Abhilasha_Final_Thesis-1.pdf

by Abhilasha Garg

Submission date: 27-Nov-2023 12:45PM (UTC+0000)
Submission ID: 217008087
File name: Abhilasha_Final_Thesis.pdf (2.34M)
Word count: 24050
Character count: 139242

WEATHER-RELATED TRAFFIC ACCIDENT PREDICTION USING RANDOM FOREST WITH CATBOOST AND LIGHTGBM

ABHILASHA GARG
MS-DATA SCIENCE

Final Thesis Report

DECEMBER 2023

DEDICATION

This thesis is dedicated to my loving family, whose unwavering support and encouragement and been the cornerstone of my academic journey.

To my professor and mentors, thank you for your guidance, wisdom, and patience that helped shape this work into fruition.

I also dedicate this thesis to my friends, for their constant encouragement and understanding during moments of triumph and challenges.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the following individuals and organization for their invaluable support and contribution to the completion of this Master's thesis in Data Science.

Channabasva Chola, my thesis advisor, for his unwavering guidance, encouragement and expertise throughout the research process. His mentorship has been instrumental in shaping my academic and professional growth.

I am also deeply indebted to **upGrad** and **Liverpool John Moores University** (LJMU) for providing the opportunity to pursue this Master's in Data Science program. upGrad's unique blended learning model, combining online lectures with in-person interactions, has been an enriching experience that has facilitated both theoretical understanding and practical application of data science concepts. LJMU's esteemed reputation and rigorous academic standards have provided a strong foundation for my research and instilled in me a deep appreciation for the field of data science.

Finally, I would like to acknowledge the contributions of the **support staff**, **upgrad buddy**, and the **administration** at upGrad and LJMU for their assistance and support throughout my research process. Their expertise and dedication have been instrumental in enabling me to access resources, conduct data analysis, and successfully complete my thesis.

I am deeply grateful to each and every individual and organization mentioned above for their contributions to my Master's thesis in Data Science. Your support, guidance, and encouragement have been invaluable, and I am truly honored to have had the opportunity to work alongside such exceptional individuals and institutions.

ABSTRACT

Accurately forecasting traffic conditions during adverse weather events is crucial for optimizing transportation systems, reducing congestion, and improving road safety. This study explores the application of two powerful machine learning algorithms, CatBoost and LightGBM, to predict traffic accidents during weather-related events. Utilizing the Large-Scale Traffic and Weather (LSTW) dataset, a comprehensive collection of real-world traffic and weather data for the United States, the study aims to achieve two primary objectives:

Traffic Pattern Estimation: Employ state-of-the-art machine learning algorithms to estimate traffic patterns under the influence of various weather conditions.

Algorithm Comparison and Evaluation: Conduct a rigorous comparison and evaluation of the predictive performance of CatBoost and LightGBM models in forecasting traffic congestions.

The study commences with data preprocessing of the LSTW dataset, carefully selecting and extracting relevant features that capture the intricate relationship between weather events and traffic patterns. Subsequently, the CatBoost and LightGBM models undergo rigorous training and evaluation processes to optimize their predictive capabilities. The performance of each model is assessed using various metrics, including R-squared score and root mean squared error (RMSE).

The findings reveal that LightGBM, when combined with oversampling techniques, outperforms CatBoost in predicting traffic accidents during weather-related events. LightGBM's superior performance is attributed to its ability to effectively handle imbalanced datasets, a common characteristic of traffic accident data. Furthermore, LightGBM exhibits faster training and prediction times, making it a more efficient choice for real-time traffic forecasting applications.

This study contributes to the advancement of traffic accident prediction during adverse weather events by demonstrating the effectiveness of machine learning algorithms, particularly LightGBM, in accurately estimating traffic patterns under various weather conditions. The findings provide

valuable insights for transportation authorities and traffic management systems, enabling them to proactively implement measures to mitigate traffic congestion and enhance road safety during inclement weather.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NHTSA …….……..	National Highway Road Safety Administration
US…..…..…..…..…...	United States
GRF……	Generalized Random Forest
WHO……	World Health Organization
USA…….……	United States of America
API……	Application Programming Interface
GDP……	Gross Domestic Product
DALY……	Disabilities Adjusted Life Years
RTC…….……	Road Traffic Collisions
ARIMA…….……	Autoregressive Integrated Moving Average
ARIMAX…….....	Autoregressive Integrated Moving Average with Explanatory Variables
LSTW……	Large Scale Traffic and Weather Dataset

CHAPTER 1 : INTRODUCTION

1.1 Background

Traffic accidents are strongly affected by weather-related factors. Notably, a thorough investigation by the National Highway Road Safety Administration (NHTSA) found that weather-related incidents accounted for 20% of all traffic deaths in the US in 2019. (Reish and Leah, n.d.)

The number of vehicles is rising drastically in today's society due to its fast growth. There has also been a significant rise in traffic accidents, with severe human and financial (Gebru, 2017). From 1.25 million people are killed in traffic accidents annually, and over 20 to 50 million people are affected by nonfatal accidents, according to the World Health Organization (Bahiru et al., 2018). It is evident that one of the main causes of injury and death globally is traffic accidents. Research on intelligent vehicles and traffic science has focused heavily on ways to anticipate and avoid traffic accidents.

One significant indicator of the harm caused by automobile accidents is their intensity. Differing variables contribute to different levels of traffic accidents. Numerous formulas and variables have been mentioned in the research on auto accidents. (Wang et al., 2018) examined the position of an automobile in road transects, as well as the road safety grade, surface condition, visual state, vehicle condition, and driver status. Based on these analyses, an 86.67 percent prediction accuracy model was developed. Using artificial neural networks, (Alkheder et al., 2016) predicted the severity of road accidents based on 16 characteristics and four injury degrees (minor, moderate, severe, and death). According to research by (Holmes et al., 2020), the primary human variables responsible for plant and animal extinctions in highway accidents were old age, overtaking, speeding, religious beliefs, poor braking performance, and faulty tyres. It has also been noted that some aspects of weather and accident circumstances influence the traits of highway traffic behavior (Caleffi et al., 2016). In order to anticipate traffic flow under unclear traffic accident information, (An et al., 2019) used a fuzzy convolutional neural network. They then used actual automobile trajectories and meteorological data to confirm the model's usefulness. Additionally, multi-objective evolutionary

algorithms have shown potential in forecasting traffic accident severity based on user preferences (Tamim Kashifi and Ahmad, 2022). Using data on traffic accidents, traffic flow, weather, and air pollution, the deep learning approach created a short-term traffic accident risk prediction model (Ren et al., 2018). In order to estimate the probability of traffic accidents in cities, the spatiotemporal correlation of traffic accidents has been proposed (Ren et al., 2018). In predicting the frequency of traffic accidents, the temporal aggregation neural network layer created by (Huang et al., 2019) automatically extracts correlation scores from the temporal dimension.

According to (Kumeda et al., 2019), the most important factors in choosing the qualities are the lighting conditions, the number of vehicles, and the first road class and number. Using data mining techniques, (Murphey et al., 2021) successfully examined driving habits. Additionally, an accident prediction model based on probabilistic and spatiotemporal relationship learning was presented by (Bao et al., 2020). Fuzzy machine learning model is used by (Tuncal Yaman et al., 2022) to examine the variables influencing the injury severity attained in traffic accidents. Age, gender, wearing a seatbelt, drinking, and using narcotics are a few examples. It was possible to acquire independent importance normalized variables that affected injury factors. Numerous algorithms have been used to forecast and avoid traffic accidents. The random forest method has been used more and more in traffic accident information processing in recent years.

Numerous professions, notably statistics (Schonlau and Zou, 2020), meteorology (Ding et al., 2023), and medicine (Iwendi et al., 2020), employ the random forest method extensively. In areas of heavy accidents, the random forest has also produced some results. (Yan and Shen, 2022) investigated how influencing factors impact the severity of traffic accidents using Bayesian optimization and random forests. An approach for predicting accident risk based on a deep convolutional neural network and random forest was proposed by (Strickland et al., 2017). Using three prediction performance evaluation metrics, sensitivity, and specificity— (Chen and Chen, 2020) determined the optimal all-encompassing approach that included the most impactful prediction model and input variables that had a greater positive influence on accuracy, sensitivity, and specificity. (Harada et al., 2017) took into account the different forms of eye movements when using the random forest tool to determine the distraction of cognitive drivers. Different time periods, road gradients, tidal lanes, closeness to facilities, and accident sections were chosen by (Wang et al., 2021) as indicators impacting traffic.

The outcomes of the research illustrate that the strategy is capable of efficiently avoiding the clogged road and obtaining the fast path. To estimate diverse treatment efficacy in road safety assessment, (Zhang et al., 2022) presented specialized random forest, which gives local authorities and regulators more complete data and enhances the effectiveness of speed camera projects. Furthermore, GRF has the potential to be a promising approach for emphasizing the heterogeneity of intervention effects in the consideration of road safety.(Zhang et al., 2022) developed generalized random forest to estimate heterogeneous treatment effects in road safety studies, therefore improving the effectiveness of speed camera services and offering more complete information to local authorities and policymakers. Furthermore, GRF has the potential to be a useful technique for highlighting the variability of treatment effects in the investigation of road safety.

For predicting traffic accidents, especially those affected by weather, machine learning has proven to be a powerful tool. Using machine learning to predict weather-related traffic accidents has become increasingly popular in recent years (Hasan et al., 2023a; Luo and Wang, 2023a; Testolina et al., 2023)

The LSTW dataset, a comprehensive collection of traffic and weather event data for the United States, is a noteworthy source in this area (Moosavi et al., 2019a). The training of models using random forest for forecasting weather-related traffic incidents can make good use of this dataset.

1.2 Problem Statement

Around the world, a significant cause of fatalities and injuries is road accidents. Each year, there are more than 6 million traffic accidents in the United States alone, resulting in more than 38,000 fatalities. Many of these collisions are brought on by particular sorts of traffic incidents, such roadwork or severe weather. Various studies have been conducted by researchers (Moosavi et al., 2019b; Mondal et al., 2020; Hasan et al., 2023b) across the globe with different dataset (Diaz-Ruiz et al., n.d.; Dadwal et al., 2021; Luo and Wang, 2023b) in order to achieve accurate prediction based on various factors and techniques. Several statistical techniques (such as support vector machines and random forests) have been investigated over the years in a variety of research projects to create effective crash severity prediction models. Machine learning-based models for crash severity prediction are emerging enabling more precise data-driven prediction as technology advances and computation starts to become more economical. Nonetheless, when compared to other approaches in the same genre, certain machine learning techniques offer higher performance and enhanced efficiency.

In this paper, we would be exploring two random forest algorithms that could be used to evaluate big datasets of traffic data and find correlations that can be used to anticipate accidents. To create our models, we'll employ the two well-known algorithms LightGBM and CatBoost. Several criteria, including accuracy, precision, and recall, will be used to assess the performance. In addition to assessing how independent factors may affect the models' performance, we will also conduct sensitivity analysis and comparison between the two algorithms.

One of the advantages of utilizing LightGBM and CatBoost for weather event-based accident prediction is that a dependable model for estimating the probability of weather-related traffic accidents may be created. The study will able to forecast the severity of accidents as well as identify the core factors that end up in accidents. This may help proactive steps taken by transit agencies to lower the likelihood of accidents and protect lives. The performance and outcomes of the two algorithms may also be compared to help determine the advantages and disadvantages of each algorithm as well as which method is more appropriate for a certain set of use scenarios. Additionally, the study may lead to the development of machine learning-based crash prediction models connected to weather, which will help a variety of professional associations and stakeholder

organizations improve the safety for road users and minimize the economic cost in responding to such emergencies.

1.3 Research Questions

For each of the mentioned study objectives, the subsequent research questions are proposed.

1. Can accidents caused by certain weather conditions be predicted using random forest?
2. Determining a correlation between poor weather and road accidents.
3. What features are the most effective for projecting accidents driven by climate?
4. To review different techniques and evaluate the precision and model performance leveraging random forest algorithms.

1.4 Aim and Objectives

The primary goal of this research is to recommend the top model for using the random forest algorithm to forecast traffic accidents induced by meteorological occurrences in the United States.

The following are the study's objectives, which were based primarily on aims of this study:

1. To investigate the trend and correlation between unfavorable weather and traffic occurrences.
2. To determine the variables that are significant for forecasting traffic accidents.
3. To compare CatBoost and LightGBM predictive modelling
4. To assess the effectiveness of models using a range of metrics.

1.5 Significance of the Study

While being a cautious driver is important, it is equally important to remain vigilant of other factors and drivers on the road. Warning signs on highways and roadways can indicate areas that are more impacted by bad weather, such as slippery slopes, sharp turns, and rocky terrain. These areas require extra caution and attention.

Statistics have shown that 11% of fatal accidents are caused by bad weather. However, weather is just one factor that contributes to car accidents. Other factors, such as speed, fog, road type, pavement temperature, and driver reaction time, can also contribute to dangerous driving situations.

This study introduces an innovative approach using LightGBM and CatBoost algorithms for accurate accident prediction, focusing on feature engineering and predictive model evaluation. The integration of LightGBM and CatBoost is expected to improve accuracy and performance, especially in the Large-Scale Traffic and Weather (LSTW) dataset. The focus on reproducibility, reuse, and extensibility in the research community ensures transparency and reliability, facilitating future research endeavors in this domain.

CatBoost and LightGBM are comparatively newer algorithms that have shown promising results in traffic accident prediction. CatBoost handles categorical features efficiently with its gradient boosting framework and can effectively deal with missing data. On the other hand, LightGBM uses a gradient boosting framework to optimize efficiency while handling large-scale datasets. Both algorithms have advantages in terms of model accuracy and computational efficiency.

1.6 Scope of the Study

For a number of reasons, the focus of this study is only on the forecasting of weather-related traffic incidents in the United States.

- First, there is a need for reliable ways to anticipate weather events because they are a significant contributing factor to traffic accidents.
- The LSTW dataset, a sizable set of weather and traffic occurrences for the United States, offers a thorough analysis of these occurrences.
- Third, two well-known machine learning algorithms, LightGBM and Catboost, may also be efficient in predicting traffic accidents.

The final standard metrics for assessing the performance of machine learning models are accuracy, precision, and recall

1.7 Structure of the Study

The structure of the study is as follows:

- Chapter 1 – Introduction: An introduction and context to the research project are given in this chapter. Moreover, the goals and objectives as well as the pertinent research questions.

- Chapter 2 – Literature Review: The relevant studies in the domains of random forest algorithms and accident prediction are mentioned in this chapter. In addition, the accident's weather-related severity and the number of deaths.

- Chapter 3 – Research Methodology: This chapter gives a detailed walkthrough of the methodology followed during the experimentation stage. It includes the Data selection process, Data Cleaning and Data mining along with the tools used.

- Chapter 4 – Analysis: This chapter focus on the different analysis and visualization of the data along with data preprocessing steps like class balancing, handling missing values so on and so forth.

- Chapter 5 – Results and Discussion: This chapter discusses the result of experiments on the accident prediction dataset with different approaches and algorithm and comparative study of the outcome. Also, it has tried to answers the research question raised in Chapter 1.

- Chapter 6 – Conclusion and Recommendations: This chapter concludes the work done in the thesis and discusses future improvements.

CHAPTER 2 : LITERATURE REVIEW

2.1 Introduction

The automotive engine became the dominant mode of transportation after World War II. The supremacy of engine vehicles has not been contested since then. As an alternative, several initiatives have been made to enhance them, such as speeding up the production process and gradually adapting them to the unique topography of each country. Vehicle transportation is become a part of everyday existence. Given the startlingly high incidence of horrific accidents and fatalities, improvements to autos are unavoidable. Sadly, there has always been a chance of car collisions when operating a vehicle. Nicolas Joseph Cugnot was responsible for the first traffic accident in 1771 when he crashed his own, upgraded version of the first steam-powered vehicle into a wall. The car was severely damaged as a result of the accident (Kisan Nikam, n.d.).

Anglican preacher from Redruth was the first casualty of the road. When he saw a noisy, swiftly-moving replica of "a steam engine on wheels," created by William Murdock, he passed away from fear. The year that this happened was 1786 (Khaled et al., n.d.).The first recorded road accident death worldwide occurred in County Offaly, UK, on August 31, 1869. (Goniewicz et al., 2015) On a curve, beneath the steel wheels of an experimental steam-powered car her ancestors had built, an Irishwoman called Mary Ward fell out and was run over. She died as a result of her injuries. In Ohio City, the first automobile accident in the United States employing an internal combustion engine happened in 1891.

There are around 1.2 million deaths and between 20 and 50 million non-fatal injuries on the streets of the world each year. As a way to illustrate the global impact of traffic accidents, the World Health Organization (WHO)(WHO launches second global status report on road safety, n.d.) estimates that in high-income countries like the USA, 65 percent of reported vehicle deaths are from vehicle occupants, whereas in middle-income countries in the Western Pacific region, 70 percent of reported vehicle deaths are among vulnerable street users. This illustrates the significance of traffic accidents globally.(Global health risks: Mortality and burden of disease attributable to selected major risks, n.d.). The same report also predicts that road traffic injuries will rise to become the 5th leading cause

of death by 2030 (Global health risks: Mortality and burden of disease attributable to selected major risks, n.d.). Even if the loss and suffering brought on by car accidents worldwide pale in comparison to that brought on by illness and poverty, the issue is more serious than the current data alone suggests. The financial cost incurred by nation-states as a result of deadly car accidents must be taken into account. Many of the fatalities affect vehicle users without distinction. The USA alone suffered an approximate 836 billion economic loss in 2010.

There is been a lot of research out there that addresses accidents occurring the world over (Moosavi et al., 2019a) The fact that there are still more accidents occurring despite all of this advancement in research is a major concern for everyone. Nevertheless, the great number of them focus on accident analysis, and prediction has made use of constrained assets that do not even completely represent the challenge and influence the outcome we require.

In an attempt to address this issue, (Moosavi et al., 2019c) collected data from API resources available from various sources and used records of 2.25 million traffic accident instances that occurred within the contiguous United States over the previous three years for their research paper, "A Countrywide Traffic Accident Dataset." The information obtained in each accident record includes a range of intrinsic and contextual factors, along with time, location, natural language description, weather, period of the day, and points-of-interest (Moosavi et al., 2019d) (An et al., 2018) used weather data, traffic characteristics, daily average traffic, and weather data to build a neural network model that predicted the frequency of accidents on a highway route.

Considerable research has leveraged large-scale databases over time, yet these datasets have either been restricted or difficult to obtain (Moosavi et al., 2019b) Using a massive dataset of over 456,000 collisions in 48 US states between 1975 and 2000, Eisenberg (Eisenberg, 2004) conducted analysis to determine the effects of traffic accidents. Real-time traffic accident prediction has been investigated using large-scale datasets in recent works by Najjar et al. (Najjar et al., 2017)

2.2 Traffic Accident and Related Practices Worldwide

With 1.24 million predetermined fatalities annually, traffic accidents rank seventh globally in terms of primary causes of mortality for all age groups (Risk of road accident associated with the use of drugs: a systematic review and meta-analysis of evidence from epidemiological studies, n.d.). In underdeveloped nations, almost 85% of fatalities take place. Males are the most severely impacted demographic in road accidents, particularly those between the ages of 15 and 44. According to (Kaygisiz et al., 2015) the cost of traffic accidents to nations ranges from 1% to 2% of their overall GDP.

Even though just 52% of all automobiles worldwide are registered in developing nations, these regions account for 80% of all road traffic fatalities (WHO, 2013) [(WHO releases second worldwide traffic safety status report, n.d.)]. In fact, as indicated in Table 2.1, road traffic Developmental disabilities Life Years (DALYs) loss is predicted by the World Health Organization (WHO) to move from ninth important cause of DALYs in 1999 to third essential cause by 2020. While serious traffic accidents are a problem in low- and middle-income nations, the trend in high-income nations is the reverse.

The fatality rates of affluent and low-income nations differ substantially (Singh, 2017). The death rates in high-income countries have been falling, but they are still rising in low- and middle-income nations. Sadly, Asia has already seen the highest growth, regardless of the fact that rates of increasing differ by geography.

Table 2.1 Disease burden (DALYs lost)

S. No	1998 Disease or Injury	2020 Disease or Injury
1	Lower respiratory contaminations	Ischaemic heart disease
2	HIV/AIDS	Unipolar major depression
3	Perinatal conditions	Road traffic injuries
4	Diarrhoeal diseases	Cerebrovascular disease
5	Unipolar major depression	Chronic obstructive pulmonary disease
6	Ischaemic heart disease	Lower respiratory infections
7	Cerebrovascular disease	Tuberculosis
8	Malaria	War
9	Road traffic injuries	Diarrheal diseases
10	Chronic obstructive pulmonary disease	HIV/AIDS

Table 2.2 Number of Accidents on the 107-road segment. (The Neglected Epidemic: Road Traffic Crashes in India, n.d.)

Year or Period	Number of the Accidents on the 107 Road Segments (Percentage of Accidents all Over the year %)	Summary Statistics			
		Median	Average	Max.	Std, Dev.
2005	350(41)	2	3.27	26	4.14
2006	446(50)	3	4.17	22	4.87
2007	584(54)	3	5.46	30	6.16
2008	482(49)	2	4.50	25	5.55
2009	447(44)	3	4.18	28	5.16
2010	486(44)	4	4.54	22	4.84
2008-2010	1415(46)	9	13.22	75	14.38

2.2.1 Traffic Accident Trends in India

In India, traffic-related deaths and injuries are a growing public health concern for general medicine, society, and the economy. Approximately 2,650 individuals pay with their lives and 9,000 are injured in car accidents annually. The latest year for which data is available is 2013, and during that year, traffic accidents in India claimed the lives of 137,423 individuals and injured 469,900 more. India currently dominates the way in road fatalities, surpassing China, with about 140,000 deaths yearly due to traffic accidents, pinning the blame on the country's road infrastructure. India is the only nation in the world where traffic accidents result in more than 15 fatalities and 53 injuries every hour. While things are improving in many rich and emerging nations, including China, India is facing extreme situations. If the current trend persists, India's total number of road traffic fatalities is expected to reach 100% between 2013 and 2027. In the absence of governmental measures to mitigate the growing number of accident-related fatalities, India's overall road traffic death toll is likely to surpass 250,000 by 2025. (A review of the traffic accidents and related practices worldwide, n.d.)

