**Earthquake Damage Prediction – Houses in Nepal**

**Data 606 – Final Project**

February 12, 2021

Authors:

Albert Leung

Yu Nakamura

Li Lam

Graeme Kempthorne

**Background**

As cities around the world become larger, and more people live in denser places, the risks and damages associated with earthquakes increases. Our group wanted to explore a dataset that detailed the results of an earthquake to see if we could identify risk factors associated with building damage. One of the deadliest earthquakes to occur in the last decade was a 7.8 magnitude earthquake that occurred in Nepal on the 25th of April 2015. Almost 9000 individuals were killed on the day as the city of Kathmandu and surrounding towns came to a complete stop. To assess the costs of the earthquake, the National Planning Committee of Nepal surveyed and compiled a dataset of all the buildings damaged. We were able to obtain this dataset from Kaggle, which was compiled by an individual named Mobius (<https://www.kaggle.com/arashnic/earthquake-magnitude-damage-and-impact?select=ward_vdcmun_district_name_mapping.csv>)

Within the dataset is an ordinal variable that corresponds to the degree of building damage (grade 1 to 5). Our group is looking to explore this dataset to see which variables can be used to accurately predict building damage. This is useful analysis for government officials in Nepal as well as other similar disaster prevention committees to limit the costs and damages of earthquakes.

**Dataset**

This dataset contains multiple .csv files that detail different variables associated with buildings and their ownership/household's information that were damaged during the earthquake.

In our study, we investigate several categorical analyses to identify and predict the degree of building damage due to earthquakes with multiple predictors. The details of the variables are presented in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Variable Name** | **Data Source** | **Response / Predictors** | **Type of Variable** |
| [1] | building\_id |  | Common key | NA |
| [2] | district\_id |  | Predictors | Categorical |
| [3] | vdcmun\_id |  | Common key | NA |
| [4] | ward\_id |  | Common key | NA |
| [5] | count\_floors\_pre\_eq | Building Structure | Predictors | Numeric |
| [6] | count\_floors\_post\_eq | Building Structure | Predictors | Numeric |
| [7] | age\_building | Building Structure | Predictors | Numeric |
| [8] | plinth\_area\_sq\_ft | Building Structure | Predictors | Numeric |
| [9] | height\_ft\_pre\_eq | Building Structure | Predictors | Numeric |
| [10] | height\_ft\_post\_eq | Building Structure | Predictors | Numeric |
| [11] | land\_surface\_condition | Building Structure | Predictors | Categorical |
| [12] | foundation\_type | Building Structure | Predictors | Categorical |
| [13] | roof\_type | Building Structure | Predictors | Categorical |
| [14] | ground\_floor\_type | Building Structure | Predictors | Categorical |
| [15] | other\_floor\_type | Building Structure | Predictors | Categorical |
| [16] | position | Building Structure | Predictors | Categorical |
| [17] | plan\_configuration | Building Structure | Predictors | Categorical |
| [18] | has\_superstructure\_adobe\_mud | Building Structure | Predictors | Categorical |
| [19] | has\_superstructure\_mud\_mortar\_stone | Building Structure | Predictors | Categorical |
| [20] | has\_superstructure\_stone\_flag | Building Structure | Predictors | Categorical |
| [21] | has\_superstructure\_cement\_mortar\_stone | Building Structure | Predictors | Categorical |
| [22] | has\_superstructure\_mud\_mortar\_brick | Building Structure | Predictors | Categorical |
| [23] | has\_superstructure\_cement\_mortar\_brick | Building Structure | Predictors | Categorical |
| [24] | has\_superstructure\_timber | Building Structure | Predictors | Categorical |
| [25] | has\_superstructure\_bamboo | Building Structure | Predictors | Categorical |
| [26] | has\_superstructure\_rc\_non\_engineered | Building Structure | Predictors | Categorical |
| [27] | has\_superstructure\_rc\_engineered | Building Structure | Predictors | Categorical |
| [28] | has\_superstructure\_other | Building Structure | Predictors | Categorical |
| [29] | condition\_post\_eq | Building Structure | Predictors | Categorical |
| [30] | damage\_grade | Building Structure | Response (Target) | Categorical |
| [31] | technical\_solution\_proposed | Building Structure | Predictors | Categorical |
| [32] | bank\_account\_present | Households Demographic | Predictors | Categorical |
| [33] | income\_level | Households Demographic | Predictors | Numeric |
| [34] | mean\_age | Households Demographic | Predictors | Numeric |
| [35] | total\_size\_household | Households Demographic | Predictors | Numeric |
| [36] | total\_vaccination\_drop\_last\_12\_months | Households Demographic | Predictors | Numeric |
| [37] | total\_pragnancy\_treatment\_drop\_last\_12\_months | Households Demographic | Predictors | Numeric |
| [38] | total\_education\_drop\_last\_12\_months | Households Demographic | Predictors | Numeric |
| [39] | total\_death\_last\_12\_months | Households Demographic | Predictors | Numeric |
| [40] | total\_injury\_loss\_last\_12\_months | Households Demographic | Predictors | Numeric |

**Data Cleaning**

The dataset consists of nine files which have more than 100 variables between them, including some that are duplicated or redundant. To simplify, we reviewed and decided to focus on 40 variables that could potentially impact building damage based on our general knowledge for the study.

Before the analysis, the following steps were used for data cleaning. All analyses are performed in RStudio.

1. Data aggregation and wrangling to merge three .csv files with common keys
2. Outlier detection and elimination

The details of each step are described as below:

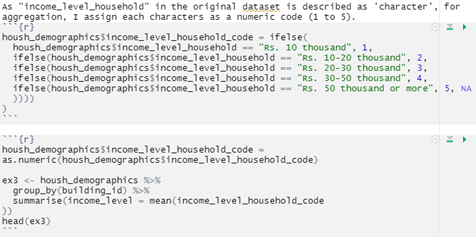
1. Data aggregation and wrangling:

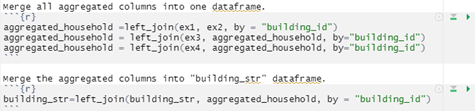
There are three .csv files (building structure, household information (such as demographics) and impact due to the earthquake) we merged into one data frame for our model study. Figure 1 provides an entity relationship (ER) diagram to show relationships between those files.

In RStudio, all files were left-joined with a unique identifier of building ID.

This is a summary of the process:

* + Check and modify data validity. For example:
    - Missing values were not loaded correctly as null into R.
    - Several duplicated IDs were listed, which caused redundant rows.
    - Empty rows in the original file sometimes mean zero. (e.g.: death counts in the dataset shows as empty in each row but mean no-death count).
    - Drop null rows of the response variable.
  + Aggregation was required to link one-to-one relationships between building id and household id. For example:
    - The average age, income level, total number of households in each building etc. were calculated then assigned to each building record.





Coding example of aggregation to merge files

Through these steps, the merged data frame was generated containing 762,000 rows and 40 variables.

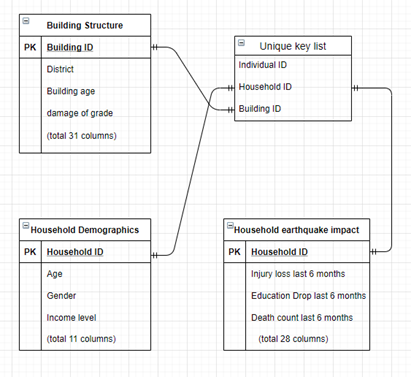


Fig. 1 ER diagram

1. Outlier detection and elimination:

There are 12 numeric variables in the merged data frame. Below are statistics of each numeric variable. There are some unreasonable ranges of values in the dataset. The box plot method is used to detect outliers which are then removed from the dataset.

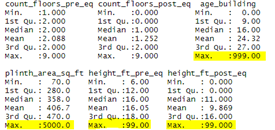
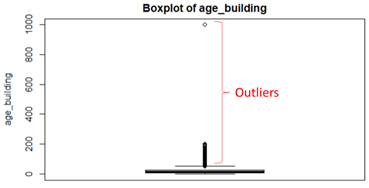


Fig. 2 Statistics summary of numeric variables

The outlier detection was implemented for six variables (age of building, area square feet of building, height feet of building pre/post-earthquake, mean age of household and total size of households.) In total, 121,000 unnecessary rows were eliminated. This reduced the number of building records from 762,000 to 641,000 rows.





Coding example of outlier elimination

**Data Exploration**

In this section, we provide an outline of how to select and remove columns in order to create a reduced data frame with only the features which are thought to be relevant to the degree of building damage an earthquake can cause.

***Response Variable: Degree of Building Damage***

As mentioned earlier, our team was investigating a model to predict the degree of building damage due to earthquake with this dataset. The response variable has five classes, labelled as Grade 1 to 5, which represent the scale of damage sustained to the building during the earthquake.

The following histogram provides the distribution of the occurrence of each grade. We can see that the number of occurrences increases from grade 1 to grade 5. This means that many buildings had severe damage due to the earthquake.

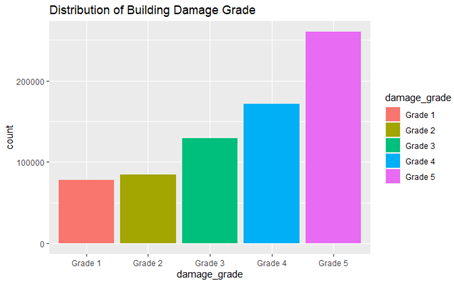


Fig. 3 Occurrence of each grade in building damage

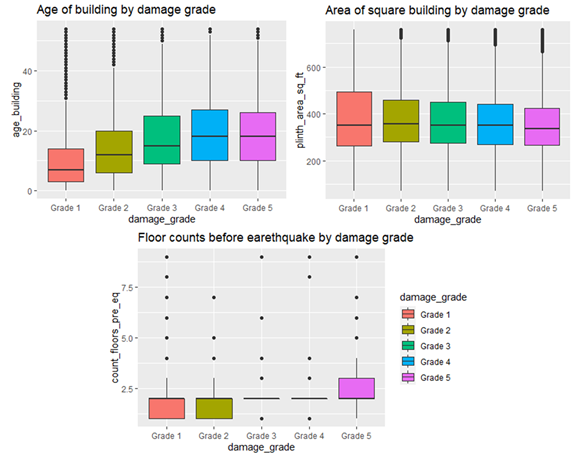
***Explanatory Variables:***

Before diving into the analysis, it was important to identify the candidate variables that could be used for the prediction model. We reviewed 12 numeric variables and 22 categorical variables.

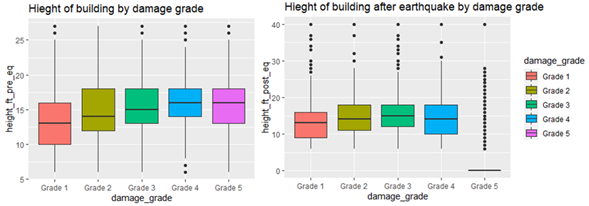
1. Numeric variables:

There are two types of numeric variables related to **Building Structure** and **Household information.**

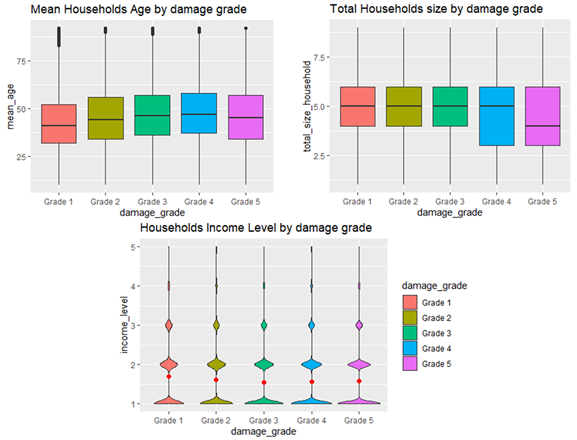
The following box plots (Fig.4) show that the mean building age is higher when building damage is more severe. The higher number of floors a building has does not appear to be corelated to the damage level. There is no clear relationship between the area and damage levels.

Fig. 4 Building structure (age, area, and floor counts) and building damage

When looking at the effect of building height before and after earthquake in Fig. 5, it is noticeable to see the significant damages caused in grade 5.

Fig. 5 Comparison of building height pre and post-earthquake by building damages

Next, we checked the household variables in each damage grade. Overall, it is hard to identify any difference between building damages (Fig. 6). There is a slight increase of mean income level in less damaged categories (grade 1 and 2) compared to the highly damaged buildings.

Fig. 6 Households variables (mean age, total size, income level) in each building damage

Other variables related to the households such as “total death count last 12 months”, “total injuries or losses last 12 months” or “total education drops last 12 months” show almost all points are zero. Therefore, those are not valid parameters for the prediction.

1. Categorical variables:

First, we looked at the district to see any distribution trends that could be identified between building damage grades. From the matrix plots in Fig. 7, there are some districts which are associated with higher building damages such as the district ID = 12, 14, 16, 22 and 30, while more than half buildings are categorized as damage grade 1 in the district = 1.

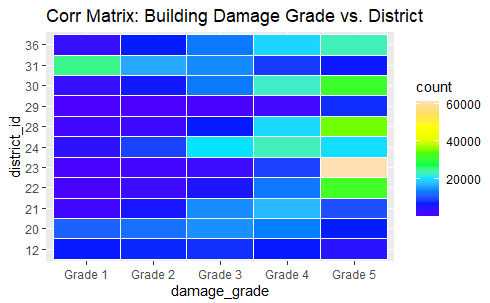


Fig. 7 Distribution of building damage grade by district

Here is the distribution percentage of each damage grade by building foundation type. Clearly, reinforced concrete (RC) has less damage, while the mud mortar-stone/brick is the worst foundation type for enduring building damage.

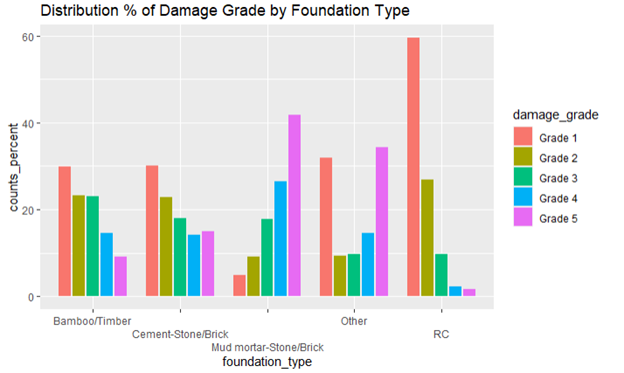


Fig. 8 Distribution % of building damage grade by foundation type

The same visualization approach is applied for roof types as well. As shown in the figure below, RC roofs have less damage compared to bamboo/timber-heavy or light roof types.

