOUNT tracked the number of pedestrians in the accident, but that detail is already tracked in COLLISIONTYPE.

Some of the data had missing data in the form of NA. I dropped all rows that had NA values for the columns ‘X’, ‘Y’, ‘ADDRTYPE’, ‘COLLISIONTYPE’, ‘ST_COLCODE’, ‘WEATHEr’, ‘ROADCOND’, and ‘LIGHTCOND’.

Some of the data had values varying from 2,1,0,Y,N,NaN to represent boolean values. In the cases of ‘SPEEDING’, ‘INATTENTIONIND’, and ‘UNDERINFL’, I converted all the values to be either 1 or 0.

Feature Selection:

Because I wanted to keep the data easily readable, I decided to make sure all my data values were simple boolean values which answered a question, for example “Was a pedestrian involved.” A 1 value means true, a 0 value means false.

To do this I had to split and simplify a lot of the columns into new columns.

My target value that I’m trying to predict with the data is the severity of the accident. This was tracked by the ‘SEVERITYCODE’ in the data which was 2 for serious injury, and 1 for non-serious injury. I converted this to a new column of ‘Severe’ which represented a 1 for a severe injury, and 0 for non-severe.

‘Intersection’ is a feature representing whether the accident occured at an intersection or a block. This data was extracted for the ‘ADDRTYPE’ column. All rows which had ‘Interesection’ as the value were assigned the value of 1 in this new feature column.

‘Inattention’ was a simple value that was already in the dataset which represented whether the accident was due to the inattention of a driver or not.

‘UnderInfluence’ was another simple value already in the dataset which represented whether the driver was under the influence of alcohol or drugs.

‘Speeding’ was another simple value already in the dataset which represented whether the driver in the accident was speeding.

‘BadRoads’ was a value that I extracted from the ‘ROADCOND’ value from the dataset. If the road condition was anything other than ‘Dry’, the value of BadRoads was set to 1.

I extracted 8 different values from the dataset’s ‘COLLISIONTYPE’ value which stored a value describing the type of collision. Each of these values I assigned to a new column as a boolean value. ‘Parked Car’, ‘RearEnded’, ‘Pedestrian’, ‘Cycles’, ‘LeftTurn’, ‘RightTurn’, ‘Sideswipe’, and ‘Angles’ were the new features I created.

The last two features I extracted were from the ‘INCDTTM’ column which stored the date and time of the accident. Using pandas’s datetime feature, I was able to convert the date into day of the week and then convert the day of the week into a boolean value of whether it was a weekend or not. This is stored as ‘Weekend’

The other datetime feature was ‘Night’ which was extracted from the time using the hour attribute. If the hour was less than 6 or greater than 20, then it was assigned 1 for ‘Night’. This means 8 PM to 6 AM was the requirement for ‘Night’.

So the final features to be used for my modeling were ‘Intersection’, ‘Inattention’, ‘UnderInfluence’, ‘BadRoads’, ‘Speeding’, ‘Parked’, ‘RearEnded’, ‘Pedestrian’, ‘Cycles’, ‘LeftTurn’, ‘RightTurn’, ‘SideSwipe’, ‘Angles’, ‘Weekend’, and ‘Night’. And the target variable was ‘Severe’. All these columns can either be 1 or 0.

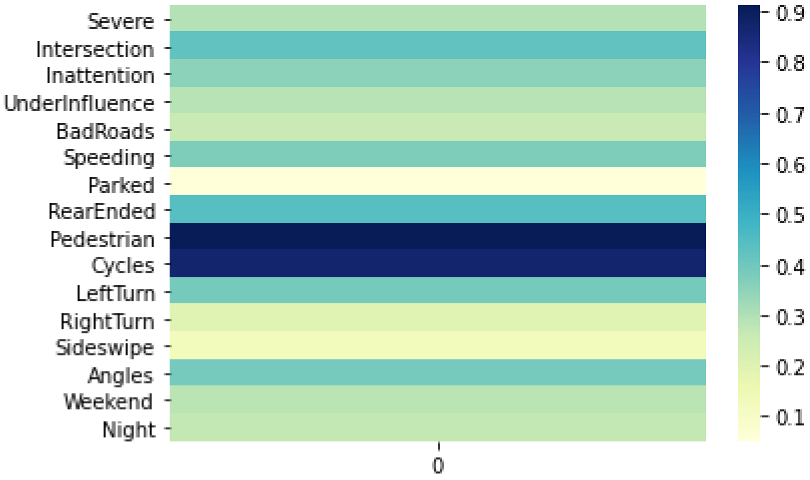
Results

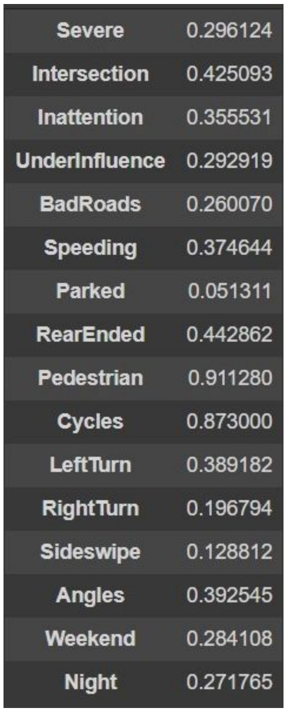
Analyzing the data:

First, I needed to find how common severe accidents were among the data. I used the ‘Severe’ column which was 1 or 0 to find the mean. The mean of this value is percent of accidents which were serious. The value I found was 29.6%. So around 30% of all accidents in the dataset were serious.

Next I wanted to find what percent of accidents were severe among cases where one of the features was true. For this, I had to filter out all rows where the feature's value was 0, and then find the mean of the 'Severe' column in the remaining rows. And repeat this for each feature.

I collected this value for each of these and you can see their comparison in the chart below:





Severe	0.296124
Intersection	0.425093
Inattention	0.355531
UnderInfluence	0.292919
BadRoads	0.260070
Speeding	0.374644
Parked	0.051311
RearEnded	0.442862
Pedestrian	0.911280
Cycles	0.873000
LeftTurn	0.389182
RightTurn	0.196794
Sideswipe	0.128812
Angles	0.392545
Weekend	0.284108
Night	0.271765

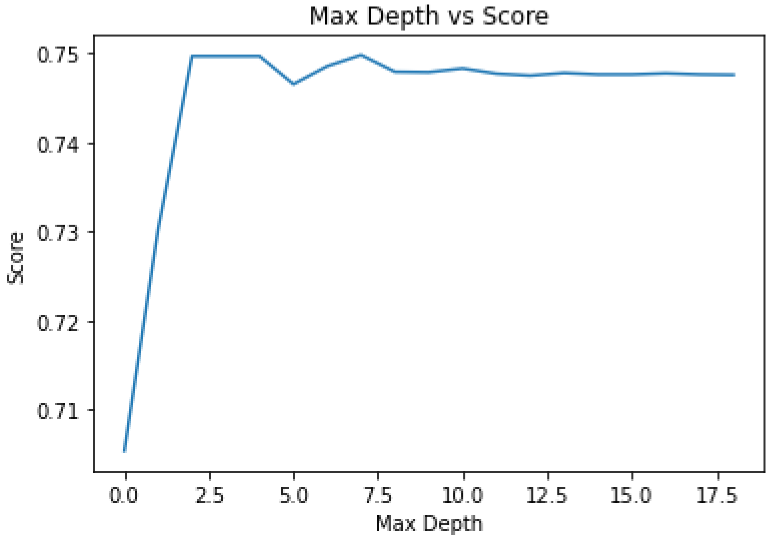
As you can see accidents involving pedestrians and cyclists were by far the most dangerous sort of accidents. 91% of accidents involving a pedestrian were serious! 87% involving cyclists were serious as well. That is compared to the 30% of all accidents which are severe.

Unsurprisingly accidents involving parked cars were the least severe. Only 5% of accidents involving a parked car were serious. Left turns accidents were about twice as dangerous as right turn accidents. Speeding, Intersection, RearEnd, and Angles were the other notable features of more severe accidents.

Modeling:

The model I chose to use for this data was a Decision Tree classification model.

To tune the model, I ran tests of it with different hyperparameter values from 1 to 20 for max depth.



From these tests I was able to determine that the optimal max_depth value for the model would be 3.

So I ran the model on the training data using sci-kit learn Decision Tree Classification model and the max_depth = 3. When I tested this model against my test data, I was able to determine a score of .75.

Discussion

Some of the results were surprising. For example how little bad roads, or being under the influence seemed to have on the severity of an accident. I would have expected that these would have led to worse accidents. The weekend and nighttime also seemed to have very little to do with the severity of accidents.

The most dangerous types of accidents were by far pedestrian and cyclist accidents. This makes sense as they have very little protection in the case of an accident relative to a person in a vehicle.

Accidents which seemed to involve very low levels of injury were accidents involving parked cars, and accidents involve sideswipes.

Conclusion

The goal of this project was to analyze the features of car accidents in Seattle which led to severe injuries and develop a model which predicts severity.

I was able to analyze some of the key features involved in accidents being severe, which were pedestrians/cyclists being involved. And the features in those least severe, which were accidents involving parked cars.

I developed a decision tree classification model which could be used to predict the severity of an accident based on the features. The features are very easily identifiable factors which can be gathered from any individual who saw the accident. This model can help find which are the most dangerous combinations of factors for driving.