Predicting Severity of accident from Weather, Road and Light Condition

David Oh

20 September 2020

1 Introduction

1.1 Background

When you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

1.2 Problem

Weather, Road and Light condition in Collision data set might help to understand relationship with Severity of accident so that the project aims to predict whether, what type of sub-condition in respective condition, how much particular or combined conditions related with Severity of accident. Furthermore, it would ever more preventive if some actions could be made on frequently accident occurring location hinging on impact of certain conditions or notifying head-up to police station or hospital near to those location as well.

2 Data acquisition and cleaning

2.1 Data sources

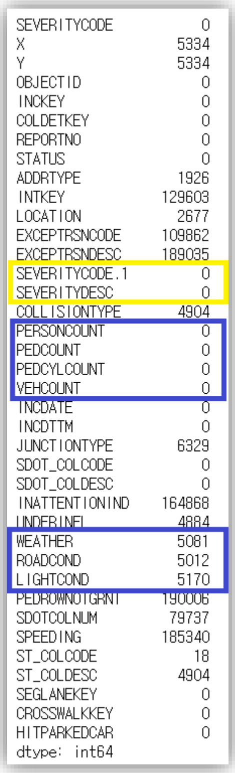
In this project, shared Collisions-All year data set provided by SPD and recorded by Traffic Records which includes all types of collisions displayed at the intersection or mid-block of a segment with timeframe: 2004 to Present.

You can find the Example Dataset by Clicking here. You can also find the Metadata by Clicking here

The first column colored in yellow is the labeled data. The remaining columns have different types of attributes. The label for the data set is severity, which describes the fatality of an accident and it is unbalanced labels. To avoid biased ML model, it needs balance the data.

2.2 Data cleaning

Download CSV file was read into a table as DataFrame. To evaluate attributes to use and quality of data in respective attribute, I calculated the number of null values in columns and value_counts to see category of value and proportion of each value group under that attribute. From the result of DataFrame.isnull().sum(), I selected attributes in blue square in Pic A. The count attributes were chose for visualized data exploration along with Weather, Road and Light condition. Weather, Road and Light condition attributes were used for training and evaluating models.



```
SEVERITYCODE           0
X                   5334
Y                   5334
OBJECTID               0
INCKEY                 0
COLDETKEY              0
REPORTNO               0
STATUS                 0
ADDRTYPE            1926
INTKEY            129603
LOCATION            2677
EXCEPTRSNCODE     109862
EXCEPTRSNDESC     189035
SEVERITYCODE.1         0
SEVERITYDESC           0
COLLISIONTYPE       4904
PERSONCOUNT            0
PEDCOUNT               0
PEDCYLCOUNT            0
VEHCOUNT               0
INCDATE                0
INCDTTM                0
JUNCTIONTYPE        6329
SDOT_COLCODE           0
SDOT_COLDESC           0
INATTENTIONIND    164868
UNDERINFL           4884
WEATHER             5081
ROADCOND            5012
LIGHTCOND           5170
PEDROWNOTGRNT     190006
SDOTCOLNUM         79737
SPEEDING          185340
ST_COLCODE            18
ST_COLDESC          4904
SEGLANEKEY             0
CROSSWALKKEY           0
HITPARKEDCAR           0
dtype: int64
```

Pic. A : Result of DataFrame.isnull().sum()

Among Count Attribute, I chose PERSONCOUNT, The total number of people involved in the collision and VEHCOUNT, The number of vehicles involved in the collision because value in the rest two count columns (PEDCOUNT, PEDCYLCOUNT) is zero in most cases.

To process null value in condition attributes (WEATHER, ROADCOND, LIGHTCOND), I replaced null value with ‘Unknown’ as the most appropriate category in terms of what the description means.

When checking value count of label value (SEVERITYCODE or SEVERITYDESC) grouping by condition attributes (WEATHER, ROADCOND, LIGHTCOND), it showed around 70% was ‘1’ in SEVERITYCODE, ‘prop damage’ in SEVERITYDESC and 30% was ‘2’ in in SEVERITYCODE, ‘injury’ in SEVERITYDESC.

