House sale price predictions

Using tree-based algorithms and recursive feature elimination (RFE)

Julian Cabezas Pena

a1785086

August 26, 2020

Introduction

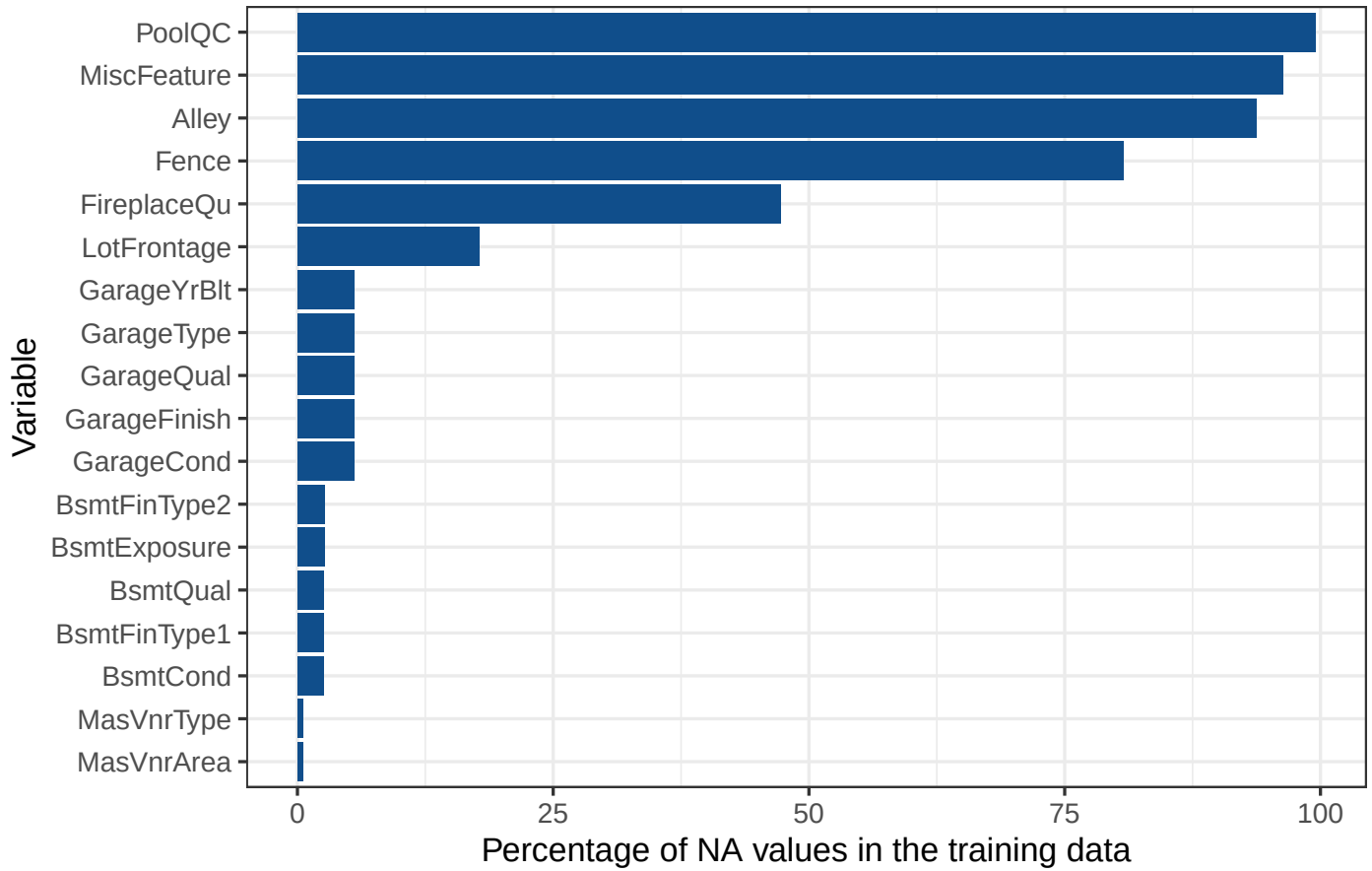
- Regression problem: Continuous target variable
- Bagging and Boosting algorithms: combine various learners (decision trees) to accomplish improved predictions
- Kaggle competition to predict house prices using the Ames, Iowa dataset

Tested methods

- Random Forest: Uncorrelated trees in parallel
- Gradient Boosting: Trees in sequence
- Extreme Gradient Boosting (XGBoost): Popular on competitions
- Categorical Boosting (CatBoost): Developed by Yandex, new (2018)

