Marketing Analytics – Homework 3

Individual Assignment
Due 8:30 AM December 3rd

This dataset gives the characteristics of applicants to a major credit card. The key dependent variable is `card`, which indicates whether a consumer was approved for a credit card. The remaining variables contain other relevant information about each consumer. The data on real-world setting that appeared in Greene, 2003.

However, I have modified the initial dataset, while keeping the relationships between variables intact. Download the training dataset that corresponds to *the last digit of your student number*. That is, if your student number ends in 4, you should download "Homework 3 Training Data – Number 4.csv".

Assignment Materials for Download:

1. An Rmarkdown template

2. A dataset corresponding to the last digit of your student number.

Submission Checklist:

To help us grade the assignments efficiently and correctly, we ask that you submit your assignments in a specific format. A complete submit the following to blackboard:

- A .rmd Rmarkdown file, based on the template for this assignment with all the code used to estimate your models.

- A .html file, generated by knitting the .rmd file in RStudio.

- An R workspace containing your two chosen models. I have provided code in the template to save the models for you. Just insert your student number on line 27 of the template

```
save(model1A, model1B, file = '[student number].Rdata')
```

- All file names should be '[student number].[file extension]', where you replace everything the square brackets with the appropriate values.

- Place all files in a single zip before submission

Data Guide:

card: Boolean. Was the application for a credit card accepted?
reports: Number of major derogatory reports.
age: Age in years plus twelfths of a year. income Yearly income (in USD 10,000).
share: Ratio of monthly credit card expenditure to yearly income.

expenditure: Average monthly credit card expenditure.
owner: Boolean. Does the individual own their home?
selfemp: Boolean. Is the individual self-employed?
dependents: Number of dependents.
months: Months living at current address.
majorcards: Number of major credit cards held.
active: Number of active credit accounts.

Predictive Analysis (16 Marks)

Now, you will estimate a predictive model to predict whether a consumer is approved for a credit card, using the dataset that corresponds to the last digit of your student number. This might be useful to a firm that is selecting which consumers to target, choosing how much to pay for the contact information of a consumer, or a firm that is simply trying to forecast demand. Firms with better predictive models will be able to more efficiently target consumers, or make better purchasing decisions. Similarly, the quality of your predictions will form part of your grade here. You will submit two predictive models:

a) The first predictive model, stored as `model1A` should use all the data *except* `expenditure`

b) The second predictive model, stored as `model1B`, can use all the provided independent variables, including `expenditure`

Save your models to an R Workspace with the code provided in the template.

To keep the computational burden low for this assignment, **you may only use linear regressions or MARS models in this section.** You can complete this section using the `runif, subset, lm, earth, predict, and mean` functions.

Your final submission will include a Rdata file with your two models. We will also look at your RMD file to see how you trained your model.

The two models will be graded out of 8 marks. The marks will be assigned as follows:

1. Correctly submitting the model will yield 2 out of 8 marks

2. I have held back a sizable portion of each dataset to evaluate your predictions. The graders will use this to evaluate the quality of your predictions, in terms of average out of sample mean-squared error. They will look at the distribution of predictions for your data set, and give marks based on the relative quality of your predictions.

3. If your predictions are in the bottom quartile, we will dive into the code you submitted. So long as your code demonstrates that you followed the recommendations below, you will receive at least 6 marks out of 8.

To improve your predictions, I have the following recommendations:

1. Run at least 10 different model specifications. Someone who assesses 20 models will find a better model than someone who assesses 2.

2. At the same time, be thoughtful about the models you are running. Look to your previous model estimates and the data exploration process to see what variables worked in your context. The best assignment estimated only a fraction of the models that others did, but they learned with each model they estimated. Other groups estimated thousands of models, but didn't think through their approach, and had worse predictions.

3. Use k-fold cross validation following the steps in the notes. Do not use the built-in cross-validation in the `earth` function as you will not get consistent results

4. Tune your model by trying different model specifications. This includes different types, formulas, and tuning parameters. Vary all three of these.

Please submit all 3 required files.

Bibliography

Greene, W. (. (2003). *Econometric Analysis, 5th edition.* Upper Saddle River, NJ: Prentice Hall.