Accident Severity Prediction using Machine Learning

Analyzing factors influencing road accident severity

Nithin (CS23B1102)
Niranjan (CS23B1076)

Course Instructor: Dr. Preeth R, Assistant Professor

IIITDM Kancheepuram

Problem Statement

Road accidents remain a critical global public health challenge, claiming millions of lives annually and causing significant economic losses. Understanding the factors that influence accident severity is essential for developing effective prevention strategies.

Key Objectives

- Enhance road safety through predictive analytics
- Support evidence-based infrastructure planning
- Identify high-risk patterns to reduce fatal incidents

Research Goal: Develop machine learning models to predict accident severity based on environmental, temporal, and infrastructure features.

Dataset Overview

Data Source

US Accidents Dataset (2016-2023)

Comprehensive nationwide traffic accident records

Sample Size

309,136 Cases

Robust dataset enabling reliable pattern detection

Target Variable

Severity (1-4)

Ranging from minor to severe accidents

Key Features Analyzed

- Temporal: Start_Time, day of week
- Location: State, coordinates
- Environmental: Temperature, visibility, weather conditions
- Infrastructure: Amenity, crossing, traffic signals
- Road features: Bumps, junctions, stop signs
- Other contextual factors

Data Preprocessing Pipeline

Data Cleaning

Handled missing values and removed irrelevant columns to ensure data quality

Encoding

Converted categorical variables using label encoding and one-hot encoding techniques

Normalization

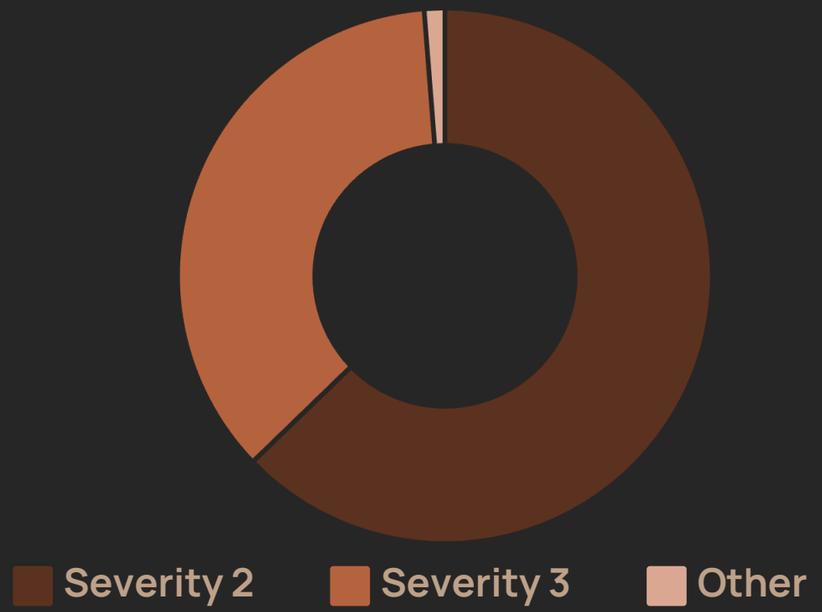
Applied StandardScaler to numerical features for consistent model input

Train-Test Split

80% training data, 20% testing data for robust evaluation

Exploratory Data Analysis

Severity Distribution Reveals Critical Patterns



Moderate accidents (Severity 2) dominate at 63%, while severe accidents (Severity 3) account for 36% of cases.

Geographic Hotspots

- **California:** Highest accident frequency
- **Florida:** Second-highest concentration
- **Texas:** Significant accident volume

Temporal Patterns

- Peak during weekdays vs. weekends
- High-risk hours: 7-9 AM and 4-6 PM (commute times)

Feature Insights

Critical Factors Influencing Accident Severity

Weather Conditions

Adverse weather significantly impacts severity:

- Fog and heavy rain reduce visibility
- Low visibility correlates with higher severity
- Clear weather shows lower accident rates

Temporal Factors

Time-based patterns reveal risk windows:

- Night-time accidents tend to be more severe
- Rush hour traffic increases collision frequency
- Weekend patterns differ from weekdays

Infrastructure Elements

Road features play a protective role:

- Traffic signals reduce accident severity
- Missing crossings correlate with higher risk
- Absence of safety features increases danger

Strong correlations observed: Visibility, humidity, and infrastructure presence show significant relationships with severity outcomes.

