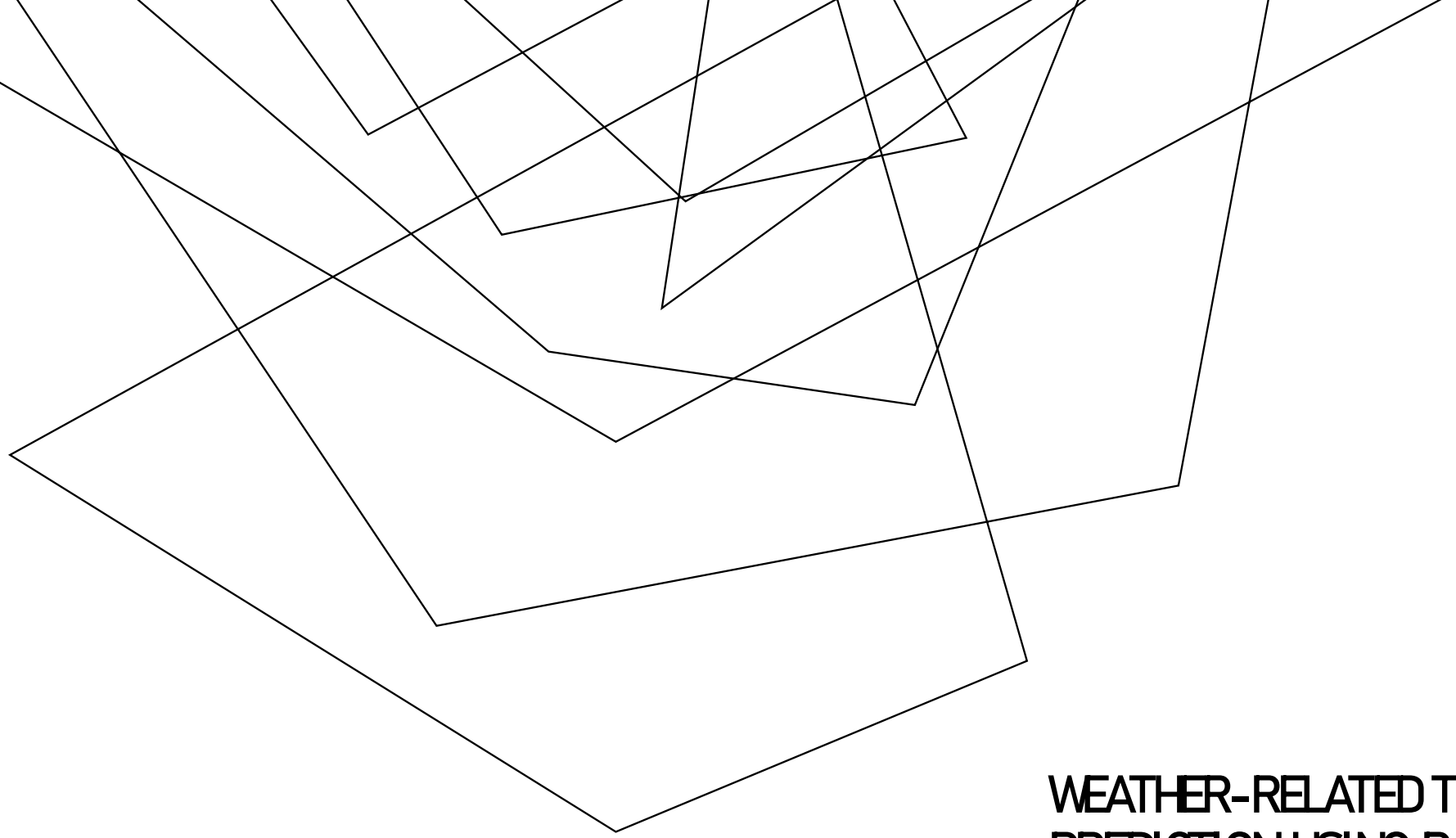


WEATHER-RELATED TRAFFIC ACCIDENT PREDICTION USING RANDOM FOREST WITH CATBOOST AND LIGHTGBM

Name: Abhilasha Garg

Course: M.S. Data Science (LJMU)

Student ID: 1088914

AGENDA

Introduction

Aims & Objectives

Literature Review

Methodology

Results & Discussions

Conclusions & Future Recommendations

INTRODUCTION

Accurate forecasting of traffic conditions amid adverse weather events stands as a pivotal factor in optimizing transportation systems and ensuring road safety.

This presentation delves into a comprehensive study focusing on the utilization of advanced machine learning algorithms—CatBoost and LightGBM—to predict traffic accidents during weather-related incidents. Through an analysis of the Large-Scale Traffic and Weather (LSTW) dataset encompassing real-world traffic and weather data across the United States.

AIMS & OBJECTIVES

Road accidents are a significant global problem, causing numerous fatalities and injuries.

Recommend the top model for using the random forest algorithm to forecast traffic.

Compare the performance of two machine learning algorithms, LightGBM and CatBoost, in predicting weather-related traffic accidents.

To review different techniques and evaluate the precision and model performance leveraging random forest algorithms.

