Richter's Predictor: Modeling Earthquake Damage

Introduction

A destructive earthquake of 7.8 magnitude occurred in Nepal in April 2015. This earthquake claimed almost 9,000 lives and around $10 billion in damages. Millions of people lost everything and became homeless in a few moments.

The goal of this project is to predict the level of damage to buildings based on building location and construction

Source Dataset

- Nepal carried out a massive household survey using mobile technology to assess building damage in the earthquake-affected districts.

- This survey is **one of the largest post-disaster datasets ever collected**, containing valuable information on earthquake impacts, household conditions, and socio-economic-demographic statistics.

- This survey data is used as source dataset for this project.

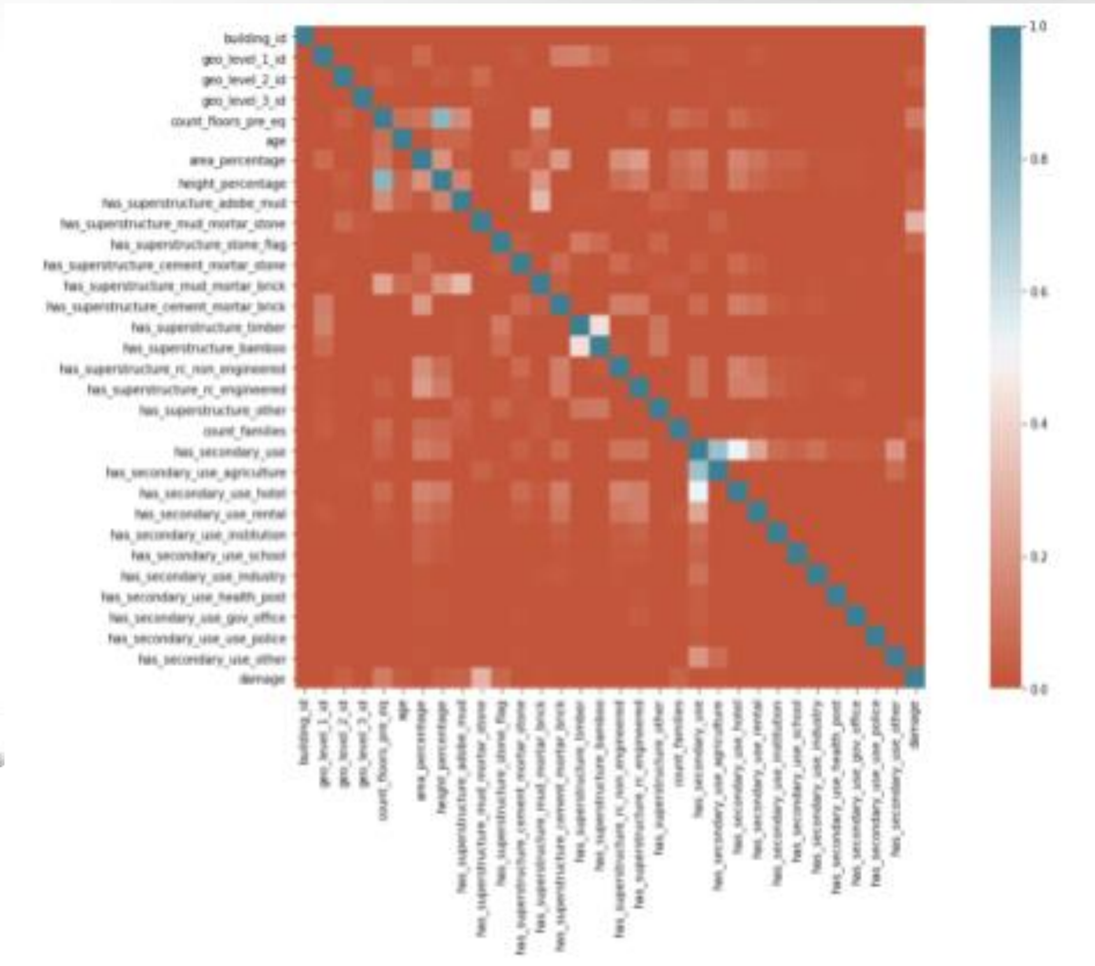
Exploratory Data Analysis

Dataset

	Field Name
0	building_id
1	geo_level_1_id
2	geo_level_2_id
3	geo_level_3_id
4	count_floors_pre_eq
5	age
6	area_percentage
7	height_percentage
8	land_surface_condition
9	foundation_type
10	roof_type
11	ground_floor_type
12	other_floor_type
13	position
14	plan_configuration
15	has_superstructure_adobe_mud
16	has_superstructure_mud_mortar_stone
17	has_superstructure_stone_flag
