Predicting car accident severity

1. Introduction

Car accidents are severe threatens nowadays for people's safety. For most cases, car accidents are related to many conditions which can be avoided in advance, thus predicting car accident severity is important for a wide variety of people. Firstly, car accident severity is related to road conditions, light conditions etc., it can be helpful for traffic administrators to take efforts to improve traffic infrastructures. Secondly, based on all conditions related to car accident, developing a model to predict car accident severity can be useful to bring up advices for drivers before they get into trouble. More importantly, with the development of auto-driving in the future, a smart model should be built to avoid car accidents, deeper understanding of all conditions related to car accident is required.

2. Feature selection

To predict car accident severity, we have analyzed all collisions provided by SPD and recorded by traffic records from 2004 to present. Firstly, we need to sort out locations of car accidents, "ADDRTYPE" can be an important condition, as it's more dangerous at intersections or narrow roads. Secondly, the number of involved people, injuries is necessary to estimate severity, here "PERSONCOUNT", "PEDCOUNT", "PEDCYLCOUNT", "VEHCOUNT", "INJURIES", "SERIOUSINJURIES" and "FATALITIES" are included in our model. Thirdly, driver's conditions also matter in accidents, "INATTENTIONIND" and "UNDERINFL" are needed. Fourthly, environmental conditions is useful to predict possibility of accident in advance, "WEATHER", "ROADCOND", "LIGHTCOND" and "SPEEDING" are included in our model. The target is to predict severity, thus "SEVERITYCODE" is used for our supervised model.

3. Data cleaning

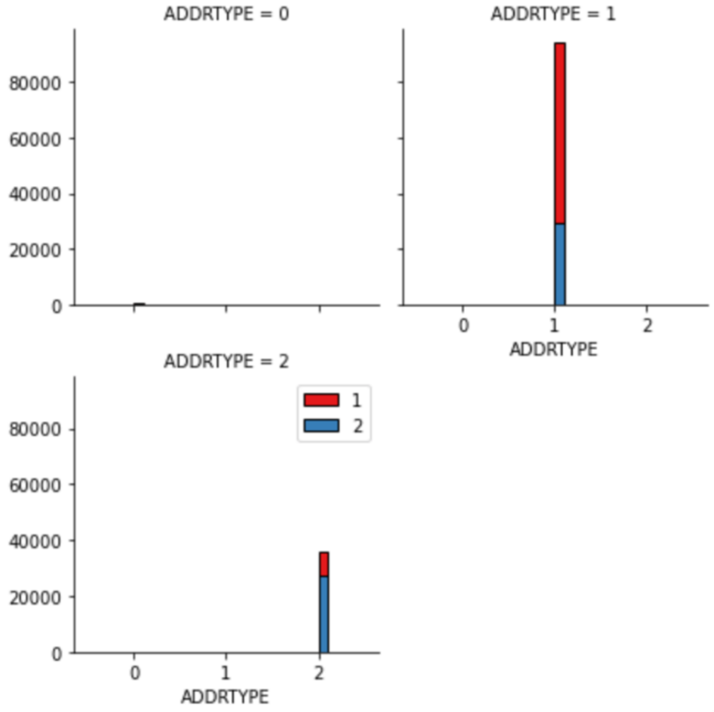
Based on the above analysis, we have looked into the data in more detail. At first, we have removed unnecessary attributes, and focused on "SEVERITYCODE", "ADDRTYPE", "PERSONCOUNT", "PEDCOUNT", "VEHCOUNT", "INATTENTIONIND", "UNDERINFL", "WEATHER", "ROADCOND", "LIGHTCOND", "SPEEDING". "ADDRTYPE" contains three types "Alley", "Block", "Intersection", we have transformed the text into numbered data "0", "1" and "2". For nan in "INATTENTIONIND" and "SPEEDING", we have replaced them into "0", thus if the accident is related to "INATTENTIONIND" and "SPEEDING", the data should be "1". For nan in "UNDERINFL", we have removed these data. For "WEATHER", "ROADCOND", "LIGHTCOND", nan data are removed at first, and then transformed into numbered data. For other attributes, nan data are removed before building a model.

