Prediction report on US-Accident

Contents

1 Introduction

1.1 Background

In recent years, road accidents have become a universal problem and are considered a significant cause of accidents. Road accidents put enormous impact on the economy of a country, public, society and environment. It is totally inappropriate and disheartening to see the people dying from road accidents. Almost every day, thousands of people lost their lives due to road accidents. It is observed that residential and shopping areas are riskier than less populated areas just because of higher exposure. Reducing the chance of road accidents is a very dominating and challenging topic of research. For handling this stagger issue, deep analysis of accident dataset is required. Accurate, huge and clean dataset of accidents is the basis of our prediction. Here we use some machine learning and deep learning models for analysis and prediction of road accidents. We also find out the relationship and effect of different variables on the occurrence of road accidents.

1.2 Problem

The overall problem is to control the road accidents, for which it is needed to find the severity of an accident. Data that might contribute to predict Severity of

accident include ID, Source, TMC, Start_Time, End_Time, Start_Lat, Start_Lng, End_Lat, End_Lng, Distance(mi), Description, Number, Street, Side, City, County, State, Zipcode, Country, Timezone, Airport_Code,Weather_Timestamp,Temperature(F),Wind_Chill(F),Humidity(%),Pressure(in),Visibility(mi),Wind_Direction,Wind_Speed(mph),Precipitation(in),Weather_Condition,Amenity,Bump,Crossing,Give_Way,Junction,No_Exit,Astronomical_Twilight,Traffic_Calming,Traffic_Signal,Civil_Twilight,Sunrise_Sunset,Nautical_Twilight, Stop, Station and predict Severity that show least and significant impact on traffic.

1.3 Interest

Almost every country is interested in predicting the severity of accidents for controlling the emergency. Accidents have a massive impact on the country's economy because there is an immense cost to fatalities and injuries. Crashes amount to approximately 1% GDP; in middle-income **countries** the cost is 1.5% of the GDP; and in high-income **countries** the cost is 2 percent of the GDP.

2 Data acquisition

2.1 Data sources

A New dataset with the name “US Accident” has been used here which includes almost 2.25 million instances of traffic accidents that took place within the contiguous United States1 between February 2016 and March 2019. This data can be found here. For creating this dataset, two important methods are used; streaming traffic reports and heterogeneous contextual data (weather, points-of-interests, etc.). so that the community may validate it, and with the belief that this process can itself serve as a model for dataset creation.

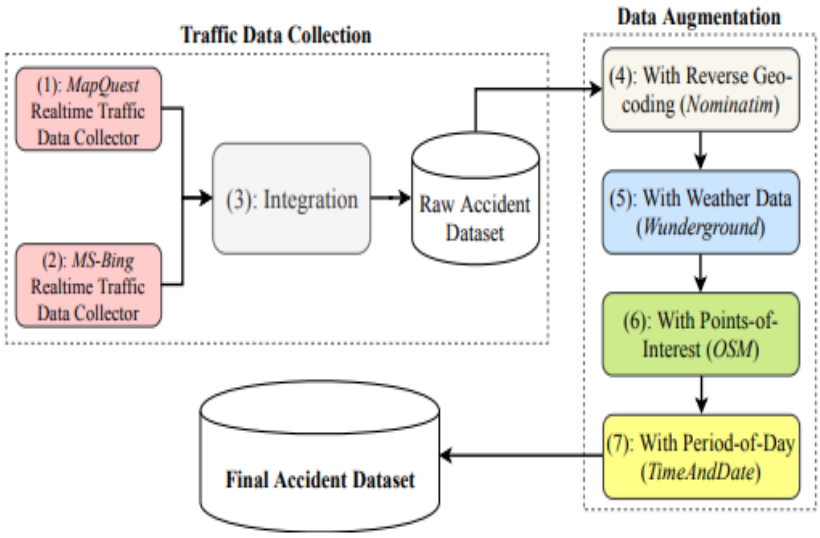

Figure 1: Data Collection

3 Exploratory Data Analysis

3.1 Accident relationship with state

Here we observe the relationship between the number of accidents and state. We observed that the highest number of accidents occurred at CA state. All other states have also been having high scores but almost 3 times less than CA. We also observed that SC and NC had almost similar number of accidents.

