7 Group 7

Predict Credit Card Default Rate with Classification Methods

Instructor: 盧子彬

Group 7 Team members: 余奇祐、黃慶杰

Outline

A Data Introduction and management

B Support Vector Machine (SVM)

- Modeling
- Comparisons
- Results

C Classification And Regression Tree (CART)

- Modeling
- Comparisons
- Results

D Conclusion and Discussion

Data Introduction

UCI

Machine Learning Repository

ABOUT DATA

- From UCI dataset
- Default of credit card clients data set in 2006
- 23 explanatory variables with 30,000 records

Data Management

Variables

- Amount of the given credit (individual consumer and his/her family)
- Gender (1 = male; 2 = female)
- Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- Marital status (1 = married; 2 = single; 3 = others)
- Age (year)

Basic information

Data Management

Variables

- The repayment status from April to September in 2005
  → sum positive months
- Amount of bill statement from April to September in 2005
  → max amount of bill statement
- Amount paid from April to September in 2005
  → max amount paid

Recoded variables

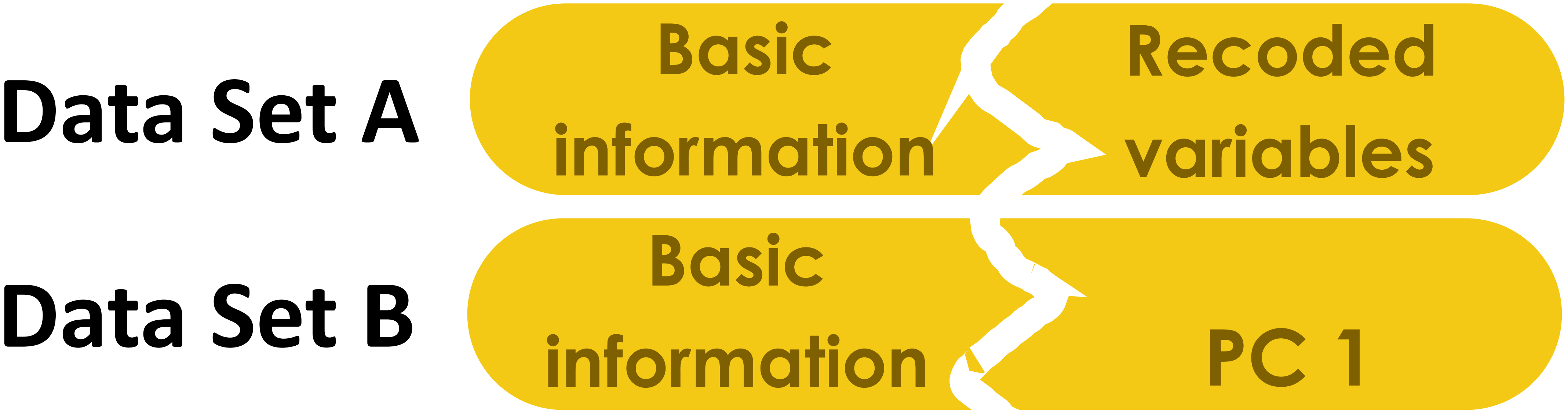
Data Management

Variables

- The repayment status from April to September in 2005
- Amount of bill statement from April to September in 2005
- Amount paid from April to September in 2005

PC 1

Data Management



Data Set A

Basic information

Recoded variables

