Predicting Car Accident Severity

Introduction

Traffic accident is very impact factor to make huge life or property damaged. Nobody want these things happen to himself or any other people. It's really terrible and horrible thing. So, if that bad thing happens, what should we do first? Especially for government such as traffic management department, hospitals, etc. What should these guys do could reduce destroy for the unlucky people or poor instruments? These government officers should read this report seriously! And use my method to PREDICTING the severity of the accident using computer, it will be more quickly than the office sirs doing!

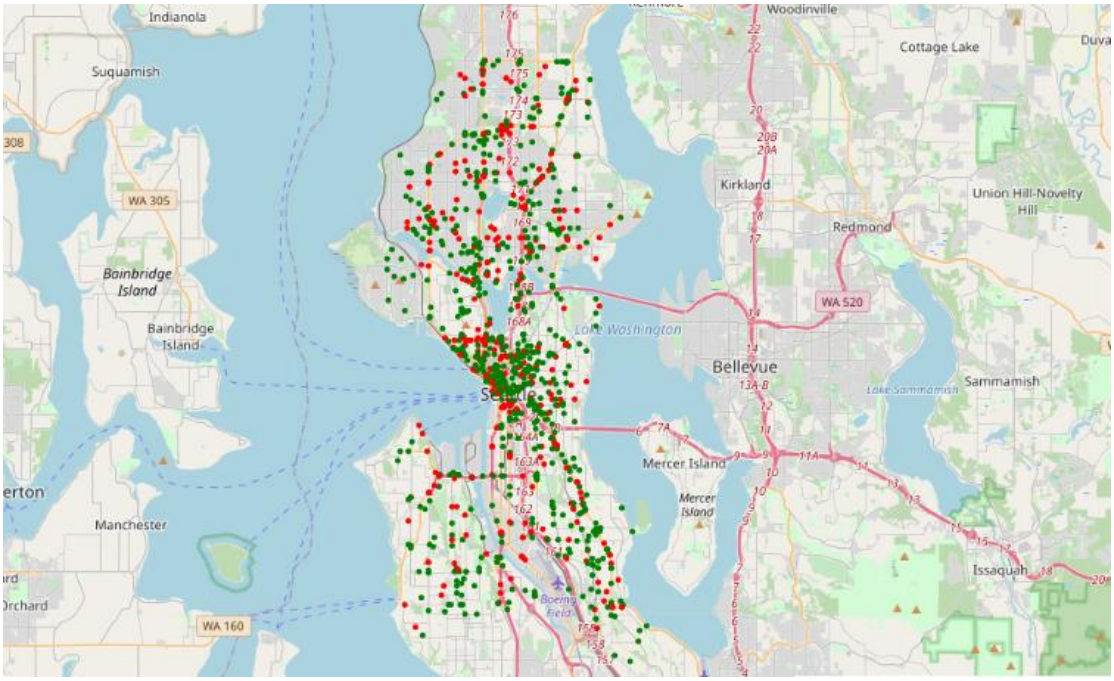

Figure 1 Distribution of sample accidents

Data

Luckily, I got the data what is the accidents happened in Seattle state. Depends on this data, using my nice machine learning technology, some useful information will be found and presented to the government guys.

In this data, there is an accident severity column, which was marked by those smart guys (looks not too bad). Besides, there are also lots of columns describe the accidents details, such as position, pedestrian count, car count, weather, collision style…

In fact, I really don't want to see those data, it's really make me unhappy, you know, those are not nice things. And another reason, maybe, is my English is NOT good! Could you understand me

please?

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDI
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	

