Stat 6242
Data Analysis Project
Due December 15, 2016 in Blackboard.

Default of Credit Card Clients Dataset

Default Payments of Credit Card Clients in Taiwan from 2005: This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

There are 24 variables:

- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default.payment.next.month**: Default payment (1=yes, 0=no)

Data Files: Use the training data to build the best model for predicting **default.payment.next.month** and the test to validate its predictive ability.

- Training dataset - UCI_Credit_Card_train
- Test dataset - UCI_Credit_Card_test

Note: The original dataset contains 30,000 observations. I randomly selected 5,000 for training and 5,000 for testing.

Objectives of the Analysis: Answer the following questions

1. How does the probability of default payment (**default.payment.next.month)** vary by categories of different demographic variables?
2. Which variables are the best predictors of default payment?
3. What is the best model for predicting default payment?

You can work in groups of 1-2 in this project!

Write-up: You will submit a write-up, via Blackboard, on December 15 by 11:59 pm. This write-up should be a polished report, with figures and R code as you deem appropriate. You do not need to submit your R code in its entirety. Your report should have the following sections (you can of course add subsections if you want), and should be no more than 6 pages.

(a) Introduction: Describe your data set. What is the problem you are trying to solve? This should be brief.
(b) Analysis: How did you make your predictions? Describe this process in detail. You can use any of the classification techniques that we learned in the course, or any other techniques as long as they are adequately described. What predictor variables did you include? What technique did you use, and why did you choose it? What assumptions, if any, are being made by using this technique? If there were tuning parameters, how did you pick their values?
(c) Summary and Conclusion: Answer the three questions in the objectives of the analysis above, emphasizing what the best model for predicting defaulting is and also whether it provides a good fit to the data.

Presentation: Your team will give a short slide presentation (10 min) on Friday December 16 summarizing your work. You can show a maximum of 6 slides. You should explain your predictive model and how you arrived at it. Your slides must be in PDF format, and must be submitted to BB by 1:00 pm on Friday, December 16. Slides with mostly (color) pictures, and not much text, are encouraged.

Acknowledgement

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

The original dataset can be found here at the kaggle website.