ISE/DSA 5103 Intelligent Data Analytics
Homework #5

Instructor: Charles Nicholson

See course website for due date

Learning objective: Perform predictive modeling using regression techniques.
Submission notes:

1. Team assignment! Include all team member names on the submitted work.

2. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader *may* view your R code, but should never *have* to in order to find your solutions.

3. In the PDF, clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.) Also, note that only *relevant* and informative computer output should be provided.

4. Make sure to *provide comments* on what your R code is doing. Keep it clean and clear!

5. You will submit your complete R script. Note: include `library` commands to load *all* packages that are used in the completion of the assignment. Place these statements at the top of your script.

6. Do not zip your files for submission. Submit exactly two files. Name the files "LastName-HW1" with the appropriate file extension (that is, .pdf for the write-up and .R or .Rmd file for the code)

1 Predicting house prices

The `housingData.csv` file in the course website is real data associated with 1,000 residential homes sold in Ames, Iowa between 2006 and 2010. The data set includes over 70 explanatory variables – many of which are factors with several levels. The file `housingVariables.pdf` provides a concise explanation of the variables and the factor levels in the data. You've examined this data before in class.

For this problem, you are challenged to make the best possible predictions of the final price of each home. You will first build an OLS model to to predict the natural log of the sale price, i.e., `log(SalePrice)`. After analyzing the OLS model, you will also use OLS variants like LASSO, PLS, elasticnet, etc. to build and compare performance.

For any model, you are encouraged to conduct any and all pre-processing that you would like, e.g., missing value imputation, factor level collapsing, outlier processing, etc. Please document these steps.

(a) OLS Model

i. Using all 1,000 observations, build a multiple linear OLS regression model using `lm` to predict the log of the sale price. You may use a stepwise regression technique, you may include interactions, etc. You must try multiple versions and then use some form of resampling (e.g., 5-fold CV) to identify the best model. Report the variables, the coefficient estimates, p-values, AIC, BIC, and VIF for the model. Report the cross-validated (or resampled) R^2 and RMSE for the best model.

ii. Provide a complete analysis of the residuals. Please provide the visualizations that you choose to best depict the residuals as well as any of the metrics we discussed in class that you prefer. Is there anything interesting in the residual pattern? How might this residual pattern influence you to change your model?

(b) Usind all 1,000 observations, create a PLS model to predict the log of the sale price. Use hyperparameter tuning to determine the number of components with RMSE as the error metric (show your chart!). Report the number of components and the CV RMSE estimate for the final model you choose.

(c) Using all 1,000 observations, create a LASSO model to predict the log of the sale price. Use hyperparameter tuning to determine the hyper parameter values with RMSE as the error metric (show your chart!). For the final model of your choosing, report the variables with non-zero coefficients (and the coefficient values) as well as the CV RMSE estimate.

(d) Using any regression technique learned in class with any combination of techniques (PCA, LDA, missing value imputation, etc.) that you prefer to predict the final log sale price. You should attempt at least 3 more model types (e.g., in addition to OLS, PLS, and LASSO, you may consider PCR, ridge regression, elasticNet, robust regression, SVR, and/or MARS).

Provide a summary of the results in a table similar to the following:

Table 1: Summary of Model Performance with 5-fold CV

Model	**Notes**	**Hyperparameters**	**CV RMSE**	**CV R^2**
OLS	`lm`	N/A	3192.4	0.5770
OLS	`lm` + 2-way interactions	N/A	4548.1	0.4910
PLS	`pls`	`ncomp` = 18	2999.6	0.5913
LASSO	`caret` and `elasticnet`	fraction = 0.43	1771.3	0.6831

elasticNet	`caret` and `elasticnet`	fraction = 0.72
		lambda=0.1
			979.9	0.7188

SVR	`caret` and `kernlab` using `svmPoly`	degree = 2
		scale=1.2
		C=0.001
			1370.1	0.6222

⋮	⋮	⋮	⋮	⋮