APPLIED DATA SCIENCE CAPSTONE PROJECT

PREDICTION OF CAR ACCIDENTS SEVERITY PREDICTION IN US

By. Sivasankari Balasubramanian

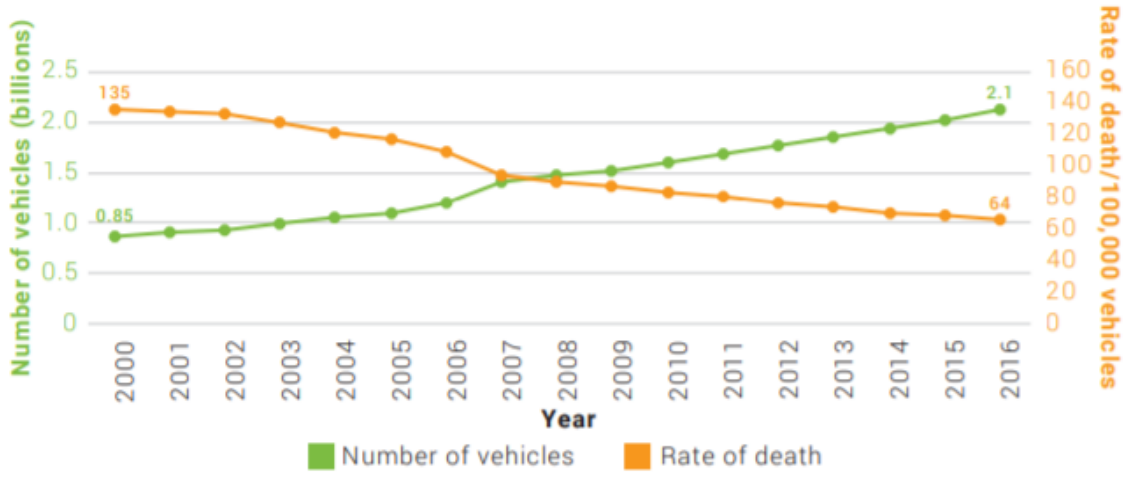
Index

Introduction

1.1 Business Understanding:

Road accidents are serious concern for most of the nations around the world because accidents can cause severe injuries and fatalities. Traffic accidents are the leading causes beyond death. According to the World Health Organization's Global status report on Road Safety 2018, number of road accident deaths are continued to climb, reaching 1.35 million in 2016. Road traffic injuries are the eighth leading cause of death for all age groups. According to the report Road traffic injuries are currently the leading cause of death for children and young adults aged 5–29 years and the second leading cause of death worldwide amongst children ages 5-14. Shockingly, crashes account for 2.2% of all deaths around the world. The World Health Organization predicts that at the current rate, car accidents are likely to become the fifth leading cause of death globally by 2030.

Figure1: Number of motor vehicles and rate of road traffic death per 100,000 vehicles: 2000–2016



Sources: GLOBAL STATUS REPORT ON ROAD SAFETY 2018

According to The National Safety Council, more than 40,000 people in the U.S. are killed in car crashes every year. They also estimate that 4.57 million people were sustained seriously enough injuries require medical. The total cost to society for these crashes comes to about $413.8 billion a year.

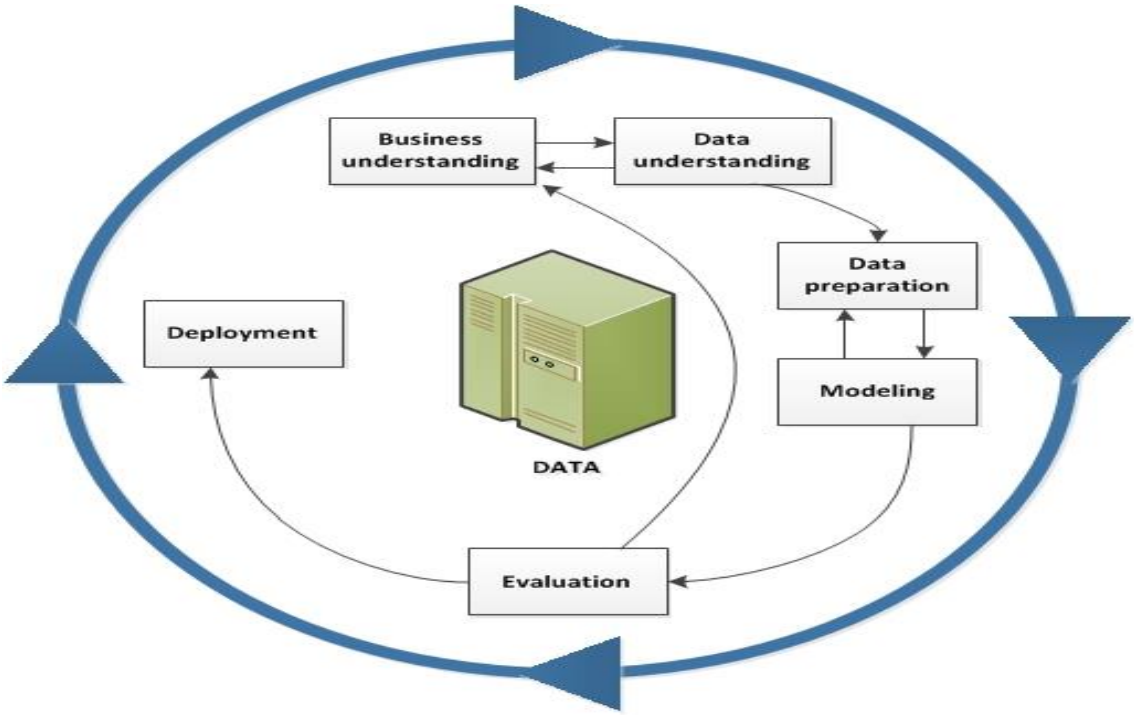
Reducing the traffic Incidents are always an important challenge. If we identify the patterns of how these severe accidents happen and the key factors, we might be able to implement the effective safety measures which are highly impact on human deaths, severe injuries and social economic loss. These safety measures will help the Traffic control Authorities, Transportation Department to impose and regulate the Traffic rules on Road safety and identify the Accident-prone areas.