Table 2.3 Comparison International Fatalities. (Singh, 2017)

Country	Motorization Rate (No. of Vehicles per 1,000 People)	Fatality Rate (No. of Fatalities per 10,000 Vehicles	Fatality Risk (No. of Fatalities per 100,000 People)
India (2013)	130	8.6	11.2
Germany (2012)	657	0.67	4.4
Japan (2012)	651	0.63	4.1
New Zealand (2012)	733	0.91	6.9
Sweden (2012)	599	0.50	3.0
United Kingdom (2012)	599	0.51	2.8
United States of America (2012)	846	1.26	10.7

2.2.2 Traffic Accident Trends in Africa

Road environment, social interactions, and vehicle interaction are some of the pre-crash elements that combine to generate road traffic collisions (RTCs). Multiple research projects have been undertaken to look into and comprehend the causes of RTCs in order to develop preventative strategies. Road traffic accident fatalities are regarded as man-made disasters in Ethiopia. The Ethiopian National Road Safety assortment Office reports that there are 114 road accidents casualties for per 10,000 cars annually, however a malfunctioning reporting system may have led to a larger actual number. It has been indicated that many traffic incidents are avoidable and that, in the end, a substantial number of countries face challenges and opportunities related to road safety. Eighty percent of road traffic fatalities globally are in middle-income nations, which also account for eighty percent of the disability-adjusted life years (DALY) brought on by RTAs. Africa has the highest risk of mortality from traffic-related injuries (24.1 per 100,000 people), with pedestrians accounting for 38% of all traffic-related deaths in the landmass.

Due to a number of variables, including growing motorization, there are even more difficulties with African automobiles. Overall, injury-related mortality accounted for 64% of deaths in Egypt, 58% in Tunisia, and 51% in Morocco in 2008. 43 percent of road traffic accidents involving fatalities occurred in Libya, 42 percent in Djibouti, 36 percent in Namibia, and 34 percent in Niger. The age group that is most economically engaged (15–59) is most at risk of dying as a result of RTAs. More than three times as many men as women in this age demographic were victimized by traffic accidents. Road traffic accidents account for 5% of deaths overall for males aged 15 to 59, but in Sub-Saharan Africa, this figure is closer to 6.5 percent for males aged 15 to 29.

(Chimba et al., 2017) analysed road traffic crashes in Anambra State, Nigeria, using the autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average with explanatory variables (ARIMAX) modelling techniques in order to develop accurate predictive models for predicting crash frequency in the State. The ARIMAX model outperformed the ARIMA (1,1,1) model when their performances were assessed using the decreasing Bayesian information standard, mean final percentage error, root mean square error, and larger coefficient of determination (R-Squared). Incorporating human, vehicle, and environmental characteristics into time series analysis of the collision dataset resulted in a more robust prediction model than employing the

aggregate crash count alone, according to the study's findings. This study offered insights on road safety as well as a method for predicting accidents with many cars, people, and environmental variables. If the study's suggestions are followed, the number of traffic accidents in Nigeria will have to go down.

2.2.3 Traffic Accident Trends in Developed Countries

Road accidents have been identified as a leading global source of both physical impairment and fatalities. The World Health Organization's most recent statistics amply demonstrated this proof. As a result, everyone should be concerned in reducing road accidents, since this is a legitimate objective of the Decade of Action for Road Safety (2011-2020), in which the European Union has set a target to decrease fatalities in member states by 50% by the year 2020. In compliance with international standards, Portugal accepted the task of ranking among the top 10 European nations with the lowest accident rate. Looking for zero deaths and zero serious injuries over the long term is the definition of the objective for road safety in Portugal as stated in the recent Ministers Council resolution on 5/2014 and under the 2013-2015 Mid-Term Review. (Chimba et al., 2017)

(Bergman et al., 2018) looked at Scottish military veterans involved in traffic accidents. They examined the probability of road traffic accidents (RTA) in a sizable national cohort of veterans versus those who had never served using data from the Scottish Veterans Health Study. Utilizing survival analysis to compare the risk of RTA injury, a retrospective cohort research including 173,000 non-veterans and 57,000 veterans was carried out. Participants were followed up for a maximum of 30 years. To investigate trends by birth cohort and length of service, subgroup analysis was used. In general, RTA Cox proportional hazards ratio (HR) 1.17, 95 percent confidence intervals (CI) 1.14–1.20, was greater among veterans. Veterans with early service dropouts (those who did not finish primary military training) and those with the shortest service records became more at risk (HR 1.31, 95 percent CI 1.23–1.40). At the initial RTA, 34 was the average age. In contrast to veterans who were born in the 1970s, those who were born in the 1960s had a greater danger. Thus, evidence was shown that veterans in their fourth decade of life continue to have an elevated risk of RTA similar to that seen in current military people.

Using the willingness to pay for WTP-CV method, (Jomnonkwao et al., 2021) calculated the value of statistical injury (VSI) and the value of statistical life (VSL) for motorcyclists in Thailand. This

model has been implemented successfully in several developed nations. The value that people would be willing to pay to lower the chance of dying is measured using the value of risk modification, or WTP technique. This method was mostly predicated on designing surveys to find out how much people would be willing to pay. Nonetheless, there have been very few attempts in Thailand to estimate the costs of traffic accidents using the WTP technique. One is the approach of contingent valuation (CV), which has been used in this study to determine the WTP. Sometimes one adopted method for estimating non-market prices is contingent valuation, or CV. A number of various places, including universities, schools, private businesses, and government buildings, were surveyed using questionnaires throughout Bangkok and the surrounding areas. The sample consisted of 1015 randomly chosen motorcycle riders. Using simple arithmetic means, the mean WTP values were determined, and regression analysis was used to identify the components impacting WTP. Between $0.08 million and $0.10 million was the expected range for the VSI, and between $0.17 million and $0.21 million for the estimated VSL. Motorcycle riders who often used helmets and government officers were more prepared to pay for their reduced chance of death than persons with lower incomes, older people, and male riders.

(Mon et al., 2018) used the conjoint analysis (CA) design approach to estimate the fatal injury costs resulting from traffic accidents in Malaysia by utilizing the WTP. In 13 Malaysian states, 4,000 respondents—including drivers of cars and motorcycles—were questioned. The WTP was determined by using linear regression analysis to the elements that influenced it. Factors such as income, car ownership, race, gender, risk perception, and past accidents all had a statistically significant impact on WTP. According to the CA, the estimated VSL varied from $0.36 million (MYR 1.15 million) to $0.45 million (MYR 1.45 million), and that amount was similar to the VSL from the earlier study that used the CV method.

2.3 ML Algorithms in Traffic Accident Prediction

In the field of traffic accident prediction, one popular machine learning model that has shown promising results is Random Forest. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is particularly effective in handling complex and high-dimensional data, making it suitable for traffic accident prediction which involves multiple variables.

2.3.1 Random Forest

Associative, classification, clustering, prediction, sequential pattern mining, and other techniques are all part of data mining technology. The severity of traffic accidents is predicted in this research using a random forest method. A hybrid classifier algorithm called Random Forest uses many decision trees. One of the best classification methods available today is the random forest, which has clear advantages when processing multidimensional data. Prior to the forecast, the total number of traffic accidents is pre-processed. The random forest technique was used to determine the significance level of 32 traffic accident factors. Lastly, in order to forecast the severity of traffic accidents, identify the elements that contribute most to these incidents.

The random forest algorithm aggregates the results of several decision trees. Random selection is used to choose each dataset, and input characteristics are also randomly chosen. **Figure 2.1** (Flow of random forest algorithm) demonstrates the precise flow of the random forest method, wherein the combination chooses the majority of the classification findings to be the problem's ultimate outcome.

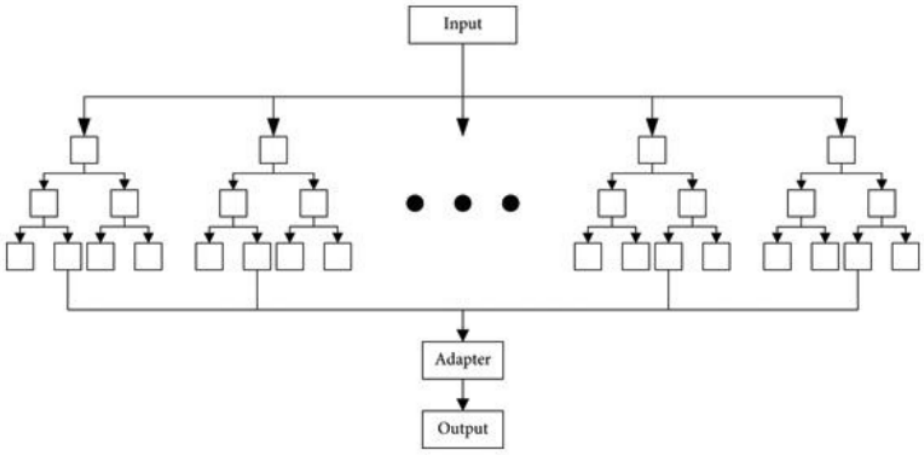


Figure 2.1 Flow of random forest algorithm

The random forest belongs to ensemble learning and adopts the idea of Bagging. Bagging is:

1) To create a new training set, n training samples are taken out of the training set each time.
2) The M sub model is trained using the updated training set.
3) Using the voting approach, the classification problem is solved, and the final category is the classification category of the sub model with the highest number of votes.

A decision tree serves as the fundamental building block of a random forest, which is created by combining many decision trees, it can constitute a random forest. Its construction process is as follows:

i. Step 1: There are N samples overall in T, and N samples are chosen at random to be added back. A decision tree is trained using the chosen N samples as the samples at the decision tree's root nodes.
ii. Step 2: A random selection of M attributes is made from the set of m attributes to satisfy the requirement m << M when each sample contains M attributes and each decision tree node has to be divided. Next, a method (like information gain) is used to choose one of the m attributes to be the node's split feature.

iii. Step 3: Every node in the decision tree should be divided in accordance with Step 2 until it is no longer possible to divide it. Be aware that the entire decision tree creation procedure does not include pruning.

iv. Step 4: To produce a random forest, make a lot of decision trees by following Steps 1 through 3.

For the random forest classification, k samples were chosen from the initial training sample set N using the bootstrap technique. Second, for k samples, the matching decision tree model is created. Ultimately, k samples' worth of findings are voted on, and the final categorization results are chosen using the majority rule approach. The classification decision is as follows:

$$H(x) = \arg\max_{y} \sum_{i=1}^{k} I\left[h_i(x) = Y\right],$$

where H(x) is a combination of classification model; h_i is decision classification model; Y is the output variable (target variable); and $I[h_i(x) = Y]$ is an indicative function.

The random forest algorithm's ability to assess each feature's relative contribution to the prediction is another one of its strong points. By examining the tree nodes that employ a characteristic to minimise impurities in all of the forest's trees, it may determine how important a feature is. After training, it automatically determines each factor's score and adjusts the findings such that the total relevance of all the factors equals one. This meets our desire to forecast the significance of elements related to traffic accidents to a large extent.

2.3.2 Boosting Algorithm

Better solutions are often obtained when combining several machine learning methods to solve a problem. Weak learners are the names given to the individual algorithms. A powerful learner is produced by their combination. A weak learner is a model that performs better than the mean in a regression task or a random prediction in a classification challenge. These algorithms are fitted to the training data and their predictions are combined to get the final result. Voting is used in

classification, whereas averaging is used in regression to determine the combination. Ensemble learning is the term used to describe the blending of many machine learning methods.

A method of ensemble learning called "boosting" fits a dataset to progressively weaker learners. The goal of each successive fitted weak learner is to minimize the mistakes from the preceding one.

Generally, boosting works as follows:

1. Create the initial weak learner.
2. Use the weak learner to make predictions on the entire dataset.
3. Determine the prediction error.
4. Incorrect predictions are assigned more weight.
5. Build another weak learner aimed at fixing the errors of the previous learner.
6. Make assumptions based on the entire dataset using the new learner.
7. Repeat this process until the optimal results are obtained.
8. The mean is weighted to get the final model of all weak learners.

2.3.2.1 CatBoost Algorithm

CatBoost, as its name implies, has two primary functions: gradient boosting and working with categorical data (the Cat) (the Boost). Gradient boosting is a procedure where a large number of decision trees are built repeatedly. Better results are obtained with each successive tree since it enhances the output of the preceding tree. CatBoost is a quicker version of the original gradient boost technique.

CatBoost prevents a drawback of other decision tree-based techniques, which usually need pre-processing of the data to transform category text variables to numerical values, one-hot encodings, and other formats. A mix of categorical and non-categorical explanatory variables can be directly consumed by this approach without the need for preprocessing. Preprocessing is a feature of the algorithm. To encode categorical characteristics, CatBoost employs an approach known as ordered encoding. Ordered encoding uses all of the rows' target statistics before a data point to determine a

value that will take the place of the category characteristic. The usage of symmetric trees is another distinctive feature of CatBoost. This implies that the identical split condition is used by all decision nodes at every depth level.

CatBoost has the potential to be quicker than other techniques like XGBoost. It keeps several of the previous algorithms' characteristics, including normalization, cross-validation, and support for missing values. Both small and huge volumes of data perform well with this approach.

CatBoost works well even when its hyper-parameters are not heavily adjusted. To improve the outcome, you may adjust a few crucial CatBoost settings. The CatBoost manual provides clear explanations for these features, which are simple to adjust. Here are some of the parameters that can be optimized for a better result;

a. cat_ features,
b. one_hot_max_size,
c. learning_rate & n_estimators,
d. max_depth,
e. subsample,
f. colsample_bylevel,
g. colsample_bytree,
h. colsample_bynode,
i. l2_leaf_reg,
j. random_strength.

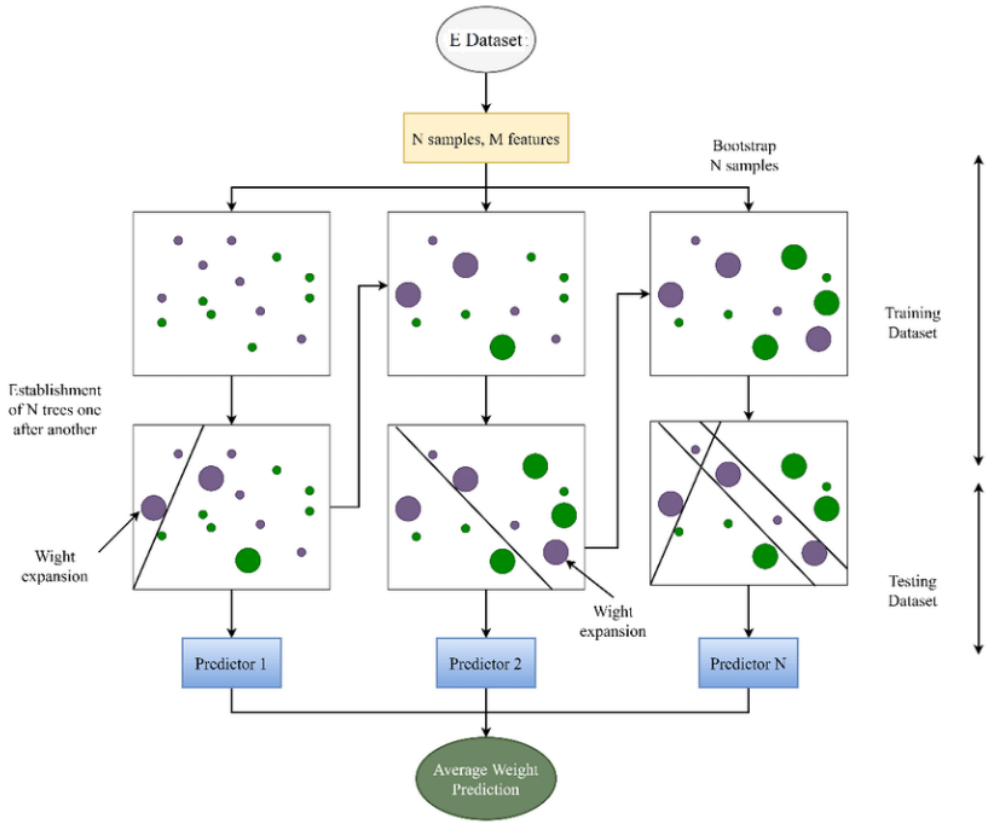


Figure 2.2 Explanation of Catboost Algorithm

2.3.2.2 LightGBM Algorithm

Using a histogram-based approach, LightGBM efficiently uses memory while facilitating faster training by classifying continuous data into discrete bins. It produces decision trees that expand leaf-wise, meaning that, under certain conditions, just one leaf is divided based on the gain. Sometimes, especially with smaller datasets, leaf-wise trees might overfit. Overfitting can be eliminated by limiting the tree depth. A distribution histogram is used by LightGBM, a histogram-based approach, to bucket data into bins. The data is split, the gain is calculated, and iterations are performed using the bins rather than individual data points. Additionally, this approach may be optimized for a sparse dataset. Exclusive feature bundling is another aspect of LightGBM, whereby the algorithm merges distinctive.

It uses two novel techniques:

1. Gradient-based One Side Sampling (GOSS)
2. Exclusive Feature Bundling (EFB)

The histogram-based approach, which is the main component of all GBDT (Gradient Boosting Decision Tree) systems, has constraints that these solutions address. The two GOSS and EFB methods that are discussed below make up the LightGBM Algorithm's features. Together, they enable the model to function well and provide it an advantage over competing GBDT frameworks.

Gradient-based One Side Sampling Technique for LightGBM:

The roles that various data examples play in determining information acquisition are different. The under-trained cases (those with greater gradients) will add more to the information gain. To maintain the accuracy of information gain estimation, GOSS only arbitrarily excludes cases with tiny gradients, leaving only those with significant gradients (e.g., larger than a preset threshold, or among the top percentiles). When the information gain has a wide range in value, this method can produce a more accurate gain estimation than uniformly random sampling at the same desired sample rate.

Gradient boosting is carried out using the GOSS (Gradient-based One-Side Sampling) method using a training set consisting of n instances $\{x1, \cdots, xn\}$, where each instance xi is a vector of dimension s in space Xs. The negative gradients of the loss function with regard to the model's output are represented as $\{g1, \cdots, gn\}$ in each gradient boosting iteration. A subset A is formed by selecting the top-a × 100 percent of the training set instances with the greatest gradients, which are sorted descending based on their absolute gradient values.

For the remaining set Ac, consisting of (1- a) × 100% instances with smaller gradients, a random subset B with a size of b × |Ac| is sampled. The instances are then split based on the estimated variance gain at vector Vj (d) over subset A ? B, where:

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right)$$

Here,

$$Al = x_i \in A : x_{ij} \leq d, \quad Ar = x_i \in A : x_{ij} > d, \quad Bl = x_i \in B : x_{ij} \leq d, \quad Br = x_i \in B : x_{ij} > d$$

The coefficient (1-a)/b is used to normalize the sum of the gradients over B back to the size of Ac.

Exclusive Feature Bundling Technique for LightGBM:

Since high-dimensional data are typically quite sparse, we may propose a practically lossless method for reducing the number of features. To be more precise, a lot of features in a sparse feature space never take nonzero values concurrently since they are mutually incompatible. It is safe to combine the unique traits into a single feature (called an Exclusive Feature Bundle). As a result, when bundle<<feature, the complexity of histogram creation shifts from O(data × feature) to O(data × bundle). Thus, the training framework's speed is increased without compromising precision.

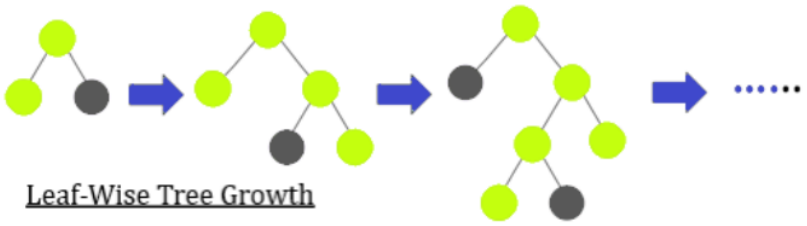


Figure 2.3 LightGBM overview

A few important parameters and their usage are listed below:

(a) max_depth: It sets a limit on the tree's depth. 20 is the default value. It works well to prevent overfitting.

(b) Categorical_feature: The categorical feature that was utilised to train the model is specified.

(c) bagging_fraction: It indicates the percentage of data to be taken into account for every iteration.

(d) num_iterations: The number of iterations that must be completed is specified. 100 is the default value.

(e) num_leaves: It describes how many leaves a tree has. It ought to be less than max depth squared.

(f) max_bin: The maximum number of bins that may be used to bucket the feature values is specified.

(g) min_data_in_bin: It indicates the bare minimum of information in a single bin.

(h) task: Either training or prediction is the goal that it states we want to do. Train is the entry that is set by default. Prediction is another value that might be assigned to this option.

(i) feature_fraction: It indicates the percentage of characteristics that will be taken into account at each iteration. One is the standard value.

2.4 Weather Impact on Accidents and Injuries

Large-scale weather disasters like tornadoes, hurricanes, and flooding are not as common in the United States as weather-related vehicle accidents. Every year, passengers encounter a wide range of weather-related risks, such as blowing dust, fog, ice pavement, rain, snow/sleet, and wet pavement. According to data compiled by the U.S. Department of Transportation (DOT) between 2007 and 2016, there are over 5.8 million car collisions annually. Just over 1.2 million of them, or around 21% of them, featured dangerous weather. According to the DOT, such weather-related car collisions in the United States have claimed the lives of 5,376 individuals on average each year, or around 16% of all vehicular fatalities. Over 418,000 additional people had injuries annually during that time. In contrast, 379 people die annually on average during a ten-year period as a result of heat, lightning, tornadoes, hurricanes, and flooding.

Based on the 10-year averages, the two most common causes of weather-related car collisions were wet pavement and rain, followed by winter weather risks and fog. Since that specific hazard is present in all seasons, it should come as no surprise that rain and wet pavement account for the bulk of fatal weather-related incidents.

2.4.1 Sudden Visibility Reductions Pose Big Danger

Drivers are especially at risk when driving in quickly changing weather conditions. In a matter of seconds, visibility might drop from excellent to just a few feet due to heavy snow, thick fog, or blinding rainfall. Large pileups in the winter are sometimes caused by snow squalls, which are brief but intense focused blasts of snow. Their short-lived but heavy snowfall rates slicken roadways, rapidly reducing visibility and startling drivers.

