Analysis of vehicle collision data and predicting the severity of collisions based on historical data using a machine learning classifier model

Introduction/Business Statement

Vehicle collision is an important issue affecting several people all over the world. A lot of emphasis is placed on good road signs and training of vehicle operators but collisions still happen. In this data science project, data has been provided which can help us analyze the severity of collisions and what factors can be responsible for the collisions. It contains several columns of data which includes defining the place of collision (Alley/Intersection), collision type, junction type, weather, road condition, light condition etc. along with a column of labels for each row representing a certain class of severity. The motor vehicles department can use the analyzed data in some of the ways described below:

1) Analyse the data for each of the features which contribute to the collision or its severity

2) Create a predictive tool which is a machine learning model based on the data provided and predict future severity of collisions

3) Improve upon existing conditions(such as road signs) for each of these areas which show high severity of collisions

4) Provide vehicle operator training to individuals based on the analysis of the data

5) Provide timely emergency services to areas which have a higher rate of severe collisions

Business Problem:

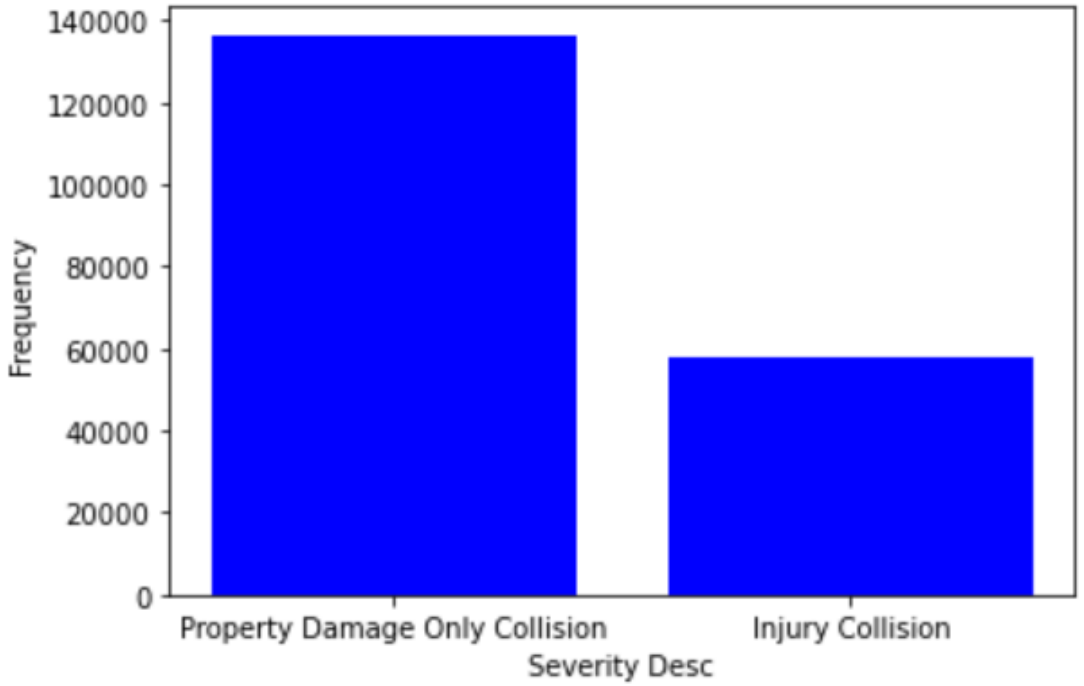
Analysis of vehicle collision data and predicting the severity of collisions based on historical data using a machine learning classifier model. The outcomes of the study will refine the motor vehicles department strategy to operate vehicles, provide better road conditions, improve vehicle operator knowledge/training and respond quicker to emergency situations.

