

Capstone Project

Weigang Wei

Introduction

Business problem

It is aimed to build a model to predict the severity of an accident based on a database which contains a set of accident severity data. This model is prepared for the local transportation department and this model may be used to set up diverted routes, and temporary speed limits etc.

Data

General description of the dataset

The dataset contains 194673 attributes and 38 columns. The general information of the dataset is shown below.

#	Column	Non-Null Count	Dtype
0	SEVERITYCODE	194673	non-null int64
1	X	189339	non-null float64
2	Y	189339	non-null float64
3	OBJECTID	194673	non-null int64
4	INCKEY	194673	non-null int64
5	COLDETKEY	194673	non-null int64
6	REPORTNO	194673	non-null object
7	STATUS	194673	non-null object
8	ADDRTYPE	192747	non-null object
9	INTKEY	65070	non-null float64
10	LOCATION	191996	non-null object
11	EXCEPTRSNCODE	84811	non-null object
12	EXCEPTRSNDESC	5638	non-null object
13	SEVERITYCODE.1	194673	non-null int64
14	SEVERITYDESC	194673	non-null object
15	COLLISIONTYPE	189769	non-null object
16	PERSONCOUNT	194673	non-null int64
17	PEDCOUNT	194673	non-null int64
18	PEDCYLCOUNT	194673	non-null int64
19	VEHCOUNT	194673	non-null int64
20	INCDATE	194673	non-null object
21	INCDTTM	194673	non-null object
22	JUNCTIONTYPE	188344	non-null object
23	SDOT_COLCODE	194673	non-null int64

24	SDOT_COLDESC	194673	non-null	object
25	INATTENTIONIND	29805	non-null	object
26	UNDERINFL	189789	non-null	object
27	WEATHER	189592	non-null	object
28	ROADCOND	189661	non-null	object
29	LIGHTCOND	189503	non-null	object
30	PEDROWNOTGRNT	4667	non-null	object
31	SDOTCOLNUM	114936	non-null	float64
32	SPEEDING	9333	non-null	object
33	ST_COLCODE	194655	non-null	object
34	ST_COLDESC	189769	non-null	object
35	SEGLANEKEY	194673	non-null	int64
36	CROSSWALKKEY	194673	non-null	int64
37	HITPARKEDCAR	194673	non-null	object

The aim of this project is to predict the severity which is described by the severity code. The description of the codes are:

- Code 1: Property Damage Only Collision, and
- Code 2: Injury Collision.

The number of different severity is shown in Figure 1.

Column is: SEVERITYCODE

1 136485

2 58188

Name: SEVERITYCODE, dtype: int64

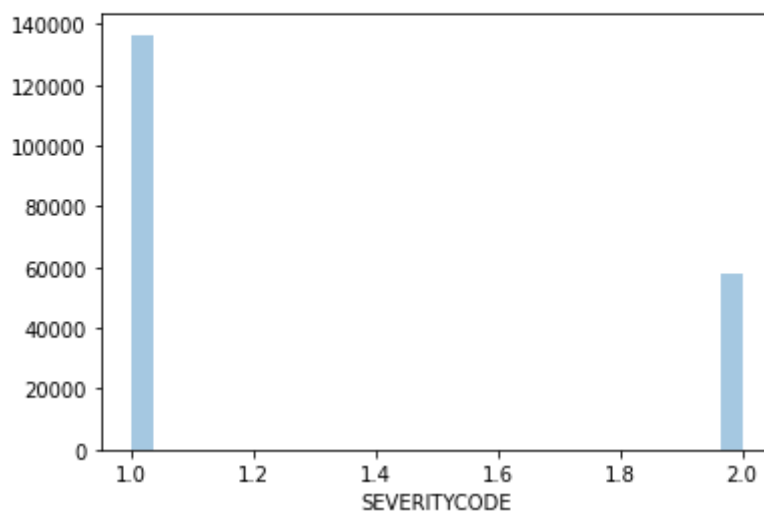


Figure 1: Number of different severity in the data set

19 number columns contains NaN value, and the percentage of NaN in 7 columns exceeds 30%, as shown in Figure 2.

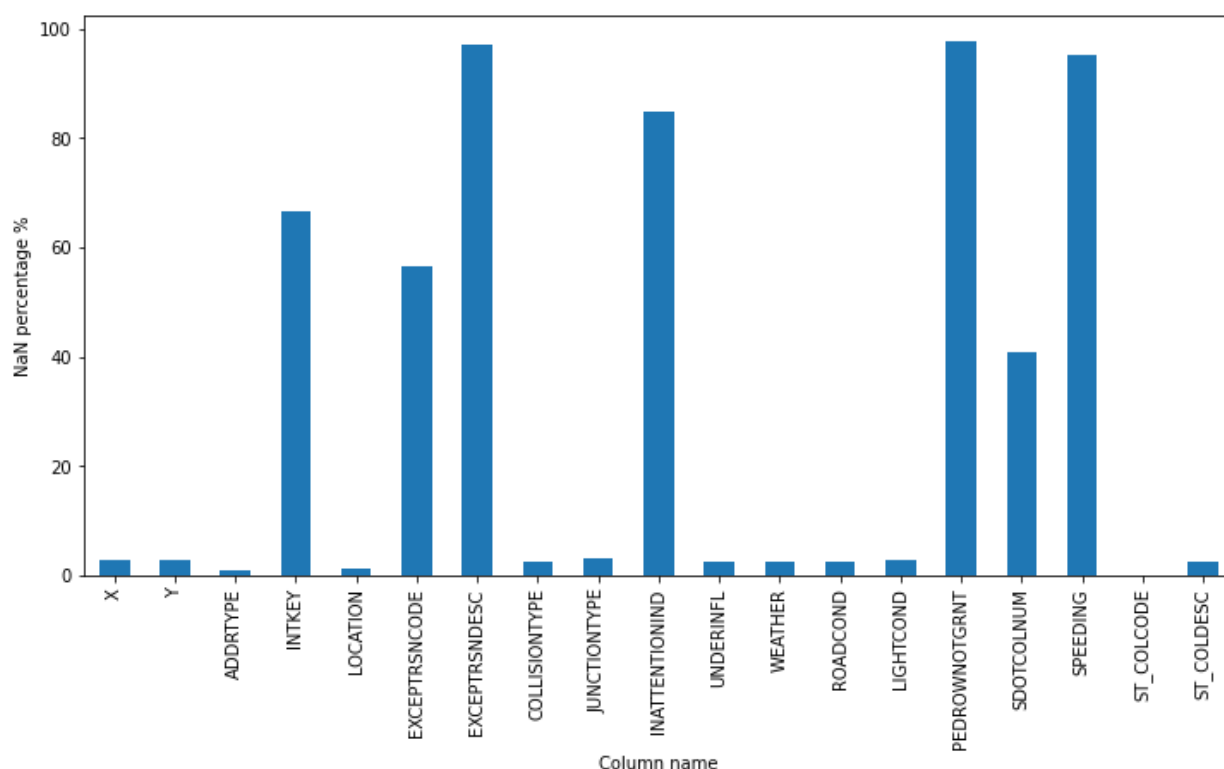


Figure 2: Columns having NaN values and the percentage of NaN

Data cleaning

Columns containing coordinates, id and unexplainable information

Columns which are clearly not related to the severity of an accident are X, Y, OBJECTID, INCKEY, INTKEY, STATUS, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, UNDERINFL, ST_COLDESC, SEGLANEKEY, UNDERINFL.

Column PERSONCOUNT, PEDCOUNT, and PEDCYLCOUNT

The involved persons column “PERSONCOUNT” includes the persons within the vehicle. This is not strongly related to an accident severity. This column will be removed from any further analysis.

