Analysis of Nepal Earthquake Damage

Nitish Goyal, April 2018

Executive Summary

This document presents an analysis of data concerning buildings that were damaged in the earthquake that hit Nepal in 2015. The analysis is based on 20,000 observations of building data, each containing specific characteristics of a building.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, potential relationships between building characteristics and building damage were identified. After exploring the data, a predictive model to classify buildings into three categories (based on severity of damage to a building) was created based on the features of a building.

After performing the analysis, the following conclusions were made:

While many factors can help indicate the severity of damage to a building, significant features found in this analysis were:

- *foundation_type*
- *roof_type*
- *ground_floor_type*
- *age*
- *height*

In the end, a recommendation has been made to group the values of categorical features providing geographical information of a building. The grouping should be done based on the distance of a geographical region from the epicenter. However, information about the epicenter's geographic location and the distance of all geographic regions from the epicenter will be required for this task.

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 10,000 observations in the training data are shown here:

Column	Min	Max	Mean	Median	Std Dev	Dcount
count_floors_pre_eq	1	9	2.147	2	0.7364	8
age	0	995	25.394	15	0.6448	31
area	6	425	38.438	34	0.2127	158
height	1	30	4.653	5	0.0179	18
count_families	0	7	0.985	1	0.0042	8

Note: The data has a column *building_id* which is a random and unique identifier. This column does not provide any information about the building, it is only used to identify observations uniquely. Thus, this column has not been described in this section and also not used for modeling.

In addition to the numeric values, the building observations include categorical features. Categorical features with more than 2 possible values are listed below:

- *geo_level_1_id* - largest geographic region in which the building exists
- *geo_level_2_id* - sub region of *geo_level_1_id* in which building exists
- *geo_level_3_id* - sub region of *geo_level_2_id* in which building exists
- *land_surface_condition* - d502, 808e, or 2f15
- *foundation_type* - 337f, 858b, 6c3e, 467b, or bb5f
- *roof_type* - 7e76, e0e2, or 67f9
- *ground_floor_type* - b1b4, b440, 467b, e26c, or bb5f
- *other_floor_type* - f962, 9eb0, 441a, or 67f9
- *position* - 3356, bfba, bcab, or 1787
- *plan_configuration* - a779, 84cf, 8e3f, d2d9, 3fee, 6e81, 0448, 1442, or cb88
- *legal_ownership_status* - c8e1, cae1, ab03, or bb5f
- *damage_grade* – 1, 2, or 3

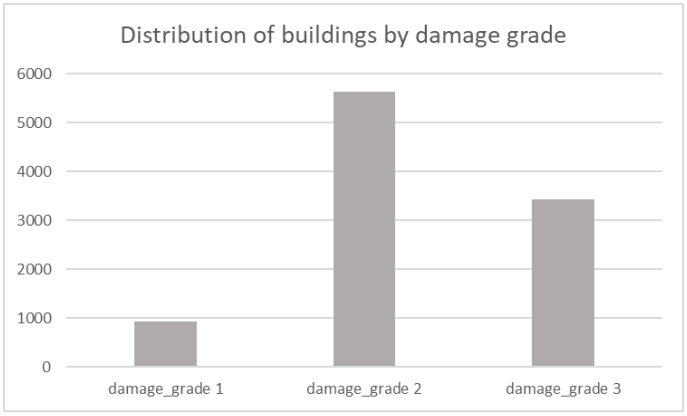
Bar charts were created to show frequency of these features, and indicate the following:

- The vast majority of buildings are in the surface condition of land d502.
- The vast majority of buildings have their foundation type 337f.
- Most of the buildings have roof type 7e76, while roof type 67f9 is very rare.
- The vast majority of the buildings have the ground floor type b1b4. Also, the ground floor types bb5f and e26c are extremely rare.
- Most of the buildings have other floor type f962.
- Most of the buildings have position 3356, while position bcab is rare and position 1787 is extremely rare.
- The vast majority of buildings have the plan configuration a779, and all other plan configurations are extremely rare.
- The vast majority of buildings have the legal ownership status c8e1, and all other legal ownership status are extremely rare. This feature is unlikely to affect the damage severity of a building. The feature importance was calculated for this feature and this feature was found out to be redundant. Therefore, this feature was not used in modeling.