Under assumption that highly frequent accident occurring location may have some relationship with condition attributes, I reviewed highly frequent accident occurring location data set filtered with number of accidents are more that 20 (mean value 3, 75% internal point value 8) but I couldn’t find meaningful difference from no-filtered case.

I assume that some aspect from location or geology may influence on occurring accident mixed with weather, road and light condition. Actually, according to data set, there are a way more accidents were happened in relatively good situation (e.g. Clear in Weather, Dry in road condition and daylight in light condition). People are likely to put more attention to drive in bed situation, so it could have driver, pedestrian and bicycle rider be more careful to surroundings and situation.

2.3 Feature selection

After cleaning the data, there were 194,673 samples, 7 attributes and 29 features from condition attributes (11 features in Weather, 9 features in Road condition, 9 features in Light condition). Upon examining the meaning of meaning of each feature and proportion of value within in feature, some of the features were less meaningful information to analyze, for instance, value ‘Unknown’ or ‘Other’ in weather, road condition and light condition attribute, and some of features contained very low, for example, value ‘Sleet/Hail/Freezing Rain’, ‘Blowing Sand/Dirt’, ‘Severe Crosswind’, ‘Partly Cloudy’ in weather.

Summary on feature selection is elaborated in table 1. below.

Kept Features	Dropped Features	Reason for dropping
Overcast, Raining, Clear, Snowing, Fog/Smog/Smoke in Weather condition	Unknown, Others Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy	Less meaningful information Very small number of cases over total cases
Wet, Dry, Snow/Slush, Ice	Unknown, Others Sand/Mud/Dirt, Standing Water, Oil,	Less meaningful information Very small number of cases over total cases
Daylight, Dark-Street Light On, Dark-No Street Lights, Dusk, Dawn, Dark-Street	Unknown, Other Dark-Unknown Lighting	Less meaningful information Very small number of cases

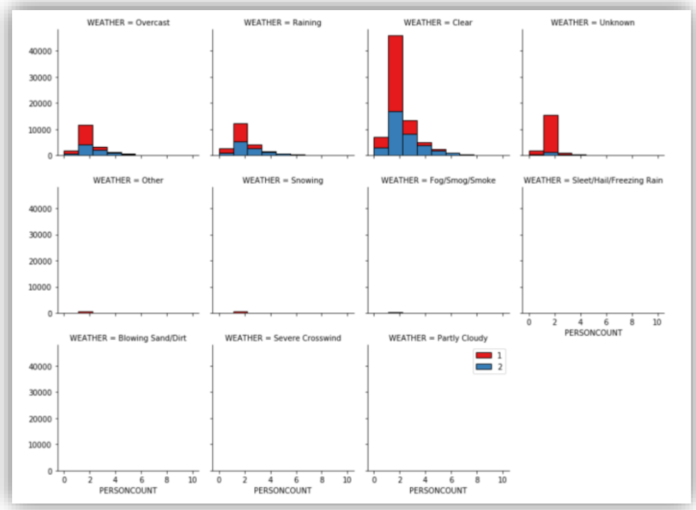
Lights Off		over total cases over total cases

