*A Mini Project report on*

Accident Casualty Severity

Prediction System

using Gradient Boosting Classifier

*submitted in partial fulfillment of the course*

CSE-1006: Foundation of Data Analytics

By

19BCE7467-Ishan Gupta

19BCN7283-Kinshuk Agarwal

Submitted to

Dr. Deepasikha Mishra



School of Computer Science & Engineering

VIT-AP UNIVERSITY, INAVOLU, AMARAVATI

OCTOBER, 2019

INDEX

|  |  |  |
| --- | --- | --- |
| S.no | Heading | PageNo |
| 1 | INTRODUCTION | 3 |
| 2 | AIM | 4 |
| 3 | WORKING WITH DATA SET | 5 |
| 4 | EXTRACTING DATA | 5 |
| 5 | DATA CLEANING | 6 |
| 6 | DATA SORTING | 7 |
| 7 | PREDICTION/ANALYSIS USING ML TECHNIQUE | 9 |
| 8 | PLOTS | 10 |
| 9 | RESULTS | 14 |
| 10 | CONCLUSION | 15 |
| 12 | FUTURE SCOPE | 15 |

**INTRODUCTION:**

The increasing number of road and traffic accidents may be a challenging issue to the transportation systems. It not only concerns with health issues but also related to the economic burden on society. Therefore, it is an important task for the safety analysis to carry out a comparative study of road accidents to identify the factors that cause an accident to happen so that preventive actions are often taken to beat the accident rate and severity of accident consequences. Therefore, a comparative study of road accident data is required to identify the several factors associated with road accidents. The main concern with road accident data analysis is to spot the foremost influential factors affecting road accident frequency and accident severity. The major problem with road accident data analysis is it's heterogeneous.

Heterogeneity in road accident data is very undesirable and unavoidable. The major disadvantage of heterogeneity of road accident data is that certain relationships may remain hidden like certain accident factors related to a particular vehicle type might not be significant in the entire data set the enormity of the effect of certain accident-related factors may be different for various conditions severity levels for an accident contributing factors may be different for different accident types. To urge more accurate results this heterogeneity of road accident data must be removed to affect this heterogeneous nature of road accident data, divide the info into groups based on some exogenous attributes e.g., accident location, road condition, cause of the accident, and analyzed every group separately to identify several influential factors associated with road accidents in each group. However, this method is unrealistic as grouping the info supported certain attributes may lead to smaller groups. Some subgroups can have a large number of samples and some can have a very low number of samples and thus restricting the application of accident severity models. Another choice is to use some approach like data processing to get rid of the heterogeneity of the road accident data.

In road accident data analysis, it's suggested that prior segmentation is extremely much useful in getting good results. However, these factors can segment the data into workable groups but this cannot be guaranteed that the subgroups will comprise of a homogeneous group of accidents.

**AIM:**

One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. However, heterogeneous nature of road accident data makes the analysis task difficult. Data segmentation has been used widely to overcome this heterogeneity of the accident data.

In this paper, we proposed a framework that the Accident Casualty of an Individual from a given dataset which includes weather conditions, temperature, lighting conditions, etc. The approach for prediction of casualty severity has two steps:

a. Cluster the data using agglomerative clustering algorithm

b. Prediction using GradientBoostingClassifier

We chose using the Gradient Boosting Classifier as Classification algorithms frequently use logarithmic loss, while regression algorithms can use squared errors. Gradient boosting systems don't have to derive a new loss function every time the boosting algorithm is added, rather any differentiable loss function can be applied to the system.

Gradient boosting systems have two other necessary parts: a weak learner and an additive component. Gradient boosting systems use decision trees as their weak learners. Regression trees are used for the weak learners, and these regression trees output real values. Because the outputs are real values, as new learners are added into the model the output of the regression trees can be added together to correct for errors in the predictions.

**WORKING WITH DATA SET:**

* Pandas
  + For Calculation metrics about our data
  + Perform basic queries and aggregations
* Matplotlib
  + Visualize our data
  + Plotting Graphs

**EXTRACTING DATA:**

Data extraction is the process of extracting the data from various sources.

This project predicts the Accident Casualty of an Individual from a given dataset which includes weather conditions, temperature, lighting conditions, etc.

We have taken the data of the road accidents in the 2015 for the different states and union territories in India.

