Introduction

Background

With the development of vehicles, the frequency of car accidents is increasing year by year. Meanwhile, the happening of car accidents shows some common characteristics like in some narrow roads or during the rainy seasons the frequency will increase because of the bad road condition. And the level of severity will also increase because the wet road will influence the speed and hit strength of the accident. And the research on what will influence the severity is meaningful. Firstly, it will warn us in which kind of condition the car accidents are tend to be more dangerous. Therefore, drivers should be more careful in those situations. Secondly, the function of the predicted result can serve as the reference of SPD officers to define the severity level.

Problems

The data of this project is obtained from the given data-set of the course,which includes all the variables of the accidents occurred in Seattle in one year recorded by SPD. The aim of this project is to predict those variables that will influence the level of severity in the accident and build a machine learning model to record this.

Interest

Drivers may be interested in this topic since they want to know the dangerous driving variables so that they can avoid. Urban planners may be interested in this topic since it depicts the detailed information of the road condition that will cause a severe accident so that they can adjust. While the police officers may can use the trained machine learning model to define the severity of an accident in the future.

Data

Data Source

This data set is obtained from the link provided by this course, which is published by SPD, recording all variables related to every accident occurred in Seattle in one year. You can download the data in 'https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv'.

Data Explaining

The data-set is given by the SPD about the characteristics of car accidents in a year in Seattle. There are 38 variables. Some of them are not that useful in this case like the OBJECTID,INCKEY and COLDETKEY that used to define the id of the accidents. I will pick up some useful data that should be applied in this project to discuss.

The SEVERITYCODE is the dependent variable that defines the severity of the accidents. It is divided in 5 levels respectively 3 for fatality, 2b for serious injury, 2 for injury, 1 for prop damage and 0 for unknown. The aim of this project is to use other variables to define and predict the SEVERITYCODE.

The ADDRTYPE defines the address of the car accidents, including alley,block and intersection. From this variable we know in what kind of areas the severe accidents are mostly likely to occur.

The LOCATION that describe the general location of the collision. The frequency of the location may be influenced in several ways like the road condition, the light condition and even the character of neighbors. From the LOCATION, we can predict the synthesis situation of those areas.

The SEVERITYDESC is used to define the type of collision, including Injury Collision and Property Damage Only Collision etc, while the COLLISIONTYPE is used to describe the detailed type of collisions including Angles and Sideswipes etc. From those two variables, we can understand the detailed information of this collision.

The PERSONCOUNT, PEDCOUNT, VEHCOUNT, INJURIES,SERIOUSINJURIES, FATALITIES record the total number of people, pedestrians, bicycles, vehicles, injuries, serious injuries and fatalities in the collision respectively. Obviously, those variables like injuries or number of cars are important factors that will influence the severity.

The INCDATE and INCDTTM that define the date and date,time of the incident. It is possible that in some dates the frequency of the accidents will increase like in some holidays festivals that people tend to go out and get drunk. Or even the time of the final exam that students are more likely to stay up till night to study which makes them tired when driving.

The INATTENTIONIND, UNDERINFL,PEDROWNOTGRNT,SPEEDING and HITPARKEDCAR which defines whether whether the accidents was due to inattention, whether the driver involved was under the influence of drugs or alcohol, whether the pedestrian right of way was granted, whether or not speeding was a factor in the collision and whether the collision involved hitting a parked car respectively which could be used to predict the cause of the accident.

The WEATHER, ROADCOND, LIGHTCOND that describe the weather, road and light condition of the accident, which could be used as natural factors of the accident.

Data Cleaning

To satisfy the need to apply the data to different models, we need to deal with the data to fit the models.

Firstly, we can see that in these 38 variables, there are some variables that is not related to the severity of the accident. Therefore, we need to drop those columns as follow.

Variable Name	Reason
'OBJECTID','INCKEY','COLDETKEY','	Used only to label the accident. No actual

REPORTNO'	meaning.
‘SEVERITYCODE.1’,'SEVERITYDESC'	Redundant columns
‘STATUS’,'EXCEPTRSNCODE','EXCEPTRSNCODE'	Unknown columns
'SEGLANEKEY','CROSSWALKKEY'	Wrong columns(with nearly all the number equal to 0)
'INCDATE','INCDTTM','SDOT_COLCODE','SDOTCOLNUM','ST_COLCODE','ST_COLDESC',’X’,’Y’	Columns that are not related

Therefore, there are 18 variables left. Then we need to replace all the NaN to 0 so that those blanks make sense. But for different columns, we should treat it differently. For example, in column ’INATTENTIONIND’, the NaN means NO so that you need to replace it, while in column ‘ADDRTYPE’, the NaN means the value is missed so that you need to delete it.

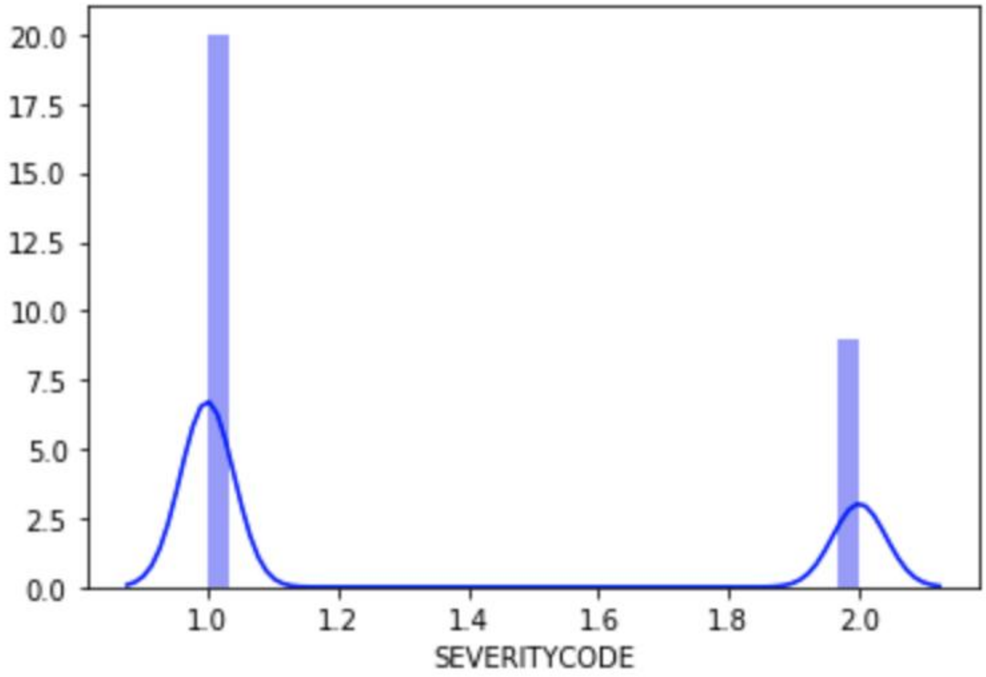
The next step is to change those categorical variables into numerical variables. For example, for the variable ‘ADDRTYPE’, it can be divided into Block, Intersection and Alley. We can use 0,1,2 to represent these 3 variables respectively.

After all these steps, we have got a data frame with 182895 rows and 18 columns.

