Predicting Building Damage Caused by the 2015 Gorkha Earthquake

Kevin Wu, Ben Fefferman, Yushu Qiu, Stephanie Wu
University of Chicago
DATA 11900

Abstract

The 2015 Gorkha Earthquake (1) was one of the worst in recent history. In this paper, we will explore data provided by the Nepalese Central Bureau of Statistics to analyze the relationship between various architectural components and building damage (2). We then perform feature engineering (3) to further prepare the data for Multiple Linear Regression (4) and several classification models (5) to classify buildings by damage grade. To conclude, several architectural stability solutions will be offered (6).

1 Introduction and Motivation

Natural disasters pose a persistent threat to communities throughout the world. In the following discourse, we present a data-intensive analysis of the aftermath of the 2015 Gorkha earthquake. With the UN Secretary General's remarks that the Intergovernmental Panel on Climate Change (IPCC) 2021 Report was "a code red for humanity,"(4) we were motivated to use the Gorkha earthquake data to yield important insights for how to build edifices that will withstand an increasingly rugged climate. With an epicenter in the Gorkha district of Nepal, this 7.8 Mw earthquake resulted in over 9,000 casualties, exacted an economic toll of roughly 7 billion USD, and damaged over 700,000 buildings in the vicinity of Gorkha (2). The following analysis considers the structural impact of this earthquake, with the goal of identifying which architectural components are most protective against natural disasters.

1.1 Data Source

Data for our analysis were obtained from the 2015 Nepal Earthquake Open Data Portal (1), which provides data from household surveys and mobile phone records regarding structural information for buildings within the 11 affected Nepalese districts. The original dataset contained 762,106 rows and 31 columns. We removed 12 rows containing missing values and capped several rows with clear misinformation (eg. building age = 999). In this paper, we will begin with an in-depth analysis of the data, focusing upon which variables are most predictive of "damage grade," measured in European Microseismic scale (EMS) from 1 (least damage) to 5 (most damage). Then, we will present results from several machine learning algorithms, including Multiple Linear Regression, Random Forest, K-Nearest Neighbors, and Neural Network, used to predict damage grade. Finally, we will propose strategies to create safer buildings in the future.

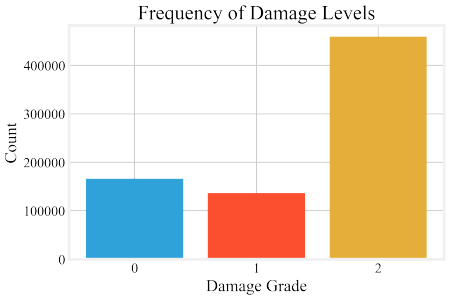

Figure 1: Frequency of each transformed damage grade.

2 Exploratory Data Analysis

To reduced imbalance and facilitate our model creation, **we first re-categorized damage grade into three classes; 0-low (grades 1 + 2), 1-medium (grade 3), and 2-high (grades 4 + 5)**. As seen in Figure 1, there is a class imbalance, with the majority of buildings exhibiting heavy damage. To account for this, we will not only use accuracy as a performance metric for our models in Section 5, but also precision and recall. We categorized potential factors into three groups: the geographical location of buildings, the materials used, and aspects of architectural design. All features pertaining to status post-earthquake (e.g. floors post-earthquake) were not considered.

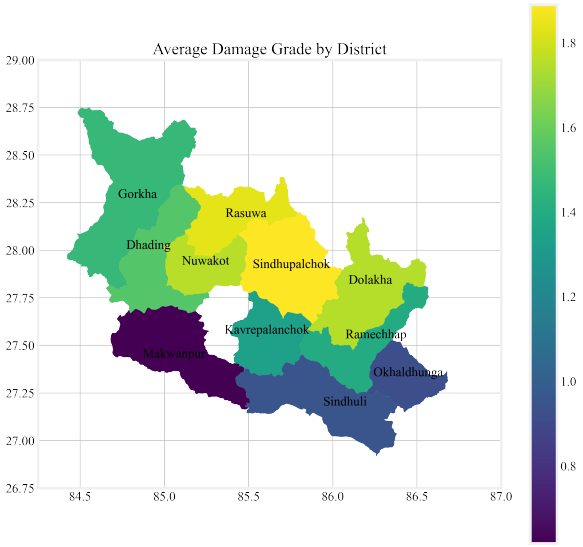

Figure 2: Choropleth of the 11 affected districts.

