Analysis of Earthquake Damage, Nepal

Microsoft Professional Capstone: Data Science
Robert Ritz, April 2018

Executive Summary

The 2015 Gorkha earthquake in Nepal caused catastrophic damage to many homes and structures in the region. This paper is an analysis of one of the largest post-disaster datasets ever collected. It contains demographic statistics and building conditions down to the level of individual structures. By applying machine learnings tools to this problem, the author hopes a model can be created that will be generalizable to other locations facing similar earthquake risk.

The author will present a clear picture of this large dataset through visualizations and descriptive statistics. After this exploration and analysis, a classification model is presented that aims to predict the damage to buildings from dataset.

The most significant features in the dataset that are useful in predicting damage categories are:

- **Geo_level_1_id** – the largest geographic region where individual buildings are located
- **Geo_level_2_id** – the next geographic region in size. Smaller than geo_level_2_i
- **Age** – age of the building in years
- **Area** – plinth area of the building in square meters
- **Superstructure type** – Adobe/Mud and Cement Mortar/Stone were the most important superstructure types
- **Foundation type** – Two foundation types were better indicators of damage categories

Key findings:

- The author theorizes that the location of the earthquake caused an outsized impact on specific geographic areas. Certain Geo 1 and 2 levels had a larger proportion of damage grades 2 & 3.
- Buildings made of mud, mortar, or stone had the highest levels of damage.
- Buildings younger than 20 years of age had a lower level of damage on average.
- Larger buildings had a lower damage grade on average.
- Damage grade 2 and 3 are difficult to separate in the model.
- The overall accuracy achieved by the model is sufficiently generalizable as to be reliable when deployed in Nepal.

Data Exploration

Data exploration will begin with a listing of features as well as some descriptive statistics. There are 40 features to begin with in the dataset (including our label feature, damage_grade). The list below is separated into categorical and numeric features.

Numeric features:

- Count of floors pre earthquake (count_floors_pre_eq)
- Age of building in years (age)
- Plinth area of building in meters squared (area)
- Height of building in meters (height)
- Count of families living in building (count_families)

Categorical features:

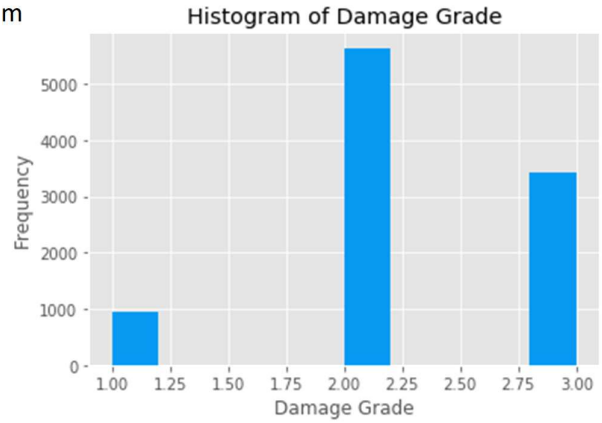
- Geographic levels 1-3 (geo_level_1_id, geo_level_2_id, geo_level_3_id)
- Land surface condition of building (land_surface_condition, 3 distinct values)
- Foundation type of building (foundation_type, 5 distinct values)
- Roof type (roof_type, 3 distinct values)
- Ground floor type (ground_floor_type, 5 distinct values)
- Type of floors for those above ground floor (other_floor_type, 4 distinct values)
- Position of building (position, 4 distinct values)
- Building plan configuration (plan_configuration, 9 distinct values)
- Legal ownership status of the land (legal_ownership_status, 4 distinct values)
- Superstructure type. This is given as 11 separate binary features. However, grouped together they create 11 distinct categorical values. (has_superstructure_)
- Secondary use. This is also given as 11 separate binary features. Grouped together they create 11 distinct categorical values. (has_secondary_use_)
- Damage grade. 1 represents low damage, 2 represents a medium amount of damage, 3 represents almost complete destruction

Descriptive Statistics and Visualizations for Numeric Features

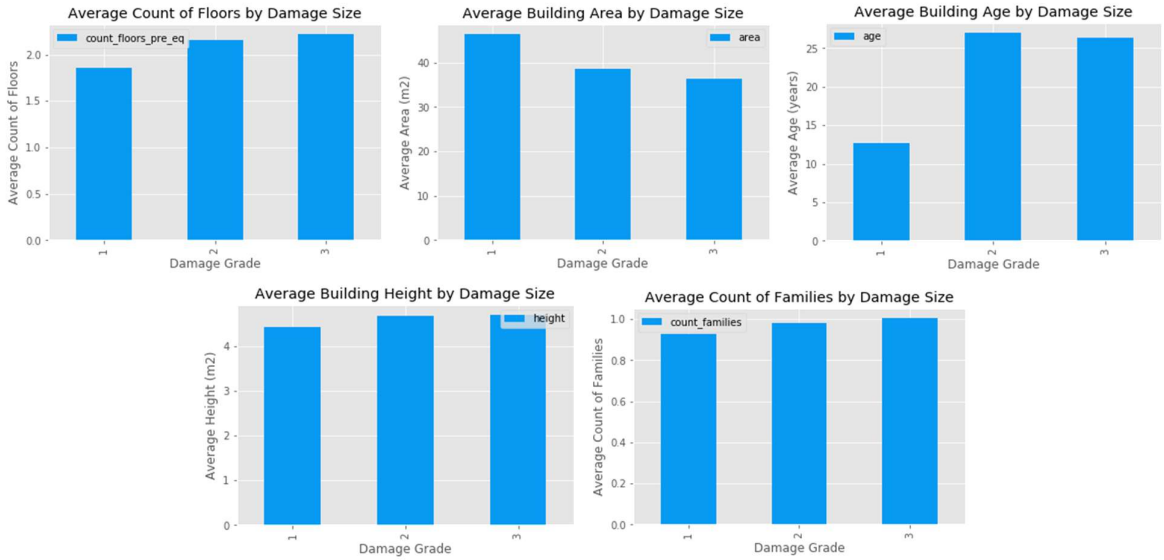
Our training dataset has a total of 10,000 records. The count, mean, standard deviation (std), minimum, quartile (25%, 50%, 75%), and maximum are given for each numeric feature.

	count_floors_pre_eq	age	area	height	count_families	damage_grade
count	10000	10000	10000	10000	10000	10000
mean	2.1467	25.3935	38.4381	4.6531	0.9846	2.2488
std	0.736365	64.482893	21.265883	1.792842	0.423297	0.611993
min	1	0	6	1	0	1
25%	2	10	26	4	1	2
50%	2	15	34	5	1	2
75%	3	30	44	5	1	3
max	9	995	425	30	7	3

Our label feature is damage_grade. The following histogram shows that over half of the records are grade 2. It is also important to note there are very few buildings with a damage grade of 1. This skew in our dataset can be found later in our model results.



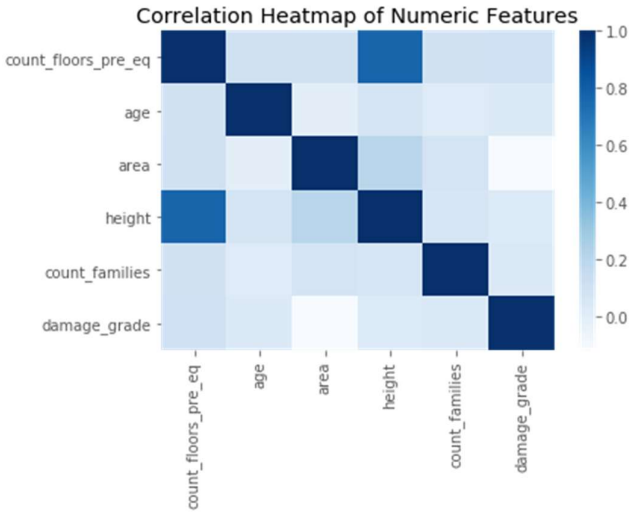
Before we move on to the categorical features we can visualize our numeric features with our damage grade feature to better understand their interaction. One simple way to do this is to find the average value of each feature at each damage grade. These are shown below.



Building area and building age both show clear trends with respect to damage grade. Younger buildings appear to have a lower damage grade. Also larger buildings with regard to plinth size have a lower damage grade.

One especially useful tool in finding relationships between variables is a pairwise plot. This visualization plots each numeric feature against every other. The diagonal line shows the histogram of the feature. In addition the hue of each damage grade is shown for each point. The lighter color is a damage grade of 1 while the darker color is a damage grade of 3.

