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 $0.99 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 – $0.99 = - $0.19.

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. Exploratory Data Analysis
2. **The Regression Problem**
3. The Classification Problem
4. The Mailing List Problem

Charity Problem — Part 2

Data Files

- `dataDict.txt`
- `trainSample.csv`

Sample Code

- `SampleCodePart2.R`

Exercises

1. Import Data

   a. Read the data into R from the CSV file `trainSample.csv` the same as you did in Part 1 Exercise 1.
   b. Subset the data to only those observations where DONR = 1. This set of observations will be the data used for this assignment (Part 2, The Regression Problem). Name this dataset `regrData`.

2. 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?

3. 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 **Regression 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 5 of this assignment.
b. Training Set
The remaining 75% of the observations will be referred to as the **Regression 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.

4. Model Fitting

Use R to develop various models for the response variable DAMT. The variables ID and DONR 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 linear regression (ISLR Section 3.1) [Recall that simple linear regression is regression with a *single predictor variable*.]
b. Multiple linear regression with subset selection (ISLR Section 6.1)
c. Shrinkage models (ISLR Section 6.2) or Principal Components Regressions (ISLR Section 6.3)
d. Another model of your choice, which may include a second model from one of the three prior categories

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 linear regression, explain how you selected which predictor to use). Explain any hyper-parameter tuning that you do (e.g. tuning the value of λ for Lasso). Report summary and diagnostic information as appropriate for each model.

5. Model Validation

Use R to perform model validation on the models you fit in Exercise 4. 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 4. The table should have three columns (at minimum): Model Name, Training Set MSE, and Test Set MSE. You can include additional columns if you would like.
b. For the Training Set MSE, predict DAMT 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 DAMT 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.

6. Model Selection

   Use the table you generated in Exercise 7 to select the best model to carry forward to Part 4 of the Charity Project.

   a. Comment on the predictive accuracy you get from your models.
   b. Explain which one of your models you select as being the best performing model and why. Note that model selection should be based on the Test Set MSE values. If two models have similar Test Set MSE 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).