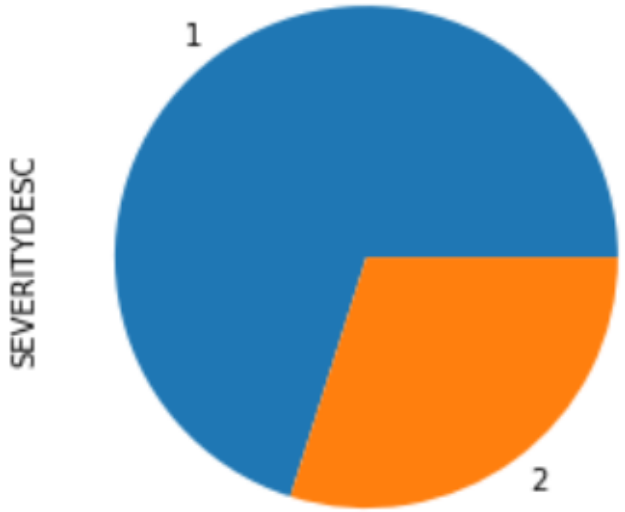
Description of Data

- The data frame contains collision data with 194673 rows and 38 columns
- The columns are listed here

'SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'

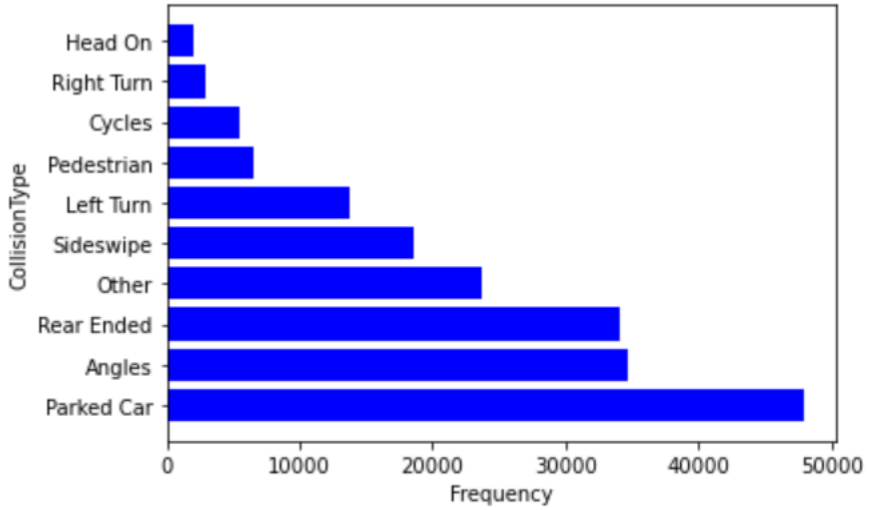
Severity of Collisions