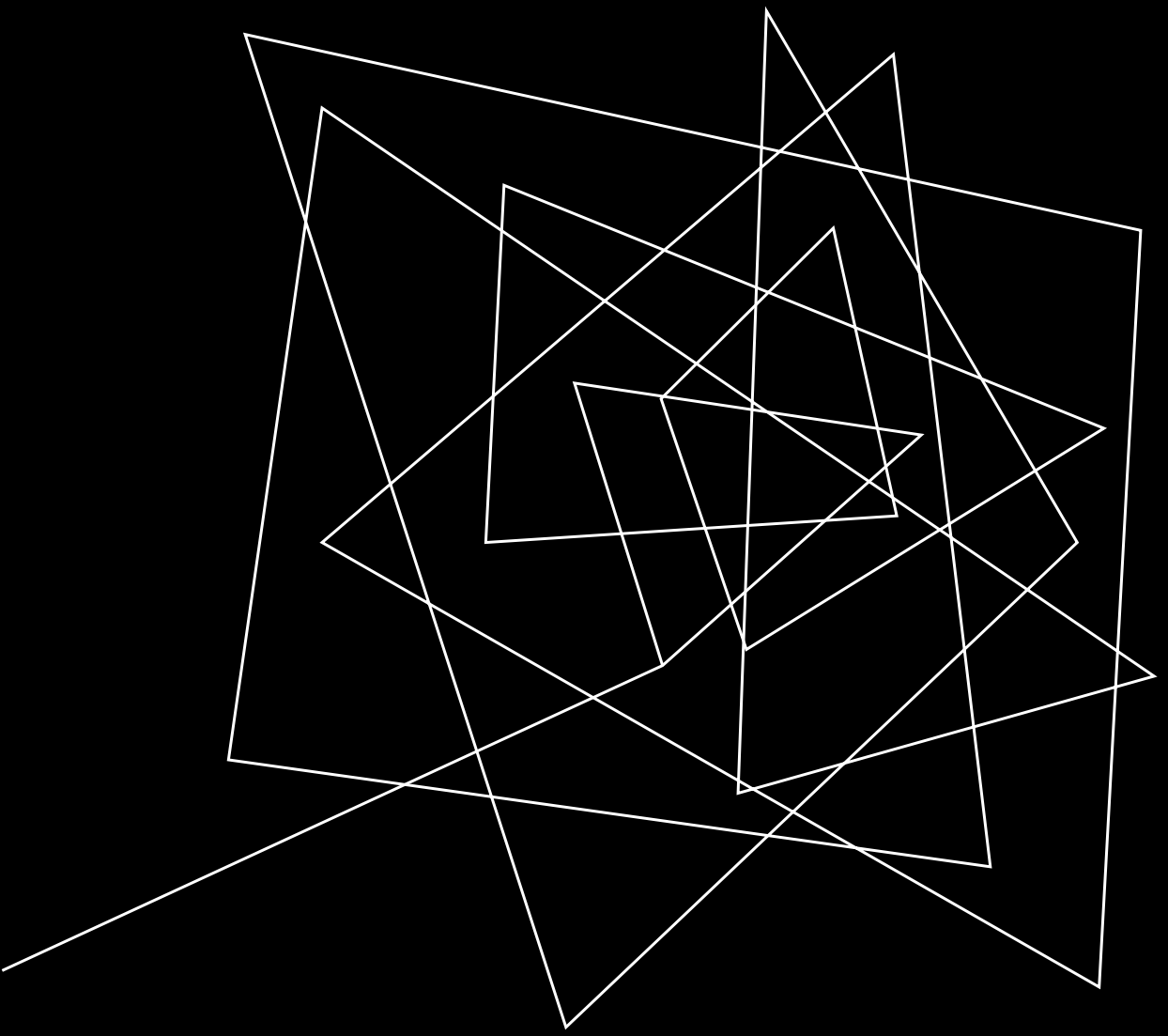
RESEARCH QUESTIONS

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 by climate?

To review different techniques and evaluate the precision and model performance leveraging random forest algorithms.



