Collision severity factors analysis and its prediction by supervised machine learning models

Abstract.

In this study, we apply the analytical method to investigate the most affecting factors on collision severity in New York City. Regarding these factors, we predict collision severity in New York City from January 1, 2001 to mid-October 2020 using machine learning models based supervised approaches such as Random Forest and Logistic Regression. The data comes from the California Highway Patrol and covers collisions from January 1st, 2001 until mid October 2020. The prepared dataset include 74 features. Among these variables, columns with correlation more than 0.25 has been chosen to predict collision severity. According to correlation analysis, seven factors are identified as the main factors affected on collision severity, including secondary road, primary road, victim safety equipment1, beat number, victim role, type of collision and victim seating position. In order to more analysis the effective principal variables, feature selection in terms have done. This plays a very important role in identifying hidden variables or factors through existed variables. Next purpose of this study is prediction of collision severity using machine learning models regarding seven above mentioned features. Supervised Algorithms such as random forest and logistic regression Models have been applyed. The highest accuracy of the model is the best classifier. In this comparative study, the accuracy on train and test data in logistic regression model is 68%, while this score for random forest model is 99%. Therefore, random forest model has higher predictive power here with 31% increase in prediction of collision severity. Besides, the good performance and validation of machine learning models is proved through confusion matrix. Consequently, random forest model indicated the most principal variable in the most important seven features is “victim role”. We also proposed two models based on ensemble learning to improve the accuracy of each learning algorithm.

Keywords: Random forest, Logistic regression, Confusion matrix, Feature selection, Ensemble learning.

2. Research Method

The present study considers supervised machine learning methos to predict “collision severity” based on dataset from the California Highway Patrol during 2001-2020. Process flow for our work is as follows:

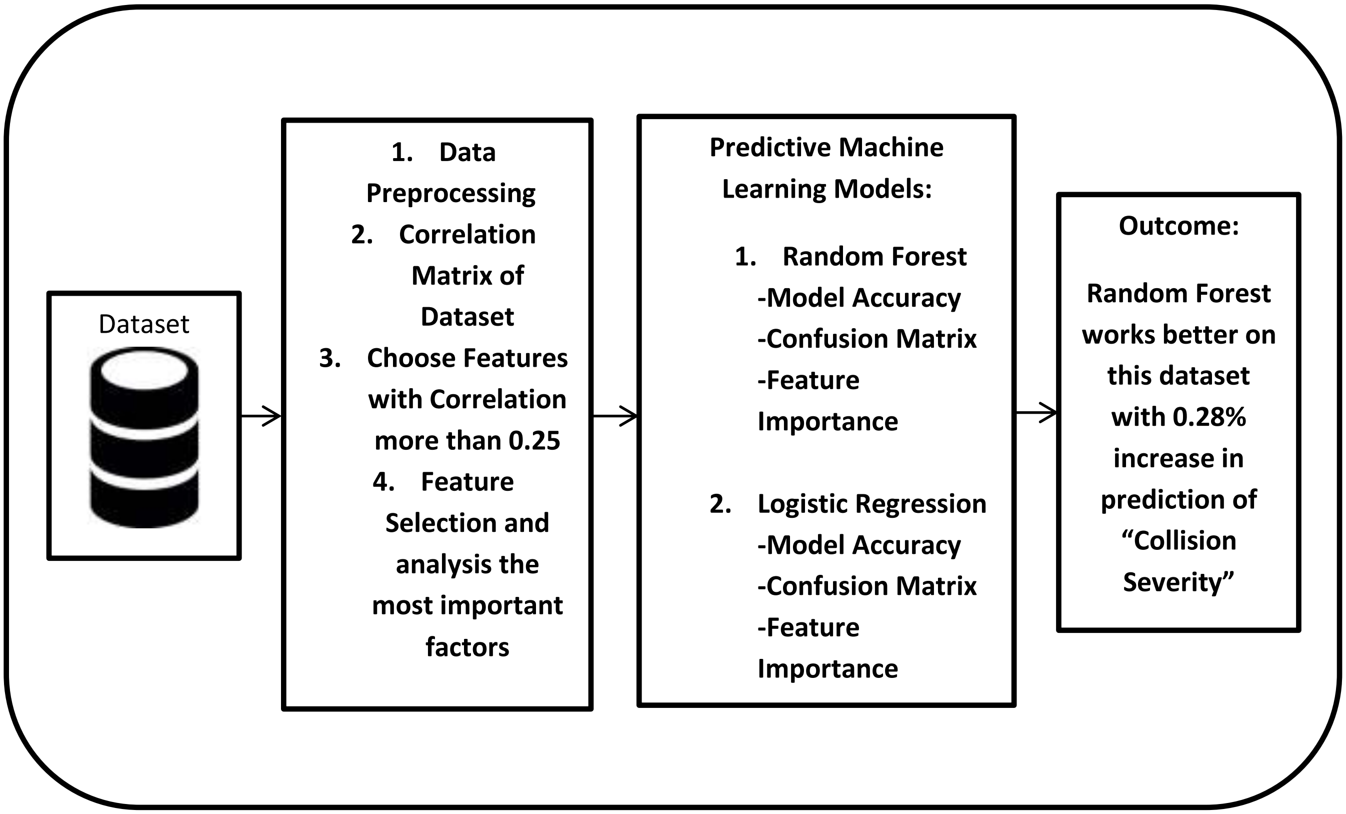
Dataset
1. Data Preprocessing
2. Correlation Matrix of Dataset
3. Choose Features with Correlation more than 0.25
4. Feature Selection and analysis the most important factors
Predictive Machine Learning Models:
1. Random Forest
-Model Accuracy
-Confusion Matrix
-Feature Importance
2. Logistic Regression
-Model Accuracy
-Confusion Matrix
-Feature Importance
Outcome:
Random Forest works better on this dataset with 0.28% increase in prediction of “Collision Severity”

