6242 Data Analysis Project:

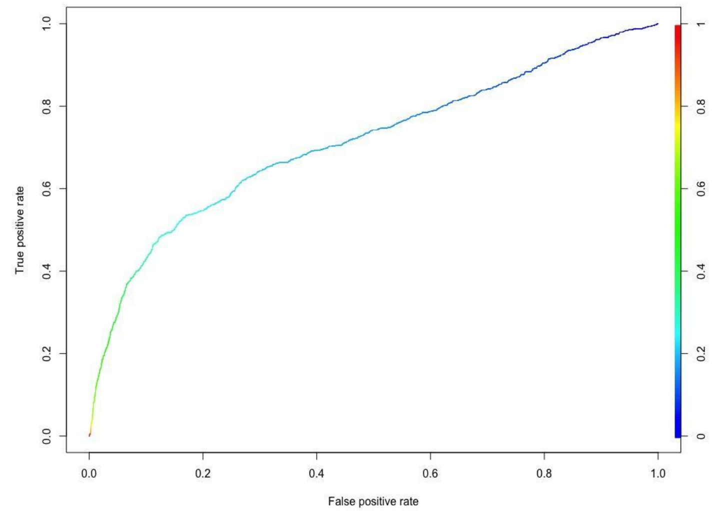
Default Payments of Credit Card Clients in Taiwan from 2005

Michael Chiang

Jian Sun

Logistic Regression

```
Call:
glm(formula = UCC$default.payment.next.month ~ MARRIAGE + AGE +
    PAY_0 + PAY_2 + PAY_3 + BILL_AMT1 + BILL_AMT4 + PAY_AMT1 +
    PAY_AMT2 + PAY_AMT5, family = binomial("logit"), data = UCC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5895  -0.6789  -0.5481  -0.3035   3.1742

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.275e+00  2.318e-01  -5.501 3.77e-08 ***
MARRIAGE    -1.896e-01  7.728e-02  -2.453 0.014162 *
AGE          1.206e-02  4.257e-03   2.832 0.004629 **
PAY_0        5.649e-01  4.363e-02  12.948  < 2e-16 ***
PAY_2        1.111e-01  4.901e-02   2.267 0.023386 *
PAY_3        1.445e-01  4.445e-02   3.251 0.001152 **
BILL_AMT1   -4.858e-06  1.317e-06  -3.688 0.000226 ***
BILL_AMT4    3.113e-06  1.490e-06   2.089 0.036728 *
PAY_AMT1    -8.345e-06  4.368e-06  -1.910 0.056070 .
PAY_AMT2    -1.650e-05  5.616e-06  -2.939 0.003293 **
PAY_AMT5    -1.066e-05  4.908e-06  -2.171 0.029895 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5207.7  on 4999  degrees of freedom
Residual deviance: 4578.6  on 4989  degrees of freedom
AIC: 4600.6

Number of Fisher Scoring iterations: 6
```



Area Under Curve = 0.7123

Result - Classification Rate = 71.23%

Linear Discriminant Analysis

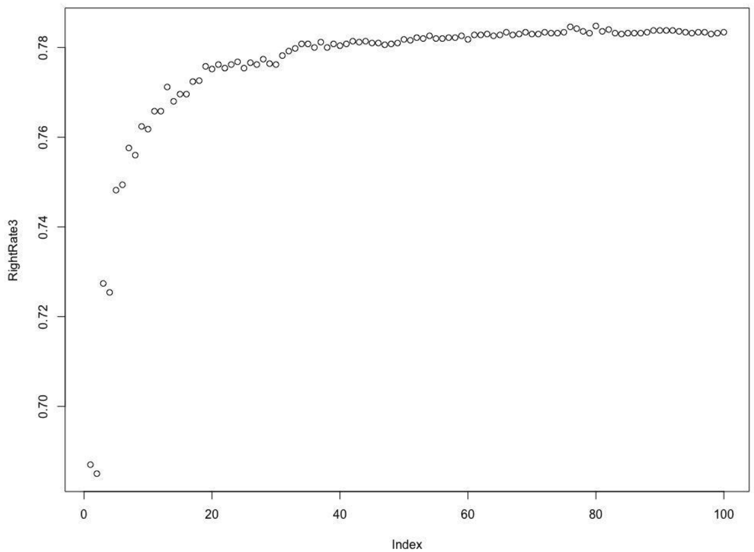
```
Coefficients of linear discriminants:
                     LD1
LIMIT_BAL -4.928587e-07
SEX       -6.585150e-02
EDUCATION -8.032433e-02
MARRIAGE  -2.488607e-01
AGE        1.581510e-02
PAY_0      6.701662e-01
PAY_2      1.601329e-01
PAY_3      1.733721e-01
PAY_4     -5.809809e-02
PAY_5      1.246299e-01
PAY_6     -8.132672e-02
BILL_AMT1 -3.876569e-06
BILL_AMT2 -6.001636e-07
BILL_AMT3 -3.147193e-06
BILL_AMT4  5.126440e-06
BILL_AMT5 -2.748228e-06
BILL_AMT6  2.587035e-06
PAY_AMT1  -4.613871e-06
PAY_AMT2  -3.912667e-06
PAY_AMT3  -5.001320e-06
PAY_AMT4   1.251129e-06
PAY_AMT5  -6.097643e-06
PAY_AMT6  -4.704967e-07
```