Figure 2 Original data

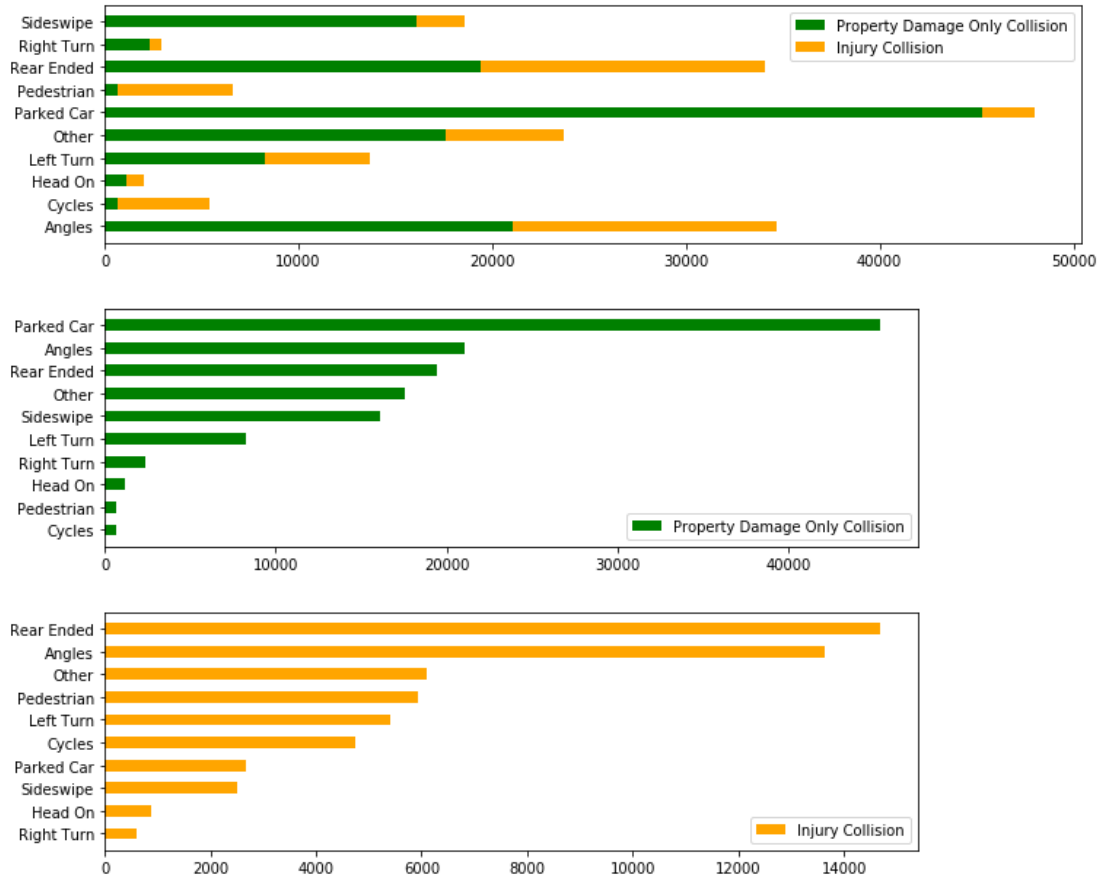

Figure 3 Severity ration in collision types

Methodology

I handle this data using cleaning. It's really not an exciting thing. Because I used a long period time to analyses it. I dropped some useless columns, and filled or modified some other YES or NO values and One-hot values. Finally, I got a nice data to train the machine learning methods. Bingo!

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INATTENTIONIND	UNDERINFL	PEDROWNOTGRNT	SPEEDING
SEVERITYCODE	1.000000	0.157984	0.185183	0.215675	-0.036399	0.033845	0.045701	0.195434	0.007403
PERSONCOUNT	0.157984	1.000000	-0.013516	-0.044032	0.419377	0.119188	0.017856	-0.046609	-0.050755
PEDCOUNT	0.185183	-0.013516	1.000000	-0.028722	-0.187881	-0.044389	0.001780	0.451463	-0.032349
PEDCYLCOUNT	0.215675	-0.044032	-0.028722	1.000000	-0.217157	-0.034827	-0.034064	0.434435	-0.036032
VEHCOUNT	-0.036399	0.419377	-0.187881	-0.217157	1.000000	0.120791	0.059796	-0.179517	-0.023892
INATTENTIONIND	0.033845	0.119188	-0.044389	-0.034827	0.120791	1.000000	-0.041258	-0.059310	-0.077872
UNDERINFL	0.045701	0.017856	0.001780	-0.034064	0.059796	-0.041258	1.000000	-0.029220	0.074061
PEDROWNOTGRNT	0.195434	-0.046609	0.451463	0.434435	-0.179517	-0.059310	-0.029220	1.000000	-0.030909
SPEEDING	0.007403	-0.050755	-0.032349	-0.036032	-0.023892	-0.077872	0.074061	-0.030909	1.000000
HITPARKEDCAR	-0.103974	-0.063080	-0.029196	-0.032520	-0.088471	-0.019598	-0.034626	-0.027897	-0.036628

Figure 4 Correlation of the features and label

Then, the most exciting thing is to train the machine learning methods including SVM, Decision Tree, K-neighbors, etc. Oh, and the score look not bad, the best one is SVM with 0.739.

	classifier	score
2	SVM	0.739
6	DecisionTree	0.739
8	RandomForest	0.737
9	AdaBoost	0.737
10	GradientBoosting	0.737
3	KNeighbors	0.736
7	BaggingClassifier	0.734
0	LinearDiscriminant	0.732
1	QuadraticDiscriminant	0.728
5	GaussianNB	0.724
4	MLP	0.701

Figure 5 Methods and scores 1

But I suddenly find out I just use the columns which are relation with the result definitely. Some one-hot columns were not used. Then I add these columns to machine learning. So, I find this best score is 0.746 by Linear Discriminant.

	classifier	score
0	LinearDiscriminant	0.746
10	GradientBoosting	0.745
9	AdaBoost	0.741
2	SVM	0.739
4	MLP	0.732
7	BaggingClassifier	0.727
3	KNeighbors	0.726
8	RandomForest	0.725
6	DecisionTree	0.699
5	GaussianNB	0.382
1	QuadraticDiscriminant	0.313

Figure 6 Methods and Scores 2

Results

All in all, according to those jobs I did above we could find the machine learning method is good to predict the severity of accident by those features we got. It must be very useful to those offers. Good luck sir, do your best!

	classifier	score
0	LinearDiscriminant	0.746

Figure 7 Result we select

```
(0.4001, 'IS_COLLISIONTYPE_Parked Car')
(0.1553, 'PEDCOUNT')
(0.1271, 'PERSONCOUNT')
(0.1114, 'PEDCYLCOUNT')
(0.0866, 'IS_COLLISIONTYPE_Sideswipe')
(0.0245, 'IS_COLLISIONTYPE_Rear Ended')
(0.0219, 'IS_JUNCTIONTYPE_At Intersection (intersection related)')
(0.0161, 'UNDERINFL')
(0.0105, 'VEHCOUNT')
(0.0085, 'IS_COLLISIONTYPE_Right Turn')
(0.0057, 'IS_LIGHTCOND_Unknown')
(0.005, 'IS_COLLISIONTYPE_Other')
```

Figure 8 some effective factors impact result much

Discussion

In the process I handle the data, I find something is useful phenomena. The special accident

types are not distributed like the total accidents.

