RoadSense - Advanced Predictive Modeling for Traffic Safety

Submitted By:

1. Pranav Harish Sharma
   NEU ID: 002851959
   Email ID: sharma.pranavh@northeastern.edu
   Percentage Contribution: 50%

2. Harvineet Singh
   NEU ID: 002814713
   Email ID: singh.harvi@northeastern.edu
   Percentage Contribution: 50%

Milestone: Master's Project Literature Review and Data Source Identification

Course Details: IE7945 53267 Master's Project SEC 05 Summer Full 2024

Date: June 1, 2024

Section 1: Introduction to the Project

1.1 Project Overview

Traffic accidents are a critical global issue, leading to numerous fatalities, injuries, and significant economic losses. With the increasing complexity of urban traffic systems, understanding the factors contributing to traffic accidents and predicting their occurrence has become more challenging and essential. The advancement of machine learning provides powerful tools to analyze vast amounts of traffic-related data and generate actionable insights to enhance road safety.

This project aims to leverage machine learning techniques to analyze and predict traffic accidents using the comprehensive US-Accidents dataset. The dataset comprises over 2.25 million records of traffic accidents across the contiguous United States, including various attributes related to accident specifics, environmental conditions, and location details.

1.2 Project Goals and Objectives

1. **Data Preprocessing**: Clean and preprocess the US-Accidents dataset to handle missing values, outliers, and inconsistent data to ensure the data is suitable for machine learning models.

2. **Feature Extraction:** Identify and extract relevant features from the dataset. This involves selecting variables such as weather conditions, time of day, road type, traffic signals, and proximity to points of interest.

3. **Exploratory Data Analysis (EDA):** Perform an in-depth exploratory data analysis to understand the distribution of data, detect patterns, and visualize relationships between different variables.

4. **Model Development**: Develop and compare multiple machine learning models to predict the occurrence and severity of traffic accidents. The models to be explored include:
    a. **Random Forest**: A learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification.
    b. **Gradient Boosting Machine (GBM)**: A technique that builds models sequentially, with each new model correcting errors made by the previous ones. GBM is known for its high accuracy and ability to handle different types of data.

c. **Deep Neural Networks (DNN)**: A deep learning approach utilizing multiple layers of neurons to model complex patterns in the data. DNNs are powerful for capturing nonlinear relationships and interactions between features.
d. **Convolutional Neural Networks (CNN):** A specialized DNN architecture particularly effective for spatial data and image processing. In this project, CNNs can be applied to analyze spatial data related to accident locations and nearby infrastructure.
e. **Long Short-Term Memory Networks (LSTM)**: A type of recurrent neural network (RNN) capable of learning long-term dependencies, suitable for time-series data. LSTMs will be used to analyze temporal patterns in traffic accidents.

5. **Model Evaluation:** Evaluate the models using various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to determine their effectiveness in predicting traffic accidents.

6. **Interpretability**: Use techniques to interpret the models and understand the impact of different features on the predictions.

1.3 Problem Statement

The project aims to address the following business questions:

1. **What are the key factors contributing to high-severity traffic accidents?**

   Details: By analyzing various attributes such as weather conditions, time of day, road type, and traffic conditions, we aim to identify the key factors that significantly contribute to high-severity accidents. This understanding can help in formulating targeted interventions, such as improving road infrastructure, enhancing traffic enforcement in high-risk areas, and launching public safety campaigns focused on specific factors like speeding or drunk driving.

2. **How can real-time weather and traffic data improve the prediction accuracy of traffic accidents?**

   Details: By incorporating real-time weather and traffic data into predictive models, we aim to assess the improvement in prediction accuracy. This question will explore the impact of dynamic data on model performance and its potential in real-time accident prevention strategies. For instance, accurate predictions can enable traffic authorities to issue timely warnings, adjust traffic signals, or deploy emergency services proactively, thereby reducing accident risks.

3. **What are the spatiotemporal patterns of traffic accidents, and how can they inform traffic management policies?**

   Details: By analyzing the spatial and temporal distribution of traffic accidents, we aim to uncover patterns and trends. This analysis will provide insights into peak accident times, high-risk locations, and seasonal variations. These insights can inform traffic management policies, such as optimizing traffic light schedules, enhancing road maintenance in high-risk areas, and implementing targeted safety measures during peak accident periods (e.g., holiday seasons, rush hours).