The data also has categorical features with only 0(false) or 1(true) as the possible values (binary features). These features provide information about either the material used for construction of a building or the usage of a building for secondary purpose. Some of the key binary features are listed below:

- *has_superstructure_adobe_mud*
- *has_superstructure_mud_mortar_stone*
- *has_superstructure_stone_flag*
- *has_superstructure_cement_mortar_stone*
- *has_superstructure_mud_mortar_brick*
- *has_superstructure_cement_mortar_brick*
- *has_superstructure_timber*
- *has_superstructure_bamboo*
- *has_superstructure_rc_non_engineered*
- *has_superstructure_rc_engineered*
- *has_superstructure_other*

Since damage grade is of interest in this analysis, a bar plot with distribution of buildings was made.



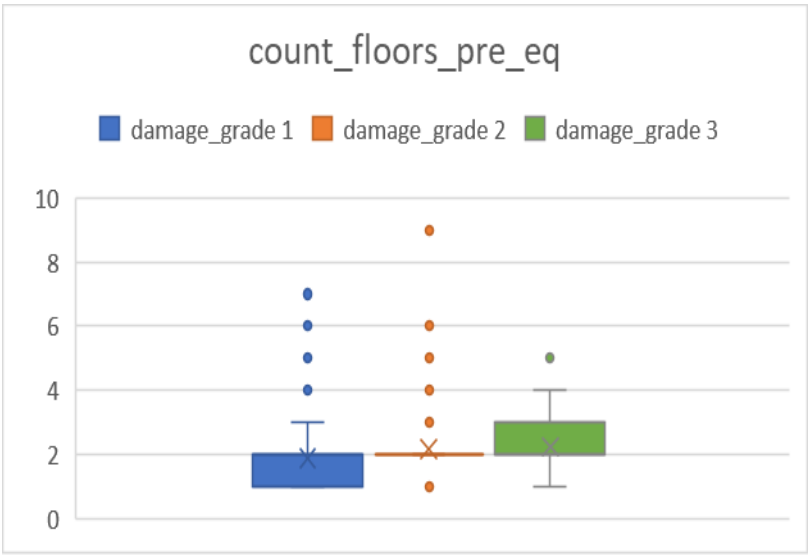
The plot shows that a vast majority of the buildings are in the groups damage_grade 2 and damage_grade 3.

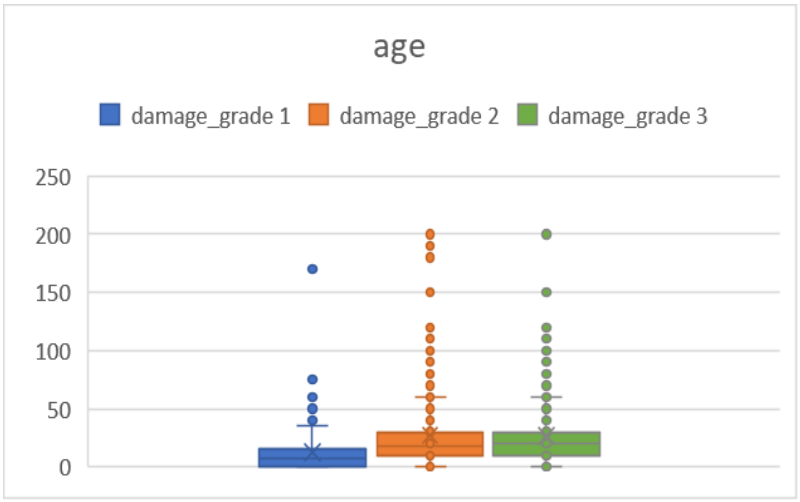
Apparent Relationships

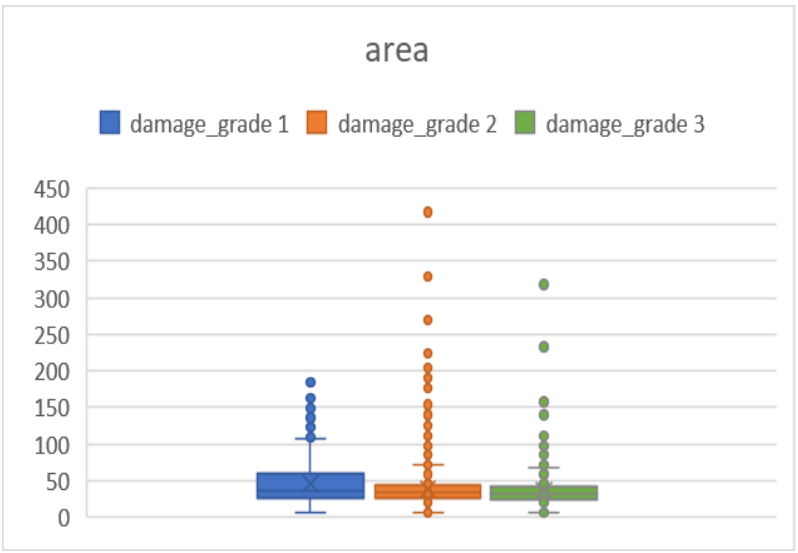
After exploring the individual features, an attempt was made to identify relationships between *damage_grade* and other features. (Please note that the y-axis for all the plots in this section is count of buildings)