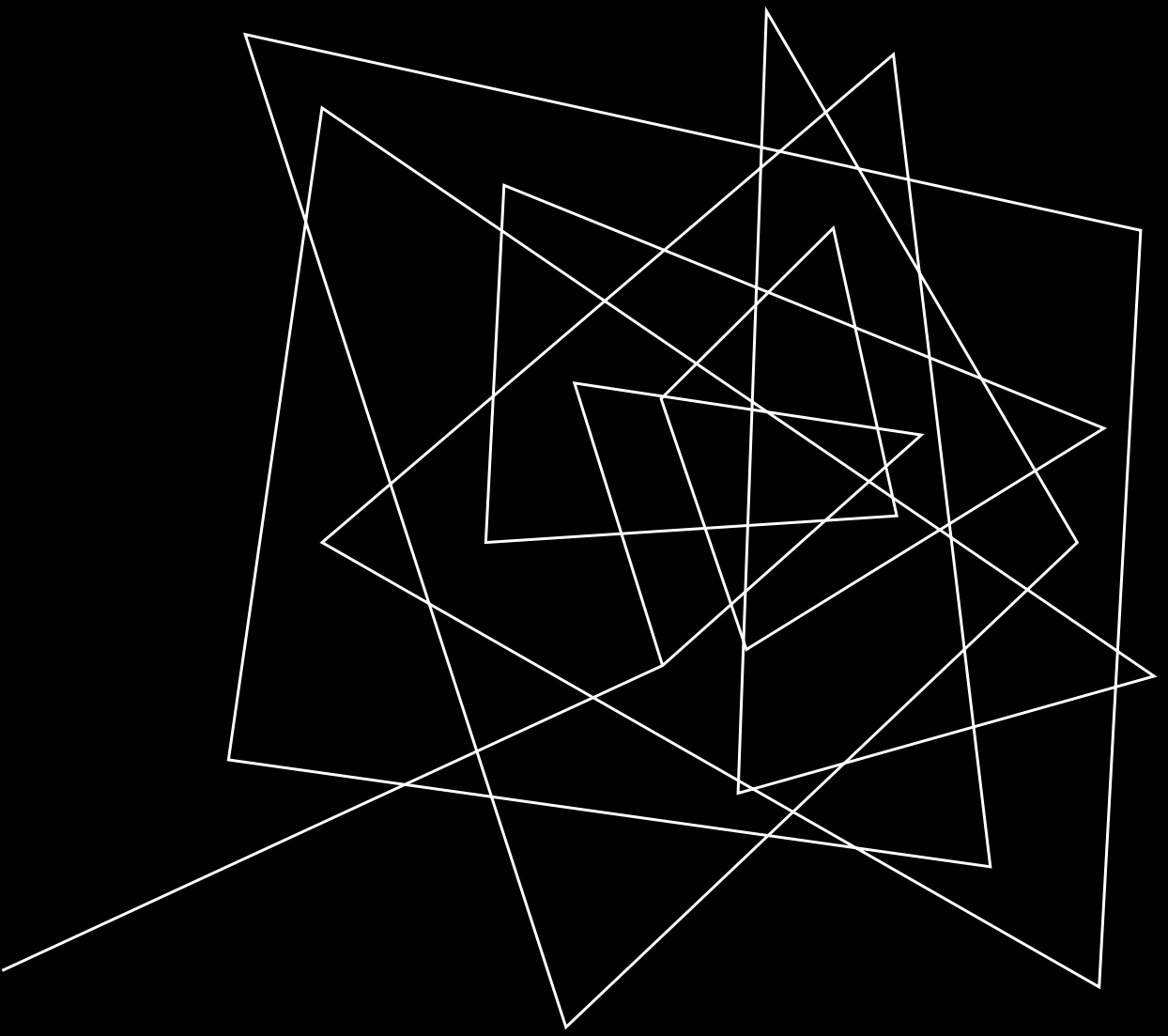
LITERATURE REVIEW

TIMELINES

Year	Events
1771	The first traffic accident happened when Nicolas Joseph Cugnot crashed his steam-powered vehicle into a wall.
1869	The first recorded road accident death occurred when an Irishwoman was run over by an experimental steam-powered car.
2003	Lyon et al. and El-Basyouny and Sayed studied accident prediction models.
2004	Eisenberg used a dataset of 456,000 collisions to determine the effects of traffic accidents.
2013	Jagannathan et al. proposed a novel decision support system.
2017	Najjar et al. used a large-scale dataset to investigate real-time traffic accident prediction.
2018	An et al. used weather and traffic data to predict the frequency of accidents on a highway route.
2018	Mon et al. used the conjoint analysis (CA) approach to estimate the fatal injury costs from traffic accidents in Malaysia.
2018	Yuan et al. used deep learning for traffic accident prediction.
2018	Sarkar et al. proposed a methodology for predicting accident risk using both structured and unstructured data.
2019	Moosavi et al. created a dataset of over 2.25 million traffic accident instances in the US over three years to analyze the various factors of an accident.
2021	Jomnonkwao et al. used the WTP-CV method to calculate the value of statistical injury (VSI) and the value of statistical life (VSL) for motorcyclists in Thailand.
2021	Ovi et al. proposed a system for real-time traffic accident prediction.
2021	Hossain et al. used a machine learning method to predict traffic accident severity in South Africa.
2022	Zhao and Deng used heterogeneous ensemble learning and millions of traffic accident data from the United States to create an accident duration prediction model.
2022	La Torre et al. proposed a European Average Prediction Model (APM) based on the Highway Safety Manual (HSM) and a new methodology for transferring the HSM to different European rural freeways.

MAJOR CONTRIBUTIONS AND OUTCOMES

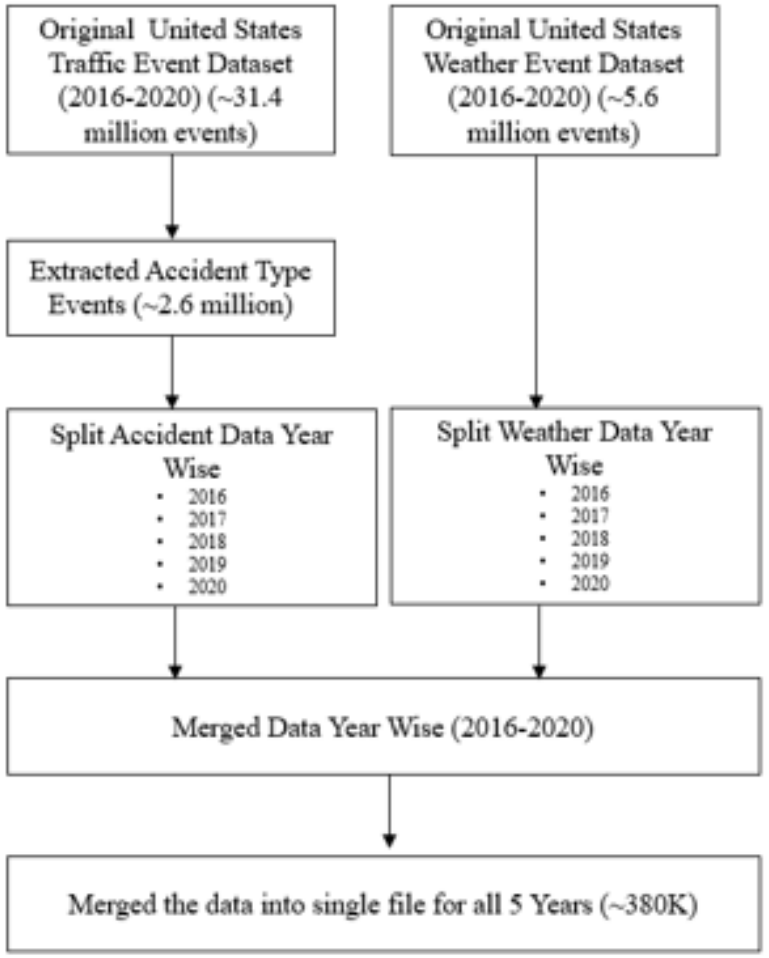
Zeng et al. [28]	Mon et al. (2018)	Chan et al. [6]	Kataoka et al. [14,22]	Herzig et al. [13]
Propose a multitask learning approach to improve accident anticipation accuracy, which also localizes risky regions associated with accidents.	Estimated fatal injury costs from traffic accidents in Malaysia using the conjoint analysis (CA) approach.	Introduced the Street Accident dataset that contain videos captured by dashcams and proposed an LSTM based model with spatial attention module to estimate the likelihood of accident occurrence in the near future for each frame	Introduced a large-scale dataset near-miss incident database (NIDB), and proposed an adaptive loss function to facilitate the earliest anticipation of an accident.	Presented the Collision dataset, which includes near-miss incident scenes in addition to accident videos. They propose spatio-temporal action graphs that effectively model the relationship between objects associated with an accident.