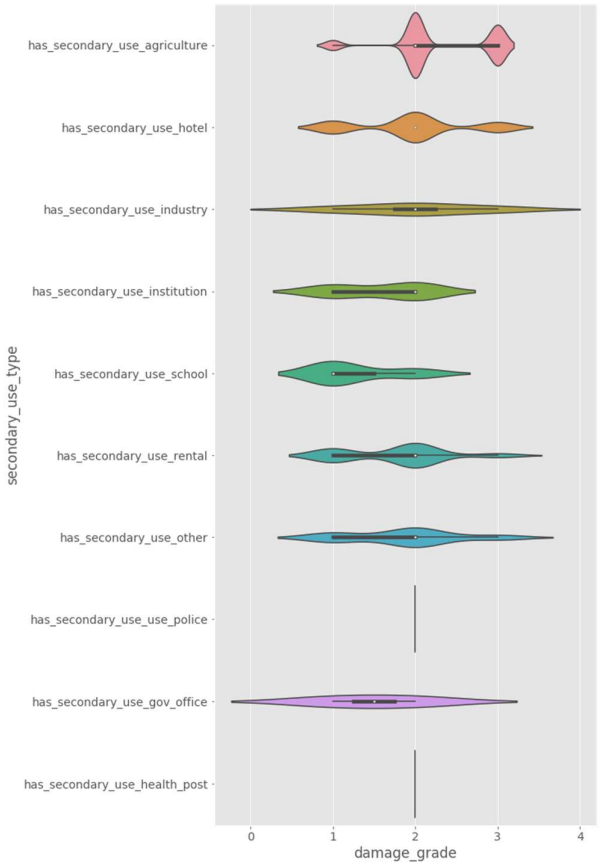
Data Analysis and Visualizations for Categorical Features

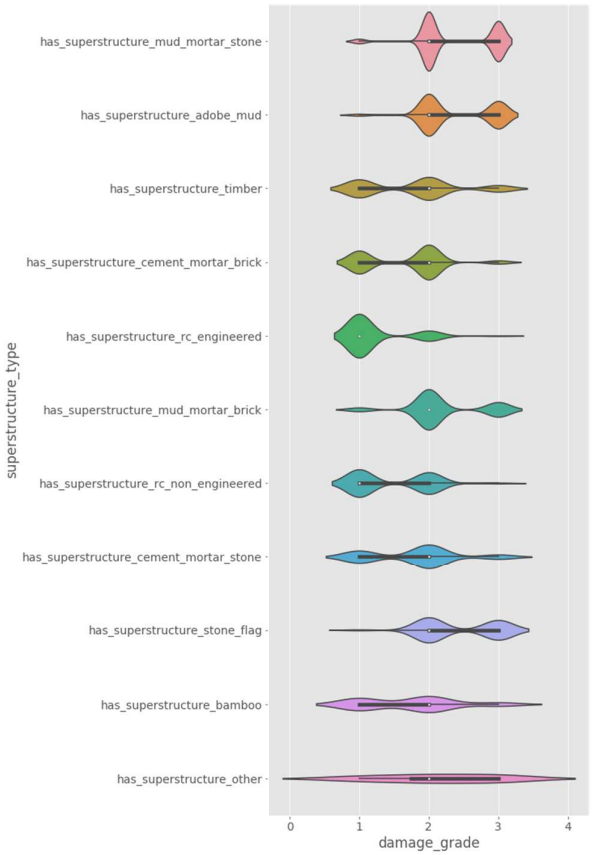


The majority of features in our dataset are categorical. As we are trying to predict damage grade, we will explore the relationship of our categorical variable to damage grade. It would also be possible to create visualizations to compare our categorical features to each numeric feature, however for the purposes of this report it is excessive.

