Counter-Strike

CE 784A Project Report: Semester 2021-22 (II)

Investigation of Machine Learning Models for Traffic Accident Analysis

Raj Kumar Yadav(190675) , Anshul Gautam(21103019)

Abstract

Road accident duration is the time between when an accident was first reported to the authorities and the time till all the obstruction to traffic on the road is completely removed and the traffic is back to normal. Reducing traffic accidents is a public safety concern all over the world. Traffic accidents are a huge risk to the safety of a large number of people on the road and the numerical value of the number of vehicle on the roads makes preventing these accidents one of the most challenging tasks and although the solutions to prevent these accidents depends largely on the road engineering and safety measure build to reduce and prevent them, accident duration and their severity analysis help in taking counter-measures to reduce the loss of life and property. Road incident duration prediction helps in preparation to reduce the severity of these accidents. These prediction models can also be used in traffic congestion prediction and also in automated or self-driving vehicles.

Keywords: Road safety; accident duration; Traffic accident analysis.

1. Introduction

In this project, we are doing a comparative analysis of different machine learning paradigms appropriate to estimate the accident duration on a dataset of car accidents collected in the U.S (49 states). The accident data was collected between February 2016 to Dec 2021. We will also examine the effect of varied factors that influence traffic accidents. like traffic calming, traffic signals, turning loops, weather, wind, humidity, precipitation, and their effect on accidents occurrence frequency and accident duration are being studied. Accident duration is the time difference between two time zones Start_Time & End_Time, Where Start_Time is the time zone when an accident occurs and End_Time is the time zone when the effect of the accident on traffic nullifies.

In this analysis we'll discuss and compare the various Accident duration prediction model some of them are:

- Regression model
- Decision Tree
- Deep Neural Network

We will be comparing the accuracy of these models using various error terms such as MAE, RMSE, etc. Along with this, we are also analyzing the severity and various factors that can have an impact on accidents such as environmental conditions, road conditions, temperature, etc.

2. Literature Review

In the past, many models had been developed to analyze accident duration for which different types of paradigms have been applied ranging from supervised to unsupervised algorithms. In the Accident duration model, we point out the relationship between accident duration and important influencing factors. But the comparative study of different models is quite challenging as the dataset used to build the models shows different characteristics. Some of the models which have been developed in past to analyze the accident duration are the following:

Bryan and Janssen (1991) [1] Purpose of this paper was to show the importance of techniques such as statistical play in accident management strategies. The time of the incident in this paper was defined from the point, the relevant in-charge receives the information of the incident till the point he/she leaves the place of the incident. Seattle Metropolitan was chosen as the area of the study and possession distribution was used for estimating the frequency of incidents and conditional probability formed the basis of the analysis as chances of accident ending in 12 sec were calculated given that it continued for 11 sec. Seasonal, weekly, special, and environmental effects were taken into the account during the study.

Yuan Wen, Qin Yuan Xiong (2012) [2] have used the KNN to analyze the accident duration by optimizing the most effective k value. And this model using the 1853 traffic accident duration dataset was taken from the Ministry of Transportation, Netherland. In this research, at first, the dataset was classified, then the distance function was modified and finally, prediction results from putting different values of k in the KNN Algorithm were compared. At last, an experiment was performed using the dataset to predict accident duration by the KNN paradigm and its outputs were properly analyzed. Also, many values of K were taken to figure out what is that "thing" that influences the K on the prediction of accuracy, and in the end, an optimal value for K for the experiment was found. Thereafter, using the result we have done an error analysis. And therefore, as a result, this method obtains higher accuracy as compared to the Decision Tree Algorithm.

Xuanqiang Wang, Shuyan Chen, Wenchang Zheng (2013) [1] has Analyzed several regression methods, PCR & PCSR, and built models between accident duration & affecting factors. In PCR principal component analysis is used to estimate the regression coefficient, all inputs were provided to the regression model. When the sample size is not enough in quantity and the dataset is highly collinear then the PLSR method is used in developing predictive models.

Different models were established for different types of incidents like stopped vehicles, accidents, and lost loads with the effectiveness of these different models respectively being 83.6%, 88.2%, and 92.7%. The maximum accuracy of the different models was seen at 20minutes of error. The analysis concludes that regression models are good in large cases for accident duration prediction.

According to previous research, it can be established that every method has its pros and cons, thus no method works for all, and the use of different methods is recommended according to the problem. To cover the complete duration of incidents, a mixture of different methods is recommended. This is the main objective of this endeavor to analyze the results of different accident duration prediction models and different influential factors

3. Data Preprocessing and Cleaning

First, we changed the data type of start time and end time to timestamps, then the duration of the incident was determined by taking the difference between start-time and end-time, then the unit of duration from timestamp format to minutes

The main objective of this project is to predict the accident duration through features that significantly affect the duration of the accident and can be measured and recorded easily when an accident occurs.Many features that are

obtained after an accident has occurred such as Severity and description of the accident are of low significance as they play an insignificant role in the prediction of the duration of the accident and are hence being dropped.

