Predicting the Severity of Collisions Occurred in Seattle, WA

Juntao Wei

September 24, 2020

1 Introductions

1.1 Background

Seattle is a seaport city on the West Coast of the United State, it is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million[1]. As the city grows and roads get more congested, the number of vehicles expands hence the collisions escalate. Collisions would vary a lot under different circumstances; for example, types of vehicles involved including motorcycles, automobiles, buses etc.; types of crashes happened on the road like read-end collisions, T-bone accidents, sideswipe collisions, or even worse, head-on crashes and so-on. All data related to collisions that were occurred in Seattle is very helpful in preventing potential accidents, optimizing traffic managements, setting up new penalties towards distracting drivers.

1.2 Problem

Although many drivers are driving as cautious as they can, accidents are still inevitable. Once a collision happens, how should we determine the severity of it hence make the most efficient response?

It is important to forecast the severity of an accident to dispatch appropriate police forces and ambulances or even fire trucks to the scene. For example, if there are pileup car accidents involving multiple injuries, the insufficient number of ambulances might lose the best timing of saving lives. By providing the on-scene description of an accident by witnesses, the call center can use the model to predict the severity of the current one, therefore provide the best help in time without overspending public resources.

1.3 Interest

9-1-1 Dispatch Control Center in Seattle would be very interested in accurate prediction of severity of potential collisions. Others like publics would also be interested by avoiding key features in the prediction to keep themselves safe.

2 Data acquisition and cleaning

2.1 Data sources

The original dataset called Seattle SDOT collisions Data collected by City of Seattle is available on Kaggle: https://www.kaggle.com/jonleon/seattle-sdot-collisions-data

[1] "Seattle." *Wikipedia*, 19 Sept. 2020. *Wikipedia*, https://en.wikipedia.org/w/index.php?title=Seattle&oldid=979176007.

The city has an open data platform found here and they update their information according the amount of data that is brought in weekly. The timeframe for this dataset is from 2004 to Present.

2.2 Data cleaning

The data contains more than 220,000 records with 40 features. Obviously, not all information would be necessary when predicting the severity.

2.2.1 Date selection

City is developing rapidly. A block today might become a mall tomorrow; hence the road condition might change significantly. In addition, manufactures have also improved cars' crashworthiness, so under same circumstances newer cars would maintain the damage as small as possible.

In order to make the most efficient prediction of the severity of collisions, the data should select as the collision date is no older than year 2010.

2.2.2 Feature selection

Except for dependent feature "SEVERITYCODE" that needs to set as the goal, there were 39 independent features recorded in the dataset. As we are making predictions from a 9-1-1 dispatcher, information we input should come from a witness's perspective of view. For example, when someone reports to the police, he/she can only mention the location, number of injuries, how many and what kind of vehicles are involved, what types of collision happened, etc. In addition, the 9-1-1 dispatcher can always ask about the road, weather and light condition of the surrounding. Other information such as case ID, report number, collision code, whether the driving is speeding or violates traffic rules, are only available when police have done their investigations after a long time. Therefore, the model should only use information that is available immediately on the site.

Features Selected	Explanation
'X' 'Y' 'LOCATION'	Coordinates and locations of a collision
'ADDRTYPE' 'JUNCTION TYPE'	Types of road when a collision happens
'COLLISIONTYPE'	Types of collision
'PERSONCOUNT' 'PEDCOUNT' 'PEDCYCLECOUNT' 'VEHCOUNT' 'INJURIES'	Number of people, pedestrians , bicycles, vehicles, total injuries involved respectively
'INCDATE'	The date of a collision
'WEATHER' 'ROADCOND' 'LIGHTCOND'	Weather, road and light conditions during the time of the collision
'HITPARKEDCAR'	If a parked car is involved

All these features have correlations regarding the severity of the collision.

2.2.3 Check NULL data

Lots of data are missing. By simply dropping all rows with NULL features, the data is heavily imbalanced, as most of data are labeled with '1', '2' as the table shown below.

