The Principals

Naren Akurati — narenakurati@ucla.edu
Ung Lik Teng — unglikteng@ucla.edu
Alex Tian — tianxinyuan23@gmail.com

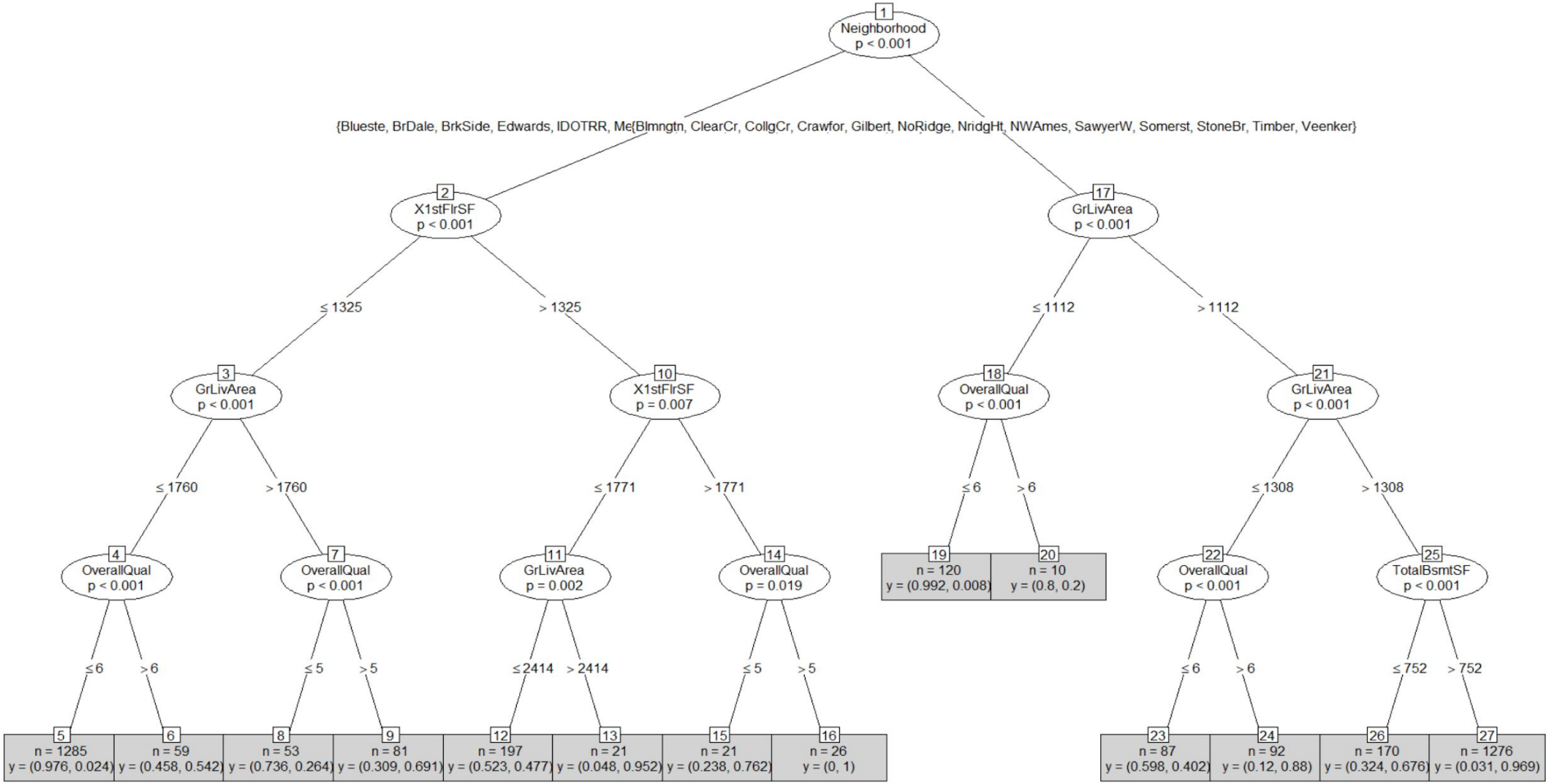
Model

Random Forest

9 features

Public test data accuracy	Private test data accuracy
0.98888	*0.97809*

Variable Name	*Mean Decrease Accuracy*
Neighborhood	39.60349
GrLivArea	38.67549
OverallQual	38.07396
TotalBsmtSF	36.62084
X1stFlrSF	35.08231
LotArea	31.96126
BsmtFinSF1	31.47820
X2ndFlrSF	28.62207
BsmtUnfSF	28.22903


1
Neighborhood
p < 0.001
{Blueste, BrDale, BrkSide, Edwards, IDOTRR, Me{Blmngtn, ClearCr, CollgCr, Crawfor, Gilbert, NoRidge, NridgHt, NWAmes, SawyerW, Somerst, StoneBr, Timber, Veenker}
2
X1stFlrSF
p < 0.001
17
GrLivArea
p < 0.001
≤ 1325
> 1325
≤ 1112
> 1112
3
GrLivArea
p < 0.001
10
X1stFlrSF
p = 0.007
18
OverallQual
p < 0.001
21
GrLivArea
p < 0.001
≤ 1760
> 1760
≤ 1771
> 1771
≤ 6
> 6
≤ 1308
> 1308
4
OverallQual
p < 0.001
7
OverallQual
p < 0.001
11
GrLivArea
p = 0.002
14
OverallQual
p = 0.019
19
n = 120
y = (0.992, 0.008)
20
n = 10
y = (0.8, 0.2)
22
OverallQual
p < 0.001
25
TotalBsmtSF
p < 0.001
≤6
>6
≤ 5
> 5
≤2414
> 2414
≤ 5
> 5
≤ 6
> 6
≤ 752
> 752
5
n = 1285
y = (0.976, 0.024)
6
n = 59
y = (0.458, 0.542)
8
n = 53
y = (0.736, 0.264)
9
n = 81
y = (0.309, 0.691)
12
n = 197
y = (0.523, 0.477)
13
n = 21
y = (0.048, 0.952)
15
n = 21
y = (0.238, 0.762)
16
n = 26
y = (0, 1)
23
n = 87
y = (0.598, 0.402)
24
n = 92
y = (0.12, 0.88)
26
n = 170
y = (0.324, 0.676)
27
n = 1276
y = (0.031, 0.969)

Procedure

- Run full model, achieved 97.65% accuracy, but scaled down due to overfitting
- Data exploration
- Remove variables with large number of NAs:
  - *Ob, Street, Alley, Utilities, RoofStyle, BsmtFinSF2, FireplaceQu, PoolQC, Fence*
- Use full model to find top 9 importance features
- Omit remaining NAs in dataset (very few remain)
- Run model with top 9 features
- mtry=3 | square root of number of features used

Limitations

- Random forest against logistic, knn, and other methods
  - Random forest robust due to correlation between features
  - Non-parametric: low bias and works well with both linear and non-linear data
- Limitations of random forest
  - Tree correlation potential issue
  - High variance without regularization
  - Easily overfit
  - Need to be aware of highly skewed variables

Improvements

- Further random forest parameter tuning
  - mtry
  - ntree
  - strata
  - nodesize, maxnodes
- Omitting outliers in data
- Use of “CARET” over “randomForest” package
  - Provide grid search function to perform hyperparameter tuning
  - More robust options
- Investigate other methods such as xgboost or using a neural network