2.1 Geography

To better understand the geographical distribution of buildings with a high damage grade, we and analyzed the proportion of buildings with each damage rating in each district. As observed in Figure 3, Sindhupalchok had the greatest proportion of buildings with a high damage grade (∼96%), followed by Rasuwa (∼90%), and Nuwakot (∼82%).

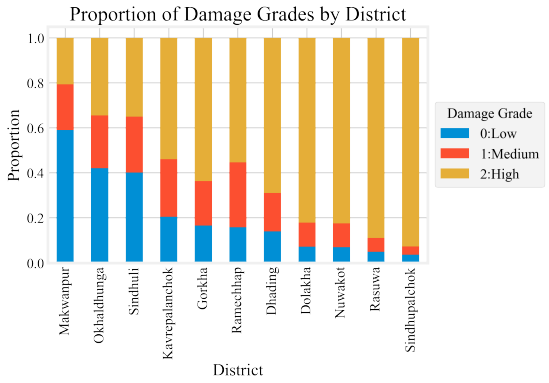

Figure 3: Damage grade distribution in 11 affected districts.

Here, we note that the damage grade distribution by district can be contextualized by the fact that Nepal is bordered on the North by the Himalayan Mountain range. Indeed, Sindhupalchok, Rasuwa, and Nuwakot all lie in the Northern sector of Nepal. Furthermore, Figure 2, which shows average damage grade by district, demonstrates a general increase of average damage grade as one progresses from South to North. Nepal is known for its mountainous terrain with mountains spanning nearly 75% of the country (7). For this reason, altitude and foundation arrangement play a critical part in the structural integrity of Nepal's buildings.

2.2 Building Materials

Ground Floor Type	Average Damage Grade
RC	0.375
Other	0.84
Timber	1.097
Brick/Stone	1.454
Mud	1.5

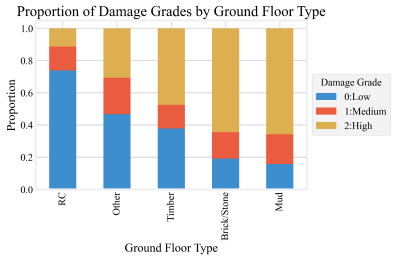

Figure 4: Damage grade stats grouped by ground floor type.

We then proceeded to analyze the impact of building materials on damage grade. One of the most pertinent factors influencing damage grade was the use of reinforced concrete (RC). As observed in Figure 4, homes with RC ground floors had the lowest proportion of high damage grade, (∼10%). In contrast, mud and brick/stone ground floors suffered from much higher proportion of high damage grades (∼65% and ∼62%, respectively). Another important point is that, aside from "Other," timber floors were the second most resilient material after RC, with ∼45% high damage. We see a similar trend in foundation type, roof type, and other floor type; that is, RC was by far the most resilient material.

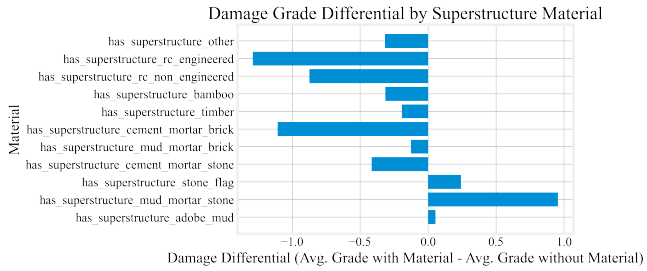

Figure 5: Damage differential for 11 superstructure materials.