Exploratory Data Analysis

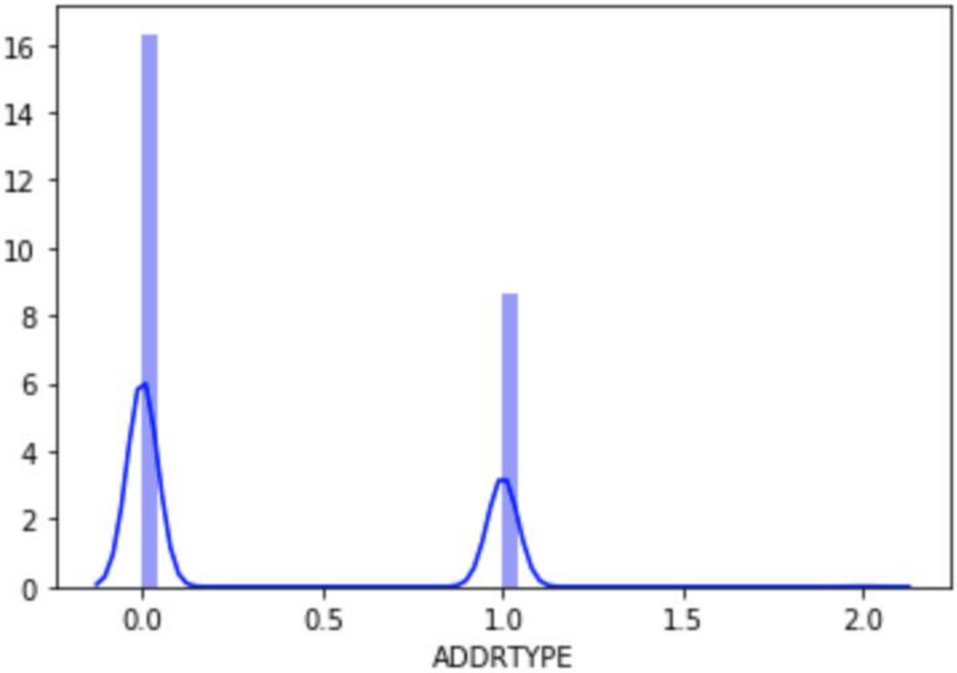
Since nearly all of the variables are transfered from categorical variables. Testing their relationship will not be that useful. Therefore, in this section we focus on the distribution of different variables.

