SEPT 2020

CAR ACCIDENT SEVERITY ANALYSIS

HOW TO REDUCE CAR ACCIDENT WITH MACHINE LEARNING?

Prepared by: Tianpeng Yu
LinkedIn: https://www.linkedin.com/in/tianpengyu
GitHub: https://github.com/Tianpeng-Yu

1. Introduction

1.1. Background

According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade. (https://www.injurytriallawyer.com/library/car-accident-statistics-seattle-washington-state.cfm) As the use of car keep increasing, it's better to understand the accidents data and find sustainable solutions to solve this question.

1.2. Problem

Car accidents have great impact on people's lives. To help reduce the severity and frequency of car collision, I want to use the Seattle car collision data to generate insights on how modeling can help reduce accidents. Given the attributes including location, weather conditions and address type, we can see which factors attributes to car accidents most and how we can alert the driver in advance.

1.3. Stakeholders

This research and report would be beneficial to the local government, people who live in Seattle and also car insurance companies. By looking into road condition factors and address type such as intersection and block, we could see if there is any possible improvement on road condition and city planning.

2. Understanding data

2.1. Data sources

I obtained the Car Collision dataset from IBM Cloud here which contains collision data, address and other different conditions. The dataset includes 38 columns and 194673 observations.

2.2.Data cleaning

There are some problems with this dataset. First, for missing values there are 10527 records and because most of the features are categorical data so I think I could build different models with/without the missing values and compare the accuracy later to decide whether I will exclude missing value.

Secondly, the major task is to predict the severity of car accident and the target variable should be SEVERITYCODE which was 1 (damage only) and 2 (injury). And the number of records 2 is twice as many as record 1. So I used resampling to down sample the majority 2 to the same amount as 1 to eliminate the unbalance.

Thirdly, to build the classification model, I will need to convert categorical values to numerical. I encode ADDRTYPE, COLLISIONTYPE, WEATHER , ROADCOND, LIGHTCOND to numeric values. Among WEATHER, ROADCOND, LIGHTCOND which has both 'Other' and 'unknown' values, I group these two into one value.

2.3.Feature selection

I selected ADDRTYPE, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND as the features.

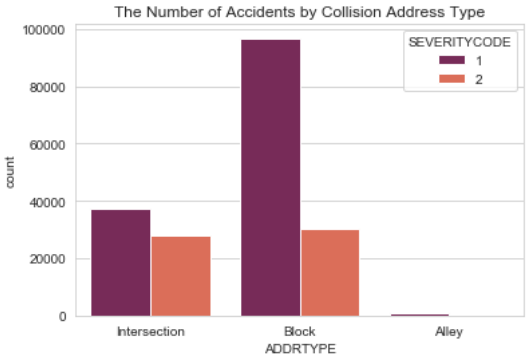
I intended to use SPEEDING, UNDERINFL (under the influence of drugs or alcohol when driving) and INATTENTIONAND (not paying attention when driving) because there is high correlation between these and car accidents, but the data is not representative enough. They only contain value 'Y' and 'Unknown'. So I did not select these as my features.

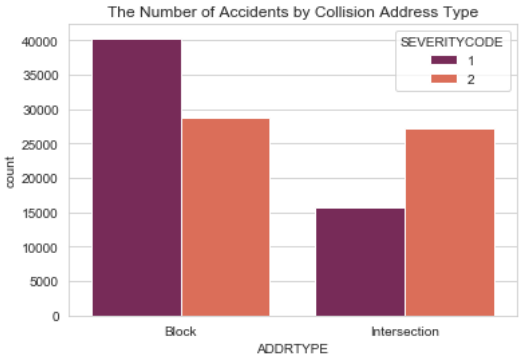
Feature	Data type	
ADDRTYPE	String to int	A description of the weather conditions during the collision
COLLISIONTYPE	String to int	Collision address type: • Alley • Block • Intersection
WEATHER	String to int	A description of the weather conditions during the collision
ROADCOND	String to int	The condition of the road during the collision.
LIGHTCOND	String to int	The light conditions during the collision.

