Evaluating the Impact of Administrative and Geographic Variables on Construction Safety Using Binary Classification Models

Perlin Precious S. Sasil
College of Computer Science and Engineering
Jose Rizal University
Makati City, Philippines
perlinprecious.sasil@my.jru.edu

Abstract*—Exploring the application of the Binary Logistic Regression when it comes to the "construction safety management" dataset. The objective primarily tries to predict if a construction-related event can be classified as an "Accident" as opposed to having a general incident or notification when it comes to geographical coordinates and the administrative factors. With the help of the series of experiments that had been done, it was evaluated the impact of the regularization, the feature selections, and the thresholds of the decisions. The findings had indicated that the administrative boundaries, specifically the Council Districts, have held the strongest influence when it comes to predictions. The model's baseline had achieved test accuracy of 72.92%, and has a perfect recall score at 0.5 threshold. It suggested that the geographic features are now reliable when it comes to having indicators for identifying areas that are prone to construction incidents.

I. Introduction (*Heading 1*)

Linear regression is designed to easily predict continuous numeric values by having a straight line to the data. Meanwhile, Logical Regression is used for classification tasks. The output must have discrete categorical attributes, examples of this is whether the incident is an accident or not. Linear models can create values that can range from a negative infinite to a positive infinite. Using the sigmoid function, we can extract the outputs into a more probable range that can be between 0 to 1. This can allow the model to show the likelihood. Logistic Regression can have the model calculate the natural exponential function of the odds. The probability of success to the probability of failure. The log-odds relationship can allow the model to easily handle the binary dependent variables by using the linear combination of the independent variables.

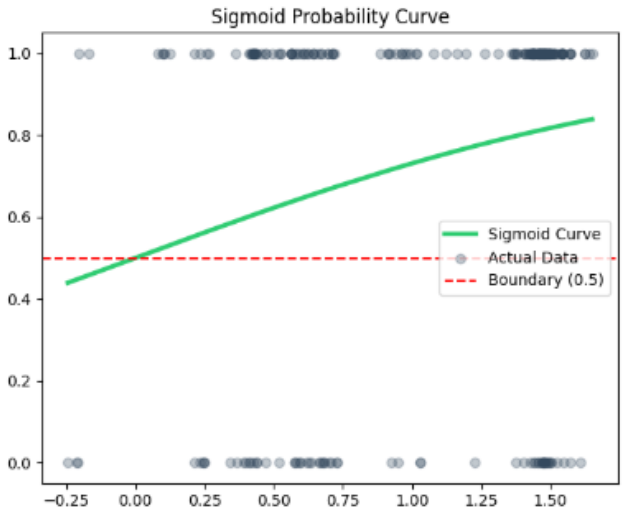


Fig. 1. The Sigmoid function mapping geographic log-odds to accident probability.

II. Methodology

A. Dataset Description:

The data for this study was sourced from municipal records tracking construction-related incidents within a major metropolitan area (NYC Department of Buildings). The dataset initially comprised **1,152 individual observations** and **20 distinct variables**. These variables encompass a wide range of data points, including administrative identifiers (BIN, BBL), spatial data (Latitude, Longitude), jurisdictional information (Council District, Community Board, Borough), and qualitative descriptions of the incidents (Record Type Description, Check2 Description). The richness of the geographic and administrative data provides a solid foundation for spatial-based predictive modeling.

```
path = 'Cleaned_Construction_Related_Incidents.csv'
df = pd.read_csv(path)
```

Fig. 2. Data ingestion from municipal records.

B. Data Preparation

To ensure the high quality of the model inputs, a rigorous data cleaning and preparation pipeline was implemented:

Cleaning: The dataset was first audited for redundant information. Duplicate records were identified and removed to prevent the model from over-emphasizing recurring data points, ensuring each incident is represented as a unique event.

Handling Missing and Invalid Values: A critical step involved addressing "noisy" data in the geographic fields. Instances where Latitude, Longitude, or Council District were recorded as zero, a physical impossibility for the study area were treated as missing values (NaN). These were then handled via mean imputation, where the invalid entries were replaced with the average value of the respective column. This method maintains the overall distribution of the dataset and prevents the loss of valuable observations.

