Abstract

Incorporating all the resources learned in Stat 159, we will dive deeper into data analytics by applying more advanced statistical learning methods to model a data set listing credit information of 400 randomly selected individuals. This project will reproduce the methods performed in Chapter 5 and 6 from the textbook, "An Introduction to Statistical Learning" by James et al. These sections cover cross-validation techniques and linear model selection/regularization methods, both of which are necessary in locating the optimal model for a given data set. The credit data set is sourced from the following url: `http://www-bcf.usc.edu/~gareth/ISL/Credit.csv`. Collaborating with a partner was mandatory and allowed us to become more comfortable with Github and Git processes.

Introduction

The motivation for this project is to perform a predictive modeling process to the Credit data set. The first step in almost all data analyses is to understand and search for unique characteristics in the data set we will be working with. Gaining insight into the relationships among the multiple variables is essential in subset selection and prior knowledge into which variables are insignificant. After completing the section covering exploratory data analysis, the Credit data set will be processed to be prepared for model fitting, which includes centering/standardizing each variable column and dealing with categorical variables. The processed Credit data set is then split into training and test sets and thereafter, five different regression models are fitted to select the best model in accurately predicting the response variable Balance.# Data

Data

Data

Data