Predicting Future Traffic Accident Severity using A Countrywide Traffic Accident Dataset [1] [2]

郭立晨 F94081076
盧湧恩 F74109040
陳均哲 F74109032

1. Introduction

1.1. Motivation

Traffic accidents pose significant risks to pedestrians and infrastructure, consuming a considerable amount of manpower and financial resources each year to address the issue. Therefore, through the analysis of various environmental conditions of long-term traffic accident events, including weather factors, geographical environment, traffic signals, and important infrastructure locations, we aim to predict the severity of accident base on multiple environment variables. This information can assist relevant law enforcement agencies in optimizing the allocation of labour for better deployment strategies.

1.2. Dataset

This dataset provides continuous records of traffic accidents covering all 49 states of the United States from February 2016 to March 2023. The data is sourced from various data providers and APIs, including traffic event data from different entities such as transportation departments, law enforcement agencies, traffic cameras, and sensors. Currently, the entire dataset contains approximately 1.5 million accident records. The dataset includes parameters related to the severity classification of accidents, occurrence time, coordinates, city names, temperature, humidity, nearby traffic facilities, and other factors associated with accident incidents.

1.3. Difficulty

There are two main challenges for this experiment. First, the dataset provider mentioned difficulties with network connectivity, resulting in the inability to guarantee recording all data for every accident. Moreover, missing values will not be updated. Therefore, before using this dataset for training, a thorough examination of the data is necessary. Handling missing values for each kind of field will be a significant challenge in the preprocessing stage. The second challenge is that the current dataset provides 47 parameters. Considering all parameters simultaneously may introduce a lot of unnecessary noise. Hence, during the training process, it's crucial to devise a method to eliminate parameters that do not significantly impact accident occurrences.

2. System Framework

2.1. Overview

The experiment will begin with an initial statistical analysis of the dataset, including calculating the mean, median, and standard deviation. Data will also be visualized to see how it's distributed. For missing values, different imputation techniques will be utilized. To prevent biases in the training results due to extremely large or small values in numerical data, normalization and standardization will be applied. Outliers will be removed if necessary to address skewness. After data preprocessing, we intend to use Regression Analysis methods such as linear regression or logistic regression to examine the correlation between accidents and factors like weather, road conditions, and time. This allows the elimination of accident factors with weak correlations and serves as a basis for feature importance analysis. Next, the chosen accident factors will be used to train a classification model. The model' s output will categorize accident severity into four levels, ranging from minor (short delay) to severe (long delay). Finally, model performance will be evaluated using metrics such as Precision, Recall, F1 score, etc, to evaluate model performance.

2.2. Dealing with Missing Values

The dataset has 22 columns with missing values, including numerical and categorical data. We addressed this by dropping columns with excessive missing values and those containing irrelevant information, such as data source and destination. Four similar columns related to twilight were merged into one for simplicity. In addition, we introduced new time-related columns to enhance the dataset. One such column is "elapsed time", derived by computing the difference between starting and ending times, as we anticipated its potential significance. Moreover, we transformed the start time into six distinct columns: Year, Month, Day,

Weekday, Hour, Minute. Simultaneously, the original start time and end time columns were eliminated. We organized the dataset by county and sorted it by timestamps. Numerical missing values were filled using interpolation, excluding rows without specific features within the same county beforehand. After filling numerical missing values, we simplified the wind direction column by reducing the number of classifications. Finally, rows with categorical missing values were dropped. These steps reduced the number of columns from 46 to 35 and the number of rows from 7,728,394 to 7,484,058.

2.3. Data Visualization

Upon conducting data visualization based on the dataset, several noteworthy patterns emerged. Notably, weather conditions categorized as Fair, Mostly Cloudy, Clear, and Cloudy were found to be the most prevalent during accidents. The data also revealed a distinct trend, with accidents of severity 2 constituting nearly 80 percent of occurrences, while accidents of severity 1 exhibited the lowest frequency. Accidents categorized as severity 4 typically entail the longer elapsed time, while those classified as severity 1 tend to have the shorter duration. Examining the temporal aspect from 2016 to 2022, accidents of severity 2 and severity level 4 demonstrated a consistent upward trajectory, whereas severity level 3 witnessed an increase from 2016 to 2018 followed by a subsequent decline. Noteworthy variations in accident frequency were observed across months, with the highest occurrences transpiring between months 10 to 12, and the lowest in months 3 and 7. Additionally, accidents of severity 1 were notably infrequent, occurring primarily between months 3 and 8. In terms of weekdays, a distinct trend is evident, with weekdays 0 to 5 consistently exhibiting a higher frequency of accidents. Further insights into temporal patterns revealed a higher likelihood of accidents during the day compared to nighttime, given the typically reduced traffic during nighttime hours. Specifically, accidents were more prone to happen during hours 6 to 8 and 14 to 17. These observations contribute to a comprehensive understanding of the dataset's temporal and severity dynamics.