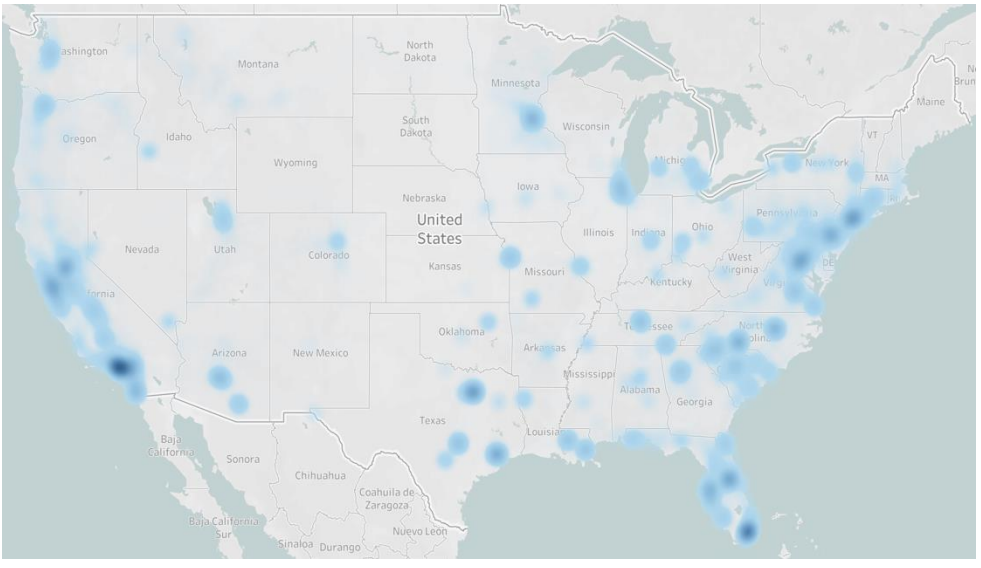
Initially, we have 2845342 accident data and 47 features of each accident, and our dataset contained outliers due to which there was a high level of asymmetry in the dataset, the minimum value of duration was 2 min while the maximum was 1682579 minutes which indicates that accident duration lasts form minutes to years which is practically not feasible to account for prediction of accident duration, so such outliers were removed(especially the data points with high accident duration) by interquartile range (IQR) method. IQR method does not assume any specific distribution of a dataset. In this method, we found lower and upper limits of accident duration, and the accident duration not in the range of upper and lower limits was removed. Our lower limit was calculated using Q1 - 1.5*IQR and upper limit Q3 + 1.5*IQR where Q1 is the 25^{th} quantile and Q3 is the 75^{th} quantile and IQR is a difference between Q3 and Q1.

So, after doing outliers removal our dataset has a minimum duration of 2 minutes and a maximum of 448 minutes.

3.1 Missing value

Although the majority of the columns had a few missing data points there were some columns with a large number of data points missing, hence columns with a missing value greater than 15 % were removed.
We were getting greater than 61 % missing value for 'Number' features and greater than 16 % for 'Wind_Chill(F)' and greater than 19 % for 'Precipitation' so we have dropped these features from our dataset