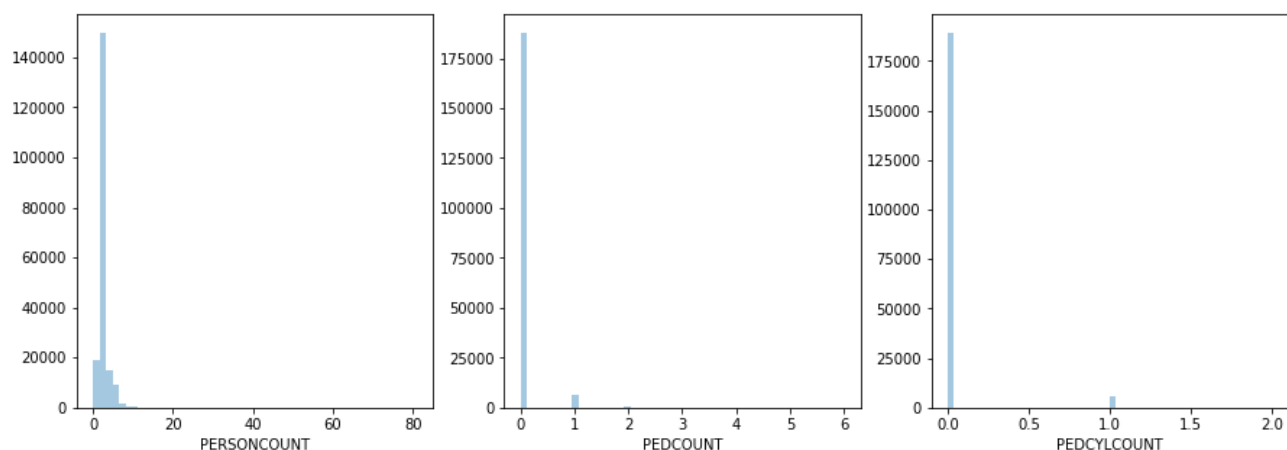


Figure 3: Attributes and distributions of *PERSONCOUNT*, *PEDCOUNT* AND *PEDCYLCOUNT*

The involved pedestrians and cyclists “*PEDCYLCOUNT*” includes the pedestrian count “*PEDCOUNT*”. Therefore the “*PEDCOUNT*” column is redundant data, and it will be removed from any further analysis.

Column *WEATHER* and *ROADCOND*

The attributes in the column *WEATHER* are “Raining”, “Clear”, “Unknown” etc. The attributes in column *ROADCOND* are “Wet”, “Dry”, “Unknown” etc. These two columns have redundant information. Therefore, only the *ROADCOND* column is considered for further analysis.