Table 1. Feature selection during data cleaning

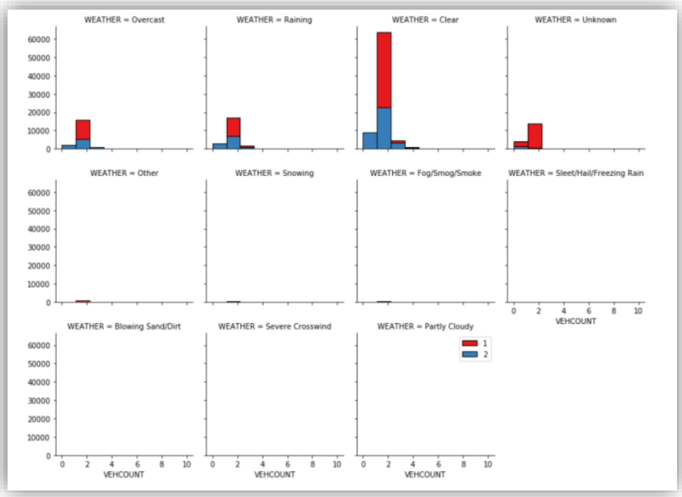
3 Exploratory Data Analysis

3.1 Relationship between Severity and Weather

Visualize histogram with PERSONCOUNT, the total number of people who involved in the collision by value in weather condition. Majority cases were happened in ‘Clear’, ‘Raining’ and ‘Overcast’, around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as ‘Unknown’, due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling. Values occurred in very low volume were excluded from feature selection as well to avoid biased model training.



Visualize histogram with VEHCOUNT, the number of vehicles involved in the collision by value in weather condition. Majority cases were happened in ‘Clear’, ‘Raining’ and ‘Overcast’, around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as ‘Unknown’, due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling. Values occurred in very low volume were excluded from feature selection as well to avoid biased model training.

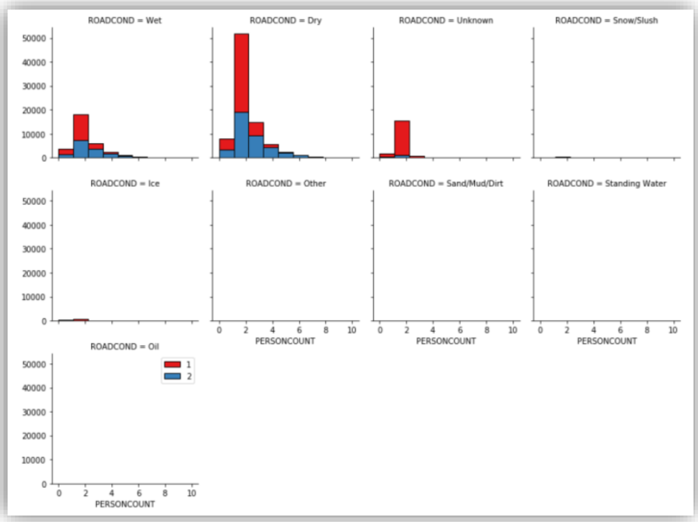


3.2 Relationship between Severity and Road Condition

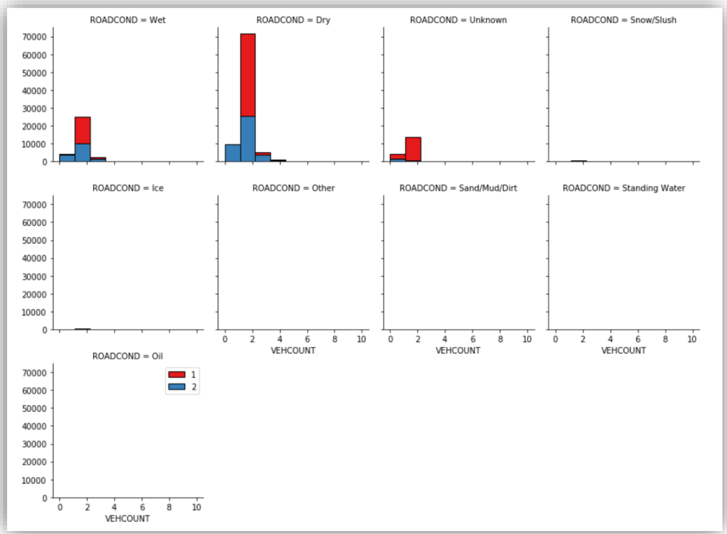
Visualize histogram with PERSONCOUNT, the total number of people who involved in the collision by value in road condition. Majority cases were happened in ‘Dry’ and ‘Wet’, around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as ‘Unknown’, due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling. Values occurred in very low volume were excluded from feature selection as well to avoid biased model training.

