Predicting Seismic Vulnerability of Buildings in Nepal

Ilona Laskowska

Abstract

The aim of this project is to investigate whether building survey data (from Nepal dataset) can be used to predict the seismic vulnerability of buildings and what are the building features which make it more earthquake sensitive.

The data analysis, visualization and machine learning task (classification model) have been performed using the pandas and scikit-learn python libraries.

Seismic vulnerability prediction turned to be at range of 65% which is quite fine result, however it might be further improved by trying out other model types.

Motivation

Most of existing buildings around the world do not meet the requirements of modern seismic design codes and are very vulnerable to the earthquake events. One of the example is a destructive earthquake of 7.8 magnitude occurred in Nepal in April 2015, after which millions of people became homeless in just a few moments.

Performing structural analysis for all buildings in the area with a high seismic risk would be very time consuming and costly. By using machine learning to identify the most vulnerable buildings and apply a strengthening solution before the earthquake event could safe some people’s houses and lives.

Dataset(s)

In order to run machine learning classification tasks, the competition dataset have been used: https://www.drivendata.org/competitions/57/nepal-earthquake/

The dataset consist of train_values, train_labels and test_values.

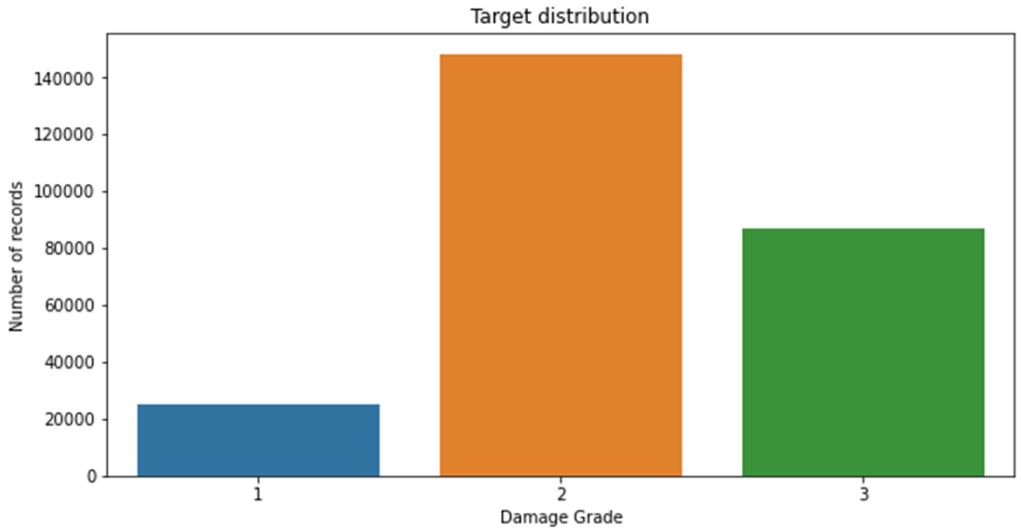
This data have been collected through the household surveys collected by the Central Bureau of Statistics and there are one of the largest post-disaster datasets ever collected, containing valuable information on earthquake impacts, household conditions and building structure (39 columns which represents different building features for approx. 260 600 buildings location)

Dataset(s)

The labels dataset contain the information about damage experienced by building due to the Gorkha earthquake:

- 1 = low damage
- 2 = medium damage
- 3 = almost complete destruction

The training data is relatively imbalanced with only 10% representing damage grade equal to 1.



Data Preparation and Cleaning, Problems

First, I have checked if any Null values exist or if there are any duplicated records. As the dataset is preprepared for the competition puropses there were no such cases.