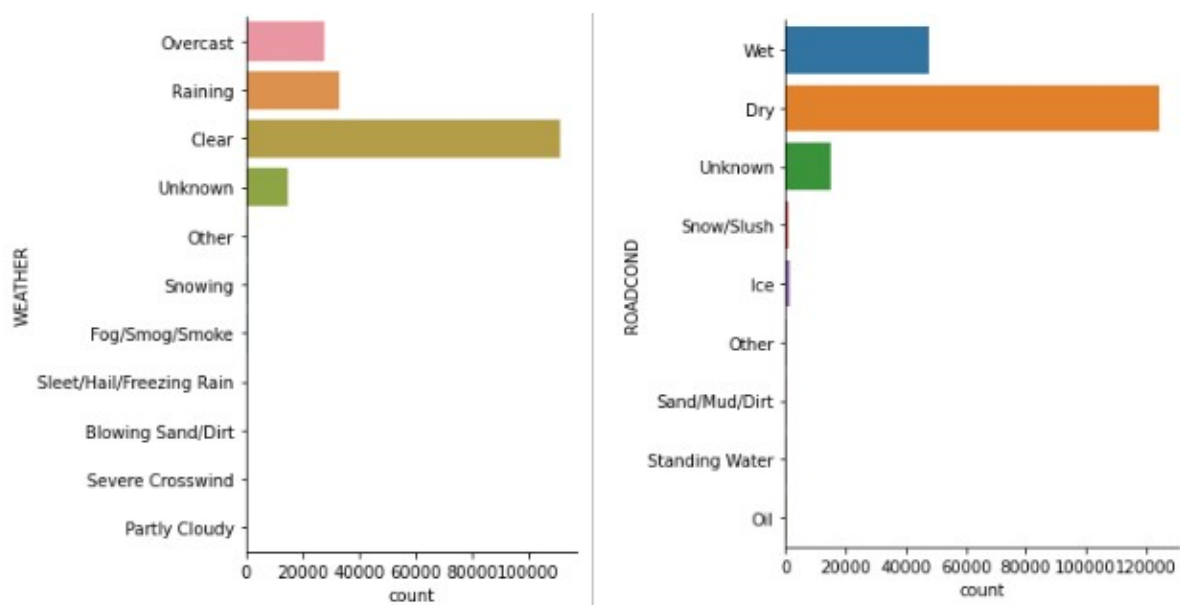


Figure 4: Attributes and distribution of column *WEATHER* and *ROADCOND*

Column SDOT_COLDESC, and SDOT_COLCODE

Column SDOT_COLDESC is the description of the column SDOT_COLCODE, which is the indication of the collision type. This is likely related to the accident severity.

Therefore, column SDOT_COLCODE will be used for further analysis. Column SDOT_COLDESC will be removed as redundant information.

Summary of the columns to be removed

As discussed above, some columns will not be considered in further analysis and modelling. They are summarised in Table 1.

Table 1: Columns not considered in further analysis and modelling to predict accident severity

Column name	Reasons not included in further analysis
X	They are the coordinates which are not related to the accident severity
Y	
SEVERITYDESC and SEVERITYCODE.1	SEVERITYCODE.1 is the same as SEVERITYCODE. SEVERITYDESC is the description of SEVERITYCODE. Therefore, these two columns are redundant information.
OBJECTID	An identification number which is not related to the accident severity
INCKEY, INTKEY	
INATTENTIONIND, INATTENTIONIND	Meaningless column, removed
STATUS	This contains binary information “Match” and “Unmatch” however what status are they are not clear. Therefore, this column is not used for further analysis
LOCATION	Accurate address of an accident, which is not related to the severity
EXCEPTRSNCODE	More than 50% data is NaN
EXCEPTRSNDESC	More than 90% data is NaN
UNDERINFL	This contains binary data “N”, “Y”, 0 and 1. However the meaning of these data is unknown. Therefore this column is not considered in further analysis
ST_COLDESC	This is the description of ST_COLCODE.
SEGLANEKEY	An identification number and not related to the accident severity
CROSSWALKKEY	
PERSONCOUNT	Redundant information of PEDCOUNT
PEDCYLCOUNT	
WEATHER	Redundant information of ROADCOND
SDOT_COLDESC	Redundant information of SDOT_COLCODE
JUNCTIONTYPE	Redundant information of ADDTYPE

Data organisation for modelling

Collision type

The collision type does not have a clear pattern related to the accident severity as shown in Figure 5.

For example, “Angles” is recorded more than 35000, however, this can be severe and not severe. Therefore, this column is removed for any further analysis.

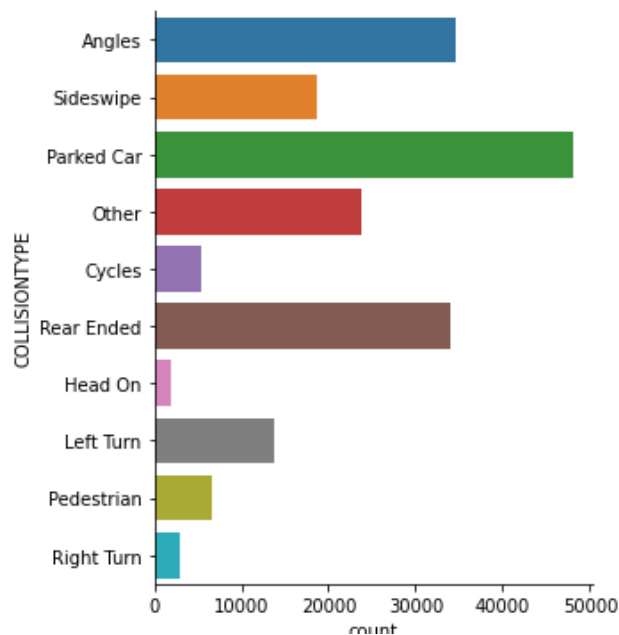


Figure 5: Collision type

Speeding

In the original data only 9333 over 194673 cases are recorded as speeding. The majority is remaining NaN. It is assumed that the remaining data is not on the condition of over speed.