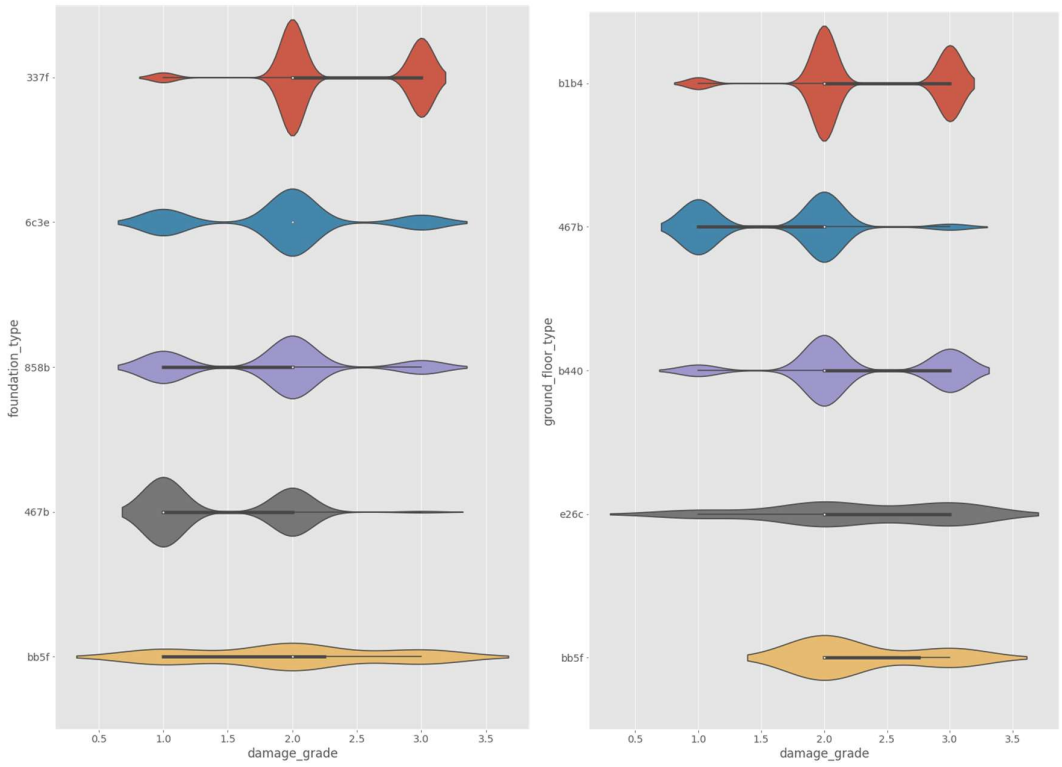
The author will present these relationships through violin plots. These are superior to standard box plots in that they also show data frequency as well as mean and the interquartile range. They can be understood as a combination of a frequency distribution plot and a standard boxplot. The following plots show the relationship between our categorical features and damage grade (excluding geo levels 1-3):