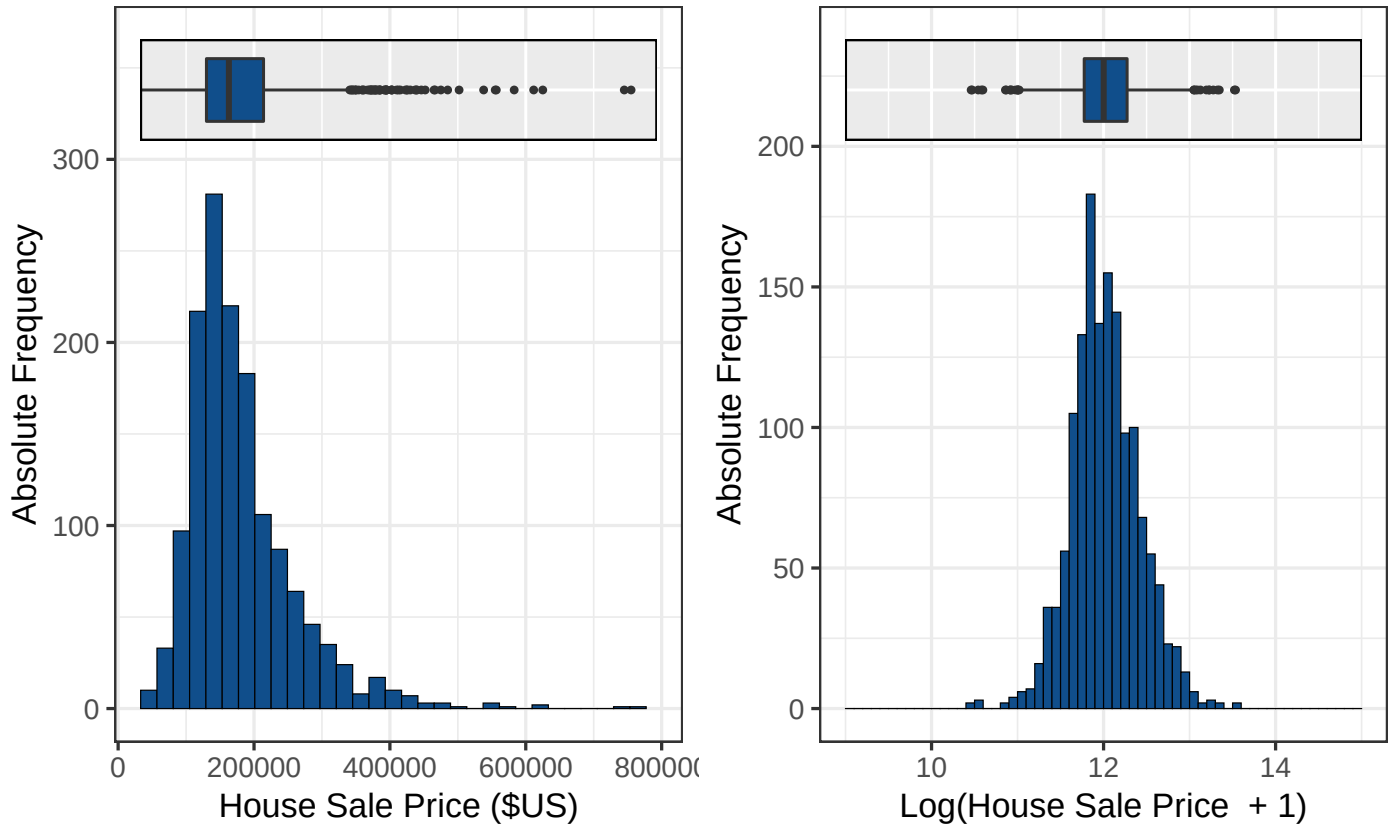
Data preprocessing

- Missing values: Handing according to documentation
- Missing values: According to neighbourhood median or mode



