ASSOCIATION BETWEEN THE DIFFERENT KINDS OF BUILDING CONSTRUCTIONS AND THEIR LEVELS OF DAMAGE CAUSED BY THE 2015 GORKHA EARTHQUAKE IN NEPAL

I) Introduction to the Research Question:

***Research question*:** This paper aims to understand the factors such as building location and construction (predictors) to find a relationship between them and the extent of damage which was suffered by the 2015 Gorkha Earthquake in Nepal (response variable).

Motivation or Rationale: Natural calamities are often unavoidable. However, the damages suffered by such calamities can be reduced by improved construction planning of infrastructure. The right choices of material and building plan can be key to not only saving buildings, but also the lives of people living in them. This serves as the motivation behind this study.

Potential Implications: By this research paper, it is possible to find the right construction needed to be used to build buildings in earthquake prone areas to reduce damage and also save life.

II) Methods:

Sample:

Population: The population is data regarding buildings and the damage levels caused to them post the 2015 Gorkha earthquake in Nepal. It was collected through surveys by Kathmandu Living Labs and the Central Bureau of Statistics, which works under the National Planning Commission Secretariat of Nepal. The survey is one of the largest post-disaster datasets ever collected, containing valuable information on earthquake impacts, household conditions, and socio-economic-demographic statistics.

Sample Size: The number of observations are 260601 records.

Description of the Sample: Nepal is a country which places importance on traditional and cultural ways of building which did not foresee the natural calamity which lead to the death of 9000 people. The structural components such as doors and window frames are generally made from wood. Furthermore, private buildings are typically three stories tall, with a load bearing central wall, and a roof made of tile. Nepal is extremely prone to natural disasters such as earthquakes, landslides, monsoons, and avalanches. Scientists are already predicting that another massive earthquake will strike soon. Rural districts and villages are especially vulnerable to natural disasters due to their location. Rebuilding rural Nepal in a sustainable and stable way is crucial for improving aspects of everyday life ranging from public health to economic development.

Measures:

Variables included: The target variable is damage_grade, which represents a level of damage to the building that was hit by the earthquake. There are 3 grades of the damage:

1 - represents low damage

2 - represents a medium amount of damage

3 - represents almost complete destruction

A total of 38 predictor varibles are used in this analysis to try and understand the statistical interpretation behind the vulnerability of buildings in Nepal. Out of the considered varaibles, 5 of them are found to be the most important in the prediction:

1) foundation_type (type: categorical): type of foundation used while building. Possible values

2) ground_floor_type (type: categorical): type of the ground floor.

3) has_superstructure_cement_mortar_brick (type: binary): flag variable that indicates if the superstructure was made of Cement Mortar - Brick

4) has_superstructure_mud_mortar_stone (type: binary): flag variable that indicates if the superstructure was made of Mud Mortar - Stone.

5) roof_type (type: categorical): type of roof used while building.

The rest of the variables used in the study is given below:

- geo_level_1_id, geo_level_2_id, geo_level_3_id (type: int): geographic region in which building exists, from largest (level 1) to most specific sub-region (level 3). Possible values: level 1: 0-30, level 2: 0-1427, level 3: 0-12567.
- count_floors_pre_eq (type: int): number of floors in the building before the earthquake.
- age (type: int): age of the building in years.
- area_percentage (type: int): normalized area of the building footprint.
- height_percentage (type: int): normalized height of the building footprint.
- land_surface_condition (type: categorical): surface condition of the land where the building was built. Possible values: n, o, t.
- other_floor_type (type: categorical): type of constructions used in higher than the ground floors (except of roof). Possible values: j, q, s, x.
- position (type: categorical): position of the building. Possible values: j, o, s, t.
- plan_configuration (type: categorical): building plan configuration. Possible values: a, c, d, f, m, n, o, q, s, u.
- has_superstructure_adobe_mud (type: binary): flag variable that indicates if the superstructure was made of Adobe/Mud.
- has_superstructure_stone_flag (type: binary): flag variable that indicates if the superstructure was made of Stone.
- has_superstructure_cement_mortar_stone (type: binary): flag variable that indicates if the superstructure was made of Cement Mortar - Stone.