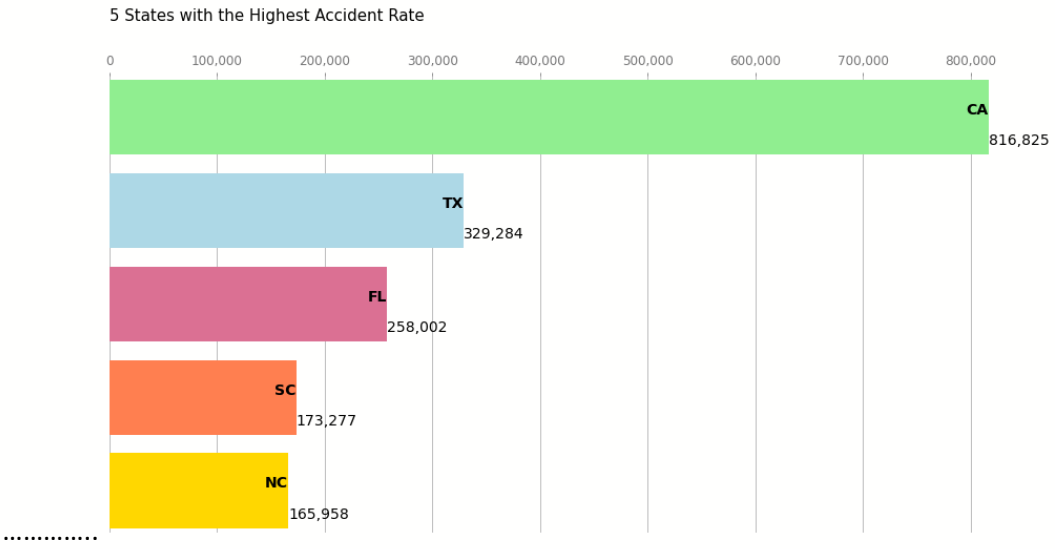

Figure 2: State Relation with Accident

3.2 Calculation of target variable

Severity is our target column that contains 4 types from 1 to 4, where 1 represents the least impact on traffic and 2 shows greater than least impact, 3 shows huge impact that is greater than 2 but less than 4 and severity 4 mean significant impact on traffic. We observed that almost 67.5% of records have severity 2, and severity 3 is 28%. In our prediction, all severity is important because we want to predict the dangerous situation for handling this in advance. Here severity 4 is only 3.2%.

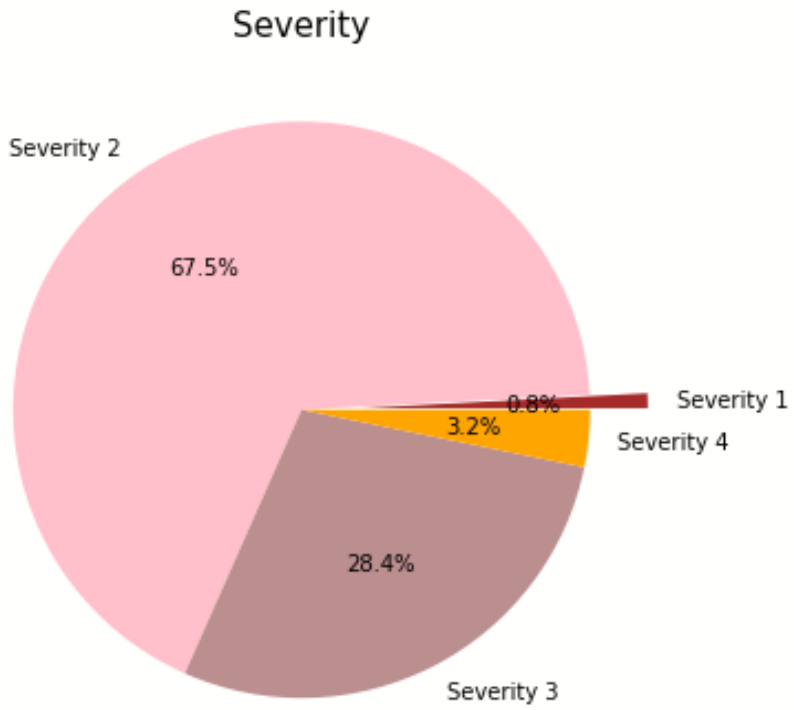

Figure 3:Target Column Analysis

3.3 Relationship between weather and Accident

Weather is the most important factor in accident prediction because mostly accidents occur in fog and cloudy weather. But according to our data visualization, a lot of accidents occur in clear weather (300000). Accidents that occur in clear weather hold some other reason. Similarly, most of the accidents occur in cloudy weather.

