Predicting Injury of Car Collisions Reported

Christian Miraglia

September 6, 2020

1. Business Case/Introduction

The city of Seattle wants to improve the response of their first responders when a car accident is reported. It is not feasible for police, EMT and fire to jointly respond to all accidents. The current procedure is for a police officer to respond and request further assistance according to the severity of the accident. This often means delay in treatment of the crash victim or victims.

The city would like to be able to predict if a reported crash is severe enough to require the response of a severe crash response team (police, fire and EMT). The Seattle Police 911 is the primary Public Safety Answering Point (PSAP). The PSAP Severity Collision Model would consider weather, road conditions, time of the day and location. The model will provide a prediction on the severity of the accident. The PSAP will coordinate with the Seattle Fire Alarm center accordingly.

2. Data

The Seattle PD, like all police departments, documents accidents in the form of vehicle accident reports. These reports are summarized in a log. The current accident log is available in .CSV format which makes it convenient as a data set for developing a model. The current dataset is made up of 38 characteristics of 194, 673 accidents. The data set characteristics vary from administrative data to specific data on the accident.

The goal of our model is to predict the severity of the accident. The data set includes the severity of the accident. This will be used as our labels. The severity is represented descriptively in text as we all coded as either 1 -not serve (no bodily injury) or 2 -severe (bodily injury). Due to possible transcribing error and just the general ease of using numerical values, the coded severity will be used as our labels and the text description will be omitted.

As mentioned previously, our data set also includes administrative data that is not characteristic of the accident. Data like report number, Object ID, Report status can be removed from our data set.

Redundant data like severity code which appears twice, as well as data that is in text as well as encoded can me reduce to just encoded. Careful review of the text and encoded data must be done to ensure the coding is adequate.

Data that describes the accident but not the environmental conditions that may have led to the accident can also be deleted. During the initial data review s, the characteristics of the accidents will be required to see if a more detailed severity code can be derived based on characteristics of the accident.

The conditions that lead to accident can be reduced to **weather, road conditions, Lighting conditions (night streetlights or without or day light), day of the week and location.** Location can is further refined to whether the accident occurred in mid-block intersection or at a driveway junction.

Characteristics that contain no data or missing data will be removed. The data set will need to be balance. No severe (code 1) out numbers Severe (code 2) by a ratio of 2.3:1. A Balanced will be done by both deleting and duplicating. Models will be creating using a mix of both balancing methods.

3. Methodology

3.1 Coded characteristics

The goal is to identify accidents that are called in that are severe enough to require medical attention. The severe cases in our data set are ones where there was bodily injury. Day of the week, weather, road conditions, lighting conditions and street locations can be analyzed by looking at the amount of severe vs not severe accidents.

Weather

Severity	1-overcast	2-raining	3-clear	4-snowing	5-fog	6-sleet	7-blow sand	8-high wind
1 (not severe)	67.8%	65.8%	67.2%	80.2%	66.5%	74.5%	71.4%	70.8%
2 (severe)	32.2%	34.2%	32.8%	19.8%	33.5%	25.5%	28.6%	29.2%

Date

Severity	1-Monday	2-Tuesday	3-Wednesday	4-Thursday	5-Friday	6-Saturday	7-Sunday
1 (not severe)	66.4%	66.4%	66.5%	66.1%	67.5%	67.7%	69.2%
2 (severe)	33.5%	33.5%	33.5%	33.9%	32.5%	32.3%	30.8%

Road Conditions

Severity	1-Wet	2-Dry	3-snow/slush	4-ice	5-Sand/mud dirt	6-Standing water	7-Oil (recently paved road or oil spill)
1 (not severe)	66.3%	67.2%	81.2%	75.7%	62.5%	73.4%	59.2%
2 (severe)	33.7%	32.3%	18.8%	24.2%	37.5%	26.6%	40.8%

Lighting conditions

Severity	1- Daylight	2- Dark- light on	3-Dark No streetlights	4- Dusk	5- Dawn	6- Dark Streetlights off	7-Dark Unknown lighting
1 (not severe)	66.0%	69.3%	77.0%	66.1%	66.1%	72.1%	62.5%
2 (severe)	34.0%	30.7%	23.0%	33.9%	33.9%	27.9%	37.5%

Location type

Severity	1- intersections	2- Block
1 (not severe)	56.2%	73.4%
2 (severe)	43.8%	26.6%

The ratio of severe to not sever is ~3:1. When we look at most of the characteristics, we see about the same ratio so there is no obvious condition that tells us if a sever accident is more likely.

The exceptions are snowing where severity is less than 20%, road conditions where these is oil normally caused by recently paved roads or some spill at 40.8% severe accident and location type intersection where the sever accidents are at 43.8%. This is the highest percentage of severe accident. While accidents mid-block are relatively lower than the 3:1 ration at ~2.6:1.