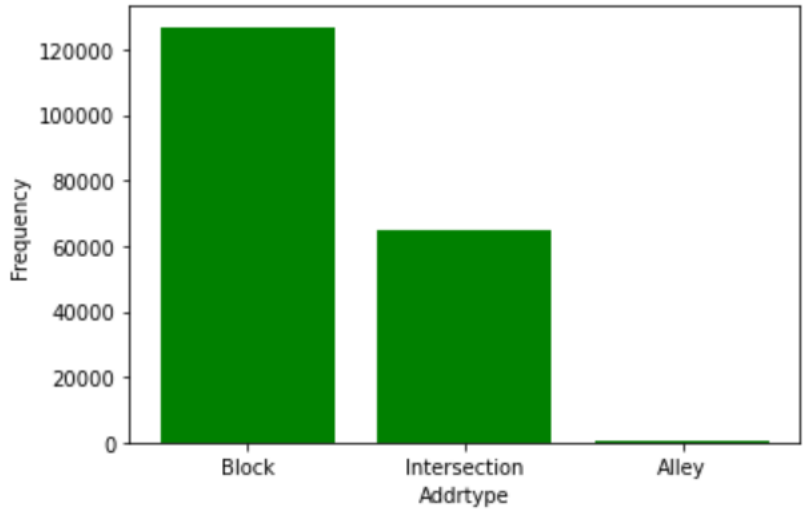


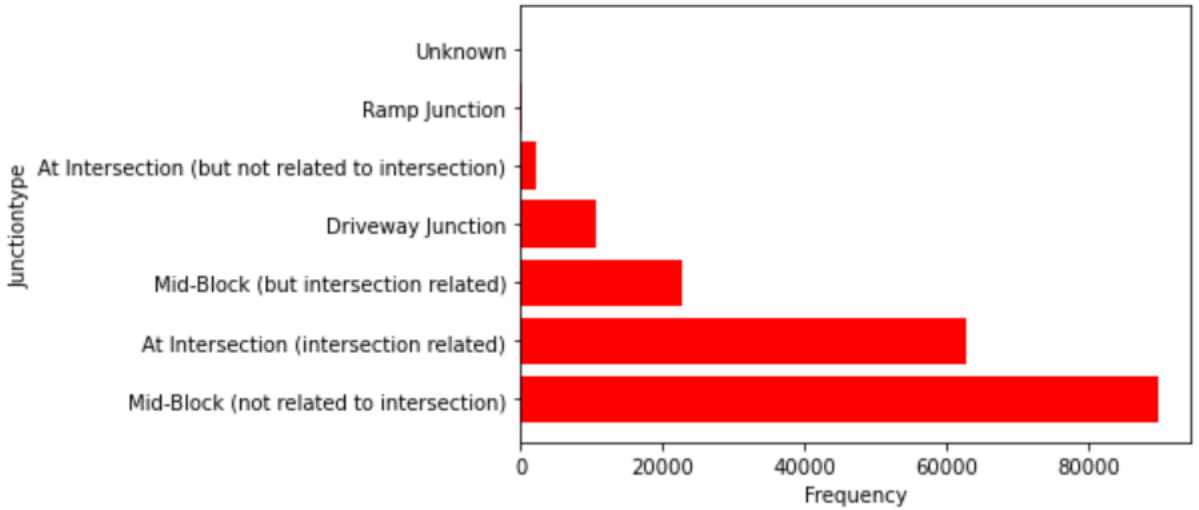


- Severity of collisions is higher for label 1 which involves property damage only collision
- Label 2 or Injury collision is much more severe has lesser frequency of occurrence but still significant

