Predicting Severity of accident from Weather, Road and Light Condition

David Oh

21~~0~~ September 2020

1 Introduction

Formatted: Font: Bold

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

Formatted: Font: Bold

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

Formatted: Font: Bold

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

Formatted: Font: Bold

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 chosed 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

Formatted: Font: Bold

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

Looking into the Weather attribute data, some of the Weather condition shows relatively small number of cases happened comparing with total number of samples.

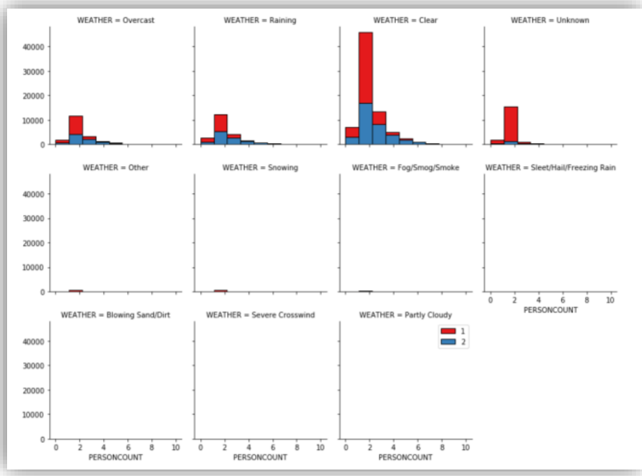
```
Clear                       111135
Raining                      33145
Overcast                     27714
Unknown                      20172
Snowing                        907
Other                          832
Fog/Smog/Smoke                 569
Sleet/Hail/Freezing Rain       113
Blowing Sand/Dirt               56
Severe Crosswind                25
Partly Cloudy                    5
Name: WEATHER, dtype: int64
```

Hypothesis here is certain weather condition may relate with car accident and severity because of driver inattention, unclear sight ahead etc. To explore this, Visualizing histogram with PERSONCOUNT, the total number of people who involved in the collision by value in weather condition. As shown below, 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. There were only two severity given data set (1-prop damage, 2-Injury). It is good to explore co-relation between more serious severity (2b or 3) and weather condition in next phase of experiment.

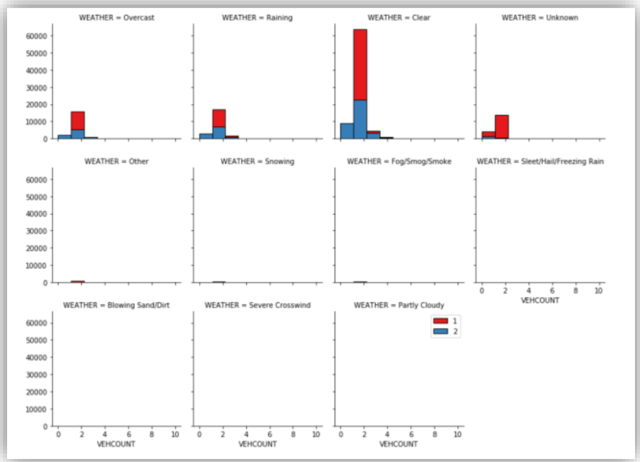
```
WEATHER                   SEVERITYDESC
Blowing Sand/Dirt         Property Damage Only Collision    0.732143
                          Injury Collision                  0.267857
Clear                     Property Damage Only Collision    0.677509
                          Injury Collision                  0.322491
Fog/Smog/Smoke            Property Damage Only Collision    0.671353
                          Injury Collision                  0.328647
Other                     Property Damage Only Collision    0.860577
                          Injury Collision                  0.139423
Overcast                  Property Damage Only Collision    0.684456
                          Injury Collision                  0.315544
Partly Cloudy             Injury Collision                  0.600000
                          Property Damage Only Collision    0.400000
Raining                   Property Damage Only Collision    0.662815
                          Injury Collision                  0.337185
Severe Crosswind          Property Damage Only Collision    0.720000
                          Injury Collision                  0.280000
Sleet/Hail/Freezing Rain  Property Damage Only Collision    0.752212
                          Injury Collision                  0.247788
Snowing                   Property Damage Only Collision    0.811466
                          Injury Collision                  0.188534
Unknown                   Property Damage Only Collision    0.905810
                          Injury Collision                  0.094190
Name: SEVERITYDESC, dtype: float64
```