'SEVERITYCODE'	# of Records
0	2
1	69325
2	33408
2b	1654
3	182

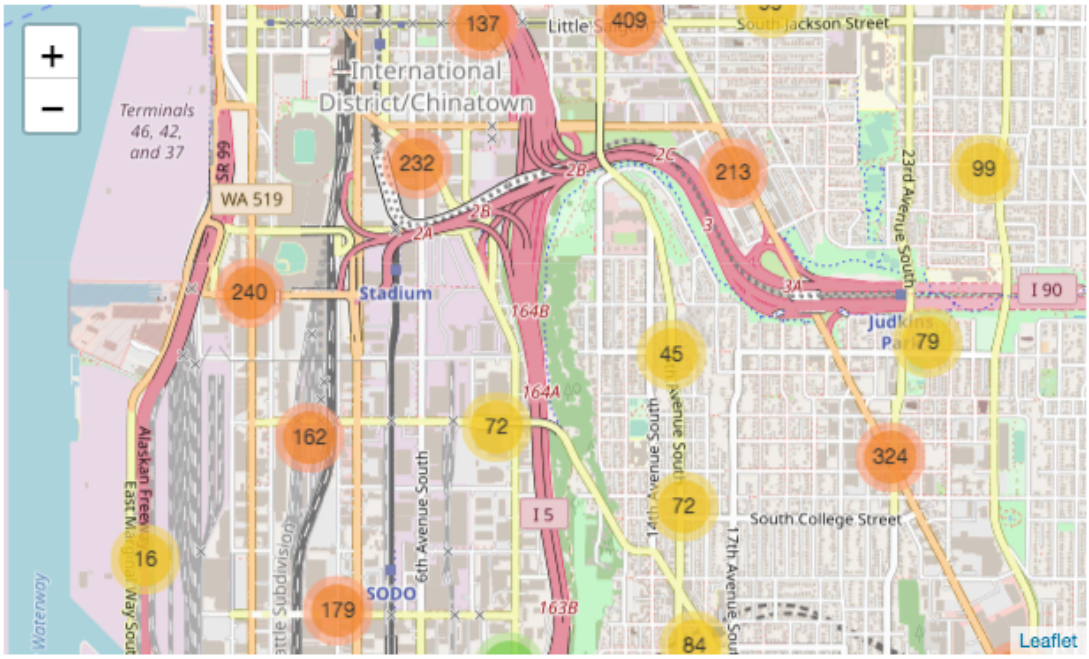
2.2.4 Data wrangling

As the training dataset should not have too many generated data, every group of same labeled data should have similar number of rows. Therefore, the final dataset being analyzed has about 1,700 rows of each group ('0', '1', '2', '2b'), except for group '3' which has only 200 rows by filling the NaN by most frequent or logical values. Group '3' is the minority group because most severe collisions do occur much less frequently.

Even though group '0' has size of 1,700, only two of them have all features filled. The rest rows are all missing 'COLLISIONTYPE', 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. Those missing values cannot be fulfilled since no other information is available, neither cannot find the pattern of them nor replace them with most frequent values.

3 Exploratory Data Analysis

3.1 Severity vs. Location



There were some multiple collisions happened in the same location, but their severity codes weren't uniform. In addition, all types of collisions happened everywhere, to simplify the model, location features will not be used.

