Analysis On US Accidents Dataset

1st N.Meghana Reddy
Computer science and engineering
PES UNIVERSITY
Bangalore, India
meghana.narpala@gmail.com

2nd Harshapriya C.R
Computer science and engineering
PES UNIVERSITY
Bangalore, India
harshapriya112@gmail.com

3rd Priyanka G
Computer science and engineering
PES UNIVERSITY
Bangalore, India
priyankagopi86@gmail.com

***Abstract*—Reducing traffic accidents is an important public safety challenge, therefore, accident analysis and prediction has been a topic of much research over the past few decades. By employing the US-Accidents dataset and through an extensive set of experiments across several large cities. The objective of the project is to analyze the US accident data from 50 states and to inform the US government agencies and the general public on trends and possible causes of traffic accidents and what could be done to reduce them.**

***Index Terms*—Accident Prediction,Descriptive Analysis,Major cause and effects,Heterogeneous Sparse Data**

I. Introduction

Road accidents are the leading cause of death world wide according to a World Health Organisation Report.More than 1.35 million lives are lost each year and some where in the range of 20 and 50 million people sustain nonfatal injuries.Globally,road accidents are the 10th leading cause of death.Reducing traffic accidents is an important public safety challenge around the world. A global status report on traffic safety notes that there were 1.25 million traffic deaths in 2013 alone, with deaths increasing in 68 countries when compared to 2010. Accident prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer. Given its significance, accident analysis and prediction has been a topic of much research in the past few decades. Analyzing the impact of environmental stimuli on traffic accident occurrence patterns , predicting frequency of accidents within a geographical region, and predicting risk of accidents are the major related research categories.Through this analysis of data we can showcase all the finding by each state like day and time safe to travel, accident prone area and zip code in each state, severity, weather conditions, also if someone wants to go from Los Angeles to San Francisco in which area accidents mostly occur. For State Government officials, this platform will help to make a decision and provide solution-based on accident issues face by each state.

II. Literature Review

A. PAPER - 1 : Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis

Source : https://doi.org/10.1371/journal.pone.0223473
Published year : October 9, 2019.

This paper aims to investigate the factors associated with road accidents in South Korea.

The rainfall data of the Korea Meteorological Administration and road accidents data of Traffic Accident Analysis System of Korea Road Traffic Authority is analyzed for this purpose.

In this connection, multivariate regression analysis and ratio analysis with the descriptive analysis are performed to uncover the catastrophic factors involved. In turn, the results reveal that traffic volume is the leading factor in road accidents. The limited road extension of 1.47percent compared to the 4.14percent per annual growth of the vehicles is resulting in road accidents at such a large scale. The increasing proportion of passenger cars accelerate road accidents as well.

56percent of accidents occur by the infringement of safety driving violations. The drivers with higher driving experience tend to have a higher accident ratio.

This study considers the data of accidents and that of their associated weather conditions to perform the analysis.

For that purpose, the data is gathered from a variety of state-of-the-art sources for more than one decade considering a number of affecting variables.

An in-depth investigation is conducted regarding the accident types that are caused by the gender of the driver during the day and night time with their driving experience and violations leading to the accidents.

Single and multivariate regression analysis along with descriptive analysis is carried out to highlight various uncovered aspects like age, gender, experience, violation type, traffic volume, etc. to understand their individual as well as the collective effects on accidents.

This study suggests propositions to mitigate the risk of accidents for elderly drivers especially during the night-time and bad weather conditions, which are equally helpful to the Governments in the design of traffic laws to govern the urban roads traffic.

a) Experiments and analysis:

- In the analysis, various factors are considered including rain, driver's age/gender, driving experience, road type, vehicle type involved in accidents, alcohol intoxication, violation type causing the accident, traffic volume, and lighting conditions so as to find their individual impact on the accidents.