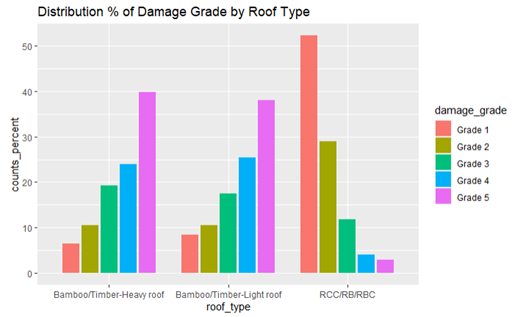


Fig. 9 Distribution % of building damages by roof type

We also reviewed other building structure attributes such as floor type, land surface condition, position (attached on 1, 2, 3 sides, and not attached), plan configuration (building shape type), the availability of RC engineered superstructure and so on with same approach described above.

Next, we again reviewed the household’s demographic information to see if households have a bank account or not. We expected that this variable indirectly represents their income level (which may affect where people live) and might have potential impact to the damage grade. However, no clear direction was observed. See figure 10.

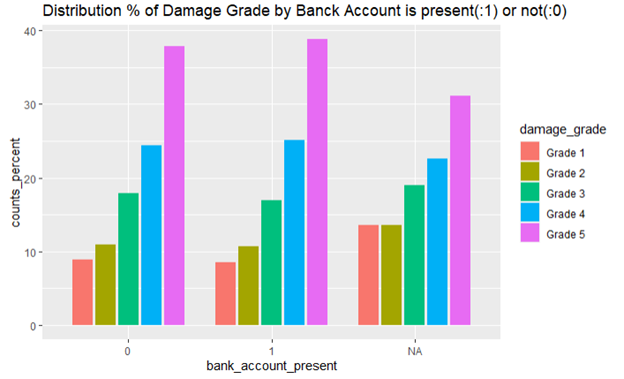
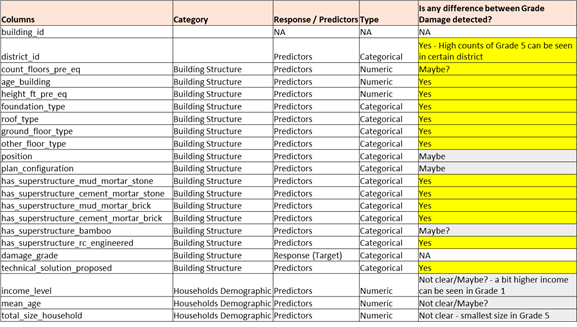


Fig. 10 Distribution % of damage grade by bank account presents

From this data exploration, a summary table of explanatory variables was created, showing the potential impacts on the building damage grade which are confirmed through data visualization.

We focus on these parameters for further data analysis and building prediction models.



**Sampling methodology**

The purpose of this section is to find out the best sampling methodology for the predictive analysis when testing the model. To check the impact by the sampling method, we estimated the population mean and the standard deviation of building age.

1. Simple random sampling (SRS):

A simple random sample of 1,000 was drawn using proportional allocation. Here is the result of estimated population mean, standard deviation and 95% confidence interval of building age.

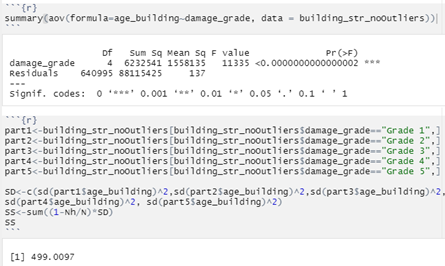


1. Stratified sampling:

A stratified sample of 1,000 was drawn using proportional allocation. Building damage grade is used as a stratum. Here is the result of estimated population mean, standard deviation and 95% confidence interval of building age.



The mean and standard deviation from the stratified sampling is closer than the one from SRS. To check the validity of stratified sampling, we checked whether the SSB >> ∑Hh=1(1−NhN)Vh or not.



From above output, SSB >> ∑Hh=1(1−NhN)Vh, therefore, we conclude that the use of stratified sampling gives us a more stable estimation compared to the SRS.

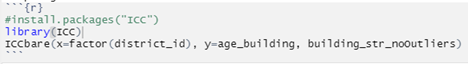
1. Cluster sampling:

For clustering sampling, we drew two clusters from “district\_id” (total 11 districts) and applied one-stage cluster sampling to estimate the population mean and standard deviation of building age.

Here is the result:

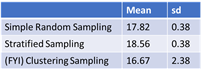


To check whether if cluster sampling is stable or not, we calculated the intraclass correlation coefficient which is 0.04.



The ICC is close to zero which suggests that cluster sampling is most likely stable.

Through the three methods used to estimate mean and standard deviation, the mean building age from each method is close, but the standard deviation with cluster sampling is quite high. Since stratified sampling is proved to provide the stableness of estimation, we decided to use stratified sampling for building damage grade for further analysis.



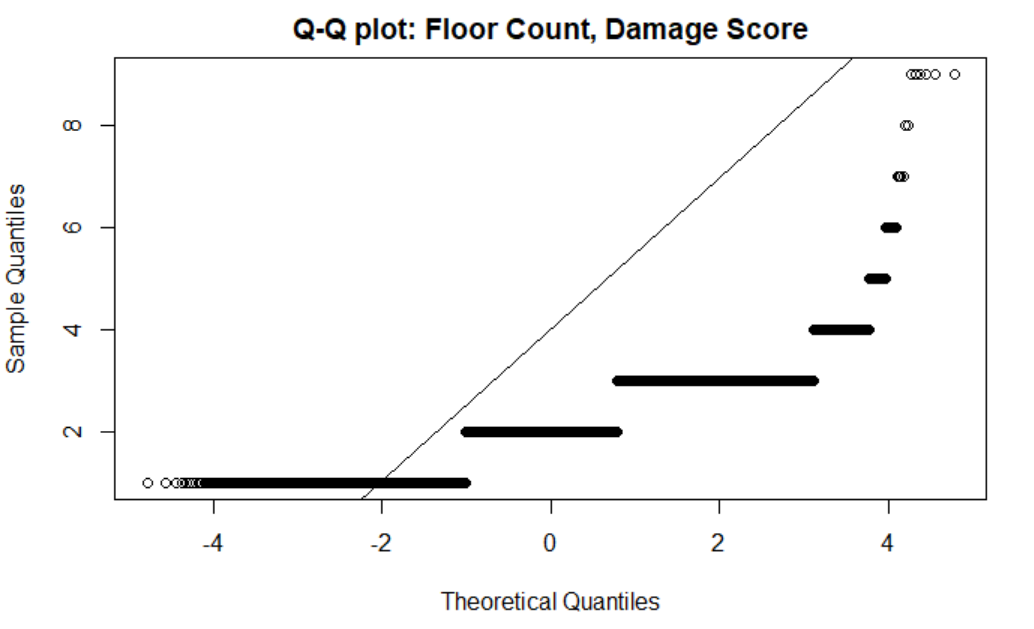
Sampling method comparison: Estimated mean and standard deviation of building age

**Predictive Model(s)**

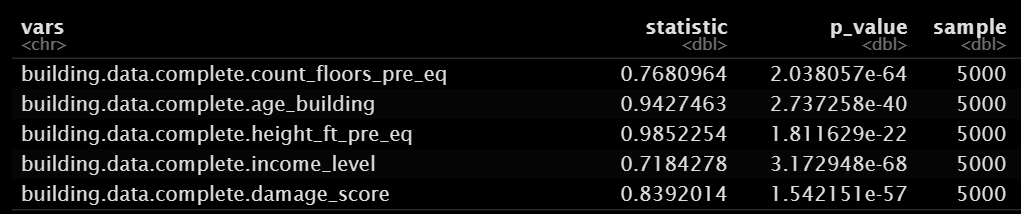
1. **Linear & Quadratic Discriminant Analysis Models**

One of the ways to model this data is to use linear discriminate analysis (LDA) or quadratic discriminant analysis (QDA). This approach uses Bayes’ theorem to predict an outcome based on the prior probability of the outcomes occurring. For our earthquake dataset, we can build a model to predict the ‘damage\_grade’ variable using the selected variables from the data exploration.

