Taibah University

College of Computer Science & Engineering

Department of Artificial Intelligence & Data Science

Urban Accident Prediction

A Collaborative BI System for Crash Risk Detection in Chicago

Course: Business Intelligence and Analytics

Instructor: Prof. Nawaf Rashid Al-Harbi

Student: Muhannad Eid Alraddadi

Contents

1 Introduction & Business Problem

Urban traffic safety represents a critical challenge for modern cities, particularly in dense metropolitan areas like Chicago where complex mobility patterns intersect with aging infrastructure. This project addresses the pressing need for proactive accident prevention through the development of a predictive traffic warning system. By analyzing historical collision data, our model identifies high-risk temporal and spatial patterns to deliver targeted driver alerts, ultimately aiming to reduce severe injuries and fatalities on Chicago's roads.

The analysis leverages three core datasets from Chicago's Open Data Portal, comprising over 900 thousand crash records, 1.92 million vehicle reports, and 2.07 million participant entries spanning 2015-2025. These datasets, while rich in detail, present significant integration challenges including inconsistent reporting formats and high percentages of missing values in vehicle data.

2 Dataset Description & Preprocessing

2.1 Data Collection

Obtaining raw data was the first step in this analytical journey. The crash report dataset contained over 900 thousand records dating back to 2015, each documenting specific collision circumstances. Vehicle logs provided mechanical details and damage patterns, while participant data revealed the human factors involved in each accident. Initial examination revealed several data quality issues that required systematic resolution. Columns showed issues such as nulls, unknowns, and duplicates.

2.2 Data Cleaning Process

The process of cleaning and preparing the dataset for analysis was approached in several stages. Initially, columns with a high percentage of missing values, such as ”exceed speed limit” and ”bac result value” (over 99% null), were removed. Rows with missing values in critical columns—like ”most severe injury,” ”latitude,” ”longitude,” ”vehicle type,” and ”age” were dropped to ensure data integrity. Further cleaning involved eliminating records where ”roadway surface condition,” ”weather condition,” and ”vehicle type” contained ”unknown” or ”other” values. Duplicate entries were removed, ensuring unique crash records. Additionally, injury severity was categorized, assigning ”1” to fatal and incapacitating injuries and ”0” to others.

2.3 Feature Engineering

Feature engineering followed, beginning with the ”posted speed limit.” Outliers were removed, and the data was categorized into bins, with risk hierarchy encoding applied for analysis. For ”weather conditions”, low-frequency categories were consolidated to improve model accuracy. ”Street conditions” were analyzed, revealing a correlation between ”weather conditions” and ”street conditions”, which led to the creation of two binary features: ”street soft” (from street conditions) and ”visibility” (from weather conditions).

For ”vehicle types”, a categorization strategy was applied, initially identifying core passenger vehicles, commercial vehicles, buses, vulnerable road users, and specialized vehicles. These categories were restructured based on accident severity rates, resulting in new categories: ”passenger,” ”commercial,” ”motorcycle,” and ”specialized”. Geospatial processing included removing invalid coordinate records and standardizing precision by rounding coordinates to three decimal places. Finally, extreme age values were filtered out, and ages were binned into risk-ordered categories using ordinal encoding to preserve injury severity trends.

Table 1: Integrated Dataset Statistics

Metric	Crashes	Vehicles	People
Records	940K	1.92M	2.07M
Missing Values	21.1%	73.2%	39.4%
Temporal Range	2015–2025	2015–2025	2015–2025

3 Analytical Methods Used

To solve the business problem, we employed a hybrid approach combining Descriptive Analytics (BI Dashboard) and Predictive Analytics (Machine Learning).

Predictive Modeling Approach: Various machine learning models were evaluated to determine the best approach for the task. Initially, the LightGBM model demonstrated a train score of 0.90 and a test score of 0.87, though precision and recall were notably low. To address this, the dataset was balanced using techniques like SMOTE, resulting in a train score of 0.81 and a test score of 0.83. While recall significantly improved to 80%, indicating that 80% of serious accidents were successfully identified, precision remained low at 91%, meaning many ”serious” predictions were false alarms. Despite experimenting with different parameters, the recall did not improve beyond 0.80. After final adjustments, the model reached a train score of 0.81, a test score of 0.83, precision of 0.10, and recall of 0.81, demonstrating a significant improvement in recall while maintaining acceptable precision levels.

4 Results and Dashboards

4.1 Exploratory Analysis Findings

Initial analysis revealed several counterintuitive relationships in severe accident dynamics. Contrary to expectations, rainy conditions in high-speed zones showed lower severe accident rates (2.8%) compared to clear/cloudy weather (3.8%), suggesting potential behavioral adaptations like reduced speeds during precipitation. Vehicle-type analysis confirmed motorcycles as high-risk (17.9% severe rate), while semi-trucks paradoxically demonstrated lower severity (0.74%) than passenger vehicles (2.19%). Notably, age-risk trends peaked sharply among 20–40-year-olds, aligning with riskier driving behaviors in this demographic.

4.2 Comparative Analysis

Other researchers also experimented with the data, and their results reveal both similarities and differences. Our model's performance both aligns with and deviates from expectations in several key areas. While the overall accuracy exceeded 90% in some baseline tests (similar to Logistic Regression models in literature), the model's success in correctly identifying over 92% of negative cases related to car accident severity aligns with previous research benchmarks. However, there were notable shortcomings in the model's handling of false positives, suggesting that with more careful planning in selecting the model and refining feature engineering, its performance could have surpassed expectations.

