Analysis of Earthquake Damage in Nepal

Vas Mandalpu, April 2018

EXECUTIVE SUMMARY

This document presents an analysis of data concerning the damages to buildings caused by the 2015 Gorkha earthquake in Nepal. The data at our disposal are composed of 10000 observations, each representing a building. Each building is described by 40 features that are examined in this report. These attributes mainly consist of information on the buildings' structure, use, localization and their legal ownership.

As first step, Data is explored by calculating summary and descriptive statistics, and by creating visualizations of the data. We identify several potential relationships between building characteristics and level of damage. Finally, a predictive model to classify buildings into three level of damages was created.

After performing the analysis, we present the following conclusions:

While many factors can help indicate the level of damage, significant features found in this analysis were:

- **Geographic regions** in which building exists. We identified several regions where the earthquake has a stronger impact on the buildings.
- **Age** of the building. Recent buildings (<15 years old) are more likely to suffer low damages.
- **Height** of the building. Higher building is more likely to suffer a complete destruction.
- **Area** of the building. A larger area building leads more likely to low damages
- **Compositions of the buildings**. Intermediates and almost complete destruction are mostly buildings made of *mud mortar stone.*

DATA EXPLORATION

Among the 40 features, one of them is a unique identifier called *'building_id'* which will not be analyzed. The feature of interest that we are trying to predict is *'damage_grade'* and corresponds to the level of damage that the building suffered during the earthquake in Nepal in 2015. Three levels of damage are considered:

- Low damage,
- Medium amount of damage,

- Complete destruction.

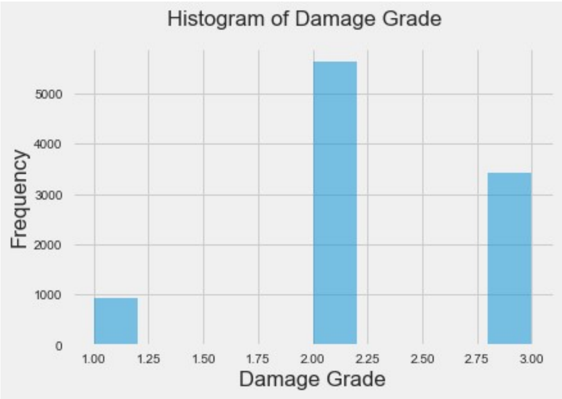
The data are of various types. For this reason, we subdivide this section in four parts:

- Numeric Features
- Geographic Regions
- Categoric Features
- Binary Features

In each subsection, we investigate the data using statistic descriptions and visualization tools.

Furthermore, we try to identify key features to determine what influence the most the level of damage that a building has suffered and also the possible correlations between features.

But first, let’s have a look at the distribution of *‘damage_grade’* by representing its histogram:



We observe an asymmetric distribution of the damages: 56% of the buildings are subject to an intermediate damage (2), 34% totally destructed (3) and the rest subject to low damages (1).

Numeric Features

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 10000 observations are shown here:

	count_floors_pre_eq	age	area	height	count_families
Dcount	8	31	158	18	8
mean	2.146700	25.393500	38.438100	4.653100	0.984600
std	0.736365	64.482893	21.265883	1.792842	0.423297
min	1.000000	0.000000	6.000000	1.000000	0.000000

25%	2.000000	10.000000	26.000000	4.000000	1.000000
50%	2.000000	15.000000	34.000000	5.000000	1.000000
75%	3.000000	30.000000	44.000000	5.000000	1.000000
max	9.000000	995.00000	425.000000	30.000000	7.000000

We note that over 10000 observations we only have 31 different building's ages with a minimum at zero and a maximum at 995. In my opinion, *age = 995* corresponds to a code which indicates that this data is unknown because there are multiple buildings with that age and no other building between 250 and 995. For the following, we'll replace those values by the mean of *'age'*. By doing this, the new mean of age is reduced to 21.5.

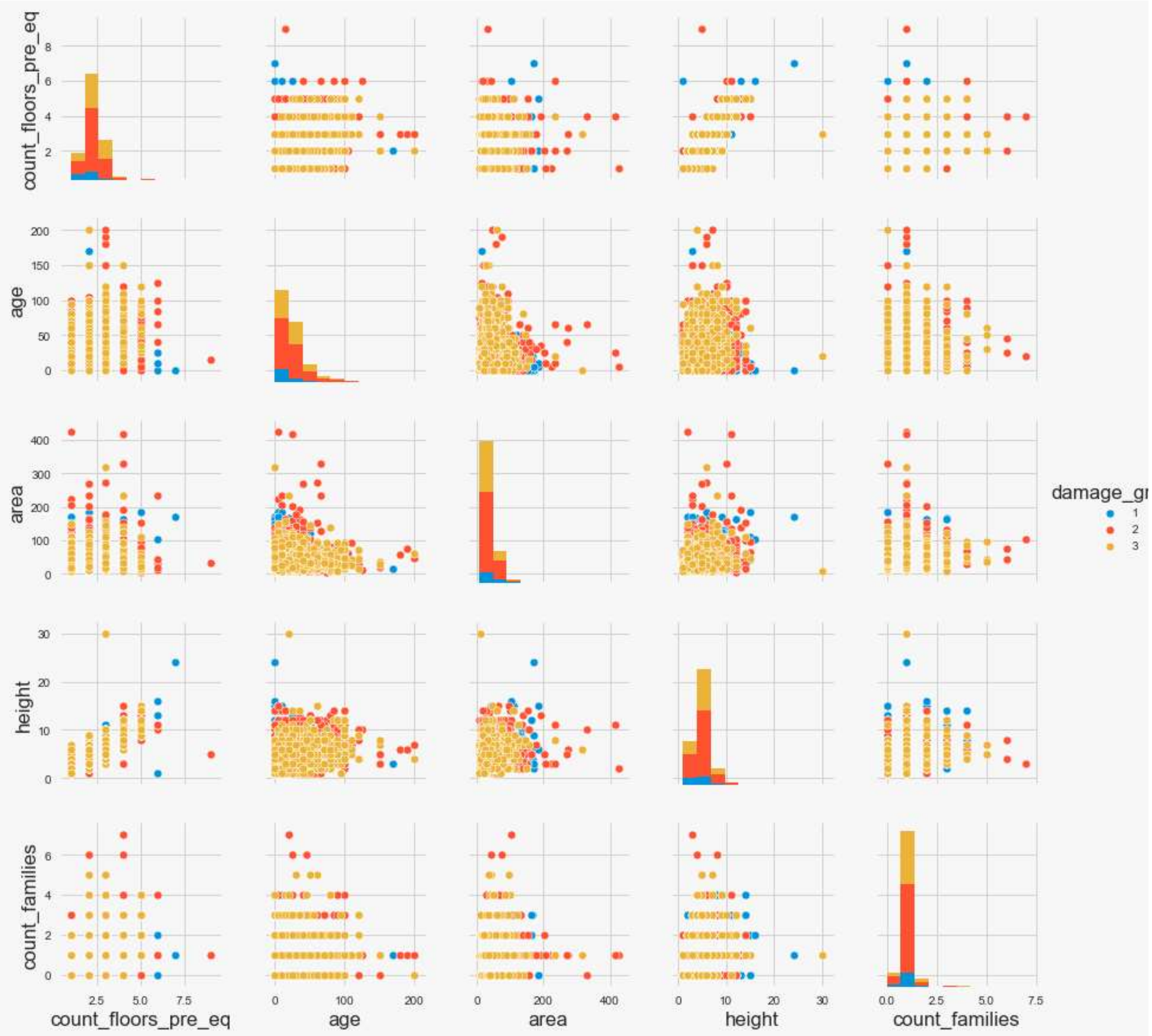
We then investigate the mean by damage grade level and obtain the following results:

Damage grade	**# floors**	**age**	**area**	**height**	**# families**
1	1.8603	10.5576	46.3209	4.4158	0.929638
2	2.1522	22.1286	38.4643	4.6703	0.982434
3	2.2156	23.4603	36.2367	4.6897	1.003211

Interestingly, we note that:

- Buildings with low damages have a significant lower average age value than the two other damage categories,
- Buildings with low damages have higher average area value than the two other damage categories.
- Buildings with low damages have a lower average floors number and height than the two other damage categories,
- The count of families' feature does not seem to be significant to determine the damage grade.

To try to identify the relationships between the numerical features, we provide a scatter-plot matrix for the numerical values with colors corresponding to the level of damage:



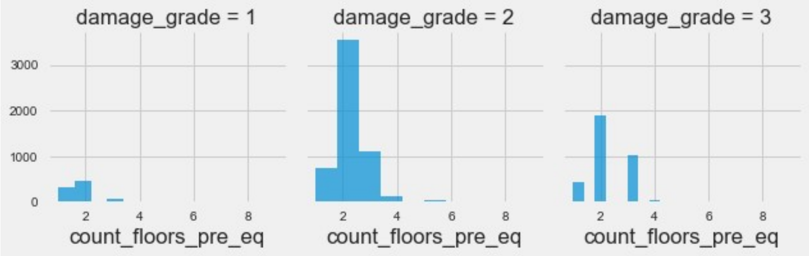
We observe the following properties:

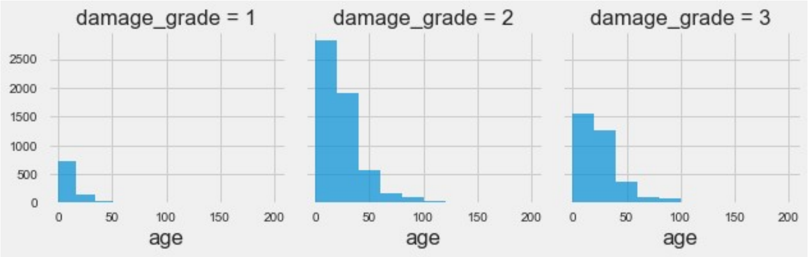
- As can be expected, we note a linear relationship between the height and the number of floors,
- The probability that a building has low damages is strongly influenced by its age,
- Possible mistakes in the data can be seen. For example, we observe a 9-floors building for about 7 m of height, which is not realistic. An example of outliers consists of the few high height/age/area buildings which appear to be isolated data on the different figures.
- A priori, no strong correlations between the other features.

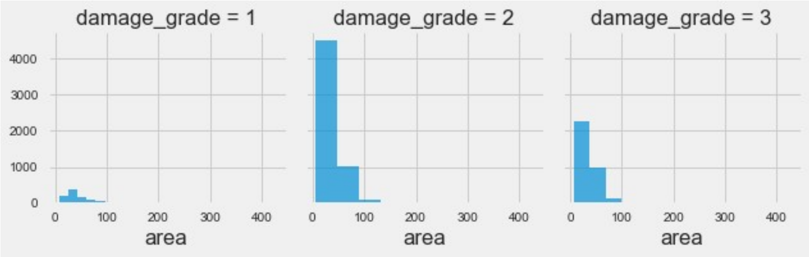
  NOTE: Something is possibly off about the height values or count floors. From the architectural design requirements of the Nepal national building code (http://wbgfiles.worldbank.org/documents/hdn/ed/saber/supporting_doc/SAR/Fina

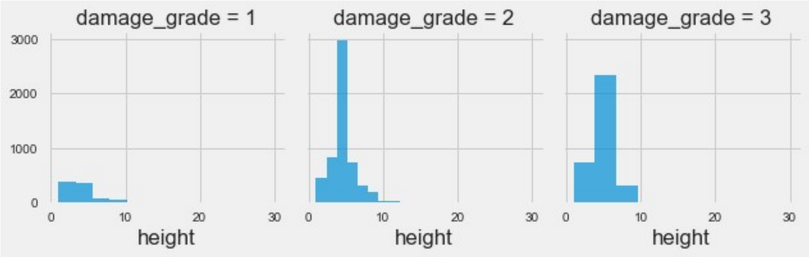
nce/Nepal/Nepal%20Public%20Building%20Codes%202003.pdf), doors must be at least 2m height. If we filter our data by keeping only the ratio *height/count_floors > 2* which corresponds to 2m/floor, our data reduces to 4117 building information.