[“Y”, “N”] are replaced by [1, 0]. The distribution plot is shown in Figure 6.

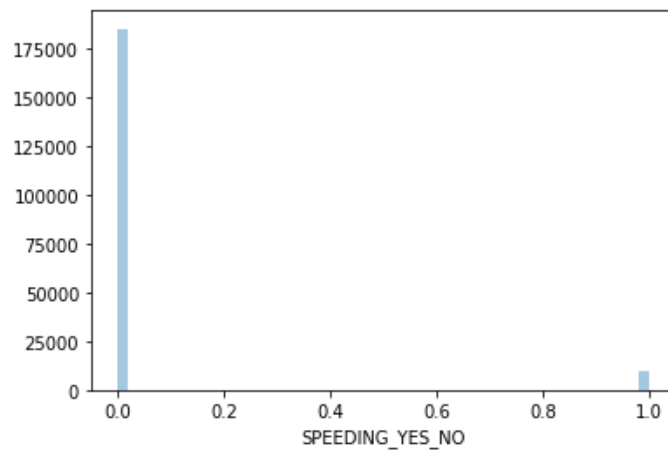


Figure 6: Speeding distribution plot

Address type

The address type in the dataset includes “Intersection”, “Block” and “Alley”, as shown in Figure 7. As the attribute “Alley” only have 751 samples, the attribute of these rows are replaced by “Block”.

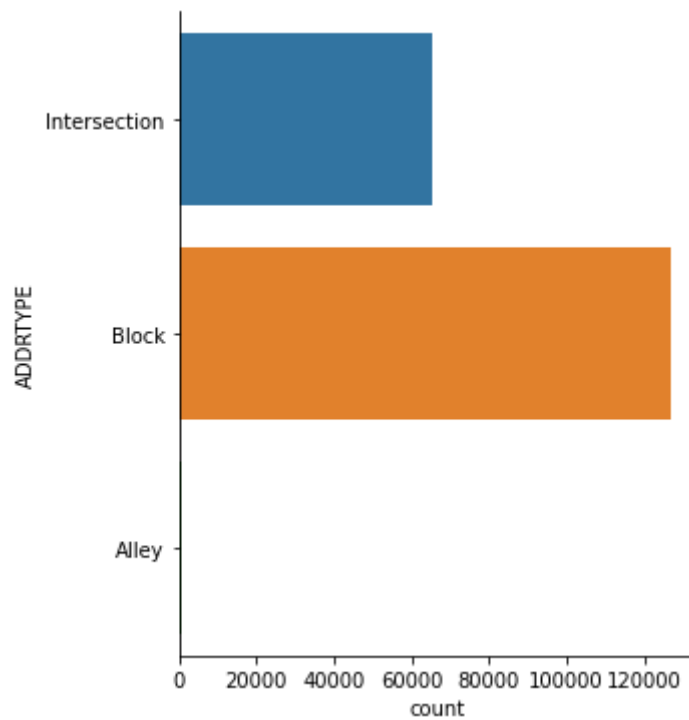
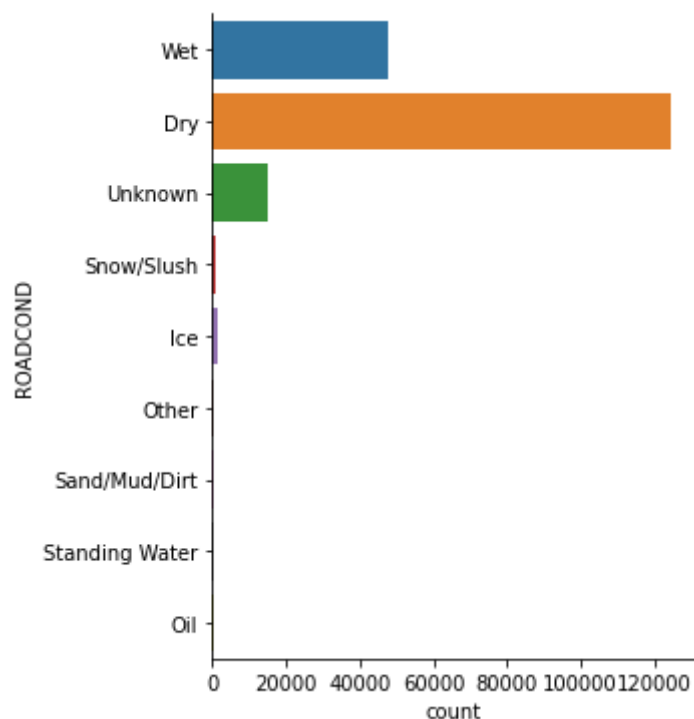