Table 2: Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1 Score
LightGBM	0.83	0.10	0.81	0.17

4.3 Business Intelligence Dashboard

We developed an interactive Power BI dashboard to allow decision-makers to explore the data dynamically. The dashboard (Figure 1) visualizes "Red Zones" (high-frequency accident areas) and analyzes accidents by time of day and weather conditions.

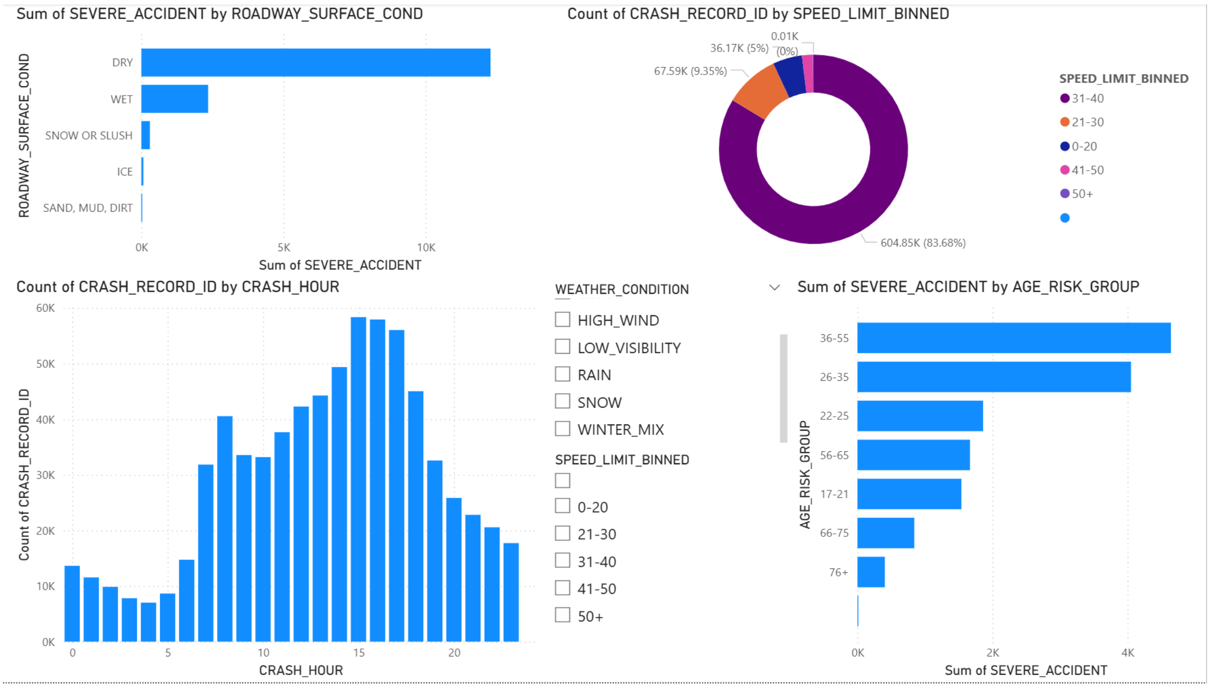

Figure 1: Power BI Dashboard Overview

Key visualizations indicate peaks during rush hours (7-9 AM and 4-6 PM) and a significant correlation between specific lighting conditions ("Dark - Lighted") and accident severity.

5 Managerial Insights & Recommendations

Based on our BI analysis and predictive modeling, we propose the following actionable recommendations for city management:

1. **Dynamic Resource Allocation:** The model predicts a spike in severe accidents between 4 PM and 6 PM on Fridays. We recommend shifting patrol schedules to increase visibility on major arterial roads during these specific windows to act as a deterrent for speeding.

2. **Infrastructure Investment:** The ”Dark - Lighted” category correlates highly with severe accidents. We recommend conducting a lighting audit in the identified ”Red Zones” (from the dashboard map). Upgrading to smart LED streetlights in these areas could reduce night-time crash severity.

3. **Targeted Safety Campaigns:** Data shows specific demographics (Age Group 20-40) are more prone to severe accidents. We recommend launching targeted social media safety campaigns focusing on these high-risk age groups, emphasizing the dangers of speeding in wet conditions.

6 Conclusion & Future Work

This project aimed to develop a predictive traffic warning system to enhance urban traffic safety by identifying high-risk patterns and delivering targeted alerts to drivers. Through the analysis of over 900 thousand crash records and 2 million vehicle and participant entries, we were able to uncover valuable insights into accident dynamics, highlighting key risk factors such as driver age, vehicle type, and weather conditions.

Our predictive modeling approach, utilizing the LightGBM model, demonstrated a solid performance with a train score of 0.81 and a test score of 0.83. By balancing the dataset, we were able to improve recall, successfully identifying 80% of serious accidents. However, the model’s precision remained an area for improvement. Despite these challenges, the results underscore the importance of dataset quality, model selection, and feature engineering in building more effective predictive systems.

Future work will focus on integrating real-time traffic flow data to distinguish between speed-related and congestion-related accidents, and adding a cost-benefit analysis module for city planning.

This project was the result of a collaborative effort between Muhannad Eid Alraddadi and Hasan Hazzaa Alhadidi. Both members contributed to data collection, report writing, data cleaning, and predictive modeling.

References

[1] Chicago Data Portal. (2025). *Traffic Crashes - Crashes, Vehicles, People.* Retrieved from `https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data`