18	has_superstructure_cement_mortar_stone
19	has_superstructure_mud_mortar_brick
20	has_superstructure_cement_mortar_brick
21	has_superstructure_timber
22	has_superstructure_bamboo
23	has_superstructure_rc_non_engineered
24	has_superstructure_rc_engineered
25	has_superstructure_other
26	legal_ownership_status
27	count_families
28	has_secondary_use
29	has_secondary_use_agriculture
30	has_secondary_use_hotel
31	has_secondary_use_rental
32	has_secondary_use_institution
33	has_secondary_use_school
34	has_secondary_use_industry
35	has_secondary_use_health_post
36	has_secondary_use_gov_office
37	has_secondary_use_use_police
38	has_secondary_use_other
39	damage

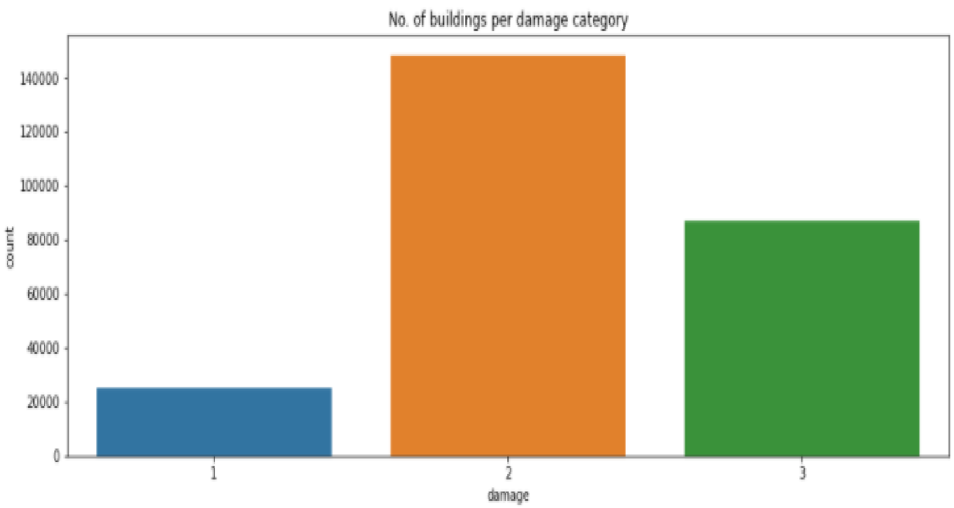
- There are 39 columns in this dataset, where
  the building_id column is a unique and random identifier.
- The data is semi-anonymized
- damage_grade represents a level of damage to the building that was hit by the earthquake.

Feature Correlation



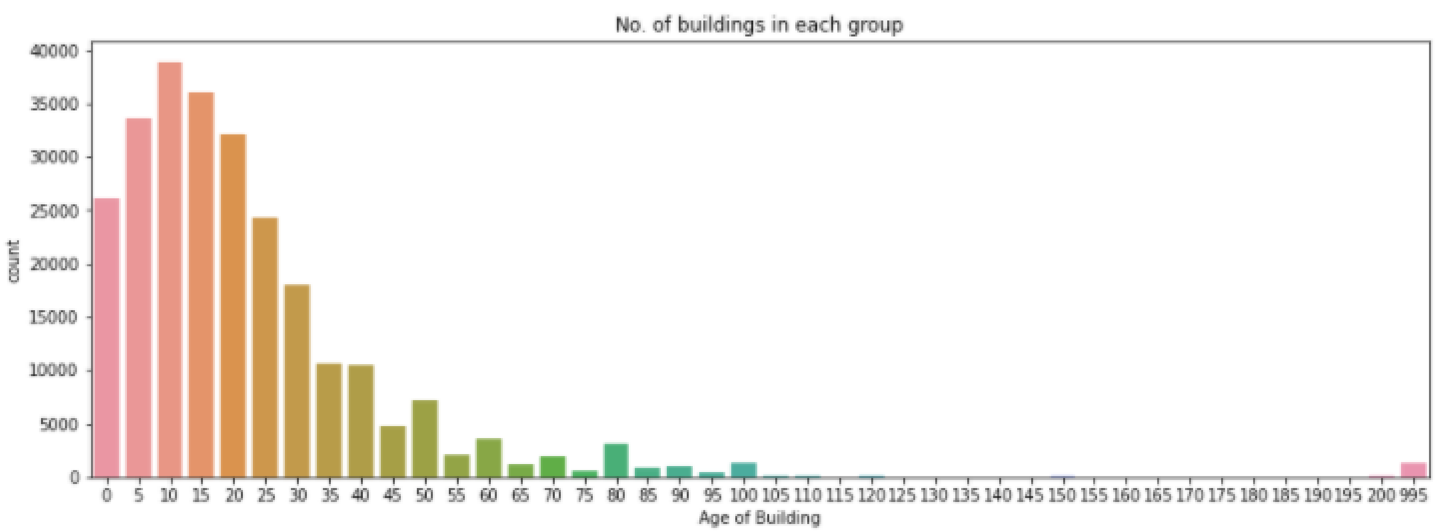
- There are not a lot of correlated fields.
- has_secondary_use is correlated with it's sub_parts and height_percentage is highly correlated with count_floors_pre_eq
- area_percentage and height_percentage are correlated with has_super_structure features and seconday use of buildings.

