Enhancing Urban Road Safety: A Data-Driven Approach for Predicting and Preventing Traffic Collisions in Chicago

Rutvi Shah[1], Tithi Dangarwala[1], Ravi Makwana[1] and Dr. Samir Patel[1]

[1] Department of Computer Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

Abstract. Using big data analytics to its full potential has become a crucial instrument for improving road safety in the field of intelligent transportation systems. Using a large dataset of past traffic collisions, this research presents a novel method for predicting traffic collisions. Our algorithm finds patterns and trends in the data through advanced data analysis techniques like Random Forest. This allows us to accurately anticipate collision-prone zones.

Our system detects regions that are vulnerable to traffic crashes by analysing data from various road segments. This enables the execution of preventive and mitigating actions to prevent and reduce traffic collisions. For transportation authorities and urban planners, this predictive modelling methodology provides priceless insights that enable targeted interventions and the best possible distribution of resources to high-risk locations.

Our study represents a breakthrough in the use of big data to improve traffic control tactics and road safety. We create the foundation for a safer and more effective transportation network by combining data analytics with intelligent transport systems, which eventually promotes community well-being and the development of smart cities.

Keywords: Big Data Analytics, Traffic Collision Prediction, Random Forest, Intelligent Transportation Systems, Road Safety, Urban Planning, Chicago

1 Introduction

In today's world, traffic congestion and auto accidents are major problems that have a big influence on public safety as well as economic productivity. These problems become more pressing as urbanization keeps growing, calling for creative solutions. Currently, traffic-related issues can be effectively addressed by incorporating Machine Learning (ML) into transportation engineering, especially when it comes to traffic collision prediction.

Moreover, the disturbing data regarding traffic collisions emphasize how urgent it is to develop reliable prediction systems to reduce the frequency and severity of these incidents. The World Health Organization (WHO) states that road accidents place a significant strain on healthcare systems and economy by significantly increasing global mortality and injury rates. Proactive actions and creative solutions are required to reduce the incidence and severity of these collisions.

To forecast traffic patterns and identify probable collision hotspots, this study sets out to conduct a thorough analysis of traffic collision prediction systems. It does this by utilizing sophisticated machine learning methods. Our goal is to provide stakeholders and users with timely insights to improve traffic management strategies and road safety measures by utilizing data analytics and predictive modelling.

The issue of traffic flow prediction has complex non-linear spatio-temporal dependencies and dependencies on external factors such as weekends, holidays, weather, and more.

This study delves into diverse approaches for traffic flow prediction. Additionally, it examines the influence of above-mentioned external factors such as weather conditions and road conditions on prediction accuracy, thus highlighting the pressing need for adaptable prediction frameworks.

Essentially, the goal of this study is to contribute to traffic safety and by provide methods and insights that support the development of evidence-based policies and decision-making in the field of transportation management.

2 Literature Survey

Traffic collision prediction stands as a pivotal attempt in advancing road safety and implementing pre-emptive measures to reduce collisions. Recent years have witnessed the rise of machine learning algorithms as potent instruments for forecasting traffic collisions grounded in historical data. This literature review delves into the utilization of Random Forest, KNN, and SVM algorithms in traffic collision prediction.

2.1 Random Forest:

The random forest model is a form of ensemble learning where predictions are generated by aggregating decisions from a series of individual base models. Mathematically, it can be represented as in equation (1):

$$g(x)=f0(x)+f1(x)+f2(x)+... \quad (1)$$

where the final model g is the sum of simple base models fi

In random forests, these base models are decision trees constructed independently using distinct subsets of the data. This approach, known as model ensembling, aims to improve predictive accuracy by leveraging the diversity of multiple models.

It constructs multiple decision trees and amalgamates their predictions to increase accuracy. In traffic collision prediction, Random Forest receives widespread praise owing to its adeptness in managing high-dimensional data and discerning intricate interrelations among features.

Research by Miaomiao Yan in [1] harnessed Random Forest to prognosticate collision severity by considering factors such as weather conditions, road geometry, and vehicle attributes. The study underscored Random Forest's efficacy in precisely categorizing collision severity levels, thus facilitating targeted interventions and resource allocation.