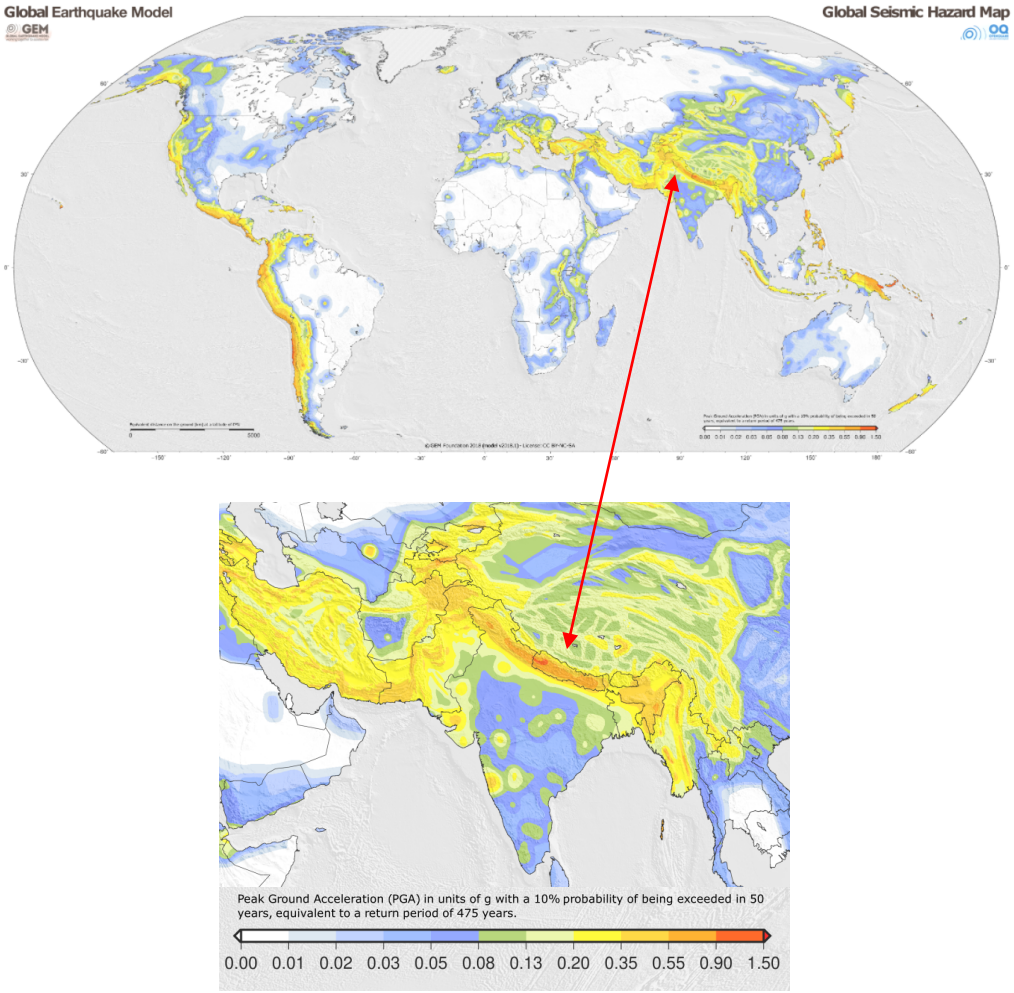
As I am a structural engineer having the domain knowledge allowed me to append to the dataset new parameter called slenderness (heigth_area_ratio) which is calculated by dividing heigth_percentage by area_percentage.

More slender buildings tends to experience higher overturning which increases the forces in the superstructure and can cause high damage.

Context

The seismic vulnerability depends both on the seismic hazard and the seismic resistance of the building.

Nepal is located at area of high seismic hazard, so the poor-quality of construction makes the buildings very vulnerable to seismic events.



Research Question(s)

1. Which building features has the highest importance for builing damage (age, area, height) ?
2. Can seismic vulnerability of buildings in Nepal be successfully predicted based on building location and construction features ?
3. Which of two supervised classification approches (kNN, Decision Tree) has a higher accuracy for such model?