Target Variable



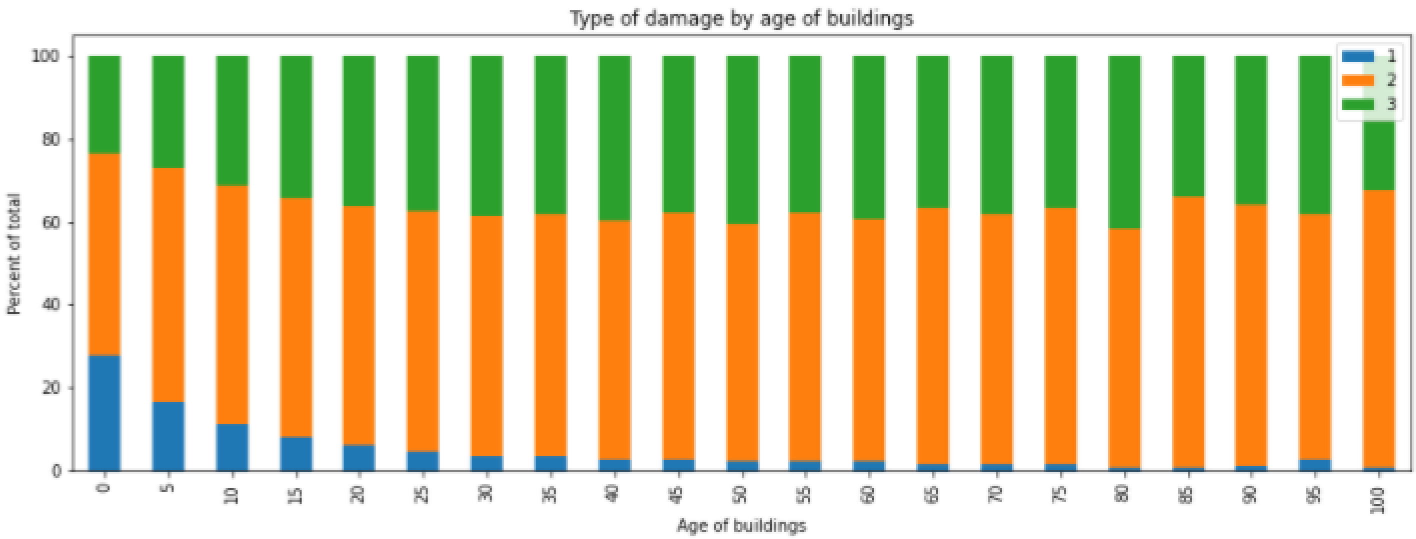
The training data is imbalanced

Age of buildings



- There are a significant no. of fairly new buildings (less than 10 year old).
- There might be something with the newer construction which was prone to more damage during earthquake or there just happened to be more new construction overall in the area.