3. Methodology

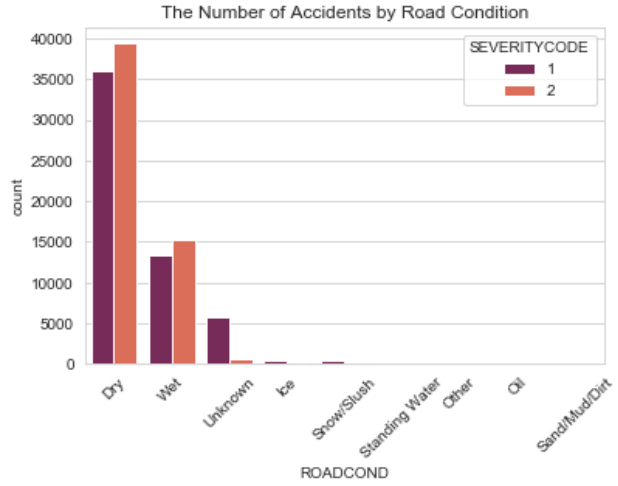
3.1.Exploratory data analysis

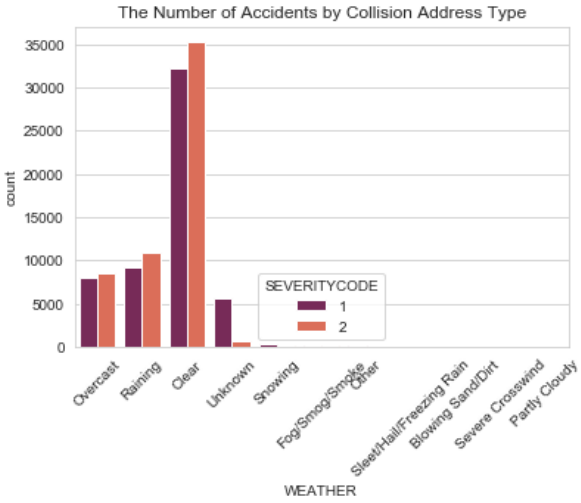
After understanding what each feature represents, I think exploratory can help me explore the relative value to understand their distribution. The following two visualizations represent the dataset with/without missing values. Based on the different distribution, we can see that the unbalanced data of target variable 'severity' affects the number of collisions among different address types. So I decided to drop all the missing values and carry out my remaining modelling.

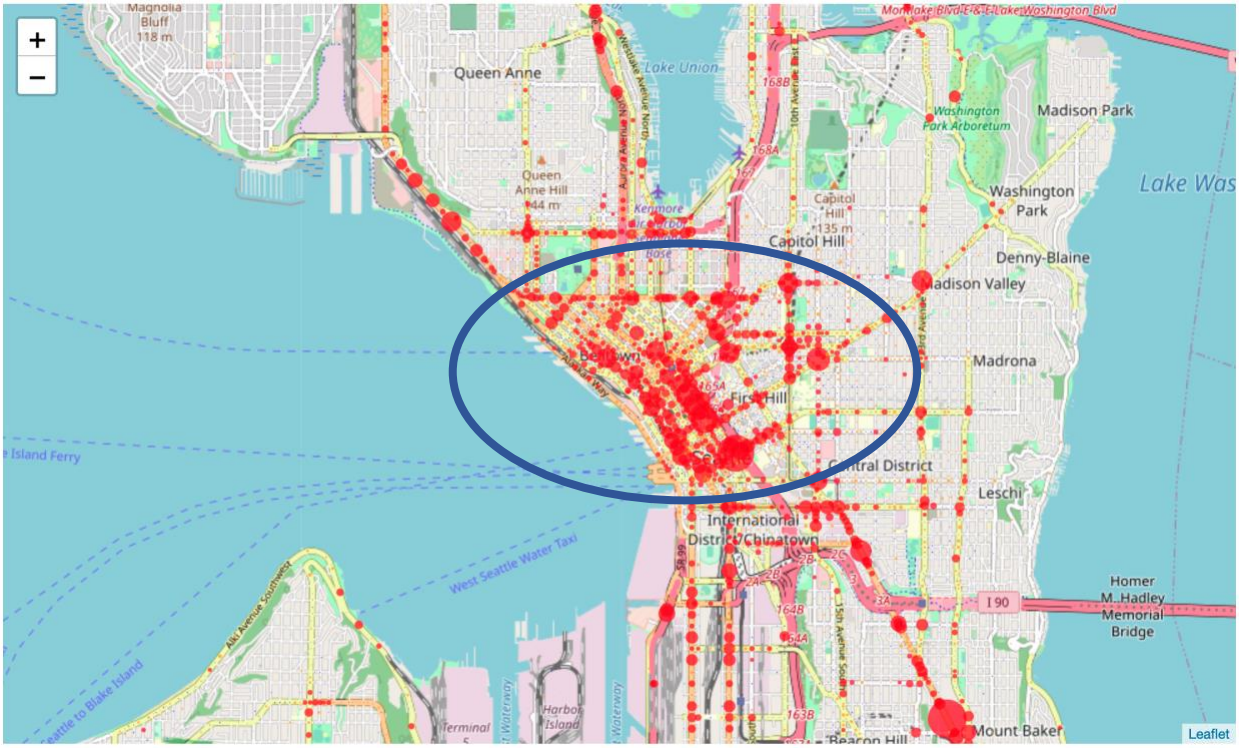




From the right chart, I noticed that in the collision severity is contrary between block and intersection address type.







From the collision map, we can also clearly see the major area that car accidents happened. And we can improve this area later.

3.2.Predictive modeling

The machine learning models used are Logistic Regression, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). The logistic regression model is the basic one to predict the binary outcome using a logistic function. KNN uses non-parametric method to classify new items based on similarity measures (distance). C-Support Vector Classification in linear kernel helps to classify new data with binary outcomes and more memory efficient by using a subset of training points.

