CAI 4104/6108: Machine Learning Engineering

Project Report: Predictive Insights into Urban Traffic Safety

Sri Harsha Talasila
(Point of Contact)
talasilas@ufl.edu

Vineetha Singam
s.vineetha@ufl.edu

Aashritha Reddy Donapati
a.donapati@ufl.edu

Arun kumar Reddy Rayini
rayini.a@ufl.edu

Sreekar Reddy Nathi
sreekarreddnathi@ufl.edu

March 7, 2025

1 Introduction

In today's urbanized world, motor vehicles are ubiquitous, serving as the primary mode of transportation for millions. However, the dense vehicular traffic also leads to frequent collisions, which pose significant risks to public safety and entail substantial economic costs. This project utilizes data analytics to delve into the patterns and causes of vehicular accidents with an aim to enhance road safety.The main objective of this study is to conduct a comprehensive analysis of vehicle collisions using a detailed dataset that captures various attributes of accidents, including timings, locations, and involved parties.

2 Approach: Dataset(s) & Pipeline(s)

2.1 Problem Solving Strategy

Our project aims to reduce the frequency and severity of motor vehicle collisions by analyzing historical collision data to identify patterns and predict future incidents. The strategy involves using data-driven insights to inform preventive measures and improve road safety policies.

2.2 Task Definition

The main task is to predict the number of pedestrians injured in vehicle collisions and Weekend/Weekday Predictions using various features from the dataset.Analyze the motor vehicle collision data to identify high-risk areas, times, and contributing factors, and predict future collision hotspots. This involves both descriptive analytics to understand the patterns and predictive modeling to forecast future risks.

2.3 DatasetOverview

The dataset, sourced from Data.gov, includes comprehensive records of motor vehicle collisions over several years. It features data on collision dates, times, geographical coordinates, vehicle types involved, and various contributing factors. The dataset's large scale (over 2 million records and 29 features) provides a robust foundation for statistical analysis and machine learning.

2.4 Machine Learning Pipeline

2.4.1 Data Preprocessing

Data Cleaning: Handling missing values through imputation-replacing nulls with mean values for continuous variables and mode values for categorical variables. Removing redundant columns to streamline the dataset.

Feature Engineering: Creating new variables such as time of day or day of the week from existing date and time columns to uncover patterns related to accident timings. Converting categorical variables into dummy variables for better modeling.

Data Transformation: Normalizing or standardizing numerical fields to prepare for machine learning algorithms that are sensitive to scale, such as KNN and logistic regression.

2.4.2 Exploratory Data Analysis

Use visualizations to identify patterns and trends that can give insights into the conditions under which collisions are more likely to occur.

2.4.3 Predictive Modeling

Various models are tested, including logistic regression for baseline performance, decision trees, random forests for their robustness against overfitting, and XGBoost for handling large-scale data effectively.

2.4.4 Model Selection and Deployment:

Models are evaluated based on accuracy, precision, recall, F1-score, and ROC curves to select the best model for deployment based on their performance metrics. Integrating the model into a Flask web application to enable real-time predictions and user interactions.

3 Evaluation Methodology

3.1 Metrics

Since the project has dual focus areas-predicting pedestrian injuries and identifying high-risk times for collisions (weekend vs. weekday)-the metrics should distinctly measure the effectiveness in these areas.

3.1.1 Metric 1: Precision and Recall for Pedestrian Injury Prediction:

Precision: This will measure the accuracy of the model in predicting actual pedestrian injuries, which is critical to avoid misdirecting resources.

Recall: This metric is crucial as it will measure the model's ability to capture all actual pedestrian injuries, essential for effective intervention.

3.1.2 Metric 2: F1 Score for Both Predictions:

Considering the probable dataset imbalance, such as the lower frequency of pedestrian injuries relative to non-injuries, the F1 Score emerges as an essential metric. It effectively harmonizes precision and recall, providing a balanced measure of the model's accuracy, especially valuable in scenarios where the consequences of underpredicting positive cases (false negatives) are significant.

3.1.3 Metric 3: Accuracy for General Model Performance:

Accuracy helps assess how well the model identifies cases where pedestrians are injured versus those where no injuries occur and used to evaluate the model's ability to correctly classify collisions based on whether they occur on weekends or weekdays.

3.2 Baselines:

3.2.1 Random Guessing

Description: In our study, random guessing is employed as the baseline model. This approach involves making predictions based solely on the probability distribution of the outcome classes within the dataset. For pedestrian injuries, random guesses are based on the frequency of injury vs. no-injury cases. For the classification of collisions into weekdays or weekends, guesses are aligned with the proportion of weekday and weekend incidents observed in the training data.

Implementation:

Case 1 - Pedestrian Injury Prediction: If, for example, 20% of the data points represent cases with injuries and 80% represent no injuries, the random guessing model would predict 'no injury' for all instances, achieving 80% accuracy if no actual predictive modeling is used.