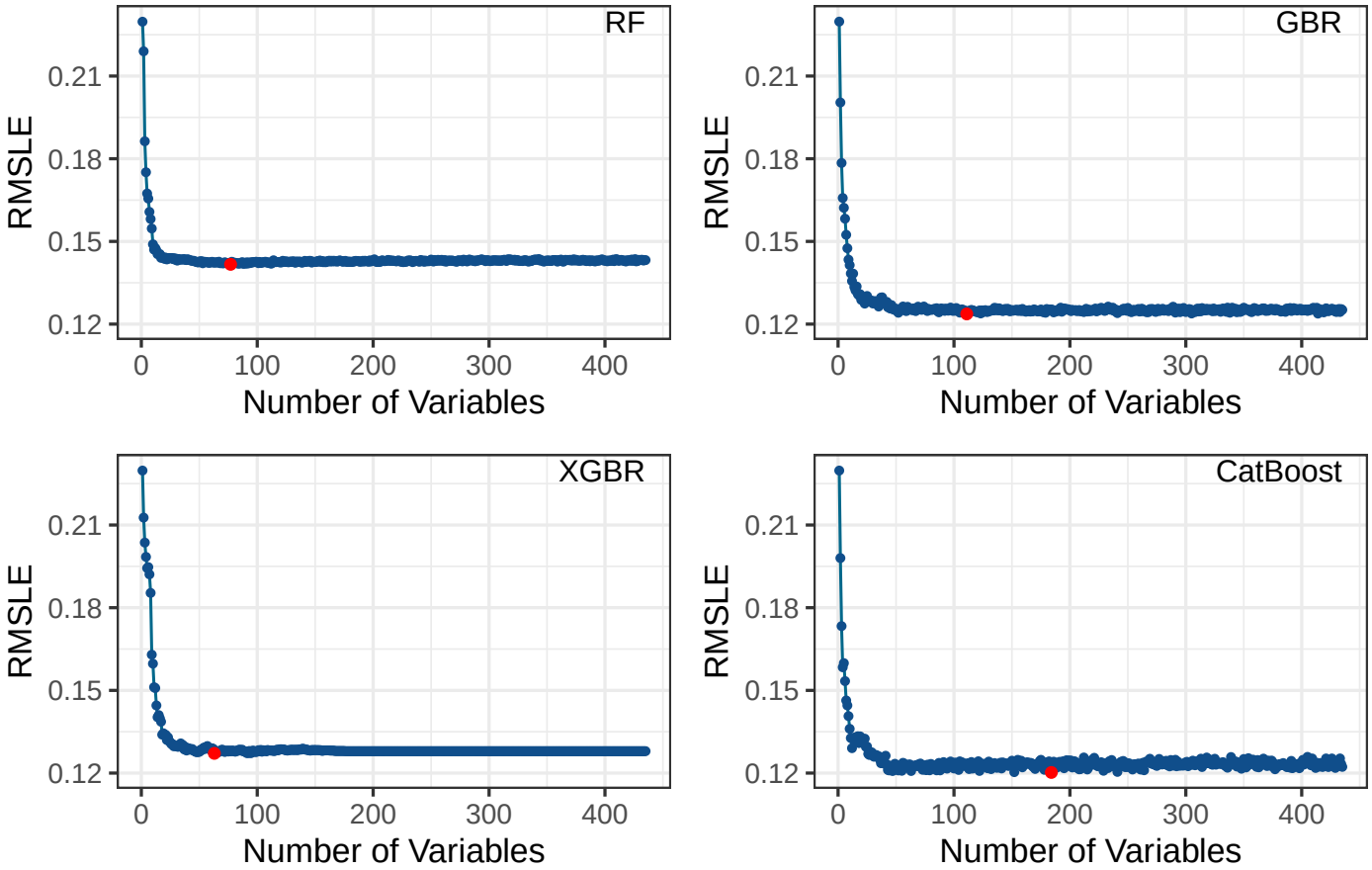
Data preprocessing

- One hot encoding of categorical variables
- Log Transformation of target variable



Recursive Feature Elimination

- Allows to drop redundant or unimportant features (5 fold CV)



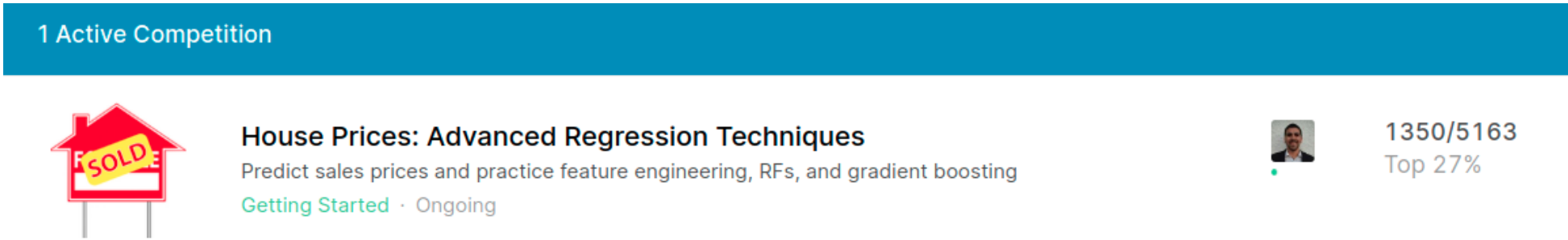
Parameter Tuning

5-fold cross validation, 3 sets of 3 parameters tested

Algorithm	Tuned parameters	RMSLE (CV)
RF	$n_{\text{trees}} = 1500$, $max_{\text{features}} = 22$ & $max_{\text{depth}} = none$	0.1356
GB	$n_{\text{trees}} = 1000$, $learningrate = 0.1$ & $depth = 4$	0.1184
XGBoost	$n_{\text{trees}} = 500$, $learningrate = 0.05$ & $max_{\text{depth}} = 3$	0.1242
CatBoost	$n_{\text{trees}} = 2000$, $learningrate = 0.1$ & $max_{\text{depth}} = 6$	0.1189

Testing

Algorithm	RMSLE (Test set)
Random forest	0.1389
Gradient Boosting	0.1372
Extreme Gradient Boosting	0.1369
Categorical Boosting	0.1269



Kaggle user: JulianCabezas

Conclusion

- The RFE-Gridsearch workflow, although time consuming, provided good results.
- The new CatBoost algorithm looks promising and can be fine-tuned to obtain even better results