The dataset contains 11 binary columns pertaining to "superstructure," (pertaining to the above ground portion of a building) where 1 indicates presence and 0 indicates absence. Thus, we computed a damage differential for each superstructure material, x: average grade with x minus average grade without x. Large negative differentials may indicate that a material is resilient towards damage, whereas large positive differentials may indicate that a material is susceptible to damage. Referring to Figure 5, we found that the most protective superstructure material was engineered RC (-1.29), followed by cement-mortar-brick (-1.11) and non-engineered RC (-0.87). Again, the presence of RC is beneficial. The least protective material was mud-mortar-stone (0.96), which also happens to be the most common material, appearing in 80% of buildings.

2.3 Architectural Elements

Visualization of average damage grade against building age in Figure 6 suggests that the average damage grade experiences a near logarithmic growth as building age increases. The graph has some fluctuations when the building is older than 40 years, which may not conform to the positive correlation. However, the histogram shows that buildings with an age of 40+ years constitute only a small proportion of the data.

Most buildings in Nepal are low lying, between 1-3 floors. Interestingly, there is a stark decline in average damage grade progressing from buildings with 3 floors to 4 floors in Figure 6, despite increased average damage grade as the number of floors increased from 1 to 3. We hypothesize that taller buildings with 4+ buildings are most likely from urban areas and are therefore more likely to contain RC than buildings with fewer floors.

Other continuous features present were plinth area (meaning the building's cover area) and building height. As both variables increased, so did damage grade, but

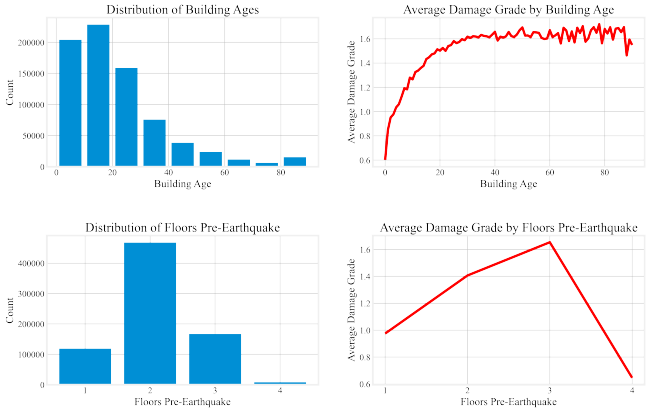

Figure 6: Plots for building age and floors pre-earthquake.

this trend was only stable at low values (until ∼25 ft. for building height and 500 sq. ft. for plinth area). After that, there is a large amount of variability.

3 Feature Engineering

We will create two new features that combine several features related to building materials previously explored in Section 2.3. This feature engineering will especially benefit the computationally intensive K-Nearest Neighbors model in Section 5.2.

3.1 Reinforced Concrete

As seen in Section 2.3, RC is the most robust building material. Hence, we engineered a new feature that counts the number of RC components in each building, aggregating foundation type, roof type, ground floor type, other floor type, engineered RC superstructure, and non-engineered RC superstructure. Values, then, range from 0-6. The proportion of high damage grade steadily decreases from ∼65% with no RC components down to ∼2% for 5-6 RC components.

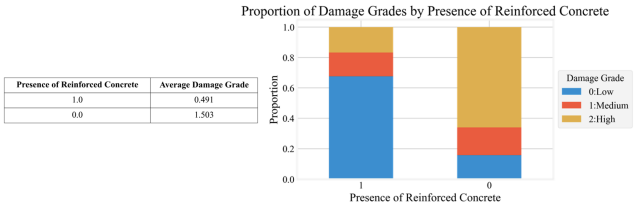

Figure 7: Damage grade stats grouped by reinforce concrete.

The problem with this feature is that it is very imbalanced, for example, only 211 buildings have all 6 RC features, under 0.0003% of all buildings. Thus, we created another feature that indicates the presence of RC with a 1 and the absence with a 0. This feature is much less imbalanced, with over 11.6% of buildings containing at least one RC component. In Figure 7, we see that buildings with RC have a much lower average damage grade (0.491) than buildings without (1.503).