Formatted: Font: Bold

Formatted: Indent: Left: 0 cm



Visualizing 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

Looking into the Road condition attribute data, some of the road condition shows relatively small number of cases happened comparing with total number of samples.

```
Dry               124510
Wet                47474
Unknown            20090
Ice                 1209
Snow/Slush          1004
Other                132
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
Name: ROADCOND, dtype: int64
```

Hypothesis here is certain road condition may relate with car accident and severity because of unexpected slippery on the road or else. To explore this, visualizing 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. There were only two severity given data set (1-prop damage, 2-Injury). It is good to explore co-relation between more serious severity (2b or 3) and road condition in next phase of experiment.

```
ROADCOND        SEVERITYDESC
Dry             Property Damage Only Collision    0.678227
                Injury Collision                  0.321773
Ice             Property Damage Only Collision    0.774194
                Injury Collision                  0.225806
Oil             Property Damage Only Collision    0.625000
                Injury Collision                  0.375000
Other           Property Damage Only Collision    0.674242
                Injury Collision                  0.325758
Sand/Mud/Dirt   Property Damage Only Collision    0.693333
                Injury Collision                  0.306667
Snow/Slush      Property Damage Only Collision    0.833665
                Injury Collision                  0.166335
Standing Water  Property Damage Only Collision    0.739130
                Injury Collision                  0.260870
Unknown         Property Damage Only Collision    0.909955
                Injury Collision                  0.090045
Wet             Property Damage Only Collision    0.668134
                Injury Collision                  0.331866
Name: SEVERITYDESC, dtype: float64
```

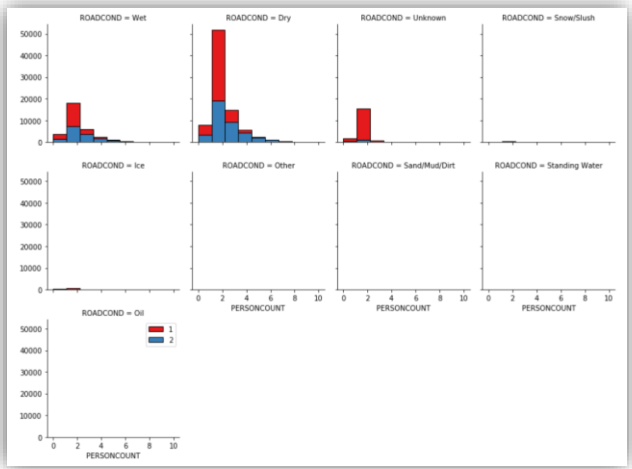
~~3.2~~

Formatted: Font: Bold

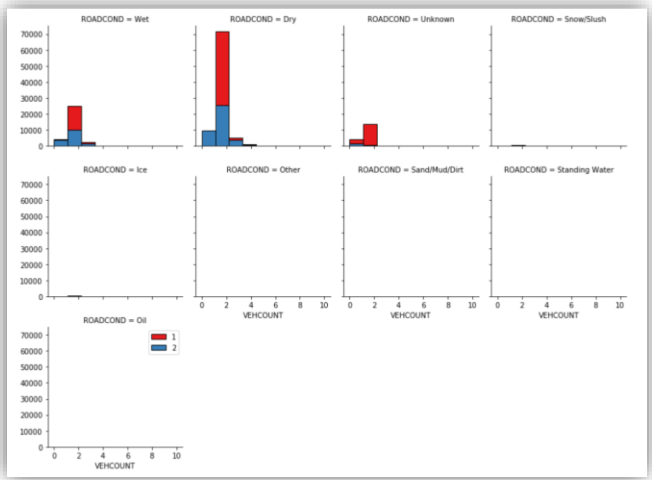
Formatted: Indent: Left: 0.75 cm, No bullets or numbering

Formatted: Indent: Left: 0.75 cm, No bullets or numbering

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

Looking into the Light condition attribute data, some of the light condition shows relatively small number of cases happened comparing with total number of samples.

```
Daylight                    116137
Dark - Street Lights On      48507
Unknown                      18643
Dusk                          5902
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Other                          235
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64
```

