Machine Learning Engineer Nanodegree
Capstone Proposal

Asmita Goswami
October 7th, 2018

Predict the damage to a building

Domain Background

Earthquake prediction is concerned with the specification of the time, location, and magnitude of future earthquakes within stated limits,and particularly the determination of parameters for the *next* strong earthquake to occur in a region.

The Objective behind this is to present a methodology for estimating the benefits of adding earthquake resistance to buildings. Determining the degree of damage that is done to buildings post an earthquake can help identify safe and unsafe buildings, thus avoiding death and injuries resulting from aftershocks. Leveraging the power of machine learning is one viable option that can potentially prevent massive loss of lives while simultaneously making rescue efforts easy and efficient.

Problem Statement

This is a classification problem. Inputs will be the building details and the output will be the prediction of the extent of damage that has been done to a building after an earthquake. The damage to a building is categorized in five grades. Each grade depicts the extent of damage done to a building post an earthquake.

Datasets and Inputs

In this dataset consist of the before and after details of nearly one million buildings after an earthquake provided by the hackerearth. This dataset is free to download.

Variable	Description

area_assesed	Indicates the nature of the damage assessment in terms of the areas of the building that were assessed
building_id	A unique ID that identifies every individual building
damage_grade	Damage grade assigned to the building after assessment (Target Variable)
district_id	District where the building is located
has_geotechnical_risk	Indicates if building has geotechnical risks
has_geotechnical_risk_fault_crack	Indicates if building has geotechnical risks related to fault cracking
has_geotechnical_risk_flood	Indicates if building has geotechnical risks related to flood
has_geotechnical_risk_land_settlement	Indicates if building has geotechnical risks related to land settlement
has_geotechnical_risk_landslide	Indicates if building has geotechnical risks related to landslide
has_geotechnical_risk_liquefaction	Indicates if building has geotechnical risks related to liquefaction
has_geotechnical_risk_other	Indicates if building has any other geotechnical risks
has_geotechnical_risk_rock_fall	Indicates if building has geotechnical risks related to rock fall
has_repair_started	Indicates if the repair work had started

vdcmun_id	Municipality where the building is located

Train and Test data are already provided from the Hackerearth. We have randomly classified data into all the five grades.

Solution Statement

The solution will be the prediction of damaged grade for a provided building. First I will do some visualization of the data to get some understanding. Then I will perform feature extraction and select the features such as area_assesed, damage grade, geotechnical risk, geotechnical risk fault crack, geotechnical risk flood, etc.

For training models I will compare logistic regression, decision trees, nearest-neighbors, and SVM since this is a classification problem. Finally I will select the best model for this problem and fine tune parameters to get best accuracy.

Benchmark Model

For this problem, the benchmark model will be Random Forest model. I will try to beat its performance.

Evaluation Metrics

Prediction results will be evaluated based on F1 Score with 'weighted' average.

Project Design

There is no so much for the preprocessing of the data so Before even start training models, I will first take glimpse of the data see what the shape and is and how they are formatted. Then I will start extracting the information such as building_id, district_id, geotechnical risk, geotechnical risk flood, repair started, etc. I may perform some graph visualization for better understanding of the data distribution.

To train models, I plan to choose 3-4 different models to compare. Because this is a classification problem, a few approaches in my head

would be regression, decision trees,KNN, and random forest. Using cross-validation I can find which model performs best, and then use that one to tweak relative parameters.For tuning parameters I will be using GridSearchCV.

After seeing the result, I am going to adjust the parameters and try to get the best result.

References -

https://en.wikipedia.org/wiki/Earthquake_prediction

http://www.iitk.ac.in/nicee/wcee/article/14_S10-032.PDF

Domain background
https://www.hackerearth.com/problem/machine-learning/predict-the-energy-used-612632a9-3f496e7f/description/

Datasets
https://he-s3.s3.amazonaws.com/media/hackathon/machine-learning-challenge-6-1/predict-the-energy-used-612632a9-3f496e7f/a490e594-6-Dataset.zip