Summary of Linear Regression Applications 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: Creation of new variables based on numerics. Includes discrete and continuous variables
- Step 2: Log-transform numerics, to align with log-transformation of the y variable
- Step 3: Create key interaction effects (e.g., "Overall Quality: High" with "Square Footage")
- Steps 4 and 5: Creating power transforms of continuous variables and then interacting. This was not fruitful, since log(x^2) = 2*log(x), so these generated linear transformations of Step 3's data

Modeling Approach

The dataset was copied after each step, which allowed for six sets of variables of increasing complexity to be used in model development. Per the assignment instructions, only OLS linear regressions were used. In total, seven models were created: one model using all of the variables from each of the six steps above, and then a seventh model which followed a forward-selection logic based on the Step 3 dataset. Models were built using the Kaggle-provided "train" dataset, which was further segmented into an 80/20 "train/validate" set. Results:

Model	Adj. R-sq (train)	RMSE (train)	RMSE (validation)	RMSE (test)
Step 0	90.3%	0.123	0.109	0.159
Step 1	90.6%	0.120	0.114	
Step 2	90.2%	0.122	0.117	
Step 3	91.0%	0.116	0.116	0.16
Step 3 - Forward Selection	90.1%	0.126	0.117	0.145
Step 4	91.0%	0.116	0.116	
Step 5	91.0%	0.116	0.116	
Avg(Step0, Step3)				0.156

Implications

- The top predictors were those that addressed house size (basement and aboveground square footage; lot size) and house quality (Overall Quality, Overall Condition). Interaction effects played a strong role here (e.g., impact of more square footage differed for higher- vs lower-quality housing)
- Of the models tested, more variables did not make for better predictions, demonstrating the risk of overfitting. The most feature-rich model demonstrated the *worst* out-of-sample performance
- Forward selection acts (somewhat) as a proxy for regularization by constraining models to drivers with the most significant p-values. While there are strong arguments against using forward selection, this is a useful first pass at identifying most important variables (Sribney)
- Mixing models can yield improved results. I averaged of the "Step 0" and "Step 3" models (the two with the highest validation RMSE), and the test score out-performed either model independently
- In the forward selection model, many of the indicators (including the intercept) are negative, while the square footage metrics are strongly positive. This likely indicates over-training on square footage metrics, which could be avoided by, e.g., binning those continuous variables to refine the contribution of each incremental square foot
- Classical OLS modeling can produce useful, easily interpretable results. The forward selection model's 0.145 score ranked in the top half of all Kaggle submissions (Rank 2412 out of 5070).
- Analysis of residuals identified the model badly predicted home prices for a certain set of smaller, lower-priced homes. Further research could identify key features to correct this.

References

Sribney, Bill. “What Are Some of the Problems with Stepwise Regressions?.” Stata. StataCorp. Accessed January 16, 2022. https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/.

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