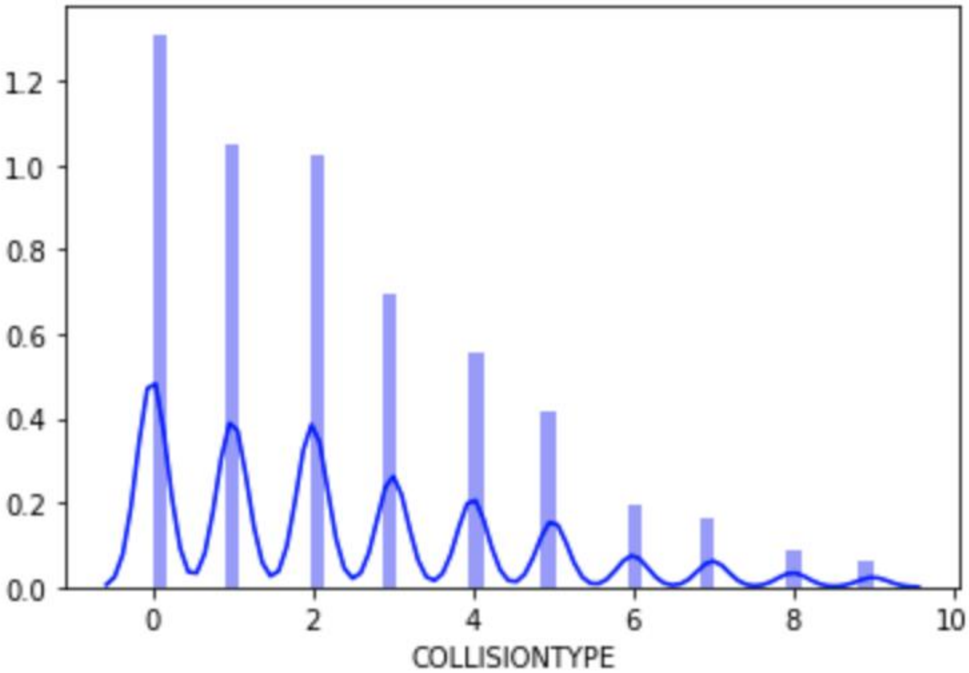
Distribution of SEVERITYCODE

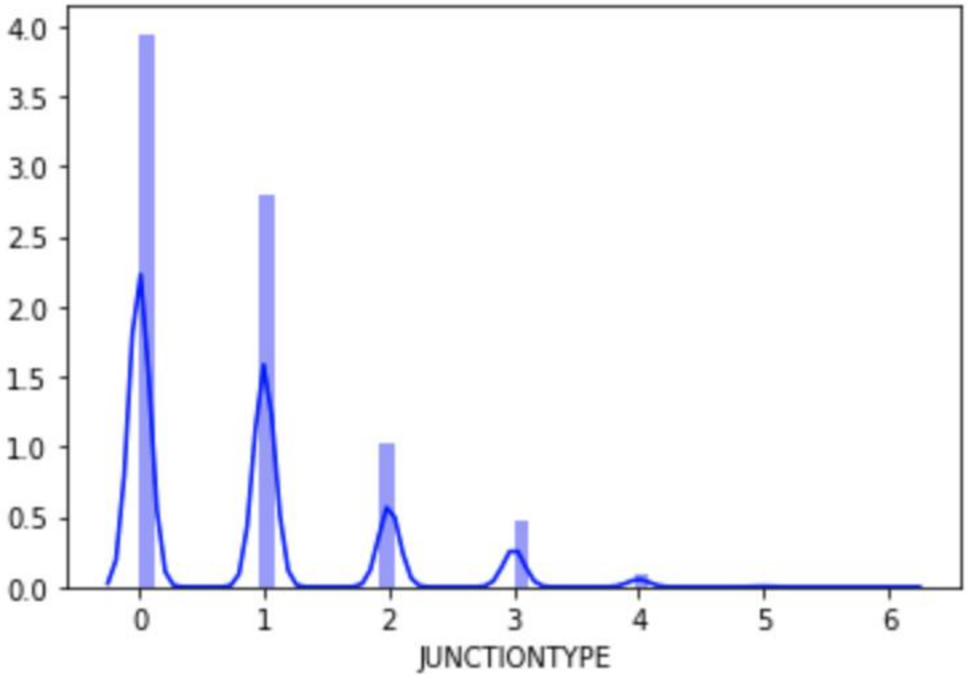


From the displot we can see the distribution of the SEVERITYCODE. Although in the metadata form, there are 5 types of SEVERITYCODE in total. However, in this form there are only 2 types of SEVERITYCODE, with 1 representing prop damage and 2 representing injury. The ratio is about 2:1.