Fig1. The workflow of this research

3. Results

3.1. Diagnosis and analysis of the most important factors

In studies with large number of variables, researchers try to do data exploration and reduce the number of variables and form a new structure and modified data for more practical and data analysis. This helps to find more accurate answers for our questions using machine learning models. The data source is Kaggle and the format is SQL. This data comes from the California Highway Patrol and covers collisions from January 1st, 2001 until mid-October 2020. There are three main tables:

- Collisions: Contains 74 columns of collision information including the location, time, severity and environmental factors.

- Parties: Contains 31 columns of demographic information for all parties involved as well as vehicle information, age, sex, and sobriety.

- Victims: Which contains 11 columns of information about the injuries of specific people involved in the collision.

In this work, the three main tables joined. In order to data cleaning, NAN values, Columns which have more than 50% unknown values, have been dropped. Also, Outliers, Rows with parties’ age of zero, have been dropped (32506 out of 351975). Among 116 columns, 74 columns including categorical and numeric features such Party Number, Victim Sex, Victim Age, Party Race, Vehicle Year, Vehicle Make, Party Age, Party Sex, At Fault have chosen. Among chosen columns, the most affecting features on collision severity computing their correlation coefficient (CC) have identified. As a result, CC for seven featurs, including are more than 0.25. Therefore, factor analysis is done to identify the principal variables and explain the correlation between the variables. There are four main reasons why feature selection is essential. First, to simplify the model by reducing the number of parameters, next to decrease the training time, to reduce overfilling by enhancing generalization, and to avoid the curse of dimensionality. Fortunatelly, random forest has emerged as a quite useful algorithm that can handle the feature selection issue even with a higher number of variables (Chen et al [4]). For this reason, after recognizing seven feature with correlation more than 0.25, Feature Selection for these more affected features have also done.

