Predicting Road Accident Risk with Machine Learning: A Data-Driven Approach to Safer Roads

Md. Mehedi Hasan
Data Scientist || AI & ML enthusiast || Researcher

October 19, 2025

Introduction

Road safety remains a global challenge, with thousands of accidents occurring daily due to a combination of environmental, infrastructural, and human factors. Many of these elements are measurable, yet often underutilized for predictive insights.

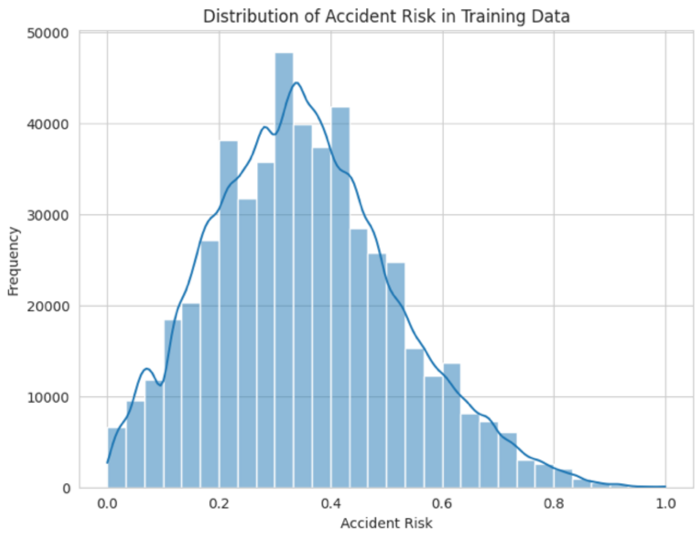
In this project, I explored how **machine learning** can transform raw traffic data into **actionable accident risk predictions** supporting smarter infrastructure design, data-informed policy, and real-time traffic interventions.

Project Overview

Using a dataset containing **over 500,000 traffic records**, I developed an end-to-end machine learning pipeline in Python to model accident risk.

The dataset included:

- **Road features:** type, number of lanes, curvature, speed limits
- **Environmental factors:** lighting, weather, time of day
- **Historical trends:** number of reported accidents per road segment



Distribution of the target variable.

The objective: **Predict accident risk for unseen road segments** using supervised ML.

Key Insights from EDA

Exploratory data analysis revealed several notable patterns:

- **Curvature matters:** A moderate positive correlation (**0.54**) was observed between road curvature and accident risk.
- **Speed influences risk:** Higher speed limits correlated positively (≈ **0.43**) with accident risk.
- **Clean but skewed data:** No missing values were found, but several features (e.g., num_reported_accidents) were heavily right-skewed.
- **Class imbalance:** Categorical variables like holiday and school_season were imbalanced, requiring careful handling to avoid model bias.

These insights guided targeted preprocessing strategies to ensure robust model performance.

Data Preprocessing

The preprocessing pipeline addressed skewness, categorical encoding, and distributional issues:

- **Log transformations** were applied to skewed numerical features to stabilize variance.
- **One-hot encoding** converted categorical variables (road_type, lighting, weather, time_of_day) into numerical representations.
- **Outliers and imbalance** were addressed to improve generalization.

This rigorous preparation formed the foundation for accurate modeling.

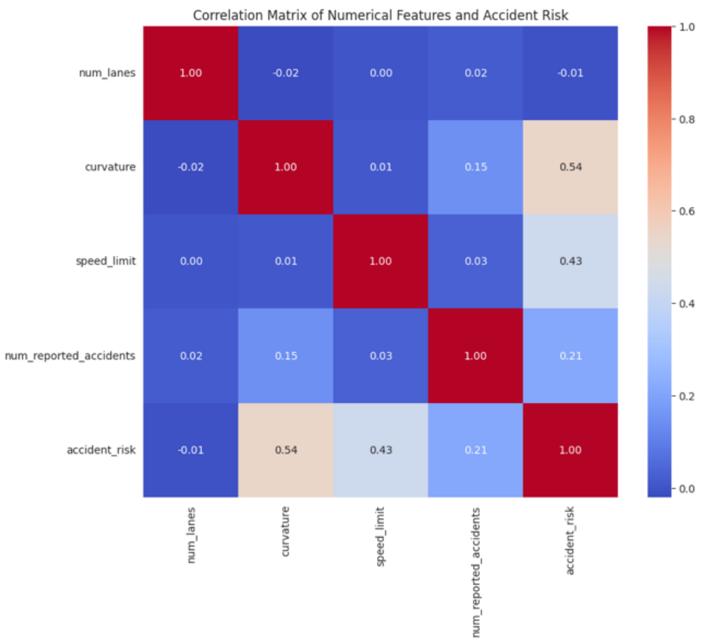
Model Development

I employed **LightGBM Regressor**, a gradient boosting framework optimized for speed and performance on large, structured datasets.

