Prediction models for estimating severity of collision in Seattle

Introduction

In our daily life, there often happens collision in our cities. The bad weather, the bad road conditions, and so on factors could lead to traffic accidents. Usually, some severe traffic accidents may cause big traffic jams. It is necessary to take timely actions to prevent or reduce traffic congestions. But how? From the brief report from the privies, the weather report, the conditions of the position, and so on, the information may help instructors make predictions. If there may happen a severe collision, traffic station could cut down the car flow and prevent more cars from driving in the accident position. Therefore, a prediction model is needed to evaluate the severity of specific collisions.

Business Audience

Those traffic managers in big cities may benefit from the analysis of collision severity since they could use predictive models to get predictions to make decisions about the traffic affairs.

Data

The data used in this capstone project is the collisions data collected from Seattle, US. It contains all the collisions happened since 2004 to present. The data set include 194673 rows and 38 columns. The attributes mainly include the location, severity, collision type, number of pedestrians, bicycles, and vehicles involved, injuries and fatalities, time, weather, road condition, light condition, and speed. The information is too much, and therefore, we should select only some of them which are highly related to the prediction.

Methodology

This project first selects several variables which are highly related to the dependent variable – severity code. Then, wash the data by dropping rows with null data. After that, we visualize several selected variables. Finally, we train decision tree model, SVM model, and logistic regression model and compare them using the final accuracies.

The variables we selected are listed as follow:

Variable	Description
SEVERITYCODE	Severity of collisions
ADDRTYPE	Address type of collisions
SEVERITYDESC	Detailed description of the severity of collision
COLLISIONTYPE	Collision type
PERSONCOUNT	Total number of people involved in the collision
PEDCOUNT	Total number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision.
JUNCTIONTYPE	Category of junction at which collision took place
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions

	during the time of the collision.
ROADCOND	Road condition
LIGHTCOND	Light condition
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

We also found that there exist plenty of null values in our data set. We directly delete those rows with null values. The value counts for the null values in each column is listed:

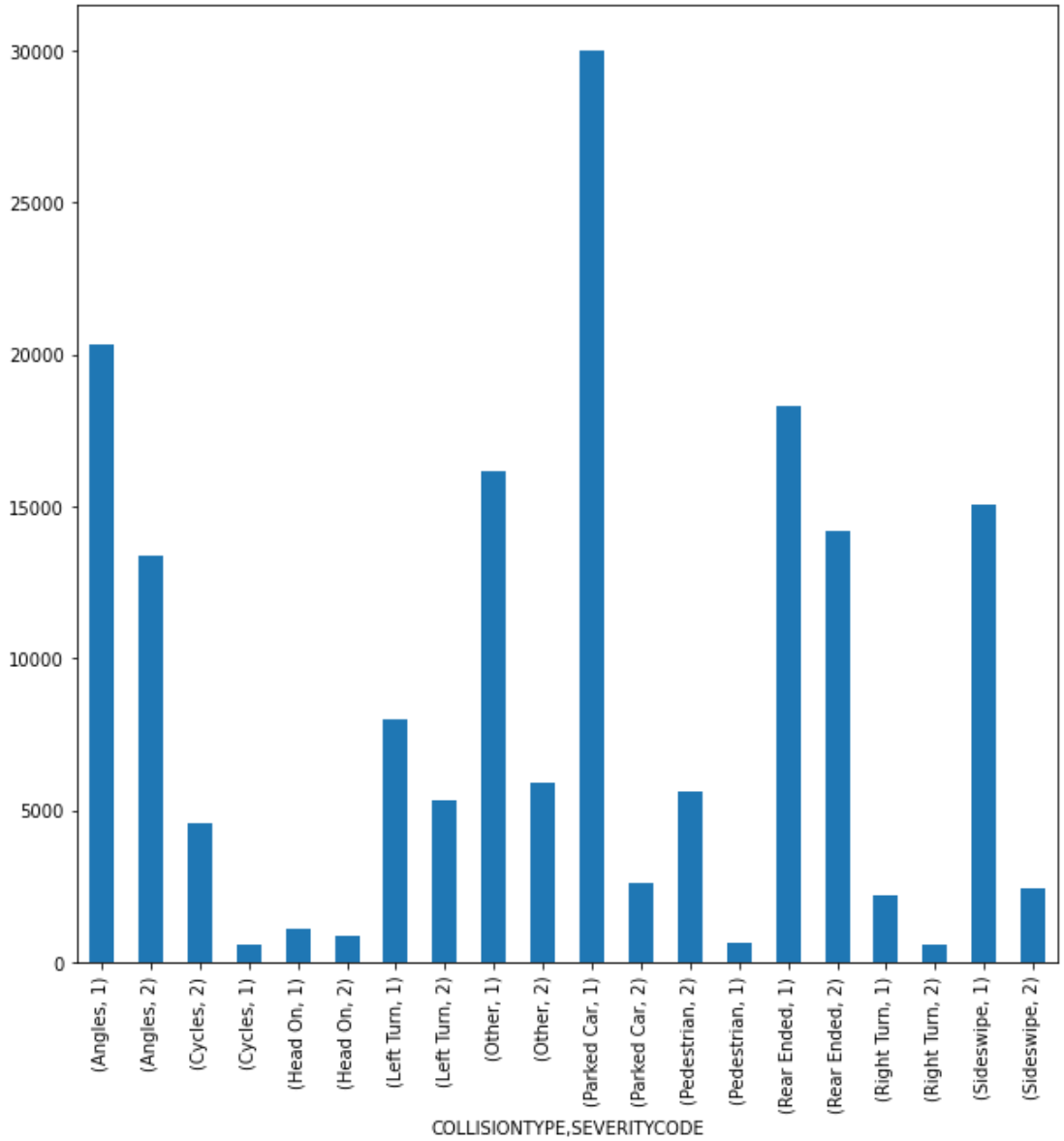
```
SEVERITYCODE          0
ADDRTYPE           1926
SEVERITYDESC          0
COLLISIONTYPE      4904
PERSONCOUNT           0
PEDCOUNT              0
PEDCYLCOUNT           0
VEHCOUNT              0
JUNCTIONTYPE       6329
INATTENTIONIND   164868
UNDERINFL          4884
WEATHER            5081
ROADCOND           5012
LIGHTCOND          5170
PEDROWNOTGRNT    190006
SPEEDING         185340
HITPARKEDCAR          0
dtype: int64
```

For the variables INATTENTIONIND, PEDROWNOTGRNT, and SPEEDING, the null values represents 0 or no. Therefore, we only manipulate on other variables.

In some variables, the value is Y or N. We transform them into 1 and 0 respectively.