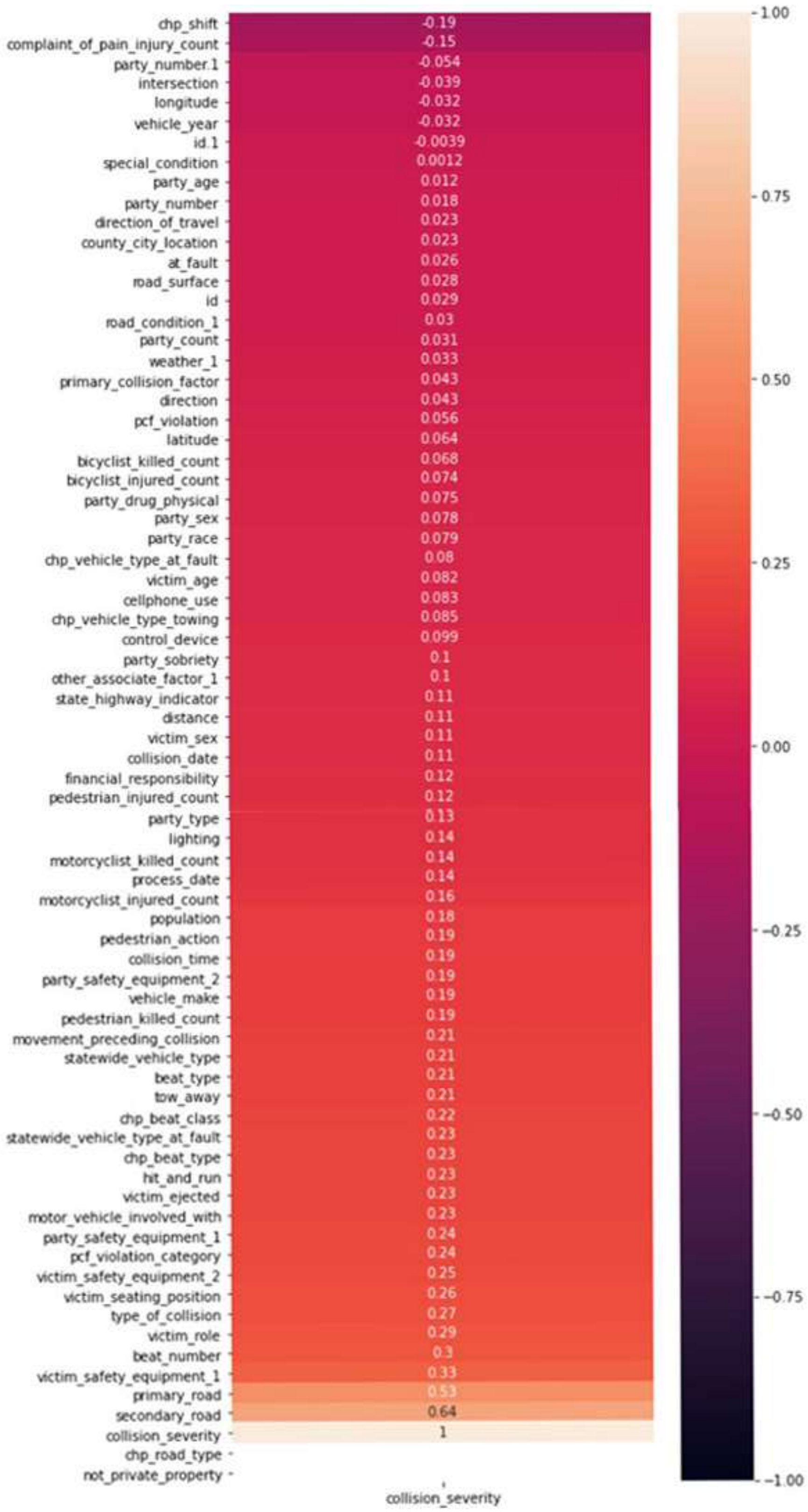

Fig2. Correlation Matrix of Dataset

Columns with correlation more than 0.25 has been chosen. In fact, variables which used in this study are secondary_road, primary_road, victim_safety_equipment1, beat_number, victim_role, type_of_collision and victim_seating_position. Figure 3 shows the situation of these features in more details.