To understand the distribution of numerical features across the damage grade, we will examine the conditioned histogram.

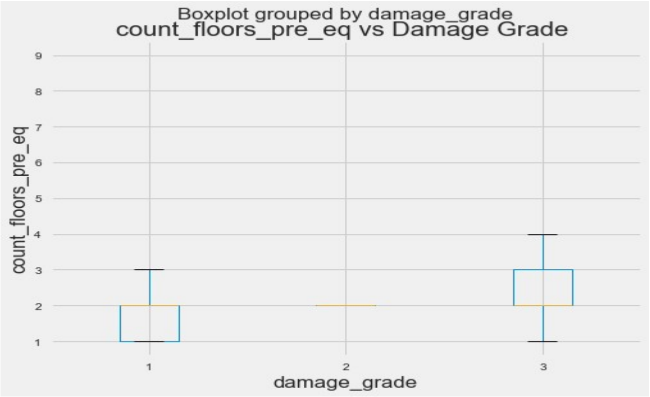


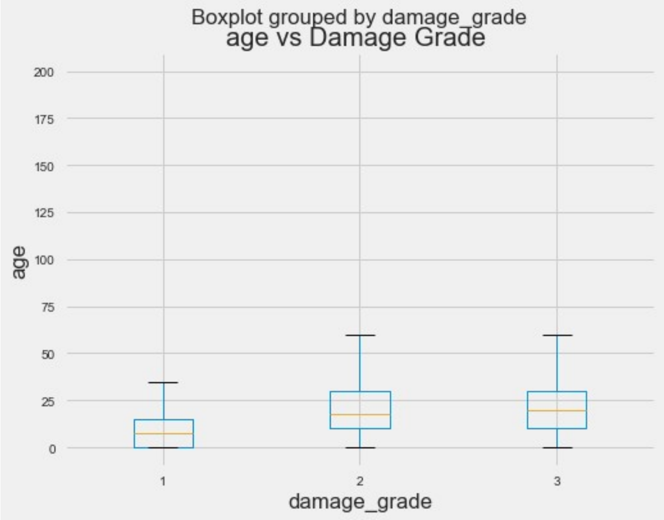


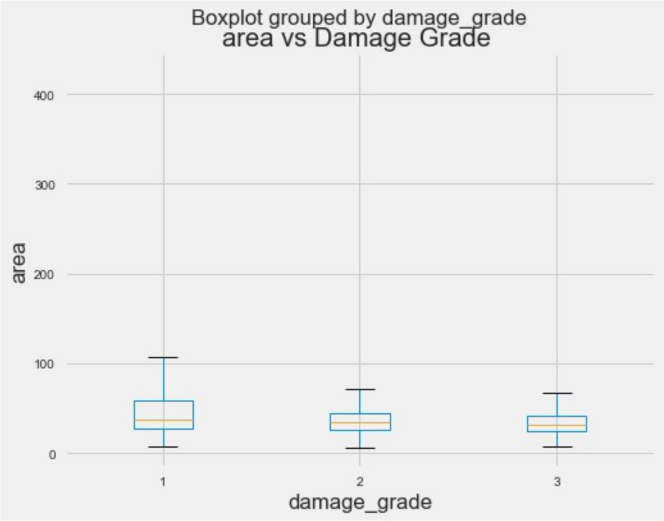


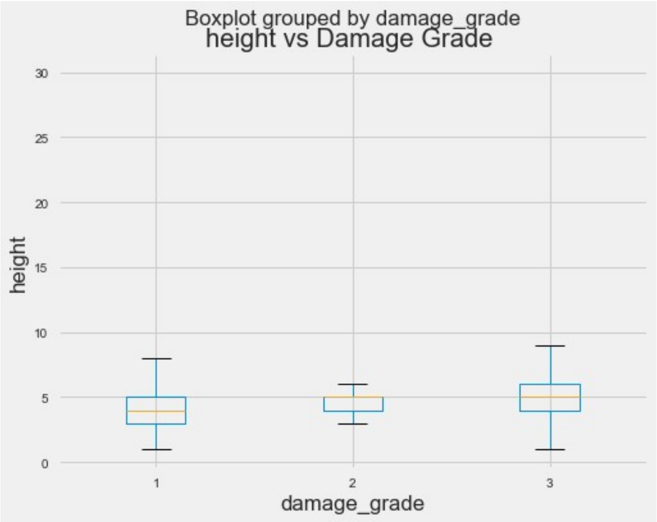


Will also examine the interquartile range with conditioned box plot to understand the outliers in numerical features.









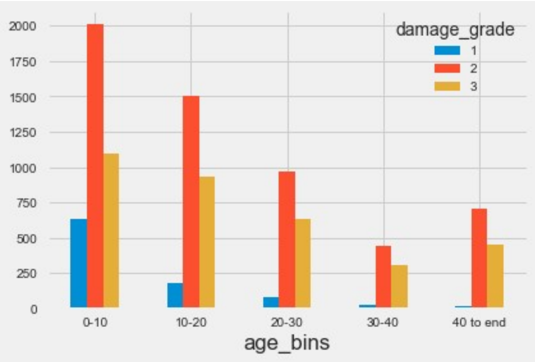
We observe following properties from the interquartile range:

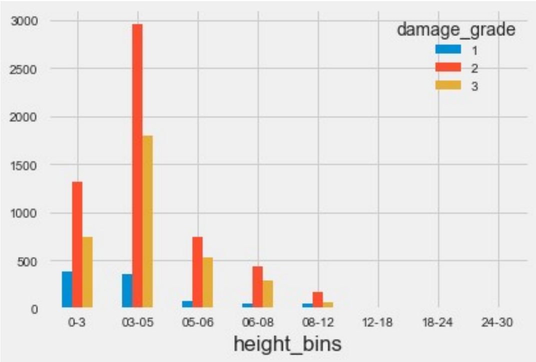
- Most buildings with medium damage grade, have two floors.
- Most building with medium damage grade have narrow IQR for height with values concentrated around 3 to 5
- More number of floors ,greater the probability of complete damage of building.