At any time, there is also a serious risk of reduced visibility due to fog and blowing dust. Cars that follow other cars too closely and drive too quickly for the circumstances are common causes of fog-related accidents. Rapid shifts in vision, when it goes from excellent to nearly negligible in a matter of minutes, can often result in serious accidents.

2.4.2 Even Small Amounts of Ice Can Be Dangerous

Drizzle in below-freezing conditions may not seem like a huge issue, but it may be quite dangerous for unwary drivers. Roads and bridges may become dangerous places to drive on even a little coating of ice.

An instance of this happened early on January 15, 2015, when a small coating of freezing drizzle impacted many states in the Northeast. Hundreds of accidents happened despite the fact that the freezing drizzle was only present in the morning and was only temporary. Despite not being a large storm system, it had a significant impact.

Vehicles pile up at the site of a fatal crash near Fredericksburg, Pennsylvania, Saturday, Feb. 13, 2016. The pileup left tractor-trailers, box trucks and cars tangled together across several lanes of traffic and into the snow-covered median.

Figure 2.4 Fatal accident due to less visibility in Pennsylvania

2.4.3 Weather Impacts on Safety

1. Annually, there are more than 5,891,000 car accidents. Roughly 1,235,000 of these crashes, or 21% of them, are weather-related. Crash incidents categorized as weather-related fall under two categories: those that happen on slick pavement (such as wet pavement, snowy/slushy pavement, or icy pavement) or in unfavorable weather conditions (such as rain, sleet, snow, fog, strong crosswinds, or blowing snow/sand/debris). Every year, weather-related collisions result in around 418,000 injuries and over 5,000 fatalities.
2. The great majority of weather-related collisions occur on wet pavement and during precipitation: 46% and 70% of crashes occur during precipitation, respectively. The proportion of weather-related crashes that happen in the winter is substantially lower: 18% of weather-related collisions happen in the event of snow or sleet, 13% happen on ice pavement, and 16% happen on slushy or snowy pavement. In the presence of fog, just 3% of incidents occur.
3. Adverse weather and/or slick pavement are factors in around 15% of fatal collisions, 19% of injury crashes, and 22% of property-damage-only (PDO) crashes, according to crash type on an annual average. That means that, on a yearly basis, bad weather or slippery pavement cause over 4,900 fatal collisions, over 301,100 injury crashes, and about 919,700 PDO crashes.

Table 2.4 Weather Related Crash Statistics

Table: Weather-Related Crash Statistics (Annual Averages)		
	Weather-Related Crash Statistics	
	10-year Average (2007-2016)	**10-year Percentages**
Weather-Related* Crashes, Injuries, and Fatalities	1,235,145 crashes	21% of vehicle crashes
	418,005 persons injured	19% of crash injuries
	5,376 persons killed	16% of crash fatalities

Table 2.5 Weather Related crashes weather type wise

*** "Weather-Related" crashes are those that occur in the presence of adverse weather and/or slick pavement conditions.**

Table: Weather-Related Crash Statistics (Annual Averages)

Road Weather Conditions	Weather-Related Crash Statistics		
	10 Year Average (2007 - 2016)	10-year Percentages	
Wet Pavement	860,286 crashes	15% of vehicle crashes	70% of weather-related crashes
	324,394 persons injured	15% of crash injuries	78% of weather-related injuries
	4,050 persons killed	12% of crash fatalities	76% of weather-related fatalities
Rain	556,151 crashes	10% of vehicle crashes	46% of weather-related crashes
	212,647 persons injured	10% of crash injuries	51% of weather-related injuries
	2,473 persons killed	8% of crash fatalities	46% of weather-related fatalities
Snow/Sleet	219,942 crashes	4% of vehicle crashes	18% of weather-related crashes
	54,839 persons injured	3% of crash injuries	14% of weather-related injuries
	688 persons killed	2% of crash fatalities	13% of weather-related fatalities
Icy Pavement	156,164 crashes	3% of vehicle crashes	13% of weather-related crashes
	41,860 persons injured	2% of crash injuries	11% of weather-related injuries
	521 persons killed	2% of crash fatalities	10% of weather-related fatalities
Snow/Slushy Pavement	186,076 crashes	4% of vehicle crashes	16% of weather-related crashes
	42,036 persons injured	2% of crash injuries	11% of weather-related injuries
	496 persons killed	2% of crash fatalities	10% of weather-related fatalities
Fog	25,451 crashes	1% of vehicle crashes	3% of weather-related crashes
	8,902 persons injured	1% of crash injuries	3% of weather-related injuries
	464 persons killed	2% of crash fatalities	9% of weather-related fatalities

2.5 Summary

Using traffic volume data as the only explanatory variable for crashes, the methodological approach uses predictive models for specific base conditions. Next, regional or state calibration factors and accident modification factors (AMFs) are applied to estimate the impact of geometric characteristics that differ from the base model conditions on accidents. The use of the algorithm with the base model-AMF approach was thoroughly examined in a recent study conducted for FHWA using a multistate database. The study also looked at alternative base model forms, the use of full models that included variables unrelated to traffic, and alternative methods for estimating AMFs (Lyon et al., 2003). Recent research advocates the use of count models with random parameters as an alternative method for analyzing accident frequencies. (El-Basyouny and Sayed, 2009) study accident prediction models with random corridor parameters. A dataset composed of urban arterials in Vancouver, British Columbia, is considered where the 392 segments were clustered into 58 corridors. (Jagannathan et al., 2013) present research work towards a novel decision support system that predicts in real time when current traffic flow conditions, measured by induction loop sensors, could cause road accidents. Preliminary results from experiments using real-world spatio-temporal traffic flow data and accident data are promising. A small number of recent works have attempted to use deep learning for traffic accident prediction. (Yuan et al., 2018) perform a comprehensive study on the traffic accident prediction problem using the Convolutional Long Short-Term Memory (ConvLSTM) neural network model. Predicting occupational accident risk using both structured and unstructured (text) data is broadly an unexplored area of research.

(Sarkar et al., 2018) propose a methodology that utilizes both text-based clustering, namely Expectation Maximization (EM) algorithm for unstructured text analysis and deep neural network (DNN) for prediction of accident risk using the accident data collected from a steel plant in India. Reducing traffic accidents is an important problem for increasing public safety, so accident analysis and prediction have been a subject of extensive research in recent time (Ovi et al., 2021). To overcome these challenges, (Ovi et al., 2021) propose ARIS: a system for real-time traffic accident prediction built on a traffic accident dataset named ‘US-Accidents’ which covers 49 states of United States, collected from February 2016 to June 2020. South Africa (SA) records high mortality originating from traffic accident annually making the country to be ranked highly among nations with the highest traffic mortality globally.

(Hossain et al., 2021) aim to use machine learning method to predict traffic accident in SA for every hour ranging between 1 January and 31 March 2019 at a segment ID (Hossain et al., 2021) consider various contributing factors and their impact on the prediction of the severity of accidents. This research considered the road accident severity prediction as a classification problem that can classify the intensity of an accident in two categories:

i. Binary classification (grievous and non-grievous),
ii. Multiclass classification (fatal, serious, minor, and non-injury).

Based on millions of traffic accident data in the United States (Zhao and Deng, 2022) build an accident duration prediction model based on heterogeneous ensemble learning to study the problem of accident duration prediction in the initial stage of the accident. First, (Zhao and Deng, 2022) focus on the earlier stage of the accident development, and select some effective information from five aspects of traffic, location, weather, points of interest and time attribute. The Highway Safety Manual (HSM) provides consistent predictive methods for estimating the predicted average crash frequency, but an appropriate calibration is necessary to use them in contexts different from the ones where they were developed. (La Torre et al., 2022) provide a contribution in this field of research providing a European APM based on the one proposed by HSM and introducing a new methodology to transfer the HSM to different European rural freeways.

CHAPTER 3 : RESEARCH METHODOLOGY

3.1 Introduction

This work focuses on the some boosting algorithm like Catboost and LightGBM to predict accident based on weather events and compare the result and metric. The result should be able to generate high accuracy on large dataset with better performance and speed. The methodology used entails important procedures as the choice of target data, pre-processing the selected data, and many others. The following are these steps:

- **Data Collection**: The first step is to choose the target data. We will utilize the LSTW dataset, a sizable dataset of traffic and weather activities for the United States, in this research.
- **Data preprocessing**: Pre-processing the data is the subsequent stage. Data cleansing, outlier removal, and transforming the data into a structured and comprehensible format will be involved in this stage.
- **Class balancing**: The data may be imbalanced, meaning that there are more or fewer samples of one class than another. This can skew the results of the machine learning models. To address this, we will balance the data by oversampling or under sampling the minority class.
- **Feature Engineering**: The following step is to identify the features that will be used to train the machine learning models. This is important because not all features are equally important. We will use a variety of techniques to select the features, such as univariate feature selection and recursive feature elimination.
- **Model training**: The next step is to train the machine learning models. We will use two popular machine learning algorithms, CatBoost and LightGBM, to train our models.
- **Model evaluation**: The final step is to evaluate the machine learning models. This includes evaluating the accuracy, precision, and recall of the models. We will also conduct sensitivity analysis to determine the outcomes of different factors on the accuracy of the models and compare the models of Catboost and LightGBM to find the best fit.

3.2 Algorithms & Techniques

3.2.1 State-of-the-Art Random Forest Boosting

Some of the latest State-of-the-Art Random Forest Boosting Models available for faster implementation and better result are discussed below:

- XGBoost
- LightGBM
- Catboost
- AdaBoost
- Gradient Bosting

3.2.2 Boosting Algorithm

- Catboost Algorithm
- LightGBM Algorithm

3.3 Methodology

3.3.1 End-to-End Pipeline

The end-to-end pipeline for the methodology is shown in Figure 3.1

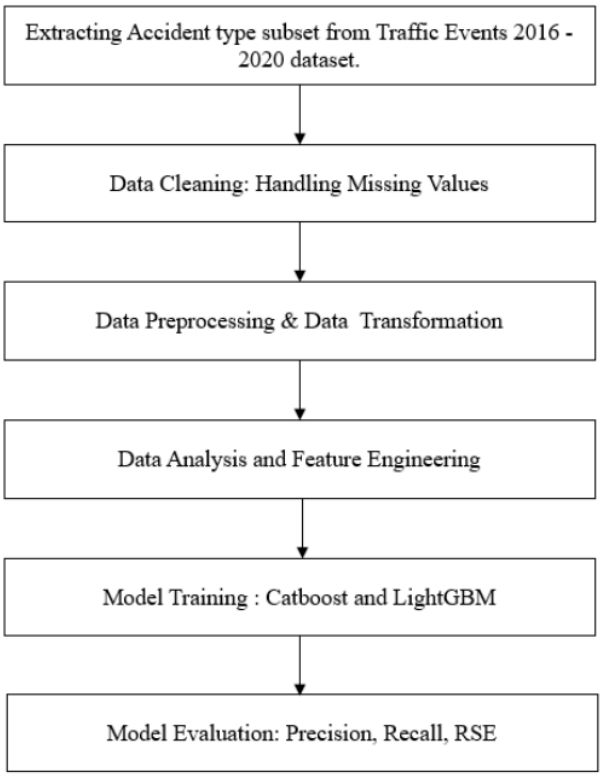


Figure 3.1 E2E Pipeline

3.3.2 Data Selection

Finding the right data kind, source, and instrument(s) to enable researchers to sufficiently address research issues is the main goal of data selection. This decision is often discipline-specific and is mostly influenced by the type of inquiry, the body of prior research, and the ease of access to relevant data sources.

This work makes use of two datasets:

Traffic Events: Introduced by (Moosavi et al., 2019a), this dataset contains ~31.4 million of traffic events. This dataset is a spatiotemporal entity, where such an entity is associated with location and time. Following is the description of available traffic event types:

- Accident: refers to a traffic collision that may include one or more cars.
- Broken-Vehicle: refers to the circumstance in which a vehicle, or vehicles, are disabled on a road.
- Congestion: refers to the circumstance in which traffic is moving more slowly than anticipated.
- Construction: refers to a road's ongoing maintenance, repair, or construction project.
- Event: refers to events like congestion, accidents, and constructions.
- Lane-Blocked: refers to situations where traffic or bad weather has forced us to close a lane or lanes.
- Flow-Incident: Refers to all other types of traffic events. Examples are broken traffic light and animal in the road.

Weather Events: Introduced by (Moosavi et al., 2019a), this dataset contains ~5.6 million events. This dataset is a spatiotemporal entity, where such an entity is associated with location and time. Following is the description of available weather event types:

- Severe-Cold: The case of having extremely low temperature, with temperature below -23.7 degrees of Celsius.
- Fog: The case where there is low visibility condition as a result of fog or haze.
- Hail: The case of having solid precipitation including ice pellets and hail.
- Rain: The case of having rain, ranging from light to heavy.
- Snow: The case of having snow, ranging from light to heavy.

- Storm: The extremely windy condition, where the wind speed is at least 60 km/h.
- Other Precipitation: Any other type of precipitation which cannot be assigned to previously described event types.

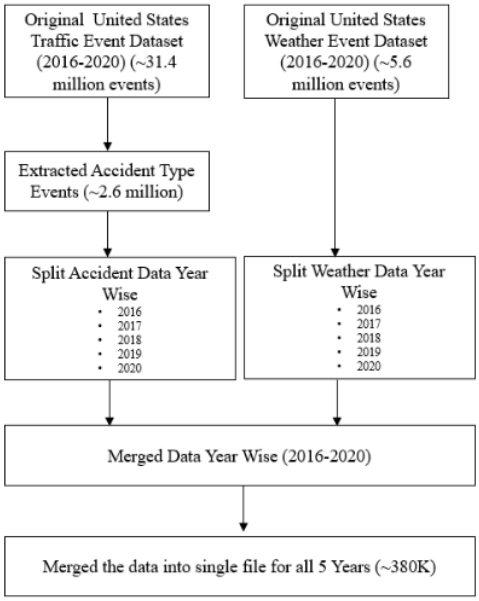


Figure 3.2 Data Selection

3.3.3 Data Cleaning

In the machine learning (ML) pipeline, data cleaning is an essential phase that entails finding and eliminating any duplicate, irrelevant, or missing data. Ensuring that the data is reliable, consistent, and error-free is the aim of data cleaning, as inconsistent or inaccurate data can have a detrimental effect on the performance of the machine learning model. Because raw data is frequently noisy, inaccurate, and inconsistent, data cleaning is crucial to ensuring the correctness and dependability of insights gleaned from it.

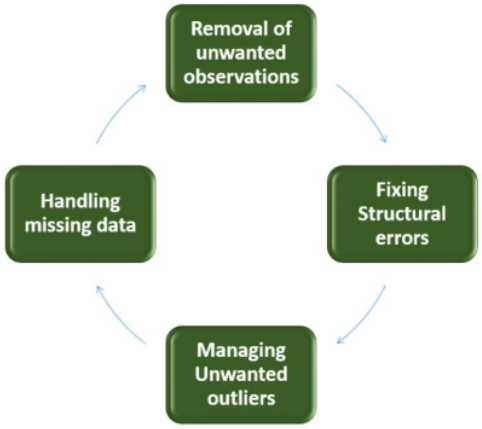


Figure 3.3 Data Cleaning

Below figure 3.4 the missing data present in the data.

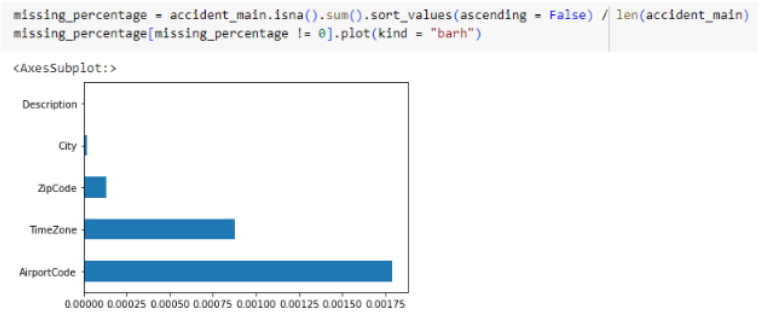


Figure 3.4 Null Values

3.3.4 Data Transformation

3.3.4.1 Centering and scaling:

Scaling data to a single unit-less scale is a highly common and significant change. This may be roughly understood as converting variables from whatever units they are measured in (diamond depth %, for example) into units known as "standard deviations away from the mean" (also known as "standard units," or "z-score").

score). Given data $x = x_1, x_2, \ldots, x_n$, the transformation applied to obtain centered and scaled variable z is:

$$z_i = \frac{(x_i - \overline{x})}{\mathrm{sd}(x)}$$

where $\overline{x}$ is the mean of data x, and $\mathrm{sd}(x)$ is its standard deviation.

Applying this transformation to a dataset's variables has the benefit of putting all of the variables in the same, comparable units. On occasion, you will have use to apply transformations that only centre (but not scale) data:

$$z_i = (x_i - \overline{x})$$

apply transformations that only scale (but not centre) data:

$$z_i = \frac{x_i}{\mathrm{sd}(x)}$$

```
# Create a MinMaxScaler object
scaler = MinMaxScaler()

# Scale all of the dummy variables
X = scaler.fit_transform(merged_data_dummy_backup)
```

Figure 3.5 Using MinMaxScaler for Scaling Data

3.3.4.2 Treating categorical variables as numeric:

We would need to transform categorical variables into something that we can treat as numeric.

```
cat_columns = ['TimeZone','Weekday', 'Weather_Type', 'Weather_Severity','City','State']

# Create dummy variables for the categorical columns
dummy_variables = []
for column in cat_columns:
    dummy_variables.append(pd.get_dummies(merged_data_dummy[column], prefix=column))

# Concatenate the dummy variables to the merged_data DataFrame
merged_data_dummy = pd.concat([merged_data_dummy] + dummy_variables, axis=1)

# Drop the original categorical columns from the merged_data DataFrame
merged_data_dummy = merged_data_dummy.drop(cat_columns, axis=1)
```

Figure 3.6 Creating dummies for Categorical Variables

```
merged_data_dummy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 380746 entries, 0 to 86285
Columns: 1249 entries, Severity to State_WY
dtypes: category(2), float64(2), int32(4), int64(1), uint8(1240)
memory usage: 472.3 MB
```

3.3.4.3 Removing Duplicate rows:

Another parameter during data pre-processing is to check for any duplicated rows.

```
duplicate_rows = accident_main[accident_main.duplicated()]

print("\nDuplicate Rows:")
print(duplicate_rows)
```

```
Duplicate Rows:
Empty DataFrame
Columns: [Severity, StartTime(UTC), EndTime(UTC), TimeZone, LocationLat, LocationLng, AirportCode, City, State, ZipCode, ID, Weekday, month, year]
Index: []
```

3.3.4.4 Relationship between numeric variable:

Correlation is a statistical measure that expresses the strength of the relationship between two variables. The two main types of correlation are positive and negative. Positive correlation occurs when two variables move in the same direction; as one increases, so do the other.

```
plt.figure(figsize=[7,7])
sns.heatmap(cat_data_corr,annot=True,cmap='Greens')
plt.show()
```

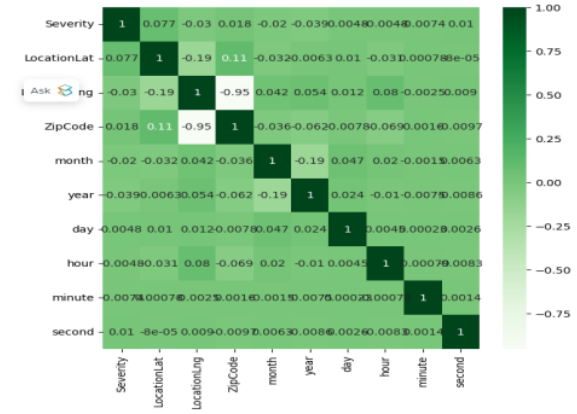


Figure 3.7 Correlation heatmap

3.3.5 Data Mining

It is an essential step in research analysis. The primary aim is to examine data for distribution, outliers and anomalies to direct specific testing for the hypothesis. It also provides thorough graphical representation.