4. Analysis of Traffic Accident

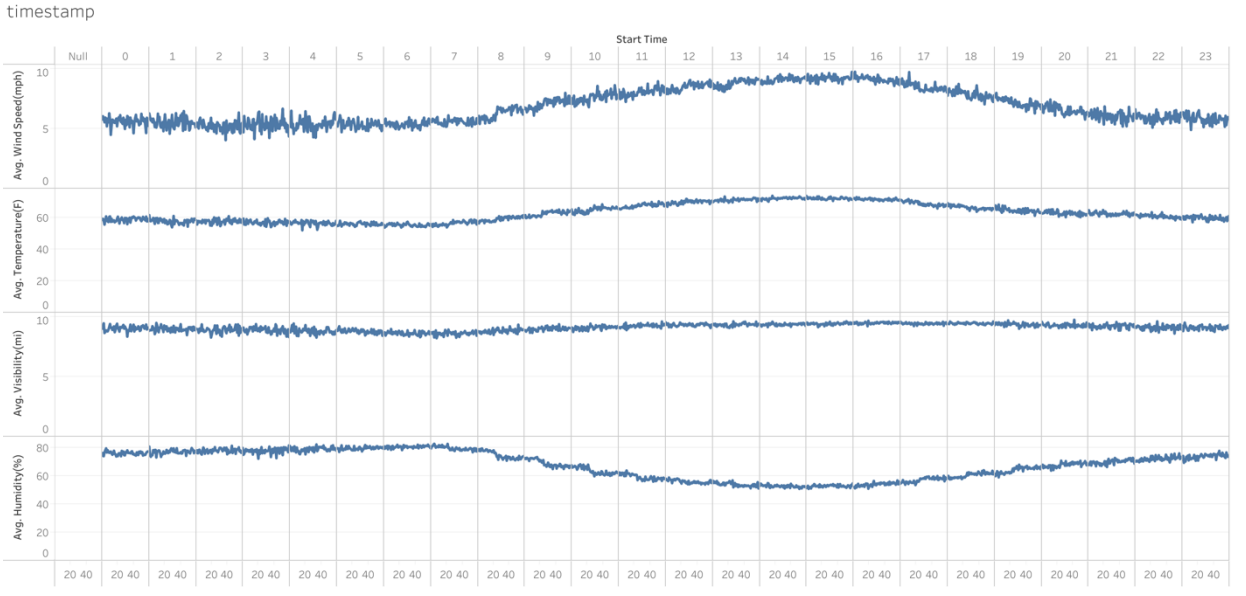


Map: 1 Shows density of distribution of accident across the US.

Max, Accidents are happening near the coastal area of US, this distribution also follows the population distribution of United States. As the number of accidents increases with the increases in population of the area.

4.1 Accident Data analysis

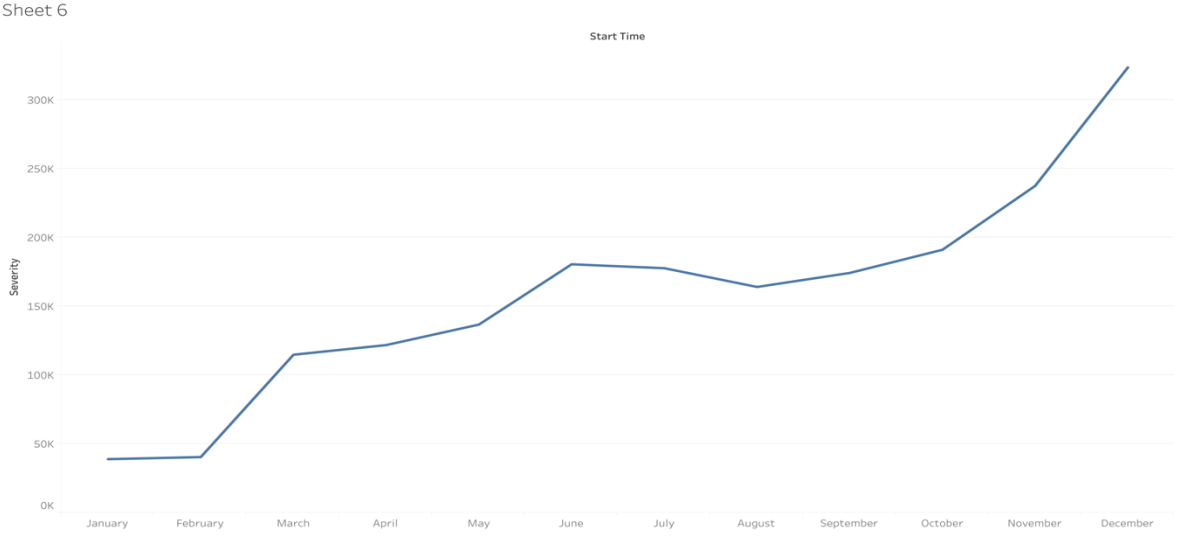
4.1.1. Analysis of weather parameters w.r.t to Time.



Graph 1

Here as shown graph 1 above the weather parameter such as wind speed, temperature, visibility and humidity trends are shown. They follow the normal pattern i.e., is Wind speed is max during the mid-day, so is the temperature, Visibility remains constant here as it is averaged over years data. Humidity fall after the mid-day as the temperature rises.