Methods

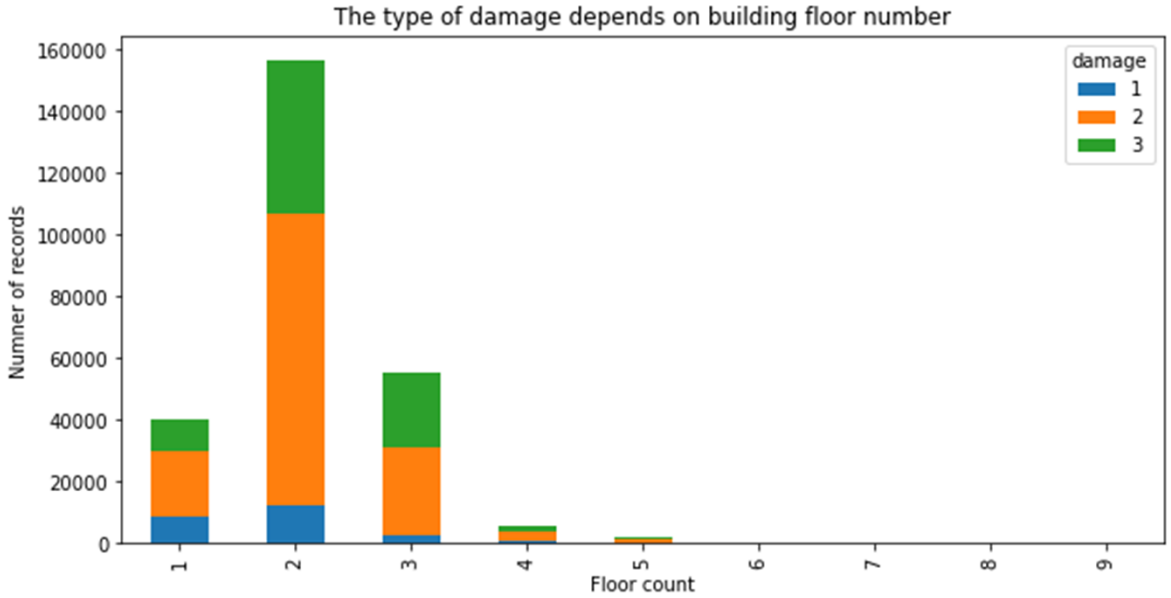
Predicting the level of earthquake impact on a building is a multiclass classification problem as there are three grades of damage (1-3). There were two machine learning models tested:

- kNN
- Decision Tree

The model performance have been measured using the micro averaged F1-score instead of the accuracy due to imbalanced classes. (as suggested at the competition website)

