Car Accident Severity

Capstone Project

Abstract

In this report, we built models to classifier the severity classes of car accidents in Seattle. The models used here are Decision Tree and Random Forest with python. We found that "At Intersection" has important impact.

1. Introduction

Road traffic injury is a serious problem and road safety is one of the major concerns. According to WHO[1], a road injury ranked in 8th in the cause of death in 2016. Even though those who had accidents survive, they may be left with physical or psychosocial disabilities. In addition, road accidents lead to financial loss. This is estimated at about US$518 billion per year[2]. To minimize damage to people and reduce economic loss, we need to analyze past data and know factors.

The purpose of this article is to analyze data about traffic collisions in Seattle and make suggestions to reduce the number of traffic accidents. If some regularities are found, it is easy to take measures against accidents. In this report, I would like to show what surrounding conditions such as weather condition, road condition and light condition are associated with the severity of the accidents mainly.

This report is useful for drivers, local residents and administration. From the drivers and local residents' perspective, the result led from data analysis will change the public awareness. From the viewpoint of administration, if we know the conditions that are likely to cause accidents, we can carry out patrols more efficiently. Furthermore, data analysis will be helpful when making a city plan to avoid traffic accidents.

2-1. Data Resource

Data used in this report were obtained from Collisions-All Years provided by SDOT Traffic Management Division[3]. This dataset contains the collisions that took place for a period(2004-2020) in Seattle, USA. The total number of collisions is 194673. SDOT investigated severity, situations of accidents, causes of accidents(accidents caused by influence of drugs or alcohol, inattention and overspeed) and surrounding conditions(weather, road condition, light condition and so on) for each collision. In this report, I will separate surrounding conditions from the original data and see the relation between severity and surrounding conditions with machine learning. I will not use data related to human factors such as overspeed. Drivers are required to obey laws. This is the only solution to reduce human factors.

2-2. Data Understanding

This dataset contains 37 attributes and 194673 cases. These attributes have not only the collision situation(the severity of the collision, the location, collision type, the number of people involved in the collision), but also human factors(whether a driver was under the influence of drugs or alcohol and so on) and surrounding conditions(address type, category of junction, the road condition, the weather condition and the light condition). The dataset is labeled and unlabeled.

We chose SEVERITYCODE as a dependent variable. This is labeled as following.

Severity Code	Status
3	Fatality
2b	serious injury
2	injury
1	prop damage
0	unknown

As independent variables, we selected surrounding conditions, ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND. The aim of this report is to reduce the severity in collision and we would like to find out the relation between the severity in collision and the surrounding conditions.

2-3. Data Cleaning

To build the model to predict the severity in collision, we need to transform the original data to the usable data. We took the following steps.

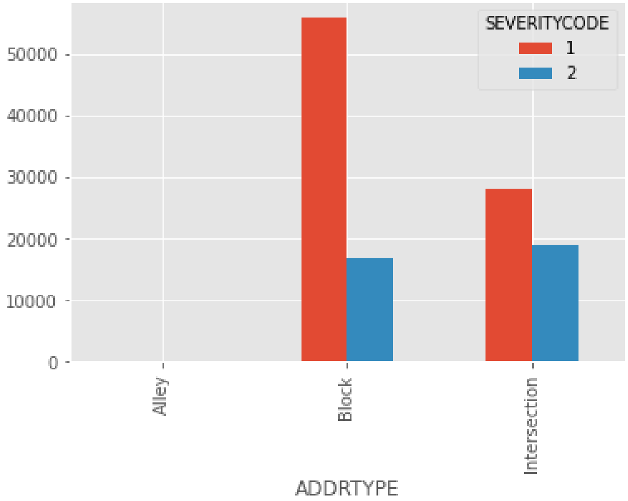
1.Retrieve SEVERITYCODE, ADDRTYPE, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, PEDROWNOTGRNT and SPEEDING from the original data.

2.Remove the cases caused by the human factors, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT and SPEEDING from the above data.

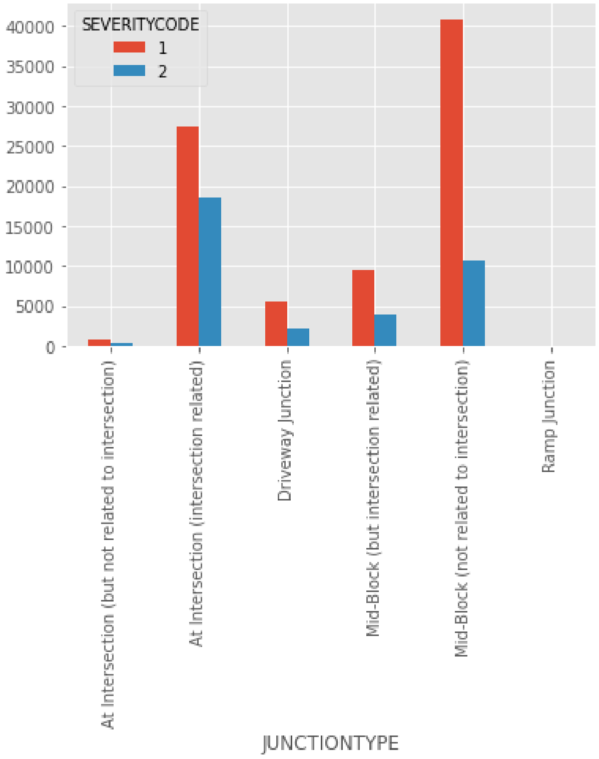
3.Drop NaN data, “Unknown” and “Other” because we would like to predict the severity in collision more accurately. In addition, we drop “Oil” from ROADCOND. We consider that this is also a human factor.

2-3. Data for Analysis

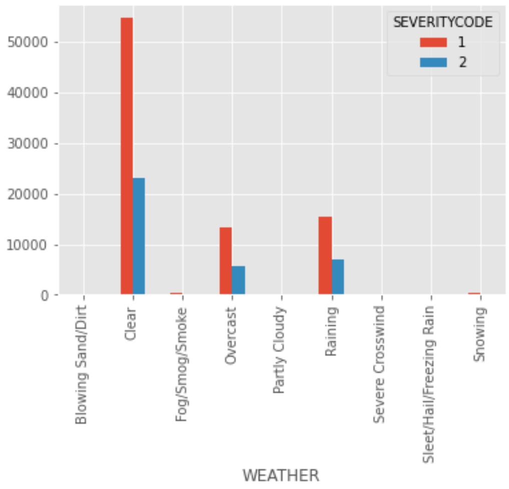
After data-cleaning, we had 120068 cases and there are only 2 severity levels, 1 and 2. We showed the bar graphs about relations between severity and independent variables.



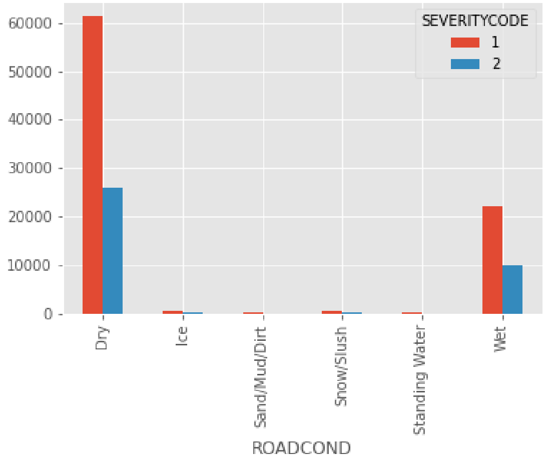
ADDRTYPE	# of collisions
Block	126926
Intersection	65070
Alley	751



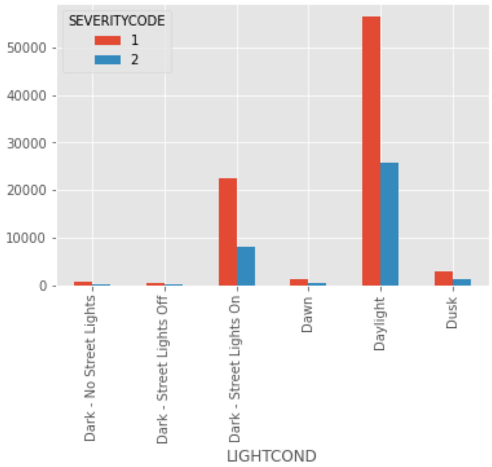
JUNCTIONTYPE	# of collisions
Mid-Block (not related to intersection)	89800
At Intersection (intersection related)	62810
Mid-Block (but intersection related)	22790
Driveway Junction	10671
At Intersection (but not related to intersection)	2098
Ramp Junction	166



WEATHER	# of collisions
Clear	111135
Raining	33145
Overcast	27714
Snowing	907
Fog/Smog/Smoke	569
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	56
Severe Crosswind	25
Partly Cloudy	5



ROADCOND	# of collisions
Dry	124510
Wet	47474
Ice	1209
Snow/Slush	1004
Standing Water	115
Sand/Mud/Dirt	75



LIGHTCOND	# of collisions
Daylight	116137
Dark - Street Lights On	48507
Dusk	5902
Dawn	2502
Dark - No Street Lights	1537
Dark - Street Lights Off	1199
Dark - Unknown Lighting	11

3-1. Model Building

We would like to find out what conditions cause “severity2” cases. In order to analyze our data we employed Decision Tree and Random Forest which are the supervised learning methods with python. Random Forest is an ensemble of decision trees. The reason why we chose these methods is that it is easy for us to understand what surrounding conditions cause more severe collisions by visualization of tree.

Before machine learning, we need to convert categorical variables into indicator variables with *pandas.get_dummies()*. Additionally, we transformed severity2 to “1” and severity1 to “0”.

Decision Tree model has parameters which we can designate. We chose entropy as “criterion” to measure the impurity. The maximum depth was determined by grid search with cross-validation which optimizes the maximum depth. Furthermore, we tried two patterns, default and balanced as “class_weight”. In general, we should make a balanced dataset when the classes are not equally distributed. In Decision Tree, we can realize a balanced dataset by designating the parameter, “class_weight”.