Likewise, Jianjun Yang in [2] employed Random Forest to forecast traffic collisions predicated on features derived from historical collision data. By incorporating temporal elements like time of day and day of the week alongside traffic volume, the model achieved commendable accuracy in pinpointing collision-prone zones and hotspots.

2.2 KNN (K-Nearest Neighbors):

KNN, is simple yet effective algorithm, sorts data points by looking at what most of their closest neighbors are doing. In predicting traffic collisions, KNN is used to spot common patterns and clusters of collisions in specific areas on maps.

Daniel Santos in [5] utilized KNN to identify traffic collision hotspots by analysing collision data and road network attributes. The study showcased KNN's proficiency in detecting clusters of collisions and prioritizing areas for targeted safety interventions.

Similarly, Darcin Akin in [6] applied KNN to prognosticate collision severity, considering factors like road conditions, weather, and traffic flow.

2.3 SVM (Support Vector Machine):

SVM, a robust classification algorithm, endeavours to discern the hyperplane that best segregates data points into distinct classes.

Xubin Sun in [7] employed SVM to forecast traffic collisions utilizing factors such as road geometry, traffic volume, and historical collision data. The study underscored SVM's efficacy in precisely classifying collision - prone zones and identifying high-risk areas for targeted safety interventions.

Likewise, Jianli Xiao in [8] utilized SVM to predict collision severity levels by considering features extracted from historical collision data. By accounting for spatial and temporal relationships between collision locations and environmental factors, the model attained commendable accuracy in classifying collision severity levels and guiding decision-making processes.

3 Methodology:

3.1 Data Collection and Preprocessing

The dataset used in this study consists of traffic collision data within the City of Chicago limits and under the jurisdiction of the Chicago Police Department (CPD).

Data preprocessing involved steps like handling missing values, standardizing date formats, and filtering irrelevant columns. Additionally, two new columns were engineered: "Number of Accidents" to capture the frequency of collisions on each day for each street and the "SAFE" column, representing the target variable indicating whether a particular day is deemed safe or not based on the occurrence of collisions, and conditions for that day.

3.2 Feature Selection and Engineering

Features relevant to the prediction task were selected based on their potential impact on traffic collisions. These features include street name, weather condition, lighting condition, roadway surface condition, collision day of the week, and collision month.

Feature engineering techniques were employed to enhance the predictive power of the model. This involved encoding categorical variables, such as street name and weather condition, to numerical values using StringIndexer. Additionally, two new columns were manufactured to provide additional insights into collision frequency and safety.

3.3 Model Selection

The Random Forest algorithm was chosen for the predictive modelling task due to its ability to handle high-dimensional data, discern complex relationships between features, and provide robust predictions.

Random Forest was deemed suitable for the problem at hand because of its effectiveness in handling both categorical and numerical features, making it well-suited for traffic collision prediction.

3.4 Model Training and Evaluation

The dataset was split into training and testing sets using an 80-20 ratio, with 80% of the data used for training and 20% for testing.

The Random Forest model was trained using the training data, with hyperparameters optimized for performance using cross-validation.

Model performance was evaluated using the Binary Classification Evaluator, which calculated accuracy as the primary evaluation metric.

3.5 Pipeline Construction

A data processing pipeline was constructed using the Spark ML Pipeline API. This pipeline encapsulates the entire workflow from data preprocessing to model training and prediction.

The pipeline includes stages for feature encoding, feature vectorization, model training, and prediction, ensuring a streamlined and reproducible workflow.

3.6 Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to gain insights into the distribution and relationships between different variables in the dataset. This involved visualizing the data in Power BI, using histograms, scatter plots, and correlation matrices to identify patterns and potential predictors of traffic collision safety.

4 Dataset Description

Collision data shows information about each traffic collision on city streets within the City of Chicago limits and under the jurisdiction of Chicago Police Department (CPD). About half of all collision reports, mostly minor collisions, are self-reported at the police district by the driver(s) involved and the other half are recorded at the scene by the police officer responding to the collision.