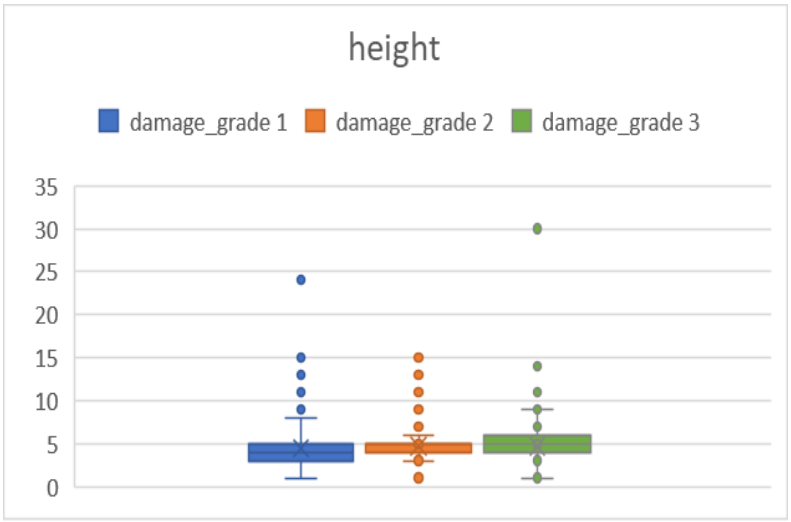
Numeric Relationships

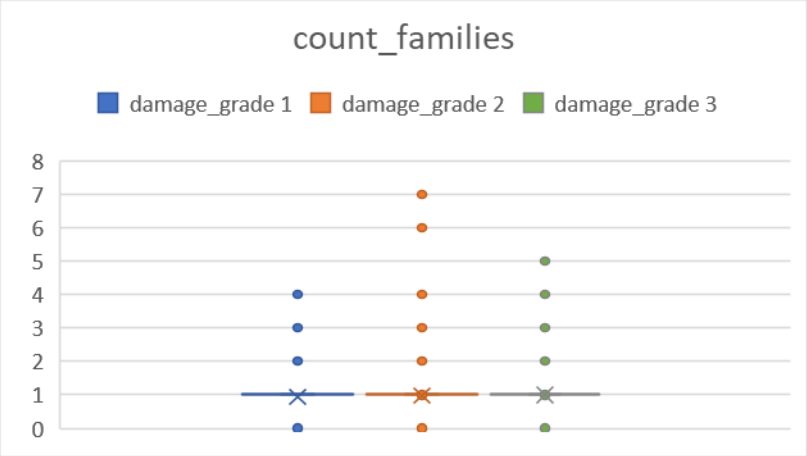
The following box-plots were generated initially to find out any apparent relationships between numeric features and damage grade.




age
damage_grade 1
damage_grade 2
damage_grade 3
250
200
150
100
50
0


area
damage_grade 1
damage_grade 2
damage_grade 3
450
400
350
300
250
200
150
100
50
0


height
damage_grade 1
damage_grade 2
damage_grade 3
35
30
25
20
15
10
5
0



The box plots show some differences in terms of median and range of numerical features for different damage grades. For example:

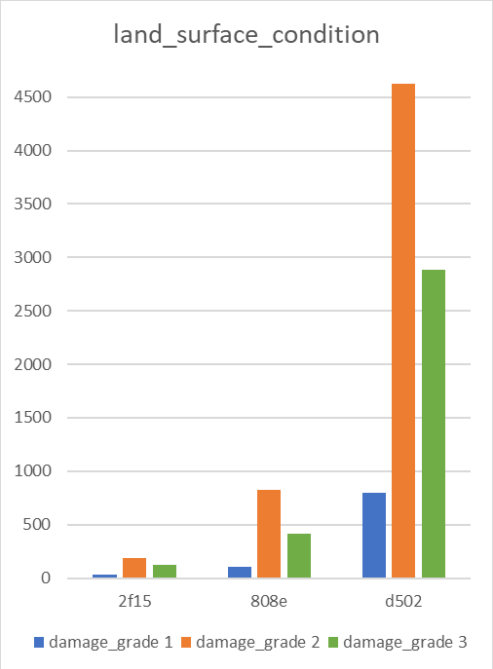
- The interquartile range for number of floors has a slight overlap, but it is evident that the range shifts upwards with damage grade. In other words, buildings with more floors tend to have higher damage grade.
- The median values of age shift upwards with damage grade, but the interquartile range has a fair amount of overlap. Particularly, buildings with damage grade 1 tend to be slightly younger than more severe damage grades (2 and 3).
- The interquartile range and median of area are very similar for all three damage grades. Therefore, it is hard to see any apparent relationship between area and damage grade.
- The interquartile range of height are fairly overlapping for the three damage grades. However, the third quartile of damage grade 3 is higher than damage grade, which means that taller buildings tend to have higher damage grade.
- The interquartile range for number of families overlap significantly and have very low variation for all three damage grades. This implies that number of families is a degenerate feature and does not provide much information about damage grade. It also makes sense, as number of families living in a building is unlikely to affect the severity of damage to a building. Based on this, the *count_families* feature was not used in modeling. This was also confirmed by calculating the feature importance

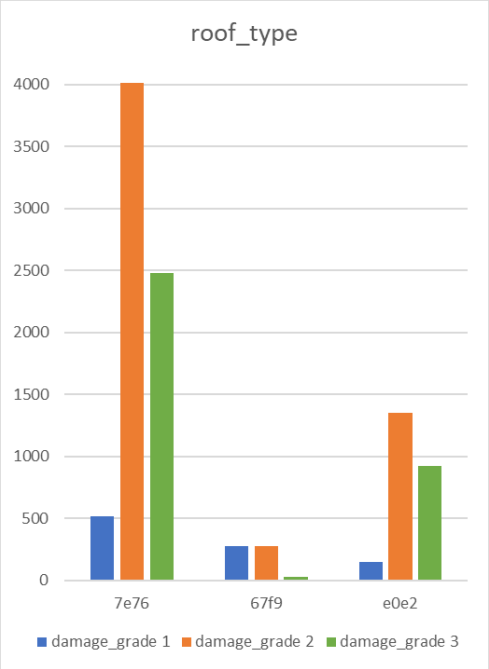
Categorical Relationships

After exploring the relationship between damage grades and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and price.

Categorical features with 3-9 category values

The following bar plots show some categorical columns conditioned by damage grade:



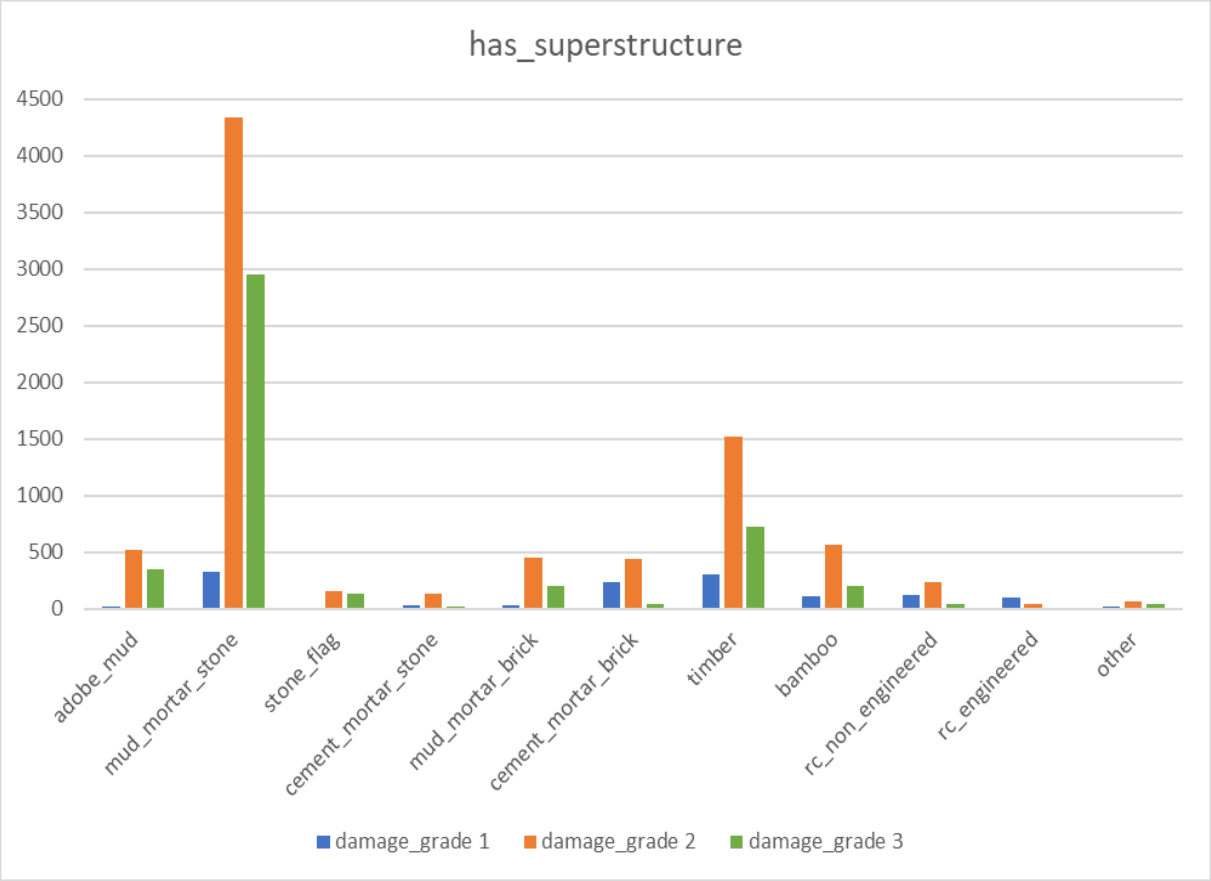


The distribution of buildings belonging to a particular damage grade with respect to the values of the categorical feature was found to be very similar for all damage grades. Similar trends were seen for all other categorical features with more than 2 category values. Therefore, no apparent relationship exists between damage grade and categorical features.

Categorical features with binary values

There are two groups of such features; one group provides information about the material of construction of buildings and the second group provides information about the usage of a building for secondary purposes.

The first group of features that provide information about the material of construction of a building was examined with the help of a bar plot conditioned by damage grade. The bar plot consists of count of buildings based on the material of construction for a super-structure of a building. These features are not exclusive and some buildings have more than one material used for construction of their super-structure.



This bar plot shows no apparent relationship between damage grade and material of construction used for a building super-structure.

The second group of features that provide information about secondary usage of a building is applicable to less than 11% of the training data set. Also, the use of a building for a secondary purpose, for example, usage as a school or rental or government office is highly unlikely to affect the severity of damage to a building. Based on this, this group of features was not used in the model. This was also confirmed by calculating the feature importance for all the features in this group.

Classification of Buildings Based on Damage Grade

Based on the analysis of the building damage grade data, a predictive model to classify buildings into three damage grade categories: 1 (low damage) ,2 (medium damage) and 3 (almost complete destruction) was made.

The model was created using the Multiclass Decision Forest algorithm and trained with the training data set of 10,000 points. Testing the model with the testing data set of another 10,000 points yielded the micro averaged F1 score of 0.699. Since the actual labels for the test data set are not available it is not possible to provide any other statistic or visualization related to the performance of the model.

Conclusion

This analysis has shown that the damage grade of a building can be predicted from its characteristics. In particular, the foundation type, roof type, ground floor type, age, height and geographic region (specifically *geo_level_id_2*) have a significant effect on the damage grade of a building. Secondary features, such as material of construction of super-structures can help further classify buildings and determine damage grade groupings to which they belong.

Recommendations

The location of any earthquake is often defined by its epicenter. Epicenter is the point on the earth's surface directly above a focus, the point where an earthquake originates. The effect of an earthquake on the earth's surface depends a lot on the distance from the epicenter. As one moves further away from the epicenter the effect reduces.

In the provided data set, there are three features that give information about the geographic location of a building: *geo_level_id_1, geo_level_id_2* and *geo_level_id_3.* These features are categorical and have a lot of values associated with them: *geo_level_id_1* has 31 distinct values, *geo_level_id_2* has 1138 distinct values and *geo_level_id_3* has 5173 distinct values. Such a large number of values for a categorical feature can lead to sub-optimal performance for the chosen algorithm.

It is possible to reduce the number of values in these three categorical features by grouping them. However, grouping leads to some loss of detail in the information contained. Therefore, grouping was not done.

The recommendation is to provide information about the geographic location of the epicenter and the relative distance of each geographic region from the epicenter. If this information is provided, then the values of the three features that provide information about geographical region could be grouped to reduce the number of distinct values. The grouping should be done in such a fashion that each group contains geographical regions with similar distance from the epicenter. This would reduce the number of distinct values for these three categorical features without causing loss of important information.