Link for the database-

<https://trello-attachments.s3.amazonaws.com/5cf2142046ceb163a0e4b189/5cf4e21e143159856a320b36/2d49903aabdaaad41a008c499314c51c/Accident_Dataset.csv>

**DATA CLEANING:**

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Data cleansing is also important because it improves your data quality and in doing so, increases overall productivity. When you clean your data, all outdated or incorrect information is gone – leaving you with the highest quality information.

Our dataset was already clean and didn’t include any sort of null values or inaccurate records.

**DATA SORTING:**

As it would have been difficult for everyone to enter the details everytime we have to predict the sevierty we have replaced the data with numbers and saved it into a new dataset.

**The condition of road surface which differ as:**

1 - DRY

2 - WET/DAMP

3 - SNOW, Frost/Ice

4 – Flood (Surface water over 3cm deep)

**The condition of lighting which differ as:**

1 - Daylight: street lights present

2 - Darkness: street lights present and lit

3 - Darkness: street lighting unknown,

4 - Darkness: no street lighting,

5 - Darkness: street lights present but unlit

**The condition of weather which differ as:**

1 - Fine without high winds,

2 - Fine with high winds

3 - Raining without high winds

4 - Raining with high winds

4 - Snowing without high winds

5 - Fog or mist – if hazard,

6 - Snowing with high winds

7 - Other, Unknown

**The Casualty Victim who are:**

1 - Driver/Rider

2 - Passenger

3 – Pedestrian

**The condition of Casualty Victim (i.e., Casualty Severity) which differ as:**

1 - Slight

2 - Serious

3 - Fatal

**The Sex of Victim:**

1 - Male

2 - Female

**The type of Vehicle:**

Motorcycle over 50cc and up to 125cc, Pedal cycle, M/cycle 50cc and under, Mobility Scooter, Motorcycle over 125cc and up to 500cc, Motorcycle over 500cc, Car, Taxi/Private hire car, Goods vehicle 3.5 tonnes mgw and under, Bus or coach (17 or more passenger seats), Agricultural vehicle (includes diggers etc.), Goods vehicle 7.5 tonnes mgw and over, Goods vehicle 3.5 tonnes mgw and under, Minibus (8 – 16 passenger seats), Other vehicle, Tram / Light rail

**Age of the victim**

Any input between 1-99

After Sorting out data like this and replacing the things in the dataset it would be easy for everyone to check the sevierty on a given road,weather condition.

**PREDICTION USING lasso lar’s regression model:**

The approach for prediction of casualty severity has two steps:

a. Cluster the data using agglomerative clustering algorithm,

**The approach for prediction of casualty severity has two steps:**

a. Cluster the data using agglomerative clustering algorithm Cluster 0

Rules: Road Surface is in average 15.31% smaller: mean of 1.10 against 1.30 globally Casualty Severity is in average 11.63% smaller: mean of 1.00 against 1.13 globally Weather Conditions is in average 15.13% smaller: mean of 1.04 against 1.22 globally

Cluster 1 Rules: Casualty Severity is in average 80.57% greater: mean of 2.05 against 1.13 globally Type of Vehicle is in average 12.33% smaller: mean of 3.32 against 3.79 globally Number of Vehicles is in average 14.70% smaller: mean of 1.67 against 1.95 globally

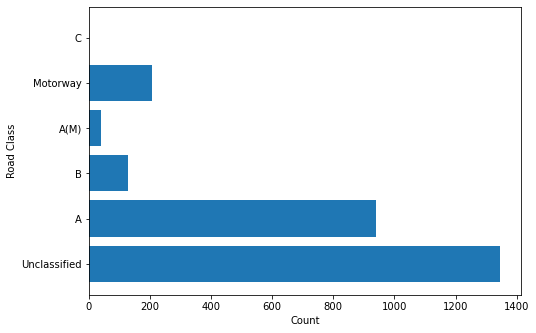
Cluster 2 Rules: Road Surface is in average 57.93% greater: mean of 2.06 against 1.30 globally Weather Conditions is in average 61.05% greater: mean of 1.97 against 1.22 globally Casualty Severity is in average 10.32% smaller: mean of 1.02 against 1.13 globally

b. Prediction using Gradient Boosting Classifier

First, the Gradient Boosting ensemble is fit on all available data, then the predict() function can be called to make predictions on new data.