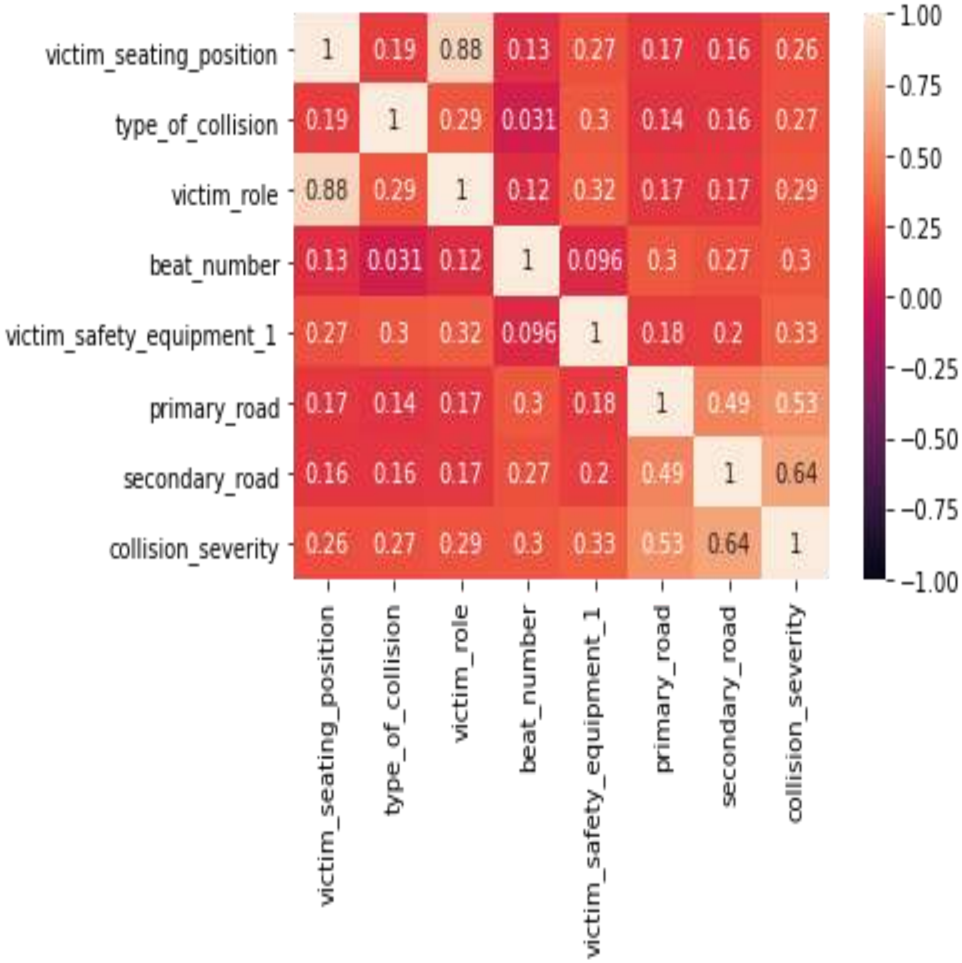

Fig3. Columns with Correlation more than 0.25

3.2. Factor Analysis

Factor analysis is a way to identify the principal variables in order to explain hidden informations in different variables of a feature. We have first describe these variables in Table 2 and Table 3. Figures 4,5,6 determine importance of variables in selected features.

Variable	Variable levels
Victim seating position	1-Driver
	2-6-Passengers
	7-Station Wagon Rear
	8-Rear Occupant of Truck or Van
	9-Position Unknown
	0-Other Occupant
	A thru Z-Bus Occupant
	--Non Stated
Type of collision	A-Head-On
	B-Sideswipe
	C-Rear End
	D-Broadside
	E-Hit Object
	F-Overturned
	G-Vehicle/Pedestrian
	H-Other
	--Non Stated
Victim role	1-Driver
	2-Passenger (includes non-operator on bicycle or any victim on/in parked vehicle)
	3-Pedestrian
	4-Byciclist
	5-Other (Single victim on/in non-motor vehicle; e.g. ridden animal, horse-drawn carriage, train or building)

	6-Non-injured party
Beat number	1-North
	2-South
	3-Mtro
	4-Valley
	5-Hill
	6-…
Victim_safty_equipment_1	A-Non in Vehicle
	B-Unknown
	C-Lap Belt Used
	D-Lap Belt Not Used
	E-Shoulder Harness Used
	F-Shoulder Harness Not Used
	G-Lap/Shoulder Harness Used
	H- Lap/Shoulder Harness Not Used
	J-Passive Restraint Used
	K-Passive Restraint Not Used
	L-Air Bag Deployed
	M-Air Bag Not Deployed
	N-Other
	P-Not Required
	Q-Child Restraint Vehicle Used
	R-Child Restraint Vehicle Not Used
	S-Child Restraint Vehicle, Use Unknown
	T- Child Restraint Vehicle, Improper Use

	U-No Child Restraint in Vehicle V-Driver, Motorcycle Helmet Not Used W-Driver, Motorcycle HelmetUsed X-Passenger, Motorcycle Helmet Not Used Y- Passenger, Motorcycle Helmet Used --Or blank-Not Stated

Table 2. Description of variables used in this study

Now we are going to clarify the most important variable level in these columns in terms of Box-Plots (See Figures 3-6). By virtue of Table 2 and Figure 4, Bus Occupant is critical component of feature “victim seating position”.

