Predicting the Severity of Possible Traffic Accidents in Seattle City

Rong Chen

(Dated: October 7, 2020)

CONTENTS

I. INTRODUCTION

Predicting whether a traffic accident may happen and its severity if happened is a very important topic [1, 2]. For a drivers, if his/her car can be equipped with a real-time alarming system which can alert the possible severity of the accident if happened, it can no doubt greatly reduce the risk of a traffic accident, thus prevent the driver, the passengers and the pedestrians from getting injured and even save their lives. Besides, such a system can also be incorporated into the navigation system, so that it can pick a route with the lowest risk to the destination.

Furtherore, more and more cars (especially those electric-powered cars) nowadays are equipped with some kind of 'auto-pilot' function based on AI, and in the future it is very possible that cars can automatically drive to the destination without a driver. For those cars, the ability to predict the possible severity or risk of possible traffic accidents in real time is of central importance. Because without such an ability, an auto-drive car is just like a moving coffin for the passengers and a moving killer for the pedestrians.

Therefore, no matter for traditional cars or auto-drive cars, we inevitably need to develop

an efficient real time method to predict the severity of a possible traffic accident, in order to alarm the risk to the driver and the passengers.

In this report, I develop such a method to predict the severity of possible traffic accidents, based on classification (the label is 'severity') which is a type of supervised machine learning. Since this report is mainly for illustration purpose, I limit the scope within Seattle city.

II. DATA

I use the example dataset provided in the week 1 of this capstone project, which is the data of the severity of traffic accidents occurred in Seattle city from 2004 to 2020. The raw data contains about 190000 records. After deleting records including vague values such as 'NAN', 'N/A', 'unknown', et al, we are left with about 160000 records in the data. There are two levels of severity, 1 and 2. The higher the level the more severe the accident is.

The first step is to pick relevant features for the machine learning. In Fig. 1, I randomly plot 1000 locations for both severity 1 abd 2 accidents on the Seattle map. The smaller green points indicate severity 1 and the bigger red points indicate severity 2. We do find some patterns of the locations which means location must be included as a one of the features in the machine learning. However, we find that the severity 1 and 2 locations have some overlaps, therefore location alone is not enough to separate severity 1 and 2. We need more features.

Although the severity of a traffic accident for sure related with factors such as how many people involved including the pedestrians, which direction the car is hit from, et al, we can not use those as features. Because those information cannot be known before the accident. We can only use the information which can be collected before the accident occurs, such as the location of the car, the date and time, the weather condition, the road condition, the light condition, whether the driver is speeding and/or under influence of alcohol, et al. After deleting features which cannot be obtained before the accident, we are left with features as shown in Fig. 2. The meaning of each feature can be found in the metadata also provided in the week 1 of the capstone project.

The second step is that, before training the data using different models, we need to do some feature engineering on the data. For the longitude X and latitude Y, since the earth is a sphere, those polar coordinates have a period of 2π, so instead of using X and Y directly,