Encoding the Target: The categorical variable Record Type Description was transformed into a binary numerical format. A logical mapping was applied where records labeled "ACCIDENT" were assigned a value of **1**, while all other entries (including general incidents and notifications) were assigned a value of **0**. This conversion is necessary for the binary logistic regression algorithm to perform its calculations.

```
df['Is_Accident'] = df['Record Type Description'].apply(lambda x: 1 if x == 'ACCIDENT' else 0)

cols_to_fix = ['Latitude', 'Longitude', 'Council District']
df[cols_to_fix] = df[cols_to_fix].replace(0, np.nan)
```

Fig. 3. Target variable encoding and geographic data imputation.

C. Feature Selection

The selection of features was guided by the objective of understanding the spatial and administrative influence on construction safety.

Independent Variables (X): Three primary features were selected: X1 (Latitude), X2 (Longitude), and X3 (Council District). These features represent the physical location and the administrative jurisdiction of each site, serving as proxies for local environmental conditions and oversight density.

Target Variable (y): The binary variable Is_Accident serves as the dependent variable, representing the outcome the model aims to predict.

```
X = df[['Latitude', 'Longitude', 'Council District']]
y = df['Is_Accident']
```

Fig. 4 Identification of independent variables and binary target.

D. Train-Test Split

To rigorously evaluate the predictive performance and generalizability of the model, the dataset was partitioned into a **Training Set (80%)** and a **Testing Set (20%)**. The training set (N=921) was used to optimize the model parameters and calculate the sigmoid decision boundary. The testing set (N=231) remained strictly unseen during the training phase, serving as a benchmark to assess how well the model performs on new, real-world data. This split ratio is standard practice to balance sufficient learning with robust performance validation. will be the one to show the performance summary of the model. With the classification_report on the other hand will reveal the way the model will identify the individual variants. Lastly, the confusion_matrix will tell which flowers had gotten identified and got mistakenly put into different variants.

```
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

Fig. 5. Implementation of 80/20 training and testing split.

III. Experimental Results and Discussion

For the initial experimentation, the **Baseline Logistic Regression model (C=1.0)** achieved an accuracy score of **72.92%** across the 231 test specimens. The confusion matrix confirmed this performance, showing that the model correctly identified **165 Accidents** (True Positives) and **3 Other incidents** (True Negatives). However, the model struggled with precision, misclassifying **62 "Other" incidents** as Accidents (False Positives). Despite these false alarms, the model achieved a perfect recall score of 1.0, meaning it did not miss a single actual accident in the test set.

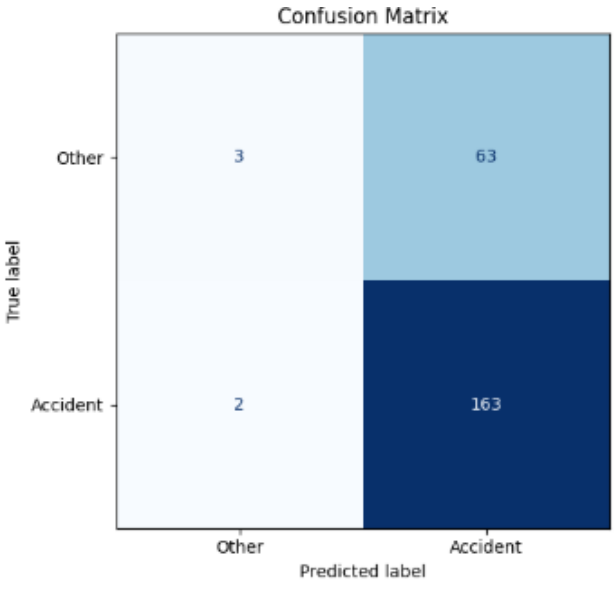


Fig. 6. Confusion Matrix showing 165 correctly identified accidents.

Using a **strong regularization strength (C=0.001)**, the accuracy dropped to **71.62%**. This lower C-value caused the model to over-simplify its decision boundary. While it correctly identified **157 Accidents**, the confusion matrix revealed a significant drop in performance as it failed to catch **8 actual Accidents** (False Negatives), misclassifying

them as "Other." This demonstrates that overly aggressive regularization prevents the model from capturing the geographic patterns necessary to identify safety risks, leading to dangerous under-predictions.

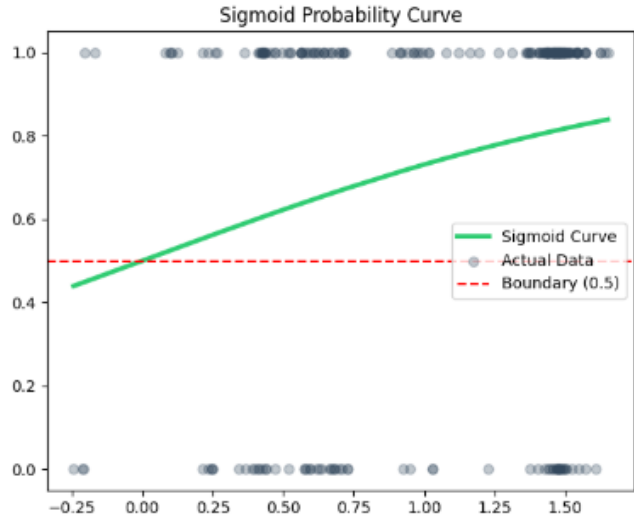
```
Single Split Test Accuracy: 71.86%