In Random Forest, we used 2 as the maximum depth because grid search in Decision Tree showed that the best depth was 2. “Criterion” was set to entropy. We set 50 as “n_estimators” which means the number of created decision trees. As we did in Decision Tree, we tried two patterns, default and balanced as “class_weight”.

In total, we established four models, Decision Tree without and with balanced and Random Forest without and with balanced.

3-2. Results

Based on four models we calculated precision, recall and f1-score for each model. The results are as follows.

1. Decision Tree without balanced

	precision	Recall	f1-score
Severity 1	0.70	1.00	0.83
Severity 2	0.00	0.00	0.00
Accuracy			0.70
macro avg	0.35	0.50	0.41

weighted avg	0.50	0.70	0.58

2. Decision Tree with balanced

	precision	Recall	f1-score
Severity 1	0.77	0.68	0.72
Severity 2	0.40	0.52	0.45
Accuracy			0.63
macro avg	0.59	0.60	0.59
weighted avg	0.66	0.63	0.64

3. Random Forest without balanced

	precision	Recall	f1-score
Severity 1	0.70	1.00	0.83
Severity 2	0.00	0.00	0.00
Accuracy			0.70
macro avg	0.35	0.50	0.41
weighted avg	0.50	0.70	0.58

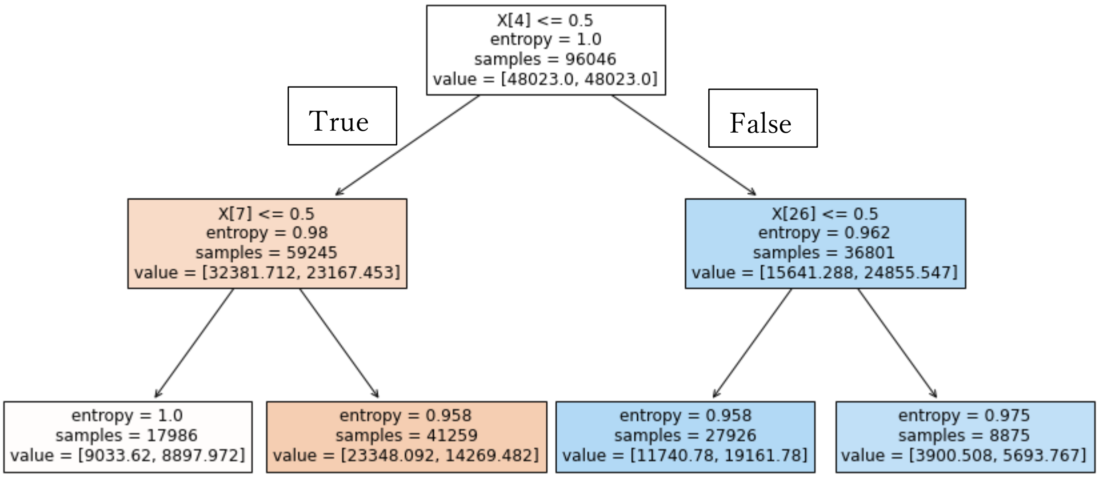
4. Random Forest with balanced

	precision	Recall	f1-score
Severity 1	0.77	0.67	0.72
Severity 2	0.40	0.53	0.46
Accuracy			0.63
macro avg	0.59	0.60	0.59
weighted avg	0.66	0.63	0.64

From the results of precision, recall, f1-score of Severity 2, the better models are Decision Tree with balanced and Random Forest with balanced.

4. Discussion

Let’s interpret the result of the Decision Tree model with balanced. The decision tree produced by machine learning is shown as:



X[4], X[7], and X[26] mean "At Intersection (intersection related)" , "At Intersection (but not related to intersection)" and "Dusk" respectively. The left side of the value is the number of severity1 collision and the right side is that of severity2 collision. Note that convert categorical variables into indicator variables. For example, "X[4]<=0.5 is True" means the collision not "At Intersection (intersection related)" and "X[4]<=0.5 is False" shows the collision "At Intersection (intersection related)". From the first step of the tree, we can decrease the entropy. In addition, we can obtain more severity2 collision samples when "X[4]<=0.5 is False". Therefore, we understand that "At Intersection (intersection related)" is important factor. At the second step, it is difficult to distinguish severity1 and severity2 case by the surrounding factors because the entropies increase.

Therefore, to reduce severe collisions(severity2), it is a good way to give drivers a warning at intersection.

To establish a better model to distinguish severity1 and severity2 case, we may need to consider other surrounding factors except information we can obtain.

Finally, although it is possible to visualize trees for Random Forest, we do not show here because there are 50 trees in the model.

5. Conclusion

In this report, to classifier severity1 and severity2 case, we built Decision Tree and Random Forest models in which the surrounding factors were selected as independent variables. By visualization of the tree we found out that “At intersection” has important impact. To build a better model, we need to collect more information about surrounding situations.

6. Reference

[1]https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2]https://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/chapter2.pdf

[3]https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data?geometry=-123.376%2C47.452%2C-121.290%2C47.776