Road Accident Analysis & Incident Count Prediction – India 2020

1. Introduction

Road accidents pose a significant threat to life and infrastructure in India, especially in million-plus cities with high vehicle density and complex traffic patterns. This project analyzes road accident data for Indian million-plus cities for the year 2020 and builds a machine learning model to predict the number of incidents (Count) based on city, cause, and outcome information.

The study supports regulatory authorities, traffic planners, and policymakers by providing insights into high-risk scenarios and a prediction tool to estimate accident counts under different conditions.

2. Objectives

- Perform detailed Exploratory Data Analysis (EDA) on 2020 accident data.
- Identify high-risk cities, causes, and outcomes related to road accidents.
- Build & compare regression models to predict incident count (Count).
- Select the best-performing model (CatBoost Regressor).
- Develop a Streamlit web application for deployment.

3. Dataset Description

The dataset contains accident records for India's million-plus cities in 2020. Key columns include:

- Million Plus Cities
- Cause Category
- Cause Subcategory
- Outcome of Incident
- Count

Data cleaning steps included removing duplicates, fixing column names, and filling missing Count values.

4. Exploratory Data Analysis (EDA)

EDA revealed trends in city-level accident counts, cause categories, outcomes, and their relationships. Cities with consistently higher accident counts emerge as hotspots. Categories such as Road Features and Impacting Vehicle/Object show the highest contributions. Fatal

incidents, while fewer, remain critical. A Cause vs Outcome heatmap reveals severity patterns for targeted safety measures.

5. Machine Learning Approach

Problem Type: Regression
Target Variable: Count
Features: All four columns are categorical, making CatBoost ideal.

Models evaluated:
- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor
- CatBoost Regressor

6. Model Performance

Final comparison:
Linear Regression – MAE 114.68 | RMSE 211.92 | R^2 0.352
Gradient Boosting – MAE 91.46 | RMSE 202.99 | R^2 0.405
Random Forest – MAE 70.27 | RMSE 186.14 | R^2 0.500
CatBoost – MAE 60.08 | RMSE 158.34 | R^2 0.638

CatBoost achieved the highest performance with 63.8% variance explained.

7. Deployment Summary

The final CatBoost model was saved as road_accident_count_model.pkl. A Streamlit app loads this model, accepts user inputs (City, Cause Category, Subcategory, Outcome), and predicts incident count.

8. Business Impact

- Data-driven insights into accident patterns.
- Better decision-making for road safety authorities.
- Helps prioritize city-level and cause-related interventions.
- Supports planning, budgeting, and risk assessment.

9. Future Enhancements

- Add multi-year datasets.
- Include weather, population, and traffic density.
- Build accident severity classification.