For the following, we will decompose ‘age’, ‘area’ and ‘height’ in ranges as follow:

- five categories for the age: [0-10], [10-20],[20-30],[30-40] and [40-end],
- seven categories for the height: [0-3], [3-5], [5-6], [6-8], [8-12], [12-18] and [18-end].

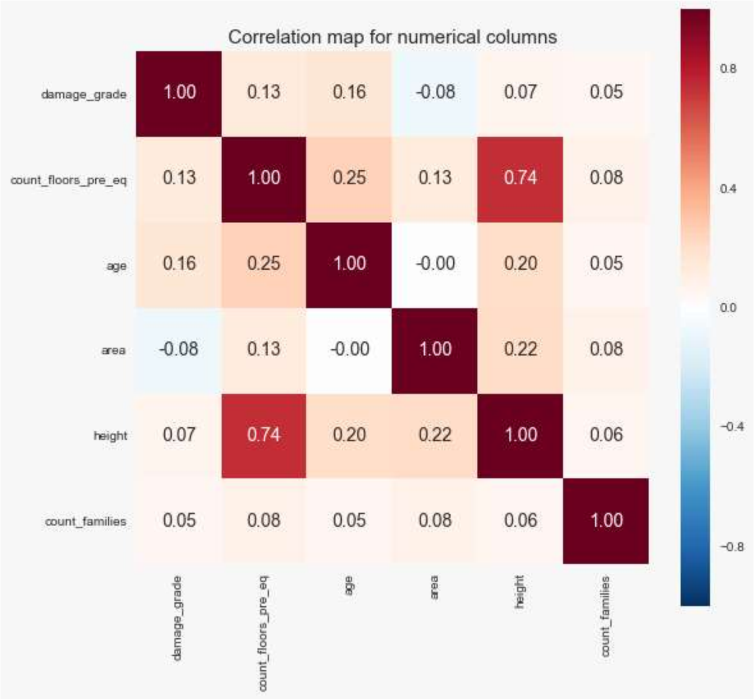
The corresponding distributions are shown in the next histograms:





We note that in this representation, most of the lowly damaged buildings are in the range 0-10.

To further study the correlations, we plot a heatmap of co-relations between numerical features:



It confirms our previous observations about the only highly correlated values being the height and number of floors. We also note a low but not neglectable correlation between area and height.

Geographic Region

The geographic regions in which the buildings exist are given by *'geo_level_1_id', 'geo_level_2 _id', 'geo_level_3_id'* from largest (level 1) to most specific sub-region (level 3). These features are categorical but we can maybe expect a certain order between the different id's. Therefore, I dedicated a subsection for them. We expect the geographic regions to have a high impact on the damages done to the building. Let's first have a look at the statistics:

General information about *geo_level_**XXX**_id*'s:

- *geo_level_1_id*: goes from 0 to 30. Most of the data are located at low id's.
- *geo_level_2_id*: goes from 0 to 1411. Right skewed distribution of the data. There is no building for some of the id's.
- *geo_level_3_id*: goes from 0 to 12151. Right skewed distribution of the data. More than the half of the id's does not contain a building.

	Geo_level_1_id	**Geo_level_2_id**	**Geo_level_3_id**
mean	7	297	2679
std	6	279	2521

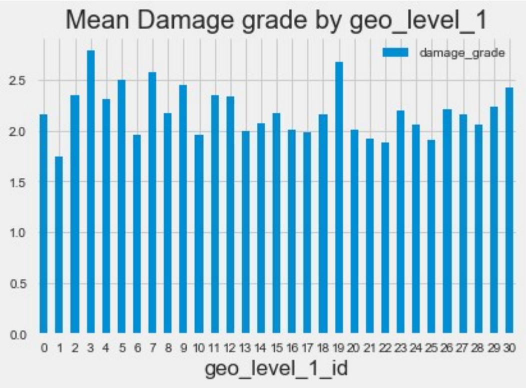
min	0	0	0
max	30	1411	12151
DCount	31	1137	5172

Here are the means of the *Geo_level_**XXX**_id* decomposed by damage grade:

The main possible useful information is the mean by damage grade of *geo_level_1_id*. We believe that the two other geographic regions are inside the *geo_level_1_id* so we can't discuss about the mean values without considering a specific *geo_level_1_id*. For this analysis we will focus only on geo_level_1_ids

damage_grade	**geo_level_1_id**	**geo_level_2_id**	**geo_level_3_id**
1	7	240	2133
2	8	318	2877
3	6	278	2502

We represent the mean damage grade by *geo_level_1_id*:



As expected, most of the mean values are above 2. We highlight the following properties:

- *geo_level_1_id* = 1 has the lowest mean *damage_grade*. We expect in this region to have a high probability of low to intermediate building's damage.
- Buildings with a *geo_level_1_id* = 3, 5, 7, and 19 have a high probability to be completely destructed