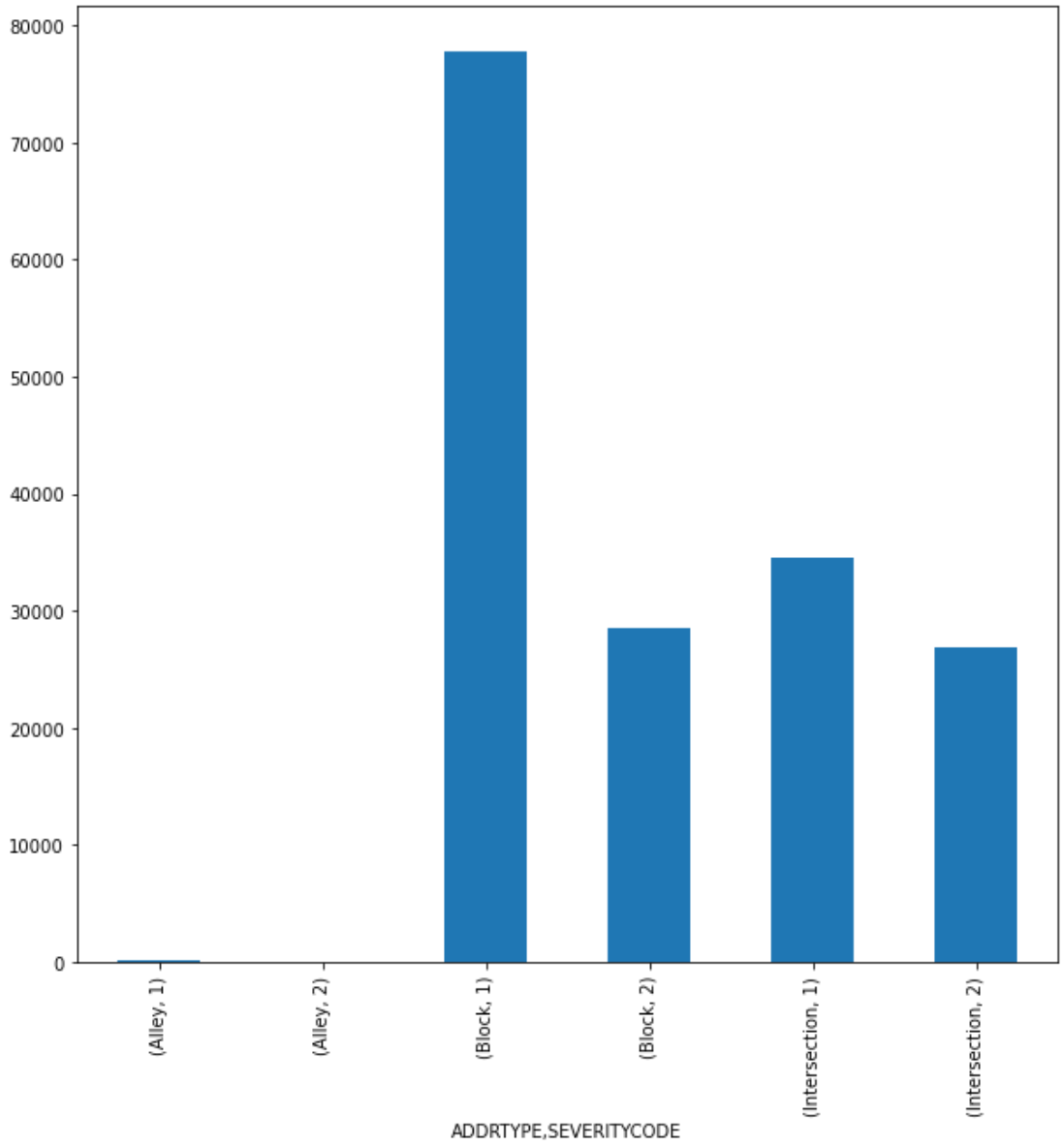
Data Visualization

We first plot the bar plot with variables collision type and severity code. The bar plot is shown as below:



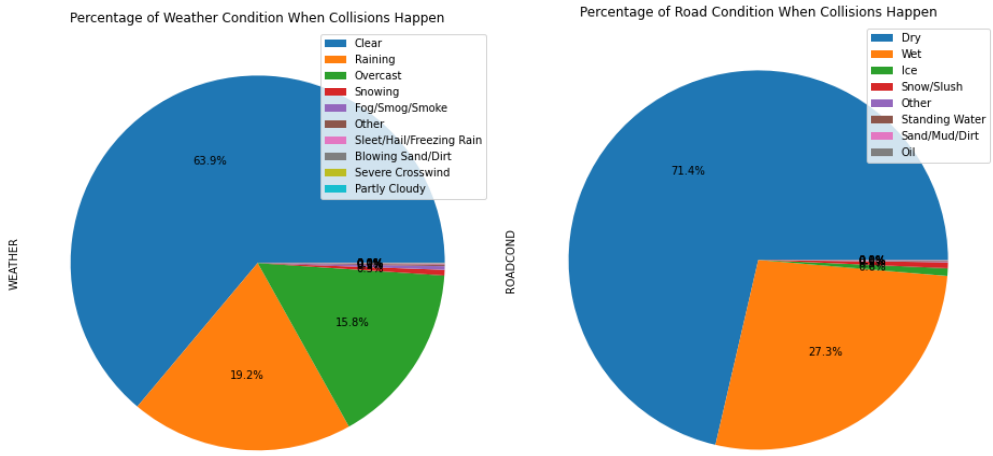
We can observe that when the collision type is cycles, the severity of collision is more likely to be 2 than 1. This means in this condition, the result is much dangerous than others.