METHODOLOGY

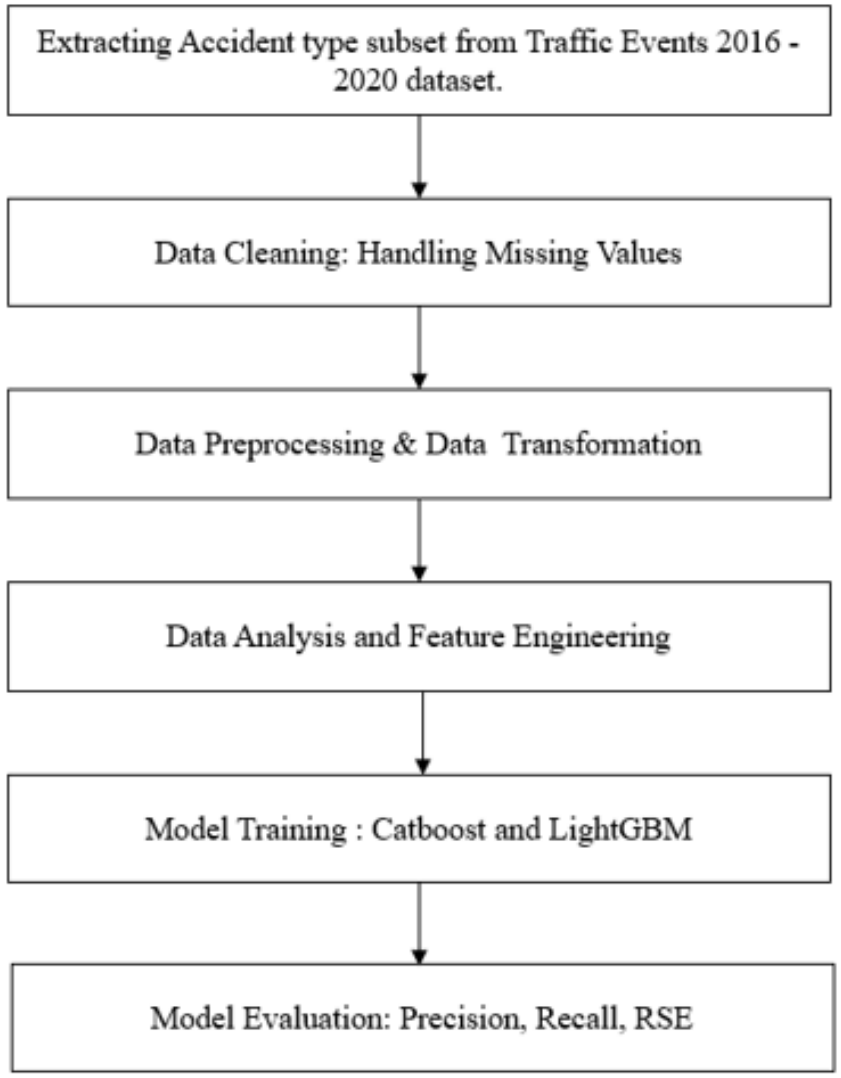
DATA SELECTION

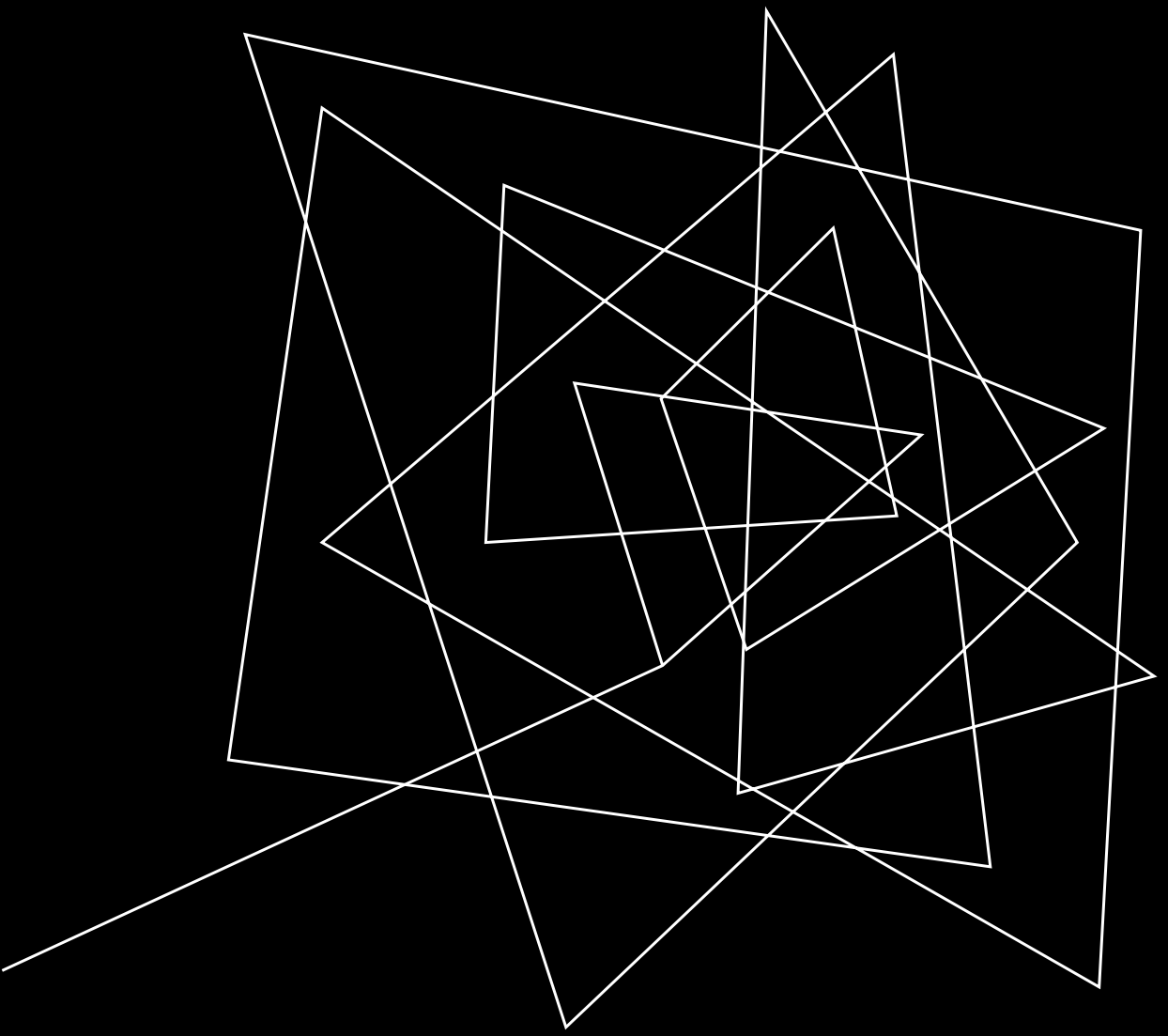
- We have used two datasets: Traffic Events and Weather Events containing around 31.4 million records in traffic events dataset and around 6 million in weather events dataset from the year 2016-2020 from all across Contiguous United States.
- Traffic event dataset contains several types of events like 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 split both the dataset year wise
- We have then merged the Weather event dataset and traffic accident dataset to form single dataset containing around 380K records.



END TO END PIPELINE

- Once we have the final dataset, data cleaning was performed by dealing with outliers, checking data skewness and handling missing values.
- With the clean dataset, data preprocessing was done and data is transformed into dummies using MinMaxScaler.
- SMOTE and Random Under Sampler Class balancing was performed.
- Regressor and Classifier Algorithm was used to train the model and calculate the relevant metrics.