Superstructure and Secondary Use Type

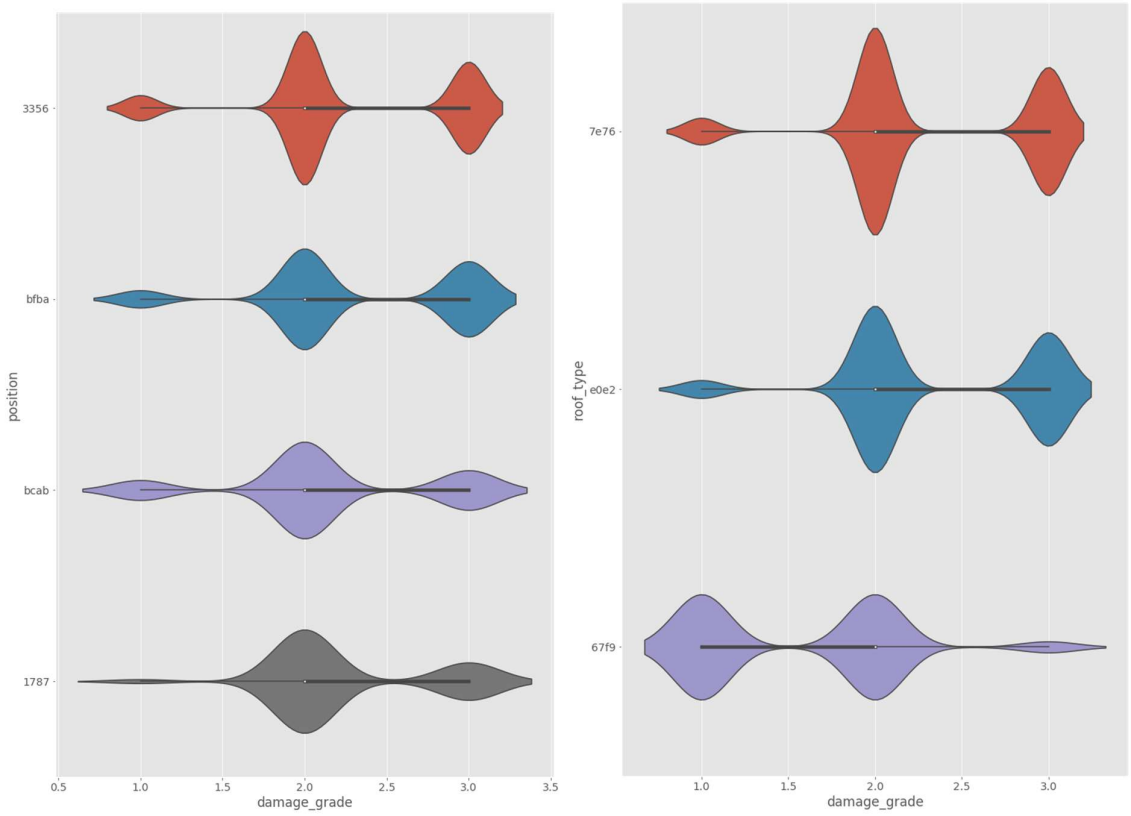




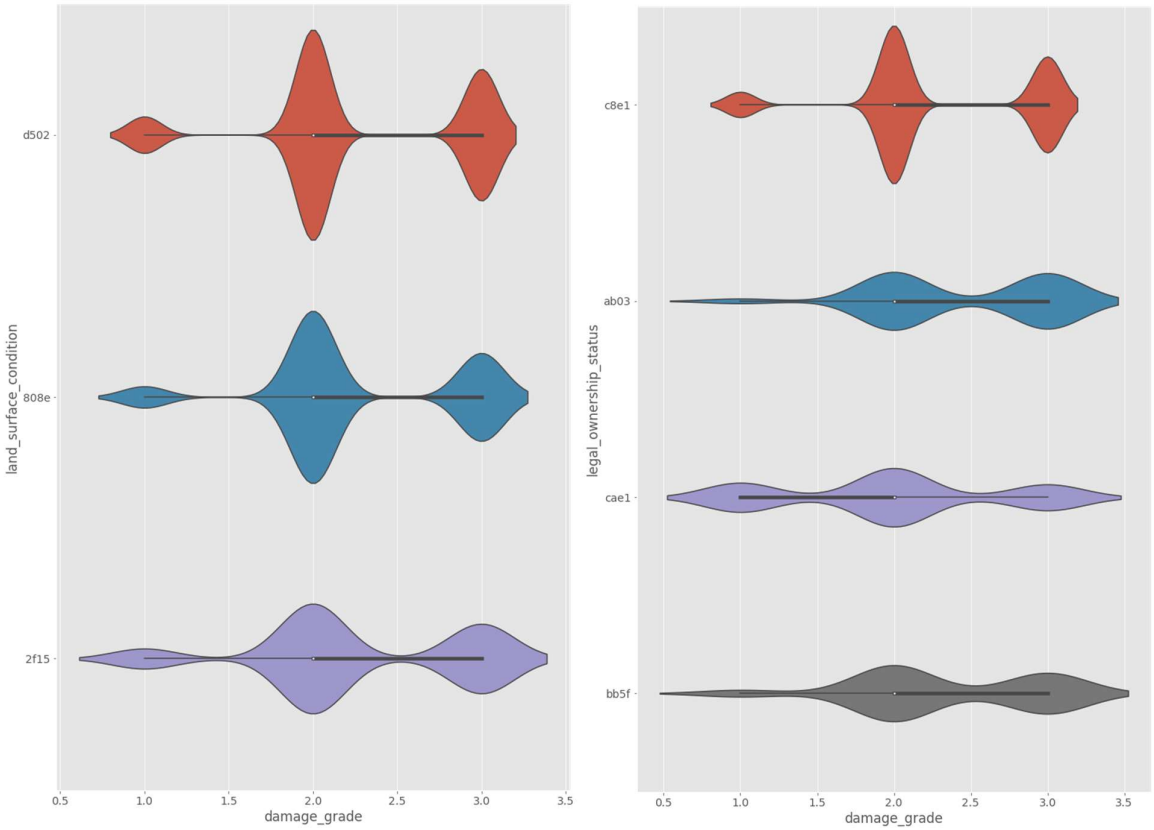
Foundation and Ground Floor Type

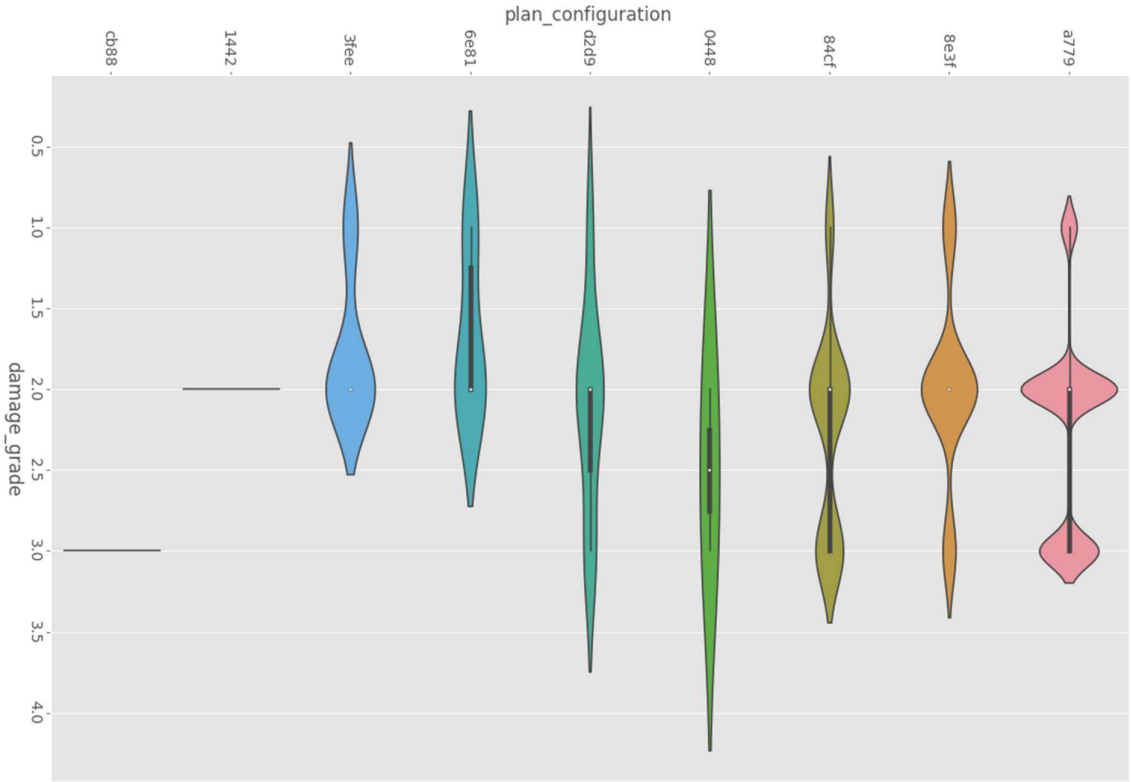


Land Position and Roof Type



Legal Ownership Status, Plan Configuration, and Surface Condition



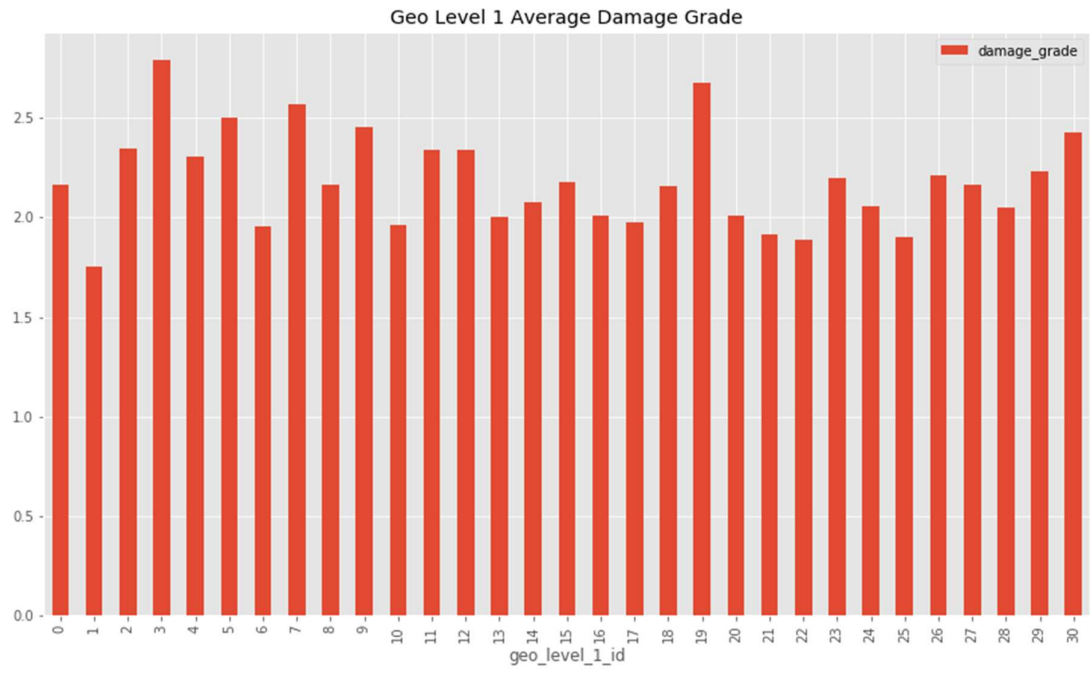


The main interpretation from these plots is that while some features do have a clear skew towards one damage grade or another, many do not. A few determinations made from the categorical feature visualizations:

- Superstructure type has a clear predictive ability. The mud/mortar/stone and adobe/mud categories are highly skewed towards damage grade 2 and 3. In contrast rc-engineered and cement/mortar/brick have a skew towards grade 1 and 2.
- Secondary use has a less clear relationship. However institutions, schools, and government buildings do trend towards a lower damage grade.
- Other categorical features are encoded in a way as to obfuscate their true values. This makes an analysis using subject matter knowledge impractical.
- These other categorical features also appear to have similar distributions as noted in superstructure type and secondary use.
- Many of the features items have very low variance. For example plan configuration cb88 and 1442 only occur at one damage grade.
- There is a clear different between damage grade 1 vs 2&3. For example the superstructure type mud/mortar/stone has a very high frequency at grades 2/3 and a very low frequency at grade 1. This is a common relationship seen through many of these features.