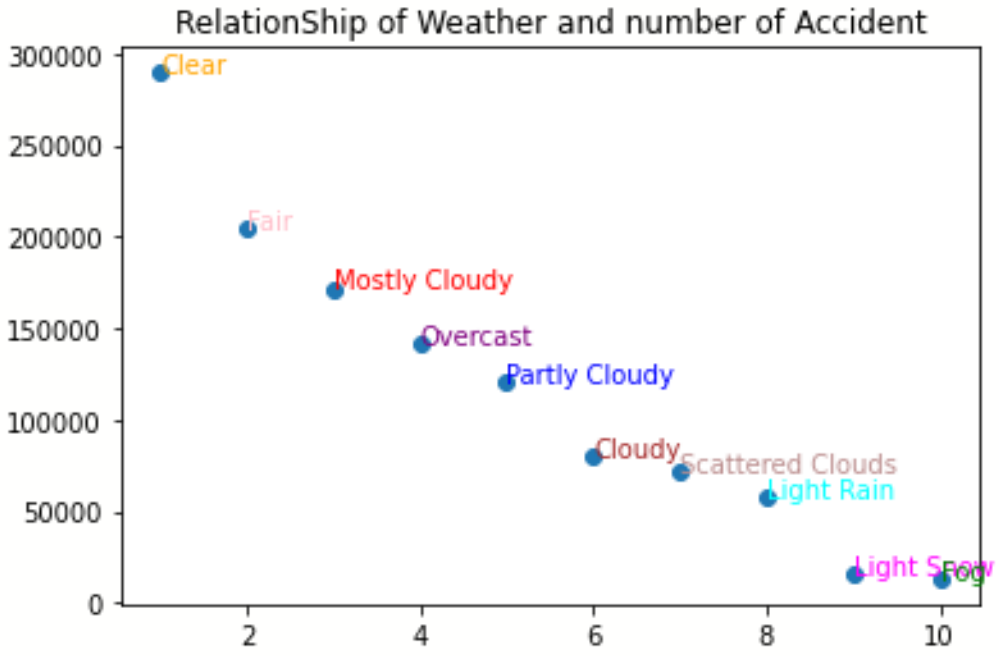

Figure 4: Weather Relation with Accident

3.4 Relationship between Number, State and severity

Here we plot three things in one plot, one is state that shows different states of US, second one is number of accidents occur in each state. Each vertical bar represents a different state. Height of each bar shows the number of accidents in the corresponding state. One more thing is two different colors in each bar, these colors divide the accident into two parts severity 1 and 4. Orange color for severity one and blue color for severity 4. We analyze that, almost each state has faced severity 4 accidents.

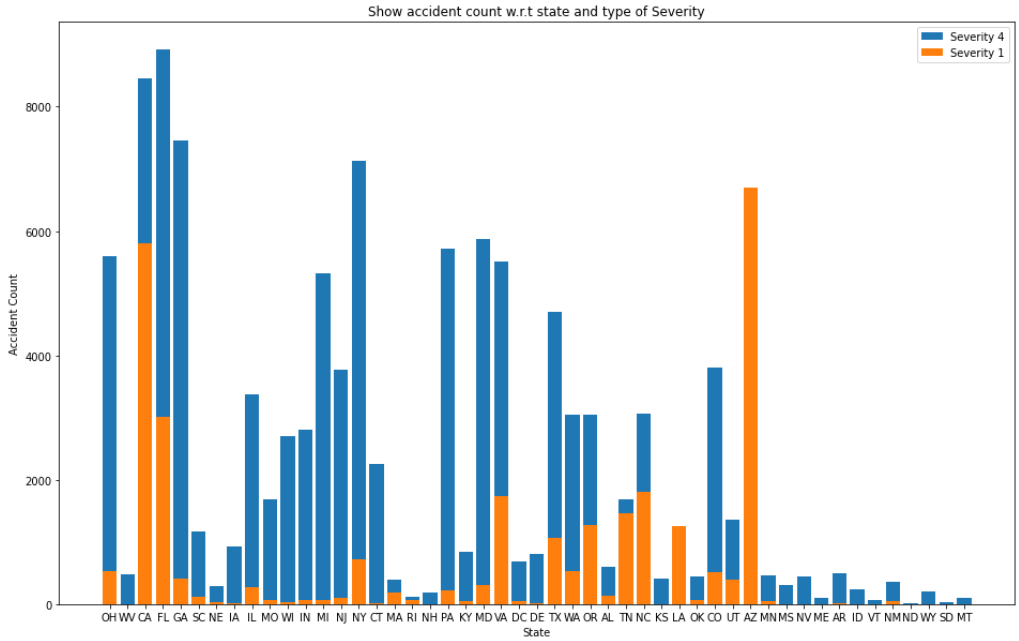

Figure 5: Severity & State & Accident

3.5 Correlation Matrix of Data

Here we try to find out the relation of our target column with multiple columns. In this correlation matrix, the diagonal cell shows correlation of similar columns. That's why we always found 1 at diagonal position. Here Dark color represents very poor relation and light color represents strong relation.