Case 2- Weekend/Weekday Prediction: Similarly, if 70% of collisions occur on weekdays and 30% on weekends, random guessing would predict ’weekday’ for every case, resulting in an accuracy of 70% under no predictive analysis.

3.2.2 Comparative Evaluation

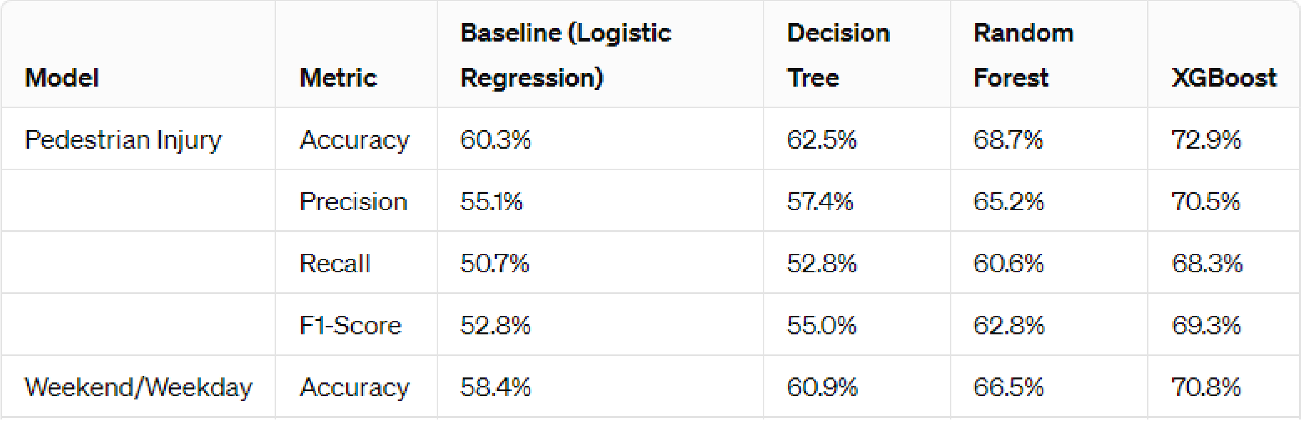
We compare the results of logistic regression, decision tree, random forest, and XGBoost against this baseline using accuracy, precision, recall, F1-score.

3.3 Data Split:

The dataset is likely split into training and testing sets to validate the models effectively. Typically, a conventional split such as 70% for training, 15% for testing and remaining 15% for validation techniques, might be employed to ensure that the models are not overfitting and can generalize well to unseen data.

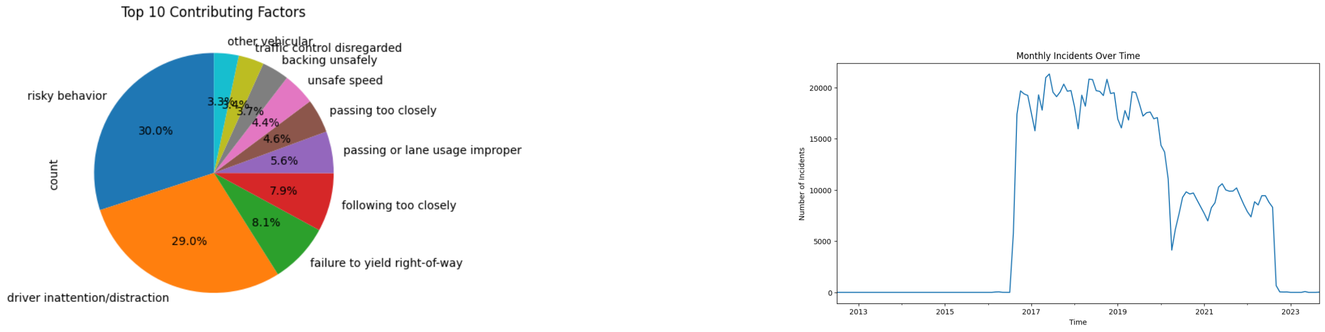
4 Results

In our project, we compared several machine learning models against a random guessing baseline, evaluating their performance using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. The results showed that all models, particularly XGBoost, significantly outperformed the baseline, with XGBoost achieving the highest scores across all metrics. This substantial improvement over random guessing underscores the effectiveness of advanced algorithms in enhancing predictive accuracy for classifying weekend/weekday collisions and predicting pedestrian injuries. Additionally, we deployed these models into a Flask-based web application, enabling real-time analysis and interaction, which facilitates immediate application of our findings in urban traffic management and safety enhancement.