This model can be powerful but is limited by the requirement that the data sampled follows a gaussian (normal) distribution. This requirement is not fulfilled by our data, as seen in the exploration, as the data collected only includes damaged buildings by the Nepal earthquake. This selection does not ensure a normal distribution and limits the predictive power of the LDA and QDA models. However, it remains a powerful model because it redraws decision boundaries to minimize the variance of the predicted outcomes. Due to the large number of categorical variables in our database the normality can only be checked on the ordinal or qualitative variables. A Q-Q plot was done to quickly test the results with a Shapiro-Wilks test done to confirm whether these variables were normal. An example of a test Q-Q plot for damage\_grade~Income level and the Shapiro-Wilks tests for each variable are shown below.



QQ normality plot between count\_floors\_pre\_eq and damage\_grade variables



Shapiro-Wilks test for ordinal and qualitative variables.

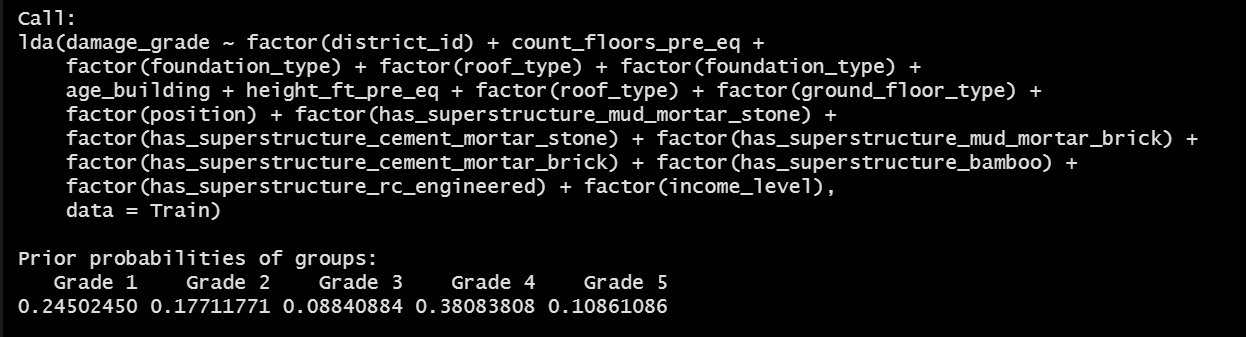
The p-values shown from the Shapiro-Wilks tests indicate that that we should reject the null hypothesis that the data is normally distributed and accept the alternative hypothesis (data is not normally distributed). This gives us an indication that the LDA and QDA models may not be appropriate for this kind of analysis, but we wanted to test and see if they were still able to provide a good fit.

To test this model, we wanted to focus just on predictors that were directly related to the buildings themselves prior to the earthquake to limit the complexity of the model. In this, the predictor variables ‘building\_id’, ‘technical\_solution\_proposed’, ‘mean\_age’, and ‘total\_size\_household’ were dropped from the model as they were measures of the individuals living in the houses. The variable ‘plan\_configuration’ was also dropped due to the low occurrence of values within the dataset that would have forced an additional stratum in order to incorporate it into the R models.

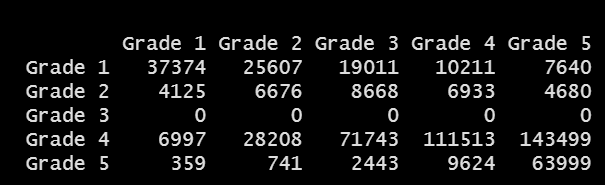
Sampling was completed using a stratified sample of 10,000 units (1.75% of total) that included the amount of each damage grade that was proportional to the overall population. The models were then trained on this data and the models tested on the remaining datapoints within the dataset.

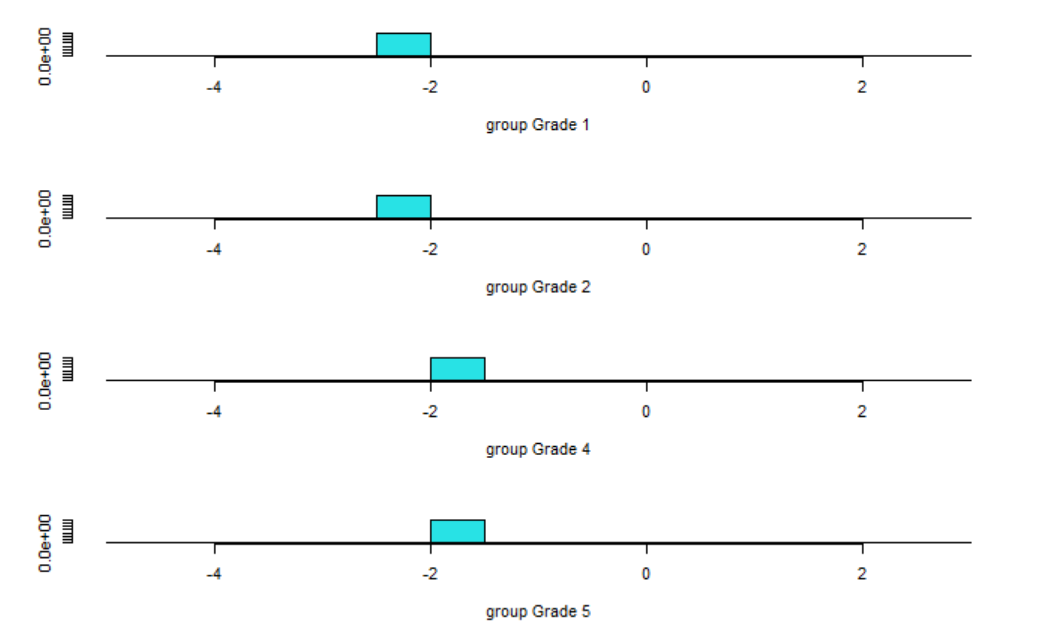
**LDA.model1**

The first LDA model used all the variables that were not removed from the selected dataset to identify the strength of this approach and to see which variables were contributing the most to the model. All the categorical variables were identified as such which greatly increased the size of the model.



In this first model, the LD1 trace accounted for 87.96% of the between-class variance and LD2 accounted for 10.6%. This model did not perform well as seen below in the chart because it failed to predict any grade 3 damage in the buildings. This was largely because the model could not find a significant difference between the damage grade means on the coefficients of linear discriminants.