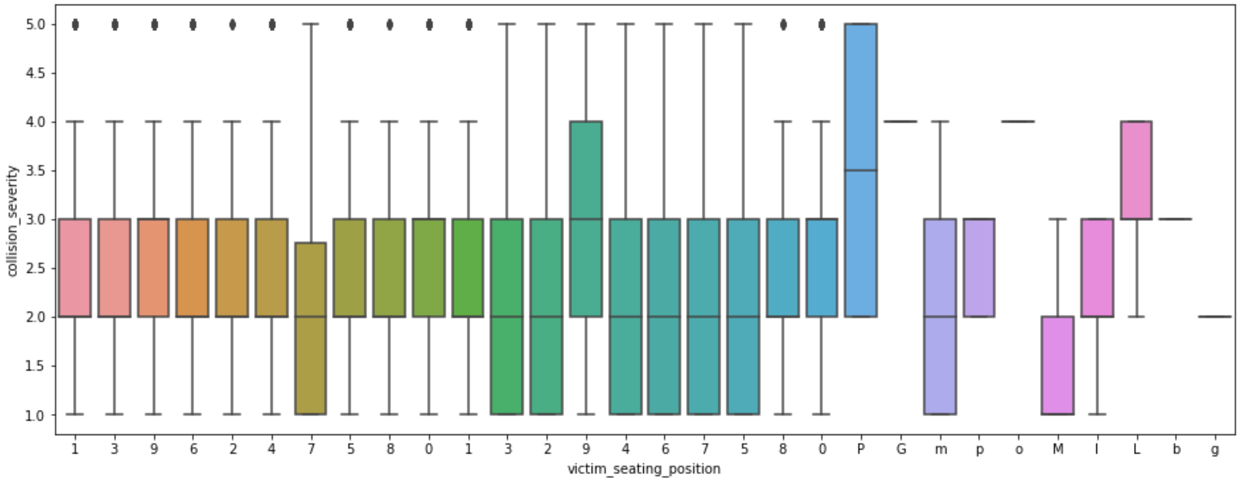

Fig4. Feature selections Victim seating position

Figure 5 shows that “pedestrain” and “overturned” are the most prominent factors for feature “type of collision”.

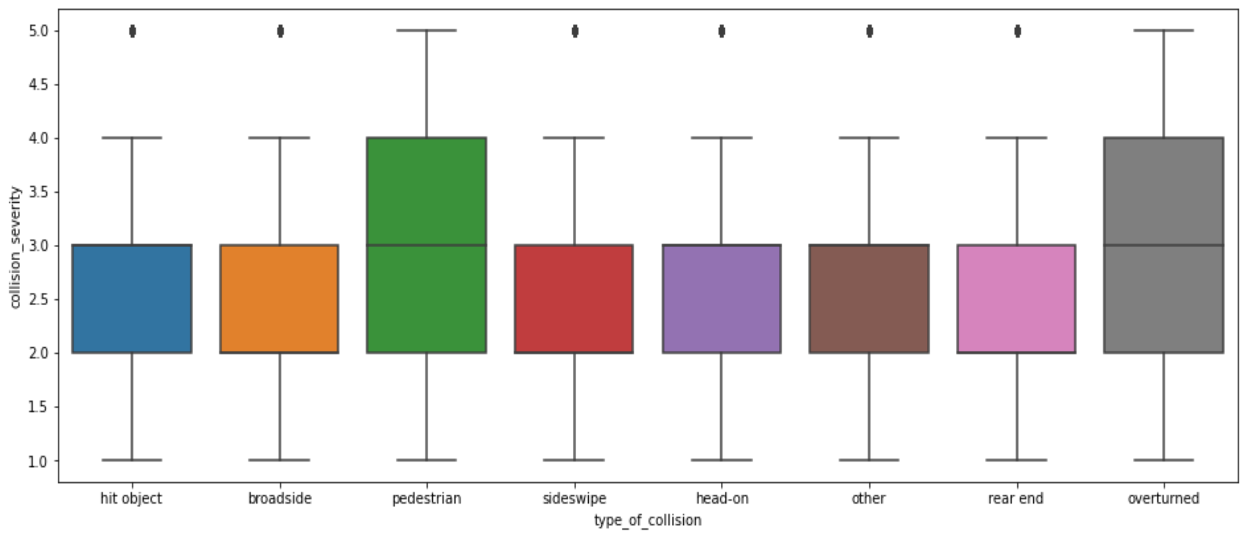

Fig5. Feature selections type of collision

Regarding of Table 2 and Figure 6, “Pedestrian” plays the most determinative role as “victim_role”.

