# Logistic Regression for Predicting Earthquake Outcomes

Ty Johnson, tjohnson@bellarmine.edu

**ABSTRACT**
This project aims to investigate the effects of an earthquake on a building based on data collected from the 2015 earthquake in Nepal, focusing on the building and land characteristics that may have an impact on whether or not a building can withstand an earthquake. Using a logistic regression, a model will be created that aims to predict the outcome of a building after an earthquake. The findings indicate that regardless of characteristics, being in the area of impact means the building was most likely destryoed. The model was able to predict the outcome of an earthquake with 86% accuracy, though bias towards a destroyed outcome may play a part in this. It can be suggested that the factors included in the dataset are not enough to predict the outcome of an earthquake and more will be needed in order to truly understand how a building can withstand an earthquake.

## I. INTRODUCTION

The dataset used in this analysis is based on the Nepal Earthquake in 2015. It will be using logistic regression to predict the outcome of a building following the earthquake, and answer the question: did it withstand the quake or was it destroyed by it? In addition to building the logistic regression model, the project will aim to answer a few questions relevant to the event through visualizations. Together, the model and visualizations should be able to assist in understanding what factors come into play for reducing the impact that an earthquake might have on a building.

## II. BACKGROUND

The Nepal Earthquake that the dataset is from happened in April of 2015. It was an earthquake that logged an initial magnitude of 7.8 and was followed by two aftershocks of 6.6 and 6.7 respectively. In the towns of Kathmandu and those nearby, there were over 9000 causalities and thousands more were injured. As for the buildings, over 600,000 were destroyed. This led to over 2.8 million people being displaced and 8 million being affected from the earthquake in Nepal and surrounding countries. Damages are estimated to have been between $5 and $10 billion dollars. The data appears to have been collected in order to assess the damage that was caused by the quake and the characteristics of the buildings that were affected. Further insights into the authors involvement in the collection of the data were not made clear, but it was provided to Kaggle for the purpose of learning logistic regression.

## III. EXPLORATORY ANALYSIS

Prior to any alterations the dataset included 762106 samples that were spread across 21 columns of various data types. That number was changed to 759428 samples with 16 columns after some changes were made. The distribution of building age, as shown in Figure 1, reveals a positively right skewed distribution which means that it can be concluded that the majority of buildings are relatively newer in the area that was affected by the earthquake. The graph removes the outliers, such as long standing buildings that were over a thousand years old, that could have affected the distribution. Figure 2 shows the count of buildings in each district and the count for how many of those buildings were destroyed. The top three districts in terms of numbers of buildings they possessed were districts 24, 31, and 30. The smallest district was district 29. In most districts, it is apparent that a majority of the buildings were destroyed. Figure 3 shows the proportion of damage grades that were given to buildings following the earthquake. 78.3% of buildings were given a damage grade of three or higher, which goes to show the destruction that was brought by the earthquake. Figure 4 shows the outcomes of buildings based on their foundations that they were constructed from. It is shown that a majority of buildings were constructed from mud/mortar-stone. There is not enough information to conclude that any foundation type was better when it came to withstanding the earthquake, although buildings constructed from RC had significantly withstood the earthquake more than its counterparts, which may be worth looking more into.

**Figure 1: Building Age Distribution**

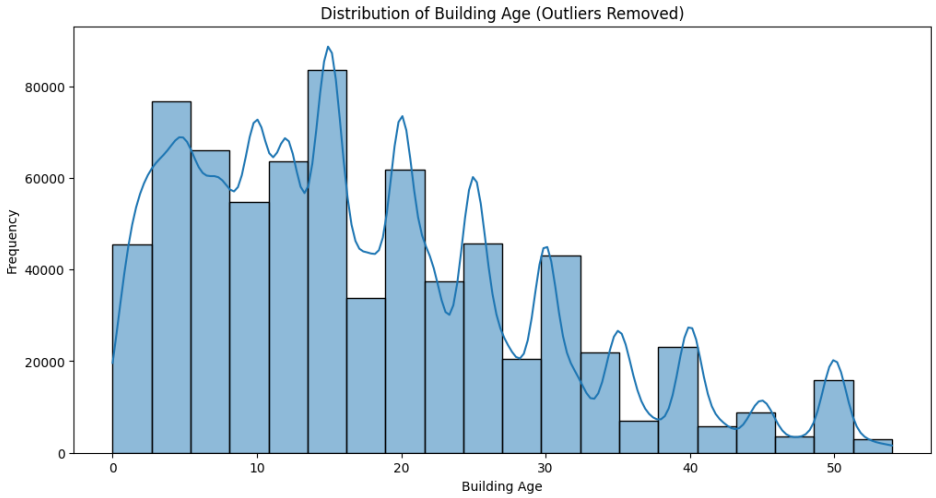



**Figure 2: Count of Buildings in Each District (Total and Destroyed)**

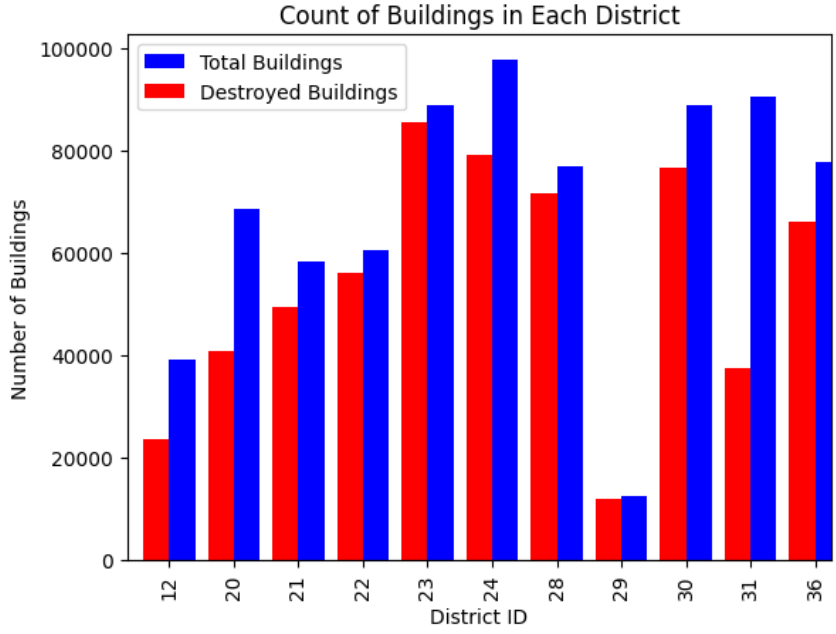


**Figure 3: Damage Grade Proportion**

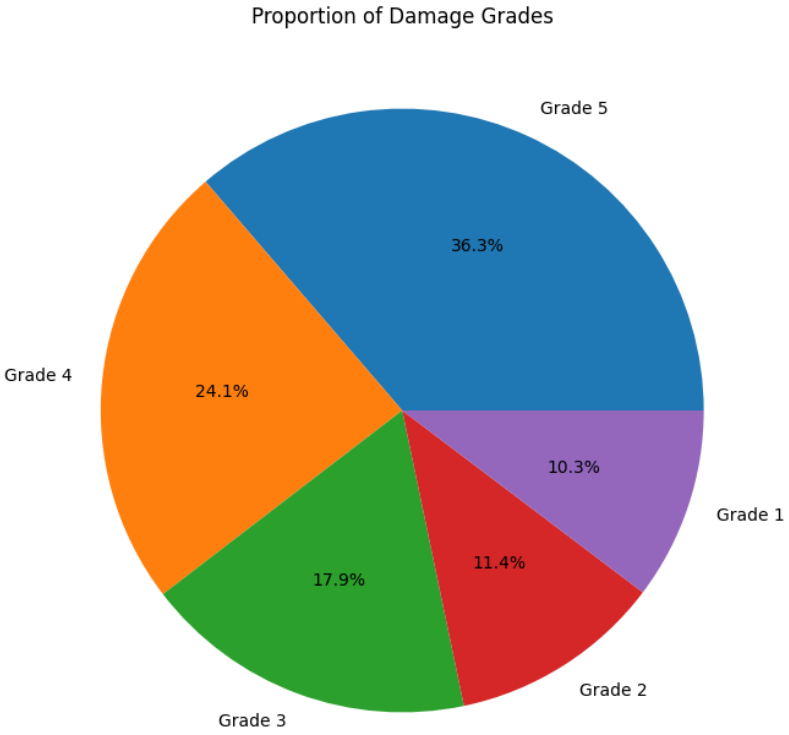



**Figure 4: Outcome by Foundation Type**

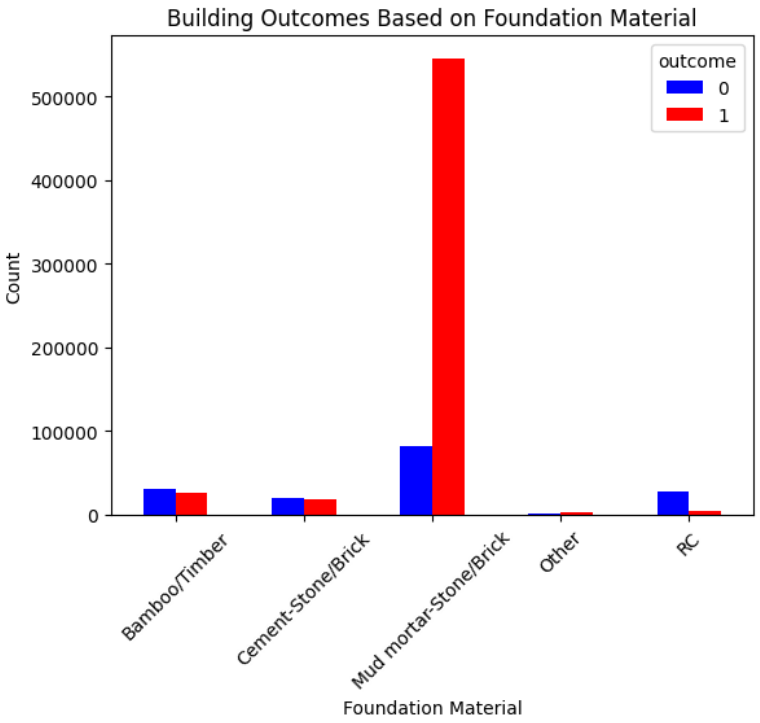


**Table 1: Data Types**

| *Variable Name* | *Data Type* |
|---|---|
| Building_id | Int64 |
| District_id | Int64 |
| Vdcmun_id | Int64 |
| Ward_id | Int64 |
| Count_floors_pre_eq | Int64 |
| Count_floors_post_eq | Int64 |
| Age_building | Int64 |
| Plinth_area_sq_ft | Int64 |
| Height_ft_pre_eq | Int64 |
| Height_ft_post_eq | Int64 |
| Land_surface_condition | Object - String |
| Foundation_type | Object - String |
| Roof_type | Object - String |
| Ground_floor_type | Object - String |
| Other_floor_type | Object - String |
| Position | Object - String |
| Plan_configuration | Object - String |
| Condition_post_eq | Object - String |
| Damage_grade | Object - String |
| Technical_solution_proposed | Object - String |
| superstructure | Object - String |

## IV. METHODS

### *A. Data Preparation*

For cleaning and preparation of the data, first I checked for any missing values in the dataset. The amount missing was very minimal, for example one column had 12 out of 700,000 values missing. The method of imputation that was used for the values that were missing was the mode, since only categorical variables contained missing values. I created two new columns to represent the change in height and number of floors before and after the earthquake, from there I was able to see erroneous values. These values showed that the buildings grew after the earthquake, which is practically impossible especially in floors and composed of a little over 2600 rows. These values were removed from the dataset given the small percentage and potential to challenge the validity of the model. There were a few columns that represented unique identifiers in the model, such as building id, ward id, and vdcmun id. These were removed from the dataset since there was no need for them in creating the model or visualizations. District Id was kept since it was used in the visualization process.