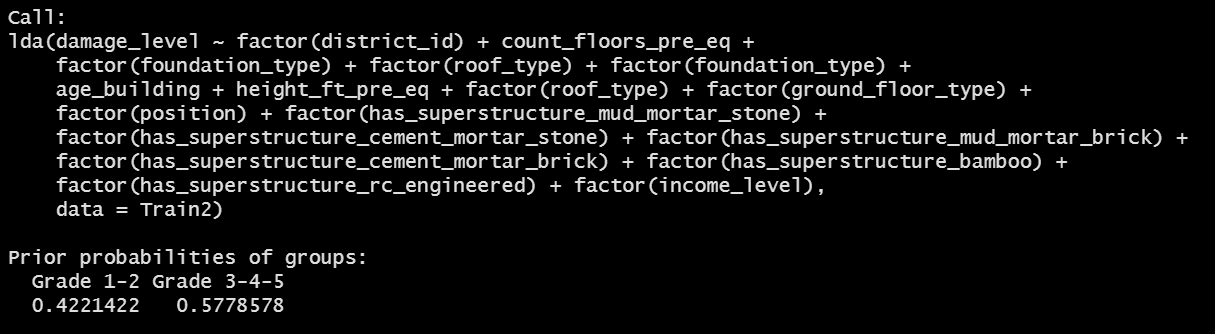




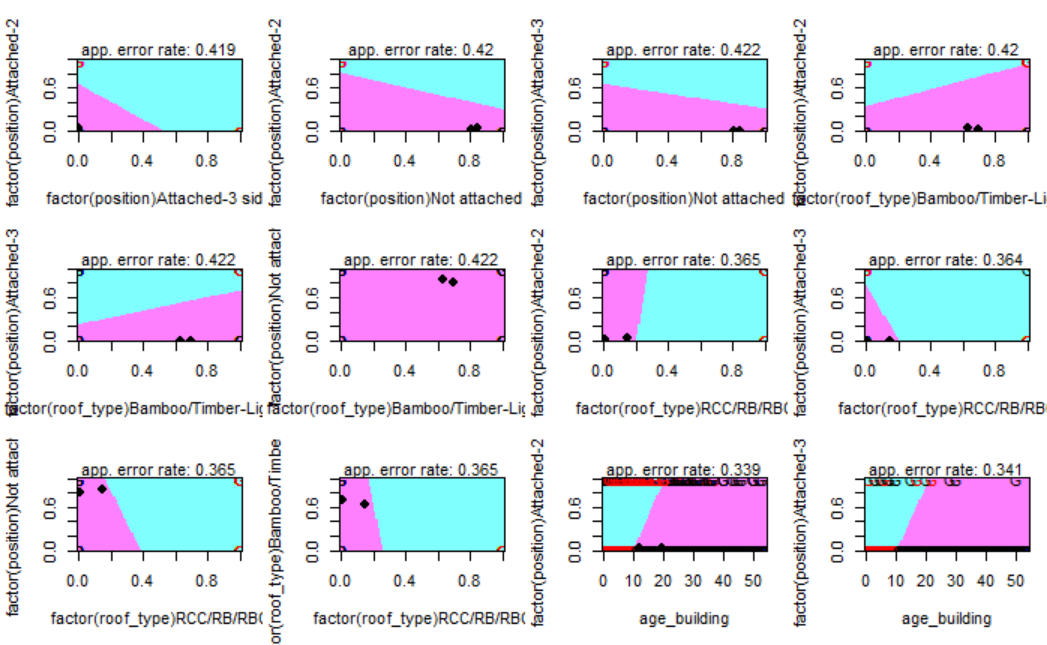
The failure to predict any grade 3 damage along with many other misclassifications lead to an overall misclassification rate of 62.2%. This was largely driven by the variance within each of these categories and how close many of the calculated means were within each of these categorical variables.

**LDA.model2**

Because of the closeness of these variables predicted by the LDA approach, and to reduce the complexity of the model, the damage grades were combined to either low (grades 1 and 2) or high (grade 3, 4, and 5). This approach reduces the precision of the model in predicting the effects of the earthquake on buildings but still allows us to explore the factors that contributed the most to the damage using a LDA approach.

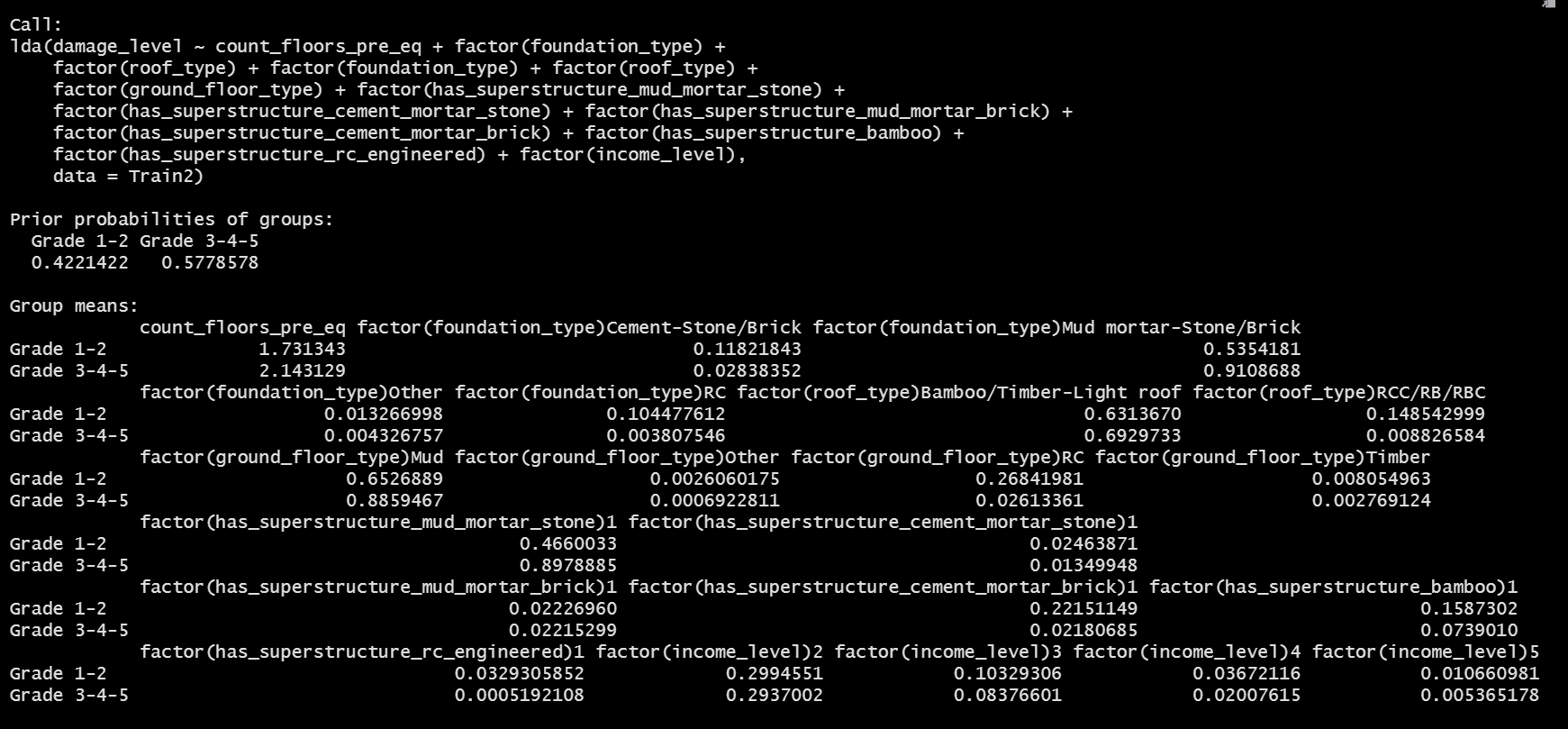


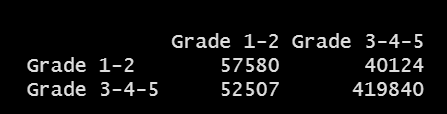
This approach reduced the misclassification rate down to 18.1% as it was clear there was some association between the building properties and the amount damage sustained. To investigate this further, a pairwise exploration of the data was done using the ‘partimat’ function in the klarR library. In this, each variable is compared to each other to determine which ones were contributing the most to categorizing the damage level based on the predicted error rate.