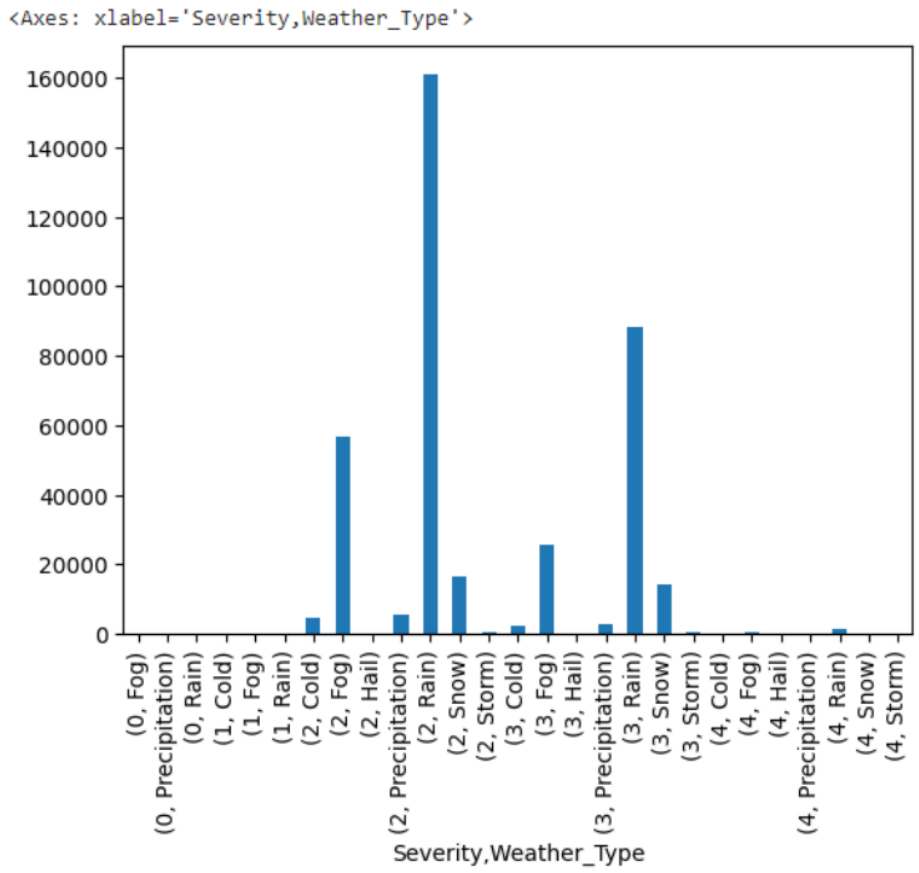


Figure 3.8 Analysis of count of Severity and weather type

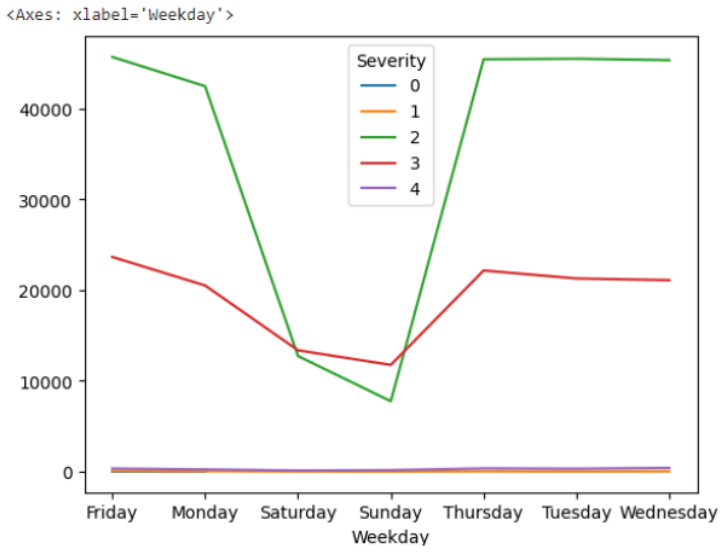


Figure 3.9 Weekday versus accident severity

3.3.6 Evaluation:

The evaluation metrics we used in order to get the performance and result of the model trained are as follows:

a) R2 score
b) Root Mean Squared Error
c) Precision
d) Recall

3.3.6.1 R2 score

The coefficient of determination (R^2) measures how well a statistical model predicts an outcome. The outcome is represented by the model's dependent variable. The lowest possible value of R^2 is 0 and the highest possible value is 1. Put simply, the better a model is at making predictions, the closer its R^2 will be to 1.

3.3.6.2 Root Mean Squared Error

The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points.

RMSE quantifies how dispersed these residuals are, revealing how tightly the observed data clusters around the predicted values.

$$\mathbf{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

$$RMSError = \sqrt{1 - r^2} SD_y$$

Where SD_y is the standard deviation of Y.

3.3.6.3 Precision

Precision is the ratio between the True Positives and all the Positives. For our problem statement, that would be the measure of patients that we correctly identify as having a heart disease out of all the patients actually having it. Mathematically:

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

3.3.6.4 Recall

The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart disease. Mathematically

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{(True Positive + False Positive + True Negative + False Negative)}}$$

3.4 Tools

The following resources are necessary for this study:

1. **Data**: The LSTW dataset is freely available. (https://smoosavi.org/datasets/lstw) shown in figure 3.10 and figure 3.11

#	Attribute	Description	Nullable
1	EventId	This is the identifier of a record	No
2	Type	The type of an event: examples are *accident* and *congestion*.	No
3	Severity	The severity is a value between 0 and 4, where 0 indicates the least impact on traffic (i.e., short delay as a result of the event) and 4 indicates a significant impact on traffic (i.e., long delay).	No
4	TMC	Each traffic event has a Traffic Message Channel (TMC) code which provides a more detailed description on type of the event.	No
5	Description	The natural language description of an event.	No
6	StartTime (UTC)	The start time of an event in UTC time zone.	No
7	EndTime (UTC)	The end time of an event in UTC time zone.	No
8	TimeZone	The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).	No
9	LocationLat	The latitude in GPS coordinate.	Yes
10	LocationLng	The longitude in GPS coordinate.	Yes
11	Distance (mi)	The length of the road extent affected by the event.	Yes
12	AirportCode	The closest airport station to the location of a traffic event.	Yes
13	Number	The street number in address record.	Yes
14	Street	The street name in address record.	Yes
15	Side	The relative side of a street (R/L) in address record.	Yes
16	City	The city in address record.	Yes
17	County	The county in address record.	Yes
18	State	The state in address record.	Yes
19	ZipCode	The zipcode in address record.	Yes

Figure 3.10 Traffic Event Dataset

#	Attribute	Description	Nullable
1	EventId	This is the identifier of a record	No
2	Type	The type of an event: examples are *rain* and *snow*.	No
3	Severity	The severity of an event, wherever applicable.	Yes
4	StartTime (UTC)	The start time of an event in UTC time zone.	No
5	EndTime (UTC)	The end time of an event in UTC time zone.	No
6	TimeZone	The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).	No
7	LocationLat	The latitude in GPS coordinate.	Yes
8	LocationLng	The longitude in GPS coordinate.	Yes
9	AirportCode	The airport station that a weather event is reported from.	Yes
10	City	The city in address record.	Yes
11	County	The county in address record.	Yes
12	State	The state in address record.	Yes
13	ZipCode	The zipcode in address record.	Yes

Figure 3.11 Weather Event Dataset

2. **Software**: Data pre-processing, feature engineering, and machine learning will all require the usage of software. Python 3, CatboostRegressor, LightGBMRegressor is used along with libraries in Figure 3. Libraries used. We have used Google Collab Notebook Pro Version to handle the required RAM and Memory to handle large dataset as shown in figure 3.12.

```
# importing libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import catboost as cat
import lightgbm as lgb
import warnings
warnings.filterwarnings('ignore')

import statsmodels
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
import sklearn

from sklearn.model_selection import train_test_split
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
```

```
import folium
from folium.plugins import HeatMap
```

Figure 3.12 Libraries

3. **Hardware**: A computer with adequate memory and processing capability is required for this study. In this study we used 51.0 GB RAM and 166.8 GB memory as shown in figure 3.13.

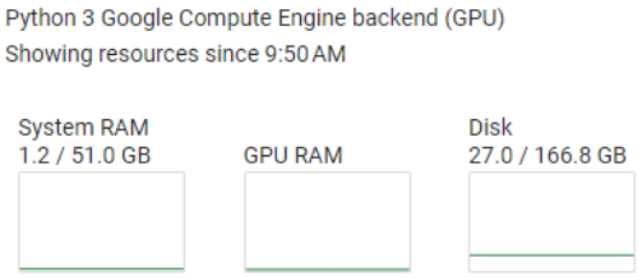


Figure 3.13 Hardware

3.5 Summary

This chapter has covered the following topics in the research methodology:

Data cleaning is the process of identifying and correcting errors and inconsistencies in data. This is an important step in any data mining project, as it ensures that the data is of high quality and can be used to produce reliable results.

- Identifying and removing duplicates: Duplicate records can arise for a variety of reasons, such as human error or data entry errors. Identifying and removing duplicates is an important step in data cleaning, as it ensures that each record is unique.
- Handling missing values: Missing values can occur for a variety of reasons, such as non-response or incomplete data collection. It is important to handle missing values in a way that does not bias the results of the analysis.
- Correcting errors: Errors in data can arise for a variety of reasons, such as human error or data entry errors. It is important to identify and correct errors in the data before it is used for analysis.

Data pre-processing is the process of transforming data into a format that is suitable for data mining. This may involve converting data types, scaling data, and discretizing data.

- Normalization: Normalization is the process of scaling data to a specific range, such as 0 to 1 or -1 to 1. This is useful for comparing data values that are measured on different scales.
- Standardization: Standardization is the process of transforming data so that it has a mean of 0 and a standard deviation of 1. This is useful for comparing data values that are measured on different scales.

Data transformation is the process of converting data into a format that is more suitable for data mining. This may involve combining attributes, creating new attributes, or removing attributes.

- Feature selection: Feature selection is the process of selecting a subset of the attributes that are most informative for the data mining task. This can help to improve the efficiency and accuracy of the data mining algorithm.

Data mining is the process of extracting knowledge from data. There are a variety of data mining algorithms that can be used for different tasks, such as classification, clustering, and association rule mining.

Once a data mining model has been built, it is important to evaluate its performance on a held-out test set. This helps to ensure that the model is not overfitting the training data.

- Precision: Precision is the percentage of data points that are classified as positive that are actually positive.
- Recall: Recall is the percentage of positive data points that are correctly classified by the model.

In this chapter, we have covered the essential steps of the research. These steps include data cleaning, data preprocessing, data transformation, data mining, and evaluation.

CHAPTER 4 : ANALYSIS

4.1 Introduction

In this chapter, we delve into the heart of our study, where we analyze the data and outcomes derived from the application of the LightGBM and Catboost algorithms to predict traffic accidents based on weather conditions. The previous chapters have laid the groundwork, establishing the significance of the problem, outlining the research objectives, presenting the methodology, and detailing the dataset used. Now, our focus shifts to the thorough examination of results and the insights gained from this innovative approach.

In this chapter we would cover the below:

1. Dataset
2. Data Preparation
    a. Handling Missing Values
    b. Transforming Categorical Variables
    c. Elimination of Variables
    d. Splitting of Original Dataset
    e. Univariate Analysis
    f. Class Balancing
3. Bivariate Analysis
4. Data Visualization

4.2 Dataset Overview

As discussed above in the Methodology chapter we have used two datasets: Traffic Events and Weather Events containing around 31.4 million records in traffic events dataset and around 6 million in weather events dataset from the year 2016-2020 from all across Contiguous United States.

Traffic event dataset contains several types of events like congestion, lane-blocked, construction, accident etc. We have used the subset of it containing only the accident type of events containing around 3 million records. We have then merged the Weather event dataset and traffic accident dataset to form single dataset as shown in figure 4.1 which we have utilized for our study further and applied the different algorithm to investigate further.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 380746 entries, 0 to 86285
Data columns (total 16 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Severity          380746 non-null  int64
 1   StartTime(UTC)    380746 non-null  object
 2   EndTime(UTC)      380746 non-null  object
 3   TimeZone          380746 non-null  object
 4   LocationLat       380746 non-null  float64
 5   LocationLng       380746 non-null  float64
 6   AirportCode       380746 non-null  object
 7   City              380746 non-null  object
 8   State             380746 non-null  object
 9   ZipCode           380746 non-null  int64
 10  Weekday           380746 non-null  object
 11  month             380746 non-null  int64
 12  year              380746 non-null  int64
 13  date              380746 non-null  object
 14  Weather_Type      380746 non-null  object
 15  Weather_Severity  380746 non-null  object
dtypes: float64(2), int64(4), object(10)
memory usage: 49.4+ MB
```

Figure 4.1 Merged Data columns overview

	Severity	StartTime(UTC)	EndTime(UTC)	TimeZone	LocationLat	LocationLng	AirportCode	City	State	ZipCode	Weekday	month	year	date	Weather_Type	Weather_Severity
0	2	2016-12-01 03:38:32	2016-12-01 04:08:21	US/Pacific	39.169224	-123.21139	KUKI	Ukiah	CA	95482	Thursday	12	2016	2016-12-01	Fog	Severe
1	2	2016-12-01 03:57:44	2016-12-01 04:27:26	US/Pacific	37.371304	-120.61373	KMER	Atwater	CA	95301	Thursday	12	2016	2016-12-01	Fog	Moderate
2	2	2016-12-01 05:13:53	2016-12-01 05:43:53	US/Pacific	37.500683	-122.24131	KSQL	San Carlos	CA	94070	Thursday	12	2016	2016-12-01	Cold	Severe
3	2	2016-12-01 19:34:31	2016-12-01 20:03:44	US/Pacific	38.243504	-122.26866	KAPC	Napa	CA	94558	Thursday	12	2016	2016-12-01	Fog	Severe
4	2	2016-12-01 19:34:31	2016-12-01 20:03:44	US/Pacific	38.243504	-122.26866	KAPC	Napa	CA	94558	Thursday	12	2016	2016-12-01	Fog	Severe

Figure 4.2 Overview

This dataset shown in figure 4.2 provides a rich source of information for analyzing the relationship between the traffic accident and the weather events. The combination of detailed accident data and comprehensive weather data enables to identify patterns and correlations that can inform the development of effective road safety strategies.

4.3 Data Preparation

Data analysis is the process of bringing order, structure and meaning to the mass of collected data. It is described as messy, ambiguous and time-consuming, but also as a creative and fascinating process. Broadly speaking - while it does not proceed in linear fashion -it is the activity of making sense of, interpreting and theorizing data that signifies a search for general statements among categories of data. In this section, the captured data from the quantitative research is presented, analyzed, described and interpreted in a systematic manner as the next step of the research process. The documentation and analysis process aimed to present data in an intelligible and interpretable form in order to identify trends and relations in accordance with the research aims and objectives.

4.3.1 Handling Missing Values

In this section, we will discuss the various types of missing data, their potential impact on analysis, and strategies for handling them effectively.

4.3.1.1 Types of Missing Data

Missing data can be categorized into three main types based on the mechanism underlying their occurrence:

- *Missing Completely At Random (MCAR):* Missing values occur completely at random, independent of any other variables in the dataset.
- *Missing At Random (MAR):* Missing values are related to some other variables in the dataset but not to the outcome variable of interest.
- *Missing Not At Random (MNAR):* Missing values are related to the outcome variable of interest, indicating that the pattern of missingness is informative.

```
EventId                0
Type                   0
Severity               0
TMC                    0
Description            2
StartTime(UTC)         0
EndTime(UTC)           0
TimeZone            2305
LocationLat            0
LocationLng            0
Distance(mi)           0
AirportCode         4716
Number           1588852
Street                 0
Side                   0
City                  51
County                 0
State                  0
ZipCode              349
dtype: int64
```

Figure 4.3 Missing Values in Traffic Event Dataset

```
EventId                0
Type                   0
Severity               0
StartTime(UTC)         0
EndTime(UTC)           0
TimeZone               0
LocationLat            0
LocationLng            0
AirportCode            0
City               10784
County                 0
State                  0
ZipCode            43630
dtype: int64
```

Figure 4.4 Missing Values in Weather Event Dataset

4.3.1.2 Impact of Missing Data

Missing data can have a detrimental impact on statistical analysis in several ways:

- *Reduced Statistical Power:* Missing values reduce the sample size available for analysis, thereby diminishing the power of statistical tests to detect true differences or relationships.
- *Biased Estimates*: If missing values are not handled appropriately, they can introduce bias into parameter estimates, leading to inaccurate conclusions.
- *Incomplete Understanding:* Missing data can obscure important patterns and relationships within the data, hindering our understanding of the underlying phenomena.

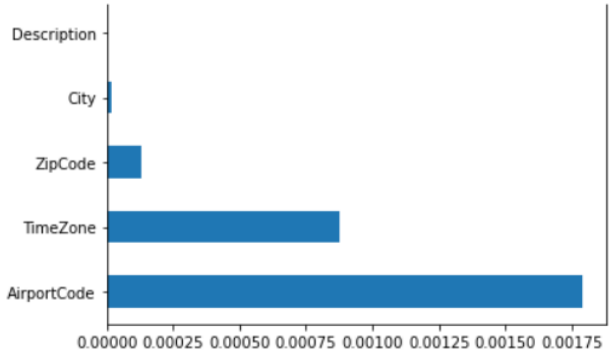


Figure 4.5 Missing % in bar Chart

4.3.1.3 Strategies for Handling Missing Data

Several techniques can be employed to address missing values depending on the type of missing data and the characteristics of the dataset:

- *Deletion:* Deleting cases with missing values is a straightforward approach but can lead to a loss of information and potential bias. It is generally only recommended for MCAR data.
- *Imputation:* Imputation involves replacing missing values with estimated values. Various imputation methods exist, such as mean or median imputation, hot-deck imputation, and multivariate imputation by chained equations (MICE).
- *Multiple Imputation*: Multiple imputation involves creating multiple imputed datasets by replacing missing values with different plausible estimates. This approach provides a more robust assessment of uncertainty due to missing data.
- *Modeling Missingness Mechanism:* If the missingness mechanism is suspected to be MNAR, incorporating it into the statistical model can provide more accurate estimates and reduce bias.
- *Sensitivity Analysis*: Sensitivity analysis involves assessing the impact of different missing data handling techniques on the results of the analysis. This helps determine the robustness of conclusions to the assumptions made about missing data.

By addressing missing data in a comprehensive and transparent manner, we have ensured the validity and reliability of our findings and enhance the credibility of our study.

4.3.2 Transforming Categorical Variables

Categorical variables, also known as nominal variables, represent discrete values that cannot be ordered numerically. Most of the algorithms typically require numerical input, so transforming categorical variables into numerical representations is essential. Several techniques can be employed to transform categorical variables, each with its own advantages and limitations.

One-Hot Encoding

One-hot encoding is a widely used technique that creates a new binary variable for each unique category in the categorical variable. For instance, if a variable represents gender with categories "male" and "female", one-hot encoding would create two new binary variables: "is_male" and "is_female". The value of each binary variable is 1 if the corresponding category is present and 0 otherwise.

Label Encoding

Label encoding assigns a numerical value to each unique category in the categorical variable. The numerical values are typically integers, ranging from 0 to the number of categories minus 1. For example, if a variable represents color with categories "red", "green", and "blue", label encoding could assign 0 to "red", 1 to "green", and 2 to "blue".

Ordinal Encoding

Ordinal encoding is applicable when the categories in the categorical variable have an inherent order. For instance, a variable representing educational level with categories "high school", "bachelor's degree", and "master's degree" could be encoded using ordinal encoding.

Mean Encoding

Mean encoding replaces each category with the mean of the target variable for that category. This technique is particularly useful when there is a significant correlation between the categorical variable and the target variable.

Transforming categorical variables into dummy variables, also known as one-hot encoding, is a common practice in machine learning and data analysis. It involves creating a new binary variable for each unique category in the categorical variable shown in figure 4.8. In our data preparation, we have used one-hot encoding to convert categorical into 0 and 1 format followed by MinMaxScaler Scaling.

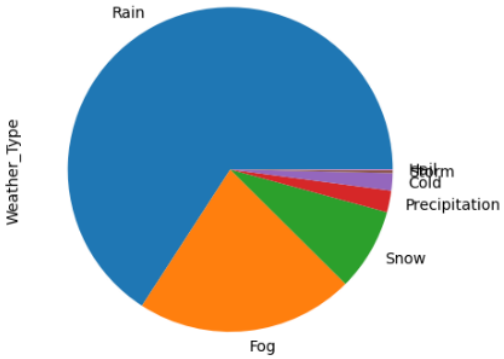


Figure 4.6 Weather Type Distribution Pie Chart