Exploratory Data Analysis – Collisiontype, Addrtype & Junctiontype

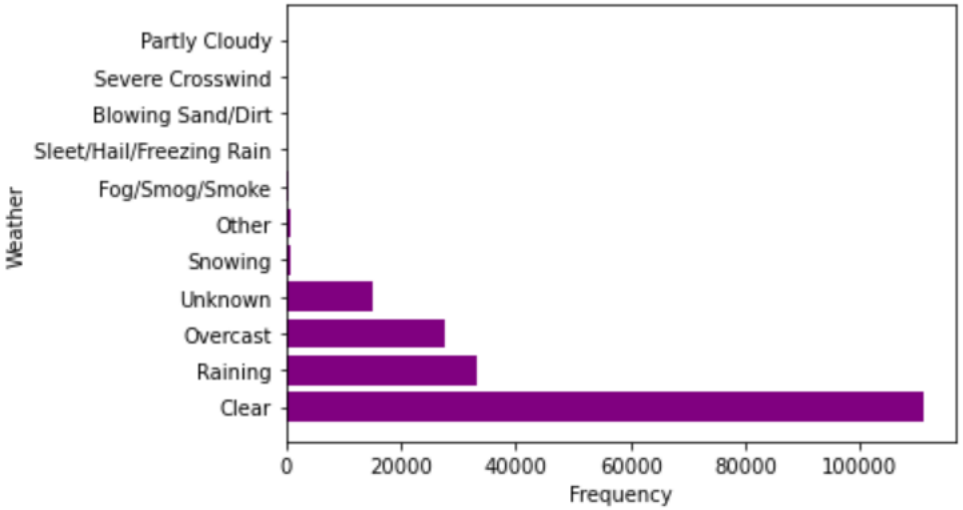


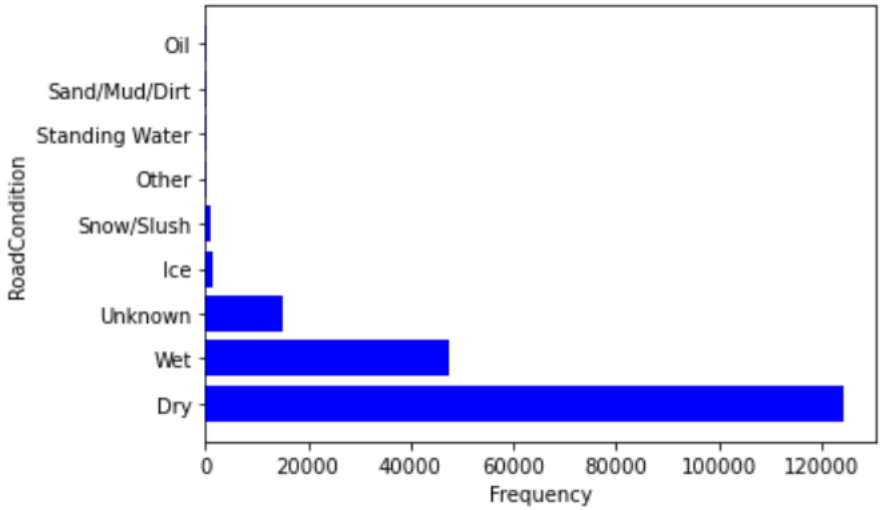


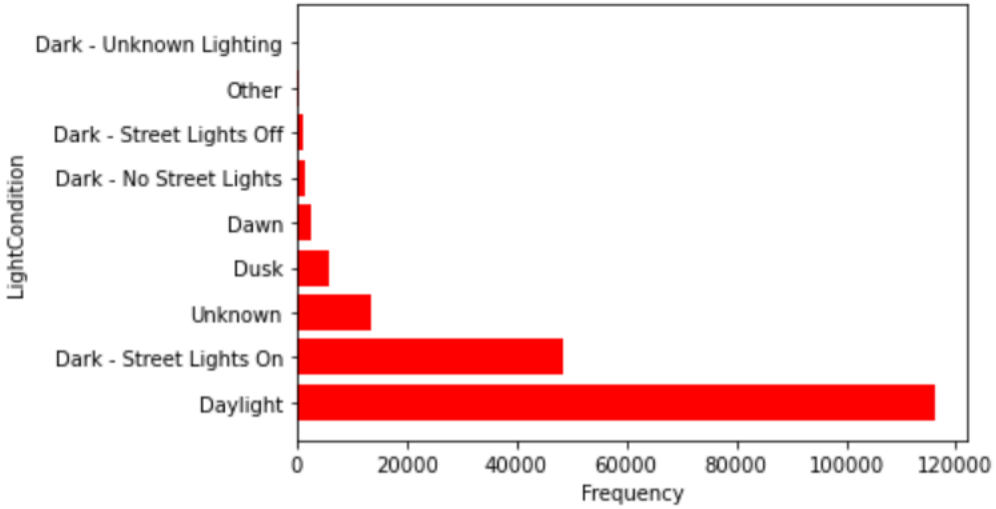


- **Block** has the highest number of collisions
- Mid-Block(not related to intersection) has the highest number of incidents in Junctiontype
- Highest number of collisions are with a parked car followed by Angles and Rear ended

Exploratory Data Analysis – Weather, LightCondition and Roadcondition



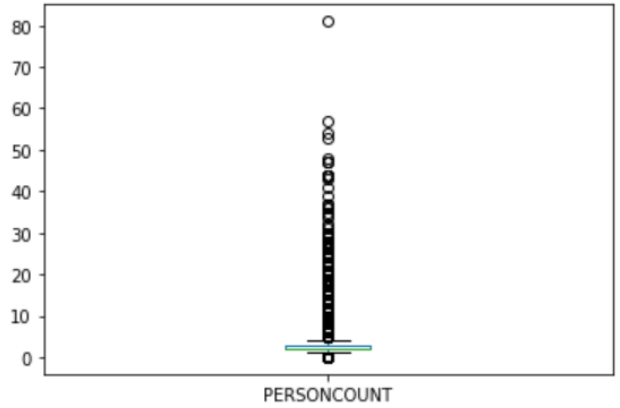




- Clear days have had most number of collisions
- Dry road condition also has had the most number of collisions
- Number of collisions in Daylight far exceeds the number on Dark –Street lights on; Also indicative of probably more number of vehicles on the road in Daylight