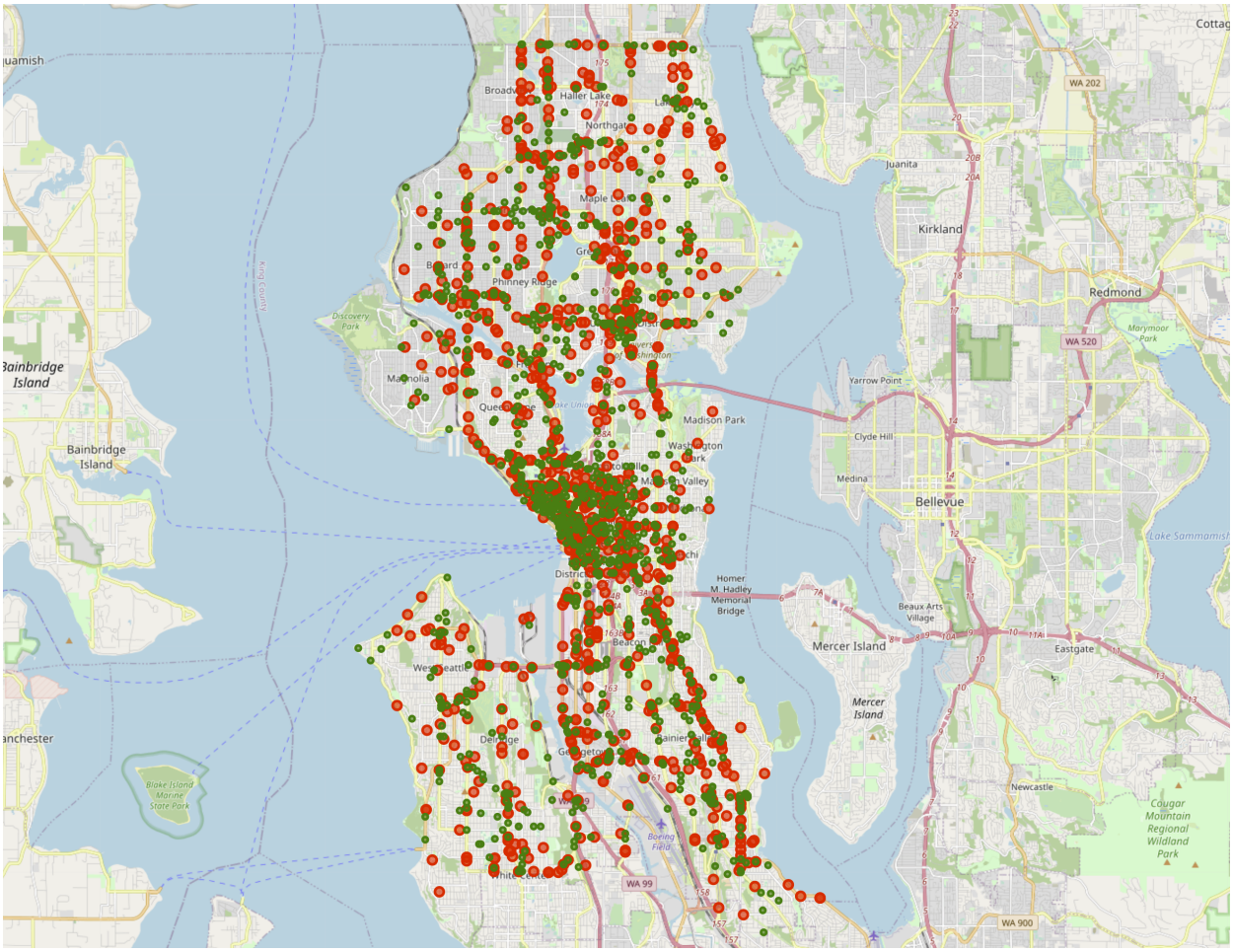

FIG. 1. Locations of severity 1 and 2 accidents in Seattle.

	SEVERITYCODE	X	Y	ADDRTYPE	INCDATE	INCDTTM	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
0	2	-122.323148	47.703140	Intersection	2013/03/27 00:00:00+00	3/27/2013 2:54:00 PM	0	Overcast	Wet	Daylight	0
1	1	-122.347294	47.647172	Block	2006/12/20 00:00:00+00	12/20/2006 6:55:00 PM	0	Raining	Wet	Dark - Street Lights On	0
2	1	-122.334540	47.607871	Block	2004/11/18 00:00:00+00	11/18/2004 10:20:00 AM	0	Overcast	Dry	Daylight	0
3	1	-122.334803	47.604803	Block	2013/03/29 00:00:00+00	3/29/2013 9:26:00 AM	0	Clear	Dry	Daylight	0
4	2	-122.306426	47.545739	Intersection	2004/01/28 00:00:00+00	1/28/2004 8:04:00 AM	0	Raining	Wet	Daylight	0

FIG. 2. The first 5 records in the data after deleting irrelevant features.

we use the value of the cosine and sine of X and Y [1] which is more reasonable. Similarly, since the day of the year, the hour of the day, and the minute of hour also have periods such as 365, 24, and 60, we also use their cosine and sine instead. For features 'WEATHER', 'ROADCOND', and 'LIGHTCOND', we use one hot method to convert each of their string values as a feature which is either 0 or 1. Finally we do feature scaling on all the values of the features so that they are distributed around their average values with variance 1.

The third step is to split the data into training set and test set and then balance the training set. As usual, we use 80% of the data as training set and 20% as testing set. We see that among all the records, 2/3 are severity 1 and only 1/3 are severity 2. So the data is not balanced in terms of severity. To proceed, we do up sampling for the severity 1 records in the training set, such that now there are the same number of severity 1 and 2 data in the training set. Note that we keep the test data unchanged, there is no need to balance the data in the test set.

III. DISCUSSION

With the training data ready, we now use models such as logistic regression, decision tree, K-nearest neighbor, support vector machine (SVM), XGBoost and random forest to train the data.

IV. SUMMARY

[1] A. Hébert, T. Guédon, T. Glatard, and B. Jaumard, CoRR **abs/1905.08770** (2019), arXiv:1905.08770.

[2] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 10.1145/3347146.3359078 (2019).