Case Study 4: Classifying Churn

CLASSIFICATION OF ASSIGNMENT

This is a Category B Assignment - Group Assignment:

Your group may not receive help from anyone outside your group. All questions concerning this assignment should be addressed to your professor. It is an honor code offense to give help to other groups and individuals or receive assistance from other groups and individuals.

DATA DESCRIPTION

This dataset of a U.S. bank's customer that contains for getting the information about if this customer will leave bank (Churn).

Input variables:

Row Number
Customer ID
Surname – Last name
Credit Score
Geography
Gender
Age
Tenure – How long a customer has been with the bank
Balance – Average balance of customer
NumOfProducts – Number of bank product customer is using
HasCrCard – Whether customer has a credit card Churned - 1, Yes and 0, No
IsActiveMember

Target Variable
Exited – "Churned" = 1, Yes and 0, No

INSTRUCTIONS

THE MAIN GOAL OF THE ASSIGNMENT IS TO CLASSIFY OBSERVATIONS BASED AS EXITED (churned) (1 = Yes, 0=No)

Data obtained from Kaggle: https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction?resource=download&select=Churn_Modelling.csv

In an R script file, load the Case4.csv dataset and using the appropriate functions to inspect the data frame. Complete the parts listed below. ***CLEARLY LABEL EACH SECTION***

Part 1: Interpreting Logistic Regression Results

Using all the data, run a logistic regression looking at the effect ***age*** and ***gender*** have on classifying exited. ***Interpret the results of the coefficients in comments and provide some insight into what the results mean.***

Part 2: Comparing Methods

- Divide the data into training and testing group.
- Create a logistic regression model using all the applicable predictor variables
- Center and scale your data and create a LDA model using centered and scaled training/test data.

- Using that same centered and scaled data, create a QDA model.
- Create a knn model that also includes centered and scaled data.
- Provide the accuracy rates of the validation set for each test conducted above.
- For each model listed above, provide a summary in comments interpreting the confusion matrix/table object regarding True Positives, True Negatives, False Positives, and False Negatives, specificity and sensitivity. If a model above would not run, give a statement as to why it would not run and what that means for your analysis.
- Finally, in comments, select the best model(s) and describe what that means in terms of the shape of the data (linear to non-parametric)
- One required R Script file per group should be submitted via blackboard by the due date listed in the system.

GROUP CONTRIBUTION AND ASSESSMENT

- Everyone in the group must contribute and write a portion of the code
- Groups will be peer evaluated through teammates.
- Participation and score through teammates will be incorporated into your final average.

Failure to complete a group evaluation on TEAMMATES will result in a 10%-point reduction

RUBRIC

There are 5 models to assess individually:

1. Logistic Regression only 2 predictor variables (age and gender) – assign name as shortlogmodel in your code
2. Logistic Regression with all the predictor variables – assign name as logisticmodel in your code
3. A LDA model with all the predictor variables – assign name as ldamodel in your code
4. A QDA model with all the predictor variables – assign name as qdamodel in your code
5. A KNN model with all the predictor variables – assign name as knnmodel in your code

Each model will be graded for correctness of code, alongside correctness of interpretation and accuracy rate provided.