Predict the Severity of An Accident

1. Introduction

1.1. Business Problem

We need a system to could warn that the weather and the road conditions about the possibility of we getting into a car accident and how severe it would be, so that we would drive more carefully or even change our travel if we are able to.

1.2. Interest

All driver personal car or public transportation should be need interested audience for this solution solving problem and maybe any factory will make device tool for warning system installed at vehicle.

2. Data acquisition and cleaning

2.1. Data Source

To provide the stakeholders the necessary information, I use dataset collision all year download from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv which this dataset collisions provided by SPD and recorded by Traffic Records including all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe dataset contain data from 2004 to present that have attribute data:

Attribute	Data type, length	Description
OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • **Alley** • **Block** • **Intersection**
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.

Attribute	Data type, length	Description
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

Figure 1: Collision Dataset

For detail attribute dataset we can find at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

2.2. Data Cleaning

Some action for data cleaning from the dataset is manage missing value with replace all null value with "Unknown", especially in attribute "SPEEDING" replace null value with "N", drop attribute “X", "Y". After cleaning process, there are 38 attribute and 196.126 rows data.

3. Exploratory Data Analysis

There is some result of exploratory data collision base on state collision code in figure 2 and severity code frequency in figure 3 as bellow of figures.

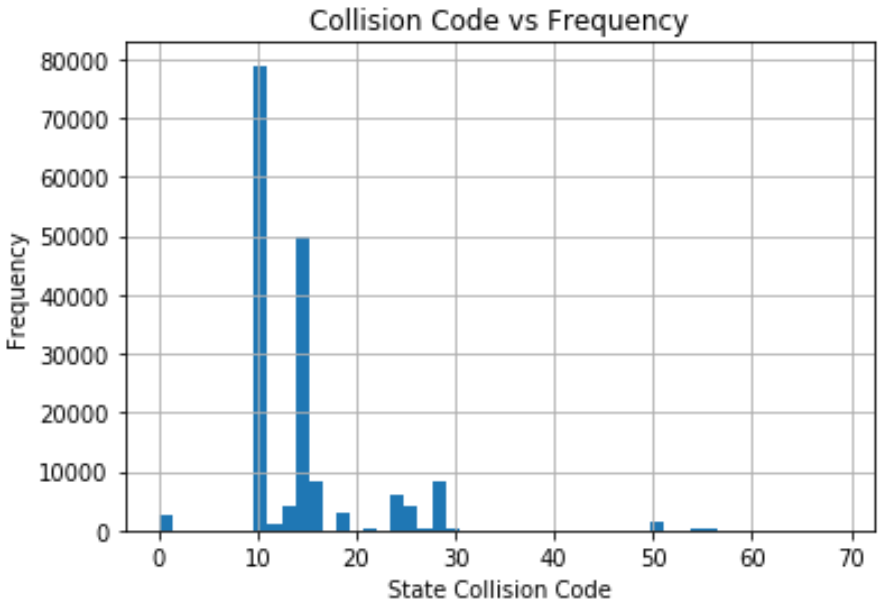


Figure 2 Collision Code vs Frequency

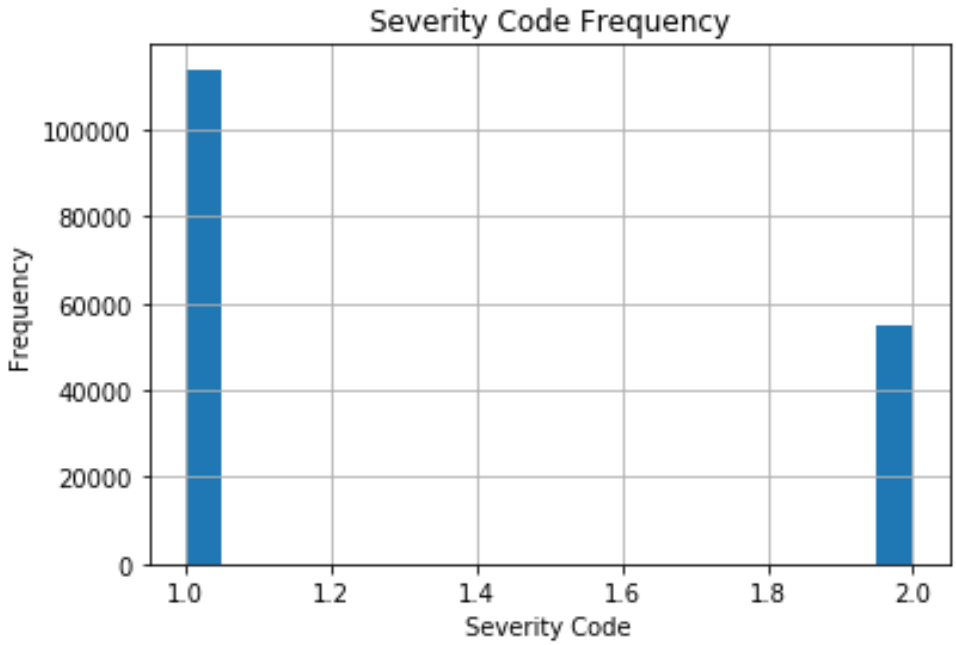


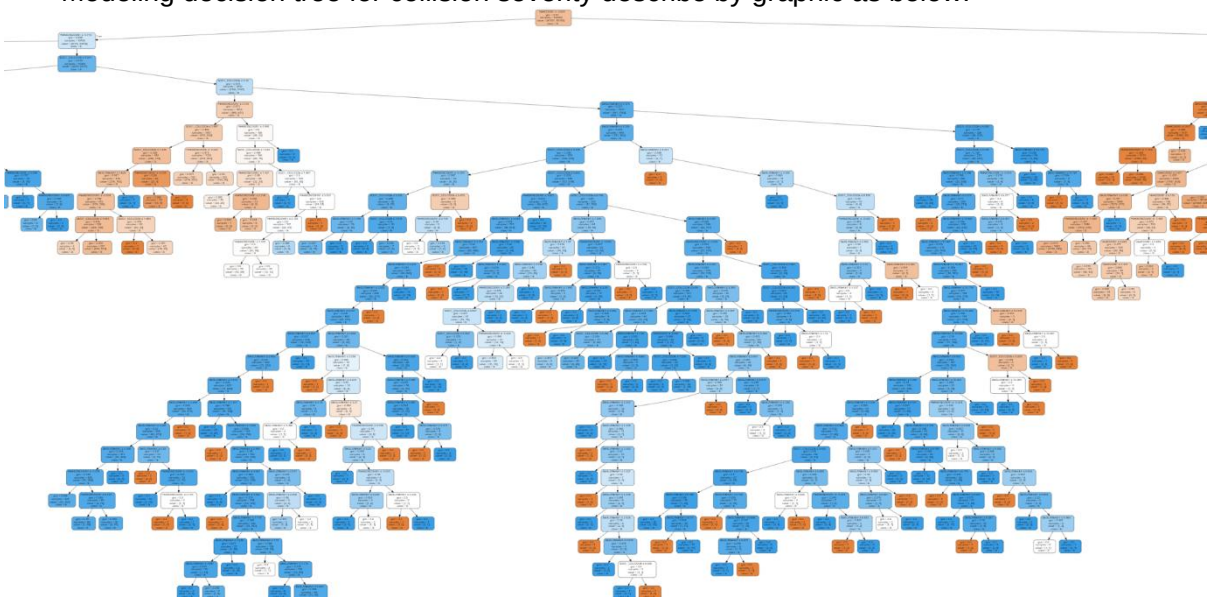
Figure 3Severity Code Frequency

4. Predictive Modeling

There are two types of models, decision tree and KNN, that can be used to predict collision. Decision tree models can give predict severity collision for new data.

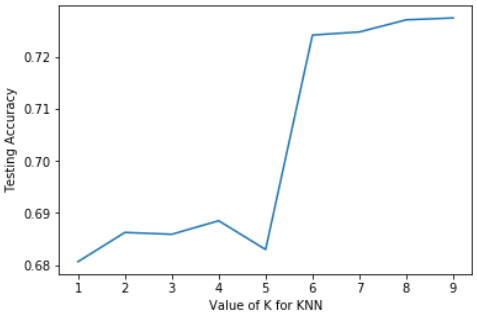
4.1. Decision Tree Model

Modeling decision tree for collision severity describe by graphic as below:



Predicted values [2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1]
F1-Score : 0.6902247634772901
Jaccard : 0.7383739837398374

4.2. KNN Model




```
Accuracy Test with K= 1:  0.6806598474546207
Accuracy Test with K= 2:  0.6862768284751375
Accuracy Test with K= 3:  0.6859220717791048
Accuracy Test with K= 4:  0.6885236208833442
Accuracy Test with K= 5:  0.6829953290368356
Accuracy Test with K= 6:  0.7241766688346243
Accuracy Test with K= 7:  0.7247679299946787
Accuracy Test with K= 8:  0.7271034115768935
Jacard: 0.05

Predicted using k = 9: [1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 2 2 1 1 2]
F1-Score                : 0.6952661826753996
Jaccard Score           : 0.7204096354197462
```

5. Conclusions

This report may be helpful for someone who drive a personal car or public transportation, so it shall not be used as a single decision-making tool or some factory will make a device tool for warn device installed at vehicle.

6. Future directions

Models in this study mainly focused on driver, vehicle and weather features. However, interactions with traffic signs might also contribute to decrease accident could bring significant improvements to the models.