Distribution of address, collision and junction type

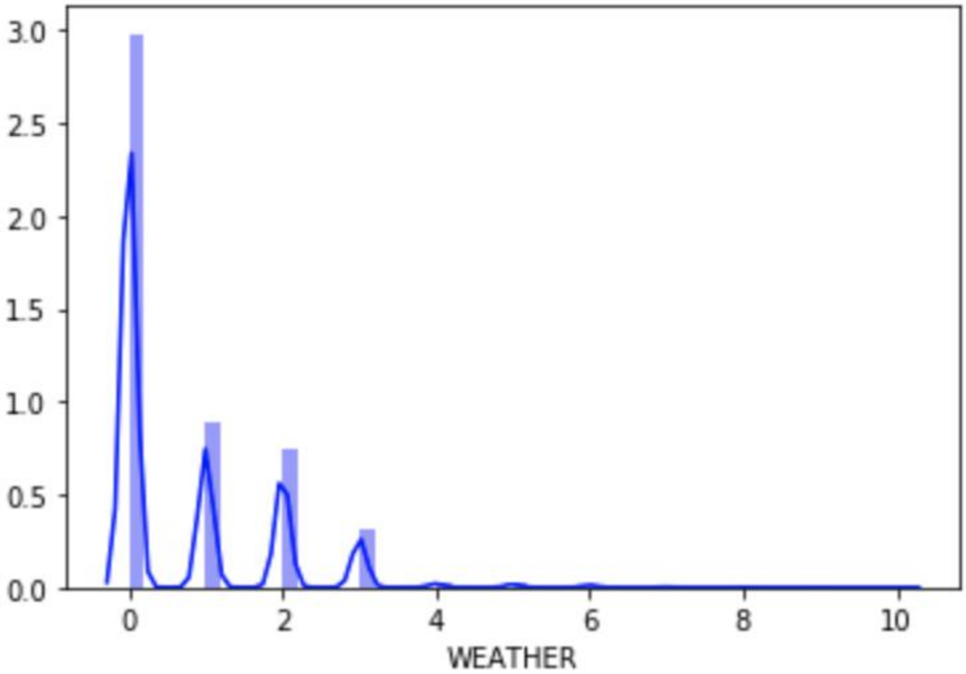


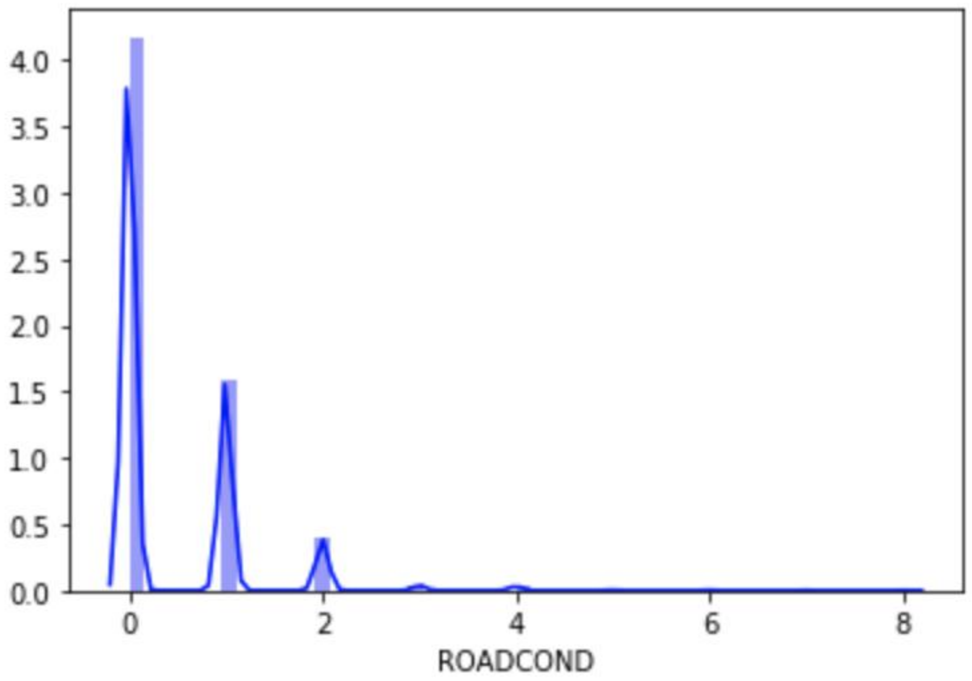


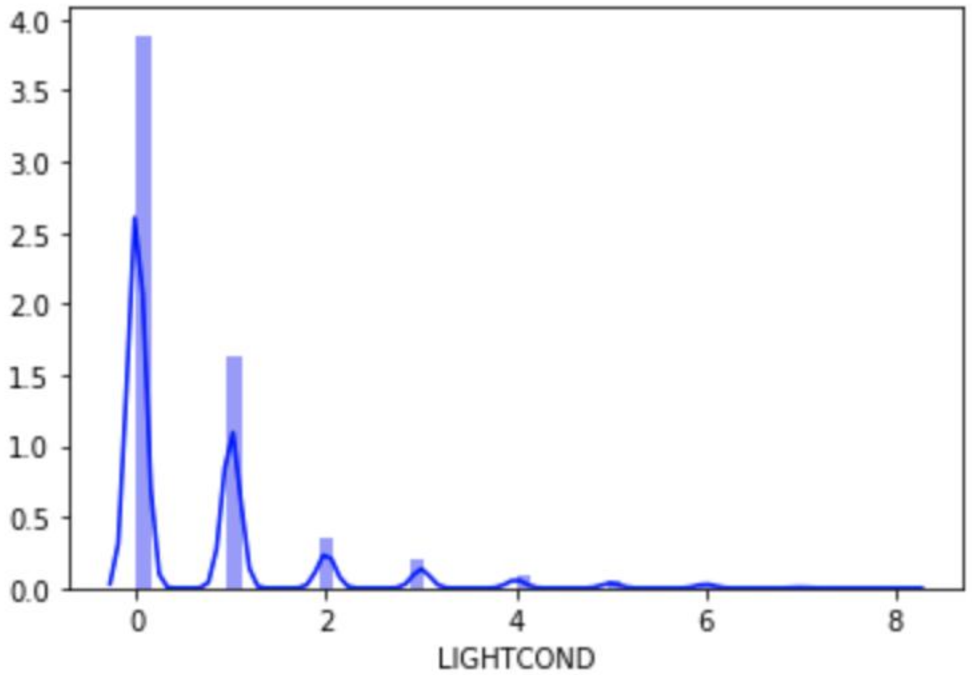


From these distribution plot we can see that in a block the accident is more likely to happen. For collisiontype, the most frequent collision is driving into a parked car, while the most frequent junction type is in the mid-block.

