Models Comparison

Gradient Boosting Regressor

Advantages: emprically best performance, no categorical data encoding needed( works with categorical data well), deals with missing data well, auto feature selection dealing well with non-informative predictors, no need to centering and scaling data

Disadvantages: less interpretable, more parameters to be tuned, slower to train, not stable if collinearity exists

Lasso Regularized Ordinary Linear Regression

Advantages: higly interpretable, auto feature selection, quicker to train using the XTX beta hat formula

Disadvantages: a linear curve, surface thus limited, can’t deal with missing values( have to remove or permute), not stable if collinearity exists, data needs to be centering and scaling, feature number has be smaller than samples number

Perfomance comparison:

Training/ Fitting vs

Predicting vs

RMSE vs training size