- has_superstructure_cement_mortar_brick (type: binary): flag variable that indicates if the superstructure was made of Cement Mortar - Brick.
- has_superstructure_timber (type: binary): flag variable that indicates if the superstructure was made of Timber.
- has_superstructure_bamboo (type: binary): flag variable that indicates if the superstructure was made of Bamboo.
- has_superstructure_rc_non_engineered (type: binary): flag variable that indicates if the superstructure was made of non-engineered reinforced concrete.
- has_superstructure_rc_engineered (type: binary): flag variable that indicates if the superstructure was made of engineered reinforced concrete.
- has_superstructure_other (type: binary): flag variable that indicates if the superstructure was made of any other material.
- legal_ownership_status (type: categorical): legal ownership status of the land where building was built. Possible values: a, r, v, w.
- count_families (type: int): number of families that live in the building.
- has_secondary_use (type: binary): flag variable that indicates if the building was used for any secondary purpose.
- has_secondary_use_agriculture (type: binary): flag variable that indicates if the building was used for agricultural purposes.
- has_secondary_use_hotel (type: binary): flag variable that indicates if the building was used as a hotel.
- has_secondary_use_rental (type: binary): flag variable that indicates if the building was used for rental purposes.
- has_secondary_use_institution (type: binary): flag variable that indicates if the building was used as a location of any institution.
- has_secondary_use_school (type: binary): flag variable that indicates if the building was used as a school.
- has_secondary_use_industry (type: binary): flag variable that indicates if the building was used for industrial purposes.
- has_secondary_use_health_post (type: binary): flag variable that indicates if the building was used as a health post.
- has_secondary_use_gov_office (type: binary): flag variable that indicates if the building was used fas a government office.
- has_secondary_use_use_police (type: binary): flag variable that indicates if the building was used as a police station.
- has_secondary_use_other (type: binary): flag variable that indicates if the building was secondarily used for other purposes.

*Managing the variables***:** The categorical variables are one-hot encoded, which is a technique in which categorical varaibles are included in a regression problem. This increased the prediction vector from 38 to 70 because of the categorical answers in the question.

Analysis:

Statistical Methods: The total of 70, one-hot encoded variable vector is given as an input to a lasso regression algorithm. The algorithm predicted that for 53 iterations of variable inclusion is an ideal predictor for the given problem statement with an average squared error of 0.3015. The important varaibles in the study predicted from the

regression were the five variables given above. The varaibles are further explored using chi-squared tests and box plots to further understand their effect on the damage levels of the buildings

Train-test split: The data was split into 70% training set and 30% testing set

Type of cross-validtion: A k-fold cross validation method with k=10 is used.

III) Results:

Univariate Analysis: The varaibles of consideration are foundation_type_r, ground_floor_type_v, has_superstructure_mud_mortar_stone, roof_type_x, and has_superstructure_mud_mortar_brick. The variables are all categorical. Hence, a frequency table is used to depict their distribution.

Figure 1: Univariate Analysis of explanatory variables

The FREQ Procedure

foundation_type_r	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	41405	15.89	41405	15.89
1	219196	84.11	260601	100.00

ground_floor_type_v	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	236008	90.56	236008	90.56
1	24593	9.44	260601	100.00

HSSMMS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	62040	23.81	62040	23.81
1	198561	76.19	260601	100.00

roof_type_x	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	244418	93.79	244418	93.79
1	16183	6.21	260601	100.00

HSSCMB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	240986	92.47	240986	92.47
1	19615	7.53	260601	100.00

Bivariate Analysis: The damage_grade varaible is the response variable in this problem statement. Box plots are used to visualize the variables relationship with the variable. Chi squared tests are conducted as well.

Figure 2: Bar chart of foundation_type_r vs damage grade:

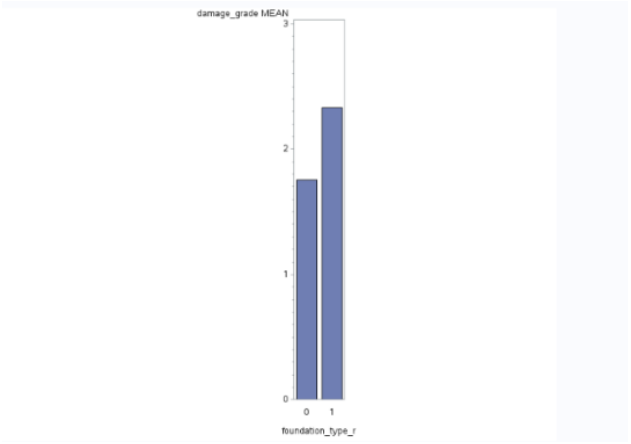


Figure 3: Bar chart of ground_floor_type_v vs damage grade:

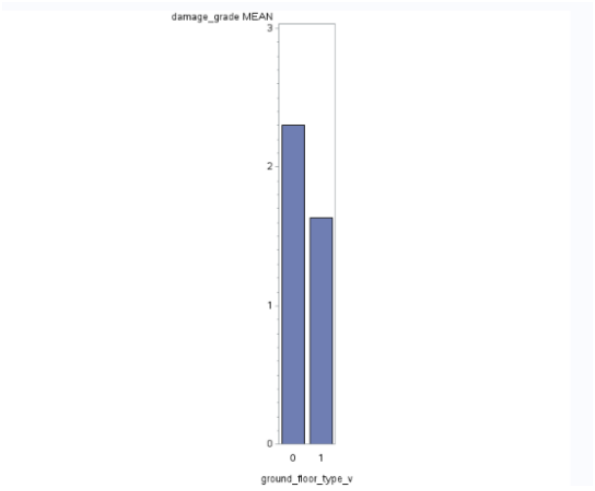


Figure 4: Bar chart of has_superstructure_mud_mortar_stone vs damage grade:

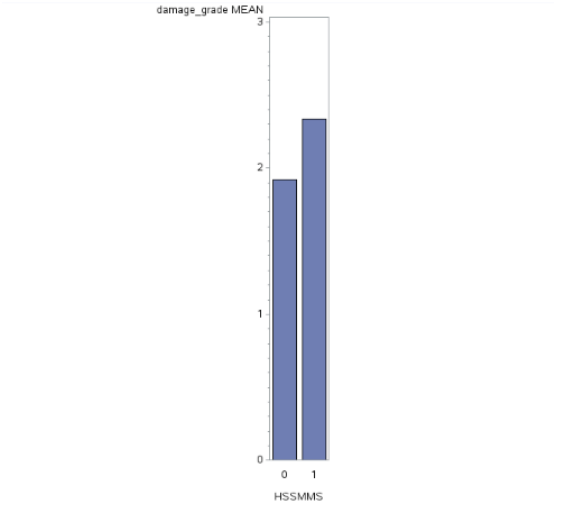


Figure 5: Bar chart of roof_type_x vs damage grade

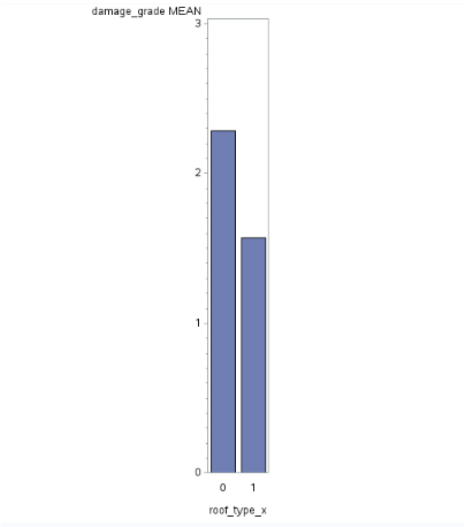
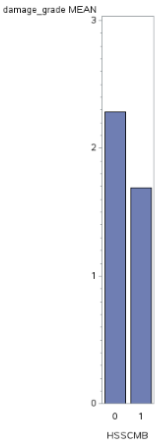


Figure 6: Bar chart of has_superstructure_mud_mortar_brick vs damage grade:



The chi-square value obtained for the variables are:

1) foundation_type_r - 40230.6283

2) ground_floor_type_v - 35848.4509

3) has_superstructure_mud_mortar_stone - 29276.0350

4) roof_type_x - 29905.7022

5) has_superstructure_mud_mortar_brick - 20491.7976

From the chi-square test, the p values obtained from all the variables in the bivariate analysis was <0.0001. ***This proves that the explanatory variables considered are statistically significant to the target variable.***

Lasso Regression: The 70 explanatory variables are given as input to the lasso regression model. The training and testing set are split in a 7:3 ratio. K-fold cross-validation with k=10 is used. The results are depicted in Figure 7 below. We can see from the figure that the five variables analyzed in the univariate and bivariate analysis are identified as the most important amongst all variables.