- Night makes one-third of total day time on average. However, 40percent of the total fatalities and injuries occur during the night time . Such statistics indicate potential danger that drivers face when driving during the night time. Furthermore, the different time during the night has a different associated risk of death and injury . Driving during 2:00 am to 5:00 am includes 5.6 times increased risk of traffic accidents.
- The researchers have various reasons for accidents rates during night time. The most general reason is the low visibility conditions and sleepiness . However, another finding is that accidents during night time are more related to the use of roads at night rather than the darkness at night.
- However, data analysis becomes complicated when multiple variables are considered collectively.

b) Theoretical framework: Theoretical framework: This study conducts the regression analysis to identify the relationships between accidents and various factors like weather, road conditions, driving behavior, etc. The regression analysis is a widely used technique to investigate the interrelationship among a group of variables. It first analyzes the functional relationship among variables and then checks the accuracy of the relationship. In this study, the single linear regression and the multivariate linear regression analysis with the descriptive analysis are performed to investigate the effect of individual factors. The simple linear regression is helpful in understanding the driving factors that could affect the output (or target) individually . However, considering the multiple variables at the same time often lead to the most meaningful knowledge from the data. For example, studying the impact of driving experience on accidents. The value, and R2 is used to evaluate the performance of the regression model and interpret the results. The value is applied to test the hypothesis in a model. It is also called the marginal significance level [32]. It tests the null hypothesis Ho, which means that there is no significant relationship between the predictions and regressors. A low value suggests that the prediction is meaningful and that the null hypothesis can be rejected. R2 is called the goodness-of-fit index, which is calculated in as follows,

where SSE stands for the sum of squared residuals or errors while SST represents a total sum of squared deviations of y from its mean value. The value of R2 varies between 0 and 1 . If R2 is close to 1 it implies that the variables are perfectly related with perfect goodness of fit. Conversely, if R2 approaches 0, it then infers that there is no relationship between the given variables and the regressors do not explain the prediction.

So we can draw out conclusions that the accident data is collected from various reliable sources to carry out the multivariate regression analysis. Results demonstrate that traffic volume and infringement of safety violations are the ultimate causes of road accidents . Therein, the widespread growth of vehicles affects the traffic volume which leads to a higher number of accidents. The age of the drivers, traffic volume and rain are critical factors for elevated road accidents. The analysis of collision type may also help to uncover accident causes, their associated risk factors and their relationship to the severity of the accident.

B. PAPER -2 : Accident Risk Prediction based on Heterogeneous Sparse Data

Title : Accident Risk Prediction based on Heterogeneous Sparse Data

Source : https://arxiv.org/abs/1909.09638

Year the paper was published : 10 Sep 2019

- This research paper relates to US Traffic Accidents Prediction.
- The analysis and results show significant improvements to predict rare accident events. Further, the impact of traffic information, time, and points of-interest data for real-time accident prediction is shown.
- Studying the impact of weather factors (e.g., precipitation) on road accidents , applying data mining techniques to extract association rules to perform causality analysis , and statistical analysis .

a) Methods used:

- A deep-neural-network model
- Causality analysis

b) The main contributions of this paper:

- A new methodology for heterogeneous data collection, cleansing, and augmentation to prepare a unique, large-scale dataset of traffic accidents. This dataset has been collected for the contiguous United States over three years, and contains 2.25 million traffic accidents.
- A variety of insights gleaned through analyses of accident hotspot locations, time, weather and points-of-interest correlations.

c) Main claims:

- Using US-Accidents data set, variety of data analysis and profiling to derive a wide-range of insights were done.
- Analysis demonstrated that about 40 percent of accidents took place on or near high-speed roadways (highways, interstates, etc.) and about 32 percent on or near local roads (streets, avenues, etc.).
- They also derived various insights with respect to the correlation of accidents with time, points-of interest, and weather conditions.
- Traffic accidents are a major public safety issue, with much research devoted to analysis and prediction of these rare events.
- The daily distribution of traffic accidents, where significantly more accidents were observed during the weekdays.
- Most of the accidents took place near junctions or intersections (crossing, traffic signal, and stop).
- MapQuest tends to report more accidents near intersections, while Bing reported more cases near junctions. This shows the complementary behavior of these APIs and the comprehensiveness.