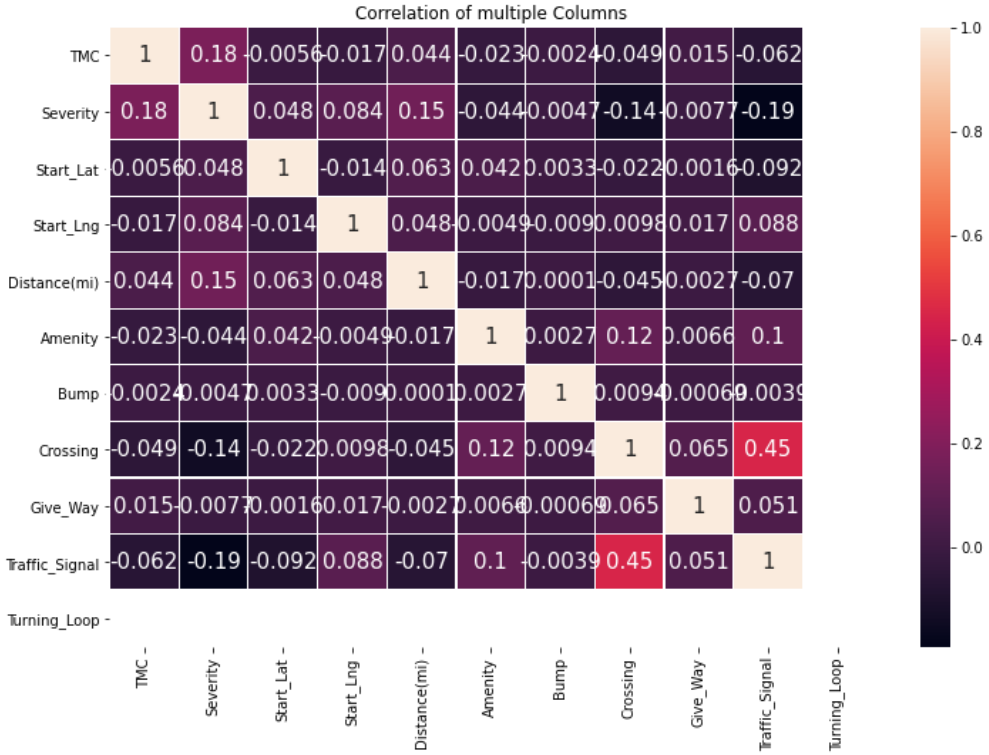

Figure 6: Correlation Matrix

3.6 Hour Wise Accidents Relation

Now we analyze the accident rate at each hour of a day. We observed that, according to our data most of the accident take place at 7 or 8 o'clock of morning.

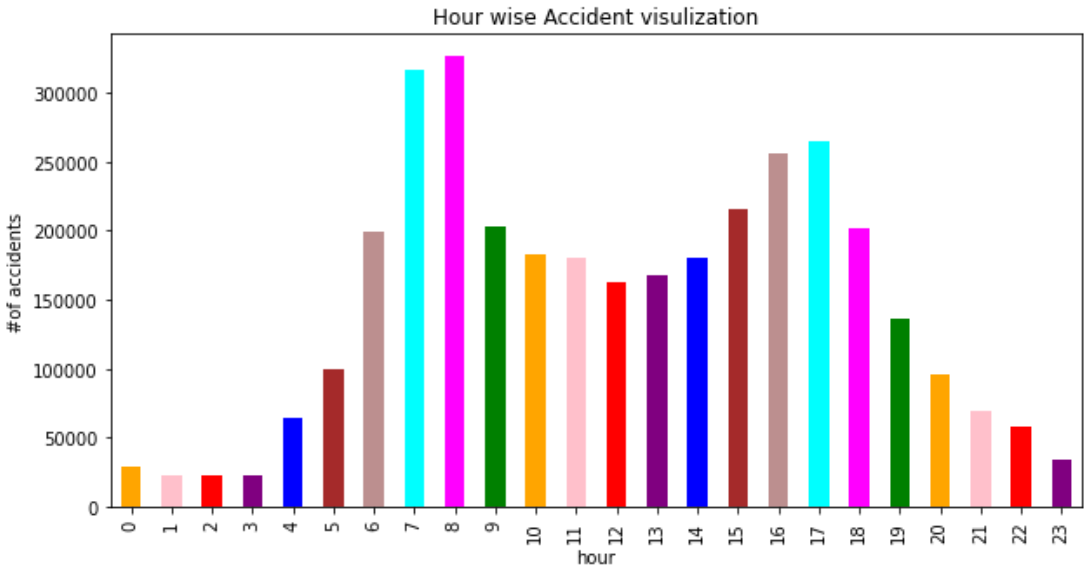

Figure 7: Hour wise Accident Analysis

3.7 Week Wise Accident Analysis

According to our data we insight that most of the accidents take place during working days because everyone is in a rush and tries to reach out to their destination as soon as possible. In most of the cases, this rushing became a reason of danger.