1.2 Objective:

Our motivation is to predict the accident severity of any road, weather conditions, and the environment which are playing key roles in the collision. Our first aim is to identify the key factors causing the accidents as mentioned in second and third phases (i.e., Data Understanding and Data Preparation) of Data science methodology (Figure 2) and the second one is developing a model that can accurately predict the severity of the accidents.

Data cleaning is performed to identify and handle the corrupt and missing records. Further we are using some classification algorithms and evaluation methods to predict the severity. Here we are using SVM, LR, KNN and Decision tree classification models to predict the severity.

Figure 2: Six phases of Data science methodology. CRISP - DM



Source:
https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/images/crisp_process.jpg

2. Data Understanding:

2.1 Dataset Understanding:

US accidents Dataset with 3.5 million records and 49 columns including weather conditions, Turning loop, wind speed etc collected from Kaggle is used in this project. It covers countrywide car accidents which includes 49 states of the USA. The accident data are collected from February 2016 to June 2020.

Dataset Link: https://www.kaggle.com/sobhanmoosavi/us-accidents

2.2 Attributes Explanation:

Traffic Attributes:

ID - Accident Record unique identifier.

Source - Source of the accident report (i.e. the API which reported the accident).

TMC - Traffic Message Channel code providing detailed description of the incident.

Severity - severity of the accident, a number between 1 and 4, where 1 is the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).

Start_Time - start time of the accident.

End_Time - refers to when the impact of accident on traffic flow was dismissed.

Start_Lat - Shows latitude in GPS coordinate of the start point.

Start_Lng - Shows 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) - length of the road where the accident happens.

Description - Description of accident.

Address Attributes:

Number - Street number in the address field.

Street - Street name.

Side - Relative side of the street (Right/Left) in address field.

City - City name.

County - County name.

State - State name.

Zipcode - Zipcode in address field

Country - Country name.

Timezone - timezone based on the location of the accident (eastern, central, etc).

Airport_Code - airport-based weather station which is the closest one to location of the accident.

Weather Attributes:

Weather_Timestamp - timestamp of weather observation.

Temperature(F) - temperature (in Fahrenheit).

Wind_Chill(F) - wind chill (in Fahrenheit).

Humidity(%) - Humidity(in percentage).

Pressure(in) - air pressure (in inches).

Visibility(mi) - Visibility (in miles).

Wind_Direction - Wind direction.

Wind_Speed(mph) - wind speed (in miles per hour).

Precipitation(in) - precipitation amount (in inches), if there is any.

Weather_Condition - weather condition (rain, snow, thunderstorm, fog, etc.).

Point-Of-Interest Attributes(13):

Amenity - A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location.

Bump - A POI annotation which indicates presence of speed bump or hump in a nearby location.

Crossing - A POI annotation which indicates presence of crossing in a nearby location.

Give_Way - A POI annotation which indicates presence of give_way sign in a nearby location.

Junction - A POI annotation which indicates presence of junction in a nearby location.

No_Exit - A POI annotation which indicates presence of no_exit sign in a nearby location.

Railway - A POI annotation which indicates presence of railway in a nearby location.

Roundabout - A POI annotation which indicates presence of roundabout in a nearby location.

Station - A POI annotation which indicates presence of station (bus, train, etc.) in a nearby location.

Stop - A POI annotation which indicates presence of stop sign in a nearby location.

Traffic_Calming - A POI annotation which indicates presence of traffic_calming means in a nearby location.

Traffic_Signal - A POI annotation which indicates presence of traffic_signal in a nearby location.

Turning_Loop - A POI annotation which indicates presence of turning_loop in a nearby location.

Period-of-Day (4):

Sunrise_Sunset - period of day (i.e. day or night) based on sunrise/sunset.

Civil_Twilight - period of day (i.e. day or night) based on civil twilight.

Nautical_Twilight - period of day (i.e. day or night) based on nautical twilight.

Astronomical_Twilight - period of day (i.e. day or night) based on astronomical twilight.

3. Data Preparation:

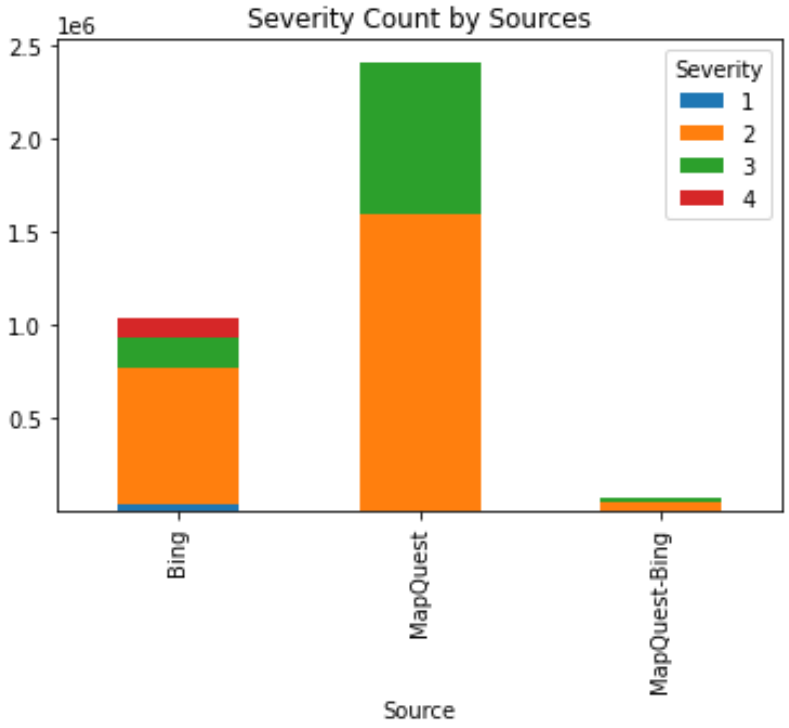
3.1 Data Pre-processing:

Main objective of this step is to get the pre-selected variable for machine learning. It includes the steps Exploratory Data Analysis, dealing with missing values, dropping features and converting the data types.

