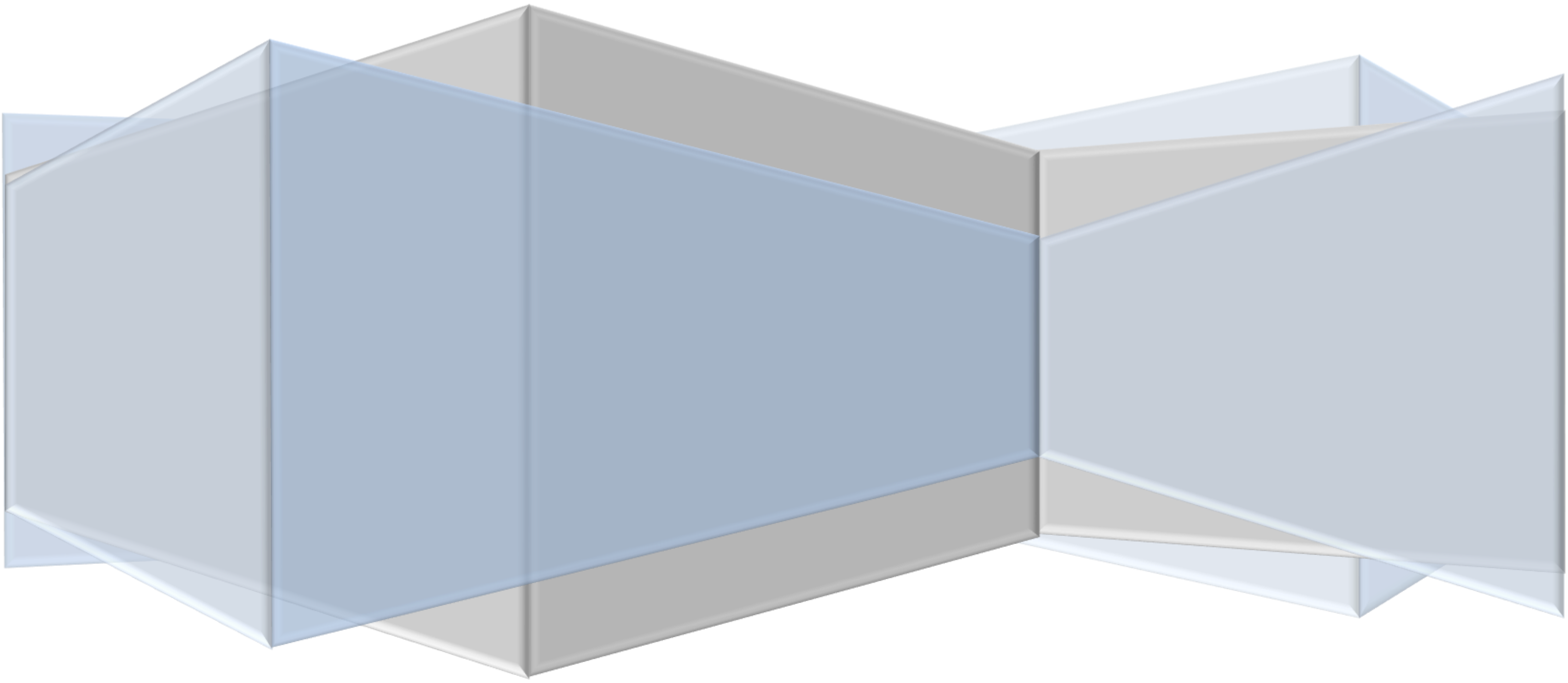
Coursera

CAR ACCIDENT SEVERITY PREDICTION

Applied Data Science Capstone

The project aims to understand the factors which play a role in the severity of accidents using Machine Learning Models



INDEX

BUSINESS PROBLEM

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. The project aims to predict how severity of accidents can be reduced based on a few factors.