Analysis of cumulative severity of accidents



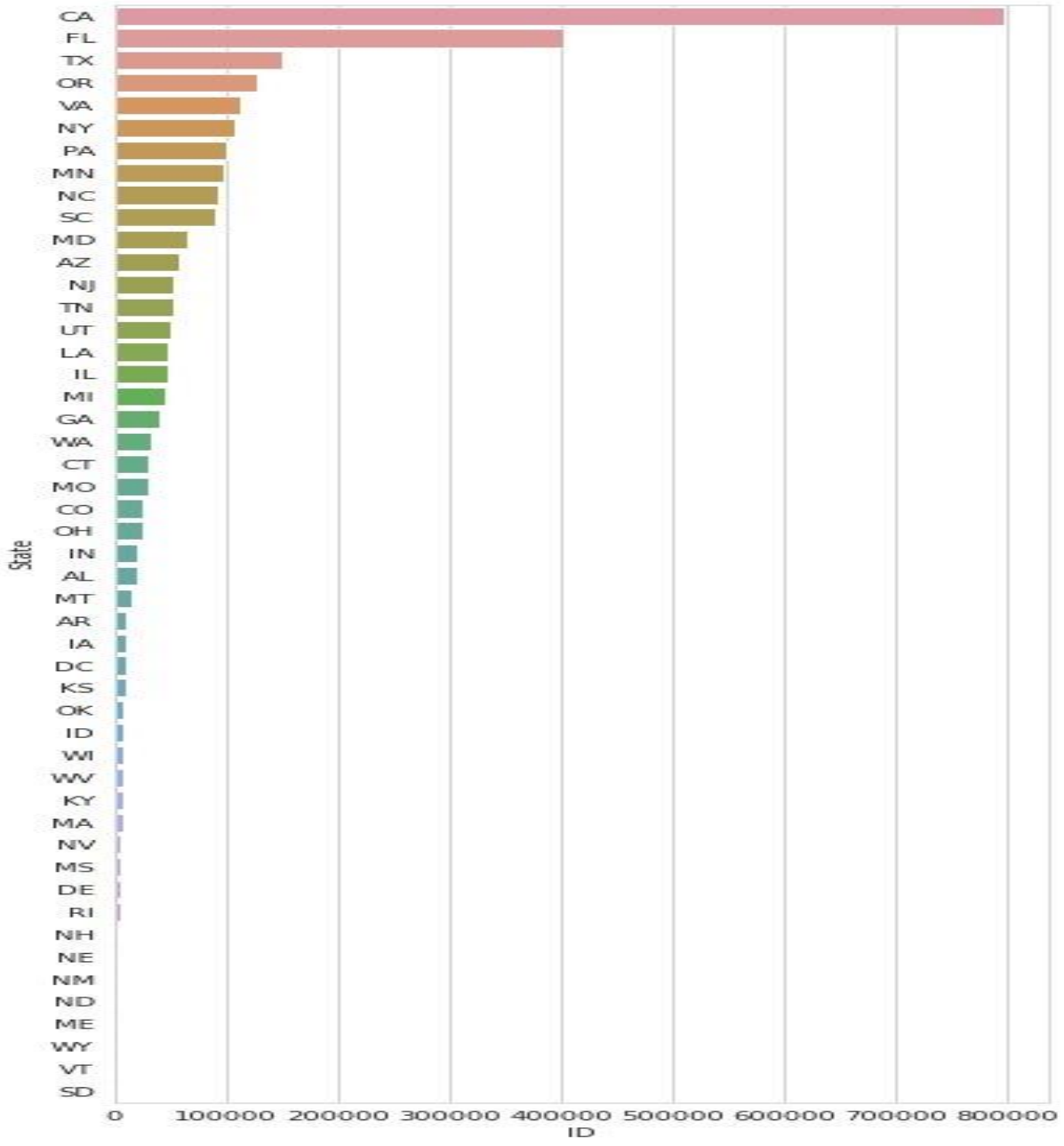
Graph 2

As visible from the graph 2 above the cumulative severity of accidents increases sharply in month of February, November and December. This increase is attributed to increased travel during the holiday season

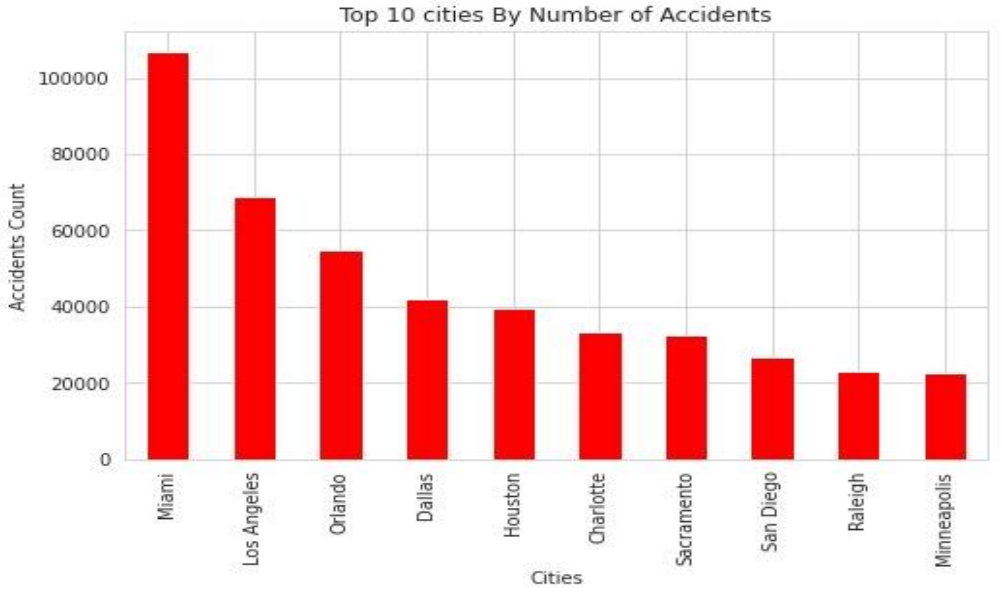
4.1.2. State wise analysis

By analyzing the dataset, we found that State of "California" has the highest number of accidents and is followed by Florida and Texas