These data mainly come from two sources MapQuest and Bing. We are trying to understand the severity cases provided by each source. MapQuest reported less accidents with severity level 4 which cannot be seen in the plot itself, whereas Bing reported almost the same number of level 4 accidents as level 2. Meanwhile, MapQuest reported much more level 3 accidents than Bing in terms of proportion.

So, it is quite hard to choose one source so we are going with handling the data from Bing.

Figure 3: Accidents severity based on sources



3.2 Dropping Features:

TMC, Distance(mi), 'End_Time' (we have start time), 'Duration', 'End_Lat', and 'End_Lng' dropped as they collected after the accident only. ID, Description also not providing any useful information.

Country and the Turning_loop is having one unique value only. Also around 60% of data is missing for the Number and Wind_Chill features.

3.3 Modifying and adding Feature sets:

Converting Start_Time, End_Time and Weather_Timestamp to the real date time columns. Also simplying the Wind_Direction and Weather_Condition features to avoid complications.

Also splitting up the Start_Time feature into Day, Month, Year, Weekday, Hour and Minute.

All the POI and POD features data types are modified to Integer dtypes for our convenience.

3.4 Handling the missing values:

Dropping the few rows as they are less in count compared to the total sample values. City, Zip code, Airport_Code, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight and etc ..

3.5 Value Imputation:

Few of the columns have very small missing part which can be filled by median and mode. Some of the weather attributes are handled in this way. (Wind_Speed, Pressure, Temperature,Visibility, Wind_Direction, Weather_condition etc.

3.6 Sampling the Data:

Through the EDA above, we can clearly notice that the class distribution in this dataset is very imbalanced. This is due to the fact that the lowest and highest severity accidents don't occur as often as compared to other two severities, so we don't have adequate data for those classes. This means if we used the data in its existing condition then the model may never give predictions which have those probabilities.

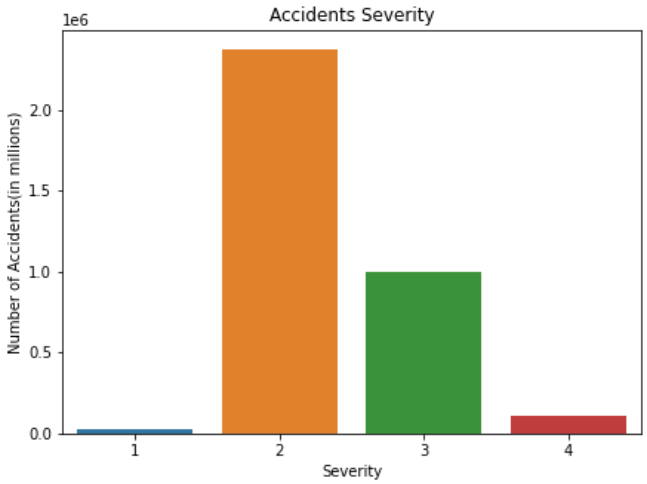
As per the previous Data Exploration, Severity 4 will be much serious than other severity lev els. So mainly focussing on the Severity 4 accidents and grouping other severity accidents. **Total values:** 1032119, **Severity 4:** 105167 **Other records:** 926952

The Data is unbalanced, it is difficult to do the analysis further. So resampling is done with the value around 100000.

4. Data Analysis and Visualization

We are performing the analysis of the available dataset using different features. Initially we are just identifying the how many accidents were happened on each severity which is the dependent variable and the values we have to predict using classification algorithms.

Figure 4: Severity Distribution



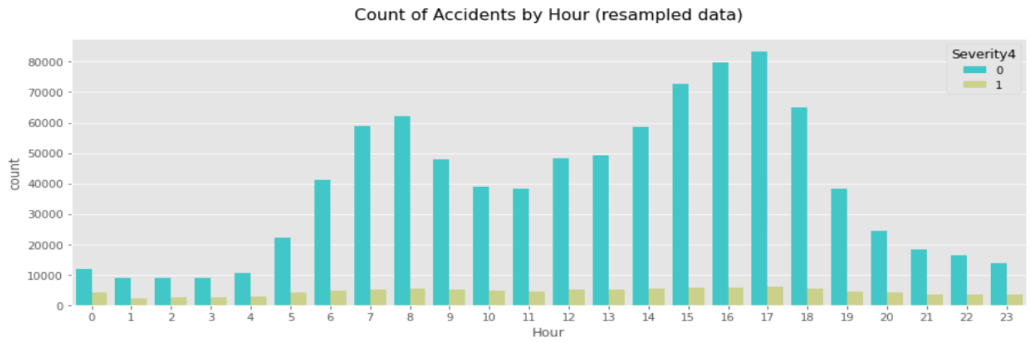
4.1 Visualization based on various Features:

Let us do the analysis based on the various features like Spatial, State, Time zone, Wind Direction, Hour, Weekday, Weather Conditions and Period of Day.

4.1.1 Hour

Most accidents happened during the daytime, especially AM peak and PM peak. When it comes to night, accidents were far less but more likely to be serious.

