Capstone project Report

Mohamed Osama

Car Accidents

Introduction

Intro

Accidents happen all the time on the road. However, each accident can vary in how damaging it is. What is the cause of the most and least severe accident? We will be looking at accidents in Seattle to see if we can figure out if an accident would be severe or not based on other statistics.

Problems

There are many causes for an accident, but how common is it? What can we do about it? If we have the right information, can we lower the amount of accidents done on the streets?

Data and cleaning

Data source

There is an open dataset for traffic in Seattle found here and a reference site attached with it. I also used this site for geodata and shape of Seattle. However, I used a much more simplified version of that dataset from the course instructor.

Data cleaning

On the CSV file I used, there was a lot of missing data. Since the dataset is pretty large, I decided to remove all of missing data, as well as columns that contained a ton of missing data. Afterwards, I looked through each data to see its values and how I personally think it'd influence the model. One of the harder things on deciding what to keep and what to not keep is the direct influence.

A lot of the data aren't exactly numbers, and they're more descriptions that we'll have to transcribe into numbers in the future for the computer to read. However, since I cannot see the correlation between the words and actual severity index, I had to pretty

much guess and play around with the features until I got a model I liked. There were also a lot of 'unknown' features, I decided to remove all of them, as I don't think they'll help my model predict it at the end.

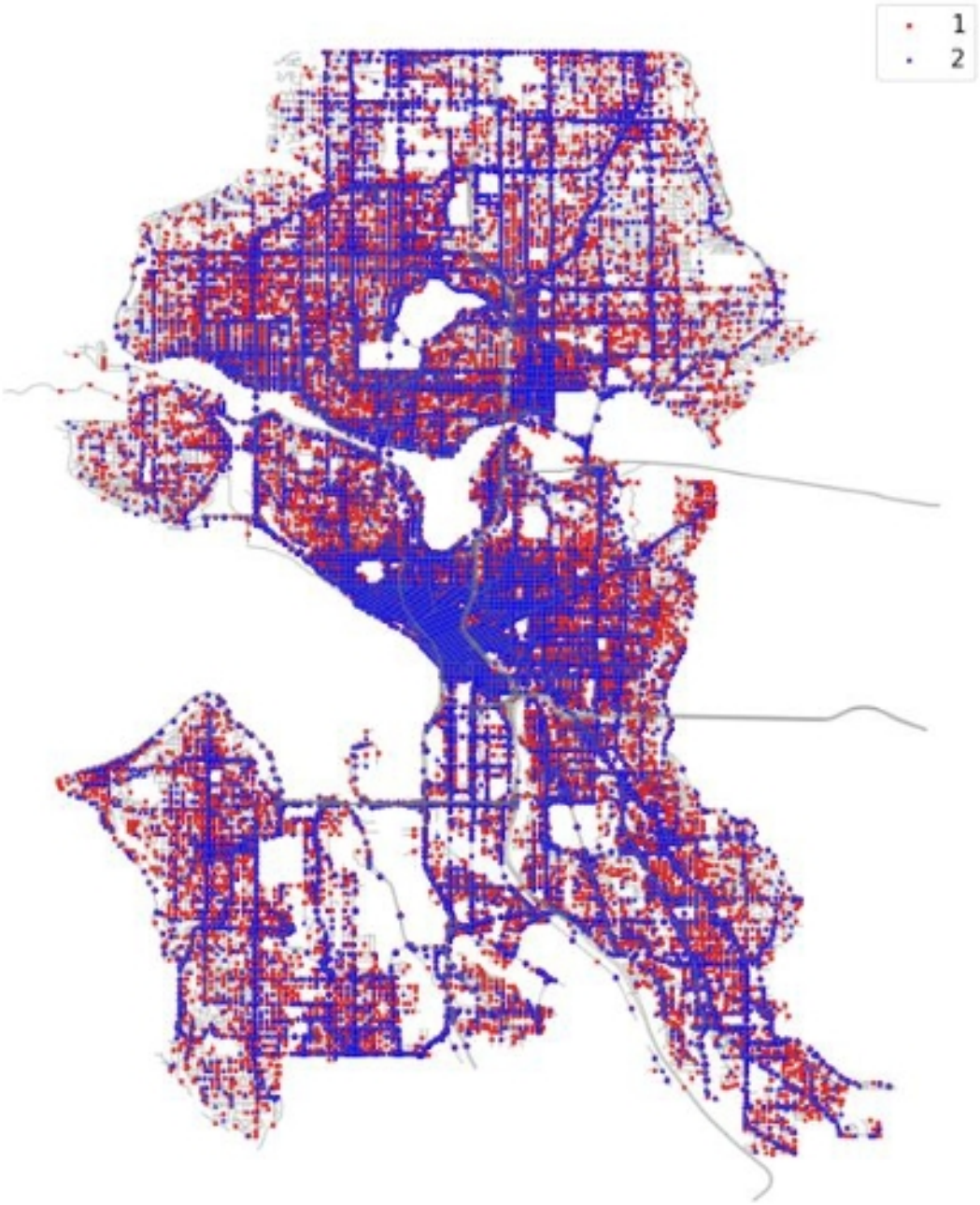
Feature selection

After cleaning up the data, there were 180062 total entries and 27 features left. Looking at some of the features, there were ones I knew that weren't needed in the actual model itself. Below is a table of all the features that I kept, discarded and the reason why I discarded them.