3.2 Severity vs. Address and Junction type

Among two address types –Block & Intersection,

and five junction types – 'Mid-Block (not related to intersection)',

'At Intersection (intersection related)', 'Driveway Junction',

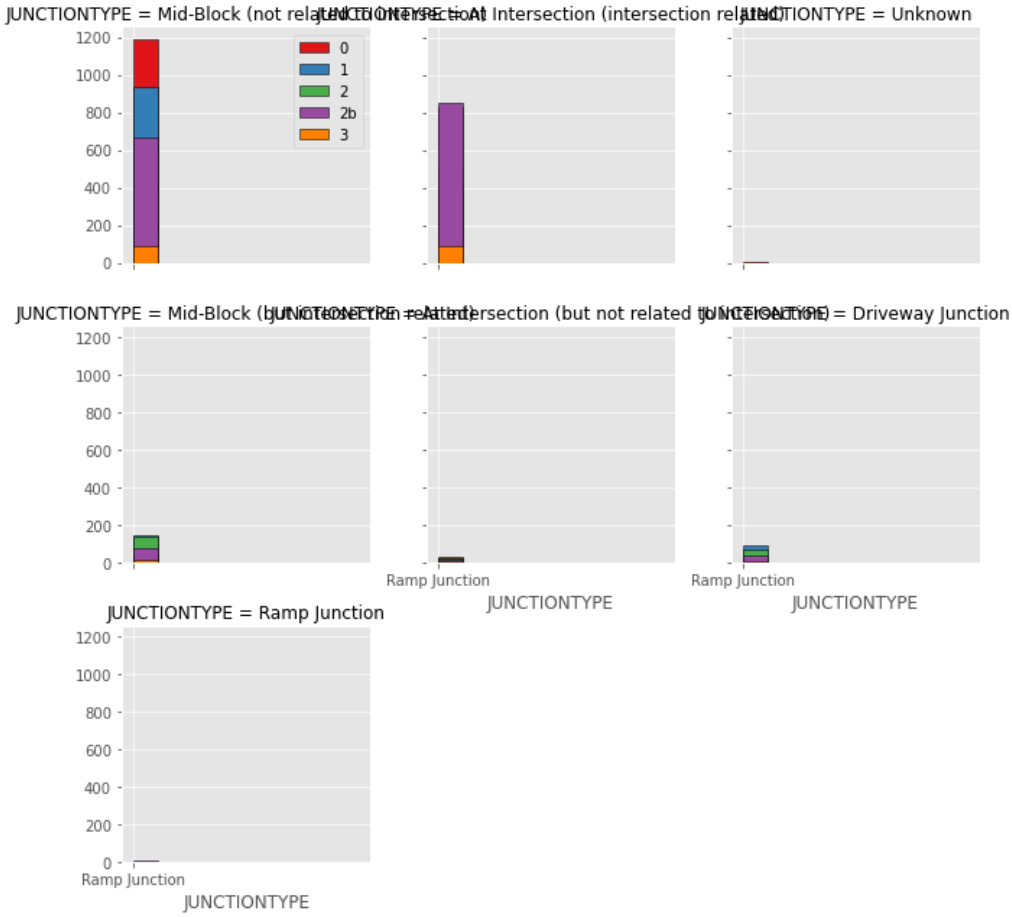
'Mid-Block (but intersection related)',

'At Intersection (but not related to intersection)',

'Ramp Junction',

'Unknown'

Compare the number of people involved in each types of junctions under different severity codes.



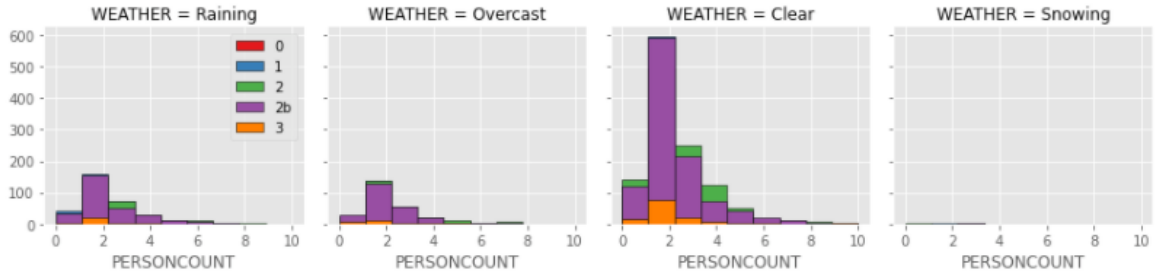
As the plot shown above, collisions with severity code 1 are more likely to happen on Mid-Block (not related to intersection. Collisions with severity code 2b are also more likely to happen at Mid-Block (not related to intersection and at Intersection (intersection related).

3.3 Severity vs. Weather, Road and Light condition

Remember that since group '0' have NaN values under these features, hence the analysis will not include this group.