Figure 5: Count of Accidents based on Hour



4.1.2 Timezone:

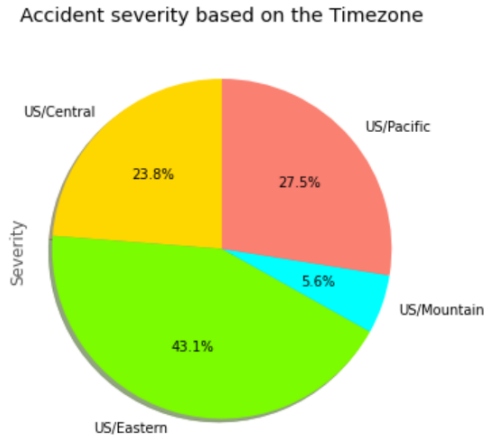


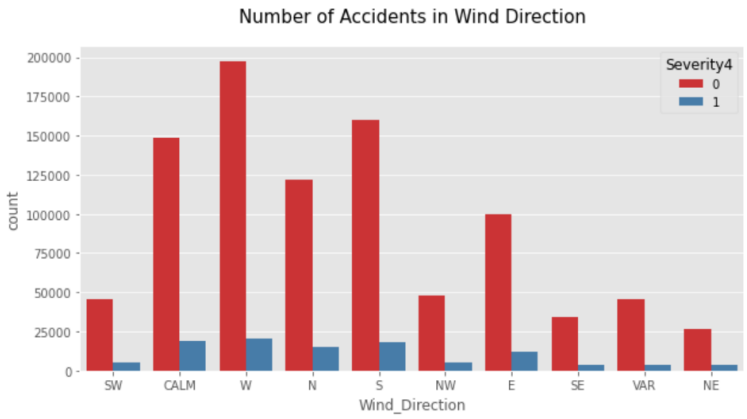
Figure 6: Severity based on Timezone

- Eastern time zone is the most dangerous one.

4.1.3 Wind Direction:

When the Wind is Calm and in West Direction severity is high.

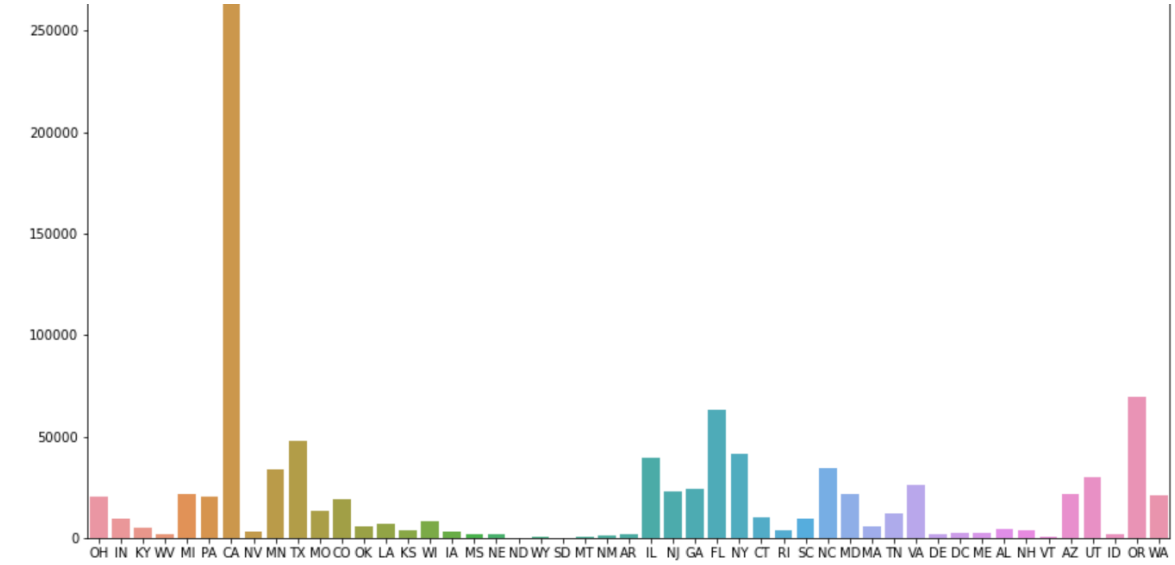
Figure 7: Count of Accidents based on Wind_Direction



4.1.4 State

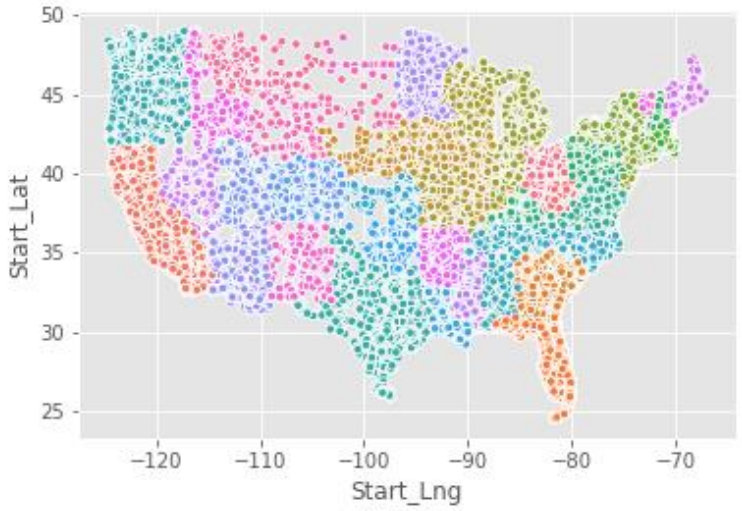
- FI, CA, and TX are the top 3 states with the most accidents.

Figure 8: Severity Distribution on States.



4.1.5 Spatial Features:

Figure 9: Distribution of Accidents based on the Spatial Features

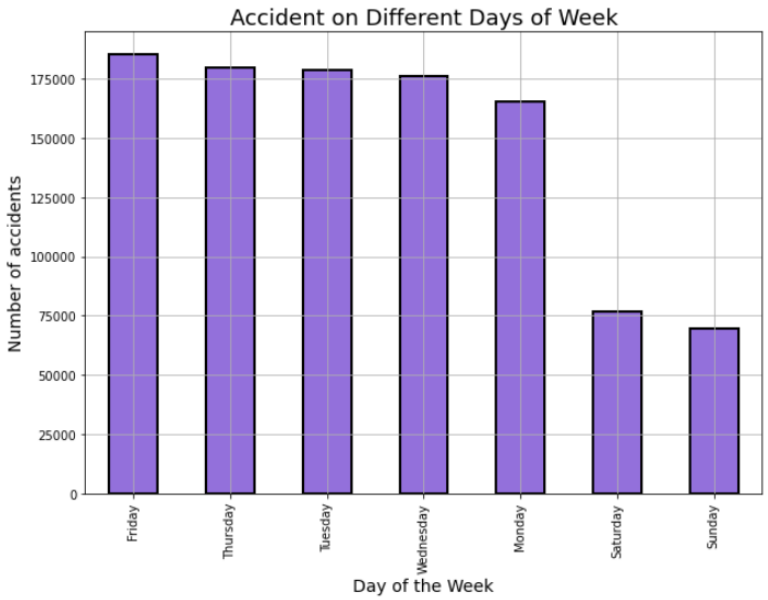


From the above graph we could see that Eastern area is highly affected by with the Accidents comparted with other zones.

