Data description

Bank Accounts data. The dataset contains information about bank customers. The goal is to predict whether a customer will close his credit card account, in that case the bank may try to propose more convenient services and prevent the client from leaving the bank.

`CLIENTNUM`: client number. Unique identifier for the customer holding the account

`Customer_Age`: customer's age in years

`Gender`: M=Male, F=Female

`Dependent_count`: number of dependents

`Education_Level`: educational qualification of the account holder

`Marital_Status`: Married, Single, Divorced, Unknown

`Card_Category`: type of Card (Blue, Silver, Gold, Platinum)

`Months_on_book`: period of relationship with bank

`Total_Relationship_Count`: total no. of products held by the customer

`Months_Inactive_12_mon`: no. of months inactive in the last 12 months

`Contacts_Count_12_mon`: no. of Contacts in the last 12 months

`Credit_Limit`: credit Limit on the Credit Card

`Total_Revolving_Bal`: Total Revolving Balance on the credit card

`Avg_Open_To_Buy`: Open to Buy Credit Line (Average of last 12 months)

`Total_Amt_Chng_Q4_Q1`: change in transaction amount (Q4 over Q1)

`Total_Trans_Amt`: total transaction amount (Last 12 months)

`Total_Trans_Ct`: total transaction count (last 12 months)

`Total_Ct_Chng_Q4_Q1`: Change in Transaction Count (Q4 over Q1)

`Avg_Utilization_Ratio`: average Card Utilization Ratio

`Income`: annual income of the account holder (in thousands of dollars)

`Closed_Account`: binary variable, 1 if the customer closed the account, 0 otherwise

Purchases data. The dataset contains information about annual customers' purchases from a Portuguese wholesale retailer. It contains the following variables:

`Channel`: channel from which a customer purchases wholesale retailer's products (Retail, Food Service)

`Region`: region where the customer comes from (Lisbon, Oporto, Other)

`Fresh`: Annual expenditure (euro) for the purchase of fresh products

`Milk`: Annual expenditure (euro) for the purchase of milk

`Grocery`: Annual expenditure (euro) for the purchase of grocery products

`Frozen`: Annual expenditure (euro) for the purchase of frozen products

`Detergents_Paper`: Annual expenditure (euro) for the purchase of detergents or paper

`Delicatessen`: Annual expenditure (euro) for the purchase of delicatessen products

By looking only at the quantitative variables, the goal is to find homogeneous groups of customers who have similar purchasing behaviours.

Questions

Part I: Classification

Data cleaning, Exploratory analysis

1. Import the data set `bank_accounts_train.csv`. The categorical features have some missing values denoted by "Unknown": decide how to deal with them (notice that the numerical predictors do not have missing values). You may want to re-code missing values using `R`'s special value `NA`. Also, it may be useful to encode categorical variables as `factors`.

2. Describe the data, measure and visualize the most relevant relations.

Logistic Regression, k-NN.

3. Fit a *Logistic Regression Model* to estimate the effect of `Income` and `Gender` on the probability of closing an account and plot the fitted logistic curves. Does `Income` have a different effect for males and females? Interpret the coefficients and comment the results.

4. Consider only the continuous predictors `Total_Trans_Amt` and `Total_Trans_Ct`. Choose the best number of neighbours k to use in k-NN by evaluating the predictive performance of the model on the validation set. Plot the scores obtained as a function of k.

Do your best!

5. Your goal is to find the best model for predicting the individual probabilities of closing an account, with respect to the AUC metric. You may use any strategy that you think is useful, including but not limited to:

   - AIC and BIC based stepwise procedures for Logistic Regression
   - *Feature Engineering*: build new predictors as a function of the existing ones (e.g. ratios)
   - *Dimensionality Reduction*: build new features as the "scores" resulting from the first components in PCA (you can choose how many components to include)

   Report and compare the results for some of the best models you can find (use again a validation set for model selection). Load the data set `bank_accounts_test.csv`. It contains the features observed in some new test data, but not the outcome. Use your best model for computing a vector of predicted probabilities on the test data.

6. Suppose you have the following information:

   - the bank's loss for every customer who closes the account is 50
   - the bank's gain for each customer who doesn't close the account is 50
   - the bank is willing to offer highly competitive interest rates to the customers who are predicted to leave. *Assume, for simplicity, that all the customers who receive the offer will accept it and won't leave.* In this case the bank's gain is reduced to 20 for each customer.

   Discuss the issue of the choice of the threshold and provide some relevant metrics.

Part II: Clustering

7. Load the data set `purchases.csv`. Using only the numerical features (i.e. exclude `Channel` and `Region`) group customers into clusters using k-means and hierarchical clustering methods, in the latter case possibly considering different dissimilarity measures (e.g. *euclidean distance*, *1 - row correlation*).

8. Verify if units in the same business sector tend to have similar spending profiles (i.e. tend to belong to the same cluster) by computing the confusion matrix between cluster labels and `Channel` labels, and possibly some appropriate association measures.

Submissions

1. Write a report/notebook and include the `R` code you used. Try to extensively comment your code to make it more readable. You may submit the report in any format you prefer, anyway the recommended option is to use a RMarkdown file in order to automatically include figures and code in the report. In this case directly submit the .Rmd file, and not the compiled version.

2. Write a `csv` file containing the predicted probabilities computed at Question 6. You can use a command of the form `write.csv(my_prob, "my_prob.csv", row.names = F)`.

The report and the predictions file must be submitted by the group leader using the dedicated form on the LUISS learn course page.

Grading

The final grade will be based on the methodological correctness and the arguments you give in the answers. Higher grades will be granted for insightful comments, smart coding and clever data visualization. Extra points for those who achieve the *best performance on the test data*!