2.4. Normalization and Standarlization

Initially, we conducted an outlier removal process by excluding instances where the distance column exceeded a threshold of 10. The number of rows is reduced to 7450438 after that. Subsequently, for categorical columns, we employed label encoding to convert them into numerical representations. Also, we transfromed boolean values into 0 and 1. Following this, standard scaling was applied to normalize all the columns, ensuring a consistent and comparable scale for the dataset.

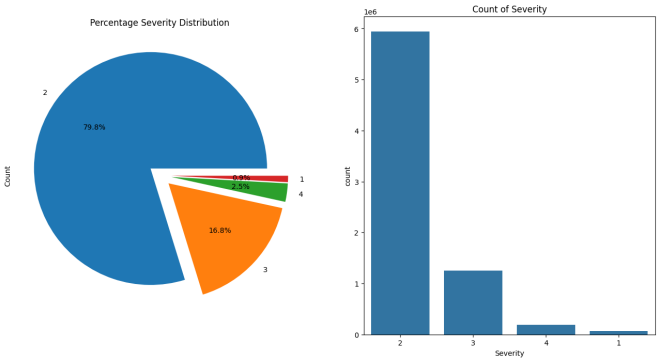

Figure 1. Severity percentage

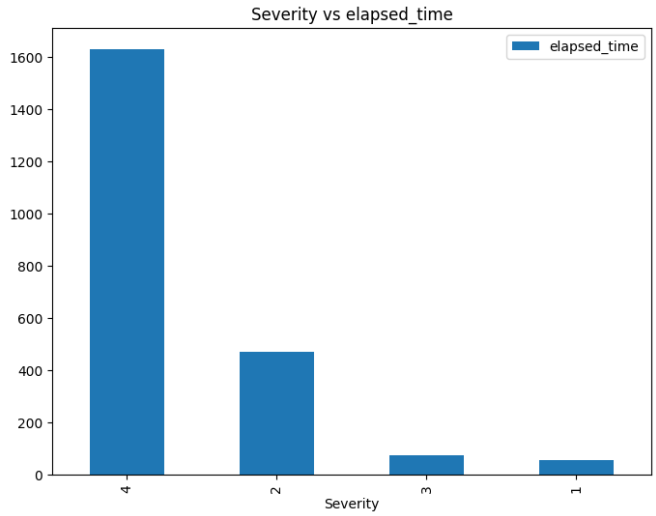

Figure 2. Severity vs elapsed time

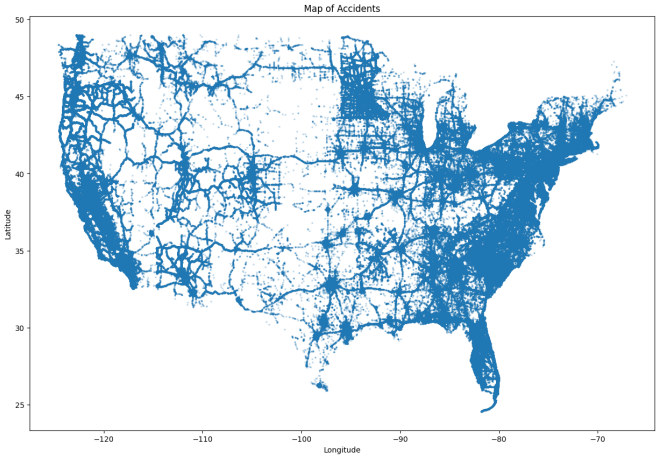

Figure 3. Map of Accidents

3. Expected Results

We anticipate that the experiment will identify several significant factors highly correlated with the occurrence of accidents. Building upon this knowledge, we aim to train a predictive model capable of accurately forecasting accident severity. In the future, real-time monitoring of various environmental factors will provide us with insights into the potential distribution of accident severity, aiding proactive measures and interventions.

References

[1] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A Country-wide Traffic Accident Dataset. *arXiv e-prints*, page arXiv:1906.05409, June 2019. 1

[2] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. *arXiv e-prints*, page arXiv:1909.09638, Sept. 2019. 1