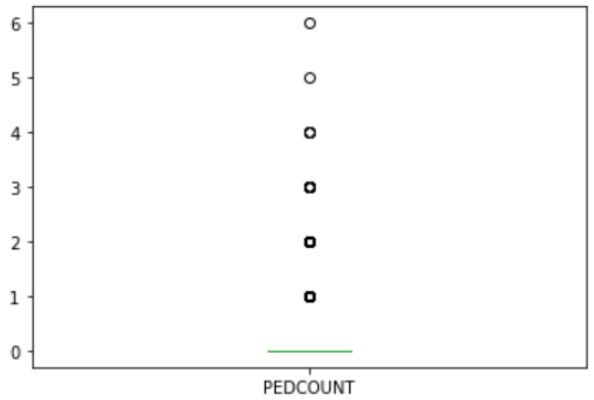
Data Exploration: Continuous variables & Statistics

```
count    194673.000000
mean          2.444427
std           1.345929
min           0.000000
25%           2.000000
50%           2.000000
75%           3.000000
max          81.000000
Name: PERSONCOUNT, dtype: float64
```



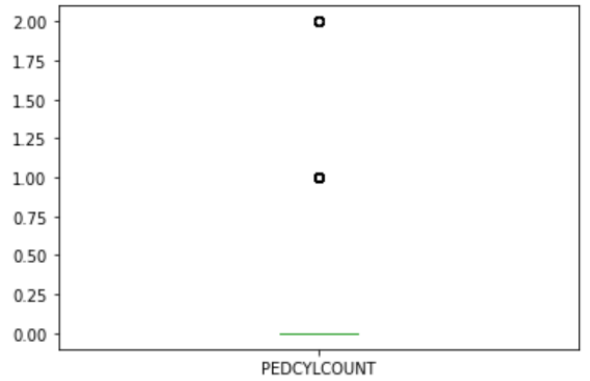
Personcount

```
count    194673.000000
mean          0.037139
std           0.198150
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           6.000000
Name: PEDCOUNT, dtype: float64
```



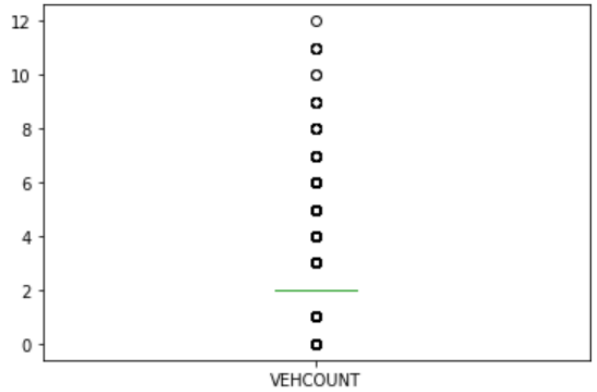
Pedcount

```
count    194673.000000
mean          0.028391
std           0.167413
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           2.000000
Name: PEDCYLCOUNT, dtype: float64
```



Pedcylcount

```
count    194673.000000
mean          1.920780
std           0.631047
min           0.000000
25%           2.000000
50%           2.000000
75%           2.000000
max          12.000000
Name: VEHCOUNT, dtype: float64
```



Vehcount

- We have box plots above for the variables of Person count, Ped count, Pedcyl count & Veh count
- Each of the variables also have their statistics above their box plots

Methodology: Data Cleaning and Arranging

- This is a classification problem with a labeled set of data. The data will be standardized and split into training and test sets based on the filtered set of features for modeling purposes.

- After cleaning the data and replacing NaN with data points or numbers using average/median values of the columns the data set will be converted into an array of the form **X**. The labelled column of severity data will be extracted as column **y**.

- All categorical variables in the array X will be replaced by numeric values and the whole array will be normalized using the preprocessing library. Upon retrieving the arrays in the desired form classification methods of KNN, Decision trees, Logistic Regression will be applied. SVM technique will not be attempted since it works best with much less data of 1000 rows. Here the data set is fairly large and may lead to poor performance of the SVM algorithm.

