STATS 503 Proposal, Team 14, 3/15

Team members: Chen Xie, Xun Wang, Xinye Jiang

Dataset Description:

The dataset is about the case of customers default payments in Taiwan from April to October in 2005. It has a binary variable, default payment (Yes = 1, No = 0), as the response variable and the following 23 variables as explanatory variables.

Variables	**Descriptions**	**Type**	**Details**
Y	Default Payment	Binary	Response
X1	Amount of the given credit	Numeric	NT dollar
X2	Gender	Binary	1 = male; 2 = female
X3	Education	Categorical	1 = graduate school; 2 = university; 3 = high school; 4 = others
X4	Marital status	Categorical	1 = married; 2 = single; 3 = others
X5	Age	Numeric	year
X6-X11	History of past payment	Categorical	History of past payment from April to September, 2005
X12-X17	Amount of bill statement	Numeric	Amount of bill statement (NT dollar) from April to September, 2005
X18-X23	Amount of previous payment	Numeric	Amount of previous payment (NT dollar) from April to September, 2005

Proposed Algorithm:

1. Data management: Clean data and deal with missing values.
2. Data analysis: Summarize the data numerically and graphically. Data transformation (PCA / ...)
3. Classification and Prediction: Explore and visualize the relationship between the response variables and the predictors (classification methods: decision tree / logistic regression /...).
4. Comparison: Compare prediction performance among classification methods

Techniques may be applied: PCA, Cross validation, classification methods (decision tree / logistic regression)

Source: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients