STA 363/663

Project 2: Shrinkage and Selection

The Goal

You have a client who works for a rental company, VRBO. This company allows individuals to put their homes or apartments, or rooms in a home or apartment, up for short term rent. This service is usually used by individuals who are looking for a place to stay for vacation.

Your client gives you a data set with information on n = 1561 rentals and asks you to build a model that can used to predict the price of a rental. This is helpful to the client so they can determine appropriate prices for new rental properties. Setting an appropriate price can make a difference in how successful an individual is in renting out their property. The data set you need can be found on Canvas.

What you will be turning in:

You will be submitting two documents for this project.

- Formal Report:
  - This is the write up that will explain your work.
  - More details on the sections needed for this report are included below.
  - In your formal report, there should be **no code showing.** This includes warnings and other stray code output – hide it all.
  - You will be graded on spelling, grammar, and writing, as well as your stats. Make sure you use spell check.
- Code Appendix:
  - This is a Markdown file or R script with **annotated** code.
  - The goal is that a person who reads your report, and wants to replicate your results, could access your code appendix and completely reproduce the results and figures in your formal report.

The Data:

The data provided contains information on the following variables. Note that the response variable is price.

UnitNumber	The number of the rental listing.
Price	The price in US dollars of a one-night rental.
overall_satisfaction	an average satisfaction score, given as numbers between 2.5 and 5
reviews	the number of reviews about the property on the VRBO website.
room_type	Is the rental for an entire house/apt, a private room, or a shared room?
accommodates	How many people the rental can accommodate
bedrooms	the number of bedrooms included in the rental price
minstay	the minimum number of nights an individual must book at the property in order to stay there
neighborhood	the name of the neighborhood in which the rental is located.
district	the name of the district in which the rental is located. Note: there are multiple neighborhoods per district.
WalkScore	a score indicating how easy and safe it is to get to areas of need/interest by walking
TransitScore	a score indicating how easy and safe it is to get to areas of need/interest by public transit
BikeScore	a score indicating how easy and safe it is to get to areas of need/interest by biking
PctRentals	the percent properties in the neighborhood that are rental properties (rather than used by permanent residents)

Your formal report should contain the following sections, which must be labelled in your final report.

Abstract / Executive Summary

This is the first section in your report, but it is actually the last thing you will write. Called an abstract in academia, and an executive summary in industry, this one paragraph summary of your entire paper is the first thing people will read.

In less than 400 words, you need to describe:

(1) The goal of your project
(2) A brief summary of the methods you used to achieve your goal
(3) A brief summary of your results

For example:

Estimating the number of pandas inhabiting a national panda preserve is critical to understanding the health of the panda species as a whole. In this report, we discuss the process of estimating the number of pandas in the Wolong National Nature Reserve. Data were collected from volunteers walking trails in the park, and we applied capture-recapture estimation process to then estimate the total number of pandas in the park. We detail the steps of the process and compare our results to two other possible methods of estimating pandas. Based on our estimation, there are between 140 and 155 pandas in the surveyed area. We discuss our findings, and limitations of the study, in the following paper.

We will refine the ability to write these as we proceed through the semester. It is an important skill when you are working in industry or writing a paper for academia!

Part 1: Introduction

In this section, you need to describe:

(1) What data are you working with?
(2) What is the client's goal with this data?

When you write your paper, you will be assuming that the **reader has not read the project assignment**, so you need to provide an overview of what you are doing and why. Yes, you already have some of this in the abstract…repeat it here.

Part 2: Data Cleaning

Data Cleaning

In this section, you will describe any necessary data cleaning steps. Do you have to contend with missing data? How much? Are there any other data cleaning steps you need to complete? If so, explain what they are and why they are needed. Note: You should NOT mention any code terms, like na.omit or for loops, in a formal paper. Why? Your reader may not know anything about code!

The client would like you to consider all the features given as possible features for modeling. Part of your job in this section is to identify any features that cannot or should not be used, and to explain to your client why these should not be used. Otherwise, you will be keeping all the features. In other words, you are not being asked to pick and choose which features you *think* are good predictors.

Note: You do not need to explore the relationship between each feature and Y. Your client has mandated that you must use linear regression for this modeling task and is not interested in considering transformations of the Y variable. Some of the relationships are not strongly linear, but there is nothing we can do about that right now (though we will be able to in our next part of the course!). In other words, you should not exclude a feature because its relationship with price is not linear.

Part 3: Ridge Regression

The client first asks you to use ridge regression to predict price. They are interested in a summary of how ridge regression works, and then they would like you to train a model using ridge regression.

If you need a random seed, use 1 as your random seed.

Make sure you clearly explain the process of choosing any needed tuning parameters and state the value of the tuning parameters you select.

Once you have trained your model, show the coefficients in a professionally formatted table. You do not have to write out the trained model, as this will be long, but you do need to show the table! Assess the predictive accuracy of the model built using these features. Explain clearly to the client how you assessed predictive accuracy and describe how this model is performing in prediction.

Part 4: Lasso

The client next asks you to use the lasso to predict price. They are interested in a summary of how lasso works and how it is different from ridge regression. They then ask you to train a model using the lasso.

If you need a random seed, use 1 as your random seed.

Make sure you clearly explain the process of choosing any needed tuning parameters and state the value of the tuning parameters you select.

Once you have trained your model, show the coefficients in a professionally formatted table. You do not have to write out the trained model, as this will be long, but you do need to show the table! Assess the predictive accuracy of the model built using these features. Explain clearly to the client how you assessed predictive accuracy and describe how this model is performing in prediction.

Part 5: Elastic Net

The client finally asks you to use elastic net to predict price. They are interested in a summary of how elastic net differs from ridge and lasso, and then they would like you to train a model using elastic net. When you train this model, they ask you to consider

α = 0, .01,. 02, .03, …, .99,1 as your sequence of possible choices.

If you need a random seed, use 1 as your random seed.

Make sure you clearly explain the process of choosing any needed tuning parameters and state the value of the tuning parameters you select.

Once you have trained your model, show the coefficients in a professionally formatted table. You do not have to write out the trained model, as this will be long, but you do need to show the table! Assess the predictive accuracy of the model built using these features. Explain clearly to the client how you assessed predictive accuracy and describe how this model is performing in prediction.

Part 6: Comparison and Conclusions

We have explored several different ways of using features to predict price. In this section, the client would like you to compare the performance of the three models in terms of the fitted models and predictive performances. Discuss whether any of the models is performing better at prediction than others, and by how much. If the client wants the absolute best value of the predictive metric you have chosen, which model should they select?

You should also discuss any concerns about the model performance in this section. Consider the predictive accuracy of this model with the best value of the predictive metric. Does this indicate strong predictive performance? Describe how you come to this conclusion.

And you are done!!!!