Predicting Levels of Earthquake Damage

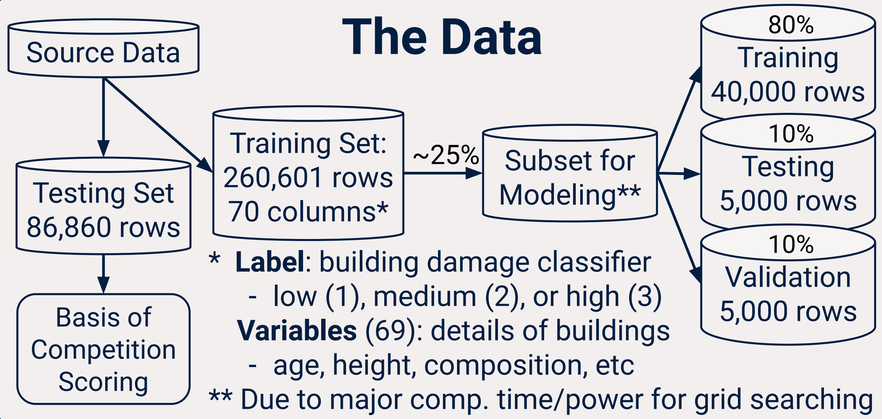
A Comparison of Classification Algorithms

Harrison Lee, Sam Pastoriza, Haley Roberts, Hyuksoo Shin

Introduction

- **Prompt**: "Richter's Predictor: Modeling Earthquake Damage" challenge by DrivenData[1]
- **Goal**: Predict level of damage sustained by buildings in the 2015 Gorkha earthquake in Nepal

The Data



* **Label**: building damage classifier
- low (1), medium (2), or high (3)

Variables (69): details of buildings
- age, height, composition, etc

** Due to major comp. time/power for grid searching

Analysis & Statistical Methods I

1 **Train 9 classification models** on the **subset** training data set using 10-fold cross validation

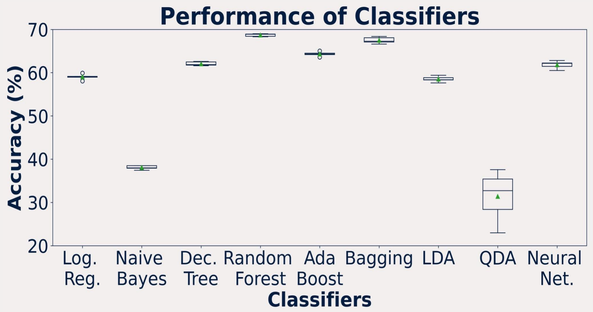


Figure 1: Accuracies of NIne Classification Models

2 **Tune top 6 classifiers** (excluding decision tree) on the **subset** training data set

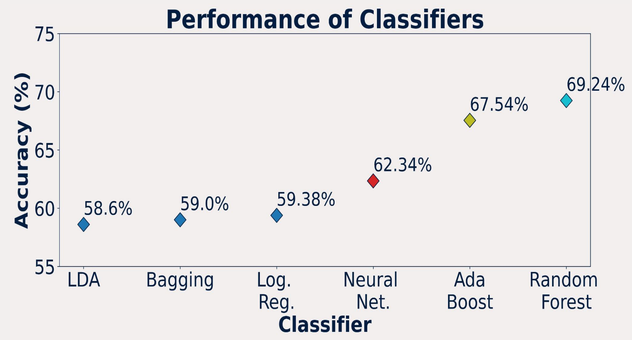


Figure 2: Accuracies of Six Tuned Models

Analysis & Statistical Methods II

2 (tuning top 6 classifiers continued)