4.1.6 Weekday

The number of accidents was much less on weekends while the proportion of level 4 accidents was higher.

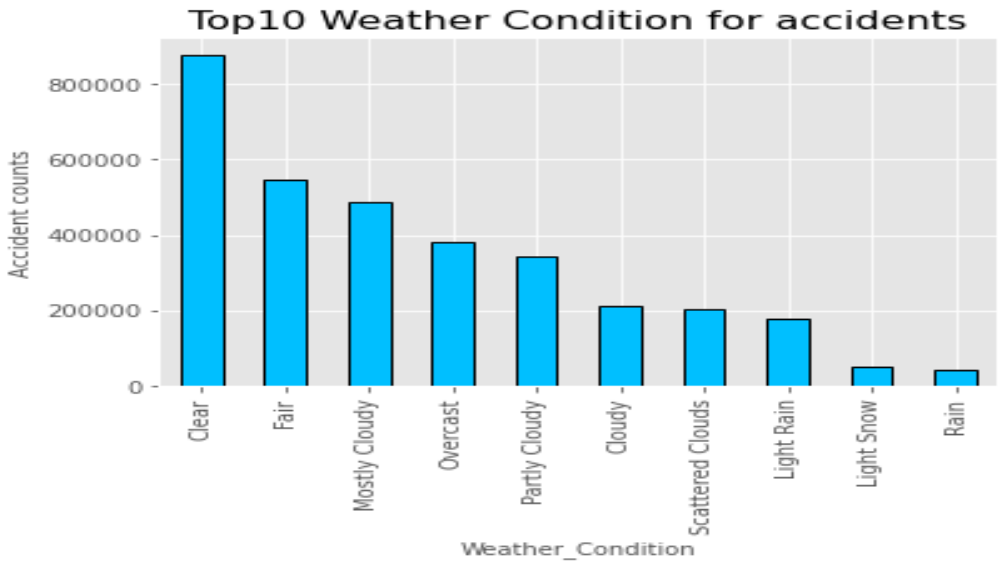
Figure 10: Accidents on Weekdays



4.1.7 Weather_Condition:

Most of the accidents occur Clear and Fair conditions.

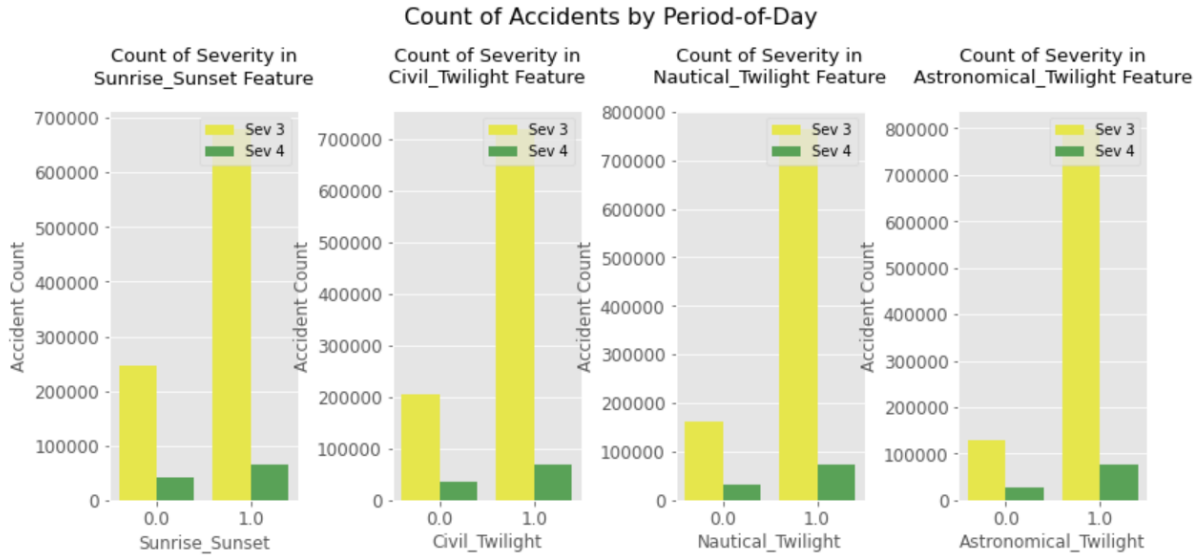
Figure 11: Highest Weather condition for Accidents



4.1.8 Period-of-Day

Accidents were less during the night but were more likely to be serious.

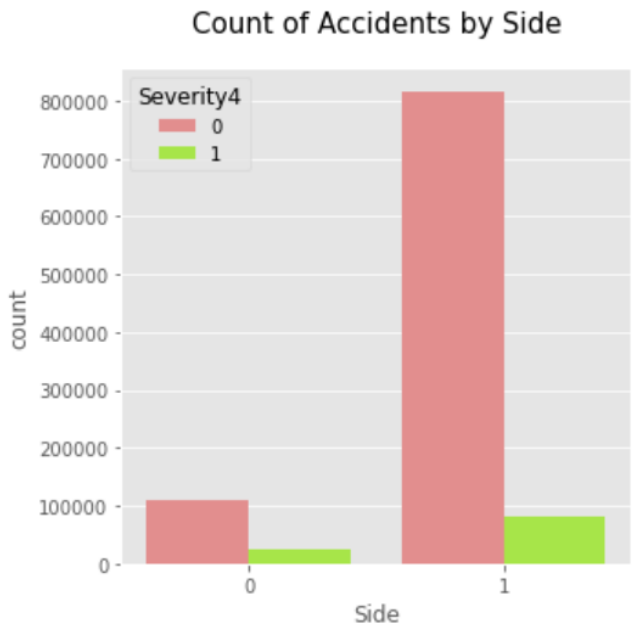
Figure 12: Accident severity by POD Features:



4.1.9 Side

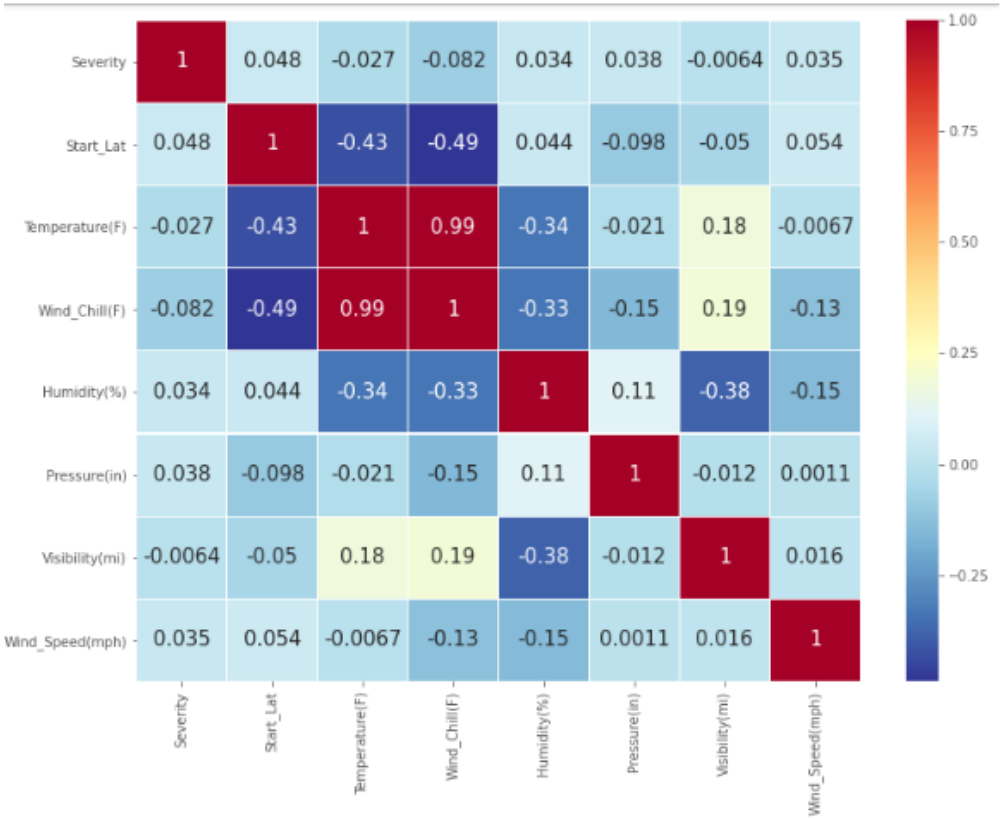
Right side of the line is much more dangerous than left side.

Figure 13: Count of Accidents based on side



4.2 Correlation of Weather Attributes:

Figure 14: Weather Feature Correlation



4.3 Impression from the Visualizations

Based on the above visualizations of US Accidents dataset after Exploratory Data Analysis highly impacted Features with values are:

- Weekdays other than weekends
- US Eastern Areas
- Right side of the Roads
- Daytime
- Clear and Fair cloud condition
- Wind is Calm and Western Side
- Evening Time accidents are severe.

5. Deployment Modelling and Evaluation

For the Modelling, we will be using supervised learning, which means learning with class labels are already given in the dataset. Based on the combination of all independent features in the dataset, classification algorithm will predict the severity of accidents which is a binary classification. Common classification algorithms which are used here is:

- K-Nearest Neighbors,
- Decision Tree
- Support Vector Machine
- Logistic Regression
- Random Forest Classifier

For Modelling and Evaluation, dataset will be split into Training and Testing subsets. The classification algorithm is trained to find the pattern that predicts the classes from the training subset, whereas testing subset performs the accuracy testing. We will use several classification algorithms for modelling and testing to determine the appropriate model for predicting Accidents severity of US.

5.1 K-Nearest Neighbors (KNN)

It is a simple classification algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

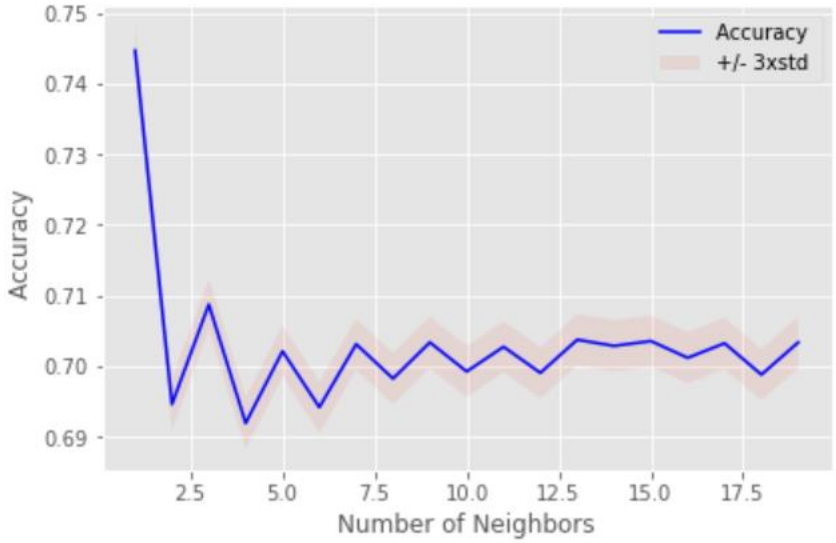
It works best when the dataset is balanced and its features are normalized. So we are building an accurate KNN model that determines the value of K, neighbours of comparison.