UNDERSTANDING DATA

We chose the unbalanced dataset provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. It covers accidents from January 2004 to May 2020. Some of the features in this dataset include and are not limited to Severity code, Location/Address of accident, Weather condition at the incident site, Driver state (whether under influence or not), collision type. Hence we think its a good generalized dataset which will help us in creating an accurate predictive model.

Data used for our purpose is downloaded from following link :

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

METHODOLOGY

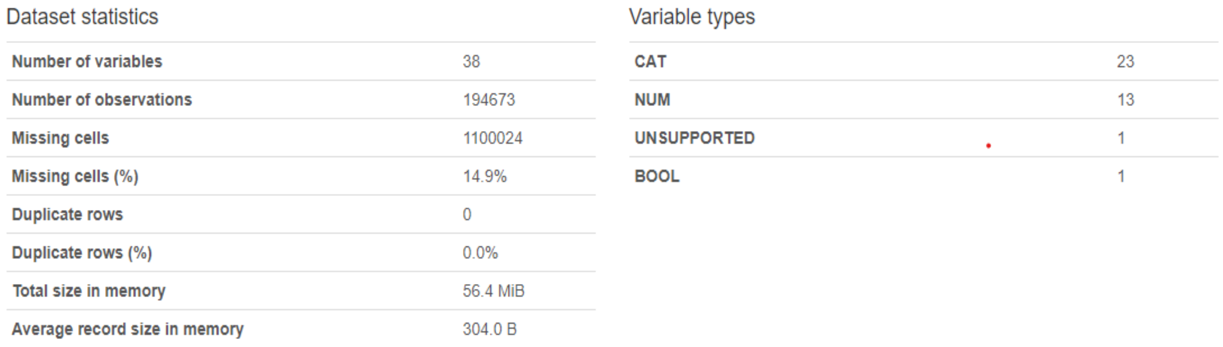
1.Data Collection

The dataset used for this project is based on car accidents which have taken place within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding car accidents the severity of each car accidents along with the time and conditions under which each accident occurred.

2.Exploratory data analysis

Use pandas profiling library to do EDA. So output.HTML would be generated. From this html file we can view statistics, No. of missing values, correlation, variable types etc.

DESCRIPTIVE STATISTICS:



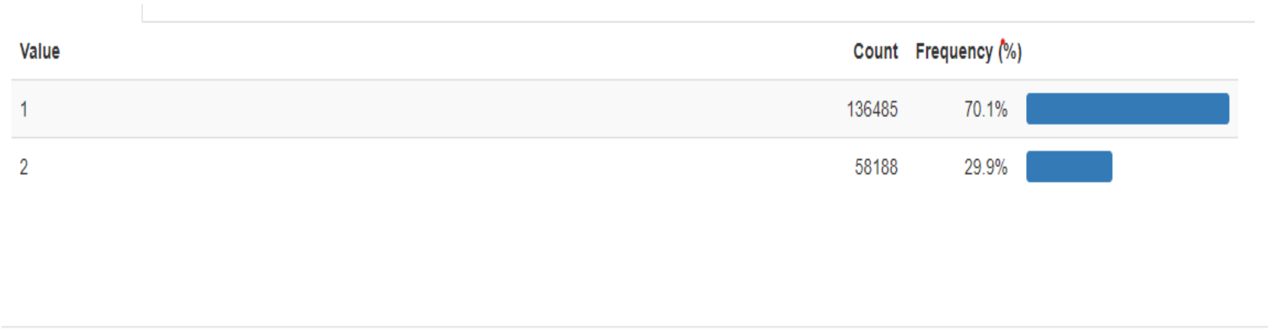
Dataset statistics

Number of variables	38
Number of observations	194673
Missing cells	1100024
Missing cells (%)	14.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	56.4 MiB
Average record size in memory	304.0 B

