Homework 3

Team member: Nandini Basu, Linyan Dai, Maxine Li

Using the soccer dataset and the code in the shared Jupyter notebook for basic reference (its purpose is just to help you get started), answer the following:

1. Ideally, how should you choose a sample out of a significantly heterogeneous dataset in order to train a model on it? For this assignment, randomly split the data into test and training sets such that only 10% of the records are in the training set. Fit a simple linear regression model to predict the overall score of a player and test your model against the test set. Calculate the R^2 for the predictions you made on the test set. How many features are used in this model?
    a. Ideally, for a heterogeneous dataset, we should use stratified sampling and then randomly select from all the stratified subsets into the training dataset. Thus the sample could be representative of the whole population
    b. The R-squared of the simple regression is **0.8904**
    c. 122, including dummy variables.
    d. Since the R squared for the base model is already high, we are not performing variables transformations.

2. Using the same training and test sets, fit a simple regression model but with 5-fold cross-validation and predict the overall scores of players in the test set. Calculate R^2 for the predictions and compare with the R^2 from question 1. Please explain your observation.
    a. The R-squared of the simple regression model but with 5-fold cross-validation is **0.8955,** hence we can see that it has increased by 0.005.
    b. In 5 fold cross-validation, the data is split into 5 sets and the model is trained on 4 and tested on 1. This process is repeated 5 times with each one of the 5 sets being used as a test set sequentially. The R squared is calculated for the 5 test data sets and averaged to calculate the overall R squared.
    c. CV is a better way of checking model predictive power than a simple split into test and train

3. Using the training data from question 1, fit a Lasso regression to predict the overall scores of players in the test set. Use the default value of alpha (alpha is used to tune the penalty. Higher the value of alpha, the fewer the number of features) parameter, which is usually 1. How many features are being used by the model? Calculate the R^2 for the predictions you made on the test set and compare with the R^2 from question 1. Please explain your observation.

   a. Number of features: 23
   b. Lasso OOS R-squared: **0.85**
   c. Liner regression OOS R-squared: **0.896**
   d. Lasso regression drops some variables, thus the R-squared decreased.
   e. Even though the R squared has dropped, using lasso ensures that the resulting model has better stability hence we would prefer the lasso model over the model using all variables.
   f. The R-square for lasso regression is smaller than the regular linear regression model. One of the reasons could be the lambda is bigger than the ideal lambda so it cut off more features than needed.

4. Do you expect your answer to question 3 to change if you are using ridge- or log- instead of lasso- penalties? Please explain.
   a. If we use ridge regression, which aims to reduce deviance but doesn't drop variables, we get a model with the same number of variables but a slightly higher R-squared. However, the incremental increase is very small and hence in this instance, this might not be an effective way of improving the model. The same holds true for log regularization. Since the betas are never zero, variable selection isn't performed but this is a possibility that the deviance will reduce.

5. Now try to fit a Lasso regression to predict the overall scores of players with an ideal value for alpha. Your code should try to test different values of alpha and use the ideal one. What, according to your code, is the ideal value of alpha? How many features are being used by the model? Calculate the R^2 for the predictions you made on the test set and compare with the R^2 from question 1. Please explain your observation.
   a. The ideal alpha is 0.01.
   b. Number of features is 59.
   c. R-squared is **0.89**, about 0.005 less than that in Question 1, as Lasso regression removed some variables.
   d. At this value of alpha, there is an optimal trade-off between model stability and deviance reduction.
   e. Even though the R square of Lasso is a little bit less than the linear model, it is a better model because it only uses fewer variables

6. Calculate AIC and BIC for the models you built in question 1 and question 5. According to each of the measures, which is the better model? Is BIC always greater than AIC? Please explain. Compare the AICs with the corresponding corrected AICs.
   a. Lasso regression model has better performance as it has less AIC

b. AIC = Deviance + 2df

   BIC = Deviance + log(n)*df

   Usually, BIC is bigger than AIC but if log(n)<2 then AIC is bigger than BIC

c. Corrected AICs are smaller than AIC in this case.

7. ICs are alternatives to CVs. Do you trust them equally? Please explain.
   a. CVs is the golden rule but could be very computationally expensive to prove. When CV is difficult to implement, we could choose AIC and BIC as alternatives. However, AIC could be a bad approximation in high dimensions, as it could overfit the data. In such cases, we can use the corrected AIC or AICc.
   b. Meanwhile, BIC adds a penalty for adding additional variables and hence can prevent overfitting. It approximates the 1 minus CV but it tends to underfit the data or choose too simple models when n is too big.