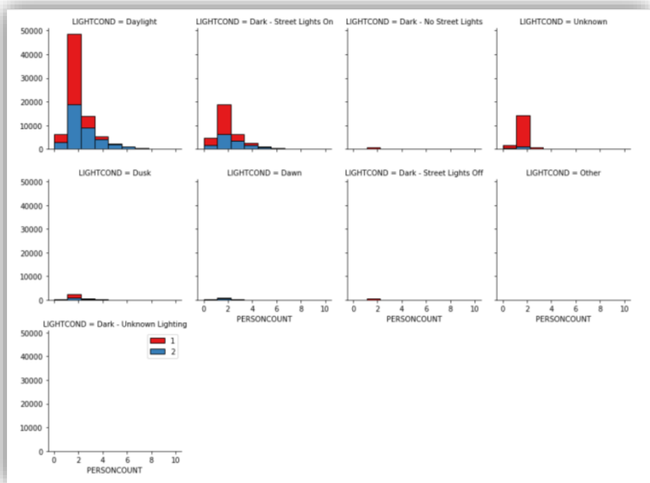
Hypothesis here is certain light condition may relate with car accident and severity because of blur sight ahead or too dark etc.. To explore this, visualizing histogram with PERSONCOUNT, the total number of people who involved in the collision by value in light condition. Majority cases were happened 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. There were only two severity given data set (1-prop damage, 2-Injury). It is good to explore co-relation between more serious severity (2b or 3) and light condition in next phase of experiment.

```
LIGHTCOND                SEVERITYDESC
Dark - No Street Lights  Property Damage Only Collision    0.782694
                         Injury Collision                  0.217306
Dark - Street Lights Off Property Damage Only Collision    0.736447
                         Injury Collision                  0.263553
Dark - Street Lights On  Property Damage Only Collision    0.701589
                         Injury Collision                  0.298411
Dark - Unknown Lighting  Property Damage Only Collision    0.636364
                         Injury Collision                  0.363636
Dawn                     Property Damage Only Collision    0.670663
                         Injury Collision                  0.329337
Daylight                 Property Damage Only Collision    0.668116
                         Injury Collision                  0.331884
Dusk                     Property Damage Only Collision    0.670620
                         Injury Collision                  0.329380
Other                    Property Damage Only Collision    0.778723
                         Injury Collision                  0.221277
Unknown                  Property Damage Only Collision    0.909081
                         Injury Collision                  0.090919
Name: SEVERITYDESC, dtype: float64
```