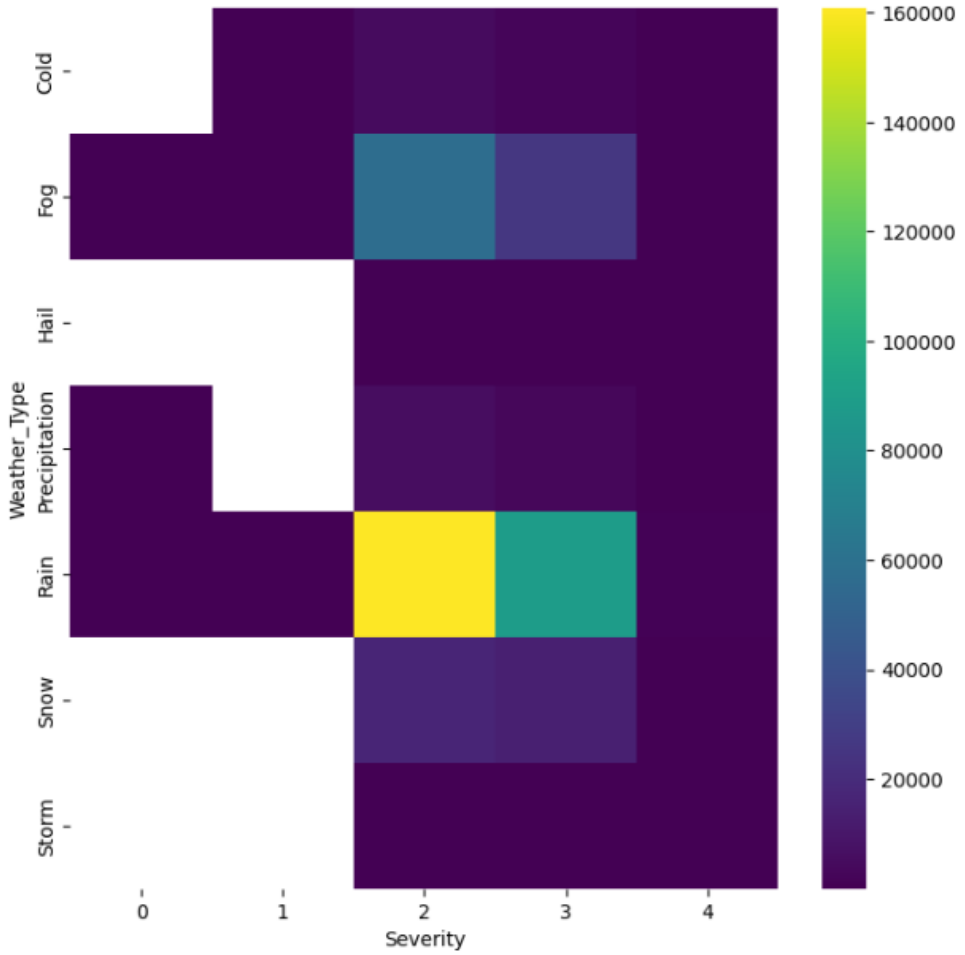


RESULTS & DISCUSSIONS

ANALYSIS RESULT

The most severe accidents occur during foggy days and rainy periods.

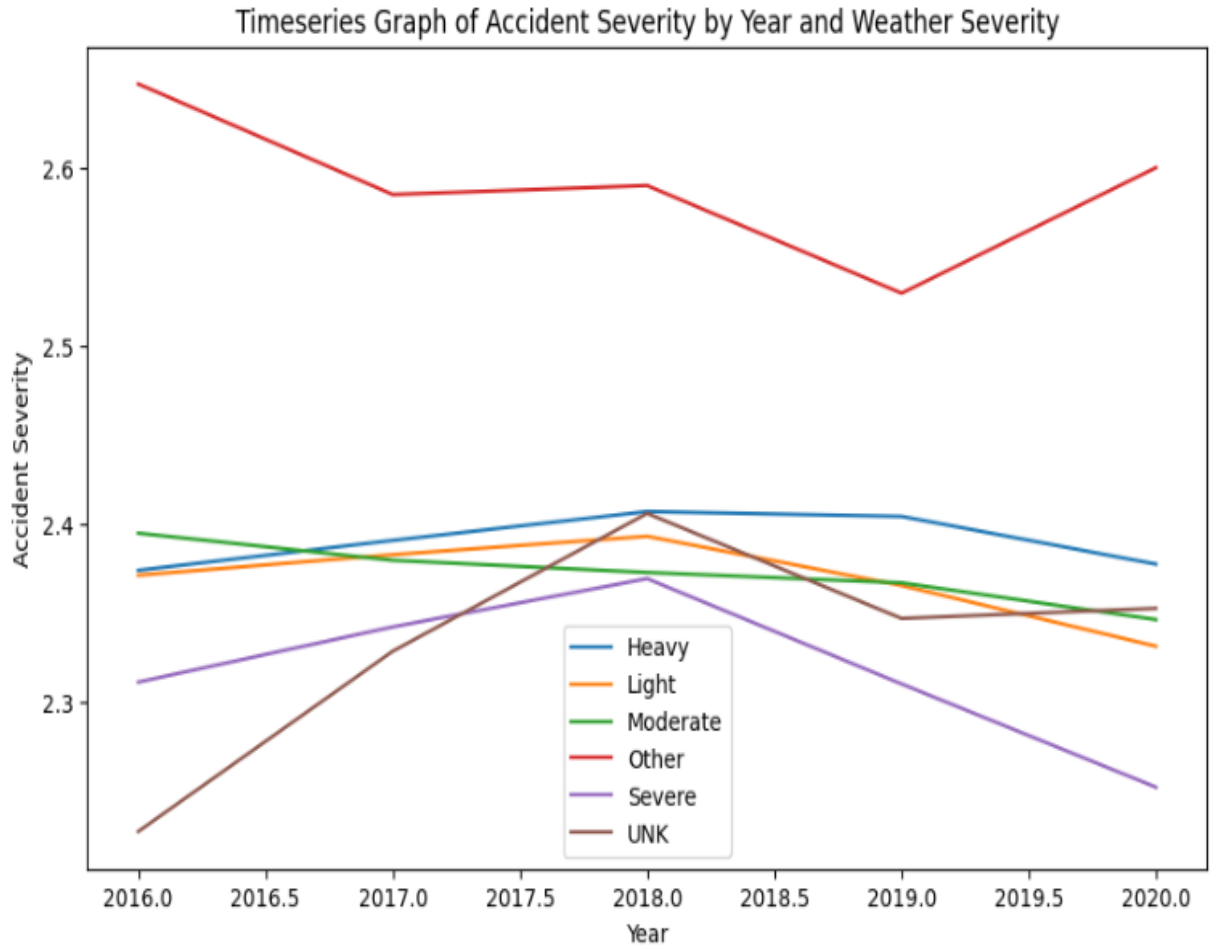
- This is likely due to reduced visibility, which can make it difficult for drivers to see hazards, while traffic congestion can lead to more opportunities for collisions.
- Interestingly, weather conditions such as storms, hail, and cold do not seem to be major contributors to accidents or traffic incidents.