3.2 Latitude and Longitude data

Along with accident characteristics the data set also includes the longitude and latitude data of each accident. At first glance the severe accidents appear more evenly dispersed throughout the map. The severe accidents appear to be more concentrated in certain areas and some areas do not show as many sever accidents. We can assume that the location of the accident will help in determining the possibility that the accident reported will be severe. For example, the south west section of the city (circled in green) seems to have a higher concentration of not sever accidents as opposed to severe accidents. It is important to keep in mind that the data set to create the heat map was balance to remove any bias.

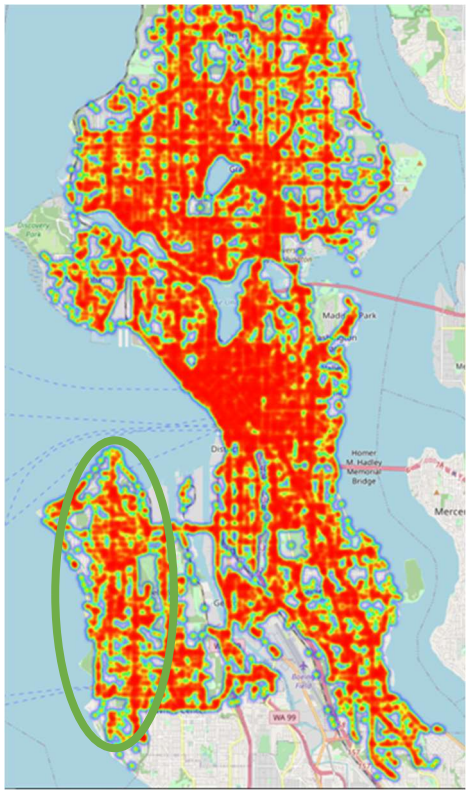


Figure 1 severe

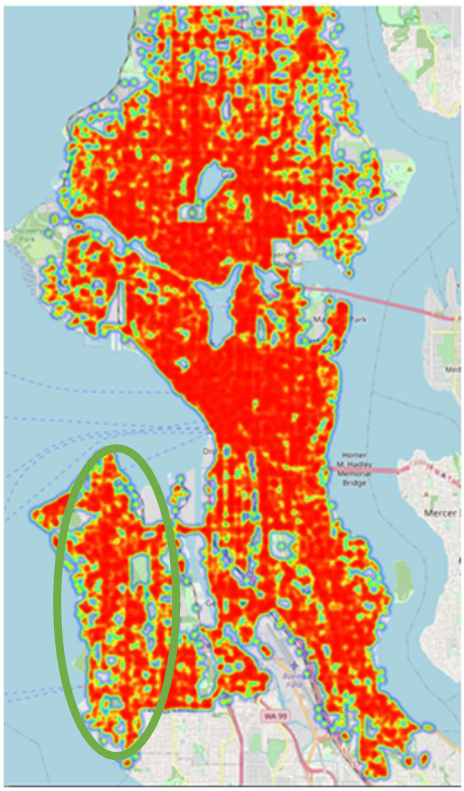


Figure 2 not Severe

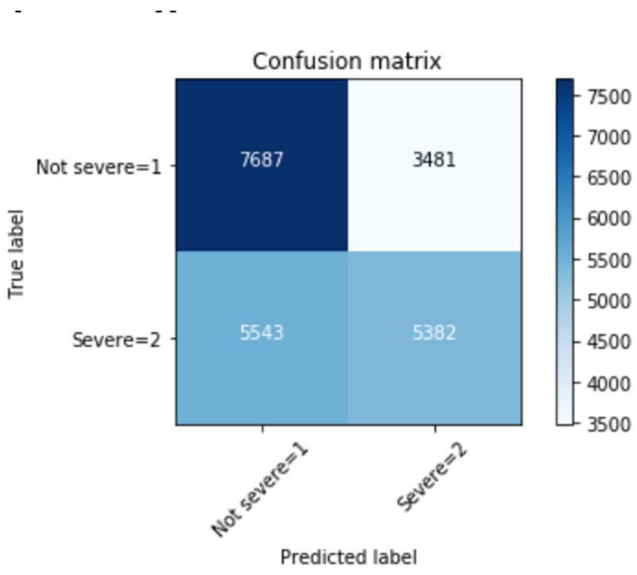
3.3 Logistic Regression Model

Our Label is based on either yes, the accident was severe or no, the accident was not severe and there were no bodily injuries. Logistic regression is a very good at categorial dependent variables.

After creating a logistic regression model using our balanced data set, we look at the accuracy of our model. Our model is trained using 80% of the data set and 20% for testing. We use Jaccard Index and Log loss to give us an idea of the accuracy.

Jaccard Index	.592
Log Loss	.676

A confusion is Matrix is also created to give us an idea as to whether a specific type of accident is more accurately predicted than another.

