NYC DATA SCIENCE ACADEMY

kaggle

Kaggle Competition

House Prices: Advanced Regression Techniques

David | Neuton | Shubh | Vineet

Feature Engineering

- Irreducible Error:
  - Missing Values
  - Outliers
- Variance:
  - Normalizing
  - Removing Columns
  - Adding Features
- Bias:
  - Transformed features

Examples

Numeric Features:

- Missing Values;
- Removed 2 outliers on *GrLivArea* (Above grade (ground) living area square feet) feature;
- Transformed *GrLivArea* using Square Root;
- Transformed *TotalBsmtSF* using Cubic Root;
- Transformed the Bathrooms features where half bathroom was inputted with .5 and full bathroom with 1;
- Transformed *LotArea* with Log;
- Transformed *LotFrontage* using Square Root;
- Created *ShedSF* based on the *MiscValue* feature;
- Dropped some features with multicollinearity.

Examples

Categorical Features:

- Joined *BasementType* 1 and 2;
- Joined *ExternalMaterial*;
- Created a "Has Fence" feature;
- Joined *Fireplaces* and *Fireplace Quality*;
- Transformed *YearRemodAdd* into range of years;
- Dropped some features with more than 90% of NA or very high frequency of the same category.

Modeling

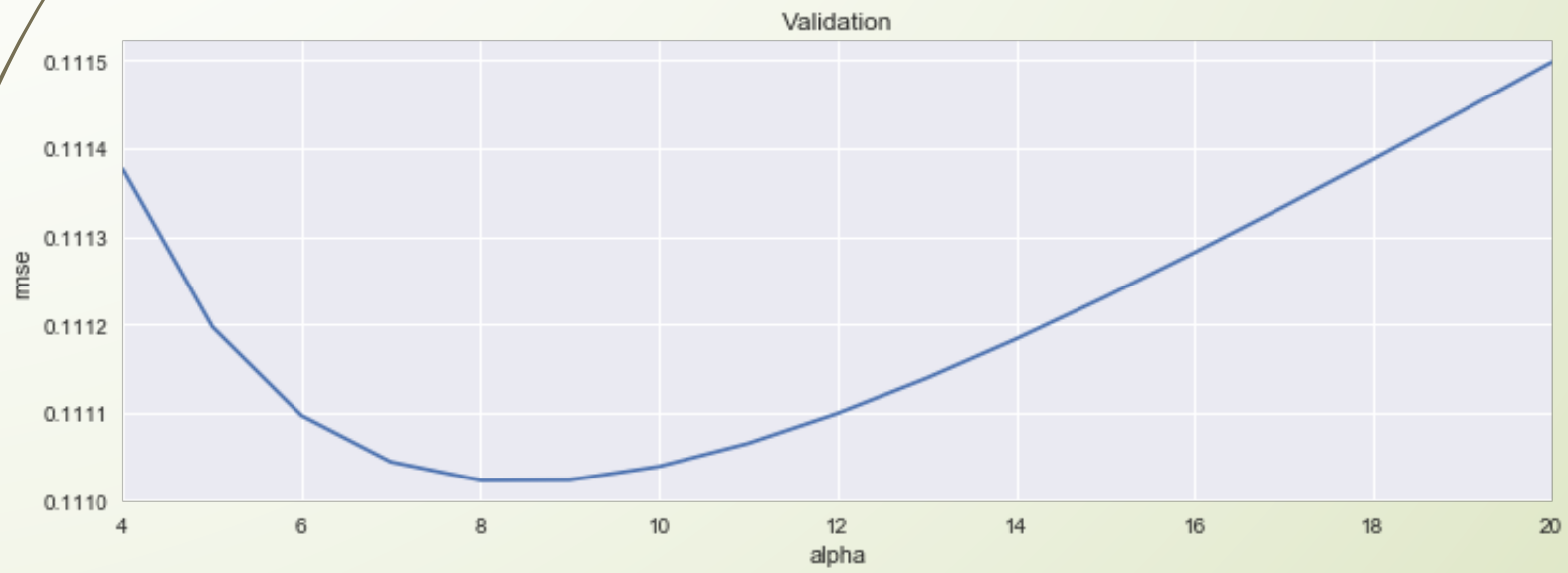
- Support Vector Regression
- Lasso
- Ridge
- Elastic Net
- Random Forest
- Gradient Boosting

Hyperparameters

- Grid Search
- Bayesian Optimization
- Graphing
- Trial/Error

An Example of Parameters Selections

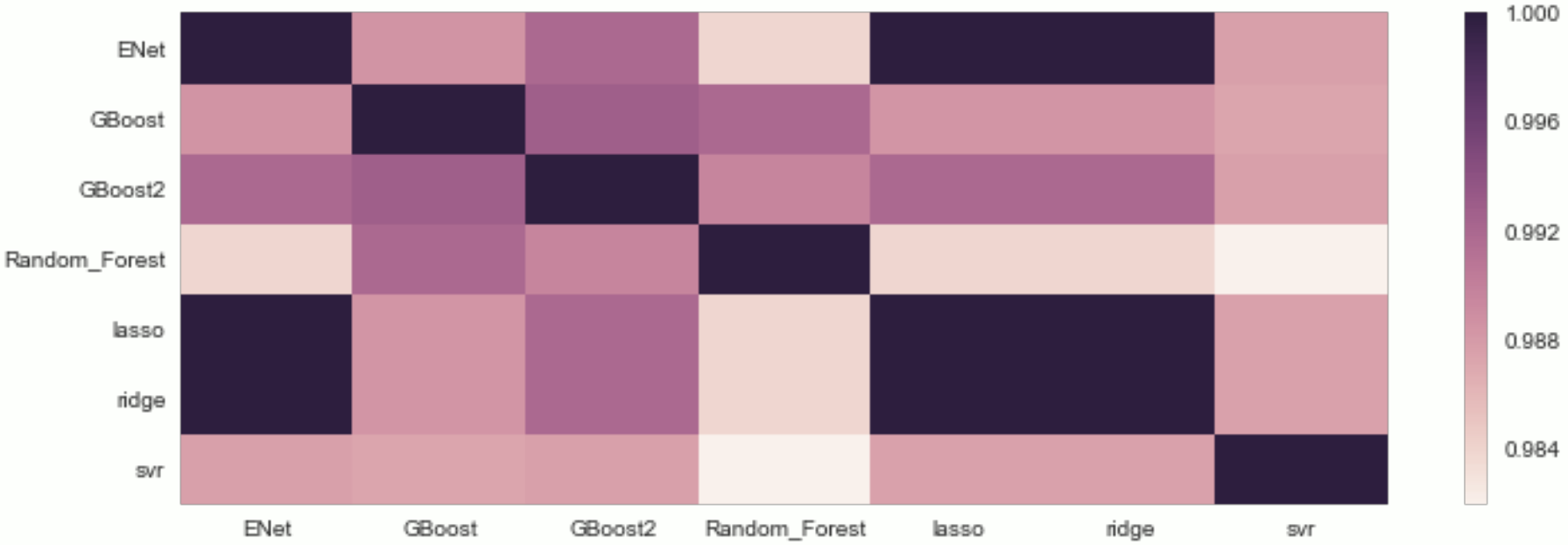
We plotted a range of alpha parameters against the RMSE to find the lowest RMSE for Lasso.



Combining Models

- Average
- Weighted Average
- Stacking

Correlation of the Models



Final Result

kaggle

Kaggle Score: 0.11696