ANALYSIS RESULT

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

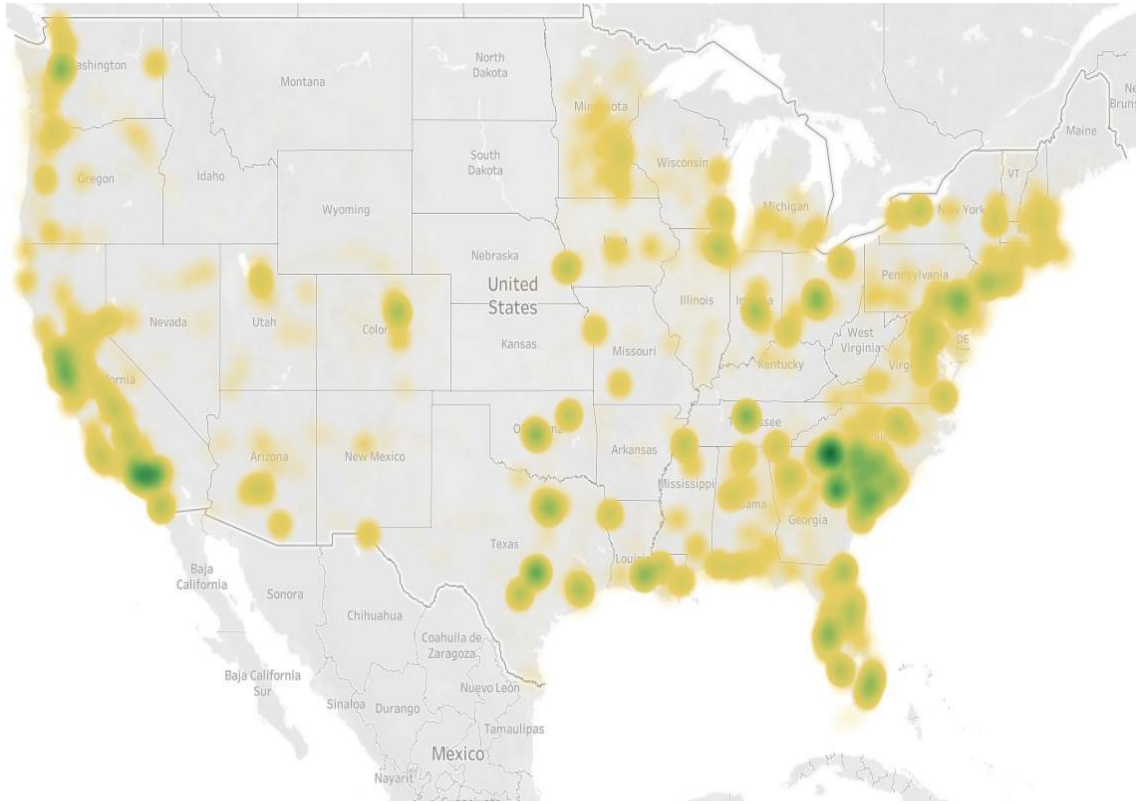
- Accident severity is highest in 2019, followed by 2018 and 2017.
- Most impacted season is monsoon (June-September).
- Accident are more likely to happen during "heavy" weather severity category.



ANALYSIS RESULT

The accident distribution map shows that accidents are more common in urban areas and along major highways. The states with the highest number of accidents are California, South Carolina, Florida, and New York.

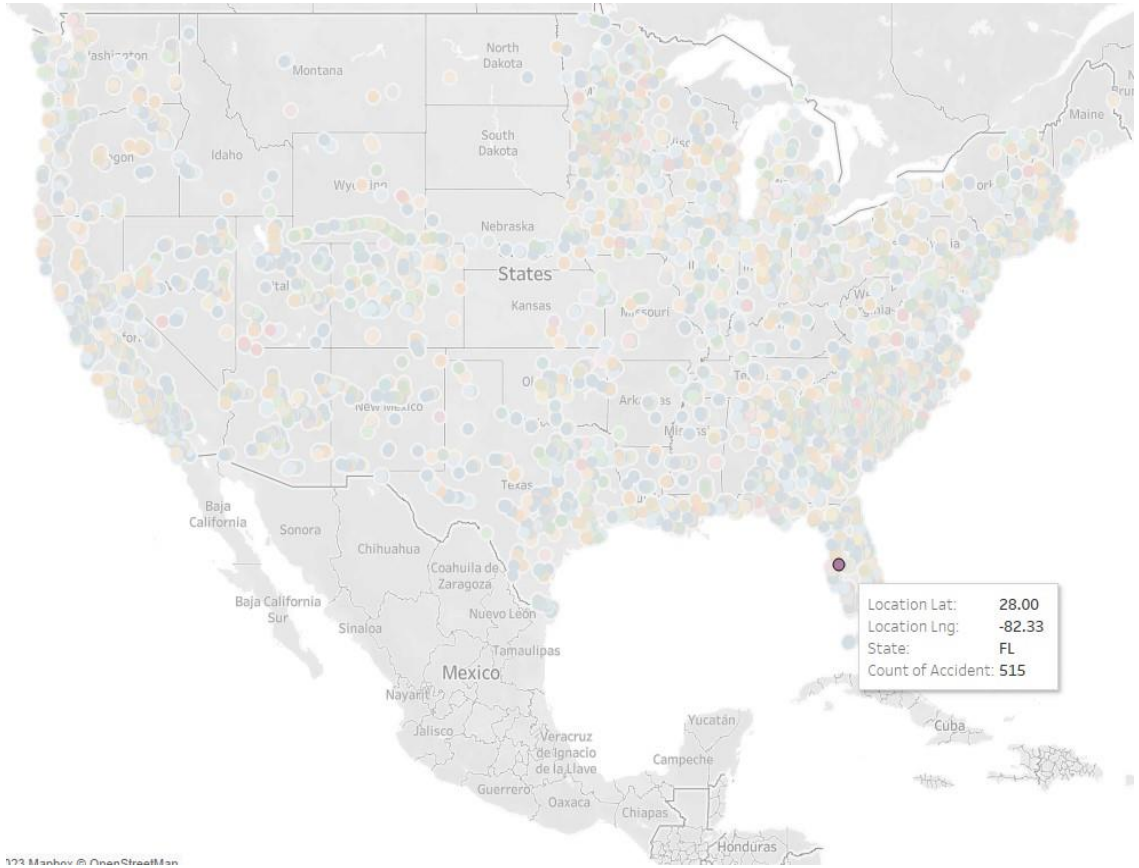
- California: Over 2.5 million accidents reported in 2020. The majority of these accidents occurred in urban areas such as Los Angeles, San Francisco, and San Diego.
- South Carolina: Over 2 million accidents reported in 2020
- Florida: Over 1.7 million accidents reported in 2020, in Miami, Orlando, and Tampa.
- New York: Over 1.6 million accidents reported in 2020, in New York City and Buffalo.