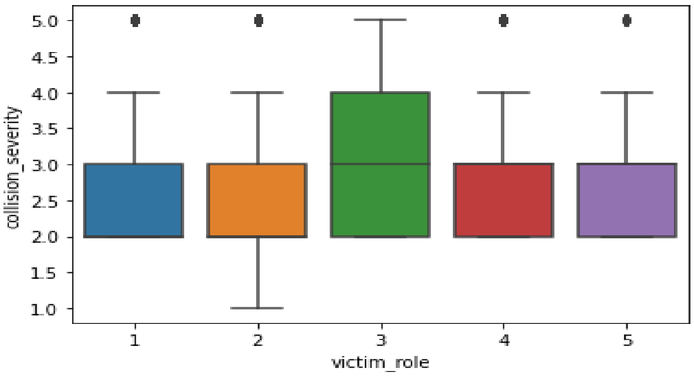

Fig6. Feature selections victim role

By view of Table 2 and Figure 7, “Lap Belt Not Used” is the most determinative element in feature “victim safety equipment1”. And also “Child Restraint Vehicle Not Used” should be pay attentioned.

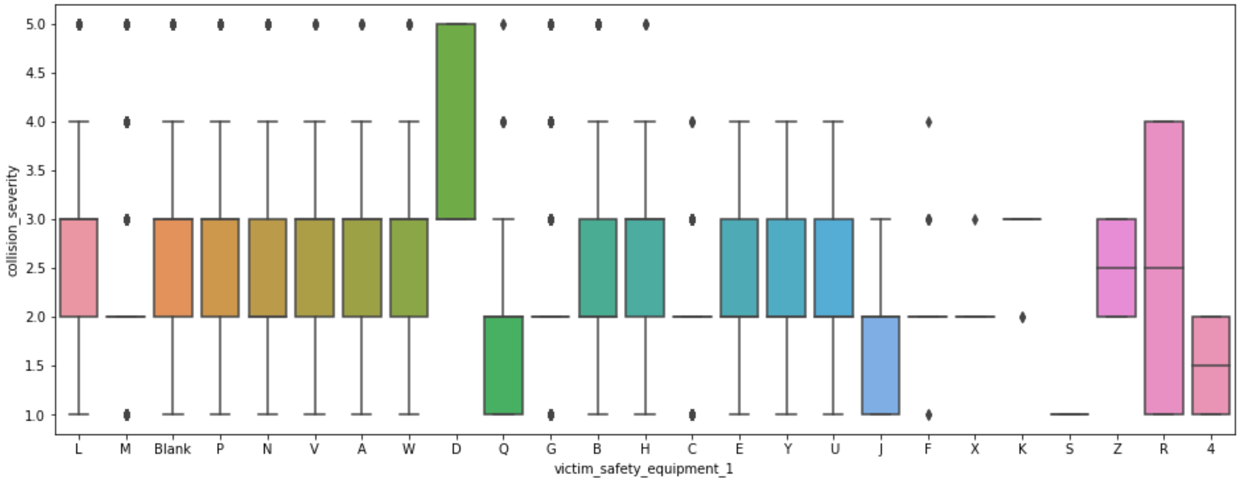

Fig7. Feature selections victim safety equipment1

Notice now that for column “primary_road”, there are so many levels such as 'NORRIS RD', 'GOSFORD RD', 'FIRST ST’,..., 'LOS ABITOS BL', 'WESTAR DR’, 'VENTURA RD FRONTAGE RD’. Also for column “secondary_road”, there is so many levels such as 'QUAIL CREEK RD', 'HARRIS RD', 'SIERRA AVE', ..., 'VANGUARD DR', 'EMERSON RD', 'GATESIDE CT’. Hence, we selected variable levels (places) in which the number of collisions happened more than 100 times.