~~3.2~~

Formatted: Font: (Default) Times New Roman, 12 pt



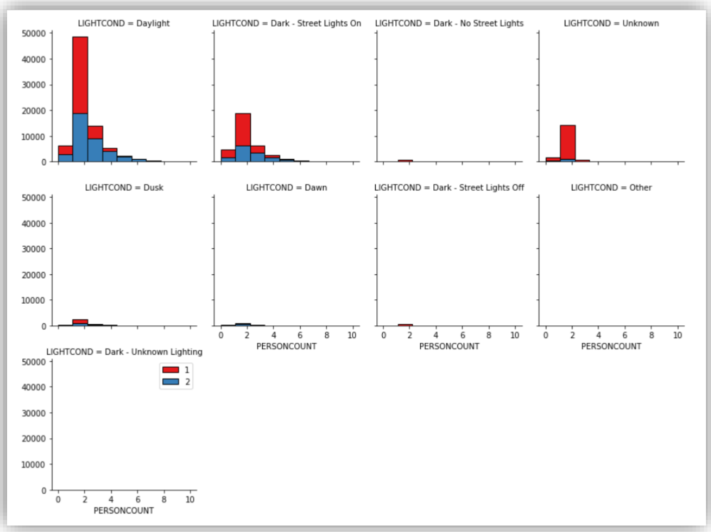
Visualize histogram with VEHCOUNT, the number of vehicles involved in the collision by value in road condition. Majority cases were happened in ‘Dry’ and ‘Wet’, around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as ‘Unknown’, due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling. Values occurred in very low volume were excluded from feature selection as well to avoid biased model training.



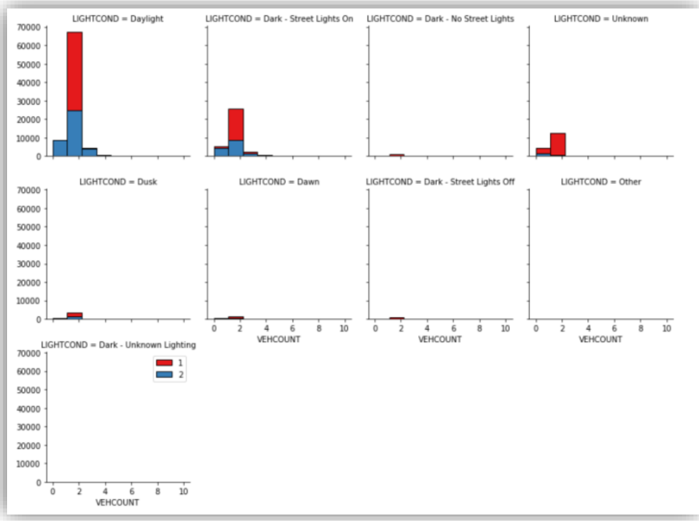
3.3 Relationship between Severity and Light Condition

Visualize histogram with PERSONCOUNT, the total number of people who involved in the collision by value in light condition. Majority cases were happened in 'Daylights' and 'Dark-Street Lights On', around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as 'Unknown', due to ambiguity of interpretation, 'Unknown' and 'Other' were excluded from feature selection for modeling. Values occurred in very low volume were excluded from feature selection as well to avoid biased model training.