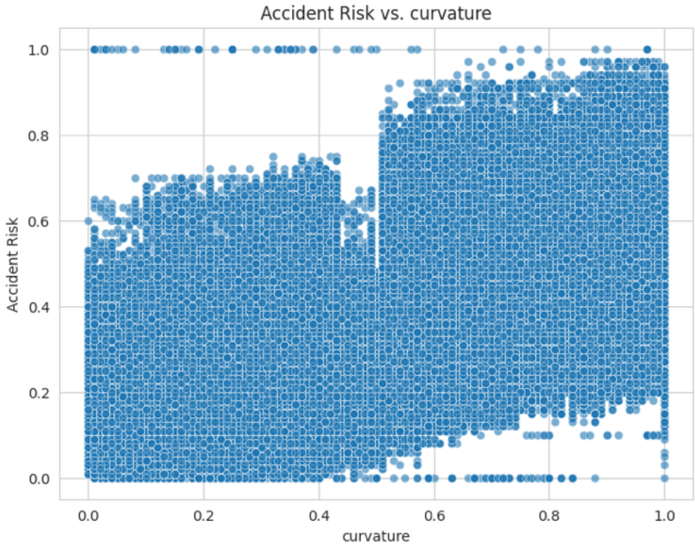
Using an 80/20 train–validation split, the model achieved:

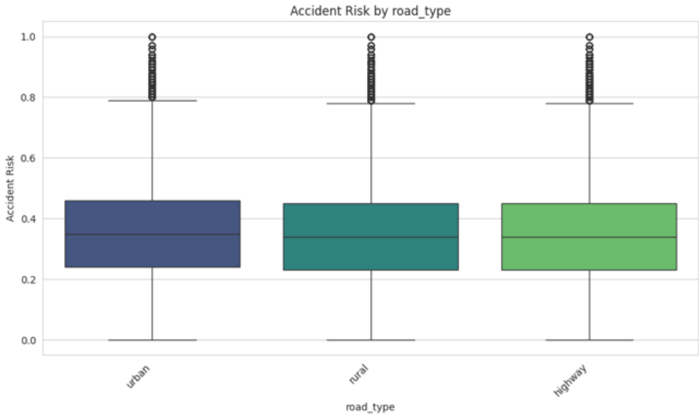
- **MSE:** 0.0032
- **RMSE:** 0.0564
- **MAE:** 0.0438

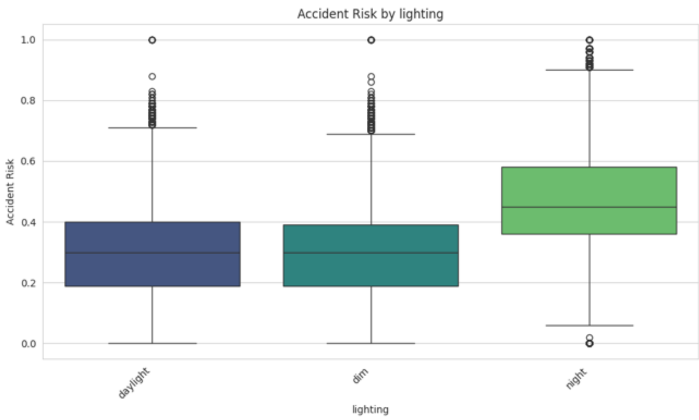
These metrics demonstrate **strong predictive power**, confirming that accident risk can be effectively modeled with the available features.

