Causal and Predictive Analytics – Homework 3

Individual Assignment

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 8 digit student number*. That is, if your 8 digit student number ends in 4, you should download "Homework 3 Training Data – Number 4.csv".

Assignment Materials for Download:

1. An R file with code for your use.
2. A training data set that corresponds to your student number, as above.

Submission Checklist:

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

- A txt or R file containing the code you used to estimate your predictive models, based on the script I provided you. Marks in the assignment are assigned based on whether the models you trained (even if you didn't submit them). I need to see these models in the code in your initial submission order to assign marks.
- Save your models model1A and model1B in an R workspace file named MyModels.RData using the provided template code.
- Answers to written questions below, submitted through blackboard.

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.

Part 1: Predictive Analysis (14 Marks)

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` can use all the data *except* `expenditure`

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

Running a predictive competition for this many students is a technical challenge. To receive full marks and keep our grading process efficient, please do the following:

1. Save your models by running the code after the comment ‘Run the following code after you've completed training.’

2. To keep the computational burden low for this assignment, you may only use linear regressions and MARS models in this section.

3. Do not use any packages other than earth, it makes the assignment more difficult to grade.

The two models will be graded out of 7 marks. The rubric is as follows:

1. Correctly submitting a zip file with your R/txt file and Rdata file following the above guidelines. 2 marks.

2. Run at least 30 different specifications (for each of the two models model1A and model1B). 1 mark.

3. Tune your model by trying both lm and earth models. 1 mark.

4. Try different tuning parameters for the mars model by changing the ‘thresh’ argument. 1 mark.

5. 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. This will be used to assign 2 marks. If you do not correctly submit the MyModels.RData file, then you will get 0/2 on this section.

Look at the telemarketing case and the predictive tuning async content for help on how to train a predictive model. To improve your predictions, 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.

Part 2: Discussion Questions

a) Suppose after seeing the results show that consumers with higher monthly expenditures are more likely to be approved for a credit card. Based on model1B, can we conclude this is a causal relationship? (4 marks)

b) Suppose a colleague proposes using a random forest (another predictive model type) to improve the quality of the predictions. In words, describe the steps and evaluation metrics you would use to assess the suitability of a random forest model for this prediction task. (4 marks)

Bibliography

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