- Damage grade 2 and 3 show a large similarity and may be difficult to separate on our model.

Geo levels 1-3 were excluded in the above plots. As there are large number of geo levels we will use a different plotting method, bar plots. Below are the average damage level by geo ID. Only geo level 1 is shown. Geo level 2 has 1,137 unique values, so showing those in a visualization is impractical for the purposes of this report.

The training set has 5,172 unique values for geo level 3. The test set contains the same number of unique values, but there are over 1,000 values in the test set that are not contained in the training set. This signals that geo level 3 is too small of a geographic area to be generalizable across regions. Therefore it was determined that geo level 3 is too granular to have predictive value and was removed from our dataset before training.



There is clearly a difference in damage by geographic area. A few assumptions can be made first:

- As an earthquake occurs in one location (the epicenter) and radiates out from that location, we can assume the immediately surrounding areas from the epicenter will be more badly damaged.
- We don't know the organization of these geo levels. They do not appear to be ordered by distance from the epicenter. As such we can assume they are randomly ordered.

With these assumptions we can determine the following:

- As we don't know the spatial locations of the geo levels, we won't be able to do any spatial analysis.
- However there is clear predictive value in the geo levels. There is significant variation in damage between these levels.

Data Wrangling

The following steps were taken to preprocess the data before training:

- Joining of the training labels and values to create one dataframe. In scikit-learn this dataframe was used for EDA and visualization. However it was again separated into features and labels for scikit-learn.
- Converting all non-numeric features to categorical variables. This step is specific to Azure ML and was not completed in scikit-learn.
- Converting geo_level_1_id and geo_level_2_id to categorical variables
- Dropping geo_level_3_id as it was determined to be too small a geographic area to be useful
- Dropping building_id. This is an erroneous column and only used for identifying the record.
- Using one-hot encoding ("Convert to Indicator Values" module in Azure ML) of all categorical variables (including geo_level_1_id and geo_level_2_id. This method creates a binary value (0 or 1) for each categorical value.

The resulting dataset contained 1,233 columns and 10,000 rows. Given the large dataset we should not fall victim to the curse of dimensionality. This dataset was fed to the classification models detailed below.

Classification Model

Now that we have a better understanding of our data, we can begin creating our machine learning model. Our objective is to create a model that can correctly identify the damage grade of our test set. Our error will be measured using the F1 micro averaged metric (our unofficial goal is to meet or exceed 0.7 F1 micro averaged score). The training set has 10,000 records and our test set has 10,000 records.

