Homework 6 Part 1

Instructions

Due: 5:00pm on Wednesday, November 3rd

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Regression Competition III

We again return to the data set from Homework 3 to build a new multiple regression model using penalized regression techniques discussed recently in class.

As before, you will construct your model using a training data set with information on 66 variables recorded for 1808 houses. I've held back the data on 600 other houses which will serve as the test data set for assessing the predictive accuracy of your model.

You should record your answers in this .Rmd file. However, you are encouraged to use a separate .Rmd file for scratchwork. The assignment is divided into several **Components** to help organize your work. Put all work you want graded between the bars of red stars in the corresponding section.

Grading

This assignment restricts you to just using `glmnet` to build your model, so your overall score on this assignment will not be based on model accuracy (most models will have relatively similar accuracy). However, you will have optional opportunity at the end of this assignment to build a better model that synthesizes feature selection along with the other work you did on Homework 3, and I will run that model on test data as well and report the results.

The Data

The data set `house_train` can be found in the hw_3 repo and can be loaded by running the following code.

```
house<-read_csv("house_train.csv")
```

Additionally, the `data_description.txt` file in the same repo gives a full description of the variables appearing in the data set.

There is one special column of note:

- `Sale_Price` is your response variable and should not be included as a predictor.

Components

Data Exploration

Create initial training and validation sets from the `house` data.

1. How many rows are in the training set? How many are in the validation set?
2. How many columns are in the training set? How many are in the validation set?
3. Create a model matrix for both training and validation data sets. How many columns are in the model matrix for the training set? How many for the validation set?
4. Explain why your answer to the previous part will pose a problem when using `glmnet` to make predictions on the validation model matrix using a model built on the training model matrix.

Data Partitioning

In this section, create a training / validation split that solves the problem identified in the previous part.

1. Use `model.matrix` to create a model matrix for the full house data set. In order to leave the response variable in the matrix, do not include it on the left side of `~` in the `model.matrix` function.
2. Convert the model matrix to a data frame using `as.data.frame()`, apply `initial_split` function from the `rsample` package, and then create training and validation data frames.
3. Use `model.matrix` once again to create separate model matrices for the training and validation sets, along with response vectors for each set.
4. Verify that both training and validation model matrices have the same number of columns.

Model Building

In this section, use `glmnet` to create both ridge regression and lasso models on the house training data. Do not perform any data processing, or include any transformations or interaction terms (i.e just do feature selection via `regsubsets`)

Model Diagnostics

In this section, use cross validation to assess the relationship between λ and model accuracy, for both ridge regression and LASSO. Display these metrics both graphically, and then explicitly compute optimal values. Which value of λ appears to produce the optimal model for each model type? How many variables are present in the optimal ridge regression model? How many are present in the optimal lasso model?

Model Assessment

Choose 1 Ridge Regression and 1 LASSO model based on the information in the previous component, and then compute rMSE for each model on the *validation* set. Additionally, compare to the rMSE for the **full** model, as well as for the model you choose using feature selection in HW 5.

Which model performed best?

Model Interpretation

What are some advantages of Ridge Regression and LASSO in model building, compared to either the full model or feature selection using `regsubsets`? Are there any disadvantages to using Ridge Regression or LASSO?

Optional Model Enhancement

Optionally, you may use this space to build a model that improves on the one from Homework 3 by incorporating penalized regression, along with transformations, interaction terms, and feature engineering. I will assess your optional model on test data, alongside the model you made in the previous part. Use the following function to create **predictions** for your model. **NOTE that the number 3 should immediately follow your last name**

```
# This function will create a vector of predictions for the Sale_Price of houses.
# This function should take only test_data as an argument.
# The function must be self-contained, so needs to include all model building and data processin steps

FirstName_LastName3_predictions <- function(test_data){
  library(tidyverse)     ## Load whatever packages you need
  ## Create your model here
  my_preds <- predict(model, test_data)   ## Make predictions based on your model.
  my_preds<- my_preds*1 ## Transform your model predictions back to the original units for Sale_Price,
  my_preds      ##return your predictions as output
}
```

To verify that your function is working as desired, open a new .Rmd file, load the `house_train` data, copy the code for your function over to the new .Rmd, and then run the following code:

```
library(dplyr)
B <- sample_n(house, size = 100) # This creates a test dry
FirstName_LastName3_predictions(B) # Change to your First and Last Name
```