fitted response values from our LDA model

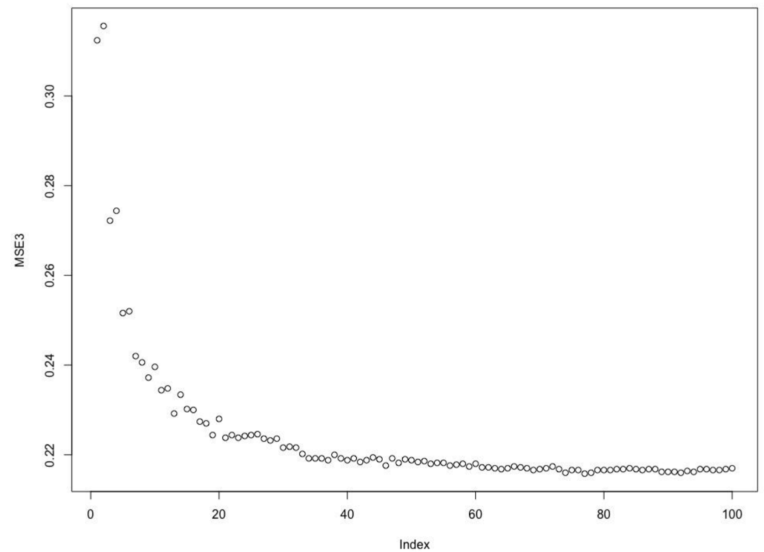
Result - Classification Rate = 80.9%

k-Nearest Neighbors Classification

classification rate vs k



MSE vs k



we select k =7

Result - Classification Rate = 75.8%

Support Vector Classifier

optimization problem
[Figure1 to elaborate]
[Figure2 to elaborate]

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

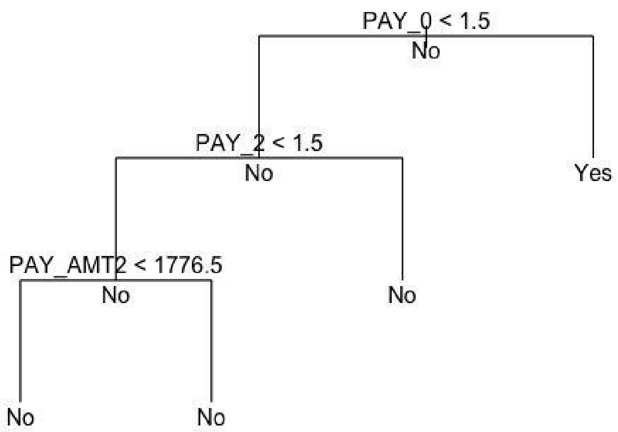
tuning parameter: we select cost = 4

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
    4

- best performance: 0.188

- Detailed performance results:
  cost  error dispersion
1  0.1 0.1896 0.02669665
2  1.0 0.1882 0.02652378
3  4.0 0.1880 0.02602563
4  5.0 0.1880 0.02602563
5  5.5 0.1882 0.02593068
6  6.0 0.1880 0.02602563
```

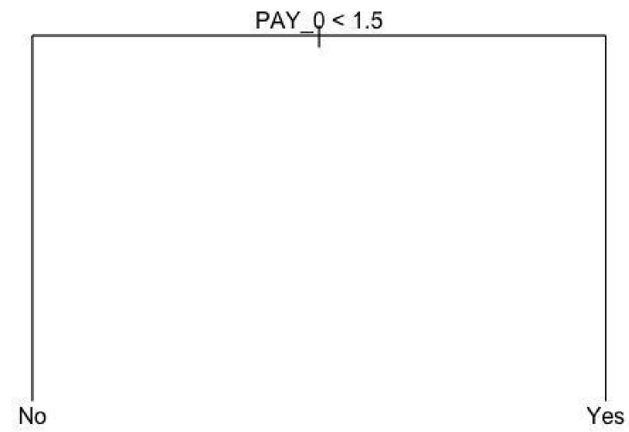
Result - Classification Rate = 80.82%

Classification Tree

classification tree prior to pruning



classification tree after pruning



Result - Classification Rate = 81.54%

Conclusion

1. How does the probability of default payment (default.payment.next.month) vary by categories of different demographic variables?

- Logistic
  - Marriage
  - Age
- LDA
  - Sex
  - Education
  - Marriage
  - Age
- Classification Tree
  - None

2. Which variables are the best predictors of default payment?

- Logistic, LDA & Classification Tree
  - PAY_0
- Logistic & LDA
  - MARRIAGE
  - AGE
  - PAY_2
  - PAY_3

3. What is the best model for predicting default payment?

- Classification Tree
  - Classification Rate = 81.54%