NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PREDICT 422: Practical Machine Learning — Charity Project: Part 1

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 four separate assignments.

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

Charity Problem — Part 1

Data Files

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

Sample Code

- `SampleCodePart1.R`

Exercises

1. Read Data from CSV File

   Read the data into R from the CSV file `trainSample.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 two fields containing strings (GENDER 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 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. General EDA
      - The two response variables are DAMT and DONR. The remaining variables (not including ID) are predictor variables.
      - How are the two response variables distributed? Look at the distribution of DAMT both with and without zero values.
      - Look at the distributions of the predictor variables. Are any of the distributions noteworthy?

   b. Regression Problem EDA
      - For this part, treat DAMT as the response variable and ignore DONR. Subset the data to observations where DONR = 1; this will correspond to DAMT > 0.
      - Perform additional EDA with respect to the Regression Problem. Which predictor variables exhibit relationships with DAMT?

   c. Classification Problem EDA
      - For this part, treat DONR as the response variable and ignore DAMT. Be sure to use the full training sample (not just the observations used in part b).
      - Perform additional EDA with respect to the Classification Problem. Which predictor variables exhibit relationships with DONR?

4. Principal Component Analysis (PCA)

   a. Apply the method of PCA (ISLR Section 10.2) to the predictor variables in the full dataset. Be sure to center (mean = 0) and scale (sd = 1) the data.
   b. Generate and include a scree plot (as shown in ISLR Figure 10.4). Discuss the proportion of variance explained by different numbers of principal components. Does there appear to be an elbow or sharp bend in the scree plot? At what number of components does the elbow occur? How many components are needed to explain “most” of the variance in the data?
   c. Generate and include a biplot (as shown in ISLR Figure 10.1). If necessary for legibility, consider showing fewer components in the biplot. Discuss any insights that can be gleaned from the biplot.

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).