Data Intensive Computing Project Phase 4 Report

Project: Eco-Friendly but is it safe? Analysis and Predicting Trends in E-vehicle Accidents

Group members:

Abhijeet Sanjiv Bonde (50624352)

Raghav Potdar (50631527)

Hui Yu (50311663)

Date: 31 April 2025

Data Sources:

The data used for this project is the NYC Motor Vehicle Collisions (Crashes) dataset, which is available for free of charge on the NYC Open Data Portal. The New York City Police Department (NYPD) keeps this data record which has extensive records of motor vehicle crashes that have been reported in New York City. The timeframe for the dataset is from June 2012 to Feb 2025 (Updated on a regular basis).

Dataset Overview

- Source: NYC Open Data (NYPD)
- URL: NYC Motor Vehicle Collisions Dataset
- Number of Records: Over 2 million crash reports (updated regularly)
- Number of Columns: 29+ features that cover various aspects of each collision

Problem Statement

With the increasing adoption of E-scooters/E-bikes as a convenient and eco-friendly mode of transportation, concerns regarding their safety have emerged. This project aims to analyze and predict accident trends using historical motor vehicle collision data. Specifically, we will investigate the severity of E-scooter/E-bikes accidents compared to other vehicle types, identify key contributing factors, and determine high-risk locations and periods for accidents.

Codebase Instructions

1. Setup Python Environment

Step 1: Create and Activate Virtual Environment

```
python -m venv ev_accident_env
```

```
.\ev_accident_env\Scripts\activate
```

Step 2: Install Required Dependencies

Install all necessary Python libraries using the provided requirements.txt file:

```
pip install -r requirements.txt
```

2. Generate the Model (.pkl file)

Step 1: Running the Jupyter Notebook

Open and run the notebook file named EV Accident Analysis.ipynb.

Execute each cell sequentially. Ensure the cell that saves the Logistic Regression model (.pkl) file executes correctly.

Upon successful completion, the notebook will generate a serialized model file

Here, we are not running the code files for phase 1 for data cleaning and pre-processing as our code uses external API (Nominatim) to fill in information related to address and zip codes and it took us 5-6 days to run. So we are using the final CSV generated from phase 1 to the modelling part.

3. Launching the Streamlit Application

Step 1: Verify Directory Structure

Your directory should look like this:

```
project_folder/
├── app.py
├── LR_model.pkl
├── ev_heatmap.html
└── requirements.txt
```

Step 2: Start the Streamlit Server

Run the Streamlit app from your terminal or command prompt:

```
streamlit run app.py
```

The application will start running locally and open automatically in your default browser.

If it does not open automatically, navigate to the URL provided in your terminal (usually http://localhost:8501).

User Instruction on EV Accident Severity Prediction App

This application predicts the likelihood of casualties in electric vehicle (EV) accidents based on user-provided inputs. It was trained on historical accident data from New York State and supports decision-making by identifying high-risk scenarios.

Step-by-Step Guide

1. Enter Temporal Information

- **Select Date**: Use the calendar picker to choose any date. While this won't affect predictions directly (as the model is not time-sensitive to future dates), it contextualizes the input.
- **Hour of Day**: Adjust the slider (0 to 23) to represent the hour when the accident occurred or is being evaluated.

2. Provide Vehicle and Location Details

- **Vehicle Type**: Select the type of vehicle involved in the collision with a small electric vehicle. This input refers to the *non-EV vehicle* in the incident. The model uses this to assess the severity of accidents where small EVs collide with different vehicle classes.
- **Contributing Factor**: Choose the most relevant cause of the accident (e.g., “driver inattention/distraction,” “unsafe speed,” etc.).
- **ZIP Code**: Enter a valid New York State ZIP code where the incident took place. The app validates this field to ensure it falls within the geographic range supported by the model.

3. Run the Prediction

Click "Predict Accident Severity" to generate results based on your inputs.

Understanding the Results

- **Prediction Outcome**: Displays whether a casualty is likely (e.g., “Casualty Likely”).
- **Probability**: Shows the model’s confidence as a percentage (e.g., 74.12%).
- **Risk Factors**: Lists any significant high-risk input values, such as a specific contributing factor or time of day.

About the Model

- Provides background information on the model's data sources, features used, performance metric, and the geographic scope of its predictions

Interactive Feature

The app also includes an interactive heatmap that visualizes small EV accident hotspots across New York City. Users can explore spatial accident patterns, identify high-risk neighborhoods, and better understand where infrastructure improvements may be needed.

Modeling

In the second phase of the project, we experimented with both unsupervised and supervised modeling techniques to understand accident severity patterns. K-Means clustering (along with DBSCAN and hierarchical clustering) was applied to the E-scooter/E-bike accident subset to identify natural groupings of accidents. This revealed distinct clusters primarily differentiated by time and context of crashes. For example, one cluster corresponded to night-time accidents (mostly non-rush hours) which had a higher average casualty count (~0.86 injuries/deaths per accident) compared to daytime clusters (~0.80–0.84). Another cluster consisted of weekend accidents (mostly daytime) with slightly elevated casualty rates. These findings suggested that time of occurrence (night vs. day, weekday vs. weekend) is a significant factor in accident severity. Such insights informed feature engineering (e.g.

creation of boolean features like IsNightTime, IsWeekend, and IsRushHour) for the predictive models.