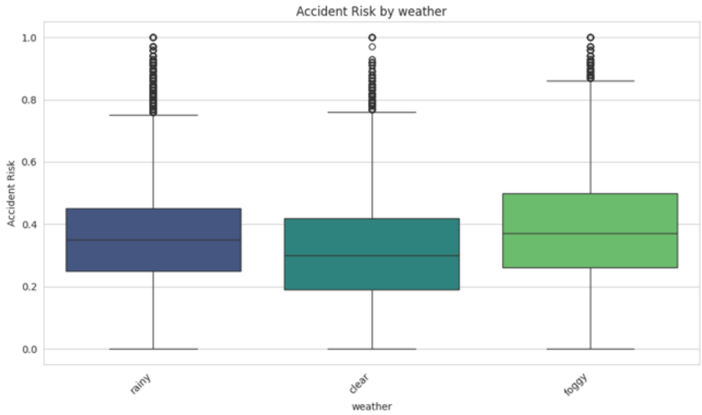


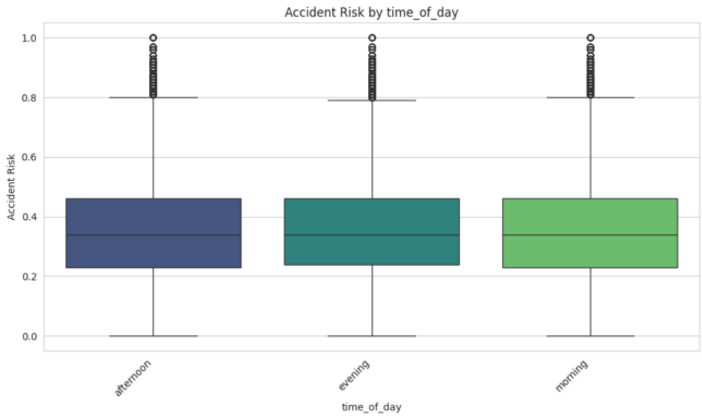
Corelation Heatmap

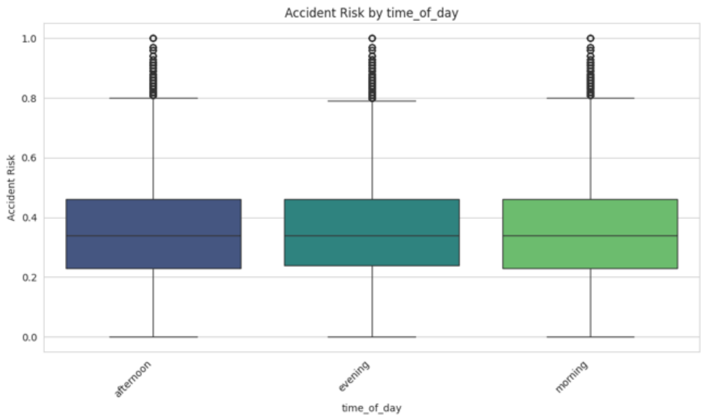

Accident Risk vs. curvature
Accident Risk
curvature
0.0
0.2
0.4
0.6
0.8
1.0

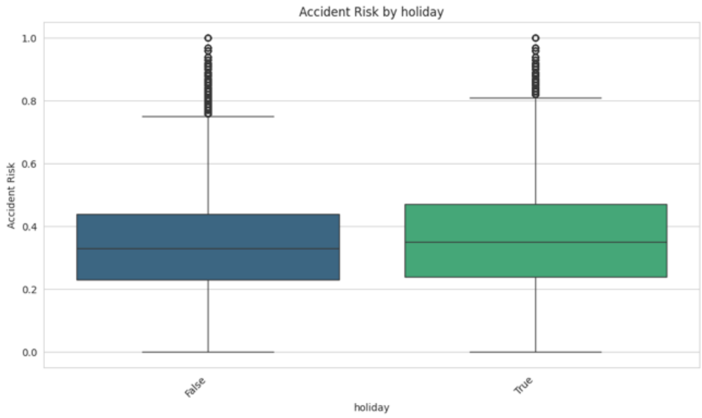

Accident Risk by road_type
Accident Risk
urban
rural
highway
road_type

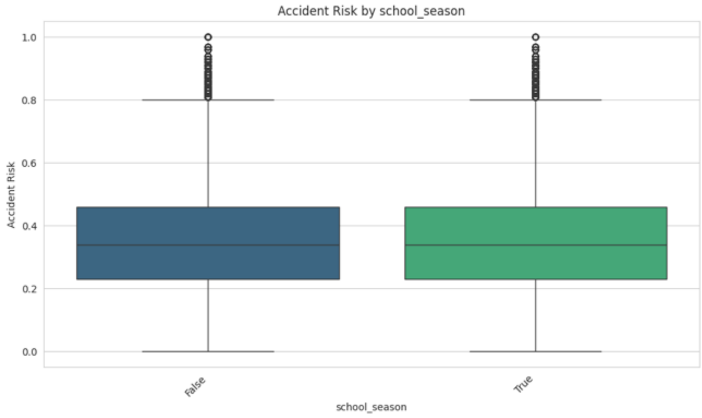

Accident Risk by lighting
Accident Risk
daylight
dim
night
lighting


Accident Risk by weather
Accident Risk
1.0
0.8
0.6
0.4
0.2
0.0
rainy
clear
foggy
weather


Accident Risk by time_of_day
Accident Risk
1.0
0.8
0.6
0.4
0.2
0.0
afternoon
evening
morning
time_of_day


Accident Risk by time_of_day
Accident Risk
1.0
0.8
0.6
0.4
0.2
0.0
afternoon
evening
morning
time_of_day





Future Directions

Potential extensions of this work include:

- Hyperparameter tuning and model comparison
- Richer feature engineering (e.g., road–weather interactions)
- Integration with real-time traffic systems for **dynamic risk-aware routing**
- Supporting urban planners with **early identification of high-risk zones**

Why This Matters

Predictive modeling for accident risk is not just a technical exercise — it's a **public safety tool**. By leveraging machine learning, we can help:

- Urban planners **design safer roads**
- Authorities **prioritize interventions**
- Navigation systems **deliver risk-aware routing** to millions of drivers

This project highlights how **data science and ML can directly contribute to saving lives** on the road.

Tech Stack

- **Languages/Libraries:** Python, Pandas, Seaborn, Scikit-learn, LightGBM

- **Workflow:** EDA → Preprocessing → Modeling → Validation → Prediction

Conclusion

This work demonstrates the impact of **well-structured data pipelines and machine learning** in addressing real-world safety challenges. By combining rigorous analysis with predictive modeling, we can build safer, smarter transportation systems.

Md. Mehedi Hasan

Data Scientist || AI & ML enthusiast || Researcher