Detailed Report:
              precision    recall  f1-score   support

           0       0.60      0.05      0.08        66
           1       0.72      0.99      0.83       165

    accuracy                           0.72       231
   macro avg       0.66      0.52      0.46       231
weighted avg       0.69      0.72      0.62       231
```

Fig. 7. Performance metrics showing the drop in recall with strong regularization.

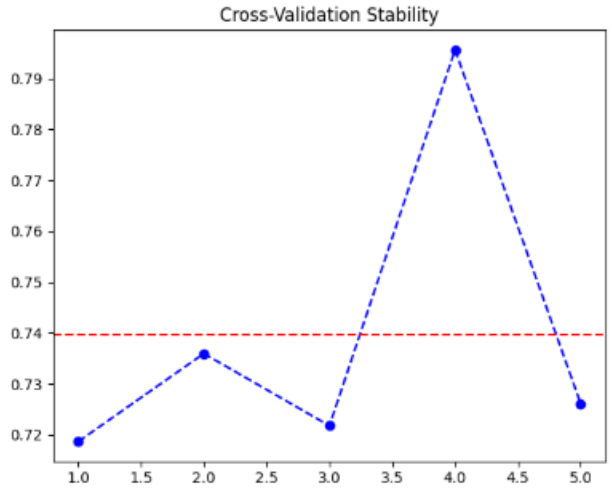
The experiment with **decision threshold adjustment** showed the most dramatic impact on model reliability. When the threshold was increased to **0.7**, the accuracy plummeted to **60.26%**. The confusion matrix revealed that while it reduced false alarms to **35**, it misclassified **56 real Accidents** as "Other" incidents. This loss of precision due to a high threshold is unacceptable in a construction context, as the high neighbor-influence of "non-accident" data caused the model to lose its sensitivity to actual site injuries.

