ORIE 4741 Project Midterm Report

Modeling Earthquake Damage

Vladia Trinh (vt95), Yoanna Efimova (yie3), Toshi Tokuyama (tt426)

Introduction

The dataset we are exploring is called "*Richter's Predictor: Modeling Earthquake Damage*," provided by the Driven Data website. [1] It is collected through surveys by Kathmandu Living Labs and the Central Bureau of Statistics, which works under the National Planning Commission Secretariat of Nepal. The dataset provides information on buildings in the region that was hit by the Gorkha earthquake. There are a total of 260,601 observations and 39 features, including building id, number of floors, age of the building, area and height of the building footprint, surface condition of the land, the position of the building, and others. [2] The features in our dataset include numerical, categorical, and binary values.

The purpose of this project is to predict the damage grade caused by the earthquake to various buildings in Nepal. To accomplish this, we first preprocess our data to make fitting a model possible. This includes handling missing values, standardizing features on a 0 to 1 scale, and converting categorical features to binary using one-hot-encoding. Next, we do some exploratory data analysis to see the correlation between the damage grade and our other variables to get a rough ranking of influential features. We also create some plots to see the distribution of specific features. Lastly, we train a basic ordinal regression model using our preprocessed dataset to get a baseline accuracy to improve upon in our final report.

Data Cleaning / Preprocessing of Data

First, we had to check if the dataset was in the appropriate format for future analysis. There were no missing or NA values in the dataset. We also checked the variable types to ensure that Python and R would interpret them correctly. In doing so, we converted the categorical variables (*land_surface_condition, foundation_type, roof_type, ground_floor_type, other_floor_type, position, plan_configuration, legal_ownership_status*) to factor variables with the appropriate number of levels and then created that many dummy variables. However, because the number of dummy variables for each categorical variable was equal to the number of levels of that variable, the columns were linearly dependent. As a result, we had to delete one of the dummy variables for each categorical variable to ensure linear independence.

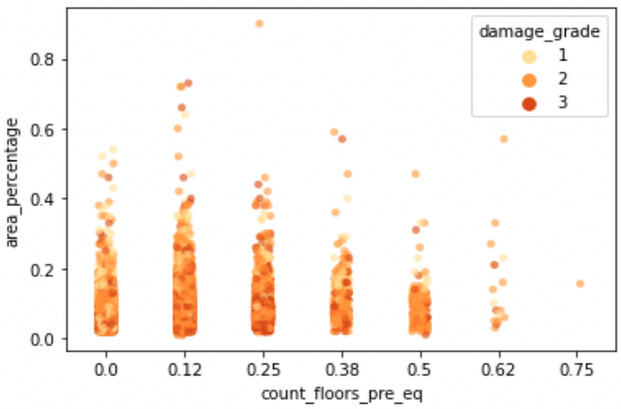
To make sure that each feature is properly weighted equally, we did some standardizing. For example, for *count_floors* and *age*, we used min-max scaling, which means that the minimum value in our dataset is 0 and the maximum is 1, and all values in between are scaled accordingly. We may want to experiment with a different type of scaling later on since the distribution of *age* is not very uniform.

Next, we checked for variables that did not have an impact on the damage grade of each building. We removed the variable *building_id* because it was clear that it was not related to the *damage_grade* variable. After removing that variable, we were ready to continue on with our preprocessing.

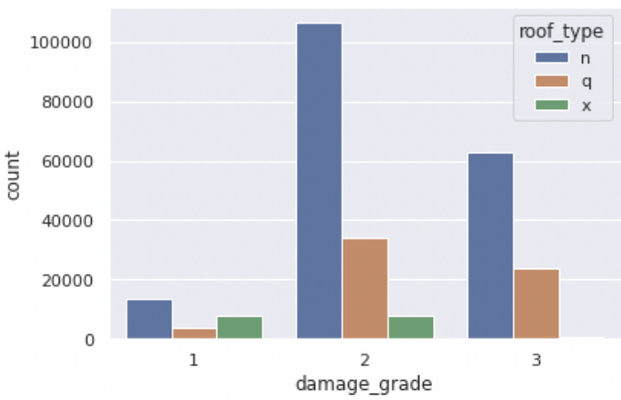
Exploratory Data Analysis

In addition to predicting the damage grade of each building, we wanted to understand how each attribute relates to the earthquake damage on the building. We wanted to see if there was any relationship between the damage grade and buildings' area, age, roof, foundation, and floor type, the position of the building, etc.