Supervised learning in phase 2 began with separate models to predict injuries and fatalities. We initially treated the number of injuries and the number of deaths as two targets. However, fatalities were extremely rare, leading to models that trivially predicted zero fatalities for almost all cases (yielding an artificially high accuracy but no practical value). The injury prediction model (a Logistic Regression) did identify some patterns, but it suffered from the class imbalance problem: a large majority of E-vehicle accidents resulted in at least one injury. About 76% of E-scooter/E-bike accidents involved an injury or fatality, compared to only ~24% with no casualties. This imbalance caused early classifiers to over-predict the majority class (injury) – for instance, an initial logistic model achieved about 76% accuracy by simply labeling almost every case as "injury", but it failed to detect the non-injury cases (recall for the no-injury class was near 0%). These preliminary results highlighted the need to combine the targets and address imbalance.

Phase 4: Modeling workflow and model selection

Data Preparation & Feature Engineering: We limited the dataset to electric vehicle accidents (18,531 records). Categorical variables like borough, zip code, and contributing factors were one-hot encoded, and the time-of-day indicators (IsNightTime, IsWeekend, IsRushHour) were added to capture the temporal risk patterns observed in clustering. The outcome CAUSALITY was defined (1 if any injury or death occurred, 0 if none). We found a severe class imbalance (approximately 76% of E-vehicle crashes had ≥1 casualty). To mitigate this, we have tried different oversampling and undersampling techniques and finally selected ADASYN oversampling on the training data, which synthetically increased the minority class (no-casualty accidents) and yielded a roughly balanced class distribution for model training.

Model Training & Hyperparameter Tuning: We evaluated a broad range of classification algorithms, including K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Extra Trees (ET), AdaBoost (ADA), Naive Bayes (NB), and Logistic Regression (LR). For each model type, we performed hyperparameter optimization using Optuna to maximize the performance metrics.

Model Evaluation & Selection: We compared models primarily by F1 score on the test data (with accuracy and recall as secondary metrics given the imbalance considerations). Logistic Regression emerged as the top performer with the highest F1 score. It achieved an F1 ~0.61, outperforming the next best model (Extra Trees, F1 ~0.58) as well as KNN, Random Forest, and others (all in the mid-0.50s F1 range). The LR model's accuracy was ~59% and recall ~59% (on the casualty class), which, while not high in absolute terms, was the best balance attained. We have also checked the balanced accuracy (For Logistic Regression) which was approximately 51%. It is still on the lower side but it's working sufficiently for our product.

Insights, Impact, and Future Work

We chose New York City as the focus of our project because it has one of the highest concentrations of small electric vehicle (EV) traffic in the United States, making it an ideal case for analyzing accident trends and proposing safety reforms. Our goal is to transform raw collision data into meaningful insights that support informed decision-making. In this section, we highlight what our target user, the New York City Department of Transportation (NYCDOT), can learn from our model, how it helps address EV safety concerns, and possible future directions to improve and extend our work.

What Can NYCDOT Learn from Using Our Product?

Our predictive model enables the New York City Department of Transportation (NYCDOT) to pinpoint high-risk areas at a granular level, including individual streets with elevated accident rates involving E-scooters and E-bikes. By visualizing these trends spatially and temporally, NYCDOT can gain a deeper understanding of when and where accidents are most likely to occur. This empowers the agency to prioritize safety interventions in the most impacted zones, ensuring efficient resource allocation and targeted policy development.

How Does It Help NYCDOT Address the Problem?

By identifying accident hotspots and analyzing contributing factors, our product equips NYCDOT with actionable insights to enhance road safety for small electric vehicles. The model highlights patterns in accident frequency, severity, and location, allowing for targeted reforms such as infrastructure redesign, safety enforcement, and public awareness campaigns.

For instance, our analysis shows that Downtown Manhattan consistently experiences a significantly higher rate of E-scooter and E-bike accidents compared to areas like Staten Island. This suggests that NYCDOT should prioritize safety interventions in higher-density, higher-risk zones such as Downtown Manhattan—where heavy traffic, limited lane space, and frequent pedestrian interactions contribute to accident risk—before expanding to lower-risk areas.

This data-driven approach ensures that policy changes and infrastructure investments are focused where they are needed most, ultimately improving public safety and reducing accident rates across the city.

Future Extensions and Avenues for Exploration

While our current model relies primarily on historical motor vehicle collision data, its predictive accuracy and scope can be significantly enhanced by incorporating additional contextual factors. In future iterations, we plan to integrate:

- Historical data on road conditions (e.g., potholes, construction zones, surface quality)
- Weather patterns (e.g., rain, snow, visibility, wind speed)

- Traffic density (e.g., congestion levels, peak hour trends)

These additions will allow for a more comprehensive understanding of the external conditions that contribute to E-scooter and E-bike accidents.

Beyond data enrichment, we also aim to scale our analysis beyond New York City. By including other major cities across the United States—and international regions such as China, where small EV adoption is highest—we can explore broader safety patterns, assess the effectiveness of differing policies, and develop more globally informed recommendations.