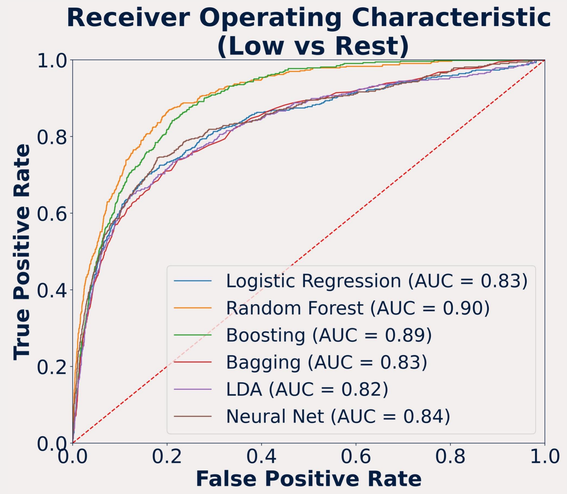


Figure 3: ROC Curve of Six Tuned Models, Damage Level Low vs Med/High

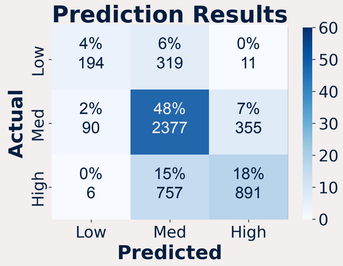


Figure 4: Confusion Matrix of Random Forest Model

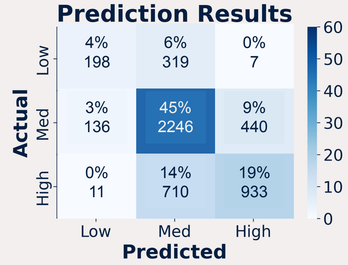


Figure 5: Confusion Matrix of AdaBoost Model

3 **Train** and **tune top 2 classifiers** (adaboost and random forest) on the **entire** training data set
- Models used for prediction of entire testing set

4 **Feature Engineering**

FEATURE	PROCESS
Location	- Engineered separate variables to better connect location to level of damage - Calculated probability of each damage level given location id
Building Materials	- Utilized PCA to focus on building materials accounting for most variability in damage

Results

Variable Importances

- Most important variables of best three models, identified based on mean decrease of impurity

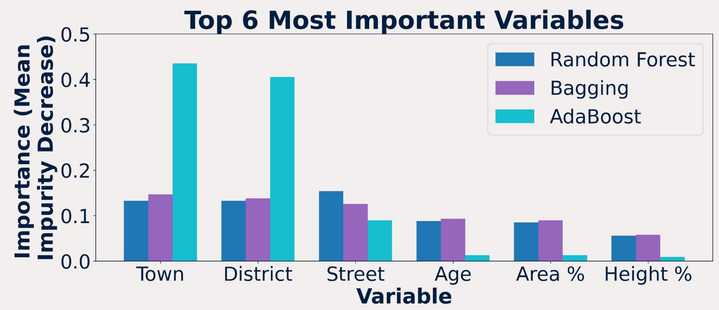


Figure 6: Top Six Important Variables

Final Results & Comparison

- Used adaboost and random forest models (tuned and trained on entire training data set) to predict building damage levels of source testing data set

MODEL	PARAMETERS	ACCURACY
Random Forest	Comp. Baseline	58.15%
AdaBoost	Hypertuned*	69.75%
Random Forest	Hypertuned**	72.68%

** Maximum of 5000 estimators at which boosting is terminated*
*** 2000 trees in forest, 5 samples required to split nodes, sqrt(total features) for consideration of best split, max tree depth of 100*

- Random forest model ranked in **top 15%** compared to models by over 5300 other competitors (top score: 75.58%)

Conclusions

- Best model's accuracy: **72.68%** (tuned random forest)
- Variable importance guides future civil engineering and architectural designs
  - Most important: building location, age, and size
  - Least important: building usage and configuration
- Geographic location relative to earthquake appears to have an impact on building damage
- Buildings constructed using cement, mortar, and brick show the best chance of sustaining minimal damage

Reference: [1] DrivenData. "Richter's Predictor: Modeling Earthquake Damage."DrivenData, https://www.drivendata.org/competitions/57/nepal-earthquake/.