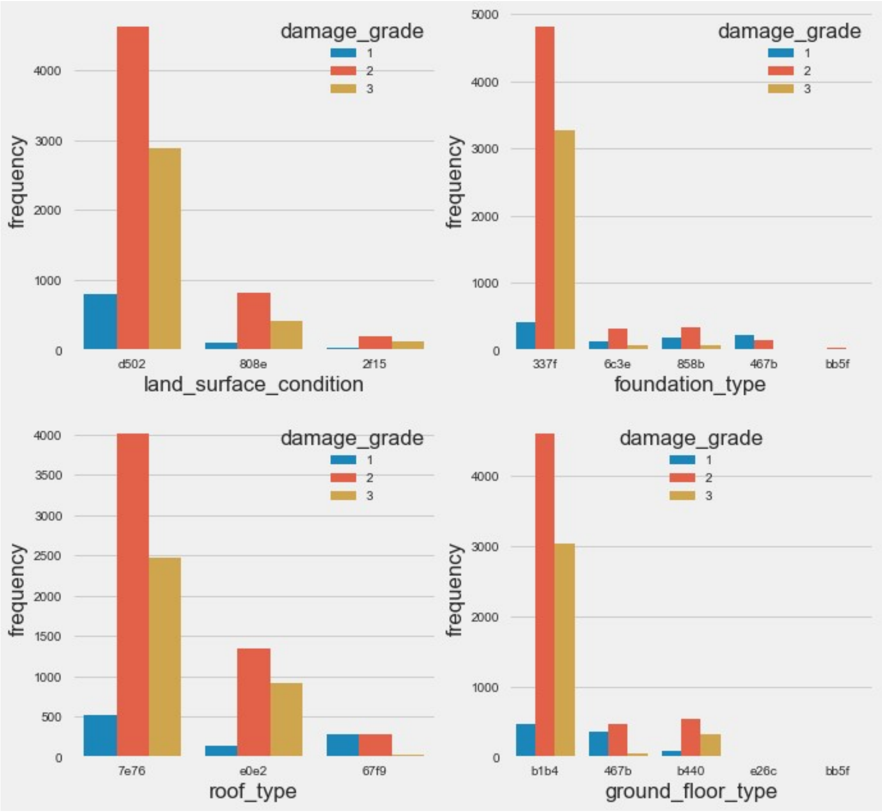
As the geo_lvl_1_id increased, we have a decrease of the number of data. It will probably be a good idea to reduce the number of geo_lvl_1_id in order to decrease the noise in our future predictions.

Categoric Features

In addition to the numeric and geographic values, the observations include categorical features including:

- Land_surface_condition: 3 different surface conditions,
- Foundation_type: 5 types of foundation,
- Roof_type: 3 types of roof,
- Ground_floor_type: 5 types of ground floor,
- Other_floor_type: 4 types of other floor,
- Position: 4 different positions,
- Plan_configuration: 9 plan configurations,
- Legal_ownership_status: 4 different status.

We create bar charts with different colors corresponding to the level of damages to gain some insight:



We observe that the ‘bb5f’ code appears in both *ground floor type* and *foundation type*. Code ‘467b’ appears in *ground floor type* and *foundation type*. Since we do not have any information about the meaning of these codes, we will not attempt to make a link.

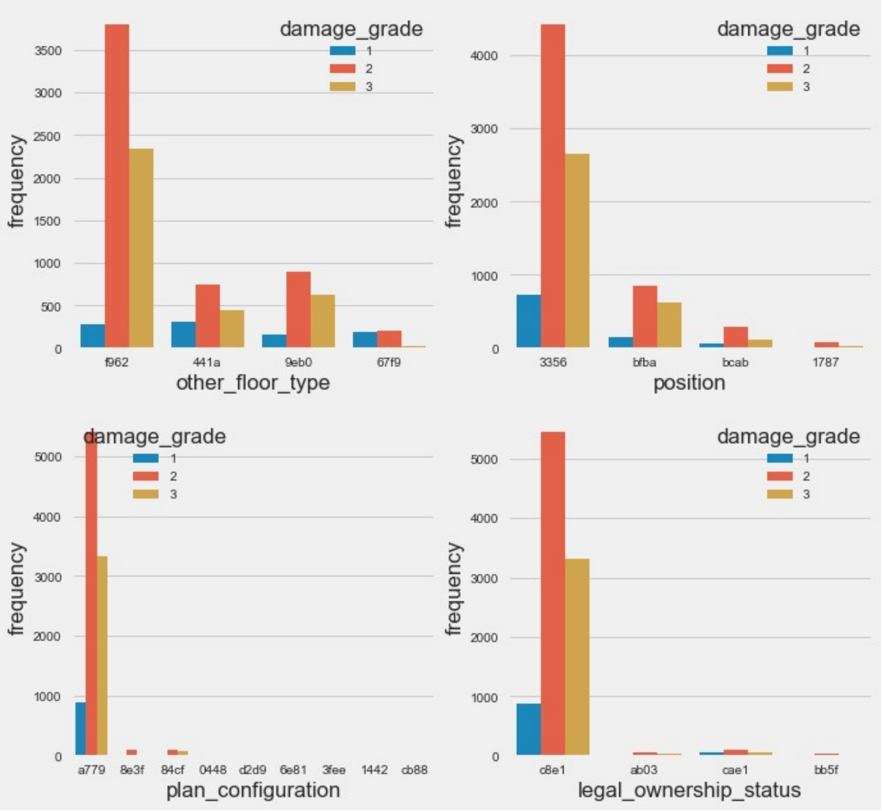
We note the following general properties:

- 83% of the *land surface condition* feature is defined as ‘d502’, only a few data are defined as land surface condition ‘2f15’.
- 85% of the *foundation type* is ‘337f’, only a few data are defined as ‘bb5f’,
- 70% of *roof type* is ‘7e46’, only a few data are defined as ‘67f9’,
- 81% of the *ground floor* are of type ‘b1b4’, only a few data are defined as ‘bb5f’ or ‘e26c’.

Other key properties:

- The *foundation type* ‘467b’ induces more likely low damage to building,
- The *roof type* ‘67f9’ leads to a higher probability of low/intermediate damage to the building than the two other categories,
- The *ground floor type* ‘467b’ have a higher probability to give low/intermediate damage to building.

We next investigate the four other categorical features.



We again note that the code 'bb5f' appears in the *legal owner ship status* feature. The legal ownership status of a land where a building is build has absolutely nothing to do with a ground floor type. It's a trap.

We note the following general properties:

- 64% of the *other floor types* are 'f962', only a few data are defined as '67f9',
- 78% of the building are at the *position* '3356', only a few data are defined as '1787',
- 96% of the *plan configuration* are 'a779',
- 96% of the *legal ownership* status are 'c8e1', only a few data are defined as 'ab03', 'cae1', and 'bb5f'. Other key property:
- The *other floor type* '67f9' leads to a more likely low to intermediate damage.