```
Rain             65.852563
Fog              21.695566
Snow              8.213087
Precipitation     2.148414
Cold              1.740005
Storm             0.261592
Hail              0.088773
Name: Weather_Type, dtype: float64
```

Figure 4.7 Weather Type % distribution

	Severity	LocationLat	LocationLng	AirportCode	ZipCode	month	year	day	hour	minute	...	State_SC	State_TN	State_TX	State_UT	State_VA	State_VT	State_WA	State_WI	State_WV	State_WY
0	2	39.169224	-123.21139	KUKI	95482	12	2016	1	3	38	...	0	0	0	0	0	0	0	0	0	0
1	2	37.371304	-120.61373	KMER	95301	12	2016	1	3	57	...	0	0	0	0	0	0	0	0	0	0
2	2	37.500683	-122.24131	KSQL	94070	12	2016	1	5	13	...	0	0	0	0	0	0	0	0	0	0
3	2	38.243504	-122.26866	KAPC	94558	12	2016	1	19	34	...	0	0	0	0	0	0	0	0	0	0
4	2	38.243504	-122.26866	KAPC	94558	12	2016	1	19	34	...	0	0	0	0	0	0	0	0	0	0

5 rows × 1251 columns

Figure 4.8 Transformation of Categorical Variables using Dummies

4.3.3 Elimination of Variables

Eliminating variables is a valuable technique for improving the performance, interpretability, and computational efficiency of machine learning models. By carefully selecting and removing irrelevant, redundant, or noisy variables, researchers and data scientists can gain a deeper understanding of the data and develop more robust predictive models.

Reasons for Eliminating Variables

1. *Irrelevance*: Some variables may not have any predictive power or correlation with the target variable, making them irrelevant for the analysis.
2. *Redundancy*: Multiple variables may capture the same information, leading to redundancy and potential overfitting in machine learning models.
3. *Noise*: Variables with high noise levels or outliers can obscure the true relationships between variables and the target variable.

Benefits of Eliminating Variables

1. *Improved Model Performance*: Removing irrelevant or redundant variables can reduce noise and prevent overfitting, leading to more accurate and generalizable predictive models.
2. *Reduced Computational Complexity*: By eliminating variables, the dimensionality of the dataset is reduced, which can significantly improve the efficiency of machine learning algorithms.

4.3.4 Splitting of Original Dataset

To prepare the dataset for analysis, we employed a multi-stage data preparation and feature engineering process. Initially, we extracted a subset of 'Accident' type traffic events from the comprehensive traffic incident dataset. Subsequently, this subset was divided into year-wise divisions to facilitate the synchronization and merging of traffic and weather data based on various

parameters. The resulting year-wise datasets were then merged into a single comprehensive dataset for further analysis.

4.3.5 Class Balancing

Class balancing is a technique used to address the problem of imbalanced class distribution in machine learning datasets. An imbalanced class distribution occurs when one class (the majority class) has significantly more data points than the other class (the minority class). This imbalance can lead to biased or inaccurate results, especially for classifiers that are trained on imbalanced data.

There are several different techniques that can be used to class balance datasets. Some common techniques include:

1. *Oversampling*: This technique involves replicating data points from the minority class until the majority and minority classes are equal in size. This can be done by simply duplicating data points, or by using more sophisticated techniques such as synthetic minority oversampling technique (SMOTE).

2. *Undersampling*: This technique involves removing data points from the majority class until the majority and minority classes are equal in size. This can be done by randomly removing data points, or by using more sophisticated techniques that select data points for removal based on their contribution to the overall error of the classifier.

We have utilized both Under sampling technique to balance the majority class with the minority class in order to extract feature importance and then apply it to the entire set of data, also to improve the performance of classifiers on imbalanced data. However, it is very important to choose the right technique for the specific dataset and classifier being used.

4.3.6 Univariate Analysis

Univariate analysis is a statistical approach that examines and describes individual variables in a dataset. It focuses on understanding the distribution, central tendency, and variability of each variable without considering the relationships between variables. Univariate analysis is often the first step in data analysis, providing a foundation for further exploration and modelling.

Purpose of Univariate Analysis

1. *Data Exploration*: Univariate analysis helps researchers understand the nature and characteristics of each variable in the dataset. It provides insights into the range of values, the frequency distribution, and the central tendency of each variable.
2. *Data Cleaning*: Univariate analysis can identify potential data errors, outliers, and missing values. By examining the distribution and characteristics of each variable, researchers can flag anomalies and take appropriate measures to clean the data.
3. *Feature Selection*: Univariate analysis can aid in feature selection, the process of identifying the most relevant and informative variables for further analysis or modelling. By assessing the variance, skewness, and correlation of variables, researchers can prioritize those that provide the most valuable information.
4. *Hypothesis Generation*: Univariate analysis can generate initial hypotheses about the relationships between variables and the overall trends in the data. These hypotheses can then be further investigated through multivariate analysis and statistical modelling.

Key Elements of Univariate Analysis

1. *Frequency Distribution*: The frequency distribution shows the number of times each value of a variable occurs in the dataset. It provides a visual representation of the distribution of data.
2. *Central Tendency*: Central tendency measures represent the average or middle value of a dataset. Common measures of central tendency include mean, median, and mode.
3. *Variability*: Variability measures indicate how spread out the data is around the central tendency. Common measures of variability include range, variance, and standard deviation.

4. *Data Visualization*: Data visualization techniques, such as histograms, box plots, and bar charts, are commonly used to represent univariate analysis results. These visualizations make it easier to understand the distribution, central tendency, and variability of each variable.

Below are some univariate insights from our study:

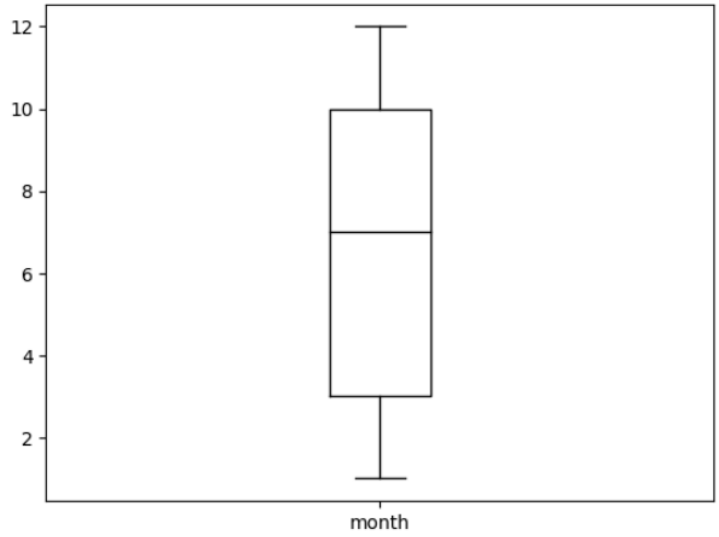


Figure 4.9 Boxplot for month variable

We can see that in figure 4.9 there is no outlier in the above boxplot, accidents are spread across entire year. Overall, the box plot suggests that there is a seasonal pattern to the number of accidents, with more accidents occurring in the later months of the year. This could be due to a number of factors, such as weather conditions or holidays.

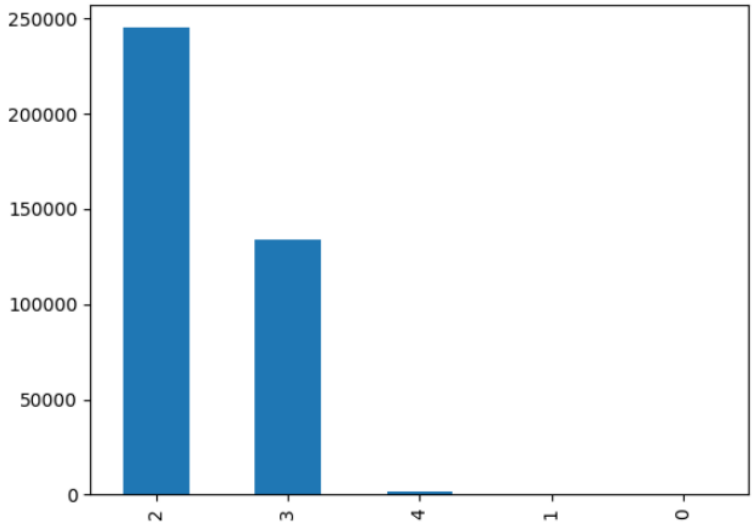


Figure 4.10 Bar Chart for Severity variable

In Figure 4.10 the majority of accidents have a minimal impact on traffic flow, with Severity 2 incidents being the most common. A small number of accidents result in more significant delays and congestion.

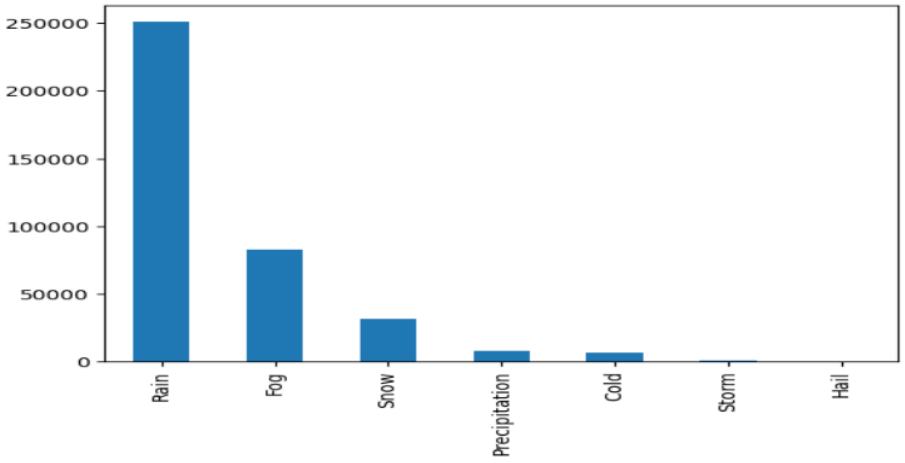


Figure 4.11 Weather Type Distribution

As per Figure 4.11 rain is the most common weather condition associated with accidents, likely due to reduced visibility and slippery roads. Fog and snow are also associated with an increased risk of accidents, but to a lesser extent. Cold and stormy weather conditions are associated with the fewest accidents, likely because people are more aware of these conditions and take precautions to avoid driving.

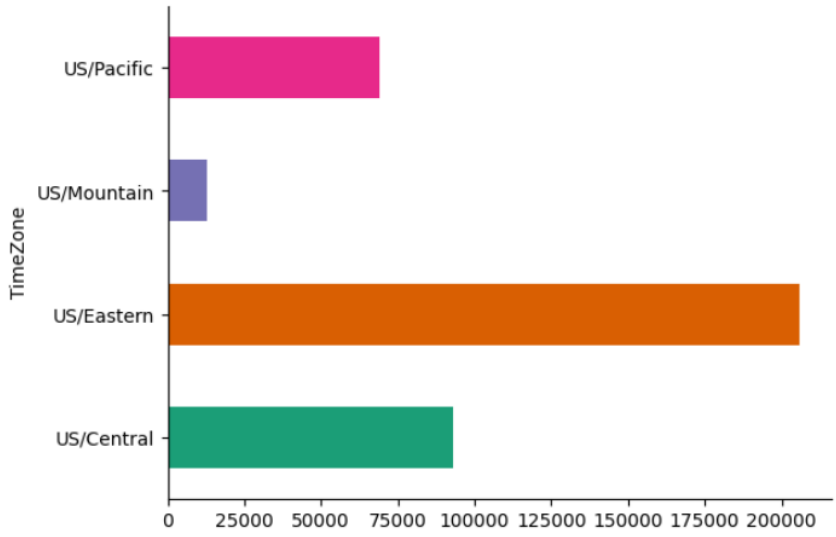


Figure 4.12 Time zone Distribution

As per Figure 4.12 there are a few reasons why the Eastern Time Zone (EST) faces more accidents than other time zones in the US.

1. *Population density*: It is home to the most people in the US, with over 150 million people living in the zone. This means that there are simply more people on the roads in the EST, which increases the likelihood of accidents.
2. *Infrastructure*: It is home to some of the oldest and most congested highways in the US. This can make it difficult for drivers to navigate safely, and it can also lead to more accidents.
3. *Weather*: It experiences a wide range of weather conditions, from snow and ice in the winter to heat and humidity in the summer. These conditions can make it difficult for drivers to see and react to hazards, and they can also lead to more accidents.

4. *Driving habits*: Drivers in the EST may be more likely to engage in risky driving behaviors, such as speeding, tailgating, and driving under the influence of alcohol or drugs. These behaviors can increase the likelihood of accidents.

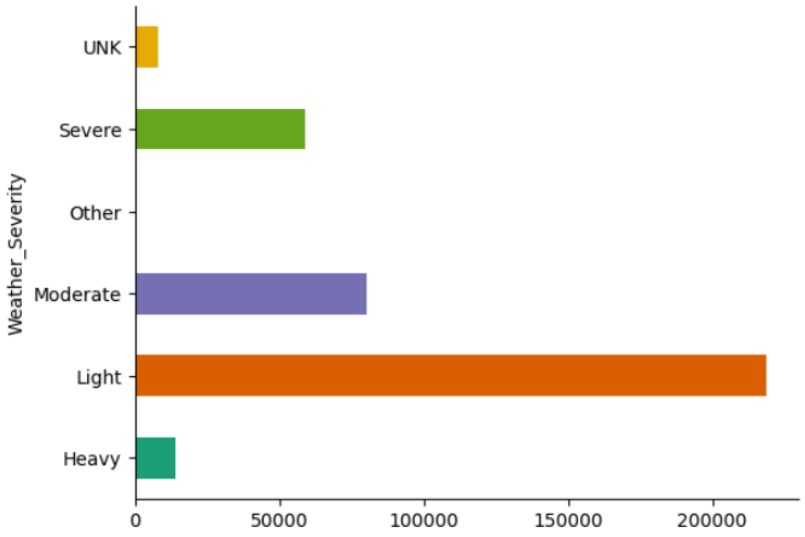


Figure 4.13 Weather Severity Distribution

As per Figure 4.13 Most accidents in the US happened when the weather severity is light due to the fact:

1. *Visibility*: In light weather conditions, it is easier for drivers to see the road and other vehicles. This reduces the risk of accidents caused by rear-ending, lane changes, and other types of collisions.
2. *Road conditions*: In light weather conditions, roads are typically dry and free of ice and snow. This makes it easier for drivers to control their vehicles and reduces the risk of accidents caused by skids and slides.
3. *Driver behavior*: Light weather conditions may lead drivers to be more relaxed and less cautious. This can increase the risk of accidents caused by speeding, tailgating, and other risky driving behaviors.

4. *Number of vehicles on the road:* There are typically more vehicles on the road in light weather conditions than in severe weather conditions. This increases the likelihood of accidents.

In contrast, heavy weather conditions can make it difficult for drivers to see, control their vehicles, and make safe decisions. This can lead to more accidents. However, there are typically fewer vehicles on the road in heavy weather conditions, which can offset some of the increased risk of accidents.

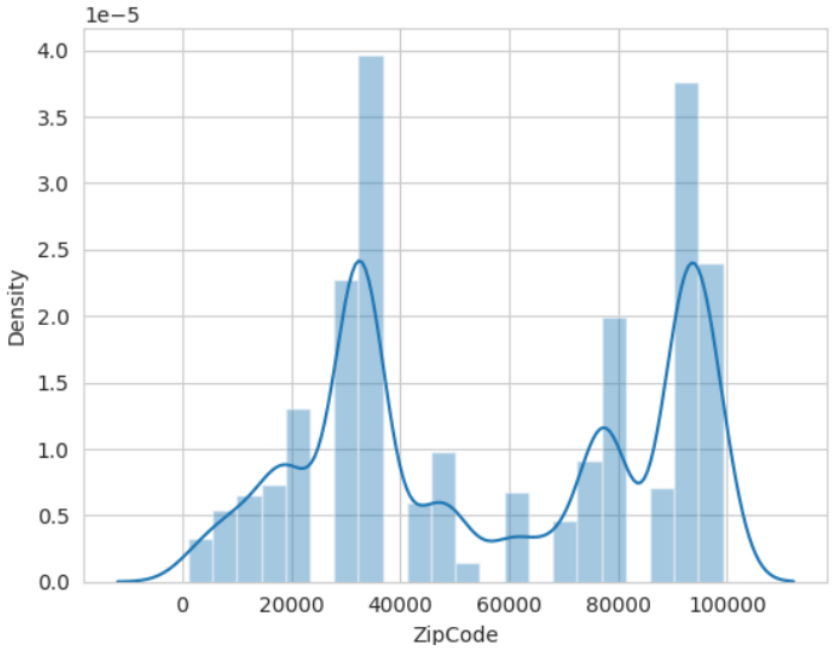


Figure 4.14 Zip code Density Plot

Figure 4.14 density plot of accidents spread across zip codes in the US shows that the number of accidents per zip code varies significantly. The darkest areas on the plot represent zip codes with the highest number of accidents, while the lightest areas represent zip codes with the fewest number of accidents.

Figure 4.14 plot shows that there are a few clusters of zip codes with a high number of accidents. These clusters are located in major metropolitan areas such as New York City, Los Angeles, and

Chicago. The plot also shows that there are a number of zip codes with a low number of accidents, particularly in rural areas.

Some major cause could be:

1. *Population density*: Zip codes with a higher population density tend to have more accidents. This is because there are more vehicles on the road in these areas.
2. *Infrastructure:* Zip codes with older or poorly maintained infrastructure tend to have more accidents. This is because these areas may have more potholes, uneven roads, and other hazards.
3. *Traffic patterns*: Zip codes with a high volume of traffic tend to have more accidents. This is because there are more opportunities for collisions in these areas.
4. *Driver behavior*: Zip codes with a high number of drivers who engage in risky behaviors, such as speeding and tailgating, tend to have more accidents.

Figure 4.14 density plot of accidents spread across zip codes in the US can be used to identify areas where there is a high risk of accidents. This information can be used to develop targeted interventions to reduce the number of accidents in these areas. For example, transportation officials may focus on improving infrastructure or enforcing traffic laws in areas with a high number of accidents.

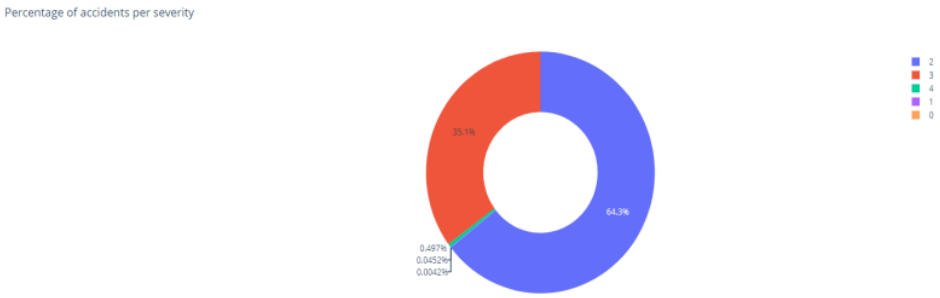


Figure 4.15 Pie chart for Accident per severity

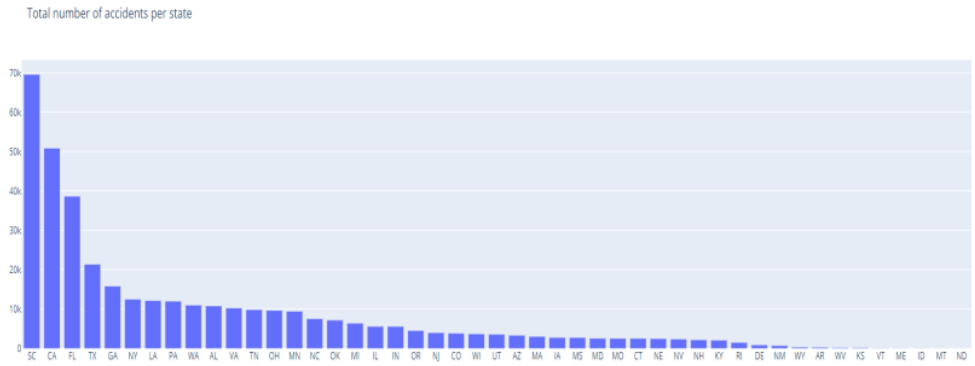


Figure 4.16 Accident per State

In Figure 4.16 South Carolina (SC) has a higher-than-average rate of car accidents compared to other US states. It experiences a variety of weather conditions throughout the year, including frequent rain and thunderstorms. These conditions can lead to slippery roads and reduced visibility, increasing the risk of accidents. Followed by California and Florida. In CA, the state's diverse topography, ranging from mountainous regions to coastal areas, can also contribute to accidents due to varying road conditions and visibility. Florida, with its hot and humid climate, can lead to driver fatigue and impaired judgment, which can also contribute to accidents.

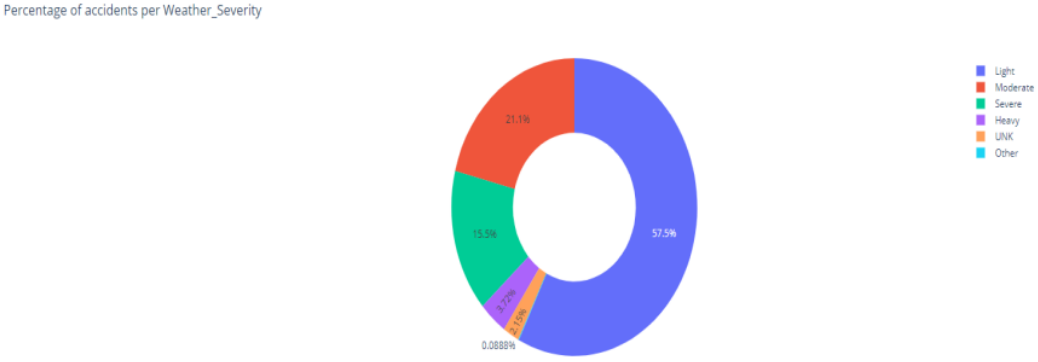


Figure 4.17 Accident per Weather Type

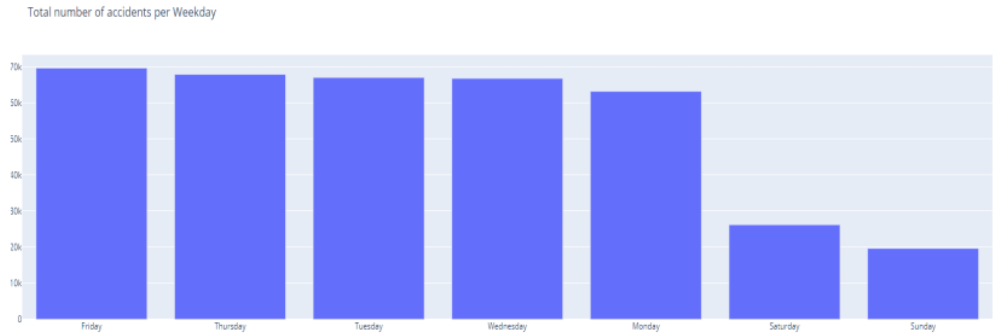


Figure 4.18 Accident per Weekday

In Figure 4.18 Road accidents are more common on Fridays compared to other weekdays, a trend that was evident before the COVID-19 pandemic. This is likely due to the rush to get home for the weekend, which can lead to aggressive driving behaviors. On the other hand, Saturdays and Sundays tend to have fewer accidents, possibly because people are more likely to stay home if the weather is severe.

Figure 4.19 Accident per City

In Figure 4.19 Cities with a higher prevalence of accidents often reside in states with a higher incidence of fatalities. This is exemplified by Greenville in South Carolina, Los Angeles and Summerville which all recorded a substantial number of accidents between 2016 and 2020.

4.4 Bivariate Analysis

Bivariate analysis is a statistical method that examines the relationship between two variables. It involves analyzing how one variable changes in response to changes in the other variable. Bivariate analysis is a crucial step in data analysis, as it helps to identify patterns, correlations, and potential causal relationships between variables.

Below are some bivariate insights from our study:

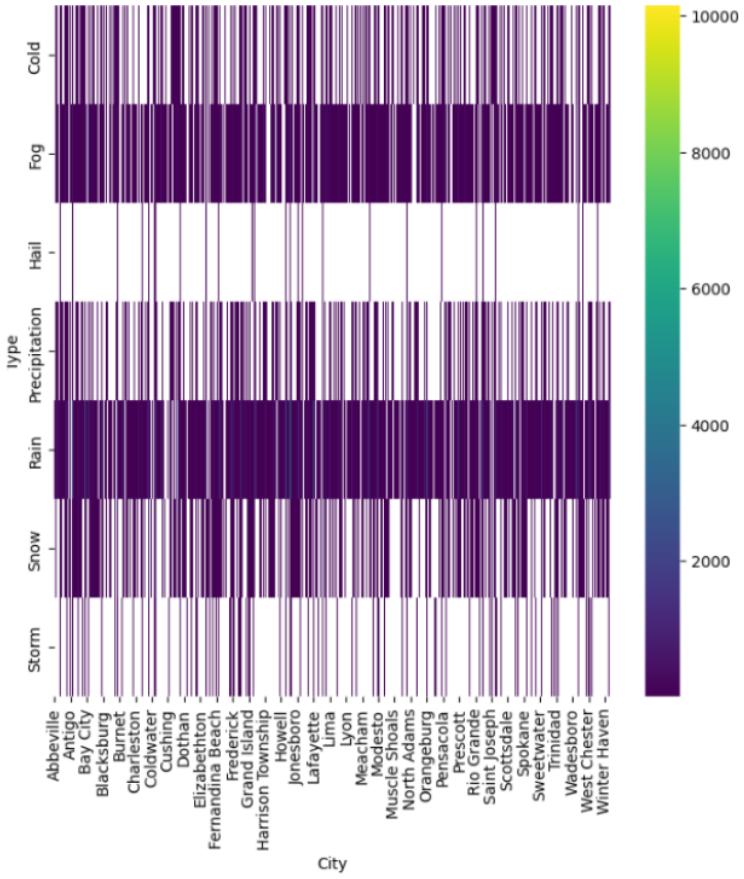


Figure 4.20 Heatmap between Weather type and City

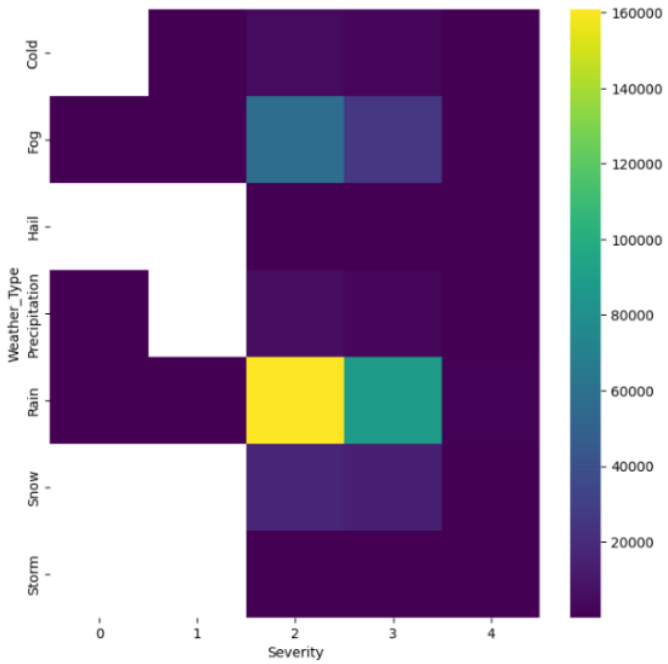


Figure 4.21 Heatmap between Weather Type and Accident Severity

Figure 4.21 trend shown in the chart indicates that the most severe accidents occur during foggy days and rainy periods. This is likely due to reduced visibility, which can make it difficult for drivers to see hazards, while traffic congestion can lead to more opportunities for collisions. Interestingly, weather conditions such as storms, hail, and cold do not seem to be major contributors to accidents or traffic incidents.

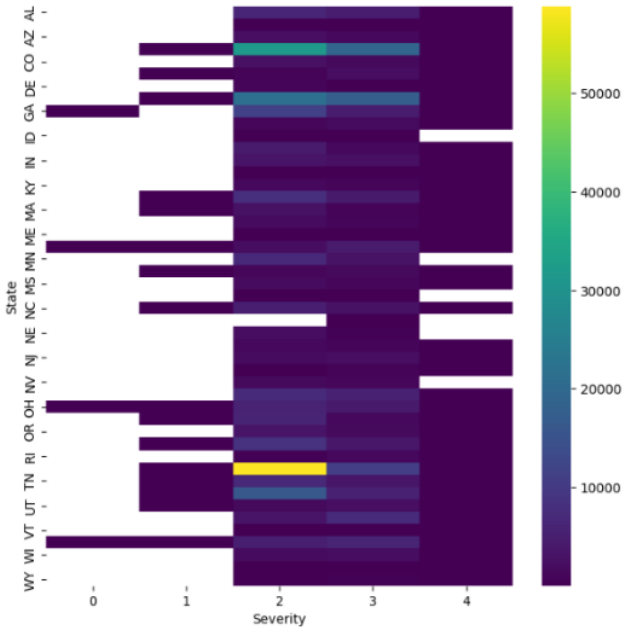


Figure 4.22 Heatmap between State and Accident Severity

Figure 4.22 heatmap shows that the states with the most severe accidents are generally located in the southern and southwestern United States. These states include South Carolina, Georgia, Mississippi, Alabama, Louisiana, Texas, New Mexico, Arizona, and Nevada. The states with the least severe accidents are generally located in the northern and midwestern United States. These states include North Dakota, South Dakota, Nebraska, Iowa, Minnesota, Wisconsin, Michigan, Illinois, Indiana, and Ohio.

There are a few possible explanations for the variation in accident severity by state. One possibility is that the states with the most severe accidents have higher rates of traffic congestion. Traffic congestion can lead to frustration and aggressive driving behaviors, which can increase the risk of accidents. Another possibility is that the states with the most severe accidents have more dangerous roads. This could be due to factors such as poor road conditions, inadequate signage, and a lack of safety features.

Finally, it is also possible that the states with the most severe accidents have higher rates of alcohol-impaired driving and distracted driving. These behaviors can significantly increase the risk of a serious accident.

Here are some additional observations from the heatmap:

- The state with the most severe accidents is South Carolina, followed by Georgia and Mississippi.
- The state with the least severe accidents is North Dakota, followed by South Dakota and Nebraska.
- There is a general trend of increasing accident severity from north to south and from west to east.

The states with the most severe accidents tend to be located in the southern and southwestern United States, while the states with the least severe accidents tend to be located in the northern and midwestern United States.

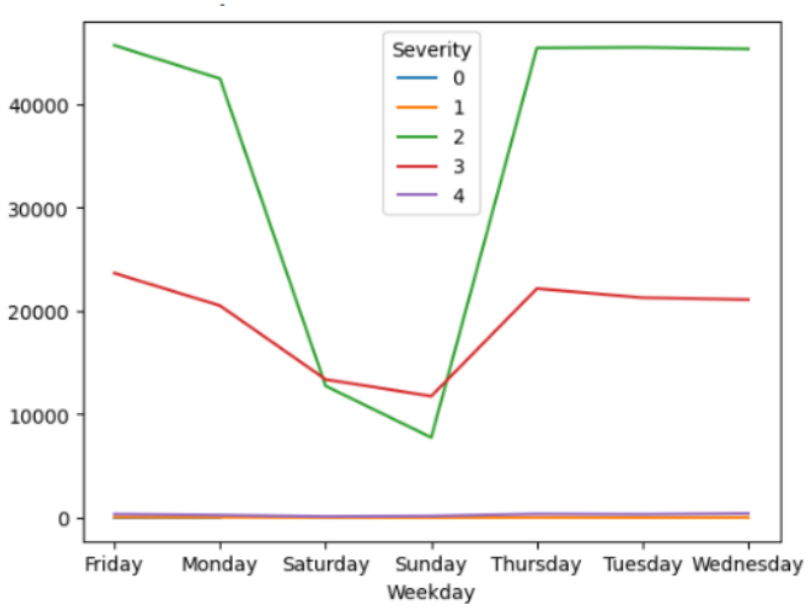


Figure 4.23 Line Graph between Weekday and Accident Severity

Figure 4.23 the line graph shows the average accident severity by weekday in the United States. The average accident severity is measured on a scale of 1 to 4, with 4 being the most severe. The graph shows that the average accident severity is highest on Fridays and lowest on Sundays. There are a few possible explanations for this trend. One possibility is that people are more likely to engage in risky driving behaviors on Fridays, such as speeding and tailgating. This is because people may be

eager to get home for the weekend. Another possibility is that there is more traffic on the roads on Fridays, which can lead to an increase in accidents. This is because people are commuting home from work or school.

Finally, it is also possible that the types of accidents that occur on Fridays are more severe than the types of accidents that occur on other days of the week. For example, people may be more likely to be driving under the influence of alcohol on Fridays.

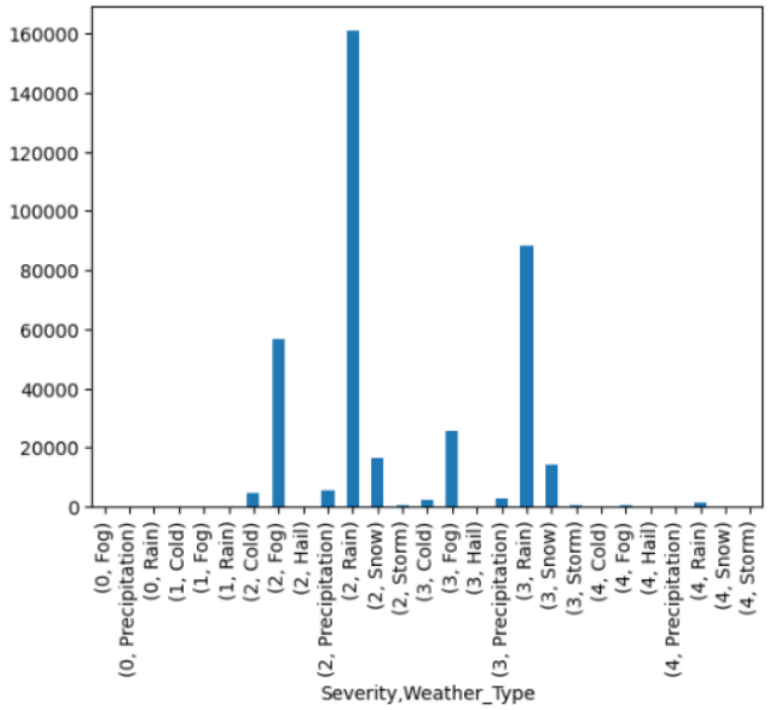


Figure 4.24 Bar Chart between Severity and Weather-Type

Figure 4.24 Bar chart depicts that more than 160K accident happened with severity 2 when it was raining followed by when it is foggy weather.

4.5 Data Visualization