d) Takeaway from this paper:

- Traffic accidents are a major public safety issue, with much research devoted to analysis and prediction of these rare events.
- However, most of the studies suffer from using small-scale datasets, relying on extensive data that is not easily accessible to other researchers, and being not applicable for real-time purposes.
- To address these challenges, new framework for real-time traffic accident prediction based on easy-to-obtain is proposed, but sparse data.
- The prediction model incorporated several neural network based components that used a variety of data attributes such as traffic events, weather data, points-of-interest, and time information.
- The impact of different categories of data attributes for traffic accident prediction, and found time, traffic events, and points-of-interest as having significant value.

e) Conclusion:

- Traffic accidents are a major public safety issue, with much research devoted to analysis and prediction of these rare events.
- Analysing the factors increasing the accident.There are factors which contribute to accidents more than others.
- The daily distribution of traffic accidents, where significantly more accidents were observed during the weekdays.
- Most of the accidents took place near junctions or intersections (crossing, traffic signal, and stop).
- MapQuest tends to report more accidents near intersections, while Bing reported more cases near junctions.
- This shows the complementary behavior of these APIs and the comprehensiveness.
- Here we note that about 32 percent of accidents happened on or near local roads (e.g., streets,avenues, and boulevards), and about 40 percent took place on or near high speed roads (e.g., highways, interstates, and state roads).
- The period of day data shows that about 73 percent of accidents happened after sunrise (or during the day).

C. Paper -3 : Analysis of US accidents and solutions

Year Published : 2020 , Based on US accidents dataset.

Worldwide status report on road safety 2009, estimates that in high income nations like the USA there are 65 of reported vehicle deaths from the Vehicle Occupants. The same report also predicts that road traffic injuries will rise to become the 5th leading cause of death by 2030 (WHO,2009).

The analysis in this paper includes number of accidents by year, number of accidents by state, best time to travel by month, day and hour, accident-prone area in each state, factors responsible of the accidents like weather, wind flow, temperature, location, etc., deaths in each state, age group of fatalities, drivers involved in accident, drivers age group, vehicles involved in accident, driver with alcohol consumption. The main recommendations from the project focus on Infrastructure, Policy, Administrative, and Human behavior-related changes that can be implemented by the state and the federal government.

a) Methodology: Overall study used here is divided in three parts.

- Understanding the current accident situation in the United States, like how many accidents are happening each year, deaths in accidents, the severity of accidents, the time and day safe to travel, and overview.
- This part tells for each state, what are the accident-prone areas and the zip code in each state.
- address the solution for the public as well as State Governments, how these accidents can be reduced, and what are solutions can be implemented by each state based on the crucial factors.
-

Python and Tableau are used in this analysis

b) Recommendations:

- Policy Recommendation
- Administrative Recommendation
- Human Behaviour Recommendation
- Infrastructure Recommendation

c) Conclusion:: Traffic accidents are a main public safety issue, with much research devoted to the analysis and prediction of these rare events. The study helped us to derive factors that are responsible for accidents.

From this dataset, a variety of insights concerning the location, time, weather, and points-of-interest of an accident are found.

The analysis helps us understand the best month, day, and hour of the day to travel.

Also, it can help us to predict what are the accidentprone areas in each state. The analysis shows that the highest death is happening between the 20-35 age group, which is impacting the US economy.

Most of the accidents occurring due to drunk driving. Finally, this study recommends infrastructure, Policy, Administrative, and Human Behavior changes, which can help to reduce US accidents.

d) The significant findings from the analysis are::

- Most of the accidents are happening in October, November, December.
- nighttime is safe to travel
- a more substantial number of deaths are from drivers in the 20-35 age group
- weather, temperature, and location are the factors responsible for 9percent of accidents
- about 60percent accidents are attributed to drunk driving.