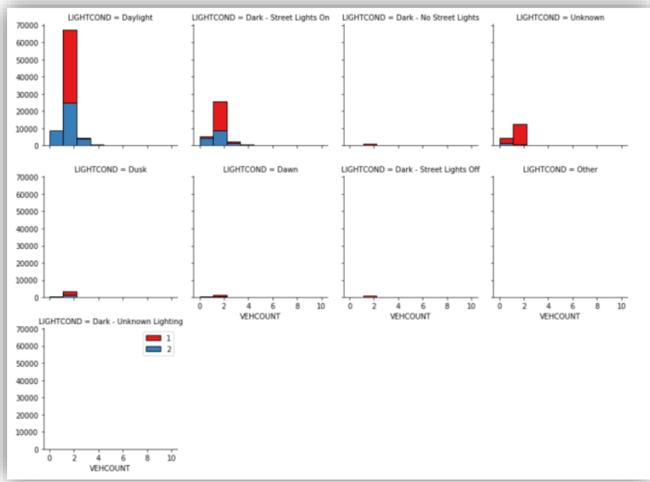
Formatted: Font: Bold

Formatted: Indent: Left: 0.75 cm, No bullets or numbering

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

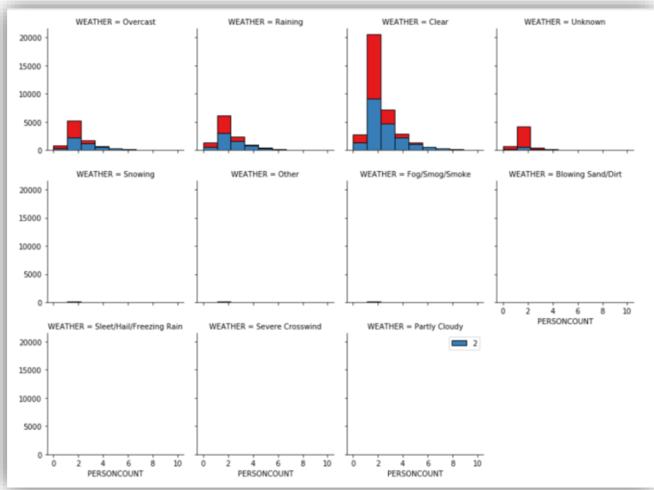


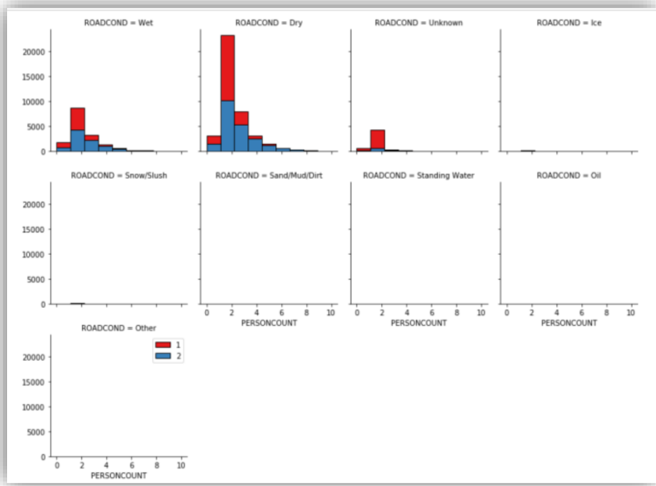
Visualizing histogram with VEHCOUNT, the number of vehicles involved in the collision by value in light condition. Majority cases were happened in 'Daylights' and 'Dark-Street Lights On, around 70% Severity code 1, 30% Severity code 2. Though a number of cases were registered as 'Unknown', due to ambiguity 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.



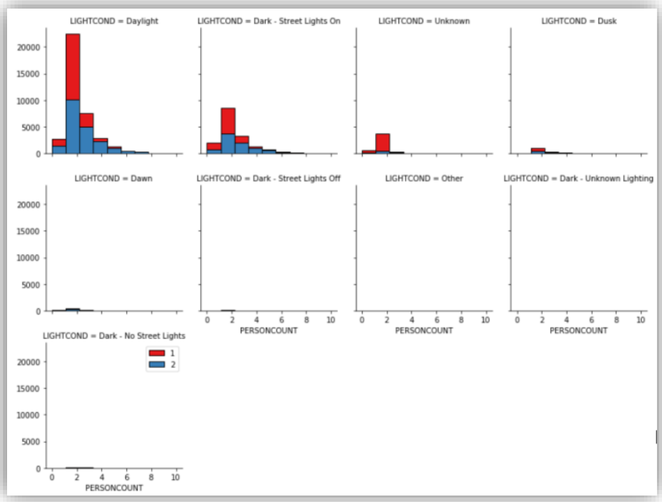
3.4 Check Impact of Frequent accident occurring location

Hypothesis here is there may be co-relation between frequently accident occurring location and condition attributes less or more. 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 turned out that no significant difference from no-filtered case in weather, road and light condition.





Formatted: Font: Bold



4 Predicting Modeling

Formatted: Font: Bold

4.1 Classification models

In light of problem, we are not going to predict particular number of cases by severity but to predict probability of severity in case of certain weather, road and light condition, so that I select predicting algorithm of classification model

4.1.1 Classification model algorithms and problems

Formatted: Font: Bold

To find out better algorithm among well-known classification model such as :

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

- k-Nearest Neighbour
- Decision Tree
- Support Vector Machine
- Logistic Regression

I spilt 70% of data set for training and 30% of data set for test purpose and conducted prediction and evaluated the accuracy classifier using the following metrics when these are applicable:

- Jaccard Index
- F1-Score

4.1.1 ● LogLoss

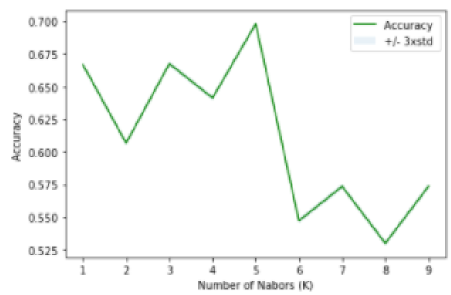
~~4.1.2 Solution to the problems~~

Formatted: Font: Bold

4.1.2 **Performance of different models**

The application of classification models was straightforward. I divided the samples into two set (70% of sample for training and 30% of sample for test). I ran each algorithm with multiple options such as the number of nearest neighbors in KNN, Maximum depth in Decision tree, different kernel functions in SVM and numerical optimizer in Logistic regression for the most accuracy within algorithm and compared the accuracy of each model and duration of execution as well.

In K Nearest Neighbor model, it showed the best accuracy when K is 5. It takes about an hour to run with range of K from 1 to 10. Roughly it takes about 8 min to complete one round.



The best accuracy was with 0.6981267764802576 with k= 5

In Decision Tree, it showed the best accuracy when Max depth is 6. Execution took a few min to complete.

The best accuracy was with 0.7034690592787918 with depth = 6

In Support Vector Machine model, it showed the best accuracy when kernel function was Linear but to apply for Non-linear problem, chose RBF. Execution took relatively longer than the rest, it took 20 ~ 25 min to complete with one function.

Kernel Function	Linear	Polynomial	Radial Basis Function	Sigmoid
Accuracy	0.70345194	0.70328071	0.70331495	0.63484127

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman

Formatted Table

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: HTML Preformatted, Automatically adjust right indent when grid is defined, Allow text to wrap in the middle of a word, Font Alignment: Baseline, Pattern: Clear (White)

In Logistic Regression model, it showed the same accuracy regardless of optimization functions, so I chose liblinear which is widely used.

Numerical Optimizer	Newton-cg	lbfgs	liblinear	sag	saga
Accuracy	0.70345194	0.70345194	0.70345194	0.70345194	0.70345194

With the best option for each model, I evaluated the accuracy with metrics like Jaccard Index, F1-Score and LogLoss if applicable. As shown table 2. Decision Tree and Logistic Regression showed the best accuracy with very small difference. I chose Logistic Regression with logarithmic loss as more impactful metric here, because the result would probably be presented with probability for each class rather than just the most likely class.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.698127	0.586849	NA
Decision Tree	0.703469	0.581031	NA
SVM	0.703315	0.580988	NA
LogisticRegression	0.703452	0.580990	0.590949

Table 2. Evaluation metrics report

~~4.1.3~~

5 Conclusions

In this study, I analyzed the relationship between Car accident severity and condition attributes (Weather, Road condition and Light condition). Accident Severity could be predicted by certain condition of weather, road and light with 59% probability by class based on logistic regression model that I built. This model can be useful in identifying accident severity and helping police station and hospital preparing accordingly.

~~5~~

6 Future Directions

~~7~~ I was able to build model to predict accident severity with 59% probability and about 70% accuracy in binary classification problem (severity 1 or 2). However, there will be more complex to apply in real situation for multiple classification (severity 1,2,2b,3 or else) and this model could be more improved on analyzing relationship among driver attention/inattention, geometry condition impact by capturing pattern of accident collision code. Models in this study mainly focused on environmental conditions (weather, road and light). However, accident occurring pattern in time and date with environmental condition might contribute more insightful prediction to prevent by alerting to driver at particular period and location.

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (맑은 고딕), 12 pt, Font color: Auto

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

Formatted: Normal, Don't adjust right indent when grid is defined, No bullets or numbering

Formatted: Font: Bold

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

Formatted: Font: (Default) Times New Roman

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

Formatted: Normal, Don't adjust right indent when grid is defined, No bullets or numbering

Formatted: Font: Bold

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

Formatted: Font: (Default) Times New Roman

Formatted: Normal, Don't adjust right indent when grid is defined, No bullets or numbering

Formatted: Font: (Asian) +Body Asian (맑은 고딕)

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