The best mean accuracy (0.745) is obtained in the initial value itself. (K = 1)

Train set Accuracy of KNN: 0.8515625

Test set Accuracy of KNN: 0.70875

Figure 15. K-Values with their Accuracy

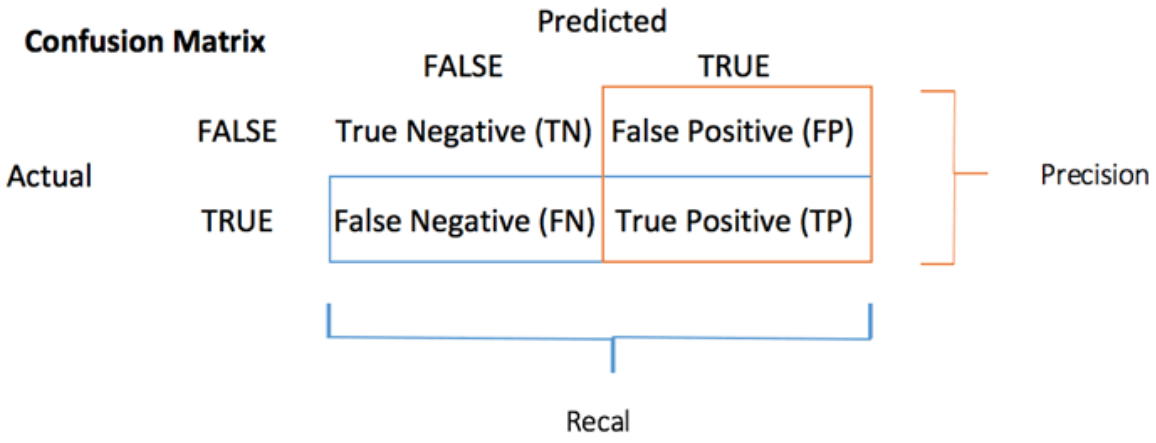


5.2 Decision Tree

Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decisions trees are the most powerful algorithms that falls under the category of supervised algorithms.

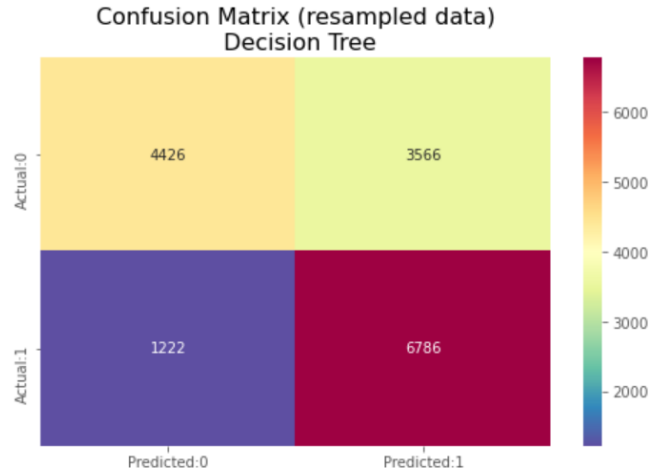
In Decision Trees, leaves represent the class labels and branches represent conjunctions of features. Entropy defines the amount of information disorder, if the node is completely homogeneous (i.e.Single class), then the entropy is 0. If it is heterogenous then the entropy is 1.

A **confusion matrix** is a performance measurement technique for Machine learning classification. It is a kind of table which helps you to the know the performance of the classification model on a set of test data for that the true values are known.



Usually the Decision Tree algorithm will run until all the leaves have become pure. Here the **general accuracy of score is 0.701** on the severity prediction of accidents.

Figure 16. Confusion matrix of Decision Tree



Train Accuracy: 0.78434
Test Accuracy: 0.7271

5.3 Support Vector Machine:

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms used for classification. They are extremely popular because of their ability to handle multiple continuous and categorical variables. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps −

- First, SVM will generate hyperplanes iteratively that segregates the classes in best way.
- Then, it will choose the hyperplane that separates the classes correctly.

Here it will produce the accuracy as below:

Train Accuracy: 0.71521
Test Accuracy: 0.704

5.4 Logistic Regression:

Logistic regression is used to predict the probability of a target variable. The nature of target or dependent variable would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

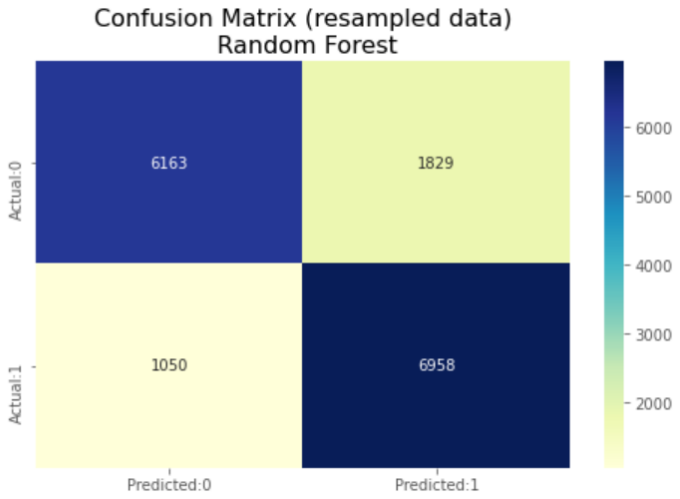
The simplest form of logistic regression is binary or binomial logistic regression in which the target or dependent variable can have only 2 possible types either 1 or 0. It allows us to model a relationship between multiple predictor variables and a binary/binomial target variable.