Graph :3



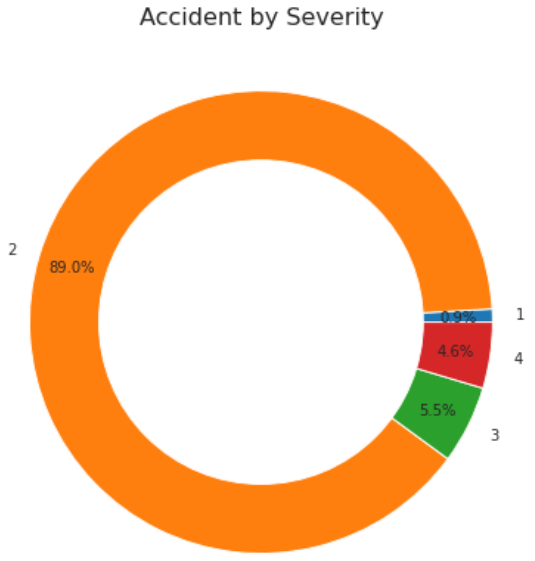
4.1.2. City wise analysis:



Graph: 4

Graph 4 Shows Miami City has the highest number of accident when compared to all major cities in the country. Miami is followed by Los Angeles and Orlando as the cities with 2nd and 3ed highest accidents in US.

4.1.3. Severity Analysis



Graph: 5

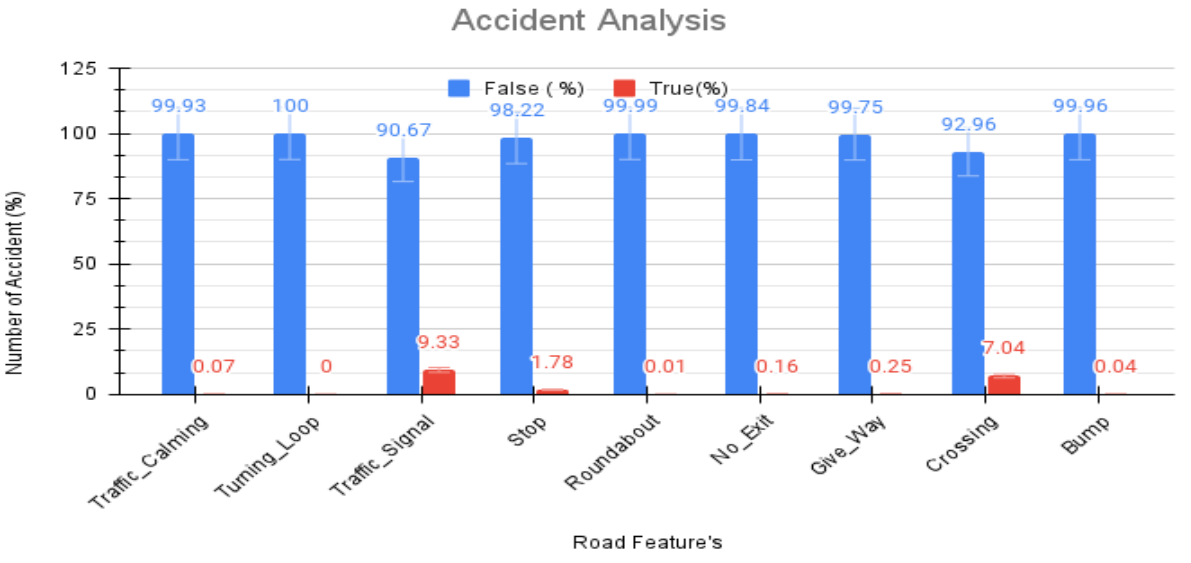
1.89 % of Accidents falls under the category of Severity level 2.

2. 5.5 % of Accidents falls under the category of Severity level 3.

3. 4.6 % of Accidents falls under the category of Severity level 4.

4. 0.9 % of Accidents falls under the category of Severity level 1.

4.1.4. Effect of Road Features on Accidents



Graph: 6

a. Traffic calming: These are the measure put in place to reduce the traffic, and the graph shows that when traffic calming measure are put in place number of accidents are less, but they increase by significant amount when these measures are not in place.

b. Traffic signals, crossing, bumps, stop, give_way, Roundabout, turning loop, no_exit are all measure to enhance the smooth traffic flow and reduce the traffic disruption and prevent traffic accidents as shown in the graph 6 above when these measures are not in place the number of accidents increases drastically.

