NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PREDICT 422: Practical Machine Learning **Charity Project: Part 2**

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 four separate assignments.

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

Charity Problem — Part 2

Data Files

- `dataDict.txt` (unchanged from Part 1)
- `projectDataPart2.csv`

Sample Code

- `SampleCodePart2.R`

Exercises

1. Read Data from CSV File

   Read the data into R from the CSV file `projectDataPart2.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. Data Quality Check

   The purpose of a data quality check is for the user to get to know the data. The data quality check is a quick summary of the values of the data. The summary can be tabular or graphical, but in general you want to know the value ranges, the shape of the distributions, and the number of missing values for each variable in the dataset.

   a. Use R to perform a data quality check on the dataset provided. Report your findings.
   b. Are there any missing values in the data?
   c. Does the data quality check indicate that there are any data anomalies or features in the data that might cause issues in a statistical analysis?

3. Exploratory Data Analysis (EDA)

   The primary purpose of EDA is to look for interesting relationships in the data. While performing the EDA, you will also uncover many uninteresting relationships. It is recommended that you focus on reporting and discussing the interesting relationships in your write-up.

   a. Use R to perform EDA for the dataset provided. The response for the regression problem is DONR. Pay particular attention to relationships between potential predictor variables and the response. Note that ID is for identification purposes only and is not to be used as a predictor.
   b. Report your findings from the EDA. Include tables, figures, and plots to illustrate relationships in the data. Which predictors show the most promise for predicting the donation amount?

4. Data Preparation

   There are several types of data preparation to consider: addressing missing values, transforming variables, deriving new variables, and re-categorizing categorical variables. You should consider performing some or all of these forms of data preparation.

   Briefly describe any data preparation steps that you take. Short sentences and bullet points are fine. From reading your response, I should understand what changes have been made to the data from its raw form (in the CSV file) to the form that you use to train your models. Items to address include:

   a. How did you handle missing values?
   b. Are there any derived or transformed variables that you added to the dataset?
   c. Did you perform any re-categorization of categorical variables?
   d. Are there any variables that you have chosen to remove from the dataset?

5. Dataset Partitioning

   For this assignment, you will employ a hold-out test dataset for model validation and selection.

   a. Hold-Out Test Set
      The first step you should take is to sample 25% of the observations in the dataset to form a hold-out test set. This data will be referred to as the **Classification Test Set** (or simply the Test Set for the remainder of this document). Report the number of observations and the distribution of response values in the Test Set. The data in the Test Set should not be used until Exercise 7 of this assignment.
   b. Training Set
      The remaining 75% of the observations will be referred to as the **Classification Training Set** (or simply the Training Set for the remainder of this document). Report the number of observations and the distribution of response values in the Training Set.

6. Model Fitting

   Use R to develop various models for the response variable DONR. The variables ID and DAMT are not to be used as predictors. Fit at least one model from each of the following four categories. Each model should be fit to the Training Set data only.

   a. Simple logistic regression (ISLR Chapter 4) [Recall that simple logistic regression is logistic regression with a *single predictor variable*.]
   b. Multiple logistic regression or Linear Discriminant Analysis (ISLR Chapter 4)
   c. Tree-based models (ISLR Chapter 8)
   d. Support vector machines (SVM) models (ISLR Chapter 9) or some other model[1] (including another one from category a, b, or c)

   For each model, report the form of the model you are fitting (e.g. the formula used to specify the model). Explain the reasoning for why you are fitting a model of that form (e.g. for simple logistic regression, explain how you selected which predictor to use). Explain any hyper-parameter tuning that you do (e.g. tuning the threshold value for logistic regression). Report summary and diagnostic information as appropriate for each model (in particular, confusion matrices and TP and FP rates).

7. Model Validation

   Use R to perform model validation on the models you fit in Exercise 6. The model validation process is outlined below.

   a. Build a table (in your document) that has one row for each model you fit in Exercise 6. The table should have seven columns (at minimum): Model Name, Training Set Accuracy, Training Set TP Rate, Training Set FP Rate, Test Set Accuracy, Test Set TP Rate, and Test Set FP Rate. You can include additional columns if you would like.
   b. For the Training Set MSE, predict DONR for all of the individuals in the Training Set, and calculate the MSE from the Training Set predictions. Note that it is expected that this MSE value will *underestimate* the test error. The Training Set MSE is included in the Model Validation table for comparison purposes only. If there is a dramatic difference between the Training Set MSE and the Test Set MSE, then that is an indication that the model has overfit the training data.
   c. For the Test Set MSE, predict DONR for all of the individuals in the Test Set, and calculate the MSE from the Test Set predictions.. Note that you do not retrain or refit the model to the Test Set data, nor do you re-tune the hyper-parameters.
   d. In addition, present the Test Set confusion matrix for each model that you build.

8. Model Selection

   Use the table you generated in Exercise 7 to select the best model to carry forward to Part 3 of the Charity Project.

   a. Comment on the predictive accuracy you get from your models.
   b. Explain which of your models you select as being the best performing model and why. Note that model selection should be based on the Test Set metric values. If two models have similar Test Set metric values, then the model with fewer predictors should be selected.

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).

---

[1] You may have some trouble fitting a satisfactory SVM model to the charity data. You are welcome to pursue building an SVM model, but I don't want you to get too stuck on it if things don't go smoothly.