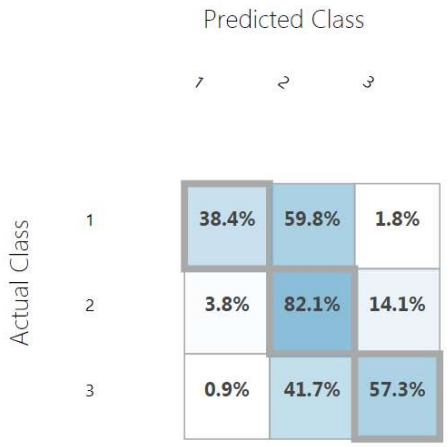
Many different models were attempted for this project. Both Scikit-learn in Python and Azure ML Studio were used. The results for each model are listed below. The Azure ML scores are with hyperparameter tuning. Scikit-learn scores have minimal tuning. All scores listed are the F1 micro averaged metric.

Algorithm	Location	Score
MLP Classifier (multi-class neural network)	Scikit-learn	.616
RandomForestClassifier	Scikit-learn	.622
AdaBoostClassifier (boosted decision tree)	Scikit-learn	.653
XGBoost (gradient descent)	Scikit-learn	.669

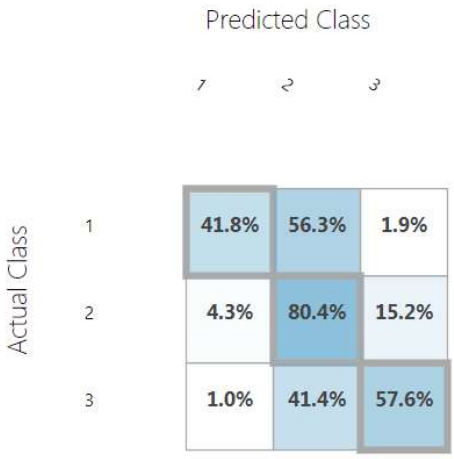
Multiclass Decision Forest	Azure ML	.688
Multiclass Decision Jungle	Azure ML	.671
Multiclass Logistic Regression	Azure ML	.690
Two Class Boosted Decision Tree (one vs. all method)	Azure ML	.658
Two Class Bayes Point Machine (one vs. all method)	Azure ML	.680
Two Class Logistic Regression (one vs. all method)	**Azure ML**	**.695**
Two Class Averaged Perceptron (one vs. all method)	Azure ML	.690

Our best model produced an F1 micro score of .695. The two highest scores were both one vs. all ensemble models. In order to compare models the F1 score is often not sufficient. As this is a multiclass problem, we will use a confusion matrix to better understand how these models performed. A confusion matrix plots the predicted class against the observed class. The resulting matrix shows how well our model performs on each class in our dataset.

Two Class Logistic Regression (best model)



Multiclass Logistic Regression



As we found in our exploratory data analysis, our biggest challenge is with damage grade 2 and 3. However the one vs all method is able to better differentiate between these and is able to get us quite close to our goal of a 0.7 F1 score.

In addition to our scores and confusion matrix, feature importance is another metric that helps understand our model performance and limitations. Below is the feature importance from our best model as calculated in Azure ML Studio through the Permutation Features importance module. Only the top 10 features are listed.

Feature	Importance Score
has_superstructure_cement_mortar_brick	0.029498
has_superstructure_rc_non_engineered	0.026283
area	0.023042
foundation_type-858b	0.016002
geo_level_1_id-3	0.012682
geo_level_1_id-2	0.009731
geo_level_1_id-17	0.009517
geo_level_1_id-21	0.007717
other_floor_type-441a	0.007547
geo_level_1_id-9	0.007019

The superstructure types are the most important predictors. Also it is clear that the geo level is quite important.

Conclusions and Recommendations

This analysis has shown that our best model is able to predict with sufficient accuracy the damage level to buildings in this earthquake. Our most predictive features are superstructure type (cement/mortar/brick and rc non engineered), area, founation type 858b, and geo level 1 ID (3, 2, 17, 21, and 9).

Damage levels of grade 2 and 3 are the most similar, and future work should include research on how to better differentiate these classes. In addition, feature engineering possibilities exist if more domain knowledge is able to be applied. This will require a better knowledge of our categorical features as our current dataset is obfuscated.

Given these recommendations the author believes the current model is able to reliably predict damage level and is suitable for deployment.