LASSO Homework

July 21, 2015

1 Predicting Loan Approval

The following data comes from the LendingClub.com. Some data has been modified (read: some variables from the raw data file has been excluded. For full list of modifications, see the Appendix). The goal of this exercise is to find the best model that predicts the actual interest rate for a loan. We will use the train.csv data set to estimate the model, and evaluate the model using the test.csv data. Whenever appropriate, exclude observations with NA. **Report the results in a well-presented table along with the write-up.**

Hint: Not all data is relevant. ID's are columns 1 and 2.

1. Provide relevant summary statistics (mean, sd, quantiles, etc).
2. Establish a baseline OLS results. Provide a brief explanation for the choice. What would happen if the train.csv data only had 15 observations?
   (a) Which variable seems to predict the model well?
   (b) Report the summary of the regression. What is the in-sample R^2? Does the baseline model perform well?
   (c) Show the residual plot.
   (d) What is the in sample $RMSE$ of the chosen model?
   (e) What is the out of sample $RMSE$ of the chosen model?
3. Use an automatic forward model based on AIC, with a full model including second order terms (incl. interaction terms)
   (a) What is the in-sample R^2 of the best fitting model?
   (b) What is the AIC and BIC of the model?
   (c) What is the in sample $RMSE$ of the chosen model?
   (d) What is the out of sample $RMSE$ of the chosen model?
4. Use an automatic backward model based on AIC, with a full model including second order terms (incl. interaction terms)
   (a) What is the in-sample R^2 of the best fitting model?
   (b) What is the AIC and BIC of the model?
   (c) What is the in sample $RMSE$ of the chosen model?
   (d) What is the out of sample $RMSE$ of the chosen model?

5. Estimate a LASSO model with $\lambda = 0.0001$ and a model including second order terms (incl. interaction terms)
   Hint: Use
   ```
   reg_temp=lm(int_rate~.^2,data=data);
   names_var=names(reg_temp$coefficients);
   names_var=names_var[-1];
   model=as.formula(paste("~", paste(names_var, collapse= "+")));
   x=model.matrix(model,data);
   reg4=glmnet(x,reg_temp$residuals+reg_temp$fitted.values);
   ```
   Hint 2: Make sure the test data lines up with the x matrix values properly.

   (a) What is the in-sample R^2 of the model?

   (b) What is the in sample $RMSE$ of the chosen model?

   (c) What is the out of sample $RMSE$ of the chosen model?

6. Estimate a LASSO model using 10-fold cross-validation.
   Hint: cv.glmnet(...) has a 10-fold cross validation as default.

   (a) What is the in-sample R^2 of the best fitting model?

   (b) What is the in sample $RMSE$ of the chosen model?

   (c) What is the out of sample $RMSE$ of the chosen model?

7. What is another way to evaluate these models rather than using $RMSE$?