Here are some of the columns that could be useful for predicting collision-prone zones:

1. **Location Description:** Description of the location where the collision occurred (e.g., Coordinates).
2. **Intersection Related:** Indicates whether the collision occurred at an intersection (Yes/No).
3. **Traffic Control Device:** Type of traffic control device at the collision location (e.g., Stop Sign, Traffic Signal, No Control).
4. **Weather Condition:** Weather conditions at the time of the collision.
5. **Lighting Condition:** Lighting conditions at the time of the collision.
6. **Traffic Way Type:** Type of traffic way where the collision occurred (e.g., One-Way, Divided, Not Divided).
7. **Roadway Surface Cond:** Roadway surface conditions at the time of the collision.
8. **Street Direction:** Direction of the street where the collision occurred (e.g., North, South, East, West).
9. **Street Name:** Name of the street where the collision occurred.
10. **Collision Day of Week:** Week Day when the collision occurred.
11. **Collision Month:** Month when the collision occurred.

5 Results

The developed Random Forest classifier model demonstrates promising performance in predicting the safety of streets in the city of Chicago based on a variety of factors. The model achieved an accuracy of **79.71%**, indicating its ability to correctly classify street safety and predicting both safe and unsafe streets for a particular day with user given conditions.

The visual analysis of the dataset using Power BI yielded important information regarding streets and found top collision prone streets in Chicago (see Fig. 1). The data showcases the top ten streets with the highest number of reported collisions in Chicago, offering valuable insights into areas requiring heightened attention for traffic safety measures at these locations.

The visualization (see Fig. 2) illustrates the variation in traffic collisions throughout the week, highlighting a noticeable increase during weekends, particularly on Fridays and Saturdays.

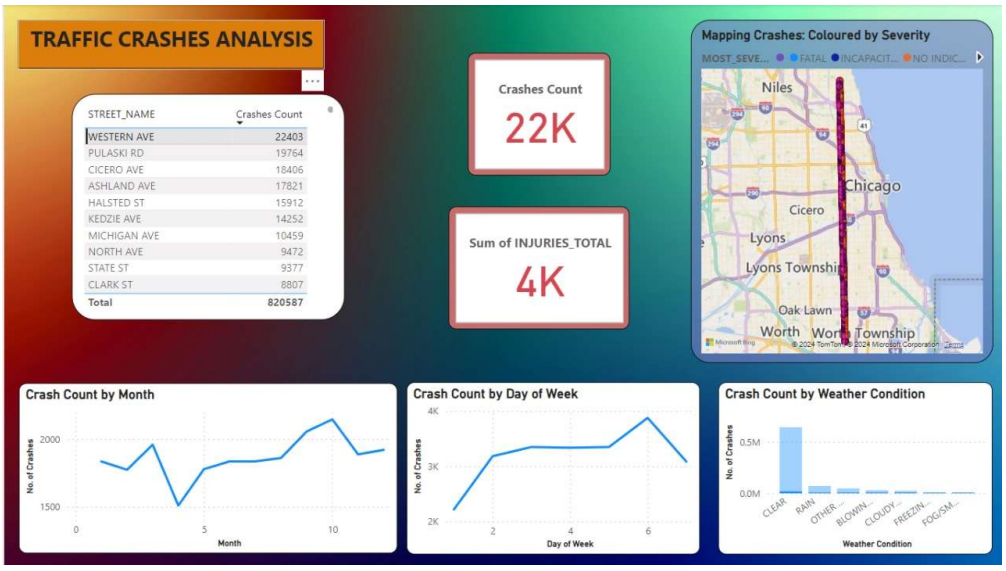

Fig. 1. Top 10 Collision-Prone Streets in Chicago

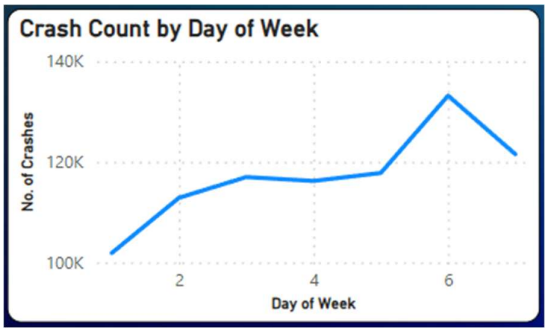

Fig. 2. Trend of Traffic Collisions by Day of the Week

Furthermore, analysis of feature importance revealed key factors influencing street safety outcomes. Among these are roadway surface condition, weather conditions and lighting conditions, the day of the week and month of the collision, highlighting the importance of temporal factors in predicting collision outcomes.