III. EDA, Visualization

A. Data

There are 3 million records in this dataset from February 2016 to December 2019.

Data related to the number of deaths in accidents from 2004 to 2018 has extracted from the National Highway Traffic Safety Administration site along with deaths by age groups

and drivers with alcohol who involved in accidents taken to analyze factors for accidents more appropriately.

- No NULL Records
- No Missing Values
- Consistent Date Format
- No Duplicate Records
- No Mismatching Columns

B. Data set description

Kaggle.com has collected streaming traffic data from two sources,"MapQuest Traffic" (MapQuest Traffic API, 2019) and "Microsoft Bing Map Traffic" (Bing Map Traffic API, 2019) respectively, "whose APIs broadcast traffic events (accident, congestion, etc.) captured by a variety of entities - the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks" (Moosavi, Samavatian, Nandi,Parthasarathy, Rajiv Ramnath,2019).

There are 3 million records in this dataset from February 2016 to December 2019.

C. Data analysis and visualization

- EDA, Visualization : Visualization for understanding the data distribution and features
- This dataset contains 49 columns which means we are dealing with 49 features in total which is a little bit too much. We will try to remove some of them and maybe combine some columns into one. Such as the features containing points of interests can be merged into one.
  a) Feature selection: : Selecting relevant features that contribute to the prediction of risk
- While the data set has 3 million records, it is not ready to use for analysis.
- There are many anomalies in the dataset like, Null records, Date format,Day is missing,Duplicate records,Mismatched column.
- To Address all these anomalies in data, data cleaning is the most important and mandatory step.
  b) Important findings: :
- There are only three API sources that reported the accidents, Bing, MapQuest, MapQuest-Bing source. Observations showed that most of the accidents were reported by MapQuest, followed by Bing.
- Number of Accidents were lower in the early years as compared with those in recent years where the number of accidents have raised.
- There are a large number of accidents during weekdays (Monday to Thursday). On the contrary, there are relatively fewer accidents on weekends (Friday, Saturday, and Sunday)
- Monthly accidents increase steadily from the lowest points in January and February, peak in October-December. First half of the year showed a smaller number of accidents compared to the last quarter of the year.
- From the visualization California is the most accident prone area.California, Texas and Florida alone make up almost 40 percent of all the accidents that took place.
- From the visualization, Pennsylvania, South Carolina and Texas have accidents with high severity.
- The Startlatitude and Startlongitude features are interesting since they can be plotted on a map, to get the exact location of the accident.
- Most accidents occur at around 8–9 am in the morning and then there is a second surge at 4 to 5 pm. As that is the time when most people travel to and from work, which results in increasing the traffic density which in turn leads to more accidents.
- From the correlation , accidents Start Lat and End Lat are the crucial factors for the accident. Temperature and wind flow are also playing an important role.

IV. Problem statement

Goal is to study accidents hotspot locations, cause and effects of accidents in US and study the impact of environmental stimuli on accident occurrence.

V. Conclusion

- Traffic accidents are a main public safety issue, with much research devoted to the analysis and prediction of these rare events. The study helped us to derive factors that are responsible for accidents.
- From this dataset, a variety of insights concerning the location, time, weather, and points-of-interest of an accident are found.
- The analysis helps us understand the best month, day, and hour of the day to travel.
- Also, it can help us to predict what are the accident prone areas in each state. The analysis shows that the highest death is happening between the 20-35 age group, which is impacting the US economy.
- Traffic accidents are a major public safety issue, with much research devoted to analysis and prediction of these rare events.
- Analysing the factors increasing the accident.There are factors which contribute to accidents more than others.
- The daily distribution of traffic accidents, where significantly more accidents were observed during the weekdays.