Summary of Linear Regression with Regularization on the Ames Housing Dataset

Background and Motivation

This analysis was performed on the Ames, Iowa housing dataset, made available via the "House Prices - Advanced Regression Techniques" Kaggle competition (Kaggle 2022). The dataset contains records of every home sale in that town from 2006-2010. Successfully building a model to predict home price confers a material information advantage during purchase negotiations.

Approaches to Pre-Processing the Data

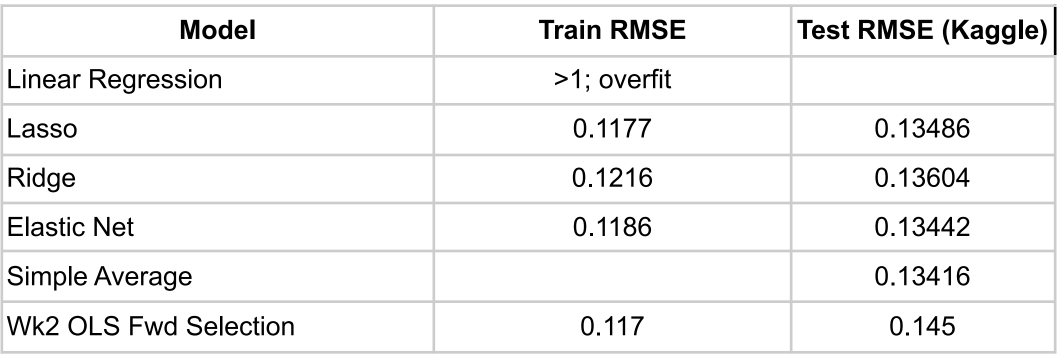
The original dataset provided by Kaggle contained 80 variables: 37 numeric and 43 categorical variables. Data was prepared in six tranches (with each step building on the last):

- Step 0: Basic preprocessing, including filling missing values, excluding outliers, and retaining some categorical variables (respecified as boolean indicators)
- Step 1: Feature engineering of numerics: create new, create zero inflation factor indicators, etc
- Step 2: Implement determination of categoricals: delete, keep, keep and interact with numerics
- Step 3: As part of the model pipeline, apply Yeo-Johnson power transformation where appropriate, and then MinMax scale the independent variables.

Modeling Approach

Once the dataset was processed, models were developed using different regularization techniques: Lasso (l1 regularization), Ridge (l2 regularization), and ElasticNet (mix of l1 and l2). Partial Least Squares - an approach that combines OLS with Principal Component Analysis - was tested but fell short of the other models' performance. A two-stage grid search was pursued for each of the three regularization models: the first stage had large steps in hyperparameter values, with the second stage finding a local optimum. Each model was developed using 5-fold cross-validation, and hyperparameters with the lowest RMSE were retained. The final model was then redeveloped on the full training set.

Results:



Model	Train RMSE	Test RMSE (Kaggle)
Linear Regression	>1; overfit	
Lasso	0.1177	0.13486
Ridge	0.1216	0.13604
Elastic Net	0.1186	0.13442
Simple Average		0.13416
Wk2 OLS Fwd Selection	0.117	0.145

Implications

- Changing up the data processing approaches between Week 2 and Week 3 likely resulted in some degradation of performance, as shown by the Week 2 Forward Selection model having a training RMSE generally lower than all of the Regularization approaches pursued this week. That said, Regularization is a tool to help prevent overfitting, and all of the Regularized models outperformed last week's winner on the test set
- While not shown in the results above, two additional tests were performed: switching the pipeline to exclude power scaling (e.g., skew was maintained, with MinMax scaling), and switching the pipeline to normalize variables rather than MinMax scale variables. These two approaches were universally inferior to the approach pursued (Yeo-Johnson power transform followed by MinMax scaling).
- Consistent with the OLS models researched last week, the top predictors were those that addressed house size (basement and aboveground square footage; lot size) and house quality (Overall Quality, Overall Condition). Surprisingly, kitchen quality emerged as an important feature once regularization was included
- Consistent with last week's results, mixing models yields better performance. The simple average of the Lasso, Ridge, and Elastic Net regressions outperformed the three models in isolation.

References

“House Prices - Advanced Regression Techniques.” Kaggle. Accessed January 9, 2022. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/.