**IBM Capstone Project – Accident severity prediction**

## 

# Introduction:

Over 37000 people die in road crashes each year in addition to 4.35 million injured or disabled. Los Angeles a popular destination among tourists has faced accidents costing to 20 billion USD per year and has been on a constant rise where steps are taken to predict and halt such incidents. But severity of road accidents has been one of the least studied area though they have a bigger impact on the casualty rate. Recently machine and statistical learning has gone hand in hand predicting solutions to mysterious questions. With advancements in technology, the process of predicting an event based on a given data set after training has led to important findings. This project aims to answer questions related to severity of road accidents in Los Angeles relating to weather and other contributing factors and take appropriate measures to predict the severity of road disasters. The data “A Countrywide Traffic Accident data set” by Sobhan needed for predictive analysis is obtained from smoosavi.org, an online data set repository. Each accident record comprises of a multitude of inherent and contextual characteristics such as place, time, description of the natural language, climate, time and interest points. This project seeks to contribute the knowledge of accidents motorway accidents by creating models such as logistic regression and other classification modules. The expected outcome from this predictive analysis is to predict the level of severity for an accident namely “Low” and “High” based on previous accidents with predictors of weather and other factors. Here, severity refers to the fatality of accidents with the underlying assumptions that more the traffic delay higher the severity. The coordinates available in the data set can be helpful in predicting further instances of accident at a given locality and various environmental factors influencing them. Predictions based on nearby traffic signal to considerations to enhance the predicted model. The predicted data with “high severity” can be helpful in setting up hospitals or amenity stations near the accident hot spots and thereby significantly reduce the fatality rates.

## **Data Collection and Preprocessing:**

The data was collected from Kaggle [and](https://www.kaggle.com/sobhanmoosavi/us-accidents) Soban Moosavi [[24].](https://smoosavi.org/datasets/us_accidents) The table below shows the original list of predictors used in the obtained data set. Predictors were filtered since most were redundant to the response ”Severity”. Moreover, 243 rows of NA values were removed as they contributed to less than 0.5% of the final data. The highlighted cells in the table shows the filtered list of final predictors used in modelling.

**Red**-Response Variable

**Purple**- Predictors used along with response

**Green**-Used to plot accident point in Exploratory Data Analysis

Table 1: List of Original Predictors.

|  |  |
| --- | --- |
| Column Name with Description | |
| ID | Unique identifier of the accident record |
| Source | Indicates source of the accident report |
| TMC | Traffic accident may have a Traffic Message Channel |
| Severity | Severity number between 1 and 4 |
| Start Time | Start time of the accident in local time zone |
| End Time | End time of the accident in local time zone |
| Start Lat | Latitude in GPS coordinate of the start point |

Table 1: List of Original Predictors.