3.3.1 Weather

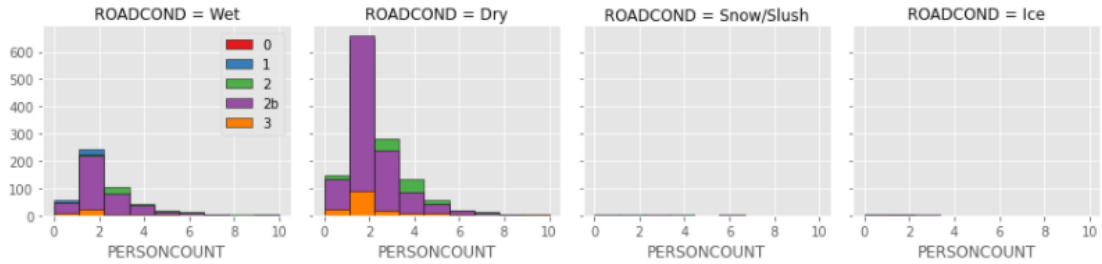
There were 11 unique weather descriptions among the 'WEATHER' column, these are four most frequent weathers: Clear, Raining, Overcast, Snowing (excludes the Unknown).



Surprisingly, clear weather caused more severe collisions and usually had two people involved, especially for group '2b'. This is on the contract of our normal cognition, as we always assume under extreme weathers people are more likely to have an accident.

3.3.2 Road Condition

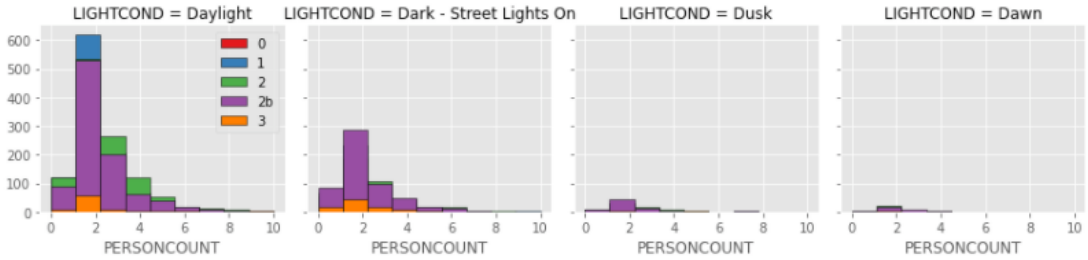
There were 8 unique road descriptions among the 'ROADCOND' column, these are four most frequent weathers: Dry, Wet, Ice, Snow / Slush (excludes the Unknown).



This result also verifies our previously observation, as clear weather has dry road and caused more severe collisions, also for group '2b'. Number of people involved in such collisions are also about two. While wet has second most people involved in a collision corresponds to the raining weather.

3.3.3 Light Condition

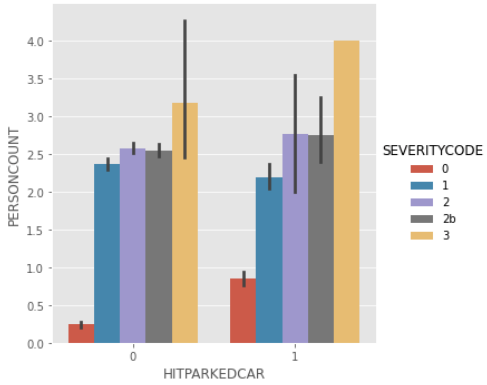
There were 8 unique light descriptions among the 'LIGHTCOND' column, these are four most frequent weathers: Daylight, Dark – Street Lights On, Dusk, Dawn (excludes the Unknown).



Daylight had more number of collisions for all severity groups. The conclusion complies with the result under the weather and road condition.

3.4 Severity vs. Number of Parked Cars Hit

In daily cognition, the collision would be more severe if a parked car is hit, because that means the impact of that collision is bigger.



'0' means no parked car was hit, '1' means yes parked car(s) was hit.

Under same severity of a collision, if a parked car is hit, more people are likely to be involved in that collision.

4 Predictive Modeling

The modeling process uses supervised training, it will use four major models to predict – KNN, Decision Tree, Support Vector Machine, Logistic Regression, and choose the best model. Since features have numerical and categorical features, and group '0' s categorical features are almost empty, here will use different features to train the model as well, then choose the features that will give most precise prediction.