Train Accuracy: 0.670
Test Accuracy: 0. 669

5.5 Random Forest:

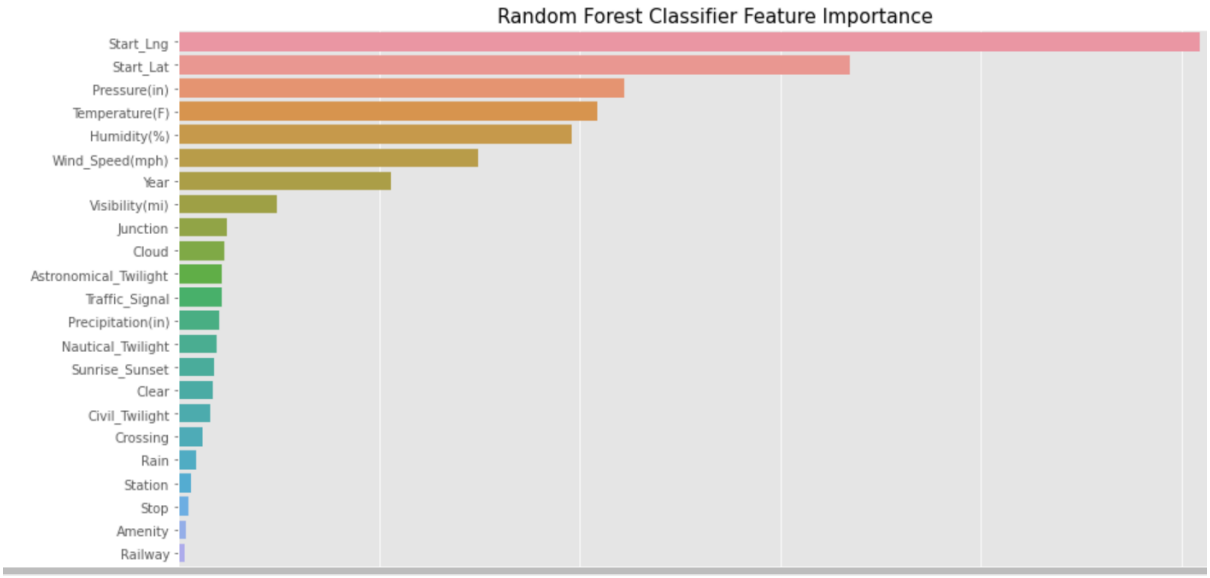
It improves the demerits of Decision Tree. Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Figure 17: Confusion matrix of Random Forest Classifier



Accuracy: 0.8200625

Figure 18. Feature importance of Random classifier

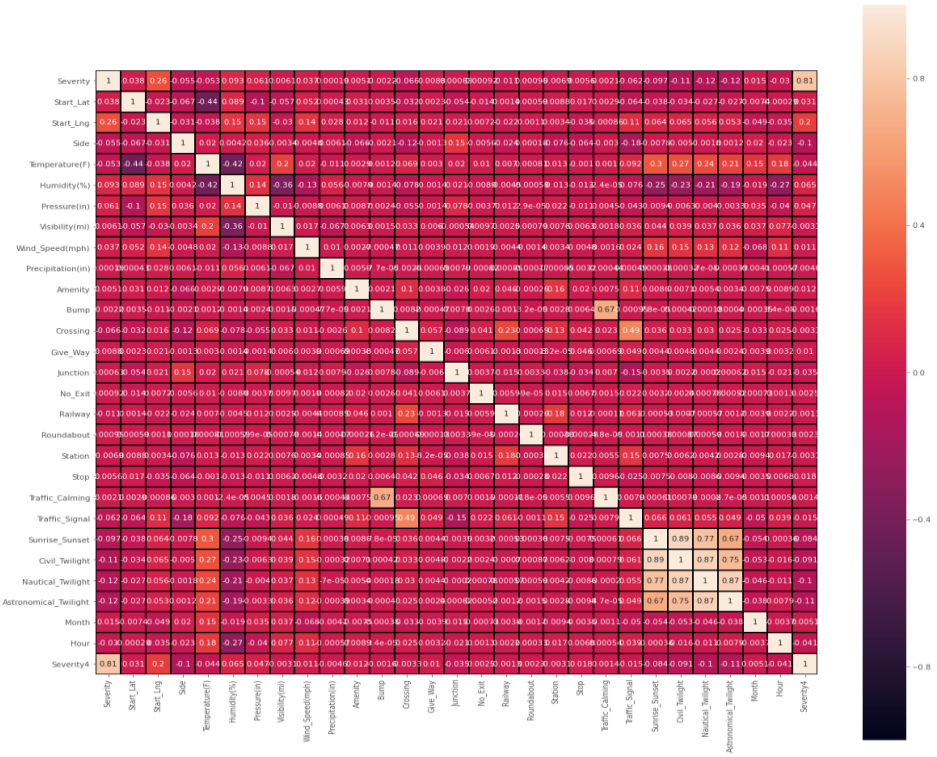


Classification Report

	precision	recall	f1-score	support
0	0.85	0.77	0.81	7992
1	0.79	0.87	0.83	8008
micro avg	0.82	0.82	0.82	16000
macro avg	0.82	0.82	0.82	16000
weighted avg	0.82	0.82	0.82	16000

Train Accuracy: 1.000
Test Accuracy: 0.8201

Figure : Correlation of Features



6. Conclusion:

6.1 Results

This section summarizes the findings on Accident Severity Prediction obtained from both EDA and the Data Analysis section of this report. Above analysis are clearly providing the counterintuitive answers to the questions. Our Analysis mainly demonstrate the top categories of the Weather conditions, Time Features, POI features shows that Clear, Calm, Traffic signal, Eastern areas and Sunlight are the ones with high accident occurring features even though the weather conditions were bad, Heavy Storm and visibility also poor.

The best evaluation is performed using Random Forest with high accuracy. Feature importance are demonstrated that would be helpful for Traffic securities to improving the Traffic safety. The best models for prediction the accidents are.

- Random Forest Classifier
- Decision Tree Classifier

6.2 Future Research and Development:

Even though the main objective of this report is to predict the accident severity model, the analysis which are performed are very helpful for improving the Traffic safety, imposing the traffic rules and regulate them.

- For future research we would be including as many features as possible by making the newly created instance from already existing one feature which would help in the predicting the accuracy high.
- Weighted XGBoost, Naïve Bayes and other similar models can be implemented instead of resampling the dataset.
- Detailed study of each feature and their correlation with dependent variable Severity will be done.

7. References:

1. https://smoosavi.org/datasets/us_accidents
2. https://medium.com/@vaibhavgope02/predicting-accident-severity-with-us-accidents-dataset-4aeaaae0b0af
3. https://www.kaggle.com/vtech6/basic-analysis-us-accidents
4. https://www.kaggle.com/deepakdeepu8978/how-severity-the-accidents-is