(Although this data has bias as the traffic flow characteristics are not available.)

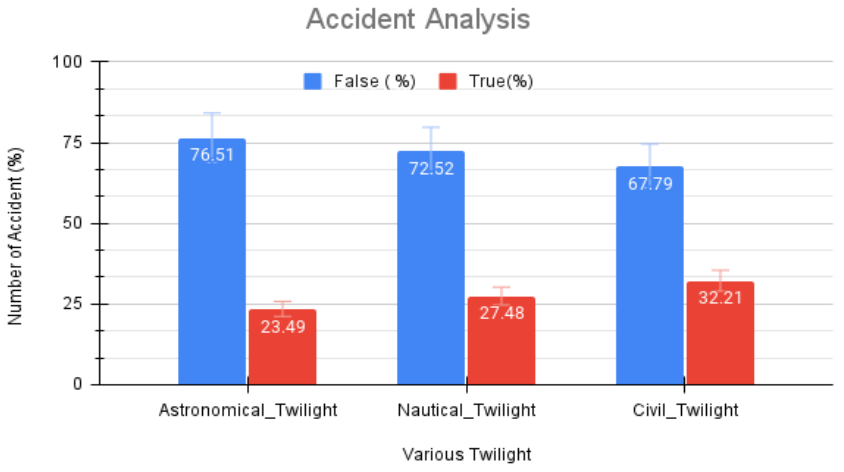
4.1.4. Effect of Various Twilight Condition:

Twilight is the visibility in the lower atmosphere of the earth due to scattering of light by upper atmosphere when sun is below the horizon and is not directly visible. Each twilight position happens twice a day i.e., before sunrise and after sunset.

Depending on the position of sun below the horizon twilight is divided in three parts and visibility is affected by these twilights creating a confusion as to turn on the headlights on or off.

Civil Twilight: Is the period when geometric centre of the sun is between horizon and 6º below it.
Nautical Twilight: Is the period when geometric centre of the sun is between 6º and 12º below horizon it.

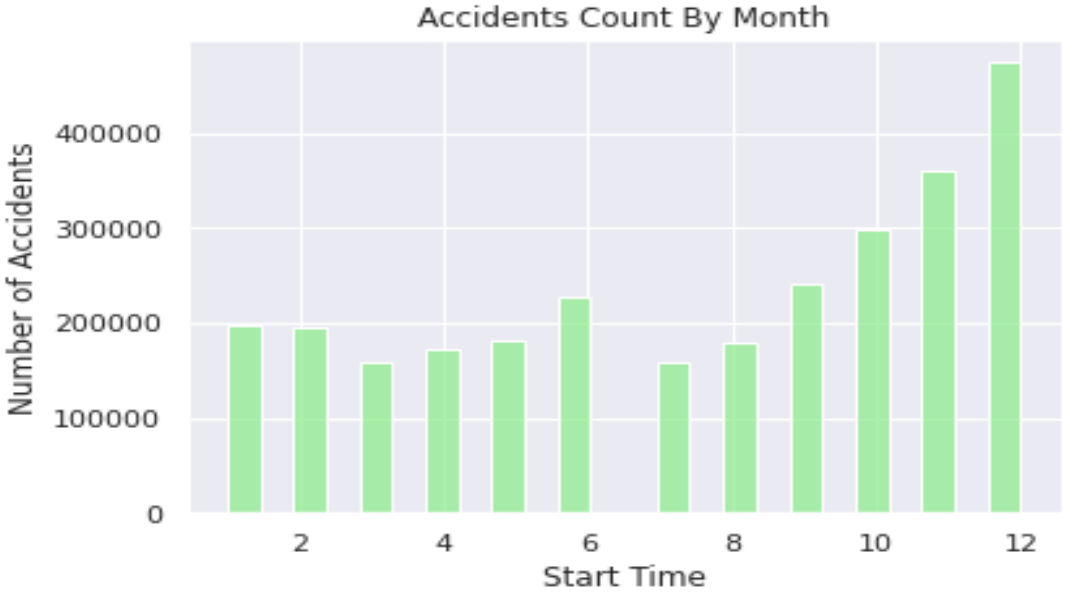
Astronomical Twilight: Is the period when geometric centre of the sun is between 12º and 18º below horizon it.



Graph 7

Analysis of graph 7 shows that when civil twilight is there number of accidents increases, this may be attributed to increased confusion whether to turn off or turn on the headlights. Further the analysis of accidents when twilights are true or false will have bias as the twilights time are of very short duration as compared to day or night duration.