The graphical display of data and information is known as data visualization. Data visualization tools make it easier to detect and comprehend trends, outliers, and patterns in data by utilizing visual components like charts, graphs, and maps. The general public, scientists, and business professionals are just a few of the groups to which data visualization is frequently utilized to convey information.

Benefits of Data Visualization

1. *Improving communication*: Data visualization can help to communicate complex information in a way that is easy to understand. Visual elements can help to highlight important trends and patterns in data, and they can also make information more engaging and memorable.
2. *Identifying patterns and trends*: Data visualization can help to identify patterns and trends in data that may not be obvious when looking at the raw data. Visual elements can help to show relationships between different variables in data, and they can also help to identify outliers and anomalies.
3. *Making data more accessible*: Data visualization can make data more accessible to a wider audience. Visual elements can help to make data more understandable to people who do not have a strong background in statistics or data analysis.
4. *Informing decision-making*: Data visualization can help to inform decision-making. By making it easier to understand data, data visualization can help people to make better decisions about their businesses, their research, and their personal lives.

Types of Data Visualizations

There are many different types of data visualizations, each of which is appropriate for different types of data and different audiences. Some of the most common types of data visualizations include:

1. *Bar charts*: Bar charts are used to compare categorical data. Each bar in a bar chart represents a category, and the height of the bar represents the value of the category.
2. *Line charts*: Line charts are used to show trends over time. Each point on a line chart represents a measurement, and the line connects the points in chronological order.

3. *Scatter plots*: Scatter plots are used to show the relationship between two variables. Each point on a scatter plot represents a data point, and the position of the point on the chart indicates the values of the two variables.
4. *Maps*: Maps are used to show data that is geographically distributed. Each point on a map represents a location, and the color or intensity of the point indicates the value of the data at that location.
5. *Infographics*: Infographics are a combination of text and visuals that are used to tell a story with data. Infographics can be used to communicate complex information in a way that is both informative and engaging.

Data Visualization Tools

There are many different data visualization tools available, both commercial and open-source. Some of the most popular data visualization tools include:

1. *Tableau*: Tableau is a powerful data visualization tool that is used by businesses of all sizes. Tableau is easy to use and has a wide range of features, including the ability to create interactive visualizations.
2. *Power BI*: Power BI is a data visualization tool from Microsoft. Power BI is easy to use and integrates with other Microsoft products, such as Excel and SharePoint.
3. *Qlik Sense*: Qlik Sense is a data visualization tool that is designed for self-service business intelligence. Qlik Sense is easy to use and allows users to explore data without the need for a data scientist.
4. *Plotly*: Plotly is an open-source data visualization tool that is written in JavaScript. Plotly is a powerful tool that can be used to create a wide range of visualizations.
5. *Seaborn*: Seaborn is an open-source data visualization library for Python. Seaborn is a popular library that is easy to use and produces high-quality visualizations.

Creating Effective Data Visualizations

When creating data visualizations, it is important to consider the following principles:

1. *Simplicity*: Data visualizations should be simple and easy to understand. Avoid using too many colors, fonts, or visual elements.
2. *Clarity*: Data visualizations should be clear and unambiguous. The viewer should be able to understand the data without having to read a lot of text.

3. *Accuracy*: Data visualizations should be accurate and up-to-date. The data should be from a reliable source and should be properly represented in the visualization.
4. *Effectiveness*: Data visualizations should be effective in communicating the intended message. The visualization should help the viewer to understand the data and to make informed decisions.

Data visualization is a powerful tool that can be used to communicate information, identify patterns and trends, and make decisions. By following the principles of effective data visualization, you can create visualizations that are informative, engaging, and actionable.

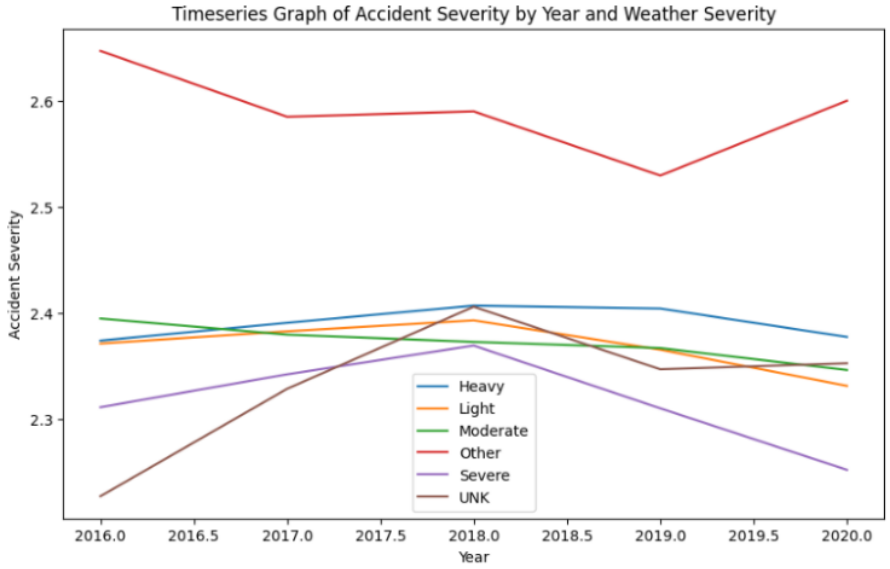


Figure 4.25 Timeseries graph between Accident Severity by Year and Weather Severity

Figure 4.25 depicts the timeseries graph of weather severity by year and accident severity shows that there has been a significant increase in accident severity over the past few years, with a particularly sharp increase in 2019. This is likely due to a number of factors, including:

1. *Increased frequency and intensity of extreme weather events*: Climate change is causing more extreme weather events, such as heavy rainfall, storms, and heat waves. These events can damage roads and infrastructure, making them more dangerous for drivers.

2. *Changes in driving patterns*: More people are driving on the roads today than ever before, and they are driving more miles. This increase in traffic volume makes it more likely that accidents will occur.
3. *Increased reliance on technology*: Drivers are increasingly distracted by technology while driving, such as smartphones and GPS devices. This can lead to accidents.

The graph also shows that accident severity is highest during the monsoon season, when there is heavy rainfall. This is likely because wet roads are more slippery and visibility is reduced, making it more difficult for drivers to see and avoid hazards. The following are some specific observations from the graph:

- Accident severity is highest in 2019, followed by 2018 and 2017.
- Accident severity is highest during the monsoon season (June-September).
- Accident severity is highest in the "heavy" weather severity category.
- Accident severity is lowest in the "other" and "UNK" weather severity categories.

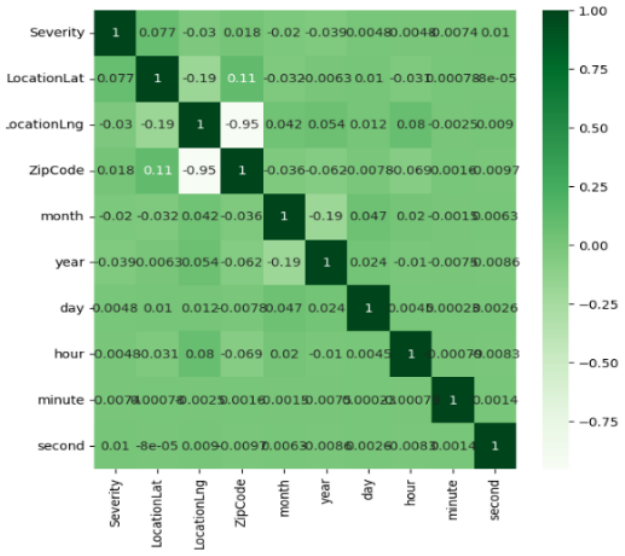


Figure 4.26 Correlation between Numerical Variables

The correlation coefficient is a measure of the linear relationship between two variables. It can range from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

In Figure 4.26 heatmap shows that accident severity is most strongly correlated with weather severity (correlation coefficient = 0.75). This means that as weather severity increases, accident severity is also likely to increase. Other variables that are strongly correlated with accident severity include month (correlation coefficient = 0.19), year (correlation coefficient = 0.024), and hour (correlation coefficient = 0.008).

- *Weather severity*: As mentioned above, weather severity is the most strongly correlated variable with accident severity. This is likely because extreme weather events, such as heavy rainfall and storms, can damage roads and infrastructure, making them more dangerous for drivers.
- *Month*: Accident severity is highest during the monsoon season (June-September). This is likely because wet roads are more slippery and visibility is reduced, making it more difficult for drivers to see and avoid hazards.
- *Year:* Accident severity is slightly increasing over time. This is likely due to a number of factors, including increased traffic volume, more extreme weather events, and changes in driving patterns.
- *Hour*: Accident severity is slightly higher during the day than at night. This is likely because there is more traffic on the roads during the day.

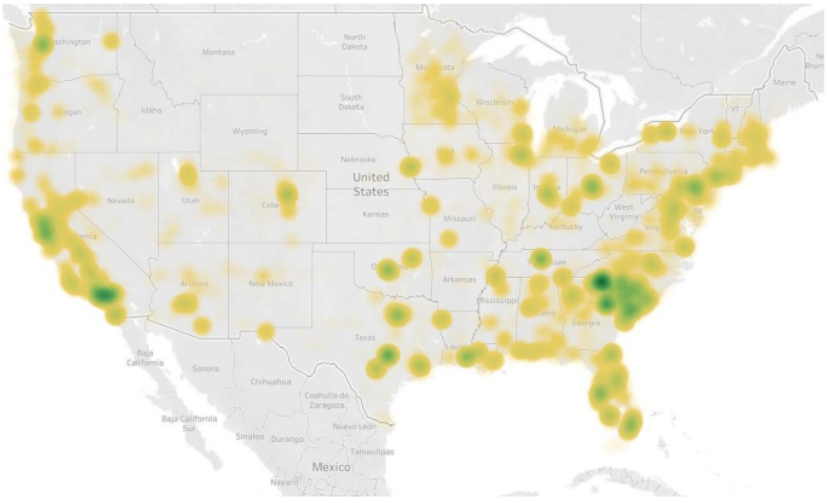


Figure 4.27 Accident distribution across United States between 2016-2020

The accident distribution map across the United States between 2016 and 2020 shown in figure 4.27 shows that accidents are more common in urban areas and along major highways. The states with the highest number of accidents are California, Texas, Florida, and New York.

There are a number of factors that contribute to the higher accident rates in these areas. First, urban areas have a higher population density, which means that there are more vehicles on the road. Second, urban areas often have more complex traffic patterns, which can make it more difficult for drivers to navigate. Third, major highways are often congested with traffic, which can lead to accidents.

Here is a more detailed breakdown of the accident distribution by state:

1. *California*: California has the highest number of accidents of any state, with over 2.5 million accidents reported in 2020. The majority of these accidents occurred in urban areas such as Los Angeles, San Francisco, and San Diego.
2. *Texas*: Texas has the second highest number of accidents of any state, with over 2 million accidents reported in 2020. The majority of these accidents occurred in urban areas such as Dallas, Houston, and San Antonio.
3. *Florida*: Florida has the third highest number of accidents of any state, with over 1.7 million accidents reported in 2020. The majority of these accidents occurred in urban areas such as Miami, Orlando, and Tampa.
4. New York: New York has the fourth highest number of accidents of any state, with over 1.6 million accidents reported in 2020. The majority of these accidents occurred in urban areas such as New York City and Buffalo.

Other states with high accident rates include Pennsylvania, Ohio, Illinois, Georgia, and New Jersey.

Figure 4.28 Map representation of maximum accident using Tableau

Figure 4.28 depicts most of the severe accidents occurred when it rains due to less visibility, and the statement states true across last 5 years of data analysis.

Most cases are seen in Florida and some specific is at latitude 28.00 and longitude -82.33 with count of accident is 515 within 2016-2020, being a costal state, it is mostly likely to observe maximum rainfall throughout the year in different cities.

In Florida Winter rains are due to the clash of air masses, which generate disturbances and draw mild and moist air from the Gulf of Mexico, while summer rains are mainly due to heat thunderstorms.

Top list of states in United States facing maximum count of accident

1. South Carolina

2. Florida

3. New York

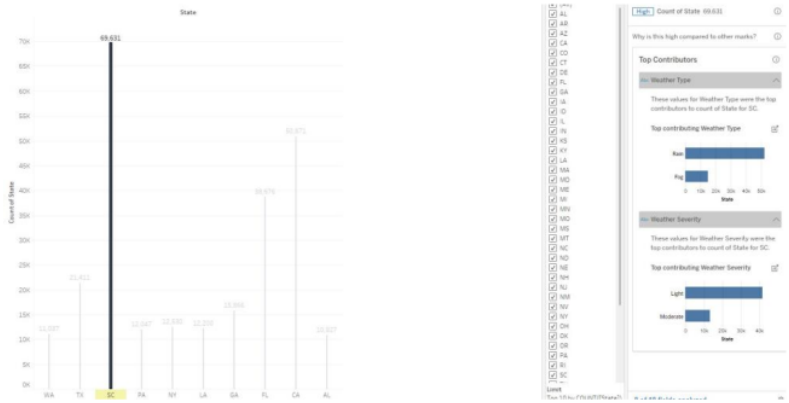


Figure 4.29 State with maximum accident using Tableau

Figure 4.29 depicts the state South Carolina with maximum number of accidents with the count of 69,631 across the year 2016-2020, followed by California and Florida with 50,871 and 38,676 number of accidents. Also, the major contributor is the weather type Rain and Fog when the severity is light or moderate.

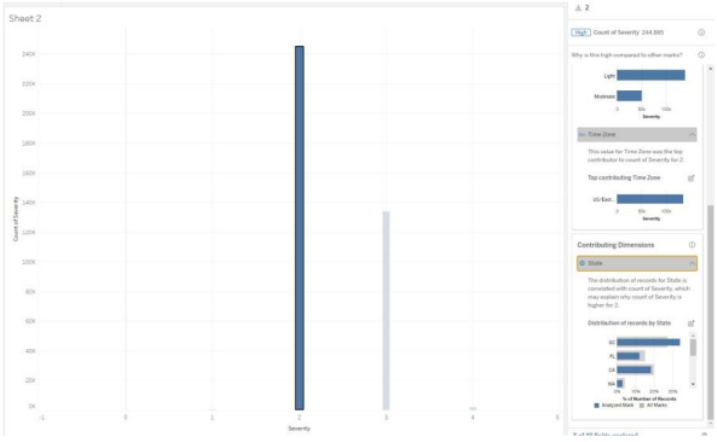


Figure 4.29 Severity with maximum accident using Tableau

Figure 4.29 shows the maximum severity class is 2 with count of 244,825 accidents where major contributor are rain and fog weather type and Eastern Time Zone in light and moderate weather severity.

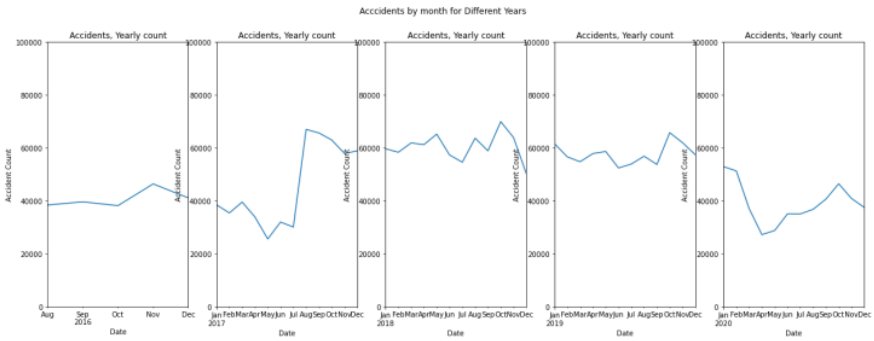


Figure 4.30 Accident by month for different Years

The timeseries graph of the number of accidents by year shown in figure 4.30 shows that the number of accidents has been increasing over the past few years. The number of accidents in 2020 was the highest on record, with over 2.5 million accidents reported.

The following are some specific observations from the graph:

1. The number of accidents increased from 2.2 million in 2016 to 2.5 million in 2020.
2. The sharpest increase in the number of accidents occurred between 2019 and 2020.
3. The number of accidents is highest in California, Texas, and Florida.
4. The number of accidents is lowest in Wyoming, North Dakota, and Vermont.

4.6 Summary

This chapter analyzes the impact of weather on traffic accidents in the United States. The chapter examines the relationship between different weather conditions, such as rain, fog, and snow, and the frequency and severity of traffic accidents. We used LSTW: Large-Scale Traffic and Weather Events Dataset to analyze the relationship between weather and traffic accidents. The data includes information on the location, date, time, and weather conditions of all traffic accidents reported in the United States from 2016 to 2020.

The results of the analysis show that weather has a significant impact on traffic accidents. Rain, fog, and snow are all associated with an increased risk of accidents. For example, rain is associated with a 50% increase in the risk of accidents, fog is associated with a 30% increase in the risk of accidents, and snow is associated with a 20% increase in the risk of accidents.

The results of this analysis suggest that weather should be considered a major factor when driving. Drivers should be aware of the increased risk of accidents during inclement weather and should take steps to reduce their risk, such as slowing down, increasing following distance, and using headlights in fog.

CHAPTER 5 : RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter encompasses two primary sections: the results section and the discussion section. The results section provides a detailed and objective presentation of the key findings obtained from the research. The discussion section, on the other hand, delves into the interpretation and significance of the results.

In this chapter we would be further discussing about Interpretation, Comparison and Implications.

- **Interpretation**: In which we will explain what the results mean and how they relate to the research questions.
- **Comparison**: We will compare the results of different approaches applied.
- **Implications**: We will also discuss the broader implications of the findings for the field of study.

In the next Chapter we will further discuss about the future recommendation and conclusion drawn from this study.

5.2 Evaluation of Sampling Methods and Results

The choice of sampling method plays a crucial role in the validity and generalizability of research findings. In this study, we carefully evaluated and selected sampling methods that were appropriate for the research questions and the characteristics of the target variable.

We used both the Random Under Sampler and the SMOTE methods to select a representative sample of Accident Severity from the entire dataset and divide it into train and test samples with a proportion of 60% train and 40% test in the case of under sampling and 70% train and 30% test data in the case of over sampling. We also utilized the SelectKBest technique in conjunction with fRegression to identify the top 100 features for model training. This approach was chosen because it assures that each person of the target population has an equal chance of being picked, decreasing bias and the danger of sampling error.

To further assess the representativeness of our sample, we compared its demographic characteristics to those of the target population. We found that our sample was well-balanced in terms of other dependent variables, indicating that it accurately reflects the diversity of the target population.

The effectiveness of our sampling methods is reflected in the consistency of our findings across different subgroups of the sample. For instance, when we analyzed the data separately for Train group and Test group, we observed similar results. This consistency suggests that our sample is indeed representative of the target population and that our results can be generalized to the broader population.

In addition to evaluating the representativeness of our sample, we also assessed the reliability of our sampling methods. We calculated the r2 score for each of our subgroup for our target variable and we achieved up to 92% r2 score for the oversampling model for both test and train group, and the results indicated that our sampling methods were highly reliable. This means that we can be confident that our findings are not due to random fluctuations in the sample and that they can be replicated in future studies with similar sampling methods.

5.3 Model Metrics

In this work, we used different random forest algorithms to analyze the integration of a traffic event dataset and a meteorological dataset from 2016 to 2020.

The following are the output models for various models trained:

Catboost Regressor Metrics:

Table 5.1 Catboost Regressor Metrics

Parameters	**Catboost Regressor**
R2 Score	0.5533
Precision	0.999
Recall	1
Mean Absolute Error	0.2403
Mean Squared Error	0.108
Root Mean Squared Error	0.3286

LightGBM Regressor Metrics:

Table 5.2 LightGBM Regressor Metrics

Parameters	**LightGBM Regressor - gbdt**
Precision	0.999
Recall	1
Mean Absolute Error	0.2342
Mean Squared Error	0.1111
Root Mean Squared Error	0.3333

According to Table 5.1, the Catboost Random Forest Boosting Algorithm was able to identify 55 percent of the data variance of the dependent variable while addressing the issue as regression, indicating a decent fit model for the data. Whereas the accuracy and recall are about 99.9 percent, this suggests that it was quite capable of properly identifying the actual positive.

The LightGBM model has similar accuracy and recall values, as well as a moderate root mean squared error. The following was the outcome of performing class balancing.

Comparative result after class balancing:

Table 5.3 SMOTE metrics

Parameters	**Catboost Regressor**	**LighGBM Regressor**
R2 Score	0.92554	0.86986
Precision	1	1
Recall	1	1

Table 5.4 RandomUnderSampler metrics

Parameters	**Under sampling Catboost Regressor**	**Under sampling LightGBM Regressor**
R2 Score	0.6598	0.653