3.2 Superstructure Material

Since the two RC superstructure features are already accounted for in the the reinforced concrete feature, we will leave them out of this feature. The other two most impactful superstructure materials (and also most frequently occurring) are cement-mortar-brick, which is damage-resilient, and mud-mortar-stone, which is damage-susceptible. We will engineer a new feature called "superstructure" which takes value 0 for neither/both, 1 for cement-mortar-brick, and 2 for mud-mortar-stone. The average damage grade for is 0.314 for 0, 0.786 for 1, and 1.578 for 2, as seen in the table in Figure 8.

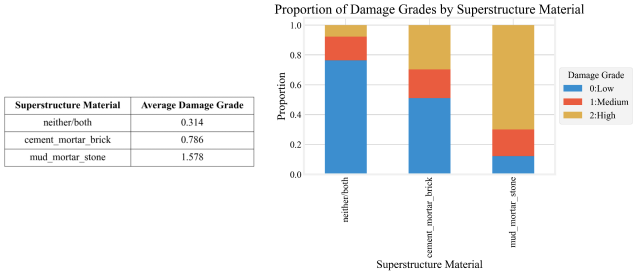

Figure 8: Damage grade stats grouped by superstructure.

4 Regression Model

To automate feature selection and determine the most salient structural elements for predicting damage grade in a high-throughput manner, we proceeded to construct several machine learning models. First, we employed Multiple Linear Regression with Elastic Net Regularization to our structural dataset, which uses a weighted sum of the L1 norm ("LASSO" penalty) and L2 norm ("Ridge" penalty) to exchange added bias with respect to the training dataset for reduced variance and enhanced generalization to unseen data. This method also asymptotically shrinks and/or eliminates (by setting to zero) the coefficients of inconsequential predictors that are merely fit to noise. This model is thus highly interpretable, as nonzero coefficients will have been automatically selected as features accounting for a significant proportion of variation in the dependent variable (e.g. damage grade).

After cleaning the dataset, removing a small number of variables known to have no relation to damage grade (e.g. building ID numbers) and one-hot encoding all categorical variables, rows were shuffled, and 80% of buildings were randomly selected for use in the training set, and 20% were retained for testing. Among the 80% of buildings used to construct the model, 10-fold cross validation was iterated over a random search of λ and α hyper-parameters until the change in the loss function was less than a pre-set threshold.

4.1 Results

Overall, this regression model yielded an R^2 coefficient of determination of approximately 0.914, indicating that roughly 91.4% of variation in damage grade could be accounted for by variation in nonzero predictors. Most significantly, upon extraction of the column names from all nonzero predictors, we determined that the number of floors and height before and after the earthquake, the age of the building (in years), condition post-earthquake, and proposed technical solution (e.g. minor or major repair) constituted the remaining, and thus the most salient, predictors.

5 Classification Models

Finally, we will implement three classification models: Random Forest Classifier, K-Nearest Neighbors Clustering, and Neural Network. For each, we will split the data into training and test sets (70-30), select input features, select hyper-parameters using 5-fold cross validation, and evaluate the model on the test set using accuracy, precision, and recall metrics.

5.1 Random Forest

The Random Forest Classifier (6) constructs a "forest" of decision trees on random subsets of the dataset. Each decision tree has one output, the class, and the Random Forest outputs the mode of all decision tree outputs. To clarify, a decision tree is a map of the possible outcomes (on the leaf nodes) of a series of related choices.

We selected the Random Forest Classifier due to its versatility, simplicity of hyper-parameters, high performance, and ability to view a feature importance plot as in Figure 9. One problem often associated with Random Forests is overfitting and time complexity (6); hence, we dropped several columns, including the 11 "superstructure" columns which were consolidated into two through feature engineering in Section 3.

