Stat 159 Project 2

Author: Lydia Maher, Ziao Liu

Abstract

In this project, we are doing a typical data analysis cycle which contains getting raw unstructured datasets, data cleaning and processment, exploratory data analysis, modeling and tuning parameters, visualization of results, report and presentation.

We are replicating work on chapter 6: Linear Model Selection and Regularization (from "An Introduction to Statistical Learning" by James et al) with the dataset "Credit". In detail, we are trying to predict "Balance", which is the response variable, from 11 predictors including Income, Limit, Rating, Cards, Age, Education, Gender, Student, Maaried, Ethinicity.

Since it is a continuous prediction, we use five algorithms: Least Square Regression, Ridge Regression, Lasso Regression, Principal Components Regression, and Partial Least Squares Regression. From our output, we found that Lasso Regression performs best, followed by Ridge Regression. Dimension Reduction Regression, in this case, do not perform as well as Shrinkage Methods.