Methods of Prediticing the Purchase of Mobile Home Policies

Ella Shafi

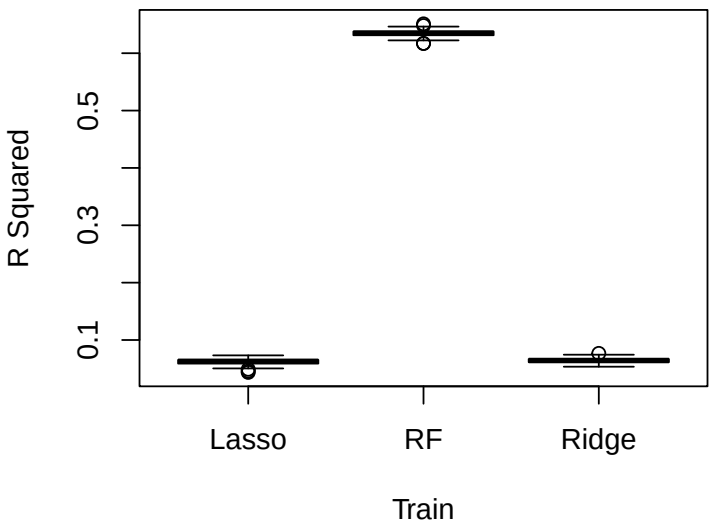
2023-06-28

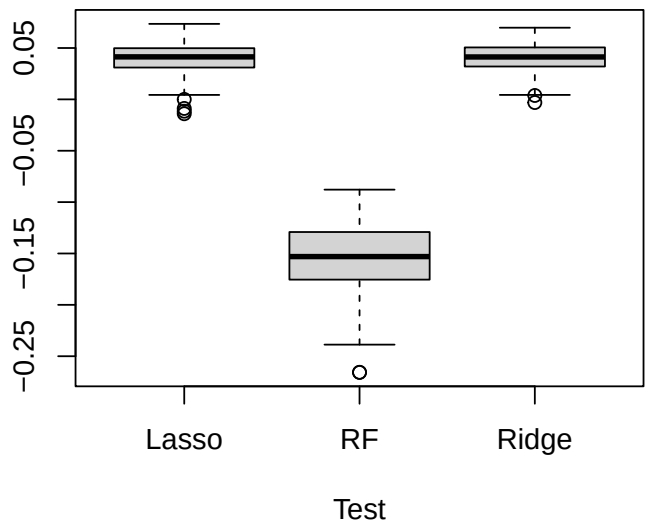
Part 4(b) Data Description

- The Insurance Company Benchmark data set contains information on customers of an insurance company.
- The main goal of this analysis is to model and predict the purchase of mobile home policies by customers.
- The data consists of 86 variables, from which, 5 are categorical and the 81 of them are numerical predictors.
- The response variable is Caravans, the indicator of mobile home policy purchase.
- The predictor variables are socioeconomic variables such as average size of household and product ownership variables such as contribution boat policies.
- From 5588 responses, %80 of the data (4657 observations) was used as training set and 930 variables as test set.

Part 4(b) Boxplots

- For lasso, ridge and random forest regression methods, the R^2 for both train and test samples are calculated.
- As we can see, for both test and train sets the R^2 value of ridge and lasso are quite close while random forest is significantly different from these two methods.
- In the train dataset, random forest method has a better R^2 value comparing to ridge and lasso methods.
- In the test dataset, random forest has the worst R^2.





Part 4 (c) - Time Logs for Model Training

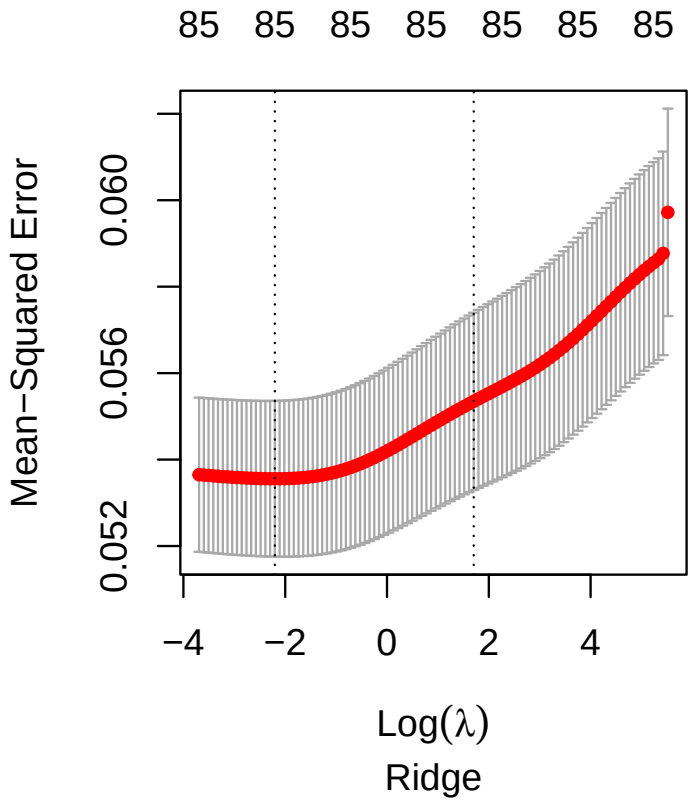
Method	Run Time
Lasso	0.39 secs
Ridge	0.33 secs

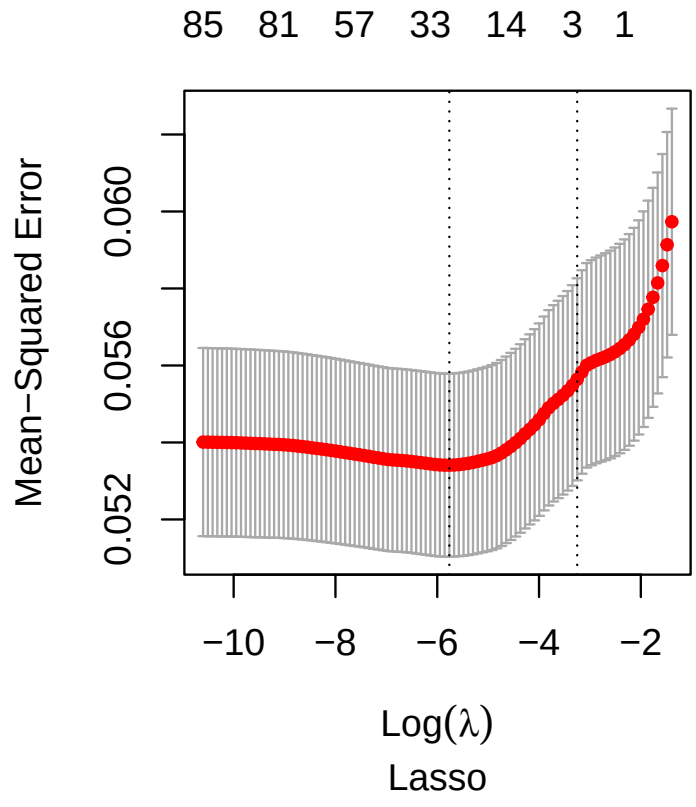
Part 5

- Based on the following outputs, we see mixed results in the trade off between the time it takes to fit the model and predictive performance.
- Specifically for ridge and lasso, the run time for the ridge was longer and yielded higher R-Squared values as well as a narrower interval than lasso.
- The random forest on the other hand produced a negative R-squared interval while also having the longest run time of the 3 procedures.
- This would suggest we would have been better off simply predicting a new sample using only the grand mean instead of the RF model, thus it performed poorly.

Part 5 - CI Plots

- Cross Validation Curves for Lasso and Ridge Regressions





Time Table

- 90% Test R^2 Interval & run time log for ridge, lasso, and random forest models for full model.

Method	CI %90 LB	CI %90 UB	Full Run Time
Ridge	0.0167288	0.0623665	0.50 secs
Lasso	0.0046207	0.0615958	0.40 secs
Random Forest	-0.2230408	-0.1059725	37.58 secs