Risk Prediction and Assessment in the Construction Industry

Zhixing Wang
Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
zw88@illinois.edu

Zain Sitabkhan
Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
zsita@illinois.edu

Deago Sirenden
Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
deagofs2@illinois.edu

Zach Da
Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
zhihuid2@illinois.edu

Abstract—This project focuses on risk prediction and assessment in the construction industry using incident and accident data from New York City. By applying regression-based models, the objective is to predict fatality and injury outcomes, as well as generate a weighted index to evaluate the severity of such events. The study contributes to understanding which attributes most strongly influence construction-related incidents and provides insights that may improve safety measures in the industry.

Index Terms—Construction Safety, Risk Prediction, Accident Reports, Regression Analysis

I. Data Description

A. File Content

The dataset consists of construction-related incidents and accidents at New York City in each of the five boroughs. It provides a large-scale CSV file suitable for predictive analysis.

B. Source

Department of Buildings (DOB) Incident Database

C. Format and Size

The dataset includes approximately 958 rows, each representing an accident or incident record, and 20 columns containing attribute fields of these records.

II. Attributes

Attribute Name	Unit/Type	Description
BIN	Integer	Building Identification Number (unique ID for each building)
Accident Report ID	Integer	Unique identifier of each accident report
Incident Date	Date	Date of the incident or accident
Record Type Description	Category (Text)	Record type, distinguishing Incident from Accident
Check2 Description	Category (Text)	Detailed category of the incident, e.g., Construction Related, Mechanical Equipment, Worker Fall
Fatality	Integer	Number of fatalities
Injury	Integer	Number of injuries
House Number	Text/Number	House number of the incident location
Street Name	Text	Street name of the incident location
Borough	Category	Administrative borough (e.g., Manhattan, Bronx, Brooklyn)
Block	Integer	Geographic block number
Lot	Integer	Lot number within the block

Postcode	Integer	Postal code of the location
Latitude	Float	Latitude coordinate of the incident location
Longitude	Float	Longitude coordinate of the incident location
Community Board	Integer	Community board identifier
Council District	Integer	City council district identifier
BBL	Integer	Borough-Block-Lot unique cadastral identifier
Census Tract (2020)	Integer	Census tract number from the 2020 census
Neighborhood Tabulation Area (NTA) (2020)	Text	Neighborhood Tabulation Area (NTA) code from 2020

Proposal for attribute usage will be made, focusing on those with predictive relevance.

III. Proposal

A. Objectives

Our main objective is to analyze different types of construction incidents at New York City that happened within 1 or 2 years from now. For this project, we would mainly be examining the nature of construction related incidents and accidents as well as performing correlations with the data by examining the prevalence of each incident and accident at each of the five boroughs of New York City. We would want to see where each type of incident has the highest probability of occurring, and where specifically measures should be implemented to prevent these types of incidents. Finally, keeping track of when these incidents occurred will also be critical as the data could also be used to calculate the frequency of accidents over time.

B. Preprocessing

Filtering may be applied to eliminate less effective or redundant variables, such as postcode or latitude, to improve model performance. Tidying and cleaning the data is also necessary before analyzing and correlating the data. We would probably need to order our data in terms of when they happen as well as categorizing the incidents/accidents that happened at each borough.

C. Output

- Incident vs accident count over time (could be in spans of 1 month)

- How many construction incidents and accidents happened at each month
- Computed severity index at each borough (e.g., grade scale from 1 = less severe to 10 = highly severe)

D. Input

Date, record type, latitude, longitude, type of incident, and BIN (business effect case, combined with other data).

E. Significance

The purpose of the model is to help avoid incidents and accidents at New York City by identifying the dominant attributes influencing outcomes, thereby guiding proactive protection measures in construction management.

References