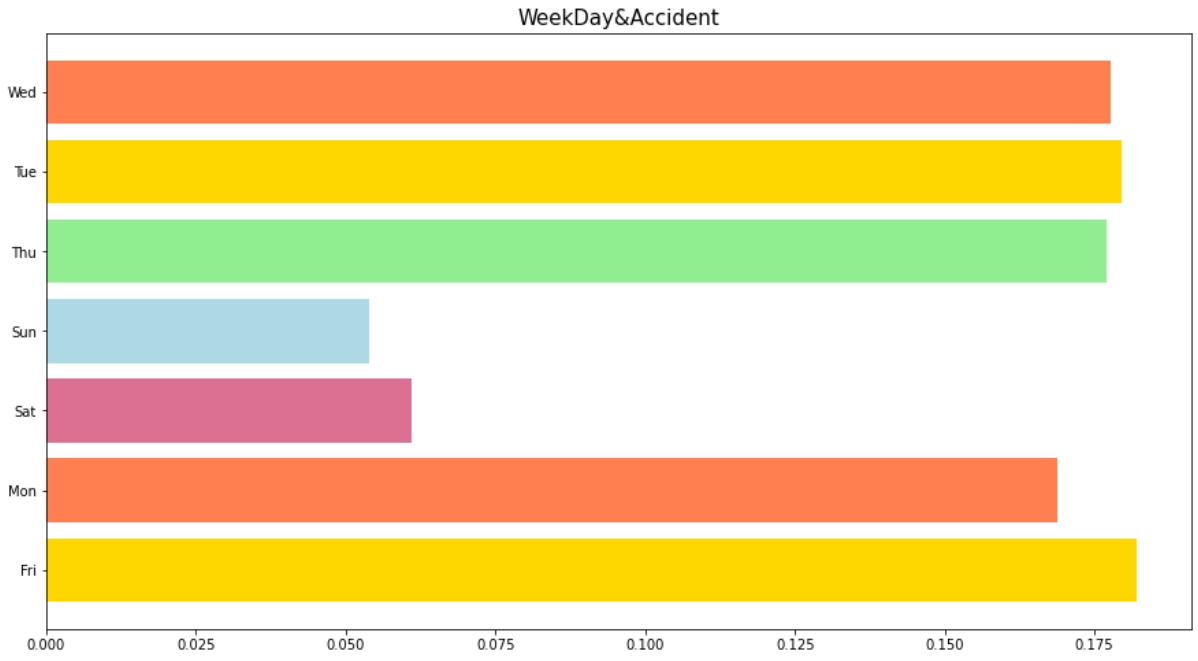

Figure 8: Week Wise Accident Analysis

3.8 Visualizing Wind Direction

Here we visualize the wind direction with its counts within the whole dataset. Here south, SSW, S, SSE, SE, SW are considered South direction. Same for all other directions.

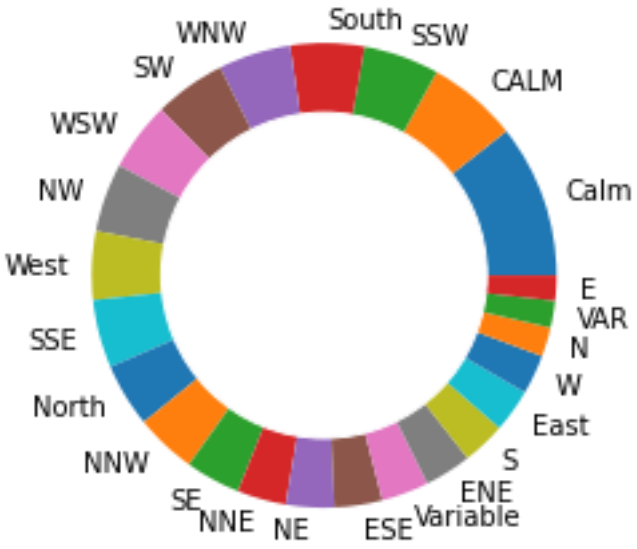

Figure 9: Wind Direction Visualization

4 Data Cleaning and Preprocessing

Data has been downloaded from a given source and contains many features that have a lot of missing values. First, we looked at the summation of missing values in our dataset. Few of the columns like 'Sunrise_Sunset','Civil_Twilight','Nautical_Twilight','Astronomical_Twilight','City','Zipcode','Airportcode' contain a small number of missing values. We simply dropped these rows because dropping does not put a huge impact on 2.25 million records.

We then looked at "Number", "Precipitation(in)" and "Windchill(F)" features contain almost 70% missing values. We simply dropped these features because these features do not have any high impact on the target column.