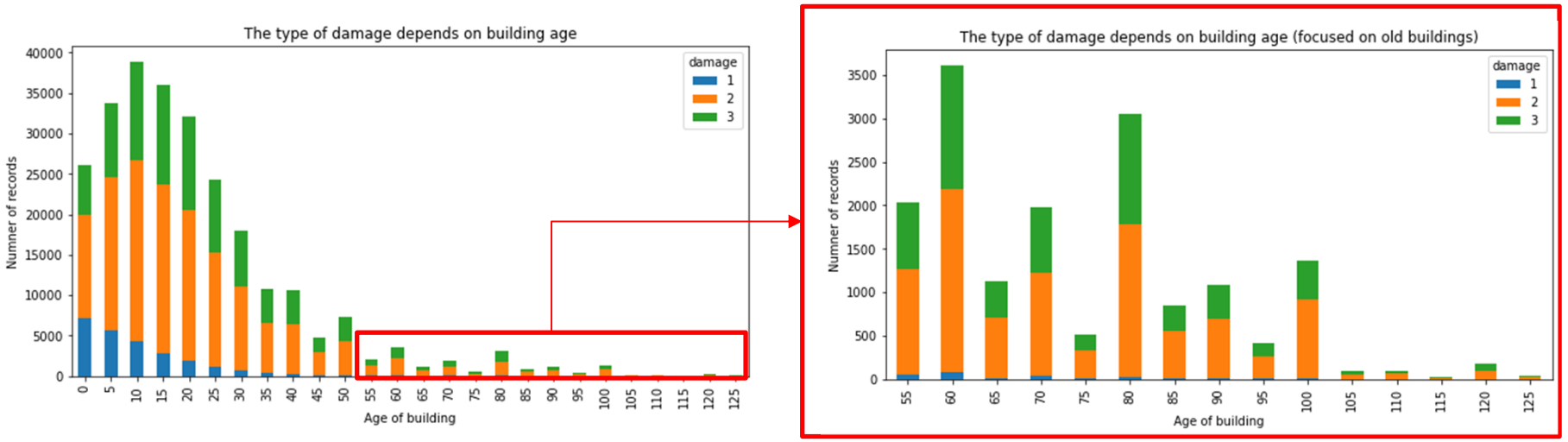
Findings

There are buildings in the dataset with up to 9 floors, but the majority the affect had only 2. Smaller buildings are more likely to have significant damage which could be due to non-engineered residential dwellings.



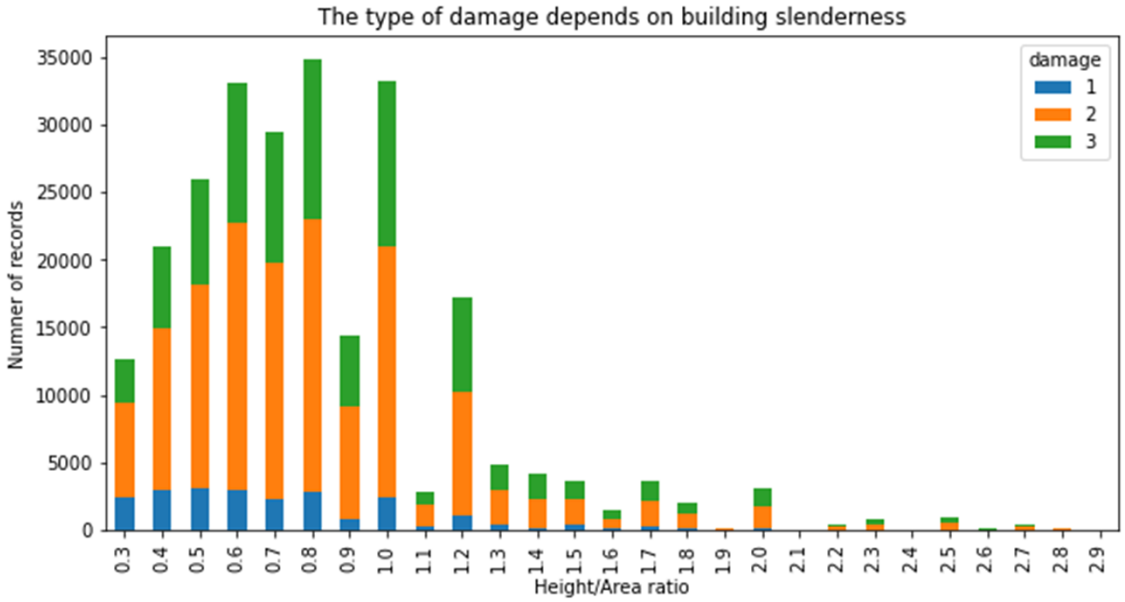
Findings

The number of buildings with low damage decrease with the building age. Less then 10 years old buildings are less likely to have significant damage. It could be due to lower degradation or improved regulations. Closer look at older buildings show that the earthquake event is for them destructive.