Model Selection Strategy

01

Logistic Regression

Baseline linear model providing interpretable coefficients and solid foundation for comparison

02

Random Forest

Ensemble learning approach using multiple decision trees to capture non-linear relationships

03

XGBoost

Advanced gradient boosting algorithm with superior handling of imbalanced data and complex patterns

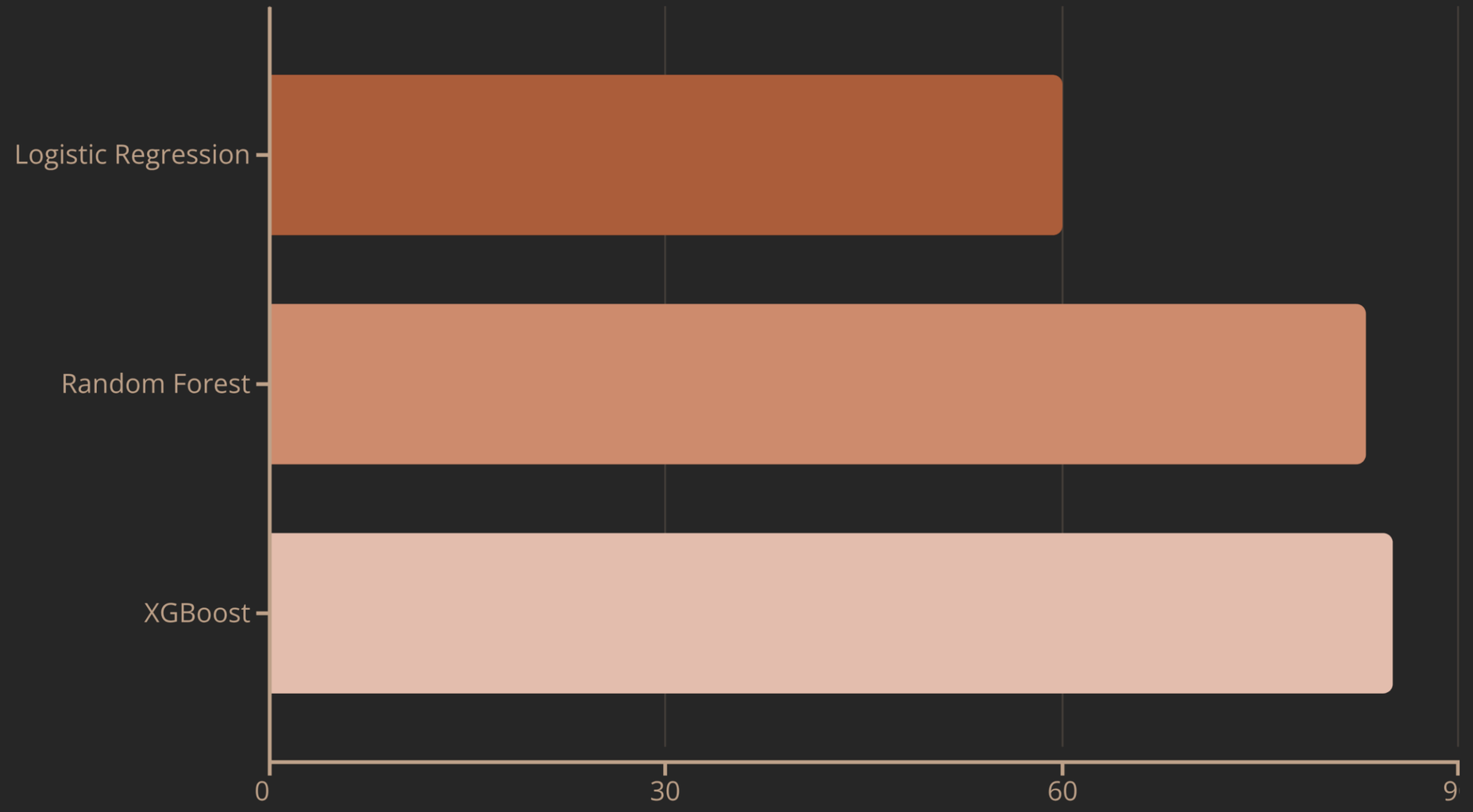
Methodology Highlights

- Applied class balancing techniques to address skewed severity distribution
- Used cross-validation for robust performance estimation
- Tuned hyperparameters for optimal results

Evaluation Metrics

- Overall accuracy
- Confusion matrix analysis
- F1-score for class balance
- Precision and recall per severity level

Model Performance Comparison



Logistic Regression

~60% Accuracy

Solid baseline with good interpretability and reasonable precision for initial modeling

Random Forest

~83% Accuracy

Strong performance with excellent feature importance insights and robust predictions

XGBoost

~85% Accuracy

Best performer with superior imbalance handling and highest predictive capability

Results & Key Findings

Model Performance Winner

XGBoost emerged as the top-performing model, achieving 85% accuracy and demonstrating superior ability to handle class imbalance while capturing complex feature interactions.

Most Important Predictive Features

Temperature

Extreme temperatures correlate with severity changes

Visibility

Low visibility is a strong severity predictor

Traffic Signals

Infrastructure presence reduces accident severity

Models successfully distinguish between moderate and severe accidents, with Random Forest and XGBoost providing reliable, actionable predictions for real-world deployment.

Conclusion & Impact

Technical Achievement

Successfully developed ML models predicting accident severity with up to 85% accuracy, demonstrating the viability of data-driven safety solutions.

Critical Discovery

Identified strong relationships between road infrastructure quality and accident outcomes, highlighting the importance of safety features.

Practical Applications

Models can assist city planners in identifying high-risk zones and prioritizing infrastructure investments for maximum safety impact.

Future Directions

- Integrate real-time traffic and weather data for dynamic predictions
- Expand analysis to include driver behavior and vehicle characteristics
- Deploy models in traffic management systems for proactive intervention