"SEVERITYCODE" is used as the target to be predicted, which contains two types "1" and "2" based on the table. The problem can be seen as a classification problem, and the model should be built on supervised learning. K nearest neighbor (KNN), decision tree, support vector machine (SVM) and logistic regression are modeled for the problem to find the best predicting.

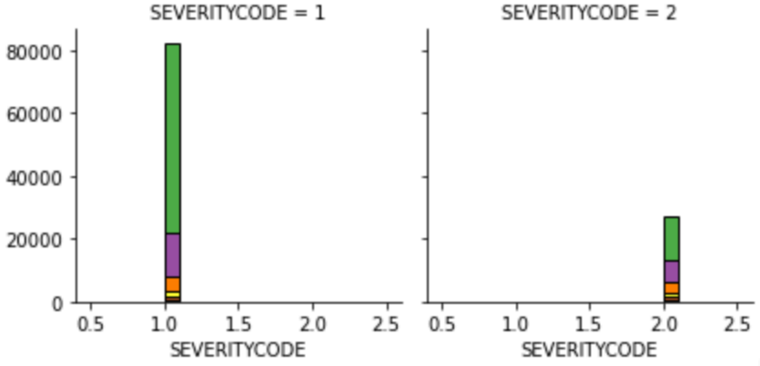
4. Methodology: data analysis

Before building the model, the relationship between selected attributes and "SEVERITYCODE" are visualized.

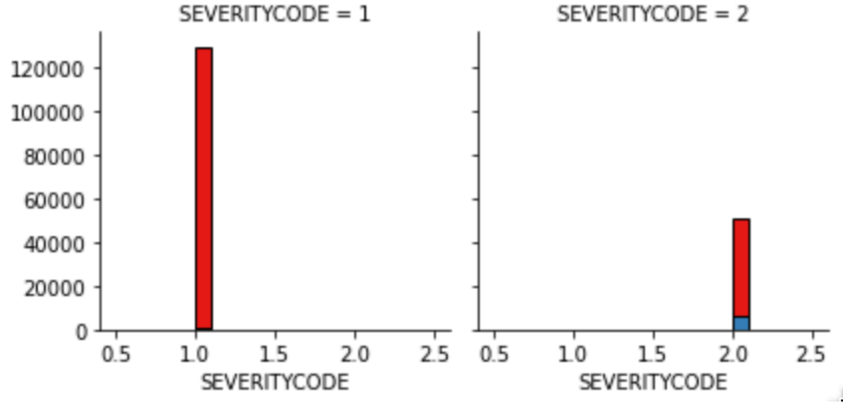
For "ADDRTYPE", most accidents take place at "Block", and "Intersection" is also more likely for car accidents to occur. Most accidents result to prop damage. At "Intersection", there's a higher probability for human injury.



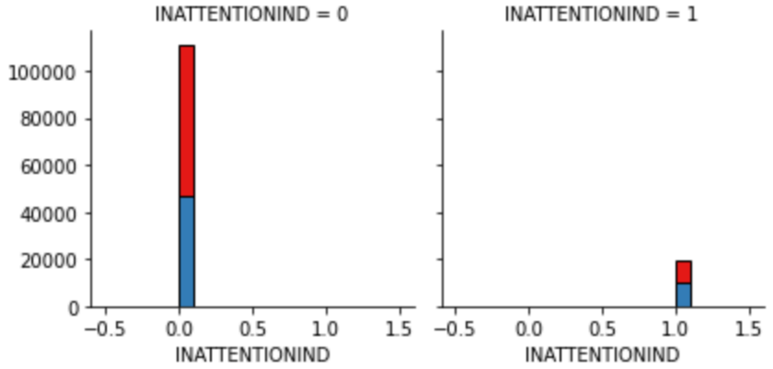
When the number of involved people becomes higher, the car accidents are less. Most car accidents involve less than 10 people.