Findings

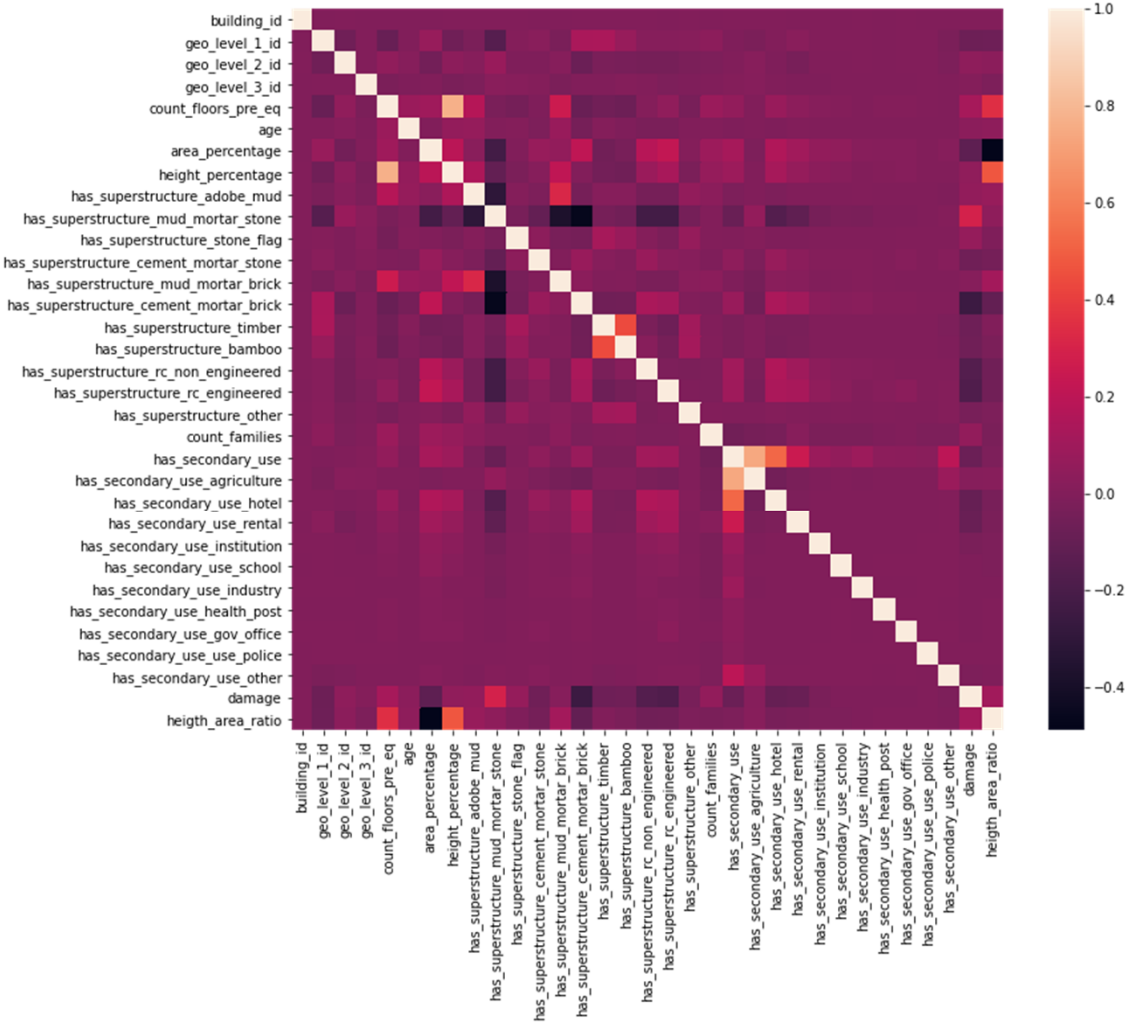
Slender buildings are more likely to have significant damage which is expected due to higher overturning.