4. Result

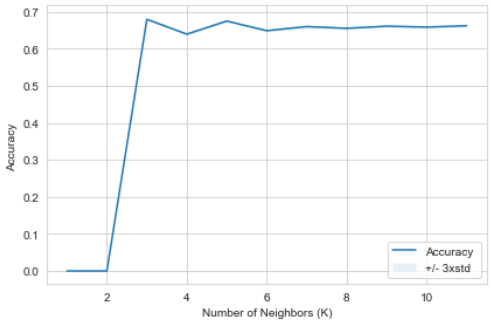
4.1. Logistic Regression

Classification report:

	precision	recall	f1-score
1	0.63	0.65	0.64
2	0.64	0.62	0.63
Accuracy			0.63
Macro avg	0.63	0.63	0.63
Weighted avg	0.63	0.63	0.63

4.2. K-Nearest Neighbor

By comparing the different performance using different k, we get the right chart and the best k =3. Then we fit the model with our trainset, we get the following confusion matrix:



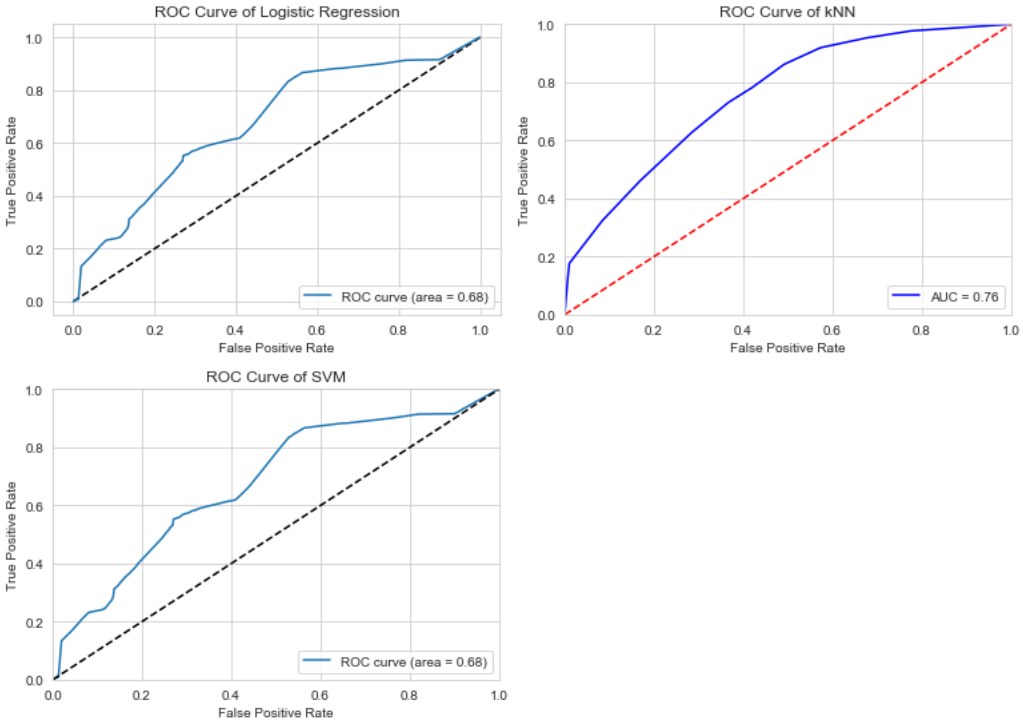
	precision	recall	f1-score
1	0.73	0.58	0.64
2	0.65	0.78	0.71
Accuracy			0.68
Macro avg	0.69	0.68	0.68
Weighted avg	0.69	0.68	0.68

4.3. Support Vector Machine

	precision	recall	f1-score
1	0.50	0.58	0.54
2	0.50	0.41	0.45
Accuracy			0.50
Macro avg	0.50	0.50	0.49

Weighted avg	0.50	0.50	0.49

To better compare these three models, I computed the roc score for each model.



5. Conclusion

From the ROC curve, the KNN model performed slightly better than the other two. And compared with F1 score, precision and recall for each model, the KNN is still better. KNN model reaches 0.76 AUC and 0.68 average F1 score.

Based on the model and visualization, the light condition and road condition should be improved especially near First Hill. And also drivers can pay more attention to those factors when driving in Seattle to stay focused and safe.

6. Future direction

The prediction of car accident severity is not completely finished. Based on the result, the dataset is underfit, which means I will need to collect more data and features such as speeding and alcohol using to better train the model.

In addition, the dataset only contains binary data for severity, however there could be more scenarios for a car accident and also the people involved. I can stay tuned with the data and keep improving my models.

7. References

https://towardsdatascience.com/predicting-crash-severity-for-nz-road-accidents-6214117e73fb

https://medium.com/@somayyeh_gh/seattle-car-accident-severity-f8a06299c460

https://medium.com/@rajivranjansingh_77828/project-for-applied-data-science-capstone-on-coursera-95ba3b39a6ca