Then, we plot a bar plot with variables address type and severity code. The plot is shown as below:



We can observe that, when the address type is intersection, the severity of collision is more likely to be 2 than 1. This means the collisions at intersections are more likely to be dangerous than other conditions.

We plot a pie chart to visualize the variables weather and road condition.



Most collisions happen when the weather is clear and the road is dry. This matches the common situations. Besides, the collisions usually happen when it is snowing or overcast with wet road condition.

Classification methods

Since the dependent variable is binary, we apply classification method to build models to predict it. We choose decision tree model, support vector machine, and logistic regression model. The decision tree model is very easy to apply and explain when facing a new case. The support vector machine model is also very simple to classify and predict. The logistic regression model can help predict the probabilities of the two possible results.

One hot coding

Before we build the models, we should first use one hot coding method to represent those variables whose values are strings by dummy variables.

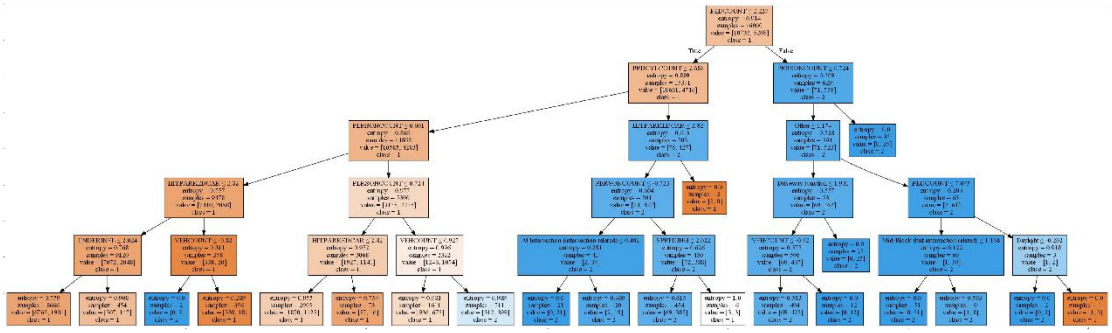
Modelling

We first normalize the data, and then split the data into training set and testing set. Here we

use 80 percent of the data to train the model and the rest 20 percent of the data to test the prediction.

For the decision tree model, we set criterion as entropy. The max depth of the tree is five. After training and prediction, the accuracy for the decision tree model is 72.25%.

The tree model graph is shown as follows:



Secondly, we build support vector machine model using the same data set. The accuracy of SVM is 71.75%.

Ultimately, we build a logistic regression model. The accuracy of logistic model is 72.4%.

Result and Conclusion

The accuracies for the three models are listed in the tables below:

	Accuracy	F1-score	Log_loss
Decision tree	0.7225	0.6722	NA
SVM	0.7175	0.6499	NA
Logistic regression	0.7240	NA	0.5734

Discussion

From the results, the accuracies are very close. Therefore, the decision tree model and the logistic regression model are preferred to be applied. Between them, the decision tree model will be preferred when a simple and easy to explain model is needed. The logistic model is very useful to predict the probability of severity of collisions.

Future Directions

In this project, many other variables such as positions, the date and time, and detailed collision descriptions are not used in building the models. Further, since the alerting model should be faster, the number of inputs should be easier to gather and record. Simpler models are needed in real cases. In this way, it can shorten the time for traffic managers to collect necessary information and make a decision in a short while.