Overall, the results of the model provide valuable insights into the factors contributing to street safety in Chicago. By analysing the model predictions, potential collision hotspots and high-risk areas were identified. This information can be invaluable for traffic management authorities in implementing targeted interventions and safety measures.

6 Conclusion

In conclusion, our study presents an approach to traffic collision prediction, by using advanced machine learning algorithms, particularly Random Forest. Through data analysis and predictive modelling, we have unveiled intricate patterns and factors influencing traffic collisions, providing stakeholders with invaluable insights for enhancing road safety.

Our model stands out as a unique contribution to the field, as no similar approach has been reported in the literature. By harnessing the power of machine learning and big data analytics, we have developed a robust predictive framework capable of accurately forecasting traffic collision outcomes.

Looking ahead, the integration of our predictive model into existing software applications represents a promising avenue for future development such as in navigation apps and in urban planning tools, to provide users with real-time insights into street safety conditions. By incorporating real-time data feeds, we can empower individuals to assess the safety of streets before venturing out, thereby enhancing overall awareness.

References

1. Yan, Miaomiao, and Yindong Shen. (2022). Traffic Accident Severity Prediction Based on Random Forest. Sustainability 14, no. 3: 1729. doi: 10.3390/su14031729

2. Jianjun Yang, Siyuan Han, Yimeng Chen. (2023). Prediction of Traffic Accident Severity Based on Random Forest. Journal of Advanced Transportation, 2023. doi: 10.1155/2023/7641472

3. Deekshetha H. R., Shreyas Madhav A. V., Amit Kumar Tyagi. Traffic Prediction using Machine Learning.

4. Sun, Xu, Hanxiao Hu, Shuo Ma, Kun Lin, Jianyu Wang, and Huapu Lu. (2022). Study on the Impact of Road Traffic Accident Duration Based on Statistical Analysis and Spatial Distribution Characteristics: An Empirical Analysis of Houston. Sustainability 14, no. 22: 14982. doi: 10.3390/su142214982

5. Santos, Daniel, José Saias, Paulo Quaresma, and Vítor Beires Nogueira. (2021). Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. Computers 10, no. 12: 157. doi: 10.3390/computers10120157

6. Akin, Darcin, Virginia P. Sisiopiku, Ali H. Alateah, Ali O. Almonbhi, Mohammed M. H. Al-Tholaia, and Khaled A. Alawi Al-Sodani. (2022). Identifying Causes of Traffic Collisions Associated with Driver Behavior Using Supervised Machine Learning Methods: Case of Highway 15 in Saudi Arabia. Sustainability 14, no. 24: 16654. doi: 10.3390/su142416654

7. Dong C, Xie K, Sun X, Lyu M, Yue H. (2019). Roadway traffic collision prediction using a state-space model-based support vector regression approach. PLoS ONE 14(4): e0214866. doi: 10.1371/journal.pone.0214866

8. Jianli Xiao. (2019). SVM and KNN ensemble learning for traffic incident detection. doi: 10.1016/j.physa.2018.10.060

9. Fiorentini, Nicholas, and Massimo Losa. (2020). Handling Imbalanced Data in Road Collision Severity Prediction by Machine Learning Algorithms. Infrastructures 5, no. 7: 61. doi: 10.3390/infrastructures5070061

10. Niyogisubizo, Jovial, Murwanashyaka, Evariste, Nziyumva, Eric. (2021). A Comparative Study on Machine Learning-based Approaches for Improving Traffic Accident Severity Prediction. doi: 10.17577/IJERTV10IS100103

11. Iranitalab, Amirfarrokh; Khattak, Aemal. (2017). Comparison of four statistical and machine learning methods for collision severity prediction. doi: 10.1016/j.aap.2017.08.008

12. Assi, Khaled. (2020). Traffic Collision Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models. International Journal of Environmental Research and Public Health 17, no. 20: 7598. doi: 10.3390/ijerph17207598

13. Chen, Mu-Ming, and Mu-Chen Chen. (2020). Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree, and Random Forest. Information 11, no. 5: 270. doi: 10.3390/info11050270

14. Kenny Santos, João P. Dias, Conceição Amado. (2022). A literature review of machine learning algorithms for collision injury severity prediction. Journal of Safety Research, Volume 80, Pages 254-269. doi: 10.1016/j.jsr.2021.12.007

15. Dataset Used: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/data_preview