Report

Sarah Nash

March 2023

1 Hyperparameters and RMSE

1.1 CatBoost

Best parameters:

- iterations: 10000
- depth: 7
- learning_rate: 0.2
- l2_leaf_reg: 3

Expected RMSE: 8.756

1.2 Random Forest

Best parameters:

- n_estimators: 1400
- max_depth : 24
- min_samples_lead : 1
- min_samples_split : 2
- max_features : 'sqrt'

Expected RMSE: 9.690

1.3 Support Vector

Best parameters:

- kernel: rbf
- degree: 12

- epsilon: 0.0
- C: 370

Expected RMSE: 9.397

1.4 AdaBoost

Best parameters:

- n_estimators : 100
- learning_rate: 1.7

Expected RMSE: 10.512

2 Confidence Intervals

For a homoskedastic and normally distributed residual, I believe that RMSE gives the standard deviation of the residuals. Based off this assumption, I'm calculating the 95% confidence interval to be within $2 * RMSE$. (Note: I'm not 100% certain that this is the correct calculation for confidence).

2.1 CatBoost

95% of the predictions should be within ±17.512 of the actual value.

2.2 Random Forest

95% of the predictions should be within ±19.38 of the actual value.

2.3 Support Vector

95% of the predictions should be within ±18.794 of the actual value.

2.4 AdaBoost

95% of the predictions should be within ±21.024 of the actual value.

3 Recommendation model

My recommended model for this regression is CatBoostRegression. The best parameters for CatBoost that I found were

- iterations: 10000
- depth: 7

- learning_rate: 0.2
- l2_leaf_reg: 3.

This combination of parameters resulted in an average RMSE of about 8.756. After running feature importance using this model, the most significant features (in order) are age, cement, water, and slag.