Formatted: Font: (Default) Times New Roman, 12 pt

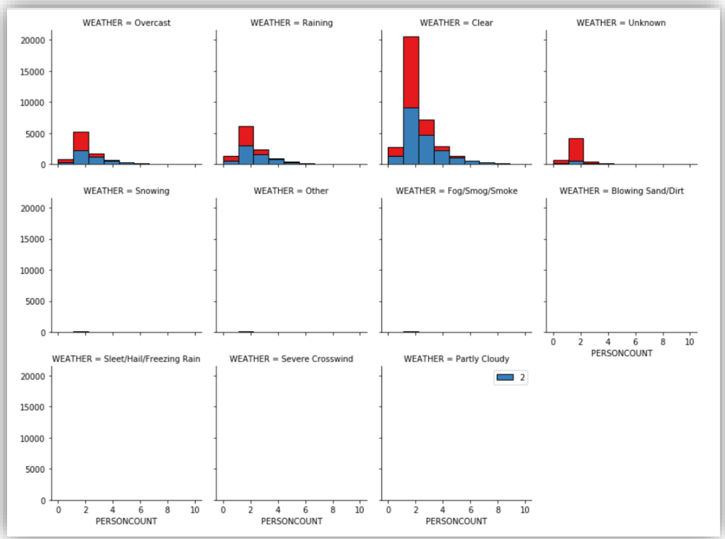


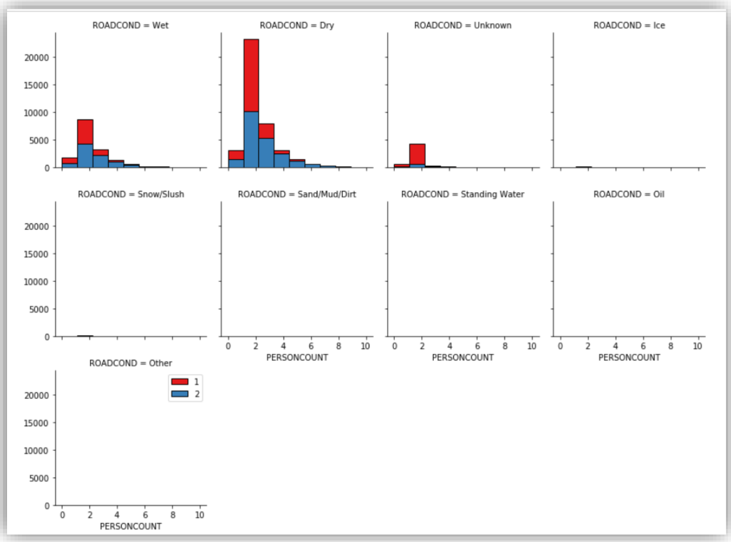
Visualize histogram with VEHCOUNT, the number of vehicles involved in the collision by value in light condition. Majority cases were happened in ‘Daylights’ and ‘Dark-Street Lights On, around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as ‘Unknown’, due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling. Values occurred in very low volume were excluded from feature selection as well to avoid biased model training.

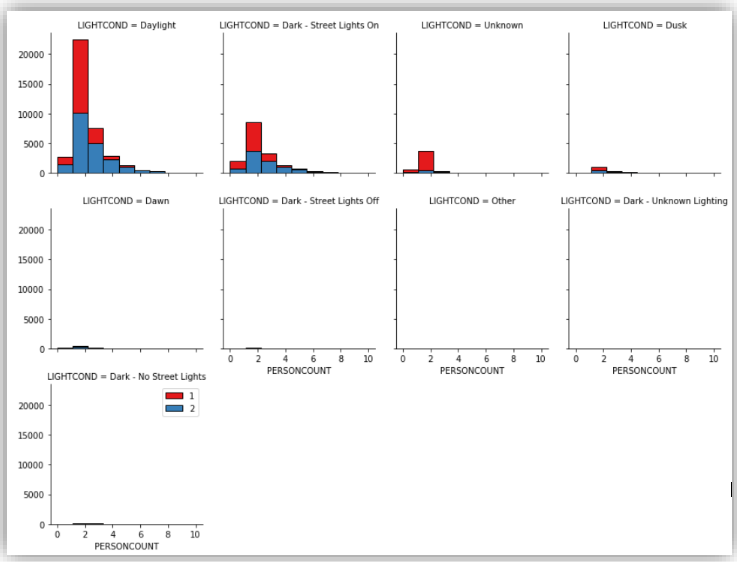


3.4 Check Impact of Frequent accident occurring location

To see any different dependency in highly frequent accident occurring location, Dataset was filtered with the number of accidents in the location > 20 (mean 3, 75% 8) and used for visualization with same condition. It showed no significant difference from no-filtered case in weather, road and light condition.







4 Predicting Modeling

4.1 Classification models

4.1.1 Applying standard algorithms and problems