Findings

Features Correlation Matrix insight:

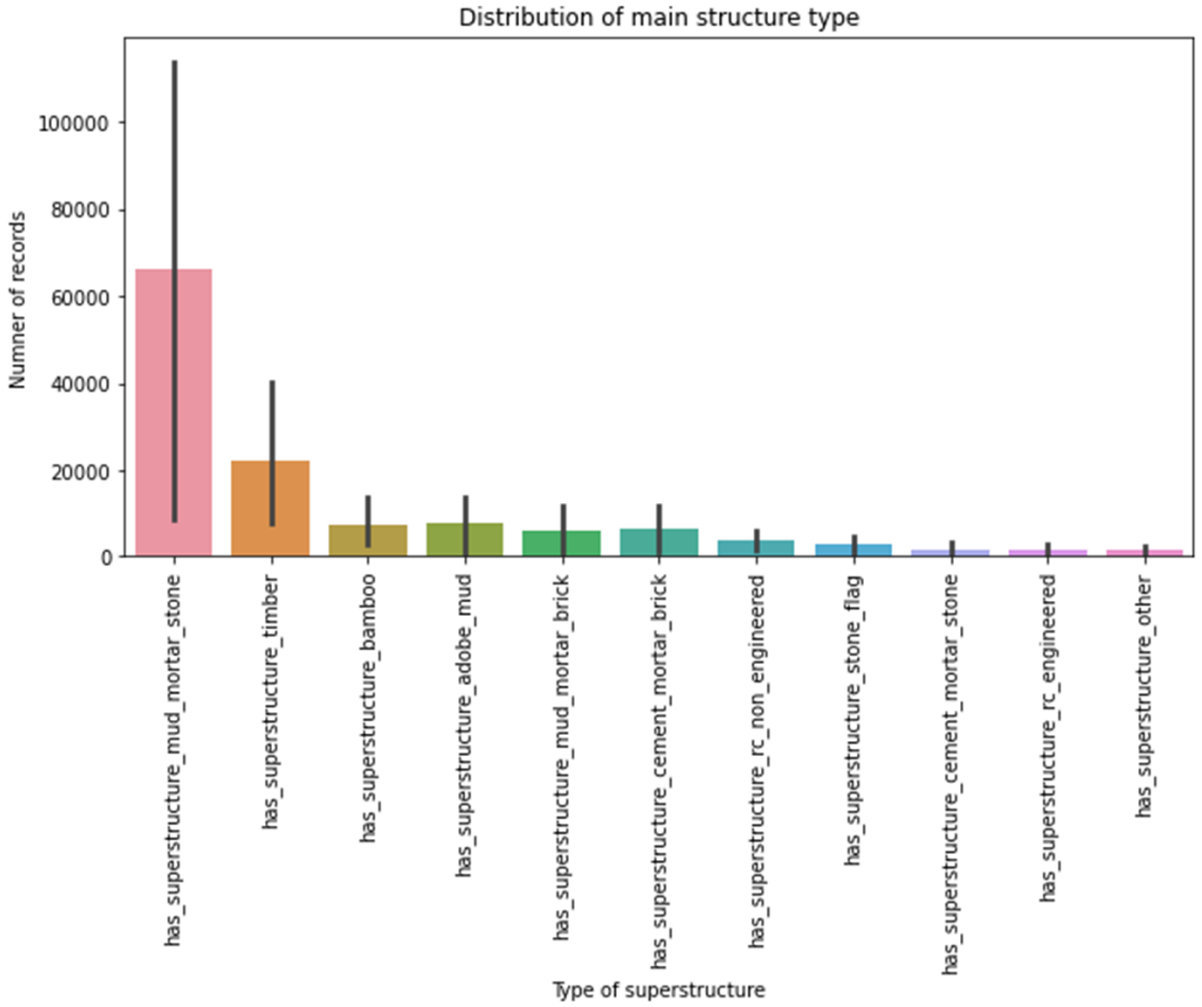
- Not many correlated fields
- Most of damaged buildings has mud mortar stone superstructure
- High correlation between number of floors and building height as they are almost the same parameter