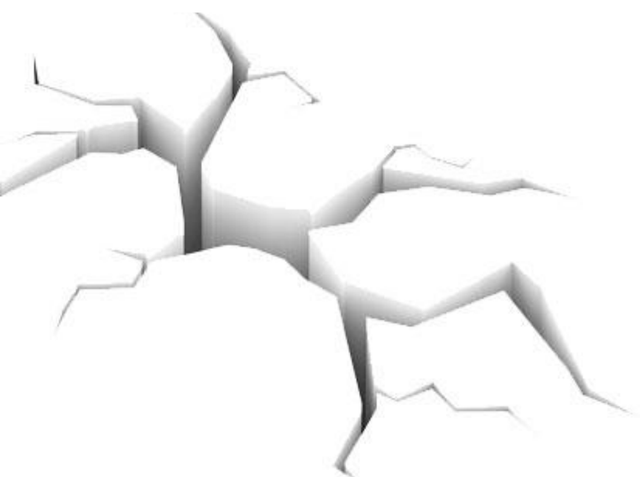
Damage by building age



It is interesting to notice that only newer constructions (buildings less than 5 year old) have lesser grade 3 damage than grade 1.

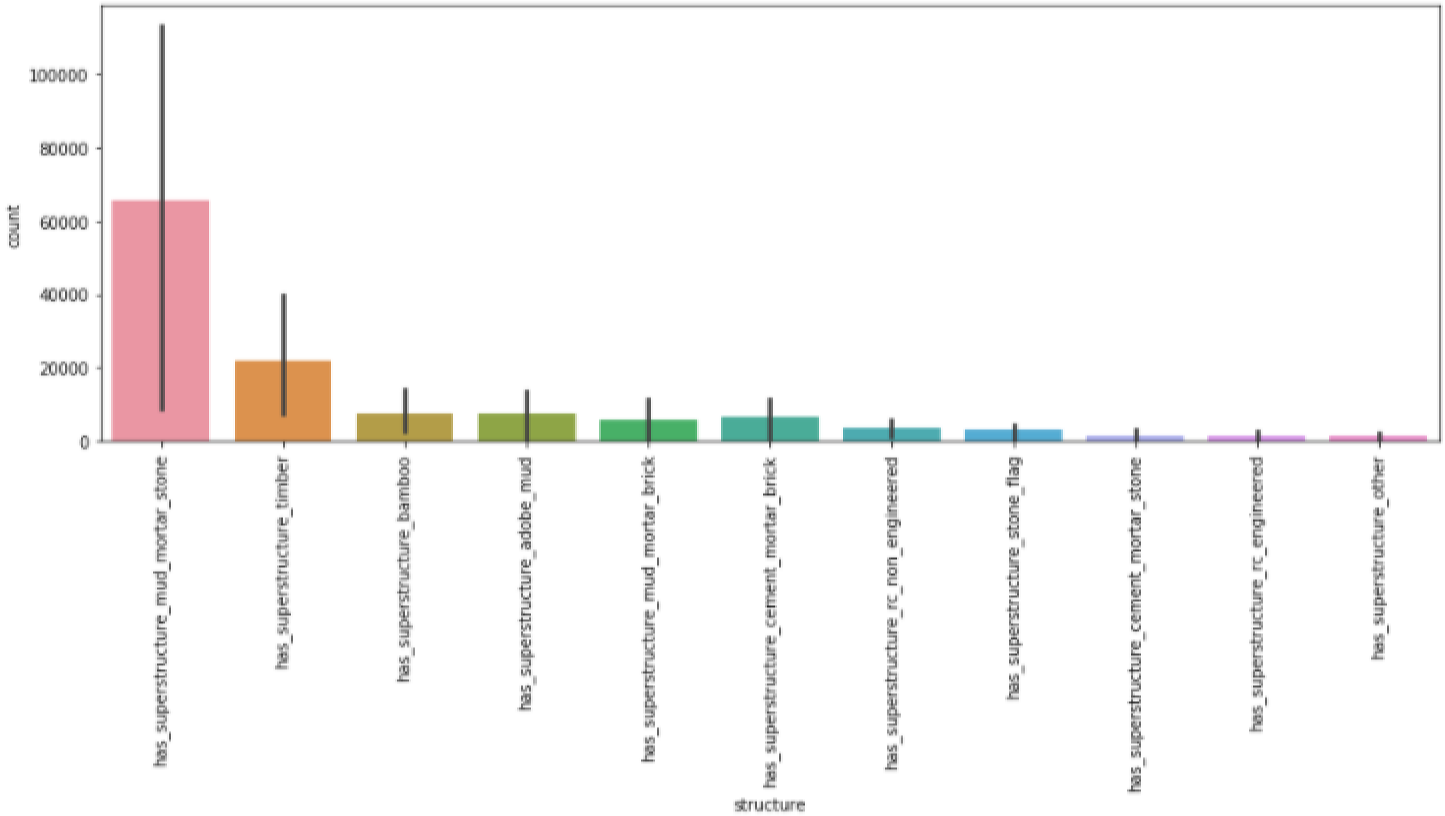
This suggests that either the newer constructions were sturdier or they were not in worse hit areas.