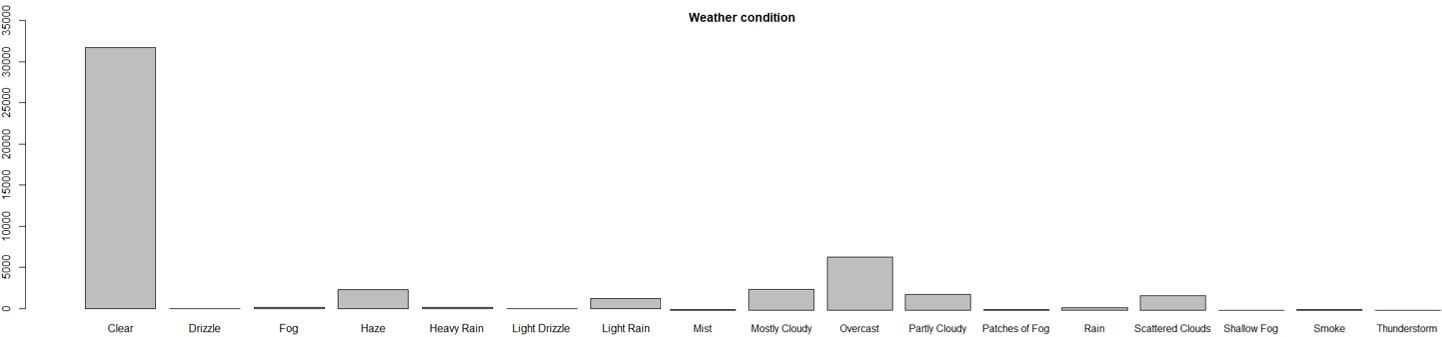
|  |  |
| --- | --- |
| Column Name with Description | |
| Start Lng | Longitude in GPS coordinate of the start point |
| End Lat | Latitude in GPS coordinate of the end point |
| End Lng | Longitude in GPS coordinate of the end point |
| Distance(mi) | The length of the road affected |
| Description | natural language description of the accident |
| Number | Street number in address field |
| Street | Street name in address field |
| Side | Relative side of the street |
| City | City in address field |
| County | County in address field |
| State | State in address field |
| Zipcode | Zipcode in address field |
| Country | Country in address field |
| TimeZone | Timezone based on the location |
| Airport Code | Denotes an airport-based weather station |
| Weather Timestamp | Time-stamp of weather observation |
| Temperature(F) | Temperature (in Fahrenheit) |
| Wind Chill(F) | Wind chill (in Fahrenheit) |
| Humidity | Humidity (in percentage) |
| Pressure(in) | Pressure (in inches) |
| Visibility(mi) | Visibility (in miles) |
| Wind Direction | Wind direction |
| Wind Speed(mph) | Wind Speed (in miles per hour) |
| Precipitation(in) | precipitation (in inches) |
| Weather Condition | Weather condition (rain, snow, thunderstorm, fog, etc.) |
| Amenity | Indicates presence of amenity |
| Bump | Indicates presence of speed bump |
| Crossing | Indicates presence of Crossing |
| Give Way | Indicates presence of Give Way |
| Junction | Indicates presence of Junction |
| No Exit | Indicates presence of No Exit |
| Railway | Indicates presence of Railway |
| Roundabout | Indicates presence of Roundabout |
| Station | Indicates presence of Station |
| Stop | Indicates presence of Stop |
| Traffic Calming | Indicates presence of Traffic Calming |

List of Original Predictors.

|  |  |
| --- | --- |
| Column Name with Description | |
| Traffic Signal | Indicates presence of Traffic Signal |
| Turning Loop | Indicates presence of Turning Loop |
| Sunrise Sunset | Period of day based on sunrise/sunset |
| Civil Twilight | Period of day based on Civil Twilight |
| Nautical Twilight | Period of day based on Nautical Twilight |
| Astronomical Twilight | Period of day based on Astronomical Twilight |

## Grouping similar Weather Conditions

The initial list of weather condition had 17 levels ranging from clear to thunderstorms and rain.



**Spread of initial Weather Conditions**

Similar weather conditions were grouped together to provide easy interpretability with 4 standard levels.

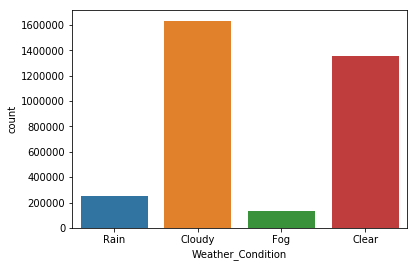
The grouping is as follows:

1.Clear

2.Rain - Drizzle, Heavy Rain, Light Drizzle, Light Rain, Thunderstorms and Rain

3.Cloudy - Mostly Cloudy, Overcast, Partly Cloudy, Scattered Clouds

4.Fog - Smoke, Fog, Haze,Mist, Patches of Fog, Shallow Fog



Grouped Weather Conditions

### Methodology:

The data is loaded into the juypter notebook. The necessary libraries such as pandas ,numpy, seaborn, sklearns are loaded. The severity is taken as the predicting label and other factors such as Traffic signal, sunrise sunset, turning loop, temperature, weather\_condition ,side, visibility(mi),precipitation(in),crossing are cleaned .The nan are dropped from all the above. The weather\_condition contains too many different variable and they are grouped under Rain, Clear, Cloudy and Fog.

The data is then converted into numpy array and training -testing datasets are separated in ratio 75-25.

The **onehotencoder** function is used to change all the string and Boolean data to numerical data for the machine learning process.

The machine learning used are:

### Logistic Regression Model

Logistic Regression is a simple linear model where the response is categorical i.e the response falls into binary classification. It is calculated based on the sum of log likelihood an event is about to happen.

*Zi* = *ln*(*Pi/*1 − *Pi*) The log of events are converted back to probability

*Pi* = 1 − (1*/*1 + *ei*) where, *Pi* = *probability of an event*

|  |  |  | Random Forest Random forests make use of the simplicity in decision trees and adds flexibility thereby increasing the accuracy. The data is first bootstrapped, random forest method randomly selects variables to create internal nodes in the tree and chooses the option with the most votes. The accuracy of the random forest is checked on every iteration using a new bootstrapped data and the variables used are changed. The testing data is run through the different trees in the random forest and choose the option with the most votes. The accuracy of the random forest is calculated with the help of “out of bag samples” correctly identified. The proportion of out of bag samples that were incorrectly classified is called “Out of Bag error”. We choose the random forest which is the most accurate by changing the number of variables in the data. Missing data in random forest are handled by using the most common value in the original data for categorical variables and by median value in case of numeric.  **Model Framework** |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |

## **Result:**

### Logistic Regression Model

Logistic Regression is a simple linear model where the response is categorical i.e the response falls into binary classification. It is calculated based on the sum of log likelihood an event is about to happen.

*Zi* = *ln*(*Pi/*1 − *Pi*) The log of events are converted back to probability

*Pi* = 1 − (1*/*1 + *ei*) where, *Pi* = *probability of an event*

The jaccard similarity score for the model is 0.6767152027967516

## Random Forest- Variable Importance Plot

A variable importance plot is taken from the random forest model to study the predictors which has the significant impact.

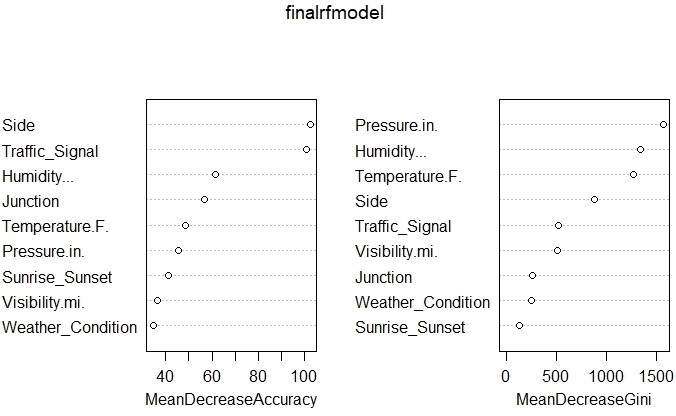


Figure 11: Variable Importance plot- Random Forest

From the variable importance plot we can infer that the two most influential predictors are presence of traffic signal and the side of the road. It is obvious that the presence of traffic signal can largely affect the likelihood of accident to be occurred.It is interesting to note that Weather condition and Visibility have low significance. So, from our results we can suggest to have more number of traffic signals and for the places where traffic signals can’t be placed we can suggest to has appropriate signboards and roundabouts to control and reduce the speed of the moving traffic. Also, another most influential predictor is side of the road while traveling and we can infer from the exploratory analysis that most of the high severe accidents have happened on the right side of the road and since most of the high severe accidents have happened on freeways we could say the accidents have happened when a vehicle is trying to make an exit from the right lane. So, restricting the speed of the vehicles and providing roundabouts for the traffic that are taking a exit will help reduce the accident fatality.