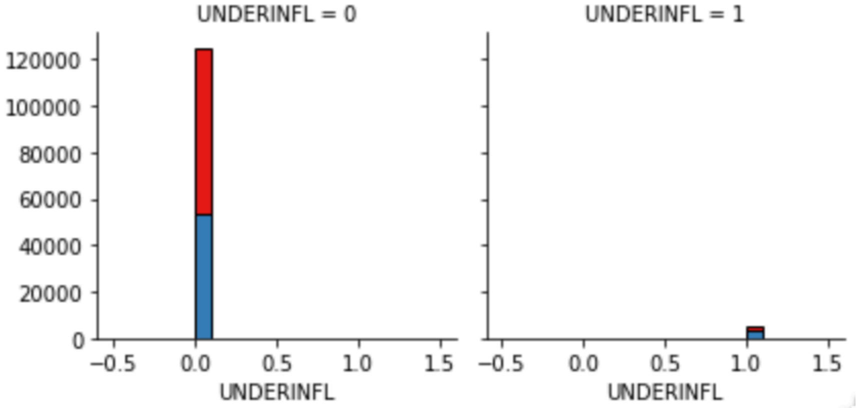
Bicycles are more likely to involved into accidents with "SEVERITYCODE" equals 1. The relationships with “VEHCOUNT” and “PEDCOUNT” are similar.



More accidents will occur when the driver is in inattention conditions. Thus, it can be helpful to avoid car accidents if the car system can remind the driver in time.



For most cases, car accidents are not related with drugs or alcohol.



Car accidents are closely related with weather, road conditions, light conditions and speeding, which are similar to the above discussions.

5. Methodology: model building

5.1 KNN

KNN classification model is applied at first to predict the severity based on the selected attributes. The test data contains 20% of the total datasheet, and different k is used for iteration. Based on average accuracy result (0.74998), the best k is 20 for the model.

k	1	2	3	4	5
Average accuracy	0.65865885	0.71534462	0.67758965	0.73491534	0.71369151
Standard deviation	0.00244839	0.00233009	0.00241348	0.00227912	0.00233414
k	6	7	8	9	10
Average accuracy	0.73411545	0.72417011	0.73894147	0.72720971	0.73363552
Standard deviation	0.00228131	0.00230779	0.00226793	0.00229985	0.00228262
k	11	12	13	14	15
Average accuracy	0.73110252	0.74475403	0.74491401	0.74755366	0.74464738
Standard deviation	0.00228949	0.00225134	0.00225088	0.00224316	0.00225165
k	16	17	18	19	20
Average accuracy	0.74907346	0.74808692	0.74859352	0.74811358	0.74998
Standard deviation	0.00223867	0.00224159	0.0022401	0.00224151	0.00223598
k	21	22	23		
Average accuracy	0.74934009	0.74912678	0.74758032		
Standard deviation	0.00223788	0.00223852	0.00224309		

5.2 Decision tree

The mean accuracy of decision tree is 0.7502199706705772, and the standard deviation is 0.0022352627543070426.

5.3 SVM

The mean accuracy of SVM model is 0.7538461538461538, and the standard deviation is 0.0022243344458415822.

5.4 Logistic regression

The mean accuracy of Logistic regression model is 0.7495800559925343, and the standard deviation is 0.00223716946766595.

Model	Average accuracy	Standard deviation
KNN	0.74998	0.00223598
Decision tree	0.7502199706705772	0.0022352627543070426
SVM	0.7538461538461538	0.0022243344458415822
Logistic regression	0.7495800559925343	0.00223716946766595

Based on the four evaluated model, SVM should be the best model to predict car accidents based on the selected attributes.

6. Conclusion

In conclusion, I have analyzed the relationship between car accident and a series of conditions including “ADDRTYPE”, “PERSONCOUNT”, “PEDCOUNT”, “VEHCOUNT”, “INATTENTIONIND”, “UNDERINFL”, “WEATHER”, “ROADCOND”, “LIGHTCOND”, “SPEEDING”. These attributes show strong influence on the severity of car accidents. I have built KNN, decision tree, SVM and logistic regression models to predict the severity of car accidents. Based on the accuracy result, SVM exhibit the best predicting accuracy based on the split dataset. This work may be helpful for traffic administrators to improve infrastructure, considering that road condition and light condition is necessary to avoid car accidents. In addition, the relationship with driver’s attention indicate the necessity to monitor and remind the driver’s psychosis. Moreover, based on the analyzed attributes, the model in this work may also improve researches into auto-driving systems in the near future.