ANALYSIS RESULT

Most fatalities 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

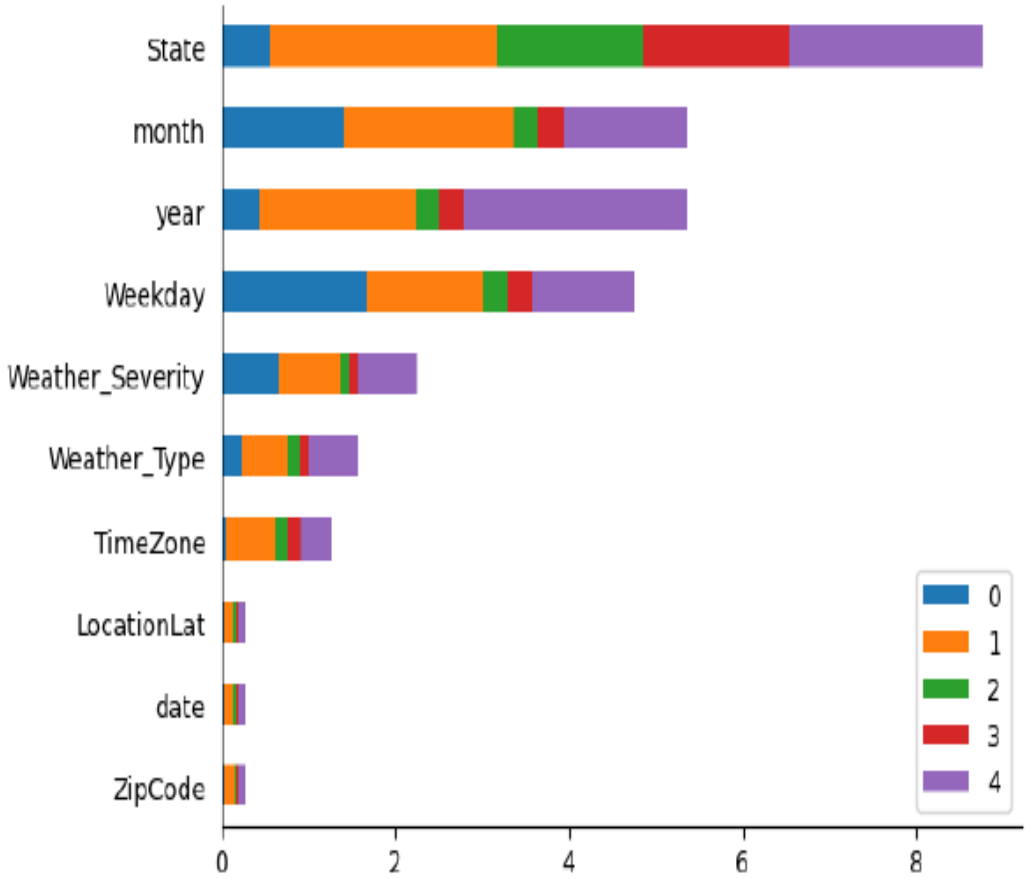
- Being a costal state, it is mostly likely to observe maximum rainfall throughout the year in different cities.
- In Florida Winter rains are due to the clash of air masses, which generate disturbances and draw mild and moist air from the Gulf of Mexico, while summer rains are mainly due to heat thunderstorms.



IMPORTANT FEATURES

Key Features for Prediction of road accident based on accident severity:

- State
- Month
- Year
- Weekday
- Weather_Severity
- Weather_Type
- TimeZone



REGRESSOR METRICS

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.

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 Random Forest Boosting Algorithm using GBDT gradient was able to identify 54% data variance and we could see the accuracy and recall are about 99.9 percent, this suggests that it was quite capable of properly identifying the actual positive. Also the root mean squared error is fairly moderate making it decent fit.

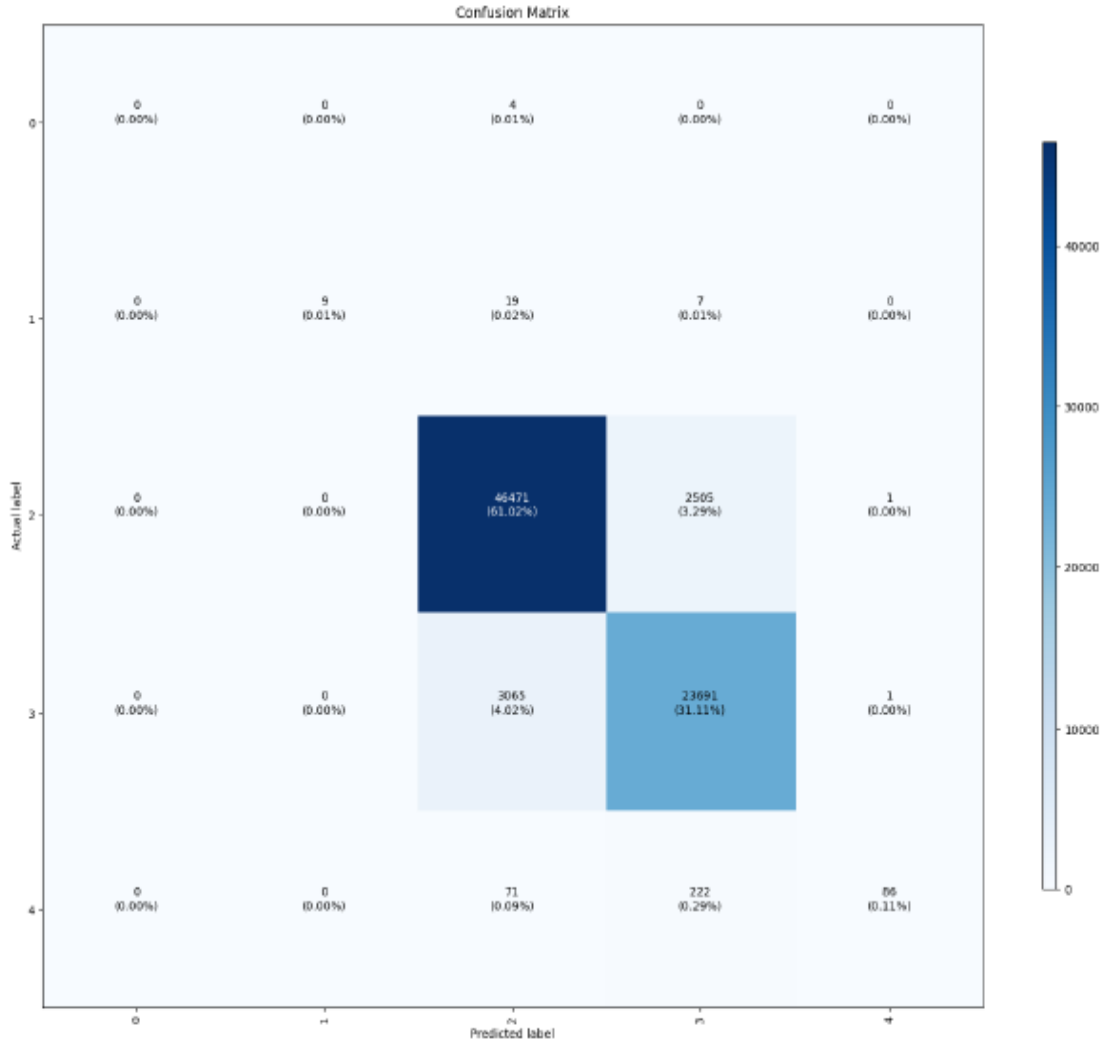
Parameters	LightGBM Regressor - gbdt
R2 Score	0.5404
Precision	0.999
Recall	1
Mean Absolute Error	0.2342
Mean Squared Error	0.1111
Root Mean Squared Error	0.3333