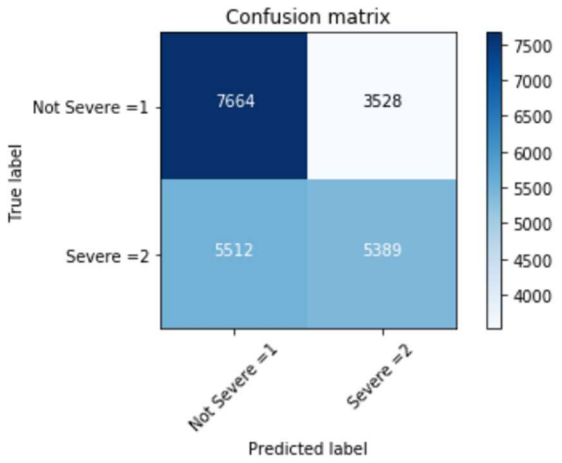


	precision	recall	f1-score	support
1	0.58	0.69	0.63	11281
2	0.60	0.48	0.54	10812
micro avg	0.59	0.59	0.59	22093
macro avg	0.59	0.59	0.59	22093
weighted avg	0.59	0.59	0.59	22093

SVM Model

A Support Vector Machine Model (SVM) is also applicable as a classifier algorithm. An SVM model was created using the same balanced data set. A Jaccard index and confusion matrix was used to check the accuracy of the model.

Jaccard Index	.591



```
                precision    recall  f1-score   support

           1       0.58      0.68      0.63     11216
           2       0.60      0.50      0.55     10877

   micro avg       0.59      0.59      0.59     22093
   macro avg       0.59      0.59      0.59     22093
weighted avg       0.59      0.59      0.59     22093
```

4 Results

Both the SVM model and logistic regression model gave us similar results.

Jaccard Index and Log Loss

	Jaccard Index	Log Loss
Logistical Regression	.59	.67
SVM	.59	-

f1-score, Precision and Recall

	f1 score		Precision		Recall	
	1-n/ severe	2- severe	1-n/ severe	2- severe	1-n/ severe	2- severe
Logistical Regression	.63	.54	.58	.60	.69	.48
SVM	.68	.50	.58	.60	.68	.50

We can see that the overall model is not very accurate. The precision is .60. When you consider there are only two possible outcomes 1-not severe and 2 severe the models do not necessarily help in predicting one accident type over the other. The recall shows better results for not severe accidents. This means if the model says it is not severe it is likely correct, or at least is a more accurate at making that prediction for that instance.

5 Discussion

The overall model is not as accurate as we would want but when you consider the goal of the model, it may still be deployable and can be improved.

It is important to remember the goal of the model. The city of Seattle wants to improve public safety by deploying EMT and ambulance when the accident is severe and not only police at first, who later assess if the accident requires EMT and ambulance.

The recall of the model when it comes to not severe accidents is a good indicator if only a police officer is required as an initial responder. Since that is more accurate, dispatchers can dispatch only police officers if the model predicts not severe. There is a good possibility that if the model predicts the accident is severe that when EMT and ambulance arrive they will find there is no need for them. This is better than the alternative of having a severe accident and having to wait for EMT and ambulance.

5.1 Reasons for in-accuracy

One possible reason is the lack of variety of data. As an example, even though we have 8 codes for weather the majority Over 70,000 of the over 110,000 accidents were coded as 3 (clear), followed by over 20,000 as code 2(raining).

Another possible reason is the fact that we only have two labeled categories, 1 not severe and 2 severe. In many cases the severity can range from a minor laceration to death or even be a minor accident where someone just claims they are injured with the hopes of making some insurance claim. There is no way to distinguish how bad the accident was with respect to the injury. This also is likely why severe accidents have a lower recall as compared to not severe accidents.

In order to create a more accurate model more detail on the type of severity would improve the ability to predict. For examples, a code for fatalities, immobilized victim, victims that can walk can be used to give better resolution.

Conclusion

In conclusion, our analysis shows that any condition can produce a sever accident, but certain conditions are likely to produce accidents that are not severe. One finding with respect tot not severe accidents is that they are more common than sever accidents on days when the weather is snow or ice and when there is slush and ice on the road. Something that is counter intuitive. Accidents that happen mid-block (not at an intersection are also less likely to be severe.

It is important to remember that the goal is to improve the response of emergency personnel in the case of a severe accident. Even though we cannot model sever accidents accurately we can predict accidents that are not sever. The model can still be deployed with the understanding that in may cases the EMT and ambulance will not be required for some accidents. Our data set was balanced but when we look at unbalanced dataset we cans see the actual not severe accident to severe accident ration is 3:1. At this rate it is likely SPD will deploy EMT and ambulance in 50% of the accidents called in and out of that only ~50% will actually require it. This is not ideal but better than the existing procedure of only deploying police at first or deploying EMT and ambulance to every accident which is not possible.