**PLOTS:**

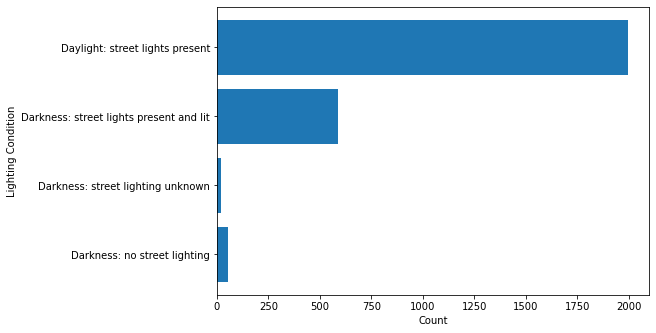
**Accidents due to different Road Class**



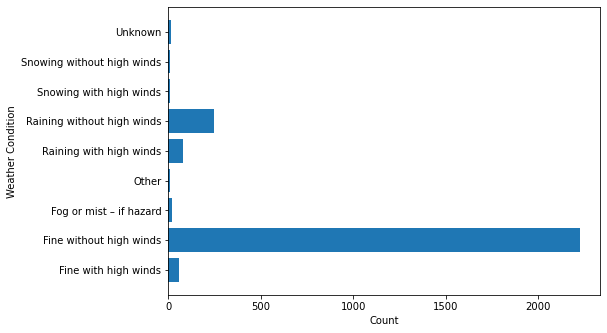
**Accidents due to different Road Surface**



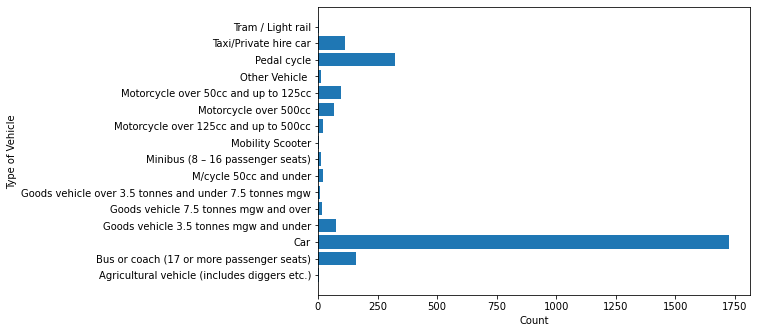
**Accidents due to different Lighting Conditions**



**Accidents due to different Weather Conditions**



**Accidents due to different Type of Vehicles**



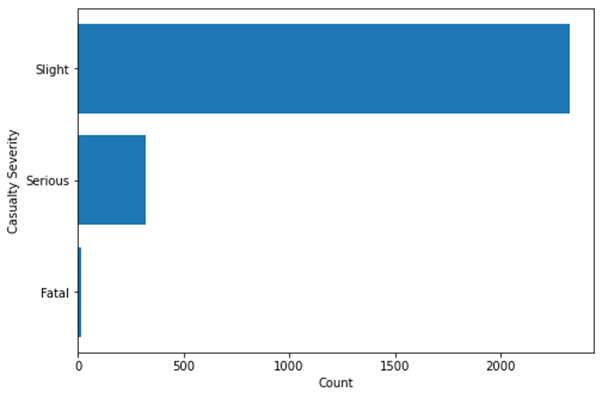
**Accidents involving Different sex of Casuality**



**Accidents including different Casualty Class**

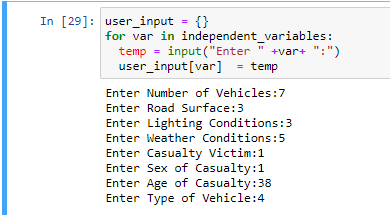


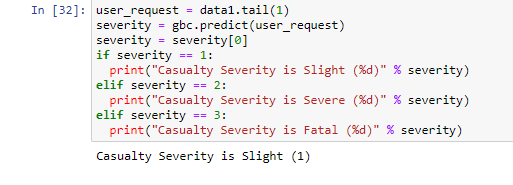
**Accident resulting to Casualties**



**RESULTS:**

**Input-**



**Output-**

**CONCLUSION:**

By predicting the Casualty Severity we can make better arrangements in case of the crash.

Heavy traffic can be avoided and at the time when the chances of fatal accidents, people could be warned and accordingly safety measures could be used.

If possible we could also restrict the entry of vehicles when fatal accidents can be predicted.

**FUTURE SCOPE:**

In the near Future, we could add a new dataset which includes the time of accidents as well and can predict the time duration of the accidents as well.

Also we could integrate the model into a website and make it open for public usage so that people could plan their journeys accordingly