Variable types

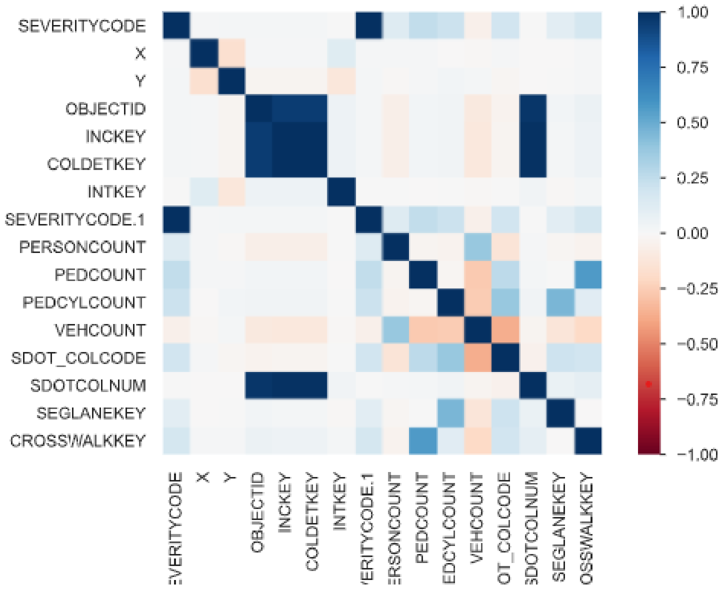
CAT	23
NUM	13
UNSUPPORTED	1
BOOL	1

FREQUENCY OF SEVERITY CODES:

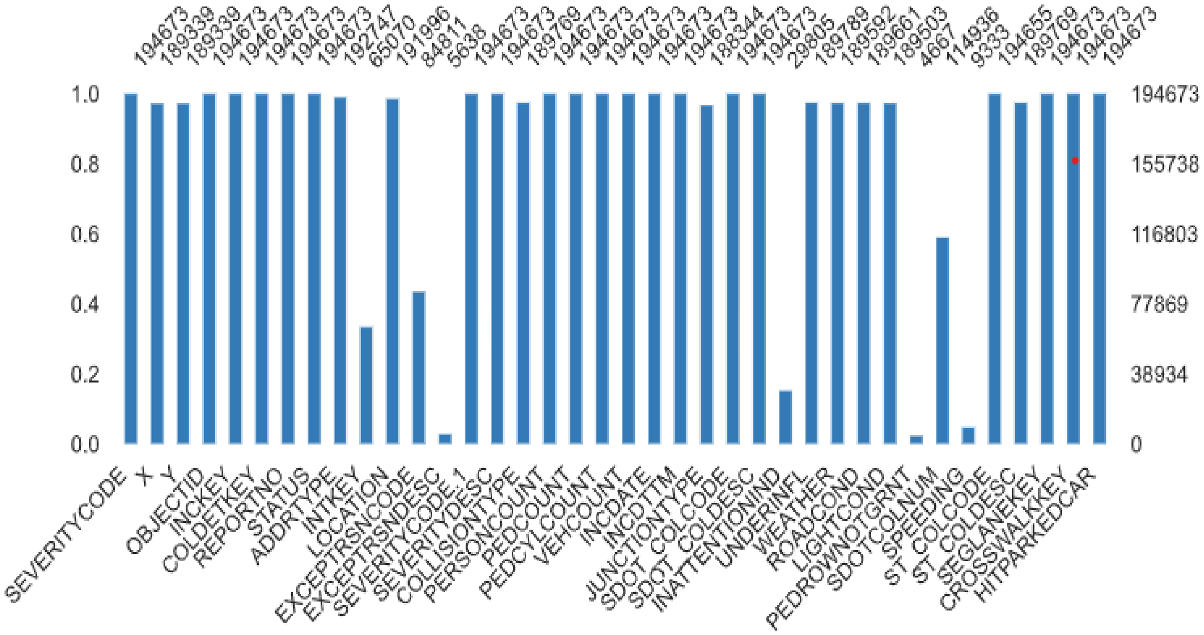


Value	Count	Frequency (%)
1	136485	70.1%
2	58188	29.9%

CORRELATION BETWEEN NUMERIC VARIABLES:



MISSING VALUES:

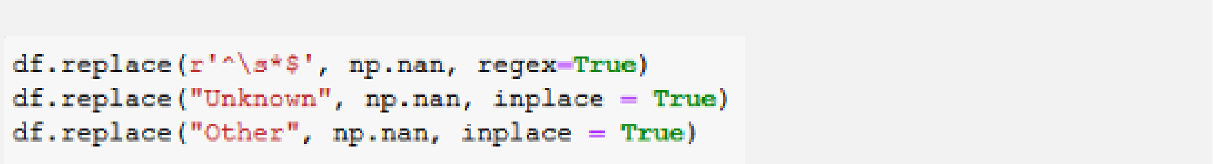


3.DROPPING IRRELEVANT COLUMNS

Columns containing descriptions and identification numbers that would not help in the classification are dropped from the data set to reduce the complexity and dimensionality of the data set. ‘OBJECTID’, ‘INCKEY’, ‘COLDETKEY’, ‘REPORTNO’, ‘STATUS’, ‘INTKEY’, ‘EXCEPTRSNCODE’ and more belong to this category. Certain other categorical features were removed as they had a large number of distinct values, example: ‘LOCATION’. After performing this step, the dimensionality dropped from 37 to 18.

4. HANDLING MISSING VALUES

To identify columns and rows with missing values is the next step. Empty boxes, ‘Unknown’ and ‘Other’ were values considered as missing values. These were replaced with NA to make the dataset uniform.

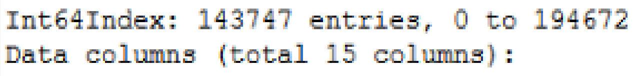


```
df.replace(r'^\s*$', np.nan, regex=True)
df.replace("Unknown", np.nan, inplace = True)
df.replace("Other", np.nan, inplace = True)
```

Replacing Missing Values with NA

Columns (“INATTENTIONIND”,”PEDROWNOTGRNT”,”SPEEDING”) which had more than 20% of its values missing were noted down and were dropped. For columns (“X”,”Y”,”COLLISIONTYPE”,”JUNCTIONTYPE”….) which had less than 20% of its values missing, the respective rows were removed since most of the columns in this dataset are categorical type, goal was to not impute the non-numerical columns; hence it did not make sense to replace the values.

Once the above two strategies were performed, the dataset reduced from having 194673 rows and 15 columns to having 143747 rows and 15 columns.

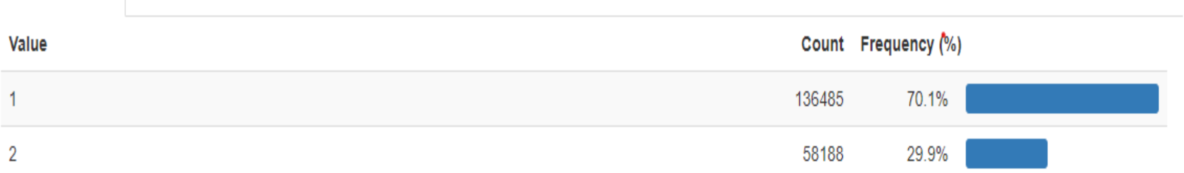


```
Int64Index: 143747 entries, 0 to 194672
Data columns (total 15 columns):
```

5.Balancing the dataset

With the above two pre-processing steps complete, a dataset (143747 rows) with 94821 rows for severity code 1 and 48926 rows for severity code 2 is obtained. Training an algorithm on an unbalanced dataset w.r.t the target category will result in a biased model. The model will have learnt more about one the category that has more data. In order to prevent this, a new balanced dataset (97852 rows) is created by randomly sampling out 48926 rows with severity code 2 and then concatenating it with 48926 rows with severity code 1. The dataset is then shuffled to randomize the rows.

FREQUENCY OF SEVERITY CODES ARE IMBALANCED HERE:



Value	Count	Frequency (%)
1	136485	70.1%
2	58188	29.9%

6. Encoding of data

The dataset is split into two datasets, X and Y, where Y contains the target feature (SEVERITYCODE) and X contains all the independent features/variables.