Variable	Variable levels	Avg_severity	Count
Primary_road	INTERSTATE 5 SOUTHBOUND	3.421965	173
	SR-39 (SAN GABRIEL CANYON RD)	3.398230	113
	US-395	3.382979	141
	SR-4	3.329341	167
	GENEVA AV	3.304348	161
	…	…	…
	INTERNATIONAL BL	1.780769	260
	WALNUT AV	1.731544	149

	WILSON WY	1.710744	121
	MARCH LN	1.699605	253
	SIERRA AV	1.630695	417
Secondary_road	FRANCISQUITO AVE	4.379747	158
	LAKEWOOD BOULEVARD	3.227941	136
	ROXFORD ST	2.950820	183
	VAN NESS AV	2.926966	178
	TYLER ST	2.898734	158
	…	…	…
	OLIVE AV	1.829268	123
	BANCROFT AV	1.771930	114
	CITRUS AV	1.767123	146
	10TH ST WEST	1.729508	122
	SIERRA AV	1.526946	167

Table 3. Description of variables used in this study

For feature “Primary_road”, the value “INTERSTATE 5 SOUTHBOUND” is the most determinative value. And for feature “Secondary_road”, “FRANCISQUITO AVE” plays the most important value.

4. Machine Learning Models

In this section, we will apply two machine learning models, random forest and Logistic Regression model in order to validate and predict “Collision Severity” in NYC for the January 2001 through October 2020. First, we will look at how to build and use the random forest in coding environment Python 3 such that it predicts collision severity well. In this way, we have investigated the accuracy of model, feature importance and confusion matrix with random forest.

4.1. Results of Random Forest Model

A forest is better than one tree in making decisions. The random forest is a model made up of many Decision Trees. Rather than just simply averaging the prediction of trees (which we could call a "forest"), this model uses two key concepts that gives it the name random:

1. Random sampling of training data points when building trees.

2. Random subsets of features considered when splitting nodes.

The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest are made by averaging the predictions of each individual tree [15].

Based on [4], a decision tree with M leaves divides the feature space into M regions Rm, $1 \leq m \leq M$. For each tree, the prediction function $f(x)$ is defined as:

$$f(x) = \sum_{m=1}^{M} c_m \prod (x, R_m)$$

where M is the number of regions in feature space, Rm is a region appropriate to m; c_m is a constant suitable to m:

$$\prod (x, R_m) = \begin{cases} 1, & if\ x \in R_m \\ 0, & otherwise \end{cases}$$