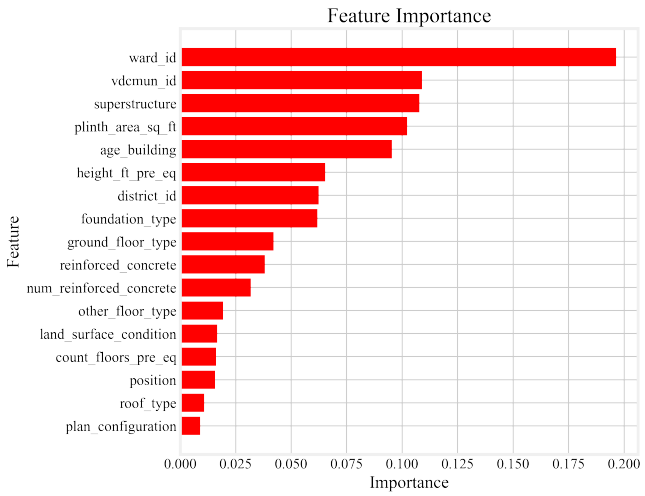

Figure 9: Feature importance in our Random Forest Model.

To prevent overfitting, our model was cross-validated on the following hyper-parameters: "n_estimators", which is the number of trees built before taking the maximum voting, "max_features," which is the maximum number of features random forest considers to split a node, and "max_depth," which is the maximum depth of each tree (3). After 5-fold cross validation on several options for each, the best hyper-parameters were max_features:'log2', n_estimators:30, and max_depth:20.

5.2 K-Nearest Neighbors

K-Nearest Neighbors is a supervised algorithm which classifies points based on a majority vote among its K closet points. In our model we select the Euclidean distance metric, defined as $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ for n-dimensional data. We select this algorithm for its simplicity and effectiveness.

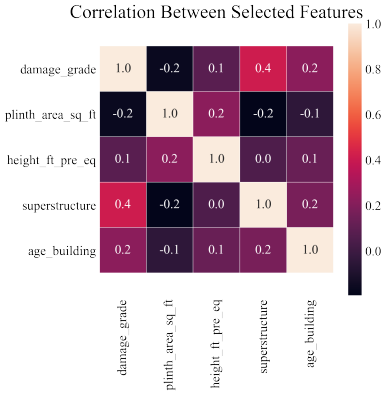

Figure 10: Correlation matrix for our selected KNN features.

Because the distance between each test point and each training point must be computed, low-dimensional data is favored for computational efficiency. Low dimensional data is also favored due to the curse of dimensionality, which states that all objects appear sparse in high dimensions. Thus, we select four features based on our data exploration, the Random Forest feature importance plot in Figure 9, and a correlation matrix between all features, which can be found in the corresponding notebook for this paper. The four selected features are plinth area, building height, building age, and our engineered superstructure feature. The correlation between these features and damage grade is visualized in Figure 10. After cross-validation, we find that $K = 100$ gives the best results.

5.3 Neural Network

A Neural Network was selected because of its high performance, fast computation time, and ability to self-select important features through combinations of matrix multiplications to compute the optimal weights an biases. Each non-input node passes a weighted sum of its inputs plus a bias: $f(b + \sum_i w_i x_i)$.

For our problem, we implemented a Sequential Neural Network with two hidden layers with the *ReLu* activation function, $f(x) = \max(0, x)$ and *Softmax* for the outputs. We round $(\frac{2}{3} * N_i) + N_o$ as a rule of thumb to estimate the number of neurons per hidden layer (5). For our data, $N_i = 28, N_o = 3$, so this value is 22. We use *categorical crossentropy* as the loss function and *Adam*, an efficient stochastic gradient descent method, as the optimizer.

5.4 Results

The results of the three models after predicting on the test set are summarized in Figure 11. The Random Forest Model achieved the best accuracy, **0.748** and precision, **0.718**. Neural Networks achieved the best recall, **0.763**. *K*-Nearest Neighbors performed the worst in all three metrics. Due to the class imbalance, precision and recall, which are defined in Figure 12 may be better metrics than accuracy.