Machine Learning models are trained only on numerical data; hence all categorical features in the dataset have to be encoded so that the algorithms can be trained on those features. The ‘get_dummies’ method from pandas library is used to convert/encode each and every categorical feature. After application, number of features in dataset X increased from 14 to 50.

7. Normalizing/ Feature scaling of data

Feature scaling of data is done to normalize the data in a dataset to a specific range. It also helps improve the performance of the ML algorithms. Standard Scaler metric is used to scale/normalize all the numerical data for both, the X_train and X_test datasets. This completes the pre-processing stage, we can move on to training our models.

8. Machine Learning Algorithms

A total of six ML algorithms were trained on the pre-processed dataset and their accuracies were compared. A brief explanation on how each of them works along with their results in shown below.

1)Logistic Regression Classifier

Logistic Regression is a classifier that estimates discrete values (binary values like 0/1, yes/no, true/false) based on a given set of an independent variables. It basically predicts the probability of occurrence of an event by fitting data to a logistic function**.** Hence it is also known as logistic regression. The values obtained would always lie within 0 and 1 since it predicts the probability.

The chosen dataset has only two target categories in terms of the accident severity code assigned; hence it was possible to apply this model to the same.

2)K Nearest Neighbours Classifier

K nearest neighbours algorithm used for both classification and regression problems. It basically stores all available cases to classify the new cases by a majority vote of its k neighbours. The case assigned to the class is most common amongst its K nearest neighbours measured by a distance function (Euclidean, Manhattan, Minkowski, and Hamming).

3)Naïve Bayes Classifier

Naive Bayes classifies objects based on Bayes' Theorem with an assumption that the predictors (features) are independent of each other. Bayes theorem is a way to calculate posterior probability P(c|x) from the P(c), P(x), P(x|c). Naive Bayes is naive because it assumes the presence of a particular feature is completely unrelated to the presence of another, and each of them contributes to the posterior probability independently.

4)Decision Tree Classifier

Decision Tree makes decision with tree-like model. It splits the sample into two or more homogenous sets (leaves) based on the most significant differentiators in the input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by category; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).

Information gain for a decision tree classifier can be calculated either using the Gini Index measure or the Entropy measure, whichever gives a greater gain. A hyper parameter Decision Tree Classifier was used to decide which tree to use, DTC using entropy had greater information gain; hence it was used for this classification problem.

5)Random Forest Tree Classifier

Random Forest Classifier is an ensemble (algorithms which combines more than one algorithms of same or different kind for classifying objects) tree-based learning algorithm. RFC is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. Used for both classification and regression.

Similar to DTC, RFT requires an input that specifies a measure that is to be used for classification, along with that a value for the number of estimators (number of decision trees) is required. A hyper parameter RFT was used to determine the best choices for the above mentioned parameters. RFT with 75 DT's using entropy as the measure gave the best accuracy when trained and tested on pre-processed accident severity dataset.

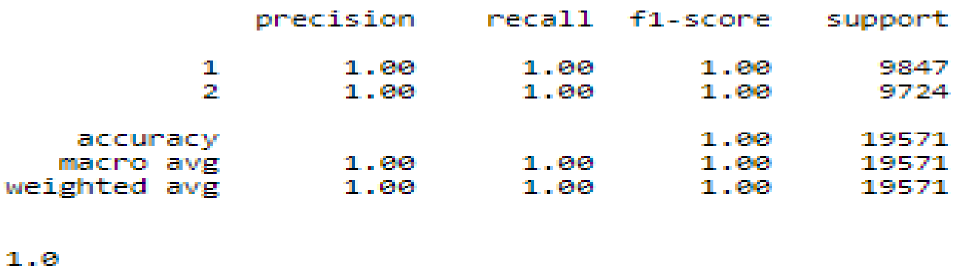
6)Support Vector Machine Classifier

Support Vector Machine is an algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes.