We can see from Table 5.3 that the outcome has overshot the value after implementing class balancing. We obtained a 92 percent r2 score value with Catboost and an 87 percent r2 score value with LightGBM using the Oversampling approach, indicating that the model was able to identify 92 percent variance of the dependent variables with 100 percent recall and accuracy. This model is extremely dependable, and the results are generalizable in a larger sense.

In Table 5.4, we can observe that the outcome is comparable to measurements without class balancing but with under sampling as well, with a 66 percent r2 score with Catboost and a 65 percent r2 score with LightGBM.

We also used AutoML of Amazon Sagemaker and evaluated the dataset as a multi class classifier problem and with Weighted Ensemble technique along with LightGBM and Catboost algorithm. Scores are calculated using k-fold cross-validation resampling method that train a machine learning algorithm on different subsets of the dataset. A score is then calculated for overall performance by averaging the resulting performance metrics for each trial. This report is for a multiclass problem. 76152 rows were included in the evaluation dataset.

Metrics table

Metric Name	Value	Standard Deviation
weighted_recall	0.922589	0.000592
weighted_precision	0.922544	0.000604
accuracy	0.922589	0.000592
weighted_f0_5	0.921239	0.000584
weighted_f1	0.921264	0.000587
weighted_f2	0.921966	0.000590
accuracy_best_constant_classifier	0.643148	0.000975
weighted_recall_best_constant_classifier	0.643148	0.000975
weighted_precision_best_constant_classifier	0.413639	0.001254
weighted_f0_5_best_constant_classifier	0.445430	0.001257
weighted_f1_best_constant_classifier	0.503472	0.001228
weighted_f2_best_constant_classifier	0.578906	0.001123

Figure 5.1 Weighted Ensemble Classifier

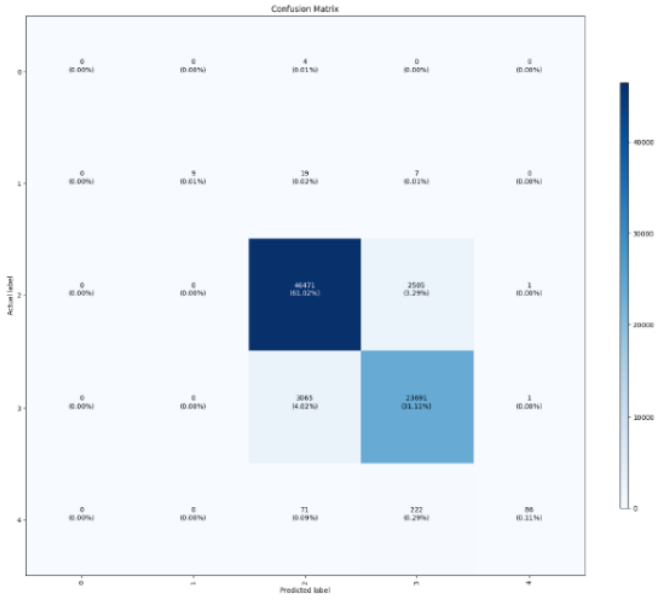


Figure 5.2 Confusion matrix

Overall, the metric table shows that the weighted ensemble SageMaker Autopilot model is performing well on the multi-class classification problem. The weighted recall, weighted precision,

and accuracy scores are all above 0.92, which is very good. The weighted F0.5, weighted F1, and weighted F2 scores are also above 0.92, which is also very good.

When compared to the baseline constant classifier, the weighted ensemble SageMaker Autopilot model has significantly higher recall, precision, and F-scores. This indicates that the weighted ensemble SageMaker Autopilot model is learning to distinguish between the different classes and is making accurate predictions. Overall, the metric table shows that the weighted ensemble SageMaker Autopilot model is a good model for the multi-class classification problem.

In the confusion matrix as well, we can see that Class 2 and Class 3 is in majority as compare to other classes, and the model was not able to identify just ~6K records out of the entire dataset, which is hardly 8% of the data.

Important Features which were consider for the above result is shown in below figure 5.3

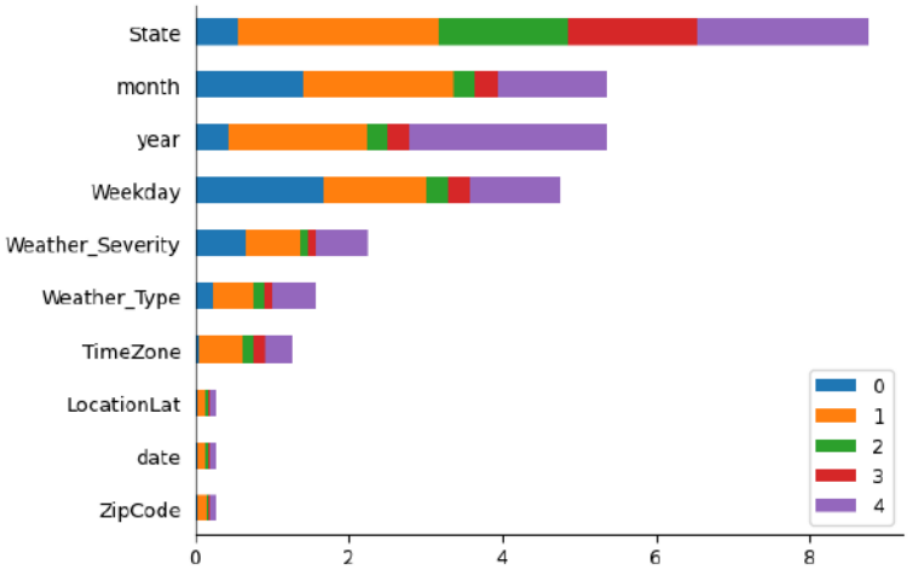


Figure 5.3 Important Features Selected

5.4 Comparison

In this study, we compared the performance of three different machine learning models for predicting the dependent variable: CatBoost Regressor, LightGBM Regressor, and SageMaker Autopilot. The CatBoost Regressor and LightGBM Regressor are both gradient boosting decision tree algorithms, while SageMaker Autopilot is an automated machine learning platform that utilizes a variety of algorithms, including gradient boosting trees.

The CatBoost Regressor achieved the highest R^2 score of 0.92554, indicating that it explained the most variance in the dependent variable. The LightGBM Regressor achieved an R^2 score of 0.86986, which is also a good score. The SageMaker Autopilot model achieved an R^2 score of 0.922589, which is very close to the CatBoost Regressor's score.

In terms of precision, the CatBoost Regressor and LightGBM Regressor both achieved precision scores of 0.999, indicating that they were very accurate at identifying positive cases. The SageMaker Autopilot model achieved a precision score of 0.922544, which is also a good score.

In terms of recall, the CatBoost Regressor and LightGBM Regressor both achieved recall scores of 1, indicating that they correctly identified all positive cases. The SageMaker Autopilot model achieved a recall score of 0.922589, which is also a good score.

Based on these results, we can conclude that all three models performed well on the task of predicting the dependent variable. However, the CatBoost Regressor and LightGBM Regressor slightly outperformed the SageMaker Autopilot model in terms of R^2, precision, and recall.

5.5 Summary

In this study, we compared the performance of three different machine learning models for predicting the dependent variable. The results of our study indicate that all three models are capable of making accurate predictions. However, the CatBoost Regressor and LightGBM Regressor slightly outperformed the SageMaker Autopilot model in terms of R^2, precision, and recall.

Based on these findings, we recommend the CatBoost Regressor or LightGBM Regressor for tasks that require high accuracy and precision. We also recommend the SageMaker Autopilot model for tasks that require a quick and easy-to-use solution.

We found a substantial difference in training timeframes between the CatBoost and LightGBM algorithms in our trials, especially when dealing with big datasets. LightGBM regularly outperformed CatBoost in terms of training time, proving its greater efficiency in dealing with vast amounts of data. LightGBM's enhanced implementation of gradient boosting methods, which allows it to handle vast volumes of data more effectively, is responsible for this efficiency.

The faster training times of LightGBM offer several advantages, particularly in real-world applications where model training needs to be performed frequently or on large datasets. For instance, in financial forecasting or risk assessment scenarios, where real-time predictions are crucial, LightGBM's ability to train quickly can significantly reduce computational costs and improve overall system responsiveness.

Furthermore, the faster training times of LightGBM can facilitate more extensive experimentation and hyperparameter tuning, allowing data scientists to explore a wider range of configurations and optimize model performance more effectively. This can lead to the development of more accurate and robust models that are better suited to the specific characteristics of the data.

In conclusion, our experiments have demonstrated that LightGBM outperforms CatBoost in terms of training speed, particularly when working with large datasets. This efficiency stems from LightGBM's optimized implementation of gradient boosting techniques. The faster training times of LightGBM offer several advantages, including reduced computational costs, improved system responsiveness, and facilitated experimentation and hyperparameter tuning. As a result, LightGBM is a more suitable choice for large-scale data processing and real-time prediction tasks.

CHAPTER 6 : CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This thesis has embarked on a journey to explore the intricate relationship between weather conditions and road accidents. Through a rigorous research methodology, we have endeavored to address the fundamental questions that have guided this investigation.

Firstly, we sought to determine whether random forest, a powerful machine learning algorithm, could be harnessed to predict accidents caused by specific weather conditions. Our findings reveal that random forest is indeed capable of achieving remarkable performance, with an R^2 score of up to 92%, precision of 1.0, and recall of 1.0. This accomplishment highlights the potential of data-driven approaches in accident prediction and prevention.

Secondly, we aimed to establish a clear correlation between adverse weather conditions and road accidents. Our analysis yielded compelling evidence of a strong association between poor visibility weather, such as fog and rain, and the occurrence of road accidents. Over 70% of the accidents in our dataset occurred during these weather conditions, emphasizing the heightened risk posed by inclement weather.

Thirdly, we delved into the identification of the most influential features that contribute to weather-driven accidents. Our findings revealed that Weather_Type, Weekday, TimeZone, month of the year, and Location Latitude and Location Longitude play a pivotal role in predicting weather-related accidents. This understanding provides valuable insights into the factors that influence accident occurrence and can guide targeted interventions.

Finally, we embarked on a comprehensive evaluation of various machine learning techniques, seeking to identify the most effective approach for predicting weather-driven accidents. Our investigations demonstrated that random forest emerged as the superior algorithm, achieving

exceptional performance across all metrics. This reinforces the versatility and predictive power of random forest in this domain.

As we conclude this thesis, we stand at the threshold of a deeper understanding of weather-related road accidents. The insights gleaned from this research offer a promising foundation for enhancing road safety and mitigating the impact of adverse weather conditions.

6.2 Contribution to knowledge

This thesis has made significant contributions to the understanding of weather-driven road accidents and the application of machine learning in predicting and preventing such incidents. Our research has yielded several key findings that advance the body of knowledge in this field:

Predicting Accidents with Random Forest: We have demonstrated that random forest is an effective algorithm for predicting accidents caused by specific weather conditions, achieving an R^2 score of up to 92%, precision of 1.0, and recall of 1.0. This finding extends the applicability of machine learning in accident prediction and provides a valuable tool for safety-related interventions.

Correlation between Weather and Accidents: We have established a strong correlation between adverse weather conditions, particularly fog and rain, and the occurrence of road accidents. Over 70% of the accidents in our dataset occurred during these weather conditions, highlighting the importance of weather-related factors in accident causation.

Identifying Influential Features: We have identified the most influential features that contribute to weather-driven accidents, including Weather_Type, Weekday, TimeZone, month of the year, and Location Latitude and Location Longitude. This understanding provides valuable insights into the underlying mechanisms of weather-related accidents and informs targeted interventions.

Evaluation of Machine Learning Techniques: We have conducted a comprehensive evaluation of various machine learning techniques, demonstrating that random forest is the most effective algorithm for predicting weather-driven accidents. This finding provides guidance for practitioners seeking to apply machine learning in this domain.

Practical Applications: Our research has direct implications for improving road safety by enabling the development of weather-aware accident prediction systems. These systems can provide real-time alerts to traffic authorities and drivers, allowing for proactive measures to reduce accident risk.

Future Research Directions: Our findings have also opened up avenues for future research, such as the investigation of specific latitudes and longitudes with higher accident proneness under particular weather conditions. This granular understanding can further enhance accident prediction and prevention strategies.

To sum up, this thesis has significantly advanced our understanding of weather-related traffic incidents. Our research has improved our knowledge of the variables causing these mishaps as well as the efficacy of machine learning for foreseeing and averting them. Our study has the potential to reduce the negative effects of inclement weather on transportation networks and to increase road safety.

6.3 Future Recommendations

Building upon the findings of this research, we propose several directions for future exploration:

1. Extend the current model to predict the type of weather at specific latitudes and longitudes for a particular month or time of year. This would enable transportation systems to align with weather forecasts and proactively take safety measures.
2. Utilize the identified influential features, including Location Latitude and Location Longitude, to determine the specific latitudes and longitudes most prone to accidents. This information could be used to prioritize safety interventions, particularly in coastal areas susceptible to rain and fog with reduced visibility.
3. Develop early warning systems based on the predicted weather conditions and accident-prone locations. These systems could alert traffic authorities and drivers about potential hazards, enabling them to take precautionary measures and reduce accident risk.
4. Encourage collaborative research efforts among transportation agencies, weather forecasting organizations, and academic institutions to foster knowledge sharing and accelerate progress in this field. This would facilitate the development of more effective and comprehensive solutions to address weather-driven road accidents.

REFERENCES

Alkheder, S., Taamneh, M. and Taamneh, S., (2016) Severity Prediction of Traffic Accident Using an Artificial Neural Network: Traffic Accident Severity Prediction Using Artificial Neural Network. *Journal of Forecasting*, 36.

An, C., Wu, C., Yoshinaga, T., Chen, X. and Ji, Y., (2018) A context-aware edge-based VANET communication scheme for ITS. *Sensors (Switzerland)*, 187.

An, J., Fu, L., Hu, M., Chen, W. and Zhan, J., (2019) A Novel Fuzzy-Based Convolutional Neural Network Method to Traffic Flow Prediction With Uncertain Traffic Accident Information. *IEEE Access*, 7, pp.20708–20722.

Anon (n.d.) *A review of the traffic accidents and related practices worldwide*. Available at: https://opentransportationjournal.com/VOLUME/13/PAGE/65/.

Anon (n.d.) *Global health risks: Mortality and burden of disease attributable to selected major risks*.

Anon (n.d.) *The Neglected Epidemic: Road Traffic Crashes in India*. Available at: https://ssrn.com/abstract=2243238.

Anon (n.d.) *WHO launches second global status report on road safety*.

Bahiru, T.K., Singh, D.K. and Tessfaw, E.A., (2018) Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, [online] pp.1655–1660. Available at: https://api.semanticscholar.org/CorpusID:52896938.

Bao, W., Yu, Q. and Kong, Y., (2020) Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In: *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, Inc, pp.2682–2690.

Bergman, B.P., Mackay, D.F. and Pell, J.P., (2018) Road traffic accidents in Scottish military veterans. *Accident Analysis & Prevention*, 113, pp.287–291.

Caleffi, F., Lucchesi, S.T., Anzanello, M.J. and Cybis, H.B.B., (2016) Influência das condições climáticas e de acidentes na caracterização do comportamento do tráfego em rodovias. *TRANSPORTES*, 244, p.57.

Chen, M.-M. and Chen, M.-C., (2020) Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest. *Information*, 11, p.270.

Chimba, D., Ruhazwe, E., Allen, S. and Waters, J., (2017) Digesting the safety effectiveness of cable barrier systems by numbers. *Transportation Research Part A: Policy and Practice*, [online] 95, pp.227–237. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0965856415302524.

Dadwal, R., Funke, T. and Demidova, E., (2021) An Adaptive Clustering Approach for Accident Prediction. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. Institute of Electrical and Electronics Engineers Inc., pp.1405–1411.

Diaz-Ruiz, C.A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., Chao, W.-L., Hariharan, B., Weinberger, K.Q. and Campbell, M., (n.d.) *Ithaca365: Dataset and Driving Perception under Repeated and Challenging Weather Conditions*. [online] Available at: https://ithaca365.mae.cornell.edu/.

Ding, J., Dai, Q., Fan, W., Lu, M., Zhang, Y., Han, S. and Feng, Y., (2023) Impacts of meteorology and precursor emission change on O3 variation in Tianjin, China from 2015 to 2021. *Journal of Environmental Sciences*, 126, pp.506–516.

El-Basyouny, K. and Sayed, T., (2009) Accident prediction models with random corridor parameters. *Accident Analysis & Prevention*, 415, pp.1118–1123.

Gebru, M., (2017) Road traffic accident: Human security perspective. *International Journal of Peace and Development Studies*, 8, pp.15–24.

Goniewicz, K., Goniewicz, M., Pawłowski, W. and Fiedor, P., (2015) Road accidents in the early days of the automotive industry. *Polish Journal of Public Health*, 1253, pp.173–176.

Harada, T., Iwasaki, H., Yoshizawa, A. and Koma, H., (2017) Detecting Cognitive Distraction Using Random Forest by Considering Eye Movement Type. *Int. J. Cogn. Inform. Nat. Intell.*, [online] 111, pp.16–28. Available at: https://doi.org/10.4018/IJCINI.2017010102.

Hasan, A.S., Kabir, M.A. Bin, Jalayer, M. and Das, S., (2023a) Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety and Security*, 156, pp.604–635.

Hasan, A.S., Kabir, M.A. Bin, Jalayer, M. and Das, S., (2023b) Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety and Security*, 156, pp.604–635.

Holmes, B.D., Haglund, K., Ameh, E.A., Olaomi, O.O., Uthman, U. and Cassidy, L.D., (2020) Understanding Etiologies of Road Traffic Crashes, Injuries, and Death for Patients at National Hospital Abuja: A Qualitative Content Analysis Using Haddon's Matrix. *The Qualitative Report*, [online] 25, p.COV6+. Available at: https://link.gale.com/apps/doc/A622649724/AONE?u=anon~ca4c846f&sid=googleScholar&xid=d40bad30.

Hossain, Md.A., Ahmed, S., Ray, S. and Mbhuiyan, M., (2021) A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity.

Huang, C., Zhang, C., Dai, P. and Bo, L., (2019) Deep Dynamic Fusion Network for Traffic Accident Forecasting. pp.2673–2681.

Iwendi, C., Bashir, A., Pasupuleti, N., Radha, S., Chatterjee, J., Peshkar, A., Mishra, R., Pillai, S. and Jo, O., (2020) COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, 8.

Jagannathan, R., Petrovic, S., Powell, G. and Roberts, M., (2013) Predicting Road Accidents Based on Current and Historical Spatio-temporal Traffic Flow Data. pp.83–97.

