

Homework 6 Programming: Cross-Validation

Due Mar 29

IST 5535 – Spring 2021, Chen

Programming Task (10 points)

The data file "UniversalBank.csv" contains a dataset of 5000 customers of the Universal Bank. Below is the description of columns in the dataset.

- ID: Customer ID
- Age: Customer's age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer (\$000)
- ZIPCode: Home Address ZIP code.
- Family: Family size of the customer
- CCAvg: Avg. spending on credit cards per month (\$000)
- Education: Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
- Mortgage: Value of house mortgage if any. (\$000)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities_Account: Does the customer have a securities account with the bank?
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Does the customer use internet banking facilities?
- CreditCard: Does the customer use a credit card issued by Universal Bank?

Complete the following tasks:

1. Import the data, show the structure of the dataset. Explain what variables in the dataset that are currently represented as quantitative variables but should be measured by nominal scale. Do not include those dummy variables.
2. Convert the data type of Education variable to dummies. Also remove the customer ID and zip code from the dataset.
3. Use logistic regression, LDA, and QDA to predict whether a customer accepts personal loan offered in the last campaign by Universal Bank. Use other variables remained in the dataset as predictors. Fit the machine learning models to the whole dataset (do not manually split the dataset into training and test sets) by using a 5-fold cross validation to calculate the performance measures including balanced accuracy, sensitivity, and specificity.
4. Compare logistic regression, LDA, and QDA. Which model performs best for predicting personal loan acceptance by the customers in terms of sensitivity? Which model performs best for predicting personal loan acceptance by the customers in terms of balanced accuracy?

Submission: Use R markdown to finish the tasks and directly type your answers or explanations on the R markdown. Submit both the R Markdown and HTML report to Canvas. The HTML report should be neatly organized with a good structure and format.