Damage to older buildings



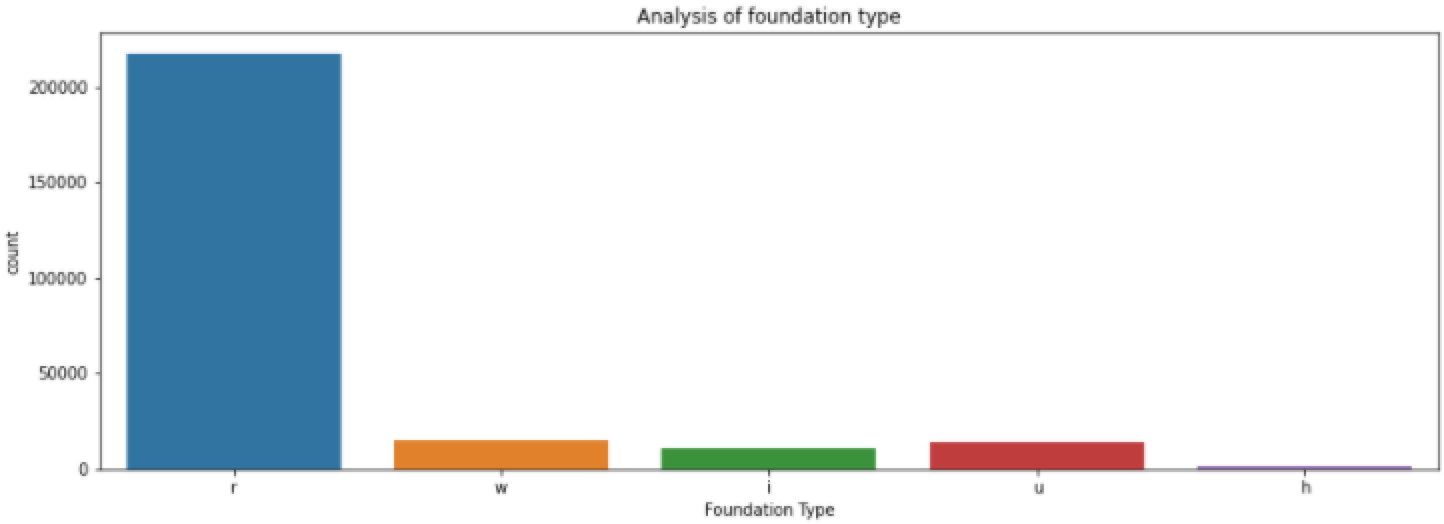
Older buildings could not tolerate the wrath of nature and bore medium to high damage.

Role of building material



Top 5 types of structures those got damaged the most were made up of timber, bamboo and some form of mud.

Role of type of building foundation



'r' type of foundation is the leading cause.

Machine Learning

I grid searched 4 model for hyper parameter tuning and XGBoost gave the highest F1 score

```
clf=XGBClassifier()

kf=KFold(n_splits=2,shuffle=True)

rs=RandomizedSearchCV(clf,param_distributions=param_grid,cv=kf,scoring='f1_micro')

rs.fit(X,y)
```

Scoring Metric

I have used F1 Score to measure the performance of the algorithms. F1 score balances the precision and recall of a classifier.

Traditionally, the F1 score is used to evaluate performance on a binary classifier, but since we have three possible labels we will use a variant called the micro averaged F1 score.

Imbalanced Class

Since the target variable is imbalanced, I tried upsampling as well as downsampling training data but it did not have any significant effect on F-1 score of test data.

Model Performance on Test data

With XGBoost model, I got F-1 score of 0.7434 on test data.

Next Steps

The data source for this project was semi-anonymized but full dataset is available on 2015 Nepal Earthquake Open Data Portal.

The full dataset can be used to apply domain knowledge and engineer more features for the model which in-turn can help with model performance.