Findings

The materials of the superstructure that got damaged the most were followed:

- mud (stone, brick)
- timber
- bamboo

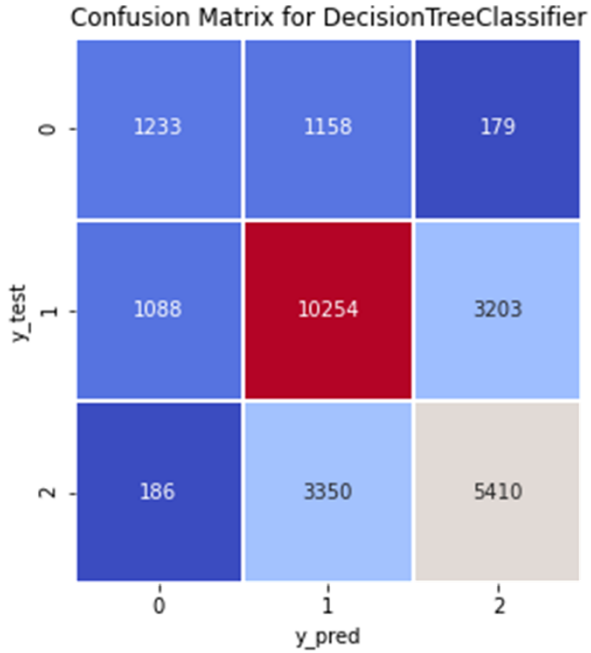


Findings

The accuracy is based on the micro average F1 score and is equal to 0.49 for kNN model and 0.64 for Decision Tree model.

However key metric for poor model performance is the number of buildings that are predicted as damage grade 1 when they actually had damage grade 3, which can be investigated using the confusion matrix.

```
KNeighborsClassifier f1
score: 0.4961
DecisionTreeClassifier f1
score: 0.6388
```



Findings

Location features have the most significant impact on damage grade, which is probably caused by the correlation with seismic hazard.

Also the building age, area and height play an important role, while the superstructure material did not have as much importance as expected.

	importance
geo_level_1_id	0.142596
geo_level_2_id	0.109629
geo_level_3_id	0.102634
age	0.082086
area_percentage	0.081816
height_percentage	0.041943
foundation_type_r	0.039917
count_families	0.018907
has_superstructure_mud_mortar_stone	0.015653
count_floors_pre_eq	0.012627
other_floor_type_q	0.009066
has_superstructure_timber	0.009051
position_s	0.008259
position_t	0.008052
roof_type_n	0.007413

Limitations

The data comes from a questionnaires submitted by building tenants and have not been verified by any engineer, which means that the accuracy is limited by different people subjective opinion and interpretation. It would be beneficial for the future to have a engineering assessment team collecting this data.

The location features has the biggest impact on the damage grade. However in order to calculate the seismic vulnerability for potential even this location data should be removed as we do not know exact earthquake location.

Conclusions

- The analysis has demonstrated that machine learning can be applied to building survey data to correctly identify approximately 65% of the buildings that were most vulnerable.
- It also possible to investigate which building features have the biggest impact on damage grade to improve building quality and design
- Comparing two different machine learning models Decision Tree performed better then kNN on such data set.

Next step could be trying out other models for example XGBoost which allowed to other people to gain the resulted accuracy of 75%.

Acknowledgements

- The dataset come from: https://www.drivendata.org/competitions/57/nepal-earthquake/

- I have not gained the feedback from any friend, so I compared my results with the competition statistics (max f1 score = 0.7558): https://www.drivendata.org/competitions/57/nepal-earthquake/leaderboard/

References

- F1-score evaluation as a better metric for imbalanced classes https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2
- Driven Data - Richter's Predictor: Modeling Earthquake Damage - Benchmark https://www.drivendata.co/blog/richters-predictor-benchmark/
- Global earthquake maps https://www.globalquakemodel.org/gem