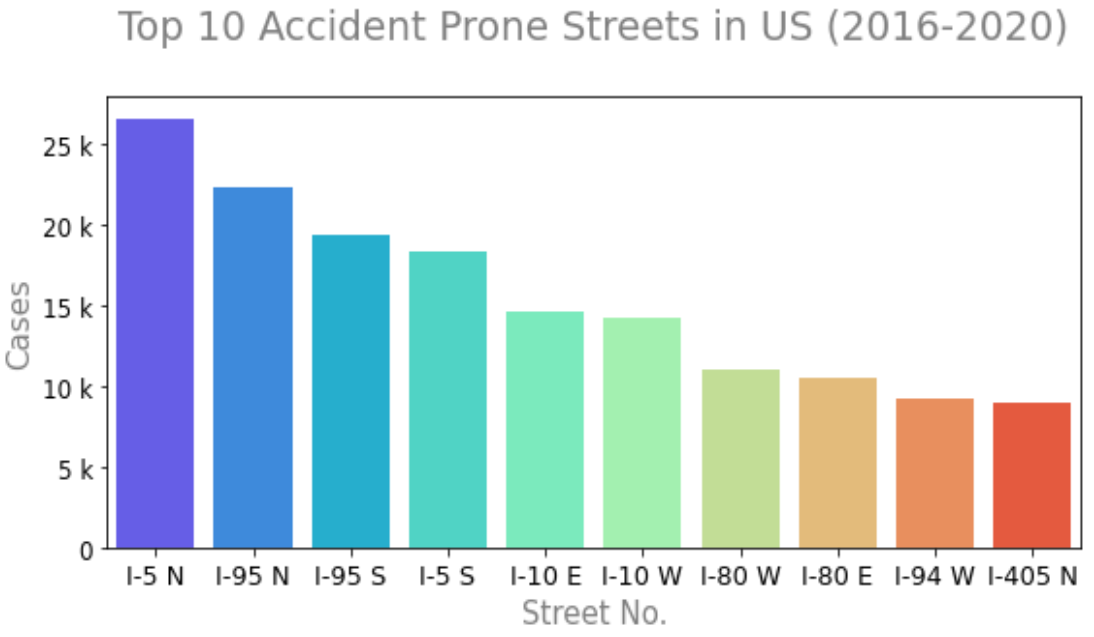
4.1.5. Accident count by month



Graph 8

As Shown in graph 8 number of accidents increases significantly in November and December this increase is attributed to increased travel (specially in groups) during the holiday season. This holiday season also increases accidents caused by bad driving behaviors like drinking and driving.

4.1.6. Direction wise streets most of the accident happens



Graph 9

By analyzing of graph 9 it was found that:

1. 28.64% accident happened on the street: I-5 N.
2. 24.08% accident happened on the street: I-95 N
3. 20.86% accident happened on the street: I-95 S
4. 19.81% accident happened on the street: I-5 S
5. 15.78% accident happened on the street: I-10 E
6. 11.25% accident happened on the street: I-80 E
7. 11.82% accident happened on the street: I-80 W
8. 15.37% accident happened on the street: I-10 W
9. 9.89% accident happened on the street: I-94 W
10. 9.70% accident happened on the street: I-405 N

Streets going to north are more prone to accidents than that of west bound or east bound.

5. Methodology

After Data was preprocessed, cleaned and outliers were removed. Features that were not found relevant to accident duration were dropped. We divided our dataset into four classes of accident duration and created four different data frames based on accident duration namely Test1(accident duration less than 60 min), Test2 (accident duration greater than 60 and less than 120 min), Test3(accident duration greater than 120 and less than 180 min) and Test4 (accident duration greater than 180 min). Data of each Test data frame was thus divided into train data and test data with duration as the dependent variable and other features as independent variables. Various models were then trained on these

datasets and results were compared. Training of these four different data frames was done on several models as listed below.

To evaluate the accuracy of the presented models, the Root Mean Squared Error and the Mean Absolute Error were adopted. The MAE measures the average magnitude of the errors, the RMSE determine their variation. In RMSE relatively higher weight is given to large errors. This implies that when large error is unavoidable then RMSE is most useful.

5.1 Multiple Linear Regression

Multiple linear regression is used to predict the dependent variable with two or more features with the assumption that the relationship of each independent variable is linearly related to the dependent variable, the residuals (a measure of how far away the predicted dependent variable from the regression line) is following normal distribution and that the correlation among the independent features are not highly co-related.