Filtered Columns

- The filtered column list of features for modeling are listed below:

"STATUS","ADDRTYPE","COLLISIONTYPE","PERSONCOUNT","PEDCOUNT","PEDCYLCOUNT","VEHCOUNT","JUNCTIONTYPE","SDOT_COLCODE","UNDERINFL","WEATHER","ROADCOND","LIGHTCOND","SPEEDING","ST_COLCODE","SEGLANEKEY","CROSSWALKKEY","HITPARKEDCAR"

Machine Learning Dataframe & Techniques

- 3 different methods applied for Machine learning & Classification; KNN, Logistic Regression & Decision Trees
- KNN – K nearest neighbors
    i. 8 nearest neighbors used for training and calculation
    ii. Model training accuracy of 0.76
    iii. Test set accuracy of 0.74 with similar Jaccard Index
    iv. Slow performance in both training and test data sets
- Logistic Regression
    i. Application of model using liblinear model and C=0.01
    ii. Training accuracy of 0.731
    iii. Test set accuracy of 0.73 with similar Jaccard Index
    iv. Probability of a label prediction done with test set data
    v. Fast performance in both training and prediction of the data
- Decision Trees
    I. Decision Tree model applied to the data set
    II. Training accuracy of 0.742
    III. Test set prediction accuracy of 0.74 with similar Jaccard Index
    IV. Fast performance in both training and prediction of the data
- SVM – Not applied as it is known to be slow in classification problems with more than 1000 rows

Results

- The data frame for collision data was cleaned of null values and rows of data were removed where categorical variables did not have any data. A subset of the data columns was selected for analysis which resulted in feature selection based on the signifance of the data and its contribution to the overall machine learning exercise of predicting the severity of collisions.
- From the data set we can observe that the number of collisions reported for label 1 are higher than for label 2. It is label 2 where the data suggests we have injury to the persons involved in the collision.
- Several histograms were prepared to analyze each of the data columns and showed the relative importance of each of the features and their subset. Also statistical analysis of some of the continuous variables was created including box plots.
- The data suggests that maximum number of collisions occurred with a parked car followed by Angles. Also the maximum number of collisions have occurred at a block followed by an intersection. Within a block the highest number of collisions are at mid block not related to an intersection. Upon analyzing external factors for collision the most number of collisions have occurred on Dry roads, in Daylight and on clear skies.
- Categorical variables were replaced with numeric values by using the preprocessing library. Once the array X was created with the features it was normalized such that the data skewness is removed for modeling. The final data frame X and the resulting labeled vector y were then fit to models of K nearest neighbor, Logistic regression and decision trees. Since this is a classification problem the above algorithms were implemented. The dataset however is large and hence SVM was not implemented.
- The data array was split into training and test sets which were used to train the models and then predict using the test data. The predicted data from the test set was used to calculate accuracy and Jaccard index. The trained model with KNN using 8 nearest neighbors showed the most accuracy of 0.76 for the trained model and 0.74 for the test data set. The alogrithm was however slow to train the model and make predictions.
- Logistic regression trained rather quickly and the training/test set accuracy was 0.731/0.73 respectively. The Decision tree classifier was equally efficient at training the model and its prediction of the test data. The accuracies were 0.742 (Trained model) and 0.74 (Test model) respectively.

Discussion

- After analyzing the data what we can observe is that the data suggests collisions are happening on clear days with dry conditions. Also majority of the accidents are occuring mid-blocks with parked cars followed by Angles and rear ending. The situations under which these collisions have happened seem more of driver negligence or improper signs with bad road conditions. The road conditions however do not consider data which may point towards potholes or barriers/restrictions, which should be made available for modeling bad surfaces as well. The data analysis can provide the city with insights on how to improve road conditions and installing proper signs for parking along the roads.

- Also, what we can observe is that a high proportion of accidents have occured at intersections which would imply that the drivers involved could prevent such collisions with improved training programs. The number of rear end collisions also seems to be high which would imply driver carelessness. The motor vehicles institute should consider improving their safety training programs for vehicle operators.

- The models available to us provide good accuracy with the data at hand. Since the model is trained on data based on blocks and intersections, the model can be skewed by relatively higher rate of collisions at certain intersections. The data for those intersections where the number of collissions is higher should be analyzed independently.

- Last but not the least the models should be updated once the street signs have improved and training programs have been put in place. With that the motor vehicles departments can reduce the number of such incidents and improve accuracy for future predictions.