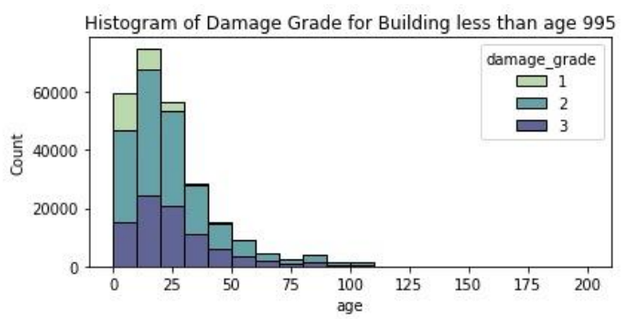
To start, we created a correlation plot for the variables to get a better sense of the correlations between each pair of features and to find which variables had the highest influence on the *damage_grade* variable. By looking at the heatmap (Fig. 1 in Appendix), we were able to notice that the dependent variable was highly correlated to the variables *area_percentage*, *has_superstructure_cement_mortar_stone*, *has_superstructure_cement_mortar_brick*, *has_superstructure_rc_non_engineered*, *has_superstructure_rc_engineered*, *has_secondary_use*, *has_secondary_use_hotel*, and *has_secondary_use_rental*. The *area_percentage* variable refers to the normalized area of the building footprint. Some of the other variables indicate if the superstructure was made of cement mortar – stone or cement mortar – brick and whether it was made of non-engineered or engineered reinforced concrete. The *has_secondary_use* variable is a flag variable that indicates if the building was used for any secondary purpose such as a hotel (*has_secondary_use_hotel*) and/or rental (*has_secondary_use_rental*).

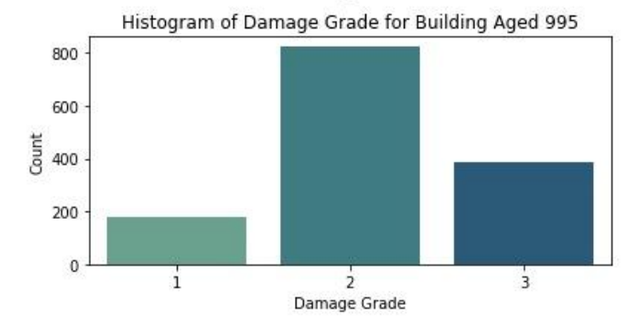


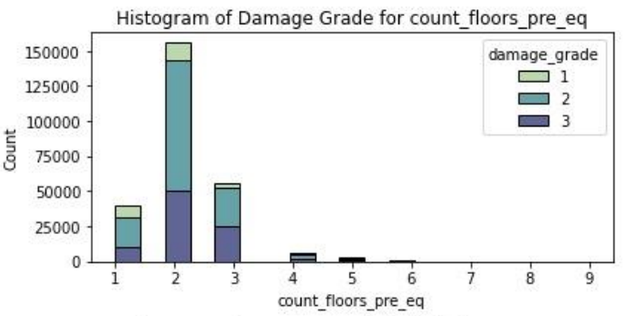
From our preliminary correlation analysis of *damage_grade* with our other features, we noticed that *count_floors* was positively correlated and *area_percentage* was negatively correlated with *damage_grade*. So, we created a scatter plot of these two features, including a dimension for *damage_grade*. From the graph, these correlations do not look too significant, but there seems to be a higher *damage_grade* centered around x = 0.25 and a lower *damage_grade* as *area_percentage* increases.

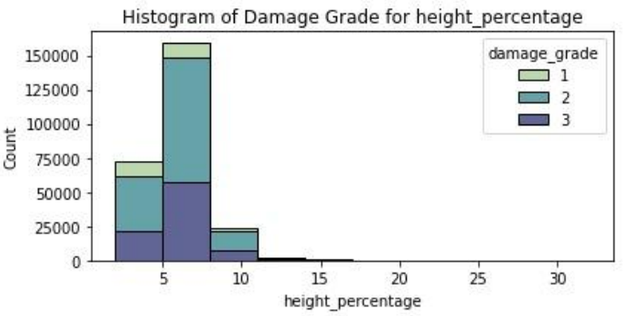


We also created a bar graph (Fig. 2) to examine the relationship between the buildings' roof type and the damage grade. By looking at the graph, we were able to see that the most stable roof type was type *x*. The most unstable one was roof type 'n' with more than 100,000 buildings of damage grade 2 and more than 60,000 buildings of damage grade 3. Moreover, we could see that there were also many buildings with roof type *q* that had more severe damages of grades 2 and 3.