CLASSIFIER METRICS

- Metrics table shows that 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.
- 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.

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

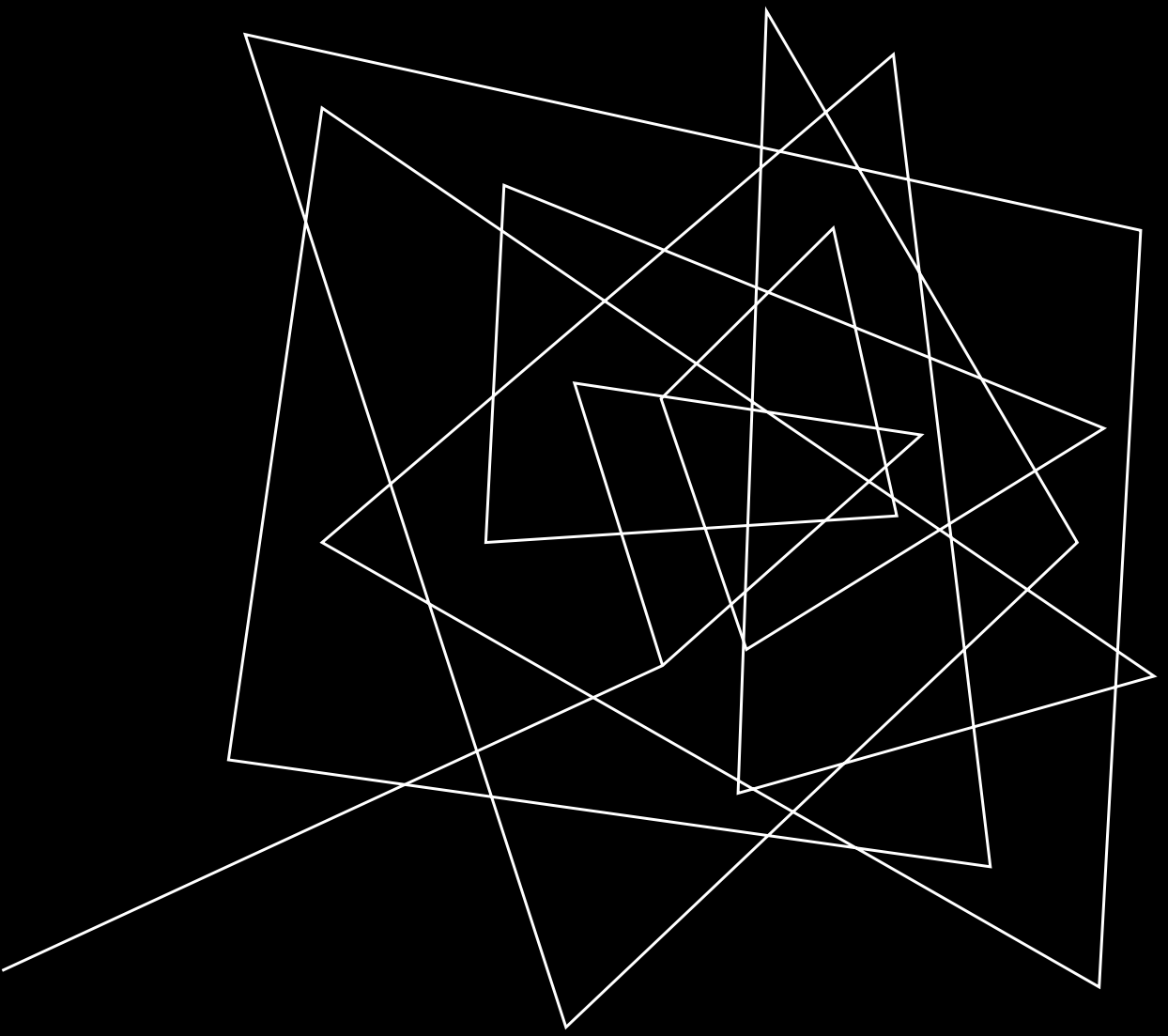


COMPARISON

- 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.
- In terms of precision, the CatBoost Regressor and LightGBM Regressor both achieved precision scores of 0.999, and SageMaker Autopilot model achieved a precision score of 0.922544, 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.

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

	Under sample	
Parameters	**Catboost Regressor**	**LightGBM Regressor**
R2 Score	0.6598	0.653
Precision	1	1
Recall	1	1



CONCLUSION & FUTURE RECOMMENDATIONS

CONCLUSION

- CatBoost and LightGBM Regressors achieve higher R^2, precision, and recall.
- LightGBM Regressor outperforms CatBoost Regressor in training speed.
- LightGBM is a suitable choice for large-scale data processing and real-time prediction tasks.
- Regressor achieved an R^2 score of up to 92%, precision of 1.0, and recall of 1.0
- Over 70% of accidents occurred during fog and rain.
- Weather_Type, Weekday, TimeZone, month, Location Latitude, and Location Longitude are the key features.
- Our analysis yielded compelling evidence of a strong association between poor visibility weather, such as fog and rain, and the occurrence of road accidents.

FUTURE RECOMMENDATIONS

- 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.
- 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.
- 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.
- 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.

THANK YOU

Author: Abhilasha Garg

Under the supervision of Channabasva Chola

with Upgrad & Liverpool John Moores University