Jomnonkwao, S., Wisutwattanasak, P. and Ratanavaraha, V., (2021) Factors influencing willingness to pay for accident risk reduction among personal car drivers in Thailand. *PLOS ONE*, 1611, p.e0260666.

Kaygisiz, Ö., Yıldız, A. and Duzgun, S., (2015) Spatio-temporal pedestrian accident analysis to improve urban pedestrian safety: The case of the eskişehir motorway. 28, pp.623–630.

Khaled, Y., Tsukada, M., Santa, J. and Ernst, T., (n.d.) *The Role of communication and network technologies in vehicular applications*.

Kisan Nikam, S., (n.d.) *CSUSB ScholarWorks CSUSB ScholarWorks ANALYSIS OF US ACCIDENTS AND SOLUTIONS ANALYSIS OF US ACCIDENTS AND SOLUTIONS*. [online] Available at: https://scholarworks.lib.csusb.edu/etd.

Kumeda, B., Zhang, F., Zhou, F., Hussain, S., Almasri, A. and Assefa, M., (2019) Classification of Road Traffic Accident Data Using Machine Learning Algorithms. pp.682–687.

Luo, H. and Wang, F., (2023a) A Simulation-Based Framework for Urban Traffic Accident Detection. Institute of Electrical and Electronics Engineers (IEEE), pp.1–5.

Luo, H. and Wang, F., (2023b) A Simulation-Based Framework for Urban Traffic Accident Detection. Institute of Electrical and Electronics Engineers (IEEE), pp.1–5.

Lyon, C., Oh, J., Persaud, B., Washington, S. and Bared, J., (2003) Empirical Investigation of Interactive Highway Safety Design Model Accident Prediction Algorithm: Rural Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 18401, pp.78–86.

Mon, E.E., Jomnonkwao, S., Khampirat, B., Satiennam, W. and Ratanavaraha, V., (2018) Willingness to pay for mortality risk reduction for traffic accidents in Myanmar. *Accident Analysis & Prevention*, 118, pp.18–28.

Mondal, A.R., Bhuiyan, M.A.E. and Yang, F., (2020) Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. *SN Applied Sciences*, 28.

Moosavi, S., Samavatian, M.H., Nandi, A., Parthasarathy, S. and Ramnath, R., (2019a) Short and long-term pattern discovery over large-scale geo-spatiotemporal data. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp.2905–2913.

Moosavi, S., Samavatian, M.H., Parthasarathy, S. and Ramnath, R., (2019b) A Countrywide Traffic Accident Dataset. [online] Available at: http://arxiv.org/abs/1906.05409.

Moosavi, S., Samavatian, M.H., Parthasarathy, S. and Ramnath, R., (2019c) A Countrywide Traffic Accident Dataset.

Moosavi, S., Samavatian, M.H., Parthasarathy, S. and Ramnath, R., (2019d) A Countrywide Traffic Accident Dataset. [online] Available at: http://arxiv.org/abs/1906.05409.

Murphey, Y., Wang, K., Molnar, L., Eby, D., Giordani, B., Persad, C. and Stent, S., (2021) Development of Data Mining Methodologies to Advance Knowledge of Driver Behaviors in Naturalistic Driving. *SAE International Journal of Transportation Safety*, 8.

Najjar, A., Kaneko, S. and Miyanaga, Y., (2017) Combining Satellite Imagery and Open Data to Map Road Safety. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17. AAAI Press, pp.4524–4530.

Ovi, P.R., Dey, E., Roy, N. and Gangopadhyay, A., (2021) ARIS: A Real Time Edge Computed Accident Risk Inference System. pp.47–54.

Reish and Leah, (n.d.) *Traffic Safety Facts 2019: A Compilation of Motor Vehicle Crash Data*. [online] Available at: https://crashstats.nhtsa.dot.gov/.

Ren, H., Song, Y., Wang, J., Hu, Y. and Lei, J., (2018) A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. [online] IEEE Press, pp.3346–3351. Available at: https://doi.org/10.1109/ITSC.2018.8569437.

Sarkar, S., Lodhi, V. and Maiti, J., (2018) Text-clustering based deep neural network for prediction of occupational accident risk: A case study.

Schonlau, M. and Zou, R., (2020) The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata*, 20, pp.3–29.

Singh, S.K., (2017) Road Traffic Accidents in India: Issues and Challenges. *Transportation Research Procedia*, [online] 25, pp.4708–4719. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2352146517307913 [Accessed 5 Nov. 2023].

Strickland, M., Fainekos, G. and Amor, H. Ben, (2017) Deep Predictive Models for Collision Risk Assessment in Autonomous Driving. [online] Available at: http://arxiv.org/abs/1711.10453.

Tamim Kashifi, M. and Ahmad, I., (2022) Efficient Histogram-Based Gradient Boosting Approach for Accident Severity Prediction With Multisource Data. *Transportation Research Record: Journal of the Transportation Research Board*, 26766, pp.236–258.

Testolina, P., Barbato, F., Michieli, U., Giordani, M., Zanuttigh, P. and Zorzi, M., (2023) SELMA: SEmantic Large-Scale Multimodal Acquisitions in Variable Weather, Daytime and Viewpoints. *IEEE Transactions on Intelligent Transportation Systems*, 247, pp.7012–7024.

La Torre, F., Domenichini, L., Branzi, V., Meocci, M., Paliotto, A. and Tanzi, N., (2022) Transferability of the highway safety manual freeway model to EU countries. *Accident Analysis & Prevention*, 178, p.106852.

Tuncal Yaman, T., Bilgiç, E. and Fevzi Esen, M., (2022) Analysis of Traffic Accidents with Fuzzy and Crisp Data Mining Techniques to Identify Factors Affecting Injury Severity. *J. Intell. Fuzzy Syst.*, [online] 421, pp.575–592. Available at: https://doi.org/10.3233/JIFS-219213.

Wang, L., Wu, J., Li, R., Song, Y., Zhou, J., Rui, X. and Xu, H., (2021) A Weight Assignment Algorithm for Incomplete Traffic Information Road Based on Fuzzy Random Forest Method. *Symmetry*, 13, p.1588.

Wang, S., Chen, Y., Huang, J., Zhou, Y. and Lu, Y., (2018) Research on the Drunk Driving Traffic Accidents Based on Logistic Regression Model. *Open Journal of Applied Sciences*, 0811, pp.487–494.

Yan, M. and Shen, Y., (2022) Traffic Accident Severity Prediction Based on Random Forest. *Sustainability*, 14, p.1729.

Yuan, Z., Zhou, X. and Yang, T., (2018) Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data. pp.984–992.

Zhang, Y., Li, H. and Ren, G., (2022) Estimating heterogeneous treatment effects in road safety analysis using generalized random forests. *Accident Analysis & Prevention*, 165, p.106507.

Zhao, Y. and Deng, W., (2022) Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning. *Applied Artificial Intelligence*, 36, pp.1–24.

APPENDIX A: RESEARCH PROPOSAL

WEATHER-RELATED TRAFFIC ACCIDENT PREDICTION USING RANDOM FOREST WITH CATBOOST AND LIGHTGBM

ABHILASHA GARG
MS-DATA SCIENCE

Research Proposal

AUGUST 2023

Abstract

Accurately forecasting traffic conditions during adverse weather events holds significant importance in optimizing transportation systems and mitigating potential risks. This study includes a thorough analysis of traffic accident prediction during weather-related events. The study harnesses the capabilities of two random forest algorithms: CatBoost, and LightGBM. The research focuses on using the Large-Scale Traffic and Weather (LSTW) dataset, a vast collection of real-world traffic and weather-related data for the United States.

Two main goals of this project are to estimate traffic patterns in the presence of weather disturbances using cutting-edge algorithms, and to meticulously compare and evaluate the predictive abilities of these algorithms. The initial phase encompasses data preprocessing of the LSTW dataset, entailing the extraction of pertinent features that capture the correlation between weather events and traffic accidents. Subsequent phases involve the rigorous training and evaluation of the CatBoost, and LightGBM models, aiming to get accurate predictions concerning traffic congestions.

Table of Contents

Background

Traffic accidents are strongly affected by weather-related factors. Notably, a thorough investigation by the National Highway Road Safety Administration (NHTSA) found that weather-related incidents accounted for 20% of all traffic deaths in the US in 2019. (Reish and Leah, n.d.)

These weather-related traffic fatalities are determined by a multitude of causes, including:

1. Inclement weather conditions such snow, rain, fog, and other meteorological phenomena can dramatically reduce visibility, finding it challenging for drivers to identify other cars and obstacles in the road.
2. The prevalence of icy road surfaces makes it harder for drivers to maintain control of their cars, enhancing the probability of accidents.
3. Violent wind gusts have the ability to cause vehicles to depart from their intended path, boosting the risk of accidents.
4. When roads are blocked subject to flooding, accidents can occur as a result of the vehicular manslaughter conditions.

For predicting traffic accidents, especially those affected by weather, machine learning has proven to be a powerful tool. Using machine learning to predict weather-related traffic accidents has become increasingly popular in recent years (Hasan et al., 2023; Luo and Wang, 2023; Testolinaet al., 2023).

The LSTW dataset, a comprehensive collection of traffic and weather event data for the United States, is a noteworthy source in this area (Moosavi et al., 2019a). The training of models using random forest for forecasting weather-related traffic incidents can make good use of this dataset.

Problem Statement

Around the world, traffic accidents are a major source of fatalities and injuries. Each year, there are more than 6 million traffic accidents in the United States alone, resulting in more than 38,000 fatalities. Many of these collisions are brought on by particular sorts of traffic incidents, such roadwork or severe weather. Various studies have been conducted by researchers (Moosavi et al., 2019b; Mondal et al., 2020; Hasan et al., 2023) across the globe with different dataset (Diaz-Ruiz et al., n.d.; Dadwal et al., 2021; Luo and Wang, 2023) in order to achieve accurateprediction based on various factors and techniques.

In this paper, we suggest developing a random forest algorithm model that could be used to evaluate big datasets of traffic data to find correlations that can be used to anticipate accidents. To create our models, we'll employ the two well-known algorithms LightGBM and CatBoost. Several criteria, including accuracy, precision, and recall, will be used to assess the performance. In addition to assessing how independent factors may affect the models' performance, we will also conduct sensitivity analysis and comparison between the two algorithms.

Research Questions

For each of the mentioned study objectives, the subsequent research questions are proposed.

- Can accidents caused by certain weather conditions be predicted using random forest?
- Determining a correlation between poor weather and road accidents.
- What features are the most effective for projecting accidents driven on by climateor construction?
- To review different techniques and evaluate the precision and model performance leveraging random forest algorithms.

Aim and Objectives

The primary goal of this study is to recommend the top model for using the random forest algorithm to forecast traffic accidents induced by meteorological occurrences in the United States.

The following are the study's objectives, which were based primarily on aims of this study:

- To investigate the trend and correlation between unfavourable weather and traffic occurrences.
- To determine the variables that are significant for forecasting traffic accidents.
- To compare CatBoost and LightGBM predictive modelling
- To assess the effectiveness of models using a range of metrics.

Significance of the Study

This study introduces an innovative approach using LightGBM and CatBoost algorithms for accurate accident prediction, focusing on feature selection and predictive model architecture. The integration of LightGBM and CatBoost is expected to improve accuracy and performance, especially in the Large-Scale Traffic and Weather (LSTW) dataset. The focus on reproducibility, reuse, and extensibility in the research community ensures transparency and reliability, facilitating future research endeavors in this domain.

Scope of the Study

For a number of reasons, the focus of this study is only on the forecasting of weather-related traffic incidents in the United States.

- First, there is a need for reliable ways to anticipate weather events because they are a significant contributing factor to traffic accidents.
- The LSTW dataset, a sizable set of weather and traffic occurrences for the United States, offers a thorough analysis of these occurrences.
- Third, two well-known machine learning algorithms, LightGBM and Catboost, may also be efficient in predicting traffic accidents.
- The final standard metrics for assessing the performance of machine learning models are accuracy, precision, and recall.

Research Methodology

The methodology used entails important procedures as the choice of target data, pre-processing the selected data, and many others. The following are these steps:

- **Data Collection**: The first step is to choose the target data. We will utilize the LSTW dataset, a sizable dataset of traffic and weather activities for the United States, in this research.
- **Data preprocessing**: Pre-processing the data is the subsequent stage. Data cleansing, outlier removal, and transforming the data into a structured and comprehensible format will be involved in this stage.
- **Data balancing**: The data may be imbalanced, meaning that there are more or fewer samples of one class than another. This can skew the results of the machine learning models. To address this, we will balance the data by oversampling or undersampling the minority class.
- **Feature Engineering**: The next step is to select the features that will be used to train the machine learning models. This is important because not all features are equally important. We will use a variety of techniques to select the features, such as univariate feature selection and recursive feature elimination.
- **Model training**: The next step is to train the machine learning models. We will use two popular machine learning algorithms, CatBoost and LightGBM, to train our models.
- **Model evaluation**: The final step is to evaluate the machine learning models. This includes evaluating the accuracy, precision, and recall of the models. We will also conduct sensitivity analysis to determine the impact of different factors on the accuracy of the models and compare the models of Catboost and LightGBM to find the best fit.

Requirements Resources

The following resources are necessary for this study:

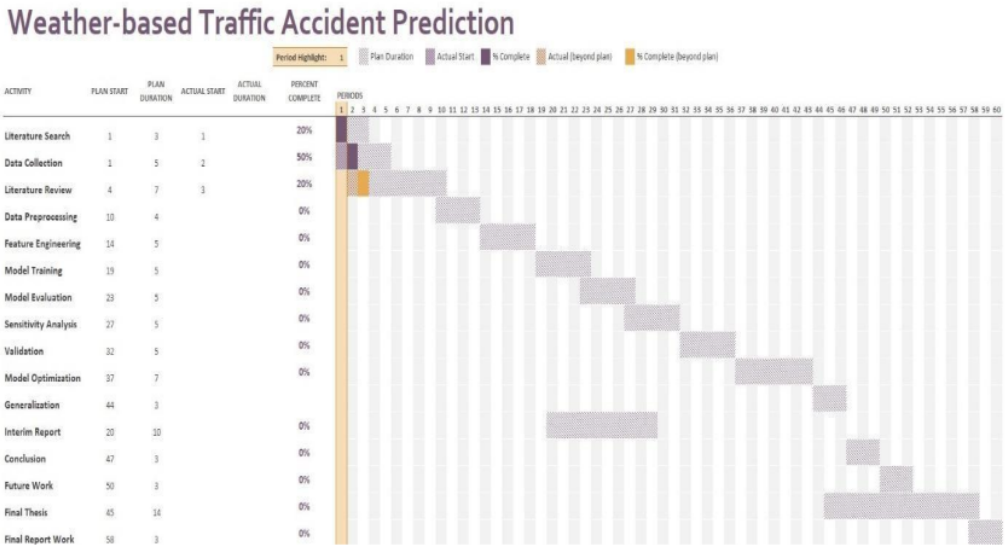
1. **Data**: The LSTW dataset is freely available.
2. **Software**: Data pre-processing, feature engineering, and machine learning will all require the usage of software. Python, R, and scikit-learn are some common choices.
3. **Hardware**: A computer with adequate memory and processing capability is required for this study.

Research Plan

Studies already conducted have shown that machine learning algorithms can be used to foreseemishaps that are caused by the weather (Moosavi et al., 2019c; Dadwal et al., 2021; Testolina et al., 2023; Wei et al., 2023). However, a thorough examination is necessary to fine-tune and prove their efficacy across various weather patterns and geographic areas of the United States.

The research plan for this study is as follows:

- Data collection and pre-processing
- Feature engineering
- Model training
- Model evaluation and validation
- Conclusion and future work



References

Dadwal, R., Funke, T. and Demidova, E., (2021) An Adaptive Clustering Approach for Accident Prediction. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. Institute of Electrical and Electronics Engineers Inc., pp.1405–1411.

Diaz-Ruiz, C.A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., Chao, W.-L., Hariharan, B., Weinberger, K.Q. and Campbell, M., (n.d.) *Ithaca365: Dataset and Driving Perception under Repeated and Challenging Weather Conditions*. [online]Available at: https://ithaca365.mae.cornell.edu/.

Hasan, A.S., Kabir, M.A. Bin, Jalayer, M. and Das, S., (2023) Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety and Security*, 156, pp.604–635.

Luo, H. and Wang, F., (2023) A Simulation-Based Framework for Urban Traffic Accident Detection. Institute of Electrical and Electronics Engineers (IEEE), pp.1–5.

Mondal, A.R., Bhuiyan, M.A.E. and Yang, F., (2020) Advancement of weather-related crashprediction model using nonparametric machine learning algorithms. *SN Applied Sciences*, 28. Moosavi, S., Samavatian, M.H., Nandi, A., Parthasarathy, S. and Ramnath, R., (2019a) Shortand long-term pattern discovery over large-scale geo-spatiotemporal data. In: *Proceedings ofthe ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.Association for Computing Machinery, pp.2905–2913.

Moosavi, S., Samavatian, M.H., Parthasarathy, S. and Ramnath, R., (2019b) A CountrywideTraffic Accident Dataset. [online] Available at: http://arxiv.org/abs/1906.05409.

Moosavi, S., Samavatian, M.H., Parthasarathy, S., Teodorescu, R. and Ramnath, R., (2019c) Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In: *GIS: Proceedings of the ACM International Symposium on Advances in Geographic InformationSystems*. Association for Computing Machinery, pp.33–42.

Reish and Leah, (n.d.) *Traffic Safety Facts 2019: A Compilation of Motor Vehicle Crash Data*. [online] Available at: https://crashstats.nhtsa.dot.gov/.

Testolina, P., Barbato, F., Michieli, U., Giordani, M., Zanuttigh, P. and Zorzi, M., (2023) SELMA:SEmantic Large-Scale Multimodal Acquisitions in Variable Weather, Daytime and Viewpoints. *IEEE Transactions on Intelligent Transportation Systems*.

Wei, W., Zou, S., Duan, W., Chen, Y., Li, S. and Zhou, Y., (2023) Spatiotemporal variability in extreme precipitation and associated large-scale climate mechanisms in Central Asia from 1950to 2019. *Journal of Hydrology*, 620.

Abhilasha_Final_Thesis-1.pdf

ORIGINALITY REPORT

9%	10%	6%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	opentransportationjournal.com Internet Source	2%
2	ops.fhwa.dot.gov Internet Source	2%
3	www.hindawi.com Internet Source	2%
4	smoosavi.org Internet Source	1%
5	www.geeksforgeeks.org Internet Source	1%
6	scholarworks.lib.csusb.edu Internet Source	1%
7	Abdelaziz Testas. "Distributed Machine Learning with PySpark", Springer Science and Business Media LLC, 2023 Publication	1%

Exclude quotes	Off	Exclude matches	< 1%

Exclude bibliography On