

Crash Risk Prediction Model using Data Science

This manuscript ([permalink](#)) was automatically generated from [uiceds/cee-492-term-project-fall-2022-swifties@427a32b](#) on October 28, 2022.

Authors

- **Lara Diab**

 [0000-0001-8489-2015](#) ·  [diablara](#)

Illinois Center for Transportation (ICT); Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

- **Renan Santos Maia.**

 [0000-0002-0877-4006](#) ·  [renanssmaia](#)

Illinois Center for Transportation (ICT); Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

- **Farid Saud**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [fsaudm](#)

Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

- **Johann J Cardenas Huaman**

 [0000-0002-4695-7639](#) ·  [Johann-Cardenas](#) ·  [transporter_pe](#)

Illinois Center for Transportation (ICT); Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

Project selection and Introduction

Road accidents are responsible for a significant number of injuries reported every year. According to the World Health Organization (WHO), approximately 1.3 million people die each year as a result of road traffic crashes (as of June, 2022). In addition, road traffic crashes cost countries 3% of their gross domestic product (Safarpour et al, 2020; WHO, 2022). Consequently, understanding what influences these accidents on roads is of utmost importance. However, it is not easy to decide which exact conditions lead to these accidents. Different road, climate, vehicle and driver conditions affect the likelihood of a driver to be in a fatal/serious accident.

The ability of predicting in an accurate way the potential occurrence of car crashes is a valuable contribution for road safety. In an approach frequently used in the literature, crash records' data are used for the development of crash prediction models, so that agencies can allocate investments to priority areas of the roadway network. However, given that the budget for infrastructure improvements is limited, adopting countermeasures for all facilities that crashes are potentially occurring is not financially feasible. Therefore, informing drivers about the potential safety risks is a way to proactively compensate the aforementioned limitations. Moreover, with the development of connected and autonomous vehicles, this information can be provided in a more optimized way, contributing for vehicles' route decision, as well as for real-time alerts that can lead drivers to take the necessary precautions to operate more safely (Yu et al, 2021).

Project Objective and Plan Proposal

The objective of this work is to use the Illinois Department of Transportation (IDOT) extensive crash data to be analyzed and, finally, be used for a crash risk prediction model based on main categorical data that can be real-time updated (such as the weather/lighting/pavement conditions). Ideally, it could be used by navigation systems to alert drivers to adopt more cautious behavior as soon as they enter higher-risk sections.

The plan to be carried out will follow the basic steps described as follows:

1. Read the IDOT's crash data CSV files available as an open data source.
2. Clean the data by deleting unwanted columns, handling missing data, and removing irrelevant observations.
3. Tidy the data by organizing the variables into columns and the observations into rows.
4. Analyze and visualize the data by finding correlations both analytically and graphically.
5. Model a prediction algorithm for crash risk according to categorical variables.

Description of the Data Set

IDOT has generated datasets with statewide crash locations produced by the Crash Information Section of the Illinois Department of Transportation (IDOT). The accident data has been collected throughout the years using Application Programming Interfaces (APIs) that provided streaming traffic incident data. There are about 300,000 accident records per year in these datasets, and each record contains attributes that include conditions like (among others that are not listed and described because these are not relevant for this study):

1. Time and date (day, month, year)
2. Coordinates (x,y)
3. Type of collision
4. A quantitative description of fatalities and injuries

5. Crash severity classification based on their impact on traffic
6. The road surface condition ("Dry", "Wet", "Snow or slush", "Ice", or "Sand/Dirt/Mud")
7. Road defects ("Debris on roadway", "Rut/Holes", "Unknown", or "No defects")
8. Light conditions (rated in a scale from 1 to 9)
9. Geometric characteristics of the road section
10. Work Zone ("construction", "maintenance", "utility", "unknown", or "N/A")
11. Possible causes of the accident.

The datasets for different years are available for download as a .CSV file in the IDOT's website:

<https://gis-idot.opendata.arcgis.com/search?groupIds=6d2862031a6d47c7a8c211e38e423e05>

Exploratory Data Analysis

Extensive data was found, each dataset containing the observations arranged in rows and the independent variables in columns. One of these datasets included up to 85 columns. Datasets from 2017,2018,2019 are included in this Exploratory Data Analysis.

Reading the Data

To carry out the first step, the datasets were imported to Visual Studio Code using the CSV library. The datasets were merged into a unified file once the total numbers of columns of each of them were matched to an homogeneous number since it was noticeable that the latest dataset included additional independent variables the first two did not include. The three files were merged into a new database after this first sanity check.

Cleaning Process

It was found that several independent variables would not provide fruitful information due to missing, unknown or incomplete data. Thus, the datasets were visually inspected to filter out irrelevant or incomplete variables. For instance, information pertaining the location (latitude & longitude, or X & Y coordinates) have not been taken into account. Columns containing codes describing the city, county or ID of the location where the crash took place has also been excluded. Columns involving duplicate information (e.g. two columns describing the same independent variable with a label and a number), traffic structures, etc were also removed. For some other independent variables, information that would have been useful was found to be significantly incomplete such as the number of lanes or type of intersection and therefore it was decided to not include it. Additionally, independent variables with a high number of labels to describe them such as Railroad Crossing Number were also filtered out since it would not provide handy information for the end-user.

After filtering out the information that will not be utilized for this analysis, the number of independent variables went from 80 to 21.

Project Objective and Plan Proposal

The objective of this work is to use the Illinois Department of Transportation (IDOT) extensive crash data to be analyzed and, finally, be used for a crash risk prediction model based on main categorical data that can be real-time updated (such as the weather/lighting/pavement conditions). Ideally, it could be used by navigation systems to alert drivers to adopt more cautious behavior as soon as they enter higher-risk sections.

The plan to be carried out will follow the basic steps described as follows:

1. Read the IDOT's crash data CSV files available as an open data source.
2. Clean the data by deleting unwanted columns, handling missing data, and removing irrelevant observations.
3. Tidy the data by organizing the variables into columns and the observations into rows.
4. Analyze and visualize the data by finding correlations both analytically and graphically.
5. Model a prediction algorithm for crash risk according to categorical variables. }

References

Illinois Department of Transportation | IDOTAdmin <https://gis-idot.opendata.arcgis.com/search?groupIds=6d2862031a6d47c7a8c211e38e423e05>

Safarpour, H., Khorasani-Zavareh, D., & Mohammadi, R. (2020). The common road safety approaches: A scoping review and thematic analysis. *Chinese journal of traumatology*, 23(02), 113-121. <https://doi.org/10.1016/j.cjtee.2020.02.005>

World Health Organization (WHO). (2022). Road traffic injuries World Health Organization Regional Office for the Eastern Mediterranean. Geneva, Switzerland. <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>

Yu, R., Han, L., & Zhang, H. (2021). Trajectory data based freeway high-risk events prediction and its influencing factors analyses. *Accident Analysis & Prevention*, 154, 106085. <https://doi.org/10.1016/j.aap.2021.106085>.