Binary Features

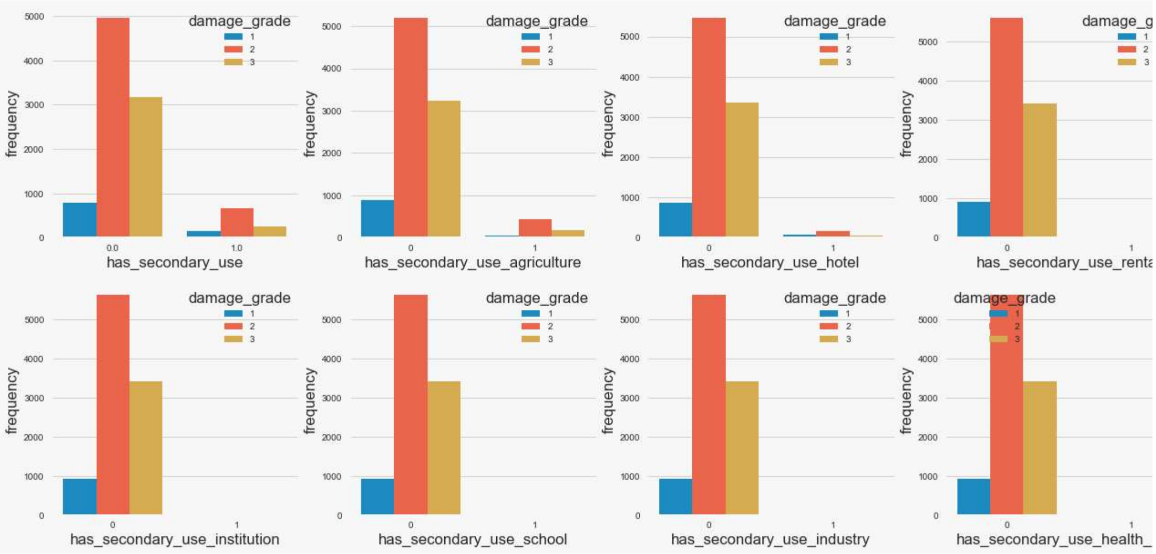
In this section, we investigate the binary features.

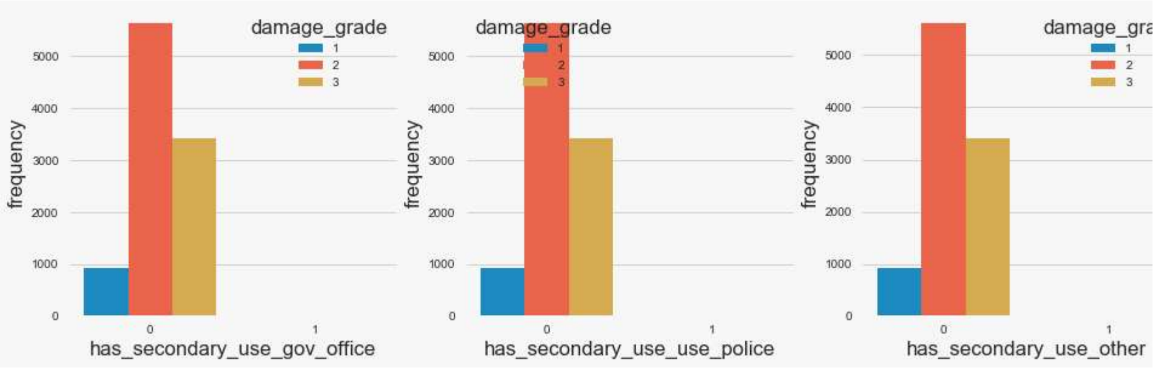
First, we study the features named *'has_secondary_**xxx**',* which corresponds to the use of the building, by giving the mean value of each variable:

Feature	Mean value
has_secondary_use	0.1086
has_secondary_use_agriculture	0.0673
has_secondary_use_hotel	0.0294
has_secondary_use_rental	0.0064
has_secondary_use_institution	0.0007
has_secondary_use_school	0.0007
has_secondary_use_industry	0.0008
has_secondary_use_health_post	0.0002
has_secondary_use_gov_office	0.0002
has_secondary_use_use_police	0.0001
has_secondary_use_other	0.0053

The mean value of a binary feature corresponds to the ratio of 1(Yes) and 0(No). For example, *‘has_secondary_use’* has 10.86% of *Yes* and the rest *No*. Most of the attributes are No. As we see in the above table, we have less than 0.1% of information with most features which is probably not enough to draw conclusions.

We now plot these features as histogram with color corresponding to the damages.





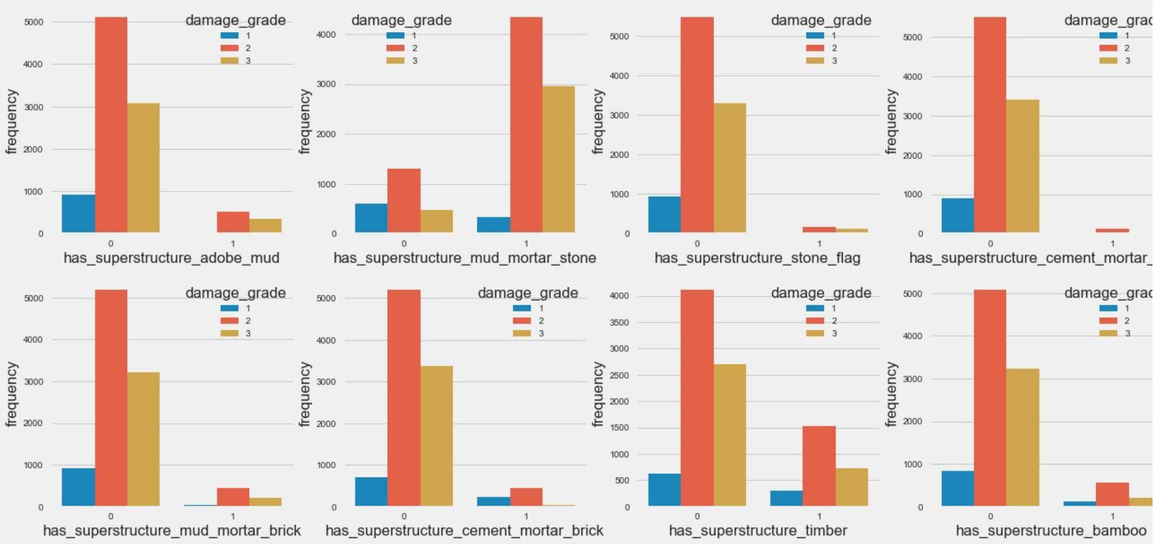
We observe that most of the buildings have no second use. Whatever the feature we consider, buildings with an intermediate level of damage are always more likely. Most of the histograms have the same global shape with almost all values concentrated at 0. Only three features may have significant number of events having a '1', i.e. *'has_secondary_use'*, *'has_secondary_use_agriculture'*, and *'has_secondary_use_hotel'*. Therefore, we consider that the other attributes are not statistically significant to perform an analysis even if we can think that such official buildings may or may not have specific building requirements.

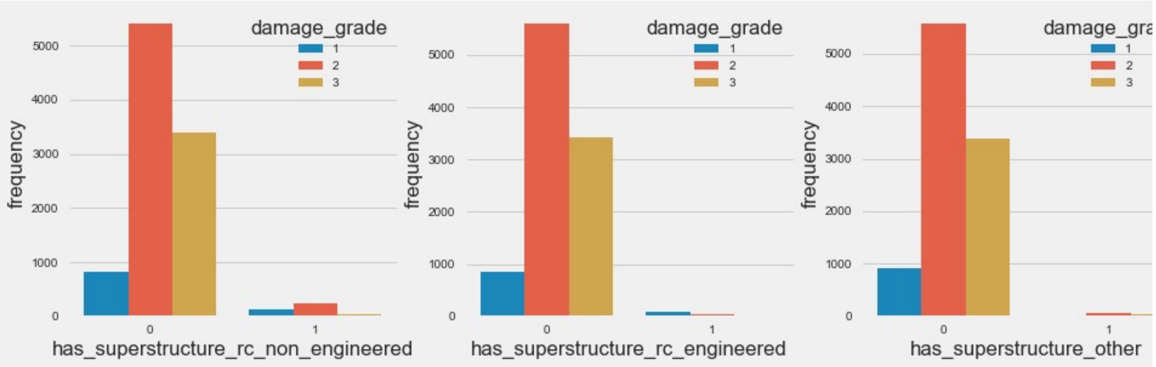
We next investigate the features named *'has_superstructure_**xxx**'*. First by giving the mean value of each attribute:

Feature	**Mean value**
has_superstructure_adobe_mud	0.0897
has_superstructure_mud_mortar_stone	0.7626
has_superstructure_stone_flag	0.0299
has_superstructure_cement_mortar_stone	0.0190
has_superstructure_mud_mortar_brick	0.0688
has_superstructure_cement_mortar_brick	0.0725
has_superstructure_timber	0.2561
has_superstructure_bamboo	0.0877
has_superstructure_rc_non_engineered	0.0400
has_superstructure_rc_engineered	0.0138
has_superstructure_other	0.0141

We note that 76% of the building have a structure composed of mud mortar stone. Building having suffered complete are mostly composed of *mud mortar stone* (86%). Buildings having suffered intermediate damages are also mostly composed of *mud mortar stone* (77%). On the contrary, low damaged buildings are mostly *not composed of a superstructure made of mud mortar stone.*

We now look at damages level distribution when buildings have or does not have a particular feature. Here are the corresponding histograms:





We note that most of the buildings have superstructure composed of mud mortar stone. In general, the probability of having intermediate damages dominates. It is then followed by the probability to have a complete destruction of the building and finally to have low damages.

Here are some observations:

- Buildings composed of *adobe/mud, mud mortar brick, stone flag* and *other superstructures* tend to rarely have low damages,

- Buildings being *rc-engineered* are more likely to low damages,

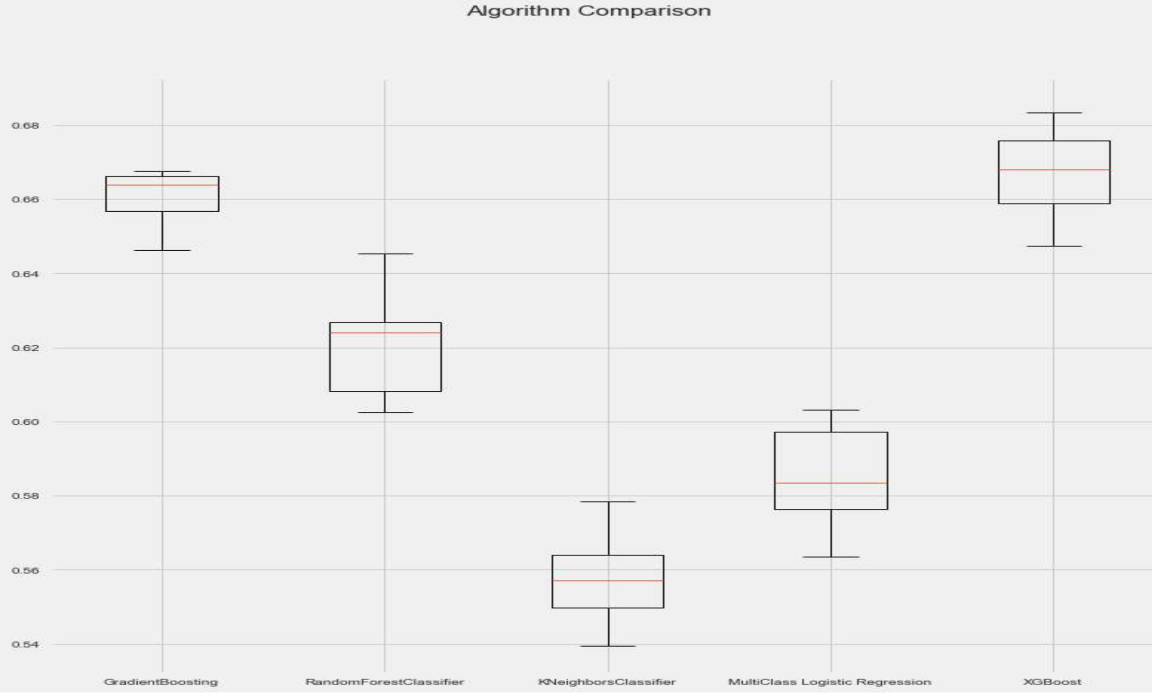
- Buildings with *cement mortar stone superstructure* have a higher probability to be low and medium damaged.

- Buildings without *mud mortar stone* tend to be more likely subject to low and intermediate damages.

CLASSIFICATION OF LEVEL OF DAMAGE

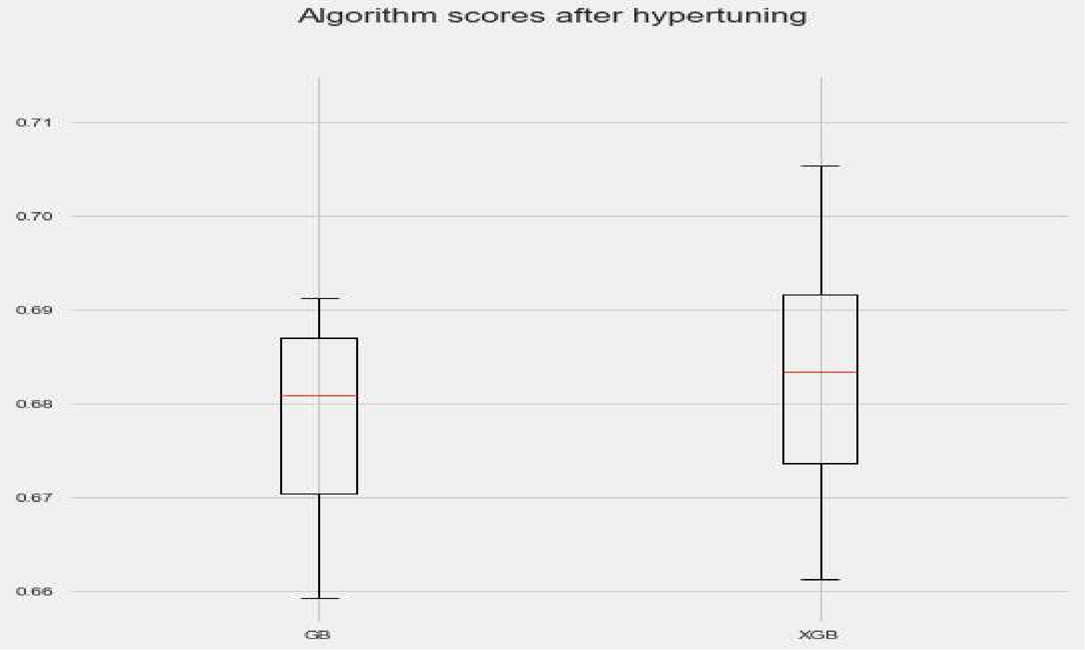
Based on the analysis of the levels of damage, we propose a predictive model to classify the level of damage into three categories: 1, 2, and 3 which represents low damage, medium amount of damage and almost complete destruction, respectively.

As part of developing predictive model , evaluated different algorithms. Here is the boxplot of F1 scores of 10-fold cross-validation for different algorithms

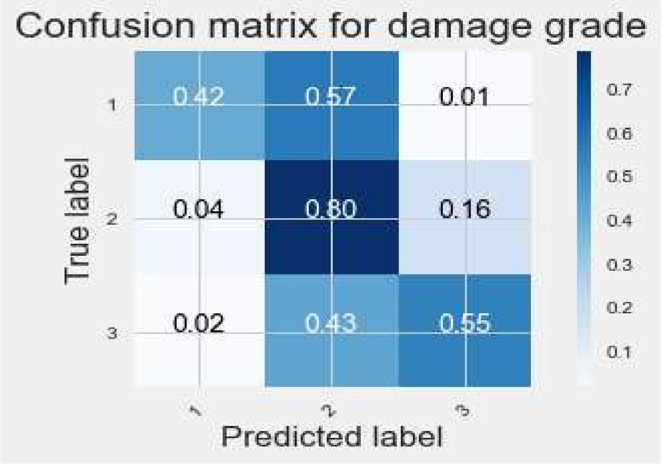


Based on this evaluation, will consider gradient boosting algorithm and xgboost algorithm only for further tuning of hyperparameters.

Here is boxplot after parameter hypertuning for gradient boosting and xgboost algorithms.



From above boxplot, we can conclude that xgboost algorithm appears to be more suited to model our problem. The model was trained with 75% of the data. We then tested the model using the remaining 25% of the data yielding to the following results:



For a micro averaged F1 score of 0.678. We observe that our model predicts with a good accuracy the intermediate damaged buildings. 55% of the time our model predicts correctly an almost complete destruction of a building. However, most of the time, our model predicts an intermediate damage level instead of low damaged level. It's worth to mention that it's very rare that a low level of damage is classified as high level of damage and inversely.

To verify that we are not overfitting, we performed a 10-folds cross-validation. The resulting micro averaged F1 score is now 0.6828.

CONCLUSION

The analysis of damages level caused to building during the 2015 earthquake in Nepal has shown that these damages can be accurately classified into two categories: intermediate damages, and complete destruction. However, we failed at identifying accurately the lowly damaged buildings. We showed that the age, height, area, structure and geographic region of the building are key elements to determine the level of damage a building suffered.