Model	Accuracy	Precision	Recall
Random Forest	0.748	0.718	0.748
K-Nearest-Neighbors	0.688	0.622	0.688
Neural Network	0.71	0.641	0.763

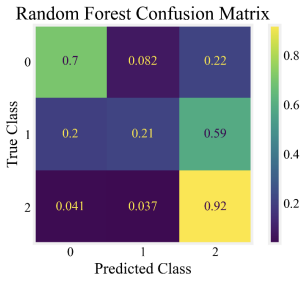

Figure 11: Results summary for all three models and confusion matrix for Random Forest Model.

The confusion matrix in Figure 11 gives more insight into our results for the overall best model, Random Forest. It performs well for low and high damage grades, predicting low when the true class is low 70% of the time and predicting high when the true class is high 92% of the time. However, the model only predicts medium for true medium 21% of the time, which is poor. Most true medium points are predicted as high (59%).

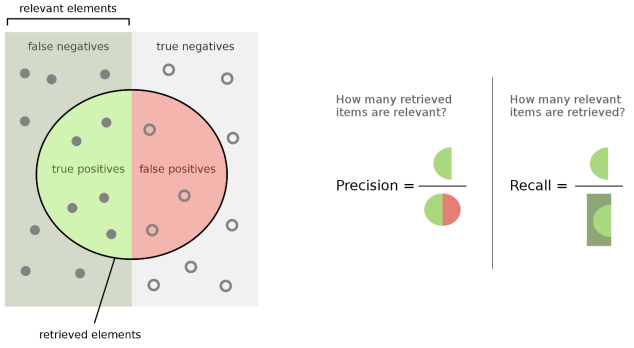

Figure 12: Graphic of precision and recall.

6 Conclusion

In conclusion, our exploratory data analysis and machine learning models have yielded several insights regarding earthquake resilience. First, the geography around a given building appears to have a significant impact on damage grade, with mountainous and coastal regions most susceptible to damage. Second, reinforced concrete components appear to add substantial protection against structural damage. Third, we acknowledge that reinforced concrete building components are often cost-prohibitive, and literature on use of this material in Nepalese society indicates that RC is not well accepted from a cultural perspective. Therefore, we highlight timber architecture as a practical alternative to RC, since we have observed that timber succeeds RC in its degree of protection, but is more protective than mud-mortar-stone materials.

Moving forward, several reforms will protect Nepalese buildings and residents from future disasters. First, we propose that RC components be integrated into vernacular (traditional) architecture via "re-engineering" by adding RC "buttresses, corner posts, corner ties, and wall braces" (2). Additionally, from a policy perspective, the Nepalese government has implemented "Mandatory Rules of Thumb" (MRT) as a form of loosely enforced building and zoning codes that have been heterogeneously implemented, focusing on urban areas to the exclusion of rural areas (2). We encourage residents and the international community to advocate for stricter and more consistently enforced MRT's to create more resilient buildings.

Finally, we note that these reforms can be applied toward increasingly frequent natural disasters due to global warming. Several structural adjustments mentioned in literature are in accordance with the resilient structures identified from our own analysis, including the use of reinforced concrete foundations atop a layer of sand to dampen vibration from earthquakes (2). Overall, we have shown that a data-intensive structural analysis can facilitate safer, more effective future structures.

References

[1] "2015 Nepal Earthquake Open Data Portal," *Government of Nepal Central Bureau of Statistics*, 2016.

[2] D. Gautam, J. Prajapati, K.V. Paterno, et al., "Disaster resilient vernacular housing technology in Nepal," *Geoenviron Disasters* vol. 3, no. 1, 2016.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *JMLR* vol. 12, pp. 2825-2830, 2011.

[4] "IPCC report: 'Code red' for human driven global heating, warns UN chief," *United Nations*, 2021.

[5] J. T. Heaton, *Introduction to Neural Networks with Java*. Heaton Research, Inc, 2005.

[6] L. Breiman, "Random Forests," *Machine Learning* vol. 45, pp. 5-32, 2001.

[7] R. R. Proud, "Nepal," *Britannica*, 2022.