**Predicting car accident severity**

**Raksha Subray**

**1.Intoduction**

In this project we will try to find Car accident severity. Specifically, this report will be targeted to the **people planning to travel or travelling**.

Since we know the weather condition can predict the severity of the accident which may cause due to bad weather so that **people traveling will be cautious or the people planning to travel may reschedule based on the severity**. which will help the people to plan safe travel rather than unexpected glitch.

We will use our data science powers to predict accident severity, alerting the ones who are traveling and helping the once’s to plan their safe travel based on weather condition.

**2. Data acquisition and cleaning**

**2.1 Data sources**

As per problem statement severity is influenced by few feature such as climatic condition, road condition, weather and speed, we can find data set from records which are available in Kaggle. But based on my requirement I have scraped data from belove link.

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

which contains information related to severity and description on main factors influencing it such as severity and road as well as weather condition.

**2.2 Data cleaning**

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping. I decided to only use data which has clear information with respect to fields such as road condition, weather and light condition.

Wherein analyzed missing data with above mentioned field and impact analysis, example if there is relationship with rainy weather and dim light and empty road condition then analyzed what happens if I drop the row, as there where no much impact because of large volume and different combination of these factors effecting severity. There where initially 194673 records after dropping rows with missing information 189337 where remaining which has complete sufficient information

**2.3 Feature selection**

After cleaning data out of 194673 records 189337 where considered for feature extraction. The whole data set had around 38 columns which had information related to location of accident, area, continent, country, number of vehicle, location condition such as cross road, people count, weather, light, road, axis of the location and many unrelated factors with severity where as which supported the exact condition and all information related to car accident. But as per use case and problem statement, there was relation between the weather condition with respect to accident and severity, same way more factors such as road condition, and light condition impacted the severity, considering these three factor as per analysis could predict the severity, so chose above three fields as in depend factors which would effect the depend factor severity. So chose the severity as label which was affected by there conditions. So, After feature selection data set would look like

Table

Description automatically generated

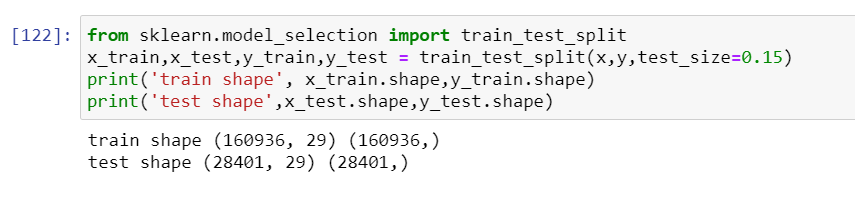
**3. Predictive Modeling**

There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. For example, severity prediction of car accident is the case based on factor either we select 1,2 which is selection problem. Therefore, in this study, I carried out classification modeling.

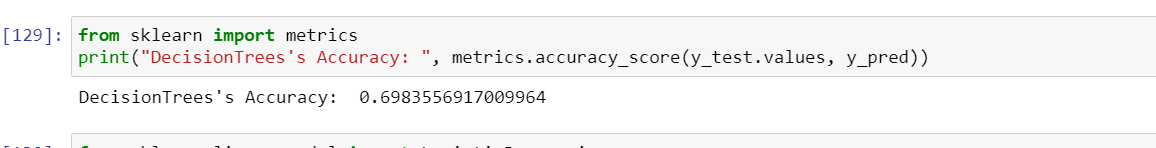
**4.1 Classification models**

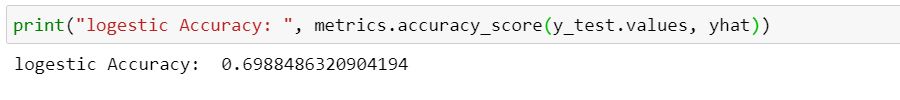
Severity of the car accident is selection problem where should select one of the category, As all the factors effecting the severity is categorical data so we had to chose classification models. As finding the severity is based on labeled data where there is the set of observation for which already severity is assigned hence it is the case of supervised learning. And under supervised learning case of classification, There several popular simpler models such as decision tree which is most promising one. There are other algorithms such as Knn classification, logistic regression( which is both regression as well as classification algorithm), SVM, random forest many more.

Divided whole data set into train as well as test. Where train data set is used to train the model. And test data is used to test the accuracy of the model. Divided the whole data set into 75:25 where 75% is train data and 25% is test data.



Chose 2 classification methods such as decision tree which divides the data into branches until all the leaf nodes are pure. So carried out decision tree which is one of the most promising model, with train data trained the model and predicted the outcome for test data and recorded the accuracy. Other model used for prediction is logistic regression which is one of the regression as well as classification model, same way prepared the model and predicted the out come and recorded the accuracy.





**4. Conclusions**

In this study, I analyzed the relationship between severity with weather, road condition and light condition. And categorized into severity 1,2 based on 3 condition using classification method. The result was evident that severity is effected by weather , road and light condition, which was evident in bad weather and night light the accident severity was 2. Based on two models which was almost of same efficiency successfully classified the unseen data into severity based on 3 conditions. Which proved initial guess with respect to factors effecting severity of accident was evident.

**5. Future directions**

Models in this study mainly focused on individual features. However car condition and speed at which the car was driving and driver skill may improve the model ,but considering the driving skill is difficult to analyze consider during modeling , could bring significant improvements to the models.