Data Analytics: Assignment 2

Gertraud Malsiner-Walli and Lucas Kook

Classification: Churn prediction

Goal

The managers of a bank are disturbed as more and more customers are leaving their credit card service. The managers would really appreciate if a data scientist could predict who will churn. Then, the bank can proactively go to this customer and provide him/her better services and turn in this way the customers' decision in the opposite direction.

Dataset

The goal of the analysis is to predict customer churn. The dataset consists of 10,000 customers including their features such as churn, age, salary, credit card limit, credit card category etc. The attributes are:

- CLIENTNUM: Client number
- Attrition_Flag: churn variable, '1' if the customer account is closed, else '0'.
- Customer_Age: Customer's Age in Years
- Gender: M=Male, F=Female
- Dependent_count: Number of dependents
- Education_Level: Educational Qualification
- Marital_Status: maritial status
- Income_Category: Annual Income Category
- Card_Category: Type of Credit Card
- 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
- 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
- Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education
  synthetic variable
- Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education
  synthetic variable

Tasks

1. Read the data into R and check whether the data are read correctly. Provide an exploratory analysis of the data (descriptive statistics, graphics). Especially:

- Eliminate the variables which are not helpful for the analysis, e.g. Client number, synthetic variables.
- Structure of data: are the variables numeric or categorical? Change categorical variables into factors. Check whether the ordering of the levels is intuitive, otherwise change the ordering of the levels (note that the first level is used as baseline in logistic regression).
- Generate the summary statistics for all variables in the data set and comment on some of them. Which percentage of customers did leave the bank?
- Generate appropriate graphs for investigating the relationship between the features and churn. Which variables seem to be related to churn?

2. Split the data into a training set (80%) and a test set (20%).
3. On the training sample, estimate the logistic regression model with all variables for the purpose of churn prediction. Which variables have a significant impact on churn?
4. On the training sample, perform stepwise model selection using AIC. Which variables remain in the final model? Interpret the effects of the final model:

- How much do the odds for churn change if a customer is contacted once more?
- How much do the odds for churn change if a customer has a Platinum card instead of a blue credit card?
- How much do the odds for churn change if the credit limit of the custumer is increased by 1000 Dollars?

5. On the training set, estimate the k-NN model with different k. In order to pick the appropriate k do a 5-fold cross-validation exercise on the training data (in R you can use the caret package to automatically select the best value for k or you can manually repeat the exercise for $k = 3, 5, 7, 9, 11, 13, 15$).
6. On the training sample, estimate the Naive Bayes model.

7. Compare all four methods (full logistic regression model, stepwise logistic regression model, k-NN approach, Naive Bayes approach) in regard to churn prediction, based on accuracy, recall and precision on the test sample (in R you can use the `predict` function available for the respective objects in R or the caret package). Which evaluation measure is the most appropriate one for the business problem? Is there a method which outperforms the others?

Submission (only one submission per team!):

The assignment should be completed in as a Notebook containing code and text explanations.

If you submit an R Notebook, submit the resulting .html file together with the .Rmd.