The flowchat of random forest

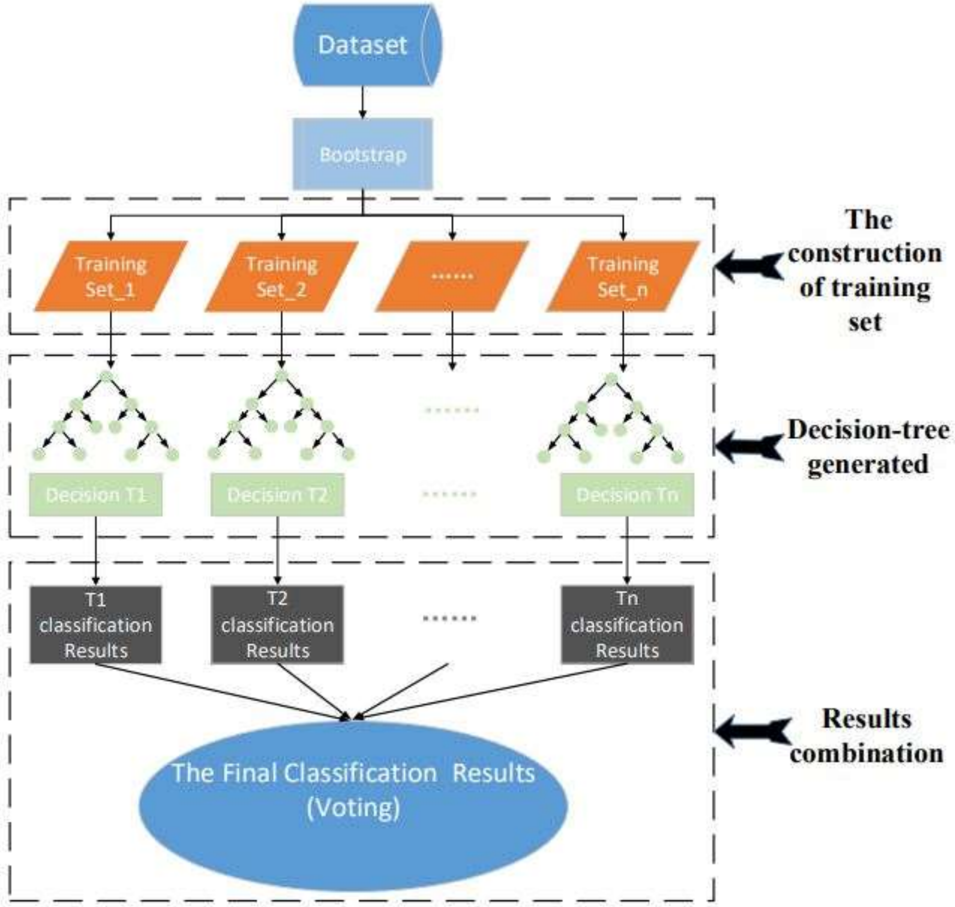

Fig8. The flowchart of Random Forest model

4.1.1. Random Forest Model Accuracy

In this section, we present how random forest model well worked on our dataset to make an accurate prediction for “collision severity”. We have trained a random forest model with 30 trees on our dataset. We limited the maximum depth of the tree to 3, otherwise it will be too large to be converted into an image. Thus, the accuracy of the predictor on the training and test data are 0.9993 and 0.9994 respectively with Mean Absolut Error 0.00094.

4.1.2. Random Forest Confusion Matrix

Confusion matrix is an useful measure for solving classification problems. To visualize more prediction result with random forest, we have show the random forest confusion matrix as follows:

N=70366	Predicted					
Actual		Property damage only	pain	Other injury	Severe injury	fatal
	Property damage only	1431	0	0	0	0
	Pain	0	17209	0	0	0
	Other injury	0	0	10287	0	0
	Severe injury	0	2	1	3261	0
	fatal	0	0	2	190	520

Table 4. Confusion Matrix for Random Forest Model

The accuracy of the classifier is calculated by the following formula:

(True Positive+True Negative)/Total=Accuracy Value.

Hence, the accuracy of the classifier is 0.994.

4.1.3. Random Forest Feature Importance

Now we will describe Feature Importances in random forest model. Infact, we figure out what predictor variables the random forest considers most important. The feature importances can be extracted from a trained random forest and put into a Pandas dataframe has seen in Figure 9.

feat	importance
Victim role	0.449560
Victim safety equipment1	0.241410
Victim seating position	0.143473
Secondary road	0.115767

Primary road	0.034933
Type of collision	0.014857

Fig9. Importance for intepretability for Random Forest

The outcome demonstrates the most important variables in order to predict “collis ion severity”.

4.2. Results of Logistic Regression Model

Logistic regression is the technique which works best when dependent variable is dichotomous (binary or categorical) [20]. Thus we have performed prediction of “collision severity” on dataset by using logistic regression model.

4.2.1. Logistic Regression Model Accuracy

In order to understand how the logistic regression model performs on our dataset, we need to obtain first the accuracy score of the model. The accuracy of the predictor on the training and test data are 0.6853 and 0.6872 with Mean Absolut Error 0.0024.

4.2.2 Logistic Regression Confusion Matrix

Table 5 shows the accuracy of the classifier by confusion matrix for logistic regression model is 0.6852.

N=70366	**Predicted**					
Actual		**Property damage only**	**pain**	**Other injury**	**Severe injury**	**fatal**
	Property damage only	3701	3013	170	2	0
	Pain	815	31109	4189	102	4
	Other injury	56	8668	11549	327	7

	Severe injury	3	1026	2703	1668	53
	fatal	0	106	364	398	333

Table 5. Confusion Matrix for Logistic Regression Model

4.2.3. Logistic Regression Feature Importance

Figure 10 determine the importance of features to predict “collision severity” by logistic regression.

feat	**importance**
Victim role	0.361475
Victim safety equipment1	0.311084
Type of collision	0.172061
Victim seating position	0.121039
Primary road	0.010951
Secondary road	0.023390

Fig10. Importance for intepretability for Logistic Regression

Conclusions

The outcome of the research shows that Victim seating position, type of collision, victim role, victim safety equipment 1, primary road, secondary road are important factors and the most important is victim role on severity of collision. In the research, two models to predict collision severity based on important features are represented. Random forest model can predict the outcome better than logistic regression model with 0.31% increase of power in prediction “collision severity”. Moreover, confusion matrices for two models approve two models accuracy. The results are important for DMV, Police, or even regular people to manage reducing severity of collisions.

Maryam_zohouri@yahoo.com