According to the dataset description TMC, "End lat", "End lag" and "Distance(mi)" all values are calculated after the accident occurred. That's why using those features for prediction of accidents is not very helpful. Thus we also got rid of these columns by dropping them.

Almost all features contain >=2 unique values but "country" is the same for all records because we perform prediction analysis on the US. That is why using this column for prediction is useless. We simply dropped this column, too, from whole records. One more column "Turning Loop" is the same for all records that indicate the presence of a turning loop. That's why we also deleted this feature from our dataset. Feature "Id" is just the representation of each record, having no relationship with prediction. So we simply ignored this feature during prediction.

Three columns that feed time as a string data type need to convert these string formats to data-time formats. After conversion of time, we observed that the difference between “weather time” and “start time” is zero. It means these two columns are almost similar. So we took only one of them.

As we know that, our dataset has been collected from two sources. One is **MapQuest** and the second is **Bing**. **MapQuest** sources give all accident data which belong to severity 3 and severity 2. There is no record that belongs to severity 1 and severity 4. On the other side Bing records belong to all four types of severities. As we observed that there is no similar proportion of **MapQuest** and **Bing** data collection. These differences show that both datasets are related to accidents, but they consider different definitions of severity. If we analyze source feature relation with different columns, we observe that **MapQuest** is given more accurate information rather than **Bing**. So, simply dropped rows that contained the source as **Bing**.

As some columns with numeric data type contained a lot of missing values. Firstly, we grouped those columns according to “Airport code” or “month”. Secondly, took **mean** of each group. At the end, missing values were replaced by the mean value of each group. But after this pre-processing, still we had some null values in these columns. We simply dropped Null rows but keeping in mind that dropping rows decrease the size of our dataset but not that much.

Similarly, “end time” is not a useful feature, because it is calculated after an accident. But “start time” is a very important feature for prediction. Unfortunately, Handling with dates is complex in machine learning. Due to this, we extracted hour, week, day, minute from start time and created 5 new columns for the representation of “start time”.

Moving towards categorical features, used **Label Encoder** and **get_dummies** technique for preprocessing Amenity, Bump, Crossing, Giveaway, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal, Side, State, Sunrise Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight features.

Divided the dataset into two parts, the training part and the test part. Severity being considered as a testing variable and all other remaining variables have been considered as Training data. After separating the data into two parts, applied **StandardScaler** machine learning technique that standardized the features by subtracting the estimation mean and scaled it to unit variance.

Finally, divided preprocessed data into train and test parts and used a different model for prediction.

Table 1: Feature Selection

Kept Feature	Dropped Feature	Reason for dropping feature
Crossing,Railway, Sunrise_Sunset	Country, turning loop	Same for all records
Traffic Signal, Junction, No Exit	Number, Precipitation(in) and Windchill(F)	Almost 70% records are null.
Temperature, Humidity	TMC, End lat, End lag and Distance(mi)	Calculate after accident occurrence, not useful for prediction.
Wind Speed, Weather condition,	weather time	Similar as start time
Side	Id	Just representation of Record

5 Predictive Modeling

There are two types of models that can be used to predict severity of accidents, one is regression and another one is classification. Regression models can provide additional information on the severity of accidents, while classification models focus on the probabilities of an accident belongs to severity. The underlying algorithms are similar between regression and classification models, but different audience may prefer one over the other. Therefore, in this study, I carried out both regression and classification modeling.

5.1 Regression Model

Regression is a form of predictive modeling, that is mostly used for extracting the relationship between input or target variable. There are a lot of regression models, but it is impossible to find out which one is superior to all other. That's why, I applied different regression models (LinearRegression, DecisionTreeRegressor, RandomForestRegressor) using MSE, MAE, RMS as the evaluation metric. All models used same training data and same test data so that there should be the standard criteria for comparing the performance of different models. In our target column, there are 4 types of values 1,2,3,4 that show the condition of severity. Although it is a classification problem, but we can solve this problem by using regression as well as classification.

Here in Table2, three regression models have been shown and analyzed based on different evaluation matrices. First, trained those models and compared those three types of errors.

Table 2: Training Error of Regression Models

	Mean Absolute Error	MSE	RMS
RandomForestRegressor	0.32	0.166	0.40
LinearRegression	0.38	0.18	0.43
DecisionTreeRegressor	7.815904571863096e-05	4.013811038597866e-05	0.006

After training, simply used trained models for prediction on test data and then analyzed the performance of those three models on unseen data as shown in Table 3 below.

Table 3: Test Error of Regression Models

	Mean Absolute Error	MSE	RMS
RandomForestRegressor	0.32	0.16	0.40
LinearRegression	0.38	0.18	0.43
DecisionTreeRegressor	0.12	0.13	0.36

Finally, it was concluded that **Decision Tree Regressor** was the best performing model among all the other models included in this scenario/comparison and had 3 times less error than all the other one's.

