DS 740 Midterm

Aditya Nanduri

AIS (Australian Institute of Sports) data

Executive Summary:

Body fat is stored as adipose tissue (the fat visible to the eye) and Intramuscular Triacylglycerols stored in the muscle itself. They have been recognized as an important source of energy during exercise. In trained athletes, this fat is stored adjacent to the mitochondria and with increased training, these numbers increase seen to have an increase too. The data in ais are measurements from high-performance athletes from the Australian Institute of Sport (ais), for 202 athletes (102 males; 100 females) on 13 variables. By predicting Body fat (quantitative variable) based on other measurements, we would be able to predict athlete’s endurance and thereby predict his ability to win.

In this analysis, we try to model/predict Athlete’s body fat using 11 different variables all of which are numeric variables except one variable “Sex” which is a 2-level factor variable. We have chosen Body fat (quantitative variable) as response variable as it would it help in identifying upcoming athletes who have more chances of being successful.

There are various Regression models to predict quantitative response variable. Based on how many mathematical assumptions are satisfied, we need to fit with an appropriate regression model. So, we need to explore data, evaluate and transform/clean data before we even start thinking about the model fitting.

Following are few statistical observations of the dataset:

1. We do not have any missing values in the dataset.
2. When we look at the distribution of each numeric variables, Weight(“Wt”) and Height(“Ht”) variables are slightly skewed with outliers (from boxplot). To resolve this, we apply Log transformation to make them normally distributed. Graphs provided in page 2.
3. When we look at correlation among variables, we see there are strong correlation between few variables like Weight (Wt) – BMI, RCC -Hc, Height (Ht) – Weight (Wt). Graphs provided in page 2.

Now that we have some statistical information about the dataset, lets proceed to for selecting model. I am considering Penalized regression (Ridge, Lasso and Elastic net) and Random Forest models. Since there are strong correlation among variables, Linear regression might not be accurate, hence penalized regression model. Also, for the same reason of collinearity among variable we choose random forest among tree regression models.

We begin our first model selection among penalized regression with cleaned/transformed data. We are using Cross Validation with 10-fold validation. With this step, we get best penalized model along best lambda value. At the end of this step we found that Lasso regression model with lambda of 5 is best model among all the combinations of lambda and Model combinations. This model produced a CV (10) error of 0.9993979 and best lambda value of 5.

In Second model selection we apply 10-fold Cross Validation for Random Forest Regression model. This produced a cv error of 2.106099.

Between Lasso and Random forest model, Lasso penalized regression model is more accurate with less cv error.

As a final step, we will perform double 10-fold Cross-Validation to assess an honest expectation of error rate for these 2 models. We first split the enter dataset into 10 folds.

One is an outer test, and the other 9 folds are sent into an inner modeling selection process where each of these models are also being fit and tested using a CV10 process. The best model from the inner fitting process is selected and used to predict the fold that was held back from the outer split. The predicted values are stored, and once the outer CV10 process is complete, we use actual values to compute the validation of the model (by explaining the percentage of variability). From this step we got cv error of 1.321632 and R squared value of 0.9653336, this means about 96% of the variability in our response values is explained by this modeling process.

Distribution of Wt Var Distribution of Ht Var

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Correlation between Ht and Wt Correlation between Wt and LBM

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

References:

<https://rdrr.io/cran/GLMsData/man/AIS.html>

<https://www.rdocumentation.org/packages/alr4/versions/1.0.5/topics/ais>

<https://www.nsca.com/education/articles/kinetic-select/sport-performance-and-body-composition/>

<https://www.healthline.com/health/exercise-fitness/ideal-body-fat-percentage>