For the model, all columns were not included. A new column named “outcome” was created and served as the y, or value that the logistic regression model was attempting to predict. This column was created by mapping the technical solution proposed into a binary outcome of 0 and 1. 0 represented the building withstood the earthquake, having had no need for repair or minor damage. 1 represented the building being destroyed by the earthquake, having major damage or the need for reconstruction. The X input for the model only included location and building characteristics. Along with the columns previously mentioned as being dropped, district ID, technical solution proposed, condition post eq, height before and after, floors after, were all not included in the model. The model included 8 variables that would be used to predict the outcome of the earthquake. Pandas get dummies feature was used to encode the categorical variables to be used in the model.

### *B. Experimental Design*

**Table 2: Experiment Parameters**

| **Experiment Number** | **Parameters** |
|---|---|
| 1 | 33% Test Size Split |
| 2 | 25% Test Size Split |
| 3 | 1% Test Size Split |

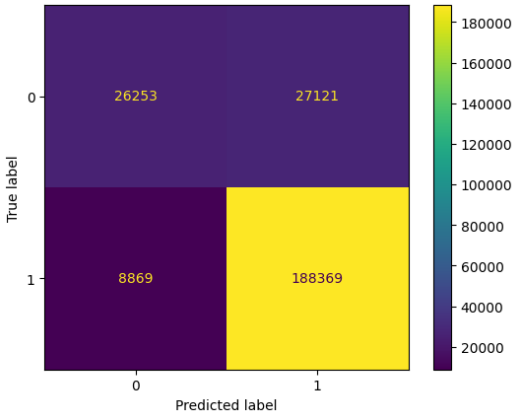
### *C. Tools Used*

The following tools were used for this analysis. Python v3.9.17 running the Anaconda environment for a Windows 11 computer was used for analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 2.1.0, Seaborn 0.12.2, Matplotlib 3.7.2, SKLearn 1.3.2. Pandas was used for working with the csv file. Matplotlib and Seaborn were used for creating visualizations. SKLearn was used to create the logistic regression model and prediction.

## V. RESULTS

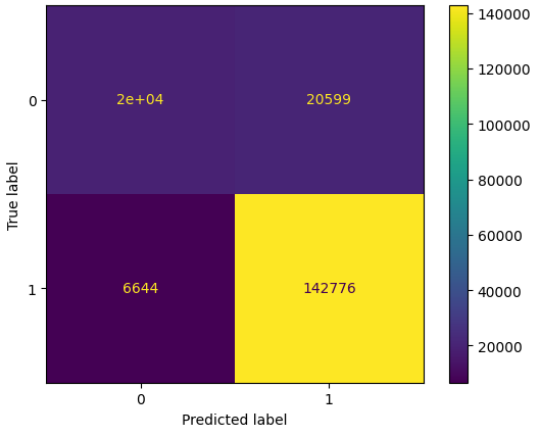
### *A. Classification Measures/ Accuracy measure*

**Figure 5: Experiment 1 Classification Matrix and Report**



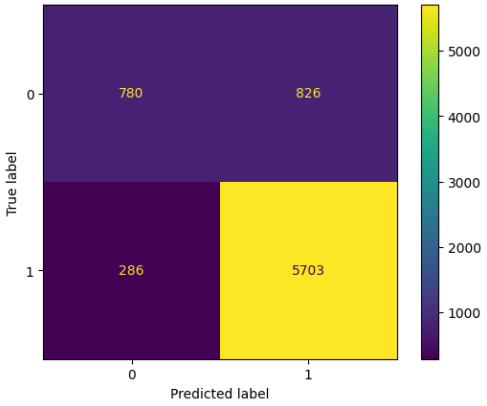

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.75 | 0.49 | 0.59 | 53374 |
| 1 | 0.87 | 0.96 | 0.91 | 197238 |
| accuracy | | | 0.86 | 250612 |
| macro avg | 0.81 | 0.72 | 0.75 | 250612 |
| weighted avg | 0.85 | 0.86 | 0.84 | 250612 |

**Figure 6: Experiment 2 Classification Matrix and Report**




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.75 | 0.49 | 0.59 | 40437 |
| 1 | 0.87 | 0.96 | 0.91 | 149420 |
| accuracy | | | 0.86 | 189857 |
| macro avg | 0.81 | 0.72 | 0.75 | 189857 |
| weighted avg | 0.85 | 0.86 | 0.84 | 189857 |

**Figure 7: Experiment 3 Classification Matrix and Report**




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.73 | 0.49 | 0.58 | 1606 |
| 1 | 0.87 | 0.95 | 0.91 | 5989 |
| accuracy | | | 0.85 | 7595 |
| macro avg | 0.80 | 0.72 | 0.75 | 7595 |
| weighted avg | 0.84 | 0.85 | 0.84 | 7595 |

Despite changes in test size, all three experiments yielded around the same result in the accuracy of the model at 86% for experiments 1 and 2 and 85% for experiment 3. Throughout all of the models, the experiments performed roughly the same when it came to the other measures of precision, recall, and f1 score. This could suggest that the model was fairly successful in making predictions, although there was some room for improvement when it comes to the accuracy.

*B. Discussion of Results*

The results of the experiments were far too close in order to conclude which model is the best in this situation. Since they all performed roughly around the same numbers and there was no significant difference between any of them, it cannot be definitive as to which is the best. For the purpose of using a model to create and validate predictions, I went with the one from experiment 1.

In order to come up with the scenarios to predict, I generated a random sample of three rows in the dataset. Using classifier predict, I inputted the corresponding values of the variables in order to predict the outcome of each scenario. The first two rows that the model was trying to predict were successful. The predicted and actual outcome was 1, or destroyed. As for the third row, the model failed to correctly predict it. The model predicted that the outcome was destroyed but the actual outcome was 0, withstood.

*C. Problems Encountered*

One of the problems that I encountered was the presence of the dummy variables in order to predict and evaluate from the model. In order to make a prediction, you have to input around 30 different numbers, which can be difficult and prone to mistakes. Furthermore, it is unusual that the numbers remain the same although the large differences in the test sizes of each experiment. I began the experiments using a different type of encoding, assigning each unique value a numerical value (for ex. Wood floor – 1, stone floor – 2, and so on…) for each column. The results were roughly the same and the accuracy did not differ across experiments. After switching to one-hot encoding, the results remained and that is what I stuck with for creating the models.

*D. Limitations of Implementation*

One of the major limitations that the model might have is the bias that is given to the destroyed outcome of a building. A majority of the buildings in the dataset were destroyed, and the model that was created may reflect that far too much. For example, how in prediction 3 the actual outcome was that the building withstood the earthquake and the model predicted that it was destroyed. There may be further alterations that need to be made in order to more accurately predict. The dataset could have included more variables in order to help predict better, such as the estimated value that went into creating the building (although that would be difficult to track). Given that I was attempting to predict a binary outcome, a logistic regression model works for the purpose. Further improvements in

the model could address the limitations. One interesting factor that could have been included is a dataset that tracks the characteristics of buildings and the effects that earthquakes had on them in different areas. From there, we could see what level of resistance a building has to different strength levels of an earthquake.

### *E. Improvements/Future Work*

In order to improve the model for future work, one of the things that could be done is better distinction between which variables that could go into the model. For example, the model used both building and land characteristics. One model could use just building characteristics, and another could just use the land characteristics, and one could use both. Different data points that weren't included in the dataset could prove to be useful in creating a better and more accurate model. Also, some variables probably could have been removed depending on what level of resistance they may provide to an earthquake (for example, if the floor type has no effect it should not be included). Further research on earthquakes prior to building a model could be done in order to create a better one.

## VI. CONCLUSION

In conclusion, the model in this project was created in order to understand how a building can withstand an earthquake, such as what characteristics will allow that to happen. Further analysis and model building would be needed in order to answer the question. The question of whether or not a building would survive an earthquake, given certain characteristics, is a smaller question that this project aimed to answer and did so somewhat successfully. In order to answer the larger question at hand, more factors would need to be tracked and further analysis would need to be made. The logistic regression model that was created is good at predicting the outcome of the Nepal earthquake, but it will need more work in order to be a reliable model for predicting beyond that.

## REFERENCES

Nepal earthquake of 2015 | Magnitude, Death Toll, Aftermath, & Facts | Britannica