5.2 Classification models

Similarly, there are a lot of classification models but it is not possible to find out that the one is superior to all other in each and every scenario. That's why, I applied different classification model (Logistic Regression, Decision Tree, Random Forest, Ada Boost, Support Vector) using classifier score as the evaluation metric. All models used the same training data and same test data, for keeping the standard criteria for comparing the performance of different models. Classification model are more straight forward. In this scenario, divided the samples into 4 classes such as severity1, severity2, severity3 and severity4. Thus the samples that were close to each other, were considered as member of the same class. In classification problem, probability of the model decides that the current record is belong to

which one class. At the end, the best one model was selected based on their respective Training and testing Score.

6 Performance of different models

Table 4: Comparison of Different Classification Models

	Logistic Regression	Decision Tree Classifier	Ada Boost Classifier	KN Classifier
Train Score	0.70%	0.99%	0.72%	0.65%
Test Score	0.70%	0.87%	0.72%	0.67%
True Positive	0,264476, 69343,3	0,282508,128657, 77	0,260679, 82780,6	0,202884 ,30819,3 32
False Positive	0,90357,48137,12	252,29430,29962,1 442	0,76710,52149,4	0,90188, 31182,6 27
True Negativ e	472129, 69882, 265472, 470995	471865, 130821, 283570, 469579	472129,83529, 261460, 471003	355905, 33979, 211718, 344497
False Negativ e	199,47613,89,376,1 318	53,29581,30062,12 44	199,51410,75939,1 315	127,2898 1,82313, 10576

Here we observed the performance of different classification models. The best performing model has been labeled as red. In above Table 4, only three values are marked as red, which belong to Decision Tree classification. Thus decision tree classifier achieved almost 100% at train time and 87% at test time.

I also evaluated the models using their ROC curves between **False Positive** and **True Positive** rate. **ROC** is a probability **curve** and AUC represent degree or measure of separability. It **tells** how much model is capable of distinguishing between classes. Each model graph contains four lines, each corresponding line belong to each severity.

6.1 ROC Curve of Logistic Classifier

In this graph, each line represents severity level. This model performed well but did not differentiate the 4 classes in a better way.

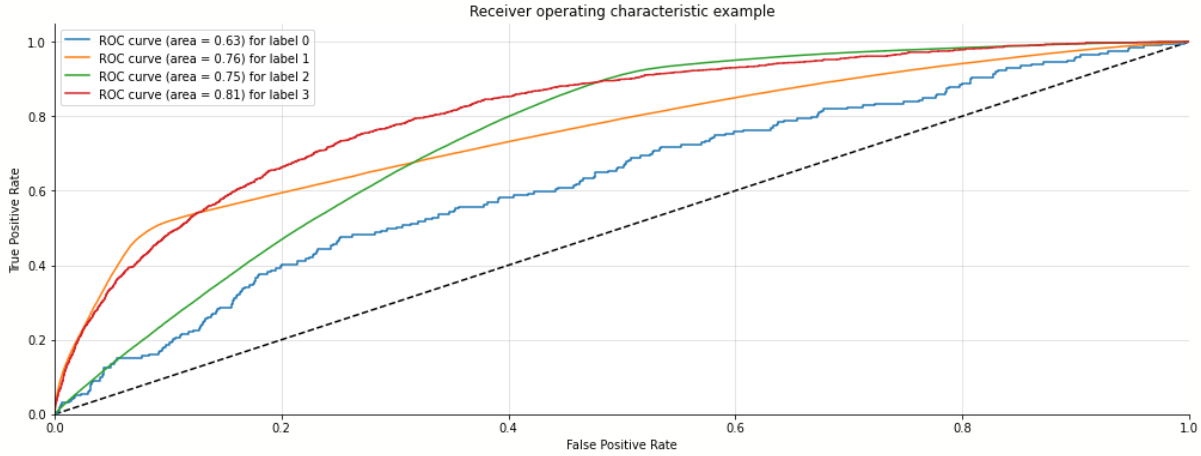

Figure 10: ROC Curve of Logistic Classifier

6.2 ROC Curve of Ada Boost Classifier

Ada Boost Classifier ROC curve is almost similar as logistic classifier and it's performance has also been proved by analyzing the accuracy score of both the models. Logistic classifier's accuracy score is 70% and Ada boost classifier's accuracy is 72%.