Distribution of environment





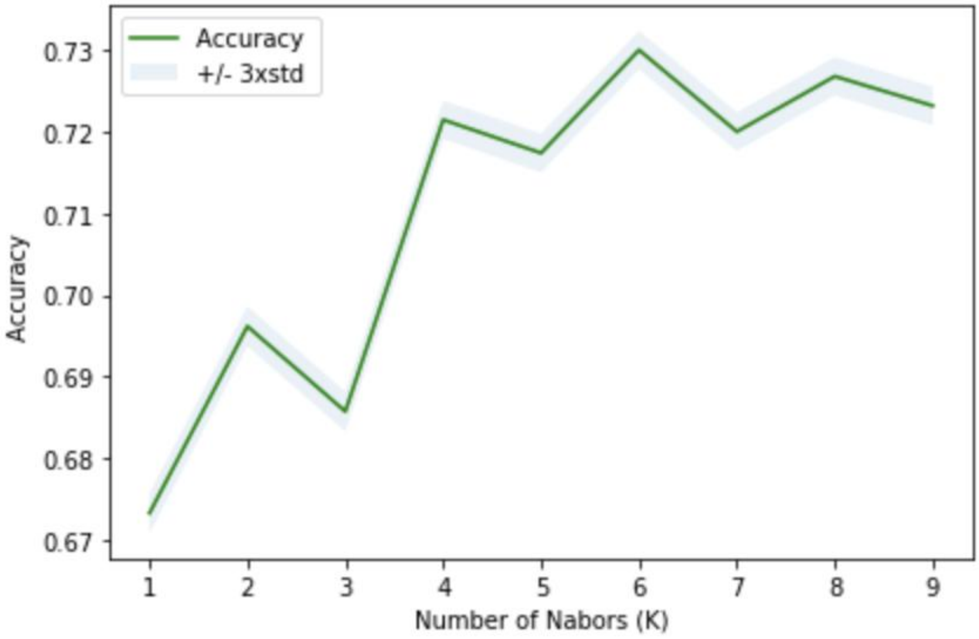


From these 3 distribution plot, we can know the condition of the accident.The most frequent weather condition is surprisingly the clear weather. The road condition and light condition are dry road and daylight respectively. Therefore, we still need to be careful in good weather and road condition.

Machine learning

K-nearest neighbors

In the K-nearest neighbors method, firstly we divide the data into dependent metric and independent metric, and then normalize the data and split the data into train set and test set. After the preparation, we can start the training process. Because we don't know a correct k, we draw the line for 1-10 to decide which k is the most accurate.



From this line, we can see that the accuracy is the highest when k=6,therefore, we chose k=6 and finish the process. And then we use the jaccard similarity score to test the accuracy. And the rate is 0.7300636977500752.

Other machine learning skills

In the decision tree process, we use the train and test sets that we defined in the previous section, and apply the decision tree method. The jaccard similarity score is 0.7466305803876541. Using the same method, the jaccard similarity score of logistic regression model is 0.7380190819869324

Machine learning result

In this section,the score of the results are

Method	K-nearest neighbors	Decision tree process	Logistic regression
Score	0.7300636977500752	0.7466305803876541	0.7380190819869324

Therefore, the K-nearest neighbors has the best result.

Future directions

In this project, there are still a lot to improve. For example, we can apply the SVM machine learning method. But when I apply the SVM, because of the huge amount of data and my old computer, there is still no result after one hour. So I have to give up the method. Secondly, we can use more data visualization method. For example, in the beginning, we can choose more visualization plots rather than only distribution plot. And in the machine learning process more visualization tools can also be applied so that the research process will be more clear.