Figure 7: Lasso Regression results:

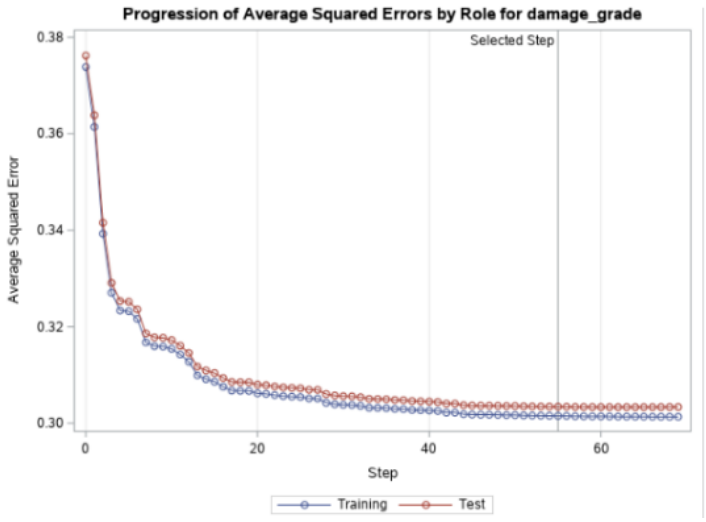
The GLMSELECT Procedure

LAR Selection Summary

Step	Effect Entered	Number Effects In	ASE	Test ASE	CV PRESS
0	Intercept	1	0.3738	0.3762	77941.4497
1	foundation_type_r_0	2	0.3614	0.3638	68736.8270
2	ground_floor_type_v_0	3	0.3393	0.3415	66612.8571
3	HSSMMS_0	4	0.3270	0.3291	66115.9665
4	roof_type_x_0	5	0.3234	0.3253	65996.6612
5	HSSCMB_0	6	0.3232	0.3252	65885.5485
6	foundation_type_i_0	7	0.3216	0.3236	65776.4632
7	count_floors_pre_eq	8	0.3168	0.3186	65096.9013
8	foundation_type_w_0	9	0.3160	0.3178	64681.0370
9	position_t_0	10	0.3159	0.3177	64392.4574
10	has_superstructure_stone_flag_0	11	0.3154	0.3172	64138.0957
11	has_superstructure_timber_0	12	0.3142	0.3161	63871.4006
12	has_superstructure_rc_engineered_1	13	0.3128	0.3146	63790.8435
13	roof_type_q_0	14	0.3099	0.3117	63631.1736
14	has_superstructure_adobe_mud_1	15	0.3091	0.3109	63498.4953
15	position_s_0	16	0.3086	0.3104	63488.8254
16	count_families_0	17	0.3076	0.3094	63414.6039
17	has_secondary_use_0	18	0.3068	0.3086	63368.2217

The algorithm identified the 55th iteration to be the best-fit model. The average squared errors for training and testing are given below:

Figure 8: Training-testing error:



Findings: The parameter estimates from the Lasso Regression for the considered variables are given below:

1) foundation_type_r : -0.102874

2) ground_floor_type_v: 0.176191

3) has_superstructure_mud_mortar_stone: -0.192428

4) roof_type_x: 0.076403

5) has_superstructure_mud_mortar_brick: 0.149577

- Clearly, a building having foundation type as category ‘r’ and having superstructure made of mud, mortar and stone decreases damages due to earthquake
- Having ground floor type as category ‘v’, roof type as ‘x’ and having a supperstructure made of mud, mortar and brick increases damages due to earthquake.

IV) Conclusions/Limitations:

From the lasso regression algorithm, it was found that buildings in Nepal having foundation type ‘r’ and the superstructure made of mud, mortar and stone have a better chance of surviving an earthquake. Also, buildings having ground floor type as ‘v’, roof type as ‘x’ and the superstructure made of mud, mortar and brick worsens the damage level caused by an earthquake. The algorithm has successfully created an association and between construction types and the damage levels that can arise from a potential earthquake to it.From these findings, buildings in the future can be built keeping in mind the factors that can improve the construction provided by this study. This can potentially save lives and even infrastructure.

There are few limitations to this study. Firstly, the population is restricted to the buildings of Nepal. There are many other regions across the globe which are vulnerable to earthquakes whose factors causing damage may differ from the buildings of Nepal. Secondly, only linear powers of the variables are considered in the regression algorithm. The results can possibly be improved by using logarithmic transformations or polynomial powers.