Predicting Traffic Accident Severity and Vehicle Count using Machine Learning

Surabhi Parab
MS in Electrical Engineering
University of North Carolina at Charlotte
sparab2@uncc.edu

Radhika Patel
MS in Electrical Engineering
University of North Carolina at Charlotte
rpate230@uncc.edu

Madhu Kiran
MS in Computer Engineering
University of North Carolina at Charlotte
mchiti@uncc.edu

***Abstract*— Traffic accidents are a major cause of injuries and fatalities across the United States, and identifying factors that contribute to accidents is crucial for improving road safety. In this project, we propose to analyze the US Accidents dataset, which contains detailed information about traffic accidents across the country. We plan to develop machine learning models that can predict accident severity and the number of vehicles involved in an accident based on factors such as weather conditions, road features, and time of day. We will also use visualization techniques to identify patterns and correlations in the data and gain insights into the factors that contribute to accidents. The proposed project aims to contribute to the development of targeted interventions to reduce the frequency and severity of traffic accidents and improve road safety across the United States.**

Keywords—traffic accidents, machine learning, regression, dataset, Kaggle.

I. Introduction

The US Accidents dataset contains information on traffic accidents that occurred in the United States between 2016 and 2020. The dataset includes features such as location, weather conditions, and severity of the accident. The aim of this project is to build a binary classification model that can predict whether an accident is likely to result in a severe or minor injury based on the available features in the dataset. In addition, a regression model will be trained to predict the number of vehicles involved in an accident.

The project involves data exploration, preprocessing, and visualization, as well as model building and evaluation using performance metrics such as accuracy, precision, recall, and mean squared error. The results of the project can be used to inform policymakers and traffic safety experts about the factors that contribute to severe accidents and the effectiveness of measures taken to reduce their frequency.

II. Details Of The Project

A. Dataset

This project will use the dataset [2] sourced from Kaggle [1]. Here are some of the features in the Dataset that we will be using:

- Severity: Indicates the severity of the accident, ranging from 1 (minor) to 4 (fatal).
- Start Time: The date and time when the accident occurred.
- End Time: The date and time when the accident was cleared.
- Start Latitude and Start Longitude: The latitude and longitude of the location where the accident occurred.
- End Latitude and End Longitude: The latitude and longitude of the location where the accident was cleared.
- Distance (mi): The length of the road segment where the accident occurred.
- Number: The street number of the location where the accident occurred.
- Side: Indicates whether the accident occurred on the left or right side of the street.
- Zipcode: The zipcode of the location where the accident occurred.

B. Model Training

The training plan for this project will involve dividing the dataset into training and testing sets, with the majority of the data used for training and a smaller portion reserved for testing the model's performance. The training plan for this project will involve the Kaggle dataset for training and testing the model's performance. Machine Learning concepts like binary classifier, linear regression or logistic regression will be trained on the dataset, and their performance will be evaluated based on metrics such as mean squared error and accuracy.

C. Evaluation Plan

The evaluation plan will involve comparing the performance of the different models, and selecting the one with the highest accuracy and lowest error. The selected model will then be used to make predictions on unseen data to evaluate its real-world performance. The model's accuracy in predicting the traffic from the dataset will be evaluated using metrics such as Mean

Squared Error, Root Mean Squared Error, and Mean Absolute Error.

D. Achievable Outcomes

Following could be the outcomes:

- A binary classification model that accurately predicts the severity of an accident based on the available features in the dataset. The model can be used to identify the factors that contribute to severe accidents and inform the design of targeted interventions to reduce their frequency.
- A regression model that accurately predicts the number of vehicles involved in an accident. This model can be used to provide early warning of potential traffic congestion and help improve response times to accidents.

III. Individual Responsibility

A. Member One

This Member, Madhu Kiran, will be responsible for Dataset Extraction and Cleaning. This includes collecting and sourcing data from Kaggle, cleaning and preprocessing data to ensure it is ready for analysis and feature engineering to derive new insights from the data. It will involve subtasks like removing any irrelevant features, normalizing the remaining features, and splitting the data into training and testing sets. This member would also be evaluating multiple machine learning regressions models such as Linear Regression, and Logistic Regression and comparing the same.

B. Member Two

This Member, Surabhi, will be the project manager. At the beginning, this Member will conduct a brief literature review of articles [2][3] related to the same topic in order to gain a better understanding of what needs to be implemented for the project. This person will be is responsible for Model Building and Optimization. This includes building and optimizing machine learning models, feature selection and parameter tuning to improve model performance. It also includes utilizing different algorithms, evaluating the tuned model on the testing set to obtain the final performance metrics.

C. Member Three

This Member, Radhika, will be responsible for Model Evaluation and Visualization. This Member will be helping Member One with evaluating multiple machine learning regressions models such as Linear Regression, and Logistic Regression and comparing the same. This includes evaluating model performance using metrics like Mean Squared Error, Accuracy, and Precision, as well as visualizing results using Histograms and other bar graphs. This person can also work on the Documentation, summarizing the project and its outcomes.

References

[1] https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

[3] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "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, ACM, 2019.