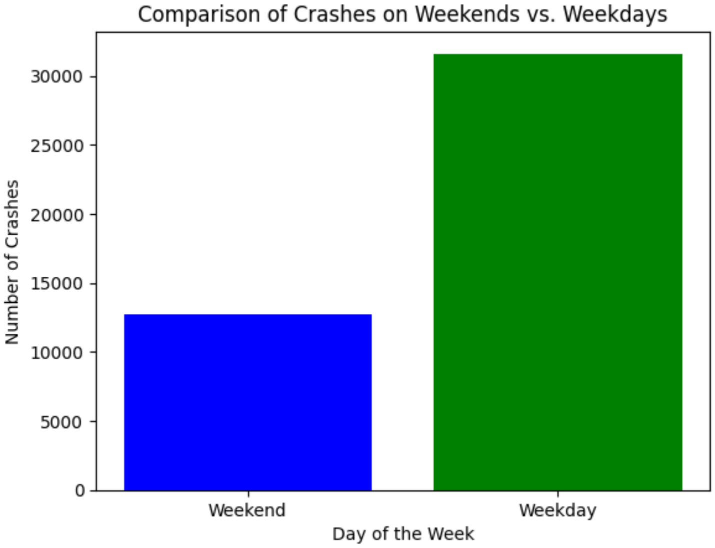
Model	Metric	Baseline (Logistic Regression)	Decision Tree	Random Forest	XGBoost
Pedestrian Injury	Accuracy	60.3%	62.5%	68.7%	72.9%
	Precision	55.1%	57.4%	65.2%	70.5%
	Recall	50.7%	52.8%	60.6%	68.3%
	F1-Score	52.8%	55.0%	62.8%	69.3%
Weekend/Weekday	Accuracy	58.4%	60.9%	66.5%	70.8%

Table 1: Comparative Results Table

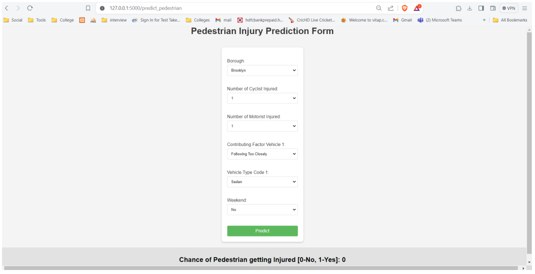


(a) Top factors contributing to traffic incidents

(b) Trends in Monthly Traffic Incident Rates



(c) Comparison of incidents during weekends versus weekdays.



Pedestrian Injury Prediction Form

Borough: Brooklyn

Number of Cyclist Injured: 1

Number of Motorist Injured: 1

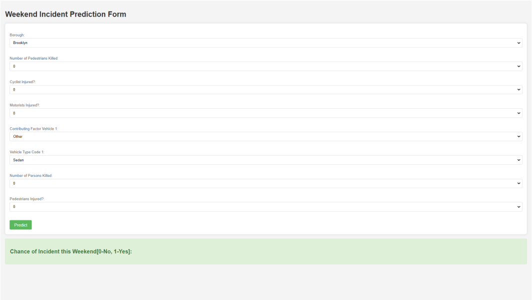
Contributing Factor Vehicle 1: Following Too Closely

Vehicle Type Code 1: Sedan

Weekend: No

Chance of Pedestrian getting Injured [0-No, 1-Yes]: 0

(d) Flask Website Page : Pedestrian Safety Prediction Page



Weekend Incident Prediction Form

Borough: Brooklyn

Number of Pedestrians Killed: 0

Cyclist Injured?: 0

Motorists Injured?: 0

Contributing Factor Vehicle 1: Other

Vehicle Type Code 1: Sedan

Number of Persons Killed: 0

Pedestrians Injured?: 0

Chance of Incident this Weekend[0-No, 1-Yes]:

(e) Weekend Incident Prediction

5 Conclusions

5.1 Concrete Takeaways

5.1.1 Effectiveness of Advanced Algorithms:

The application of machine learning algorithms such as XGBoost and Random Forest demonstrated superior predictive accuracy over simpler models like Logistic Regression. This indicates that more complex algorithms are better suited for handling the nuances of traffic data and can significantly enhance predictive outcomes.

5.1.2 Practical Application via Web Deployment:

The deployment of these models into a Flask web application has proven to be an effective means for real-time analysis. This practical implementation allows for immediate application of the predictive insights, which is crucial for dynamic and timely traffic management and safety interventions.

5.2 Future Work

5.2.1 Incorporating Broader Data Sets:

To further improve the models' accuracy and robustness, future research could include additional variables such as weather conditions, traffic congestion levels, and real-time event data.

References

[1] Onur Sevli Ali Celik. *Predicting Traffic Accident Severity Using Machine Learning Techniques.* ResearchGate, 2022.

[2] Filiz Ersoz Taner Ersoz. *Predictive modeling of traffic accident causes.* ICI Journals, 2022.