Table 1. Errors obtained for Multiple Linear Regression

Errors	Duration (<60 min)	Duration (60 - 120 min)	Duration (120-180 min)	Duration (>180 min)
Mean Absolute Error (MAE)	10.067	13.758	14.229	62.233
Root Mean Squared Error (RMSE)	12.803	15.745	16.878	68.684
R2-Square Error(r^2)	0.00351	0.00308	0.00455	0.0332

5.2 Decision Tree

The decision tree algorithm can be used in both classification and regression problems, but it is used extensively in classification problems. It is a type of Supervised learning algorithm. Classification in a decision tree is structured as a tree where the nodes and branches are represented by features and decision methods respectively and Output is represented by leaf nodes

Table 2. Errors obtained for Decision Tree

Errors	Duration (<60 min)	Duration (60 - 120 min)	Duration (120-180 min)	Duration (>180 min)
Mean Absolute Error (MAE)	9.341	15.041	18.704	57.527
Root Mean Squared Error (RMSE)	13.534	21.122	25.192	87.853

5.3 Deep Neural Network

A DeepNeural Network (DNN) is a class of artificial neural network (ANN) that is consist of multiple layers between the output and input layers. Neural Networks are made of different components neurons, synapses, weights, biases, and functions. Backpropagation is used to update the weights. Here kernels are initialized AS “normal” and activation function Relu is used in all dense layers except the output layer where linear activation function is used.

Table 3. Errors obtained for Decision Tree

Errors	Duration (<60 min)	Duration (60 - 120 min)	Duration (120-180 min)	Duration (>180 min)
Mean Absolute Error (MAE)	10.29	13.2	13.76	56.03
Root Mean Squared Error (RMSE)	14.12	16.44	17.74	72.59

6. Results

DeepNeural Network is best performing when capmared with multiple linear Regression and Decision tree. Multiple Linear regression is performing alsmost as good as DeepNeural Nwtwoek for accident duration less than 60 min and the decision tree is also performing good for accident duration less than 60 min.

From the following, we can compare which model is more efficient for which accident duration

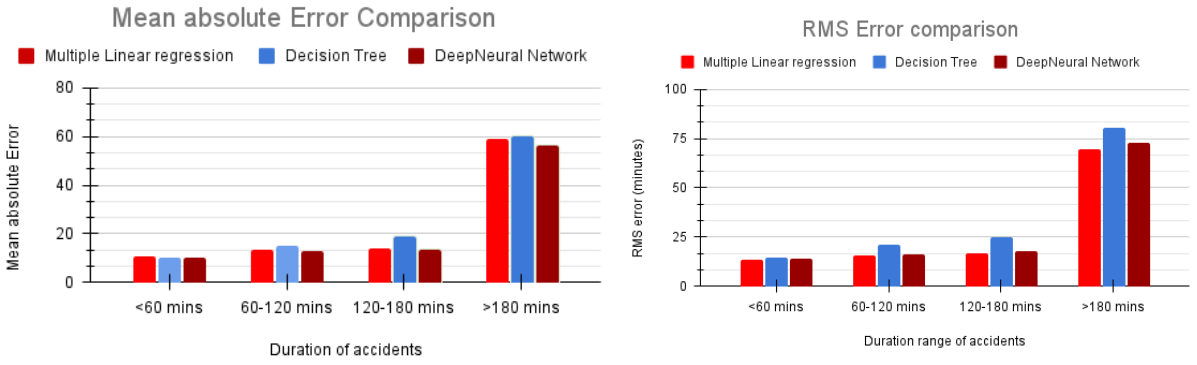


Figure 1. (a) MAE comparison of MLR, DT and DNN (b) RMSE comparison of MLR, DT and DNN

Here all the models are performing good for accident duration less than 60min as the variation is less and this creates a bias. Further as the numerical value of accident duration slab increases the bias also increases hence the error increases significantly.

7. Work Distribution

Raj Kumar Yadav - Has written the Introduction section, has read research paper Analysis of Regression Method on Traffic Incident Duration Prediction [1] & Traffic Incident Duration Prediction Based On K- Nearest Neighbor.[2]

data preprocessing and cleaning, data exploration through graph, and done the analysis of accident duration prediction using Multiple linear regression, decision tree algorithms and deep neural network and written code and has done data preprocessing, outlier removal and missing value section, writing report & presentation

Anshul Gautam - has read the research paper Estimating Magnitude and Duration of Incident Delays by A. Garib, A. E. Radwan, H. Al-Deek and provided the summary.[3] & Traffic Incident Duration Prediction Based On K- Nearest Neighbor.[2]
,
data preprocessing and cleaning, done data exploration and analyzed dataset using relevant graphs using tableau, label encoding, written code for analysis of graph and for models, writing report & presentation,

Overall, both members has equally participated in this project

8. Turnitin similarity = 6 %

9. References

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019)

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[3] Xuanqiang WANG, Shuyan CHEN, and Wenchang ZHENG." Analysis of Regression Method on Traffic Incident Duration Prediction".(2013)

[4] Yuan Wen, Shuyan Chen, Qinyuan Xiong, Rubi Han,Shiyu Chen. "Traffic Incident Duration Prediction Based On K- Nearest Neighbor"(2012)

[5] A. Garib, A. E. Radwan, H. Al-Deek. "Estimating Magnitude and Duration of Incident Delays" (1997)