1.4 Expected Outcomes and Impact

The expected outcomes of this project include:

1. **Accurate Predictive Models:** Development of highly accurate models capable of predicting the occurrence and severity of traffic accidents based on historical and real-time data.
2. **Actionable Insights:** Identification of key factors contributing to traffic accidents, providing actionable insights for policymakers, traffic authorities, and urban planners.
3. **Enhanced Road Safety:** Implementation of predictive models and insights to enhance road safety measures, reduce accident rates, and save lives.
4. **Scalable Solutions:** Development of scalable solutions that can be adapted and applied to different regions and traffic conditions, leveraging the robustness and flexibility of machine learning techniques.

Section 2: Annotated Bibliography

---

1. **Title: A Countrywide Traffic Accident Dataset**
   - **Authors**: Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Rajiv Ramnath.
   - **Publication Year:** 2019.
   - **Citations:** Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409.
   - **Key Insights:** This paper introduces the US-Accidents dataset, comprising 2.25 million traffic accident records in the contiguous United States. It includes detailed attributes like location, time, weather conditions, and points-of-interest, enriching traffic accident analysis. The dataset's breadth and depth are invaluable for our project, facilitating feature extraction and model training essential for accurate accident prediction.
   - **Relevance:** This dataset is critical for our project as it offers extensive data necessary for analyzing and predicting traffic accidents.
   - **URL:** https://arxiv.org/abs/1906.05409

2. **Title: Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.**
   - **Authors:** Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, Rajiv Ramnath.
   - **Publication Year:** 2019.
   - **Citations:** Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 33-42).
   - **Key Insights:** The authors introduce DAP, a deep neural network model for real-time accident prediction, leveraging sparse data. DAP integrates recurrent and fully connected components, capturing temporal and spatial features, validated on the US-Accidents dataset, showing improved rare event prediction and emphasizing attribute impact on accuracy.
   - **Relevance**: The paper offers insights into applying deep learning for accident prediction, with directly applicable methodology and architecture, informing our project's development and evaluation.
   - **URL:** https://arxiv.org/abs/1909.09638

3. **Title: Exploring the relationship between alcohol and the driver characteristics in motor vehicle accidents.**
   - **Authors:** Mohamed A. Abdel-Aty, Hassan T. Abdelwahab
   - **Publication Year**: 2020
   - **Citations**: Abdel-Aty MA, Abdelwahab HT. Exploring the relationship between alcohol and the driver characteristics in motor vehicle accidents. Accid Anal Prev. 2000 Jul;32(4):473-82. doi: 10.1016/s0001-4575(99)00062-7. PMID: 10868750.
   - **Key Insights:** The study explores the link between alcohol involvement and driver demographics in Florida accidents, identifying high-risk groups using statistics and models. Younger drivers, especially aged 25-34, exhibit the highest rates of alcohol/drug involvement, prompting the need for tailored education and awareness initiatives.
   - **Relevance:** Understanding how driver characteristics influence accident involvement is crucial for our traffic accident data analysis.
   - **URL**: https://www.sciencedirect.com/science/article/pii/S0001457599000627

4. **Title: Highway crash detection and risk estimation using deep learning.**
   - **Authors**: Tingting Huang, Shuo Wang, Anuj Sharma.
   - **Publication Year**: 2021.
   - **Citations**: Huang T, Wang S, Sharma A. Highway crash detection and risk estimation using deep learning. Accid Anal Prev. 2020 Feb;135:105392. doi: 10.1016/j.aap.2019.105392. Epub 2019 Dec 13. PMID: 31841865.
   - **Key Insights**: The paper explores deep learning for highway crash detection and risk estimation, using real-time data from Interstate 235 in Des Moines, IA, with CNNs and LSTMs showing high accuracy. It emphasizes real-time data importance and deep learning potential for traffic safety.
   - **Relevance**: This research provides a practical approach to using deep learning models for real-time crash detection and risk estimation. The techniques and findings are relevant to our project, especially in developing models that incorporate real-time data for enhanced prediction accuracy.
   - **URL:** https://www.sciencedirect.com/science/article/pii/S000145751930555X