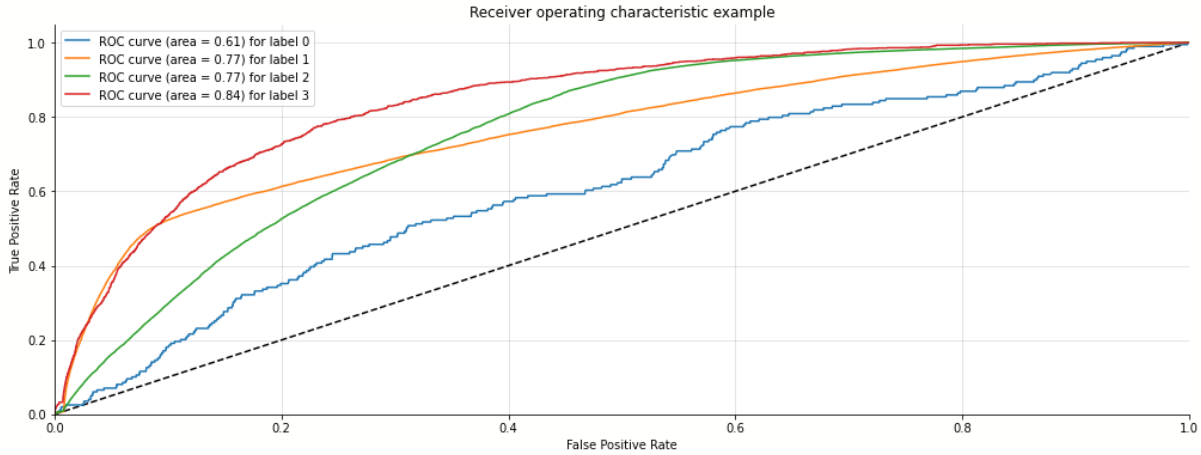

Figure 11:Ada Boost Classifier

6.3 ROC curve of Decision Tree Classifier

Here has been discussed the separability measure of our best classifier Decision Tree. This classifier creates two groups. One group is a combination of severity 1 and severity 4. Second group is a combination of severity 2 and severity 3. Severity proportion in our dataset has already been shown in Figure 3 above. Proportion of severity 1 and severity 4 is somehow similar to each other while proportion of severity 2 and severity 3 is similar to each other. As we know that size of records

that belong to severity 2 and 3 is very high w.r.t severity 1 and severity 4 and thus this model classified correctly.

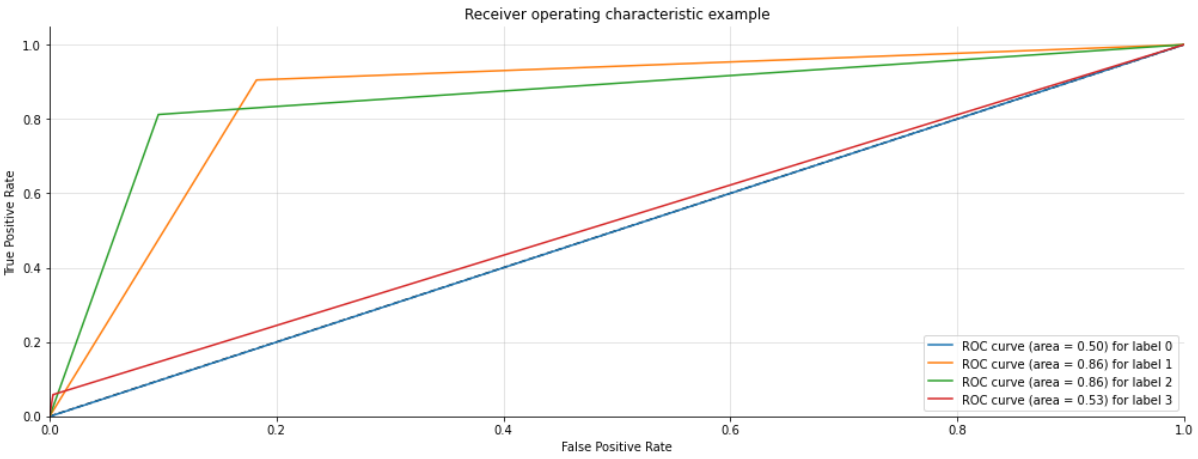

Figure 12: ROC Curve of Decision Tree Classifier

7 Conclusions

As per the requirements, all the tasks have been performed properly. After leniently visualizing and preprocessing the dataset, severity was predicted using different prediction models including regression and classification models. Based on the above calculations, it has been concluded that the Decision Tree Regressor and Classifier performed well in this scenario. And achieved almost 100% accuracy score while using as a classifier. Thus severity measure can easily and accurately be predicted using Decision Tree Classifier.

8 Future directions

Severity prediction of accidents can help a lot towards safety solutions not just in case of road accidents but almost in each and every dangerous/challenging situation. And can be implemented as an important decreasing factor in many accident avoiding systems.