Figure 7: Address type

Road condition

There are 9 different road conditions in this column. These road conditions can be grouped to wet and dry surface road. The “Unknown” samples will be grouped into dry surface road group.



Light condition

Light condition will be included in the predicting model. The values of the light condition are also grouped to “Bright” and “Dark”. Where "Daylight", "Unknown", "Dusk", "Dawn", "Other" are categorised as “Bright” and "Dark - Street Lights On", "Dark - No Street Lights", "Dark - Street Lights Off", "Dark - Unknown Lighting" are categorised as “Dark”.

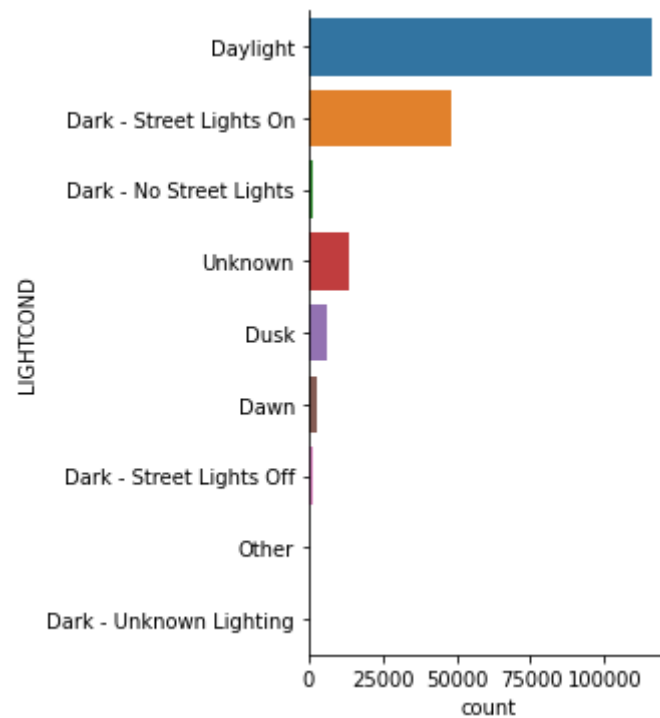


Figure 8: Light condition

Overall view of the data for modelling

The overall view of the data for modeling is shown below.

#	Column	Non-Null Count	Dtype
0	SEVERITYCODE	194673 non-null	int64
1	PEDCOUNT	194673 non-null	int64
2	VEHCOUNT	194673 non-null	int64
3	SDOT_COLCODE	194673 non-null	int64
4	SPEEDING_YES_NO	194673 non-null	int64
5	Block	194673 non-null	uint8
6	Intersection	194673 non-null	uint8
7	Dry	194673 non-null	uint8
8	Wet	194673 non-null	uint8
9	Bright	194673 non-null	uint8
10	Dark	194673 non-null	uint8

dtypes: int64(5), uint8(6)