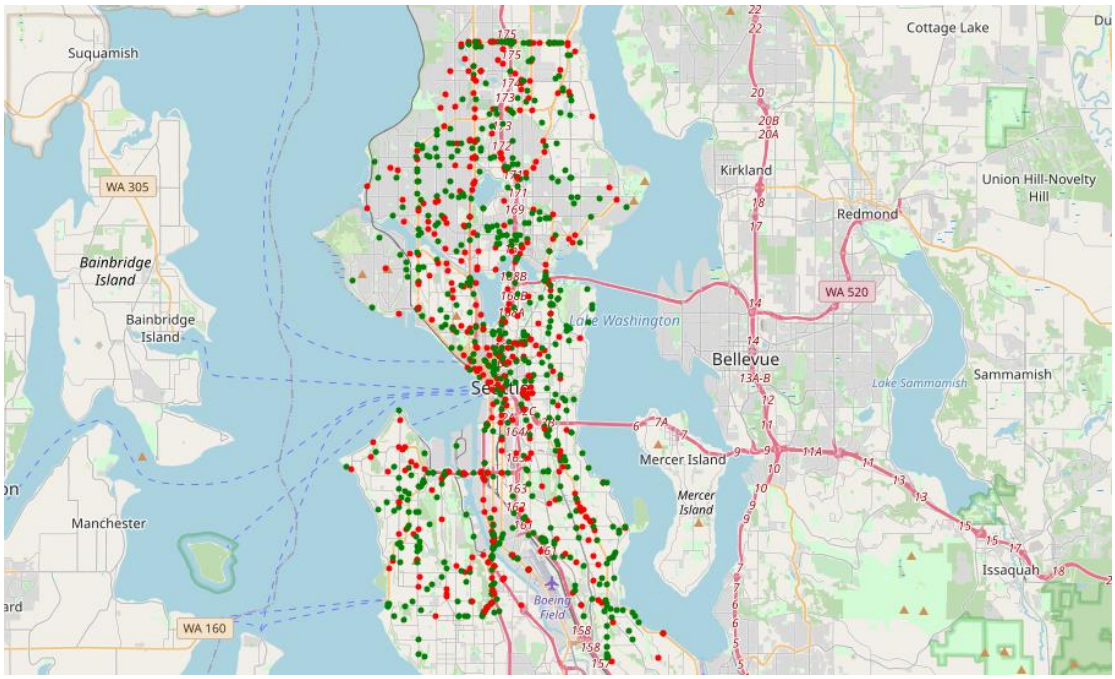

Figure 9 speeding in those roads and places

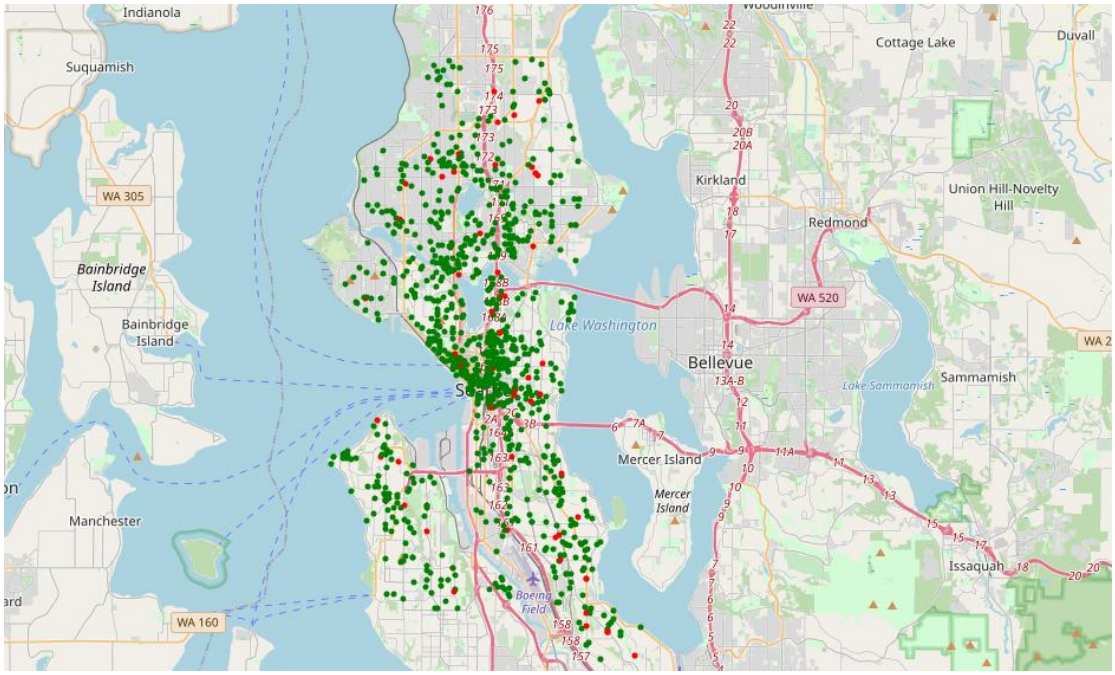

Figure 10 Parked car accident in some places

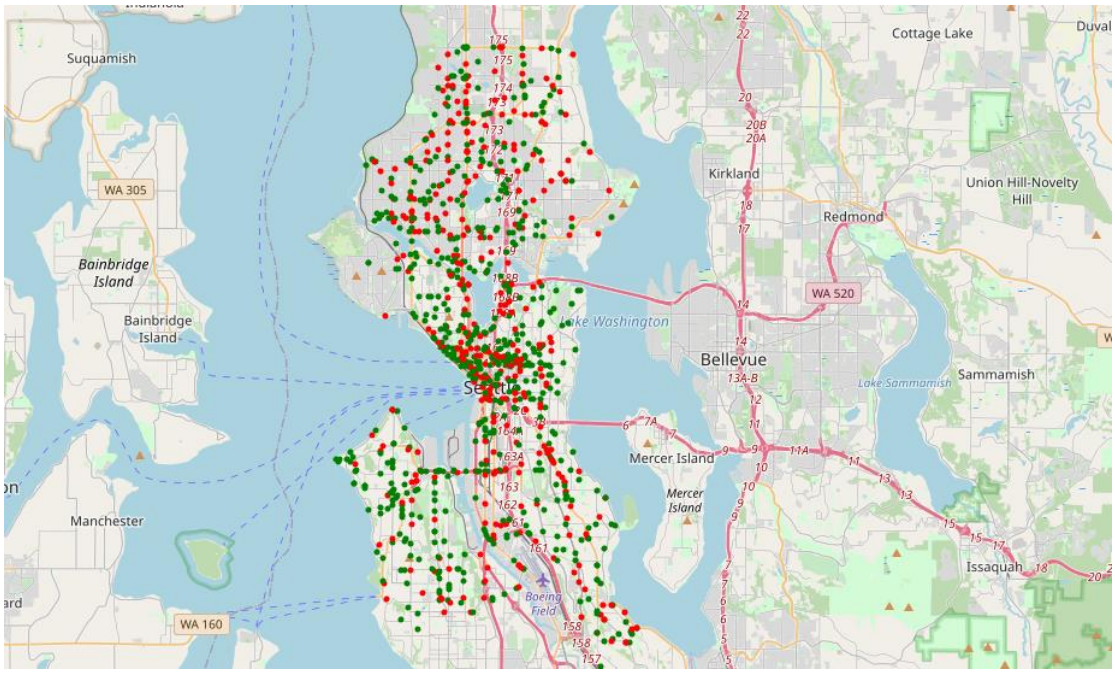

Figure 11 under influence accidents happen in these places

Conclusion

OK, everything ran good. I finished the job. But what we could do according my report? Government should improve the plan of roads and set more cameras at key point to reduce the accident happens or damage. And hospital could use machine learning method to predict severity of accident and then action ASAP with best resource distribution.
Good luck guys, hope there will be less accident happens and less people get hurt.