5. **Title: Traffic Accident Analysis Using Machine Learning Paradigms.**
   - **Authors:** Miao Chong, Ajith Abraham, Marcin Paprzycki
   - **Publication Year**: 2004.
   - **Citations**: Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. Informatica, 29(1).
   - **Key Insights**: The paper investigates four machine learning paradigms for modeling injury severity in traffic accidents: neural networks with hybrid learning, support vector machines, decision trees, and a hybrid decision tree-

neural network model. Experiment results show the hybrid model's superiority over individual approaches.

- **Relevance**: This research offers valuable insights into machine learning techniques for traffic accident analysis, guiding model selection, feature engineering, and evaluation strategies to enhance predictive accuracy.
- **URL**: https://informatica.si/index.php/informatica/article/viewFile/21/15

6. **Title: Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique.**
   - **Authors:** Omar Ibrahim Aboulola, Abel C. H. Chen (Editor).
   - **Publication Year**: 2024.
   - **Citations**: Aboulola OI. Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique. PLoS One. 2024 Apr 9;19(4):e0300640. doi: 10.1371/journal.pone.0300640. PMID: 38593130; PMCID: PMC11003624.
   - **Key Insights**: The study aims to develop predictive models for injury severity prediction in traffic accidents using various transfer learning techniques. Models like Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Residual Networks (ResNet), EfficientNetB4, InceptionV3, Extreme Inception (Xception), and MobileNet were utilized, with MobileNet achieving the highest accuracy of 98.17%. Additionally, Shapley values are employed to discern the most influential factors affecting accident prediction models.
   - **Relevance**: T This research tackles the necessity for interpretable predictive models to grasp the factors behind traffic accidents. By unraveling the complexities of machine learning and deep learning models, it aims to create more reliable models, informing interventions to prevent accidents effectively.
   - **URL:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11003624/

7. **Title: Evaluating the Impact of Weather Conditions on Traffic Accidents in Urban Areas.**
   - **Authors:** Qiang Zeng, Wei Hao, Jaeyoung Lee, Feng Chen
   - **Publication Year**: 2020.
   - **Citations**: Zeng Q, Hao W, Lee J, Chen F. Investigating the Impacts of Real-Time Weather Conditions on Freeway Crash Severity: A Bayesian Spatial Analysis. International Journal of Environmental Research and Public Health. 2020; 17(8):2768. https://doi.org/10.3390/ijerph17082768
   - **Key Insights**: The study explores real-time weather impacts on freeway crash severity using a Bayesian spatial model with data from 1424 crash records. It

incorporates hourly weather variables and other factors, revealing correlations between precipitation and crash severity.

- **Relevance**: The study enhances understanding of real-time weather impacts on freeway crashes, guiding strategies for mitigating severity, especially during rain. Its Bayesian spatial analysis framework offers robust modeling, informing engineering countermeasures effectively.
- **URL**: https://www.mdpi.com/1660-4601/17/8/2768

Section 3: Data Source Identification

Dataset: US-Accidents Dataset (2016-2023)

Source: Kaggle, Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Rajiv Ramnath.

Description: This extensive dataset comprises countrywide traffic accident records spanning 49 states of the United States. Collected ranges from February 2016 - 2023, the data sources include diverse APIs offering real-time traffic event data from entities like transportation departments, law enforcement agencies, and traffic cameras.

Dataset Size: 500,000 Records x 46 Features Columns

Attributes: Accident ID, timestamp, location details (latitude, longitude, street, city, state, county, country), weather conditions (temperature, humidity, pressure, visibility, wind speed, wind direction, precipitation, weather condition), traffic infrastructure (amenity, bump, crossing, give way, junction, no exit, railway, roundabout, station, stop, traffic calming, traffic signal, turning loop), lighting conditions (sunrise/sunset, civil twilight, nautical twilight, astronomical twilight), and accident description.

Availability: Publicly available.

URL: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download, https://smoosavi.org/datasets/us_accidents

GitHub Link: https://github.com/Pranavsharma13/RoadSense-Advanced-Predictive-Modeling-for-Traffic-Safety