Feature kept	Feature dropped	Dropped reason
ADDRTYPE, PERSONCOUNT,	OBJECTID, REPORTNO, INCKEY, COLDETKEY	All keys
PEDCOUNT, PEDCYLCOUNT,	INTKEY, SPEEDING, PEDROWNOTGRNT, EXCEPTRSNCODE, EXCEPTRNSNDESC,	Many values missing
VEHCOUNT, SDOT_COLCODE,	SDOTCOLNUM, INCDATE, INCDTTM	Missing numbers, incomplete data.
UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, HITPARKEDCAR, ST_COLCODE	SEVERITYCODE.1 LOCATION	redundancy to Severitycode feature Not needed to predict how severe the model is.

Exploratory Data Analysis

Map of Seattle with accidents data:



Red represents lower severity, and blue represents higher severity.

While there are a lot of accidents throughout the entire city, a lot of blue seems to be concentrated on the center area of Seattle.

We can note that the dense area contains a lot of intersections, which probably causes a larger amount of severe accidents. Also, most of the red dots seem to be very sporadic

in where they are. All over the city there are red, and blue is mainly where roads are.

Relationship between weather and severity code:

Severity level 1:

Clear	70858
Raining	20608
Overcast	17690
Snowing	648
Fog/Smog/Smoke	353
Other	161
Sleet/Hail/Freezing Rain	80
Blowing Sand/Dirt	30
Severe Crosswind	17
Partly Cloudy	2

Severity level 2:

Clear	34954
Raining	10749
Overcast	8469
Fog/Smog/Smoke	180
Snowing	162
Other	74
Sleet/Hail/Freezing Rain	27
Blowing Sand/Dirt	12
Severe Crosswind	7
Partly Cloudy	3

When comparing the weather to severity levels, it doesn't seem to have too much of a difference. Granted, since the data is really skewed towards severity level 1, there would be a lot more datapoints in general for it. However, I would assume that other trends would increase the severity (like rain, or other abnormal weather conditions) but this is just untrue. The majority of all accidents seem to occur when the weather is completely clear. Let's compare these two levels with other values.

Relationship between Collision type and severity level:

Severity level 1:

Parked Car	29806
Angles	20274
Rear Ended	17796
Other	15391
Sideswipe	14700
Left Turn	7980
Right Turn	2185
Head On	1082
Pedestrian	629
Cycles	604

Severity level 2:

Rear Ended	13817
Angles	13358
Other	5614
Pedestrian	5610
Left Turn	5316
Cycles	4565
Parked Car	2569
Sideswipe	2355
Head On	844
Right Turn	589

The relationship for this is a lot more different than weather. The highest amount of accidents for severity 1 is hitting a parked car. Whereas for severity 2 the highest is getting rear ended, but for both 1 and 2 the 2nd highest value are angles. One big thing to note is that accidents with pedestrians involved seem to lead to a higher level of severity.

Relationship between under the influence and severity level

Severity level 1:

0	105160
1	5287

Severity level 2:

0	51213
1	3424

By checking the ratios between the two, for severity level 1 there are roughly 20 accidents that doesn't involve intoxication to 1 accident that does. Whereas for severity

level 2, there are roughly 14 accidents that doesn't involve intoxication. There isn't too much to reference here in terms of severity. You would think that more people would be intoxicated vs not for severity two but the data seems to prove otherwise. That would mean that the higher severity levels involve things outside of alcohol.

Relationship between Light conditions and severity level

Severity level 1:

Daylight	72127
Dark	33050
Dusk	3634
Dawn	1524
Other	112

Severity level 2:

Daylight	37382
Dark	14560
Dusk	1868
Dawn	791
Other	36

Just like the previous comparison, this doesn't seem to have too much influence over whether the severity level is one or two. However, we can see that the ratio between daylight and dark is roughly the same.

Relationship between Address type and severity level

Severity level 1:

0	76182
1	34265

Severity level 2:

0	27908
1	26729

where 0 represents block and 1 represents intersection

This is where things seem more interesting. There is a clear distinction of the address type effecting the severity level. Where most of the level 1 accidents occur on the block, it's about split even on level 2. This means that more severe accidents can occur on either one

Model Selection

We will be mainly focusing on classification modeling for this as we want to

predict the outcomes as labels (level 1 or 2). While regression models can be used, it generally wouldn't show the full picture (as we will get to see in the later stages when we try it). The main model that we'll be using is a decision tree. It would look through all the features and go through each one until it reaches a level 1 or 2 rating at the end, kind of like a flowchart.

Decision tree vs Regression

The decision tree's accuracy was the highest of the models I've tried. The dataset is a little large for SVM and Kmeans so we will be ignoring those for this case in particular. Below is the classification table between the two models:

Model	Accuracy	F1-score
Decision Tree	68.5%	68.8%
Logistic regression	65.1%	65.4%

Conclusion

Of the features that I have on the database, if you input the same amount of features inside it'll predict whether the severity level would come out to be one or two. Given the dataset though the accuracy level makes a lot of sense. However, since the models both have a decently high f1-score, this means with just a little extra in the dataset we can improve on the prediction of severity level based on a few features.

Things to do in the future

Based on the graphs, the highest percentage of high severity accidents are from angles. I believe that if we were to try and improve on the viewing of angles at intersections, we would be able to lower the overall severity of accidents. The highest percentage for severity one is parked cars. This is something that can be fixed if drivers had to undergo more practice with parking and/or driving around cars before obtaining a license.

Also, there are some features that I unfortunately couldn't use. One of the features I really wanted to use is time. I wanted to see if the day of the week mattered, as well as time of day. There's probably a larger amount of accidents caused during rush hour. However, since the data isn't fully complete for that, I couldn't use it as one of the

features. If the time stamp is better updated in the future datasets, we may be able to see a clear distinction between time and accident severity.