Group '0', '1', '2', '2b' are balanced, group '3' with much fewer rows. Hence we over-sample the dataset before training it.

Train-test split is 0.15.

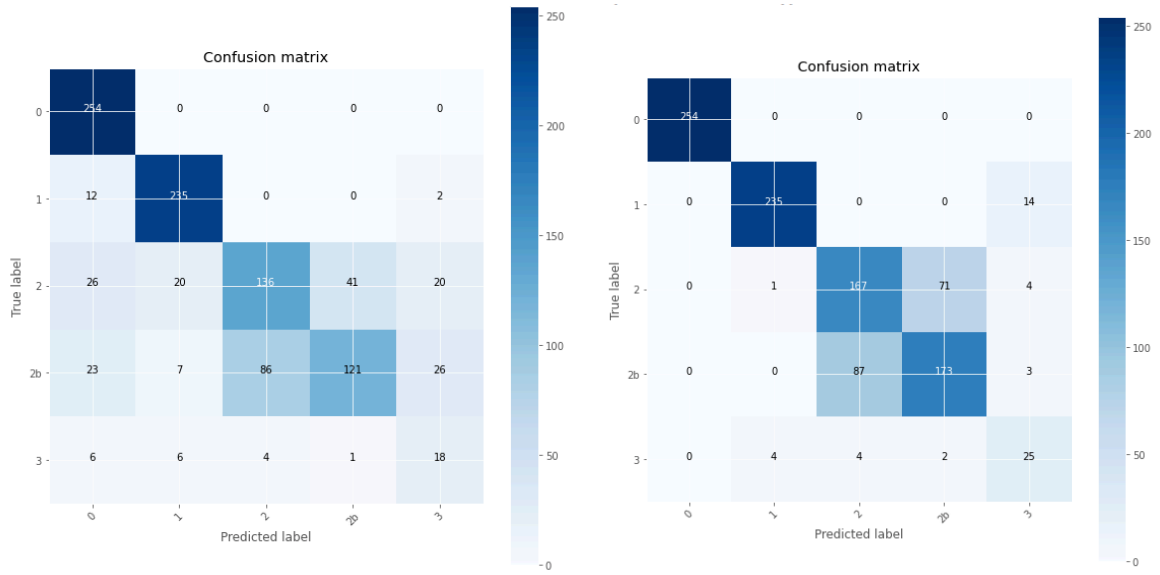
4.1 Using Numerical Features

'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'HITPARKEDCAR'

Performances:

	Algorithm	Accuracy-Score	F1-Score	LogLoss
0	KNN	0.731801	0.721816	NaN
1	Decision Tree	0.818008	0.819644	NaN
2	Support Vector Machine	0.818966	0.820371	NaN
3	Logistic Regression	0.821839	0.819647	0.865351

SVM has the best performance, as it can correctly predict 82% of the data. However, Logistic Regress also has good outcomes. Two confusion matrices (Left-Logistic Regress, right-SVM) are shown below. SVM did an excellent job in predicting group '0' and '1'.



4.2 Using Categorical Features

'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'DATE'

'Date' if certain weekdays are more likely to have specific group of collisions?

Group '0' is missing, what if we only build the model on group '1', '2', '2b', '3'?

	Algorithm	Accuracy-Score	F1-Score	LogLoss
0	KNN	0.624842	0.631821	NaN
1	Decision Tree	0.700887	0.695062	NaN
2	Support Vector Machine	0.648923	0.649762	NaN
3	Logistic Regression	0.434728	0.425248	1.130136

As we compared the models, the accuracy is much lower than previous models training using numerical features. Therefore, using categorical features itself does not give better prediction.

