NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PREDICT 422: Practical Machine Learning — Charity Project: Part 3

Problem Description

A charitable organization wishes to develop a machine learning model to improve the cost-effectiveness of their direct marketing campaigns to previous donors. According to their recent mailing records, the typical overall response rate is 5.1%. Out of those who respond (donate) to the mailing, the average donation is $15.62. Each mailing costs $0.68 to produce and send. The mailing includes a gift of personalized address labels and an assortment of cards and envelopes. It is not cost-effective to mail everyone because the expected profit from each mailing is $15.62 x 0.051 – $0.68 = $0.12.

We will address this problem over a series of assignments. The overall goal for this problem is to maximize the net profit of the next direct marketing campaign. Our approach will be two-fold:

1. We would like to build a **regression** model to predict expected gift amounts from donors.
2. We would like to develop a **classification** model that can effectively capture likely donors.

The overall problem will be broken down into <u>four</u> separate assignments.

1. The Regression Problem
2. The Classification Problem
3. **The Mailing List Problem**

Charity Problem — Part 2

Data Files

- `projectDataPart3.csv`
- `projectDataTEST.csv`

Sample Code

- `SampleCodePart3.R`
- `SaveYourModels.R`
- `DataPreparation.R`
- `RankedDonorOutput.R`

Exercises

1. Read Data from CSV File

   Read the data into R from the CSV files `projectDataPart3.csv`. As part of this step, consider the following factors:

   a. All missing values will be encoded as ‘NA’ in the CSV file. Therefore, the default setting for the argument `na.strings` of the `read.csv` function is sufficient to correctly encode missing values in R.
   b. It is recommend that you use the default setting of `stringsAsFactors = TRUE` for this dataset. This recommendation is made for the reason that the three fields containing strings (GENDER, RFA_97, and RFA_96) are truly categorical variables. A different decision might be made if there were a field such as Name that contained strings that did not belong to categories.

c. Using information in the data dictionary, identify which variables are categorical in nature. Convert these variables to factor variables. If you used `stringsAsFactors = TRUE` in the previous step, then GENDER, RFA_97, and RFA_96 will already be factor variables. Note that several categorical variables are represented by integer categories in the CSV file. Those variables will need to be converted to factor variables manually.

2. Predictions on Validation Set

In this exercise, you will apply your chosen model from Part 1 of the project to predict DAMT on the data in `projectDataPart3.csv` and apply your chosen model from Part 2 of the project to predict DONR and PDONR on the data in `projectDataPart3.csv`.

Here we define PDONR to be the probability or score (a number between 0 and 1) produced by the classification model[1]. For example, with a logistic regression model, PDONR is the value that comes directly from the regression equation (e.g. 0.61) and DONR is the value (0 or 1) that you obtain after comparing PDONR to the threshold value. While PDONR is a new concept, it is the value that will be of more use to us in this part of the project.

a. Review the data preparation steps you took in Part 1 of the project. Apply those same data preparation steps to the data in `projectDataPart3.csv`.
b. Using the model you chose from Part 1 (as trained on the training data from Part 1), predict DAMT on the data coming from Step 2a.
c. Review the data preparation steps you took in Part 2 of the project. Apply those same data preparation steps to the data in `projectDataPart3.csv`.
d. Using the model you chose from Part 2 (as trained on the training data from Part 2), predict DONR and PDONR on the data coming from Step 2c.

3. Mailing List Selection

The purpose of this exercise is to test various strategies for selecting whom to mail in order to obtain the maximum profit for the charity. Recall that there is a cost of $0.68 for each person that you choose to mail. Using the validation data provided in `projectDataPart3.csv`, you can calculate the donations received from a particular mailing list (selected from within the individuals in `projectDataPart3.csv`).

You will be provided with sample code that gives examples of ranking individuals by their PDONR values or by EXAMT=PDONR*DAMT (the expected amount of donation = predicted likelihood × predicted donation amount). The ranked scores are binned (e.g. into deciles) and a score cut-off (corresponding to a number of bins to mail) is selected.

a. The mailing list selection strategy illustrated in the sample code requires you to choose a score to rank and select a cutoff to use on that score. Evaluate this strategy by ranking various scores and calculating the profit obtained on the validation dataset. Scores that you might consider using include the predicted values of DONR, PDONR, and EXAMT. Summarize your findings with tables and figures as appropriate.
b. Select a single mailing list selection strategy to be applied to the test data. Explain your reasoning for why you chose that strategy.

4. Predictions on Test Set

In this exercise, you will make predictions on the Test Set data provided in `projectDataTEST.csv`. You will then select individuals from the Test Set to be mailed in the upcoming charity mailing campaign.

a. Repeat Exercise 1 of this assignment applied to the data in `projectDataTEST.csv.`
b. Repeat Exercise 2 of this assignment applied to the data in `projectDataTEST.csv.`
c. Write your predictions out to a CSV file called `projectPredictionsTEST.csv`. This CSV file should contain the following columns: ID, DONR, PDONR, and DAMT.

[1] Each model type (logistic regression, random forest, SVM) has a means of producing a probability or score in addition to a class assignment of 0 or 1. In R, begin by looking at the documentation for the model building and prediction functions to determine how to obtain the predicted probability values. Use the project Q & A to get additional guidance for a particular model type.

d. Apply the mailing list selection strategy that you chose in Exercise 3b to the Test Set.
e. Write the ID numbers of individuals selected for the mailing list to a CSV file called `projectListTEST.csv`. This CSV file needs only a single column: ID.

Submissions

Submit the following files in Canvas:

1. PDF or Word document that details your findings from the exercises. Include figures and tables as applicable. Clearly indicate the exercise number in your document.
2. Your R code (if more than one .r or .R file, zip them into a single file for upload).
3. Two CSV files: `projectPredictionsTEST.csv` and `projectListTEST.csv`.