Hyper parameter SVC was used to choose between Linear SVC and a Kernel SVC and the latter arrived on top with a greater accuracy when applied on the dataset in question. It used the ‘radial basis function’ kernel for performing the classification.

Result:

For me each algorithm generate same accuracy. But , many times each algorithm had different accuracy.



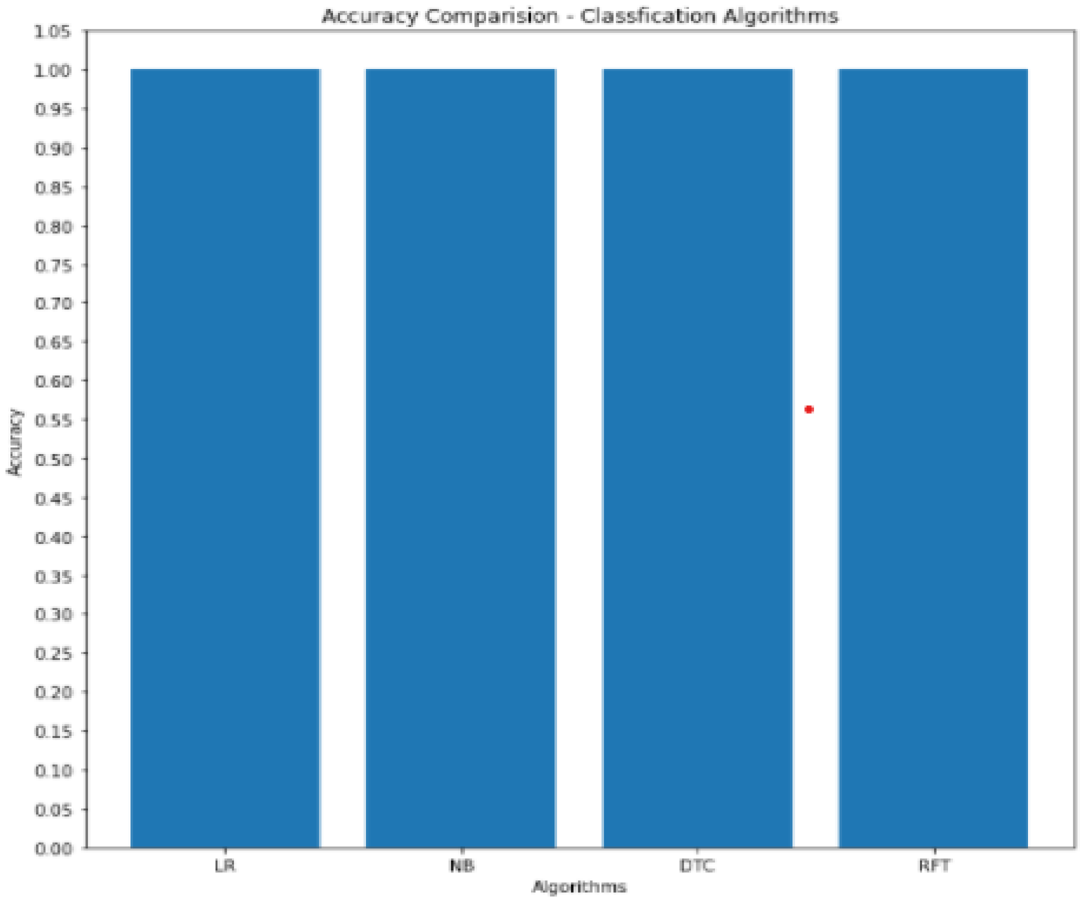
```
              precision    recall  f1-score   support

           1       1.00      1.00      1.00      9847
           2       1.00      1.00      1.00      9724

    accuracy                           1.00     19571
   macro avg       1.00      1.00      1.00     19571
weighted avg       1.00      1.00      1.00     19571

1.0
```

So comparison of these algorithms is shown below:



Conclusions

Now we can compare all of the above models and can choose model with maximum accuracy. After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.