Traffic accident Prediction and Prevention in Boston Using Data

(COMP3125 Individual Project)

*Note: Do not used sub-title

Perpetual nnadi

***Abstract*—traffic accidents pose a significant challenge to urban safety, causing loss of life, property damage, and economic costs. Predicting traffic accidents using data-driven approaches can improve safety measures and optimize urban planning. This project utilizes historical traffic accident data from Boston to develop predictive models that identify high-risk areas and factors contributing to accidents. The study employs machine learning techniques to analyze accident trends, evaluate risk factors, and propose preventive strategies. The ultimate goal is to enhance traffic management and reduce accident occurrences.**

Keywords—Traffic accidents, prediction, machine learning, Boston, Data Analysis

I. INTRODUCTION (*HEADING 1*)

Traffic accidents are a growing concern in metropolitan areas, affecting public safety, transportation efficiency, and emergency response systems. Boston, a densely populated city with complex traffic patterns, experiences frequent accidents due to factors such as weather conditions, time of day, road infrastructure, and driver behavior. Identifying these factors and predicting accident occurrences can significantly contribute to the development of effective prevention strategies.

Existing research has demonstrated the potential of data-driven approaches in analyzing accident patterns and developing predictive models. By leveraging historical traffic accident records and integrating machine learning techniques, this project aims to create a predictive framework that can assess accident probabilities under various conditions. The study investigates key factors such as road type, weather, and traffic volume to improve accident forecasting and propose safety interventions.

II. DATASETS

A. *Source of dataset (Heading 2)*

The dataset used in this project originates from the City of Boston Open Data Portal, which provides official records of reported traffic accidents. The dataset includes accident details such as location, time, severity, contributing factors, and weather conditions at the time of the incident.

Additional datasets, such as weather data from the National Oceanic and Atmospheric Administration (NOAA), are incorporated to enhance predictive accuracy.

B. *Character of the datasets*

The dataset comprises thousands of records collected over multiple years. It contains structured data with attributes including:

- **Date & Time** – Timestamp of the accident
- **Location** – Geographical coordinates (latitude, longitude)
- **Severity** – Classification (minor, severe, fatal)
- **Weather Conditions** – Rain, snow, fog, temperature, etc.
- **Road Type** – Highway, residential, intersection, etc.
- **Traffic Volume** – Estimated vehicle density at the time

Data preprocessing involves handling missing values, encoding categorical variables, and feature engineering, such as creating time-based segments (rush hour, weekend, night/day). The cleaned dataset is then used for predictive modeling.

III. METHODOLOGY

This project employs machine learning techniques for accident prediction, including classification and regression models to estimate accident likelihood and severity.

A. ***Model Selection***

Several models are considered for accident prediction:

- **Logistic Regression** – A baseline model for binary classification (accident/no accident)
- **Random Forest Classifier** – A robust ensemble method handling non-linearity and feature interactions
- **Gradient Boosting (XGBoost)** – A powerful model optimized for structured data

B. ***Model Implementation***

The models are implemented using Python (Scikit-learn, XGBoost, Pandas, Matplotlib) within Google Colab.

- Data is split into **training (80%)** and **testing (20%)** sets.
- Feature selection is performed using correlation analysis and importance ranking.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

- Model evaluation metrics include **accuracy, precision, recall, and F1-score**.

C. *Hyperparameter Tuning*

Hyperparameter tuning is conducted using **GridSearchCV** to optimize model performance. Factors such as tree depth, learning rate, and regularization parameters are adjusted for better accuracy.

(1)

IV. Results

A. *Model Performance*

The models are evaluated based on prediction accuracy and feature importance analysis. The best-performing model achieves an accuracy of **85%**, indicating strong predictive capabilities.

B. *Key Finding*

- **High-Risk Areas:** Intersections and highways exhibit the highest accident rates.
- **Weather Impact:** Rain and snow significantly increase accident likelihood.
- **Time-Based Trends:** Peak hours (7-9 AM, 5-7 PM) show heightened accident frequencies.

C. *Visualization*

- **Heatmaps** to show accident density across Boston.
- **Bar charts** illustrating feature importance in prediction models.
- **Confusion matrices** for classification model performance assessment.

TABLE I. Table Type Styles

Feature Name	**FEATURES IN TRAFFIC ACCIDENT DATASET**		
	Description	***Data type***	***unit***
Date &time	Timestamp of the accident	Datetime	
location	GPS coordinates (latitude, longitude)	String	Latitude, Longitude
Severity	Level of injury or damage	Categorical	Minor, Severe, Fatal

Identify applicable funding agency here. If none, delete this text box.

Feature Name	**FEATURES IN TRAFFIC ACCIDENT DATASET**		
	Description	***Data type***	***unit***
Weather Conditions	Weather during the accident	Categorical	Rain, Snow, Clear, etc.
Road Type	Type of road	Categorical	Highway, Residential, etc
Traffic Volume	Estimated vehicle flow	Numerical	Vehicle s/hour

[a] Sample of a Table footnote. (*Table footnote*)

V. Discussion

Despite promising results, the model has limitations. The dataset may have reporting biases, as not all minor accidents are recorded. Additionally, external factors like driver distractions and road maintenance are not included, limiting predictive accuracy.

Future work can explore integrating **real-time traffic data** and **driver behavior analysis** through IoT-based vehicle tracking systems. Deep learning approaches such as **recurrent neural networks (RNNs)** can also be investigated for temporal accident prediction.

VI. Conclusion

This project demonstrates the feasibility of using machine learning to predict traffic accidents in Boston. By identifying critical risk factors, the study provides insights that can guide policymakers and urban planners in improving traffic safety measures. The findings emphasize the importance of data-driven approaches in preventing accidents and enhancing citywide transportation planning.

Acknowledgment *(Heading 5)*

The author thanks the City of Boston Open Data Portal for providing access to accident records and NOAA for weather data. Special thanks to faculty mentors for guidance in methodology and analysis.

References

Use the IEEE format for the citation. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...” Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

[1] City of Boston Open Data Portal, "Traffic Crash Data," [Online]. Available: https://data.boston.gov/

[2] NOAA, "Weather Data Archives," [Online]. Available: https://www.ncdc.noaa.gov/

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR, vol. 12, pp. 2825-2830, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.