**LDA.model3**

With the ‘partimat’ function, a third LDA model was produced that had the lowest misclassification rate of all other alternatives attempted. Using 13 variables the misclassification rate was lowered to 17.7%.

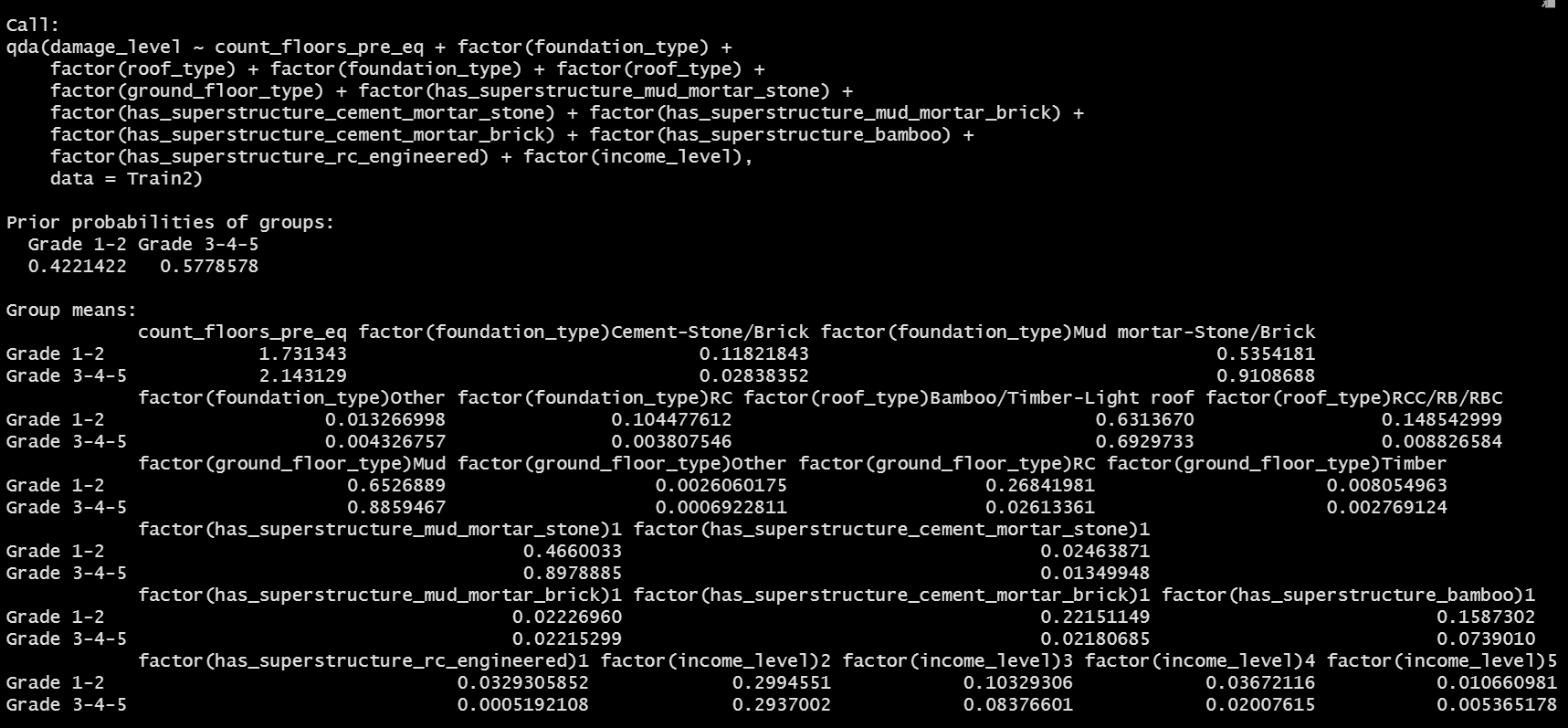




However, a 17.7% misclassification rate is still very inaccurate for this kind of analysis given the simplification done to the damage grades. This kind of analysis was useful for identifying important variables but does not provide the predictive power this type of analysis needs. It is important to note that district\_id was dropped from the LDA approach in this analysis. This was surprising as there are certain districts that have significantly more grade 5 or grade 1 damage due to their relative proximity to the earthquake. We assumed that the location of the buildings would have a significant impact on the predictive models, but this analysis reveals it is not as useful as a predictor. The LDA analysis reveals that the house structure variables are more powerful predictors of the damage level than the proximity to the earthquake’s epicenter.

**QDA.model1**

This approach was also attempted using a QDA function to determine the damage grade or damage level. QDA has an advantage over LDA in that it doesn’t assume that the covariances among all the predictor variables is the same. This allows the sigma squared variable in the function to fluctuate with each of the predictor variables. The QDA is still limited by the normal distribution assumption but we were hopeful it would improve the predictive power of this approach.



However, the model performed worse in all the cases that the LDA model was used in. The lowest misclassification rate achieved was 19.7% when the damage level was simplified, and the reduced variables were used as in the LDA.model3. Though useful in attempting to simplify some of the analysis, this predictive model approach was put aside to pursue better models in establishing useful predictors for earthquake damage.

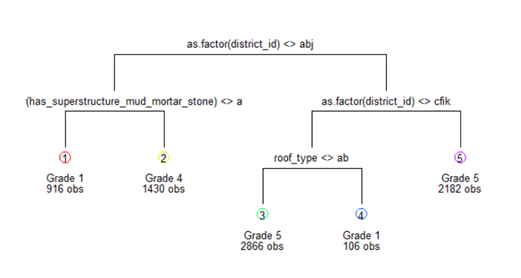
1. **Classification Tree Models**

The next model used to determine the response variable, damage levels to structural buildings, in the Gorkha earthquake in Nepal is the classification tree. Since the entire dataset is large, a stratified sample of 10,000 is taken with proportional allocation based on the damage grades. The sample is further divided via simple random sampling into a training set to create the tree regression model and a test set to evaluate the validity of the model.

The following 13 explanatory variables are used in the classification tree model:

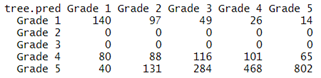
|  |  |
| --- | --- |
| * district\_id | * height\_ft\_pre\_eq |
| * land\_surface\_condition | * foundation\_type |
| * roof\_type | * ground\_floor\_type |
| * has\_superstructure\_mud\_mortar\_stone | * has\_superstructure\_cement\_mortar\_stone |
| * has\_superstructure\_mud\_mortar\_brick | * has\_superstructure\_cement\_mortar\_brick |
| * has\_superstructure\_timber | * has\_superstructure\_bamboo |
| * age\_building |  |

When the tree function is performed on the 13 explanatory variables with damage grade as the response variable, the resulting model is:

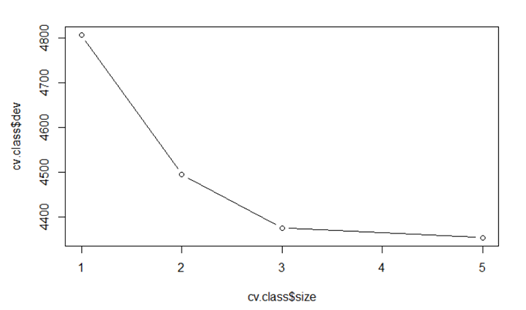


The three predictors used are “district\_ID”, “has\_superstructure\_mud\_mortar\_stone”, and “roof\_type”. There are five terminal nodes and only three of the five damage grades.

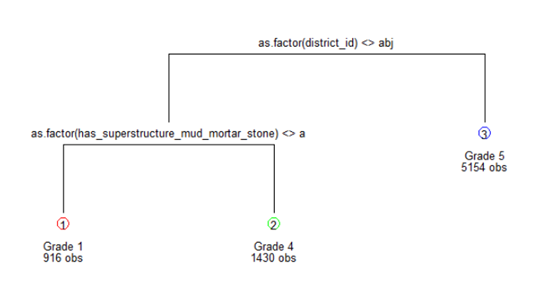
To test this model, the tree is applied against the testing data set. The misclassification rate is 58.30%. This is not surprising since there are no branches for Grade 2 and Grade 3. The following table shows the classification rate.



The next step in the process is to prune the tree. When the cross-validation error is plotted against the size of the tree, there is not a large difference in cross validation error between five and three terminal nodes.

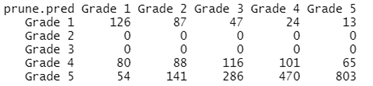


A tree size of three is chosen to prune the tree because it is easier to interpret the model compared to five terminal nodes. Also, reducing the tree to less than three nodes increases the cross-validation error. A smaller tree with fewer splits may lower the amount of variance resulting in an easier to interpret model with minimal bias.



The pruned tree uses the explanatory variables “district\_ID” and “has\_superstructure\_mud\_mortar \_stone” which results in three terminal nodes that can predict Grades 1, 4 and 5.

The classification table shows the number of correct and incorrect responses predicted by the model.



The misclassification rate yields 58.82% indicating that the pruned tree model and the full tree model yields a similar accuracy rate.

Models are based on the available input. A possible explanation of why the model performs poorly may be caused by the predictors chosen to predict the response variable. When the initial stratified sampling is taken, each layer is assumed to have similar characteristics. Likewise, the different layers are assumed to have significantly different characteristics. This effect is accumulated when the sample is partitioned into a training and testing set. Another possible contribution to the poor model may be the number of classes in the response variable. Ideally, one or two categories would perform better than five categories. When there are fewer classes, the purity of each region increases. Since the classification tree model is inadequate, let’s move onto another method.

1. **Multinomial Regression Model - Baseline-Category Logit Model**

In the quest of creating a model that is more efficient, a baseline-category logit model was chosen since it can build a model that has categorical data and numerical data. Since the entire dataset is large, a stratified sample was taken based on the proportional allocation among the categories in the response variable with a sample size of 5,000 due to computer limitations.

**Base-category Logit Model Full**

A model was created based on using the stratified sample and the following 12 predictors: age\_building, height\_ft\_pre\_eq , foundation\_type, roof\_type, ground\_floor\_type, other\_floor\_type, has\_superstructure\_mud\_mortar\_stone, has\_superstructure\_cement\_mortar\_stone, has\_superstructure\_mud\_mortar\_brick, has\_superstructure\_cement\_mortar\_brick, and has\_superstructure\_rc\_engineered. This gave the following coefficients:

(Intercept):1   
 2.615089e+00   
 (Intercept):2   
 1.497272e+00   
 (Intercept):3   
 1.244304e+00   
 (Intercept):4   
 9.628623e-02   
 age\_building:1   
 -4.800527e-02   
 age\_building:2   
 -1.300067e-02   
 age\_building:3   
 -7.722725e-03   
 age\_building:4   
 9.738804e-05   
 height\_ft\_pre\_eq:1   
 -3.444945e-02   
 height\_ft\_pre\_eq:2   
 -3.737256e-02   
 height\_ft\_pre\_eq:3   
 1.027616e-03   
 height\_ft\_pre\_eq:4   
 -4.281300e-04   
 foundation\_typeCement-Stone/Brick:1   
 -1.109048e+00   
 foundation\_typeCement-Stone/Brick:2   
 -9.170881e-01   
 foundation\_typeCement-Stone/Brick:3   
 -1.047831e+00   
 foundation\_typeCement-Stone/Brick:4   
 -8.328362e-01   
 foundation\_typeMud mortar-Stone/Brick:1   
 -2.142199e+00   
 foundation\_typeMud mortar-Stone/Brick:2   
 -1.752129e+00   
 foundation\_typeMud mortar-Stone/Brick:3   
 -1.656304e+00   
 foundation\_typeMud mortar-Stone/Brick:4   
 -1.168647e+00   
 foundation\_typeOther:1   
 -2.204272e+00   
 foundation\_typeOther:2   
 -1.942541e+00   
 foundation\_typeOther:3   
 -2.338063e+00   
 foundation\_typeOther:4   
 -1.637562e+00   
 foundation\_typeRC:1   
 -3.027082e-01   
 foundation\_typeRC:2   
 -3.458066e-01   
 foundation\_typeRC:3   
 -2.086602e-01   
 foundation\_typeRC:4   
 -1.747945e+00   
 roof\_typeBamboo/Timber-Light roof:1   
 5.197747e-01   
 roof\_typeBamboo/Timber-Light roof:2   
 1.727890e-01   
 roof\_typeBamboo/Timber-Light roof:3   
 1.128181e-01   
 roof\_typeBamboo/Timber-Light roof:4   
 2.547471e-01   
 roof\_typeRCC/RB/RBC:1   
 2.723848e+00   
 roof\_typeRCC/RB/RBC:2   
 2.325141e+00   
 roof\_typeRCC/RB/RBC:3   
 1.312354e+00   
 roof\_typeRCC/RB/RBC:4   
 2.585609e+00   
 ground\_floor\_typeMud:1   
 -1.125448e-01   
 ground\_floor\_typeMud:2   
 1.942299e-01   
 ground\_floor\_typeMud:3   
 5.704004e-02   
 ground\_floor\_typeMud:4   
 1.956017e-01   
 ground\_floor\_typeOther:1   
 1.359532e+01   
 ground\_floor\_typeOther:2   
 1.487736e+01   
 ground\_floor\_typeOther:3   
 1.522601e+01   
 ground\_floor\_typeOther:4   
 1.381297e+01   
 ground\_floor\_typeRC:1   
 3.044099e-01   
 ground\_floor\_typeRC:2   
 8.460498e-01   
 ground\_floor\_typeRC:3   
 2.392229e-01   
 ground\_floor\_typeRC:4   
 2.385954e-01   
 ground\_floor\_typeTimber:1   
 1.263337e-01   
 ground\_floor\_typeTimber:2   
 1.223918e+00   
 ground\_floor\_typeTimber:3   
 -3.431613e-01   
 ground\_floor\_typeTimber:4   
 1.420145e+00   
 other\_floor\_typeRCC/RB/RBC:1   
 -1.492925e+00   
 other\_floor\_typeRCC/RB/RBC:2   
 -6.765963e-01   
 other\_floor\_typeRCC/RB/RBC:3   
 -9.877621e-01   
 other\_floor\_typeRCC/RB/RBC:4   
 -5.406456e-01   
 other\_floor\_typeTimber-Planck:1   
 -3.870012e-02   
 other\_floor\_typeTimber-Planck:2   
 -2.051592e-01   
 other\_floor\_typeTimber-Planck:3   
 -1.382362e-01   
 other\_floor\_typeTimber-Planck:4   
 2.635269e-01   
 other\_floor\_typeTImber/Bamboo-Mud:1   
 -2.141312e-01   
 other\_floor\_typeTImber/Bamboo-Mud:2   
 7.910396e-03   
 other\_floor\_typeTImber/Bamboo-Mud:3   
 6.542335e-02   
 other\_floor\_typeTImber/Bamboo-Mud:4   
 4.747681e-01   
 has\_superstructure\_mud\_mortar\_stone1:1   
 -1.912050e+00   
 has\_superstructure\_mud\_mortar\_stone1:2   
 -8.902968e-01   
 has\_superstructure\_mud\_mortar\_stone1:3   
 -6.087495e-01   
 has\_superstructure\_mud\_mortar\_stone1:4   
 -2.008282e-01   
 has\_superstructure\_cement\_mortar\_stone1:1   
 -8.701245e-01   
 has\_superstructure\_cement\_mortar\_stone1:2   
 5.120742e-02   
 has\_superstructure\_cement\_mortar\_stone1:3   
 3.792692e-01   
 has\_superstructure\_cement\_mortar\_stone1:4   
 -4.092235e-01   
 has\_superstructure\_mud\_mortar\_brick1:1   
 1.598284e-01   
 has\_superstructure\_mud\_mortar\_brick1:2   
 5.025459e-01   
 has\_superstructure\_mud\_mortar\_brick1:3   
 9.651412e-01   
 has\_superstructure\_mud\_mortar\_brick1:4   
 9.244072e-01   
 has\_superstructure\_cement\_mortar\_brick1:1   
 1.893116e+00   
 has\_superstructure\_cement\_mortar\_brick1:2   
 1.741677e+00   
 has\_superstructure\_cement\_mortar\_brick1:3   
 1.611991e+00   
 has\_superstructure\_cement\_mortar\_brick1:4   
 8.674456e-01   
 has\_superstructure\_rc\_engineered1:1   
 1.406689e+01   
 has\_superstructure\_rc\_engineered1:2   
 1.120482e+01   
 has\_superstructure\_rc\_engineered1:3   
 1.317406e+01   
 has\_superstructure\_rc\_engineered1:4   
 -9.728167e-01

When performing the goodness of fit test, the deviance was 3282.15, and the degrees of freedom was 19,916 resulting in a p-value of 1. This means that the null hypothesis could not be rejected therefore, the model is a good fit. To test the accuracy of the model k-fold stratified cross validation with k=5 was implored. The average accuracy rate was 43.26% giving an average misclassification rate of 56.74%.

**Base-category Logit Model with 2 predictors**

Since the accuracy rate was low, a second model was created that only contained the following two predictors: age\_bulding and height\_ft\_pre\_eq with the hope that the second model would increase the accuracy rate. The coefficients of this model were:

(Intercept):1 (Intercept):2 (Intercept):3 (Intercept):4   
 1.203802717 0.059711900 -0.341423746 -0.545791982   
 age\_building:1 age\_building:2 age\_building:3 age\_building:4   
 -0.092173051 -0.029722865 -0.015176074 -0.001983580   
 height\_ft\_pre\_eq:1 height\_ft\_pre\_eq:2 height\_ft\_pre\_eq:3 height\_ft\_pre\_eq:4   
 -0.091702998 -0.051586359 -0.008874666 0.008987534

The deviance and degrees of freedom of this model was 14321.99 and 19,988, respectively. Once again, the p-value was 1, meaning that the model is a good fit. However, the accuracy rate was rate did not improve but decreased. The average accuracy rate of this model was 38.88% resulting an average misclassification rate of 61.12%. Since the accuracy rate did not improve, the first model was proposed model for the Baseline-Category Logit Model.

**Conclusions & Recommendations**

|  |  |  |  |
| --- | --- | --- | --- |
| Response variable (# of Categories) | Model | Misclassification Rate | Accuracy Rate |
| 2 | LDA | 17.7% | 82.3% |
| 5 | LDA | 62.2% | 37.8% |
| 5 | Classification Tree | 58.8% | 41.2% |
| 5 | Baseline-Category Logit | 56.7% | 43.3% |

Table 1 The misclassification and accuracy rate of the different models

From Table 1 one can see that the model with the highest predictive accuracy for all 5 damage grades is the baseline-category model. However, having a 43.3% accuracy is not very useful for predictive purposes and would not be recommended for being applied in other scenarios. This investigation is useful in that it revealed which variables are not very useful for predicting damage. Several columns were dropped throughout this investigation as they did not provide useful indicators in predicting damage. In addition, the data used is very coarse in that there are many different factors that have been grouped into a single category for almost all the categorical variables in the dataset. This issue is present in majority of the categorical variables in this dataset. This is especially problematic for the damage\_grade variable as there is likely a wide range of different types of damage that are grouped into these categories. Having a more detailed dataset with a wider range of variables would help in improving model accuracy. Therefore, more data is needed to create a predictive model for damage to buildings. Other models and groupings could also be used to improve the predictive accuracy of the model. For example, what was done for the LDA.2 model or to use other modelling techniques like the random forest.

**Responsibilities**

Data cleaning, data exploration and sampling methodology: Yu Nakamura

Prediction model analysis:

* Linear Discriminant Analysis: Graeme Kempthorne
* Tree Classification Model: Li Lam
* Multinomial Regression Model: Albert Leung

Editing and reviewing: Yu Nakamura, Graeme Kempthorne, Li Lam, Albert Leung

**References**

Goda, K., Kiyota, T., Pokhrel, R. M., Chiaro, G., Katagiri, T., Sharma, K., & Wilkinson, S. (2015). The 2015 Gorkha Nepal earthquake: insights from earthquake damage survey. *Frontiers in Built Environment*, *1*, 8.

Grandin, R., Vallée, M., Satriano, C., Lacassin, R., Klinger, Y., Simoes, M., & Bollinger, L. (2015). Rupture process of the Mw= 7.9 2015 Gorkha earthquake (Nepal): Insights into Himalayan megathrust segmentation. *Geophysical Research Letters*, *42*(20), 8373-8382.

Hall, M. L., Lee, A. C., Cartwright, C., Marahatta, S., Karki, J., & Simkhada, P. (2017). The 2015 Nepal earthquake disaster: lessons learned one year on. *Public health*, *145*, 39-44.

Mobius. Kaggle. Earthquake Magnitude Damage and . Impact. Retrieved September 20, 2021: (<https://www.kaggle.com/arashnic/earthquake-magnitude-damage-and-impact?select=ward_vdcmun_district_name_mapping.csv>)

Sheppard, P. S., & Landry, M. D. (2016). Lessons from the 2015 earthquake (s) in Nepal: implication for rehabilitation. *Disability and rehabilitation*, *38*(9), 910-913.

Shrestha, B., & Pathranarakul, P. (2018). Nepal government’s emergency response to the 2015 earthquake: a case study. *Social Sciences*, *7*(8), 127.