DSA211 Statistical Learning with R

Homework 6

Use R functions and data file (InsuranceFraud.csv) to solve the following problems:

1. An automotive insurance company wants to predict which filed stolen vehicle claims are fraudulent, based on the mean number of claims submitted per year by the policy holder and whether the policy is a new policy, that is, is three years old or less. Data from a random sample of 98 automotive insurance claims, organized and stored in InsuranceFraud.csv, show that 49 are fraudulent and 49 are not.

   (a) Develop a logistic regression model to predict the probability of a fraudulent claim, based on the number of claims submitted per year by the policy holder and whether the policy is new.

   (b) Explain the meaning of the regression coefficients in the model in part (a).

   (c) Develop a logistic regression model that includes only the number of claims submitted per year by the policy holder to predict the probability of a fraudulent claim.

   (d) Develop a logistic regression model that includes only whether the policy is new to predict the probability of a fraudulent claim.

   (e) Develop a logistic regression model to predict the probability of a fraudulent claim, based on the number of claims submitted per year by the policy holder, whether the policy is new and their interaction term.

   (f) At the 0.05 level of significance, are there evidence that each of logistic regression models in parts (a), (c), (d) and (e) of predicting the probability of a fraudulent claim is a good fitting model?

   (g) Compare the models in (a), (c), (d) and (e). Which one model is the best?

   (h) Based on the best model obtained in part (g), predict the probability of a fraudulent claim given that the policy holder has submitted a mean of 1.5 claims per year and the policy is not new.

   (i) Based on the best model obtained in part (g), predict the probability of a fraudulent claim given that the policy holder has submitted a mean of 0.5 claim per year and holds a new policy.

   (j) Based on the best model obtained in part (g), find the confusion matrix with the threshold value being 0.5 for classifying as fraudulent claim and its overall error rate.

-END-