4.3 Using All Features

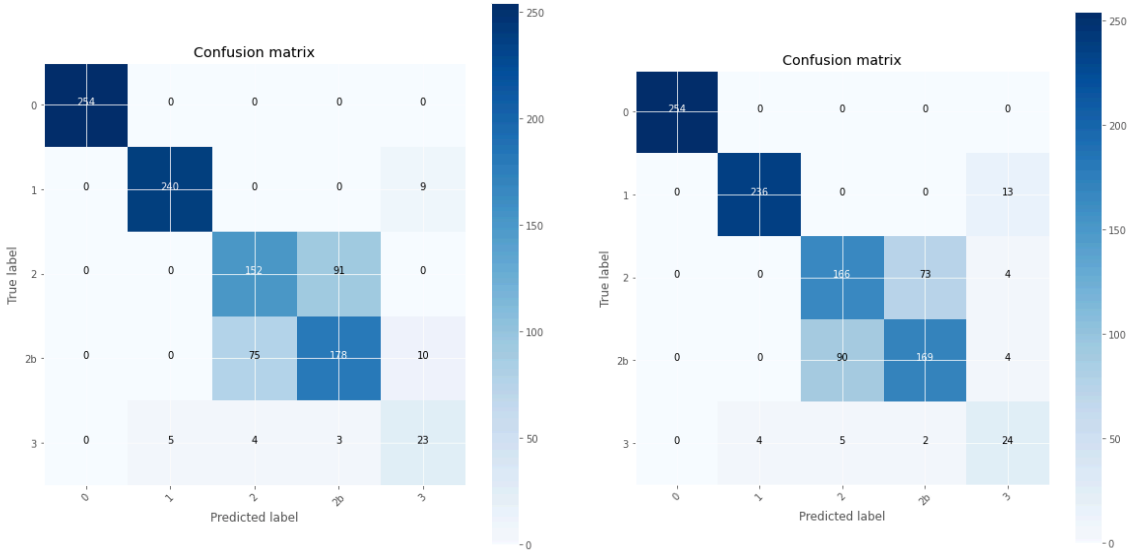
'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'HITPARKEDCAR', 'DATE'

Randomly generate values in the group '0' s NaN filed.

	Algorithm	Accuracy-Score	F1-Score	LogLoss
0	KNN	0.753831	0.754773	NaN
1	Decision Tree	0.811303	0.811979	NaN
2	Support Vector Machine	0.813218	0.814893	NaN
3	Logistic Regression	0.710728	0.706891	0.810925

SVM again has the best predicting results among other models. It again did an excellent job in predicting group '0' and '1'; however, the results for group '2', '2b' and '3' did not improve.

Two confusion matrices (Left-Decision Tree, right-SVM) are shown below.



However, compared with the model trained with numerical values, extra categorical features do not improve the performance of the prediction. The reason for this might be the randomly generated values for group '0' do not comply with realistic situations, or extra features cause the model to over-fitting. It is impossible to figure it out for now because we cannot retrieve correct group '0' s missing values.

5 Conclusions

In this study, we analyzed the relationship between a severity code of a collision and its related features, including the descriptive attributes such as number of people, vehicles, bicycles, parked cars hit etc., and

environmental factors such as the weather, road and light conditions. All these factors could be helpful in predicting the severity of a collision hence help 9-1-1 dispatchers to send most efficient help to the scene. In general, if the witness does not give too many details about the surroundings of the scene, even with only numerical features, the SVM model would provide a quite accurate prediction about the severity.

6 Future Directions

Based on the available dataset, models can best achieve about 82% of the accuracy. Of course they can still be improved if missing values can be fulfilled with actual statistics.

Models in this study mainly focused on features at one glance. In fact, there are more features on-scene can also be helpful in predicting the severity of the collision which might need the witness to have more knowledge of cars. For example, car types, as SUV are recognized as more reliable vehicles and might reduce the severity of the accident compare to sedans[2]. Also car sizes might be also important, as a full size SUV might absorb more impact during a collision compare to a compact size[3]. Even a fuel car, a hybrid or an electric car would also influence the severity. There are more features that will obviously improve the accuracy of our prediction model.

[2] M. Hernandez, "What's Safer in an Accident: Is an SUV Safer than a Car or Sedan?," Personal Injury Lawyers of Tampa, 18-Sep-2018. [Online]. Available: https://injurylawyersoftampa.com/whats-safer-in-an-accident-is-an-suv-safer-than-a-car-or-sedan/. [Accessed: 24-Sep-2020]

[3] [1]2020. [Online]. Available: https://www.edmunds.com/car-safety/are-smaller-cars-as-safe-as-large-cars.html. [Accessed: 24- Sep- 2020].