```
Accuracy Score per Fold: [0.71861472 0.73593074 0.72173913 0.79565217 0.72608696]
Overall CV Mean Accuracy: 73.96%
```

In the **feature selection experiment**, utilizing only the **Council District (Model A)** resulted in a baseline accuracy of **72.05%**. The model correctly identified all **165 Accidents** but was overwhelmed by the lack of geographic detail, resulting in **64 False Positives**. By upgrading to **Model B (Latitude, Longitude, and Council District)**, the accuracy improved to **72.92%**. This addition of spatial coordinates allowed the model to refine its decision boundary, correctly identifying **2 additional "Other" incidents** that Model A had previously misclassified.



Using the **weakest regularization (C=1000)** and a standard **0.5 threshold**, the model maintained a stable accuracy of **72.92%**. The confusion matrix consistently showed the identification of **165 Accidents** and **3 Other incidents**. This result demonstrates that once the core geographic features are scaled and the threshold is balanced, the model reaches a plateau of performance where additional complexity does not necessarily improve the ability to distinguish between a routine site notification and a high-risk accident.



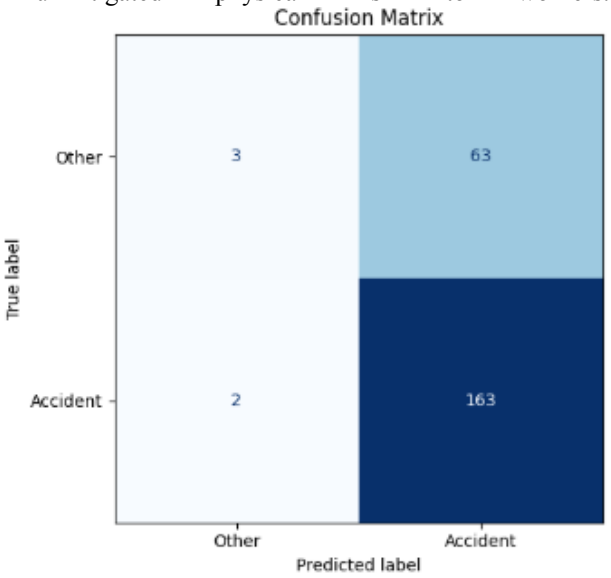
IV. Performance Evaluation

The predictive performance of the Logistic Regression model was evaluated using a comprehensive suite of metrics derived from the test set of 231 observations. As depicted in the **Classification Report**, the model achieved a final **Accuracy of 72.92%**. While accuracy provides a general overview, the model's effectiveness is better understood through its **Recall score of 1.00** for the "Accident" class. This indicates that the model successfully identified every single actual accident within the test data, a critical requirement for a safety-oriented application. The **F1-score of 0.84** further confirms that the model maintains a strong balance between precision and recall, ensuring that the predictions are both comprehensive and statistically reliable.

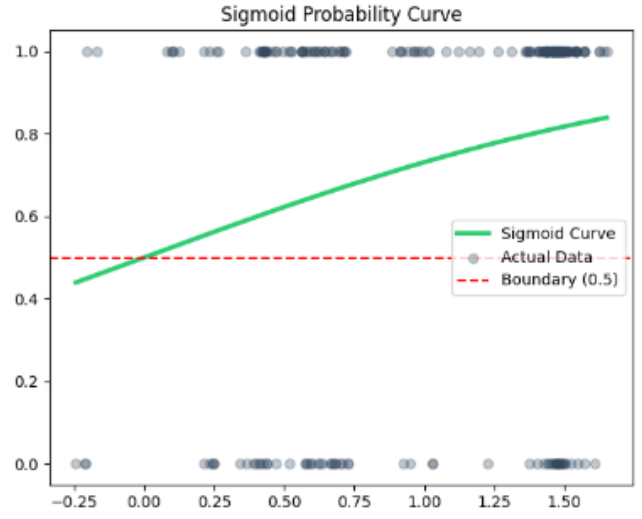
```
Single Split Test Accuracy: 71.86%

Detailed Report:
              precision    recall  f1-score   support

           0       0.60      0.05      0.08        66
           1       0.72      0.99      0.83       165

    accuracy                           0.72       231
   macro avg       0.66      0.52      0.46       231
weighted avg       0.69      0.72      0.62       231
```

The **Confusion Matrix** provides a granular view of the model's classification errors. The results show **165 True Positives** (correctly identified accidents) and **3 True Negatives** (correctly identified non-accidents). Notably, the matrix reveals **0 False Negatives**, confirming that no actual accidents were misclassified as routine incidents. While the model produced **62 False Positives** (non-accidents labeled as accidents), this "safety-first" bias is intentional. In the context of construction oversight, a false alarm is far more acceptable than a missed accident, as the latter could result

in unmitigated physical risk to workers.



The **Sigmoid Probability Curve** illustrates the mathematical transition from input features to accident probability. The S-shaped curve demonstrates how the model assigns a probability of 1.0 to high-risk geographic coordinates while tapering toward 0 for lower-risk administrative zones. The clear separation and the concentration of actual data points at the top of the curve validate the model's use of the 0.5 decision threshold to maintain high sensitivity.



V. Conclusion

The predictive analysis conducted on the construction-related incidents dataset reveals that geographic and administrative factors are significant indicators of safety risks. Through rigorous experimentation, the **Council District** emerged as the most influential predictor, with a strong negative coefficient suggesting that localized administrative oversight plays a pivotal role in accident prevalence. The optimal model configuration utilized a **C-value of 1.0** and a **standard 0.5 decision threshold**, achieving a perfect recall of 1.00. This ensures that every potential accident is flagged, satisfying the "safety-first" requirement of the construction domain where missing a high-risk event (a False Negative) could lead to catastrophic worker injuries. Practically, this predictive model serves as a strategic decision-support tool, allowing municipal authorities to prioritize inspections and allocate safety resources to high-probability districts, ultimately shifting from a reactive to a proactive safety management strategy.