In addition, we created 4 more histograms for the three variables: *age*, *count_floors_pre_eq,* and *height_percentage* (The top right histogram is for when age is 995). The histogram shows that our data is highly skewed; therefore, we should apply a transformation, so the variables become normally distributed.

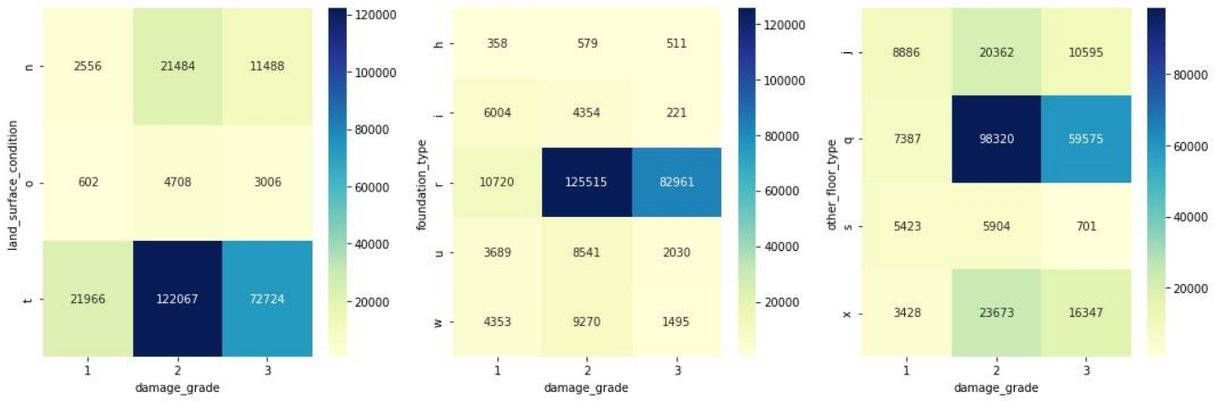
The histogram of *age* shows some buildings that are very old (greater than 100) but do not have a large count. Some buildings were aged 995, which are most likely historic buildings. We will need to consider whether to remove buildings that are aged 995 because the second oldest building is around 300 years old. The large difference between these values can add more complexity to our models.

The histogram of *count_floor_pre_eq* suggests that the variable can be turned into a categorical variable. Since a house with 2 floors seems to be the most common, dummy coding can be used in which *count_floors_pre_eq=2* is set as the base. Then if we have a linear model, we can compare the coefficients with the 2nd floor to make interpretations.

The histogram of *height_percentge* shows that most of the *height_percentage* is between the range 2% ~ 13%. The data is highly skewed, so transformation should be applied.

Overall, the histograms show that most of the damage grade is at the 2nd and 3rd level. Therefore, the challenge would be to predict houses that have *damage_grade = 1*.

To further explore some of the categorical variables, we focused on three variables: *land_surface_condition*, *foundation_type*, and *other_floor_type* with respect to the damage grade. From the three mosaic plots, we can see clear trends that certain types of floors, foundations, and land surfaces are more susceptible to earthquakes than others. However, the plot does not provide a clear picture of the relationship between the damage grades. The ratios between the damage types are similar for each variable.



Preliminary Regression Analysis

For our initial model, we chose to use an ordinal regression model because the variable that we wanted to predict (*damage_grade*) is categorical with ordered categories: 1, 2, 3. [3] We first split the data into 75% training set and 25% test set and preprocessed the data as described previously. We only trained the model on the subset of features we found to be important during our exploratory data analysis to prevent overfitting and reduce computational complexity. To measure the performance of the model, we used accuracy, which is just the number of correct predictions divided by all predictions. In the future, we may use a different performance measure since predicting *damage_grade* 1 when it is actually 3 should be graded more harshly than when the *damage_grade* is actually 2. Our model performed similarly on both the training and test data, with an accuracy of around 57%. This is quite low, which is expected since we only used the default parameters of the python package. Later on, we will perform some validation to tune parameters and allow us to include more features while preventing overfitting.

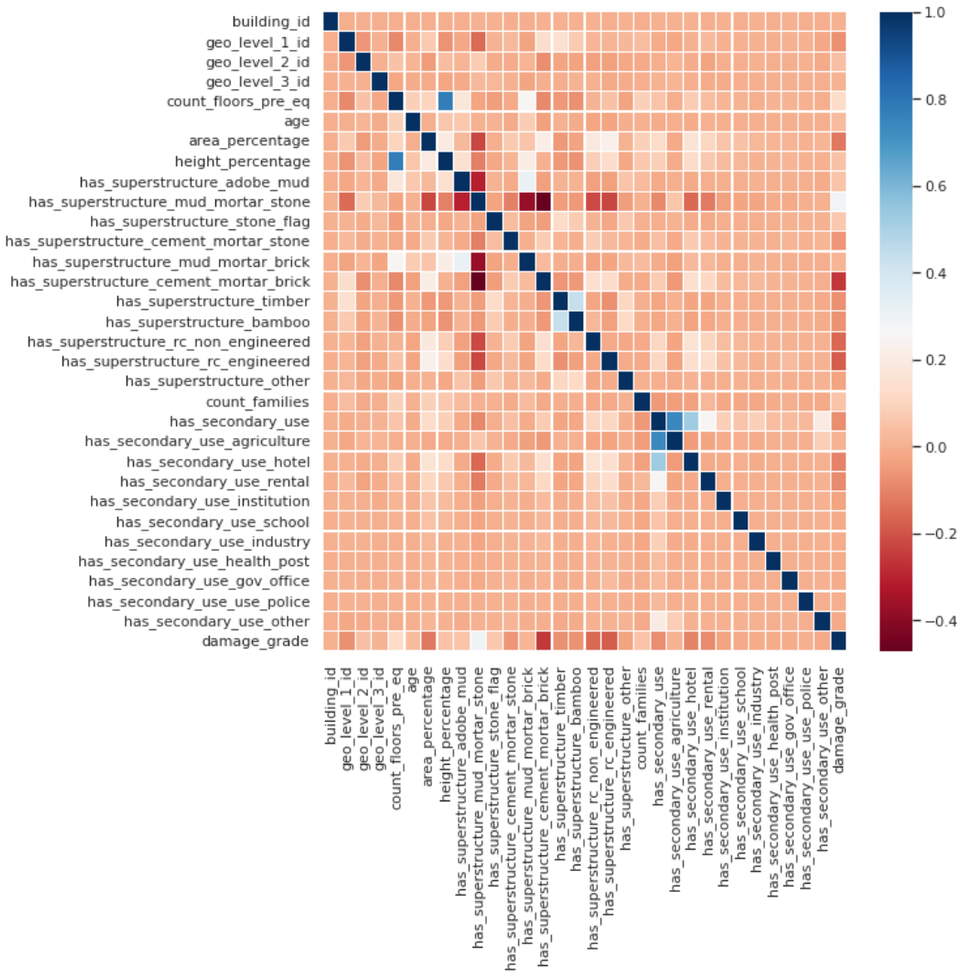
Further Steps

Going forward with our analysis, we will compare models with different numbers of features and then compare the metrics we get from the various models in order to avoid overfitting/underfitting. We can do this using the k-fold cross validation discussed in class. We will partition the data into k subsets and train the data on k-1 folds while using our remaining fold as the test set for validation. To further avoid overfitting, we hope to add regularizers to the models. We will try both the ridge and lasso regularization model and see which has the best accuracy while remaining sparse. Regularizers will also help with reducing the variance of the models.

Since the target variable is an ordinary variable, we used an ordinary regression model in our preliminary regression analysis. Another more simple possibility is using a multi-class perceptron. However, it is difficult to encode ordinal data in a way that preserves its rank instead of just being categorical, so this model may have poor prediction power. We can explore further by using other models such as random forests. Our group believes that random forest will work well because random forest allows us to get the variables that are contributing to the model the most (feature importance). Being able to understand significant variables will not only improve our prediction but will also allow us to understand the crucial factors of earthquake damage.

Appendix:

[Figure 1]



Bibliography:

[1] “Richter's Predictor: Modeling Earthquake Damage”

https://www.drivendata.org/competitions/57/nepal-earthquake/page/134/

[2] Features descriptions, “Modeling Earthquake Damage”

https://www.drivendata.org/competitions/57/nepal-earthquake/page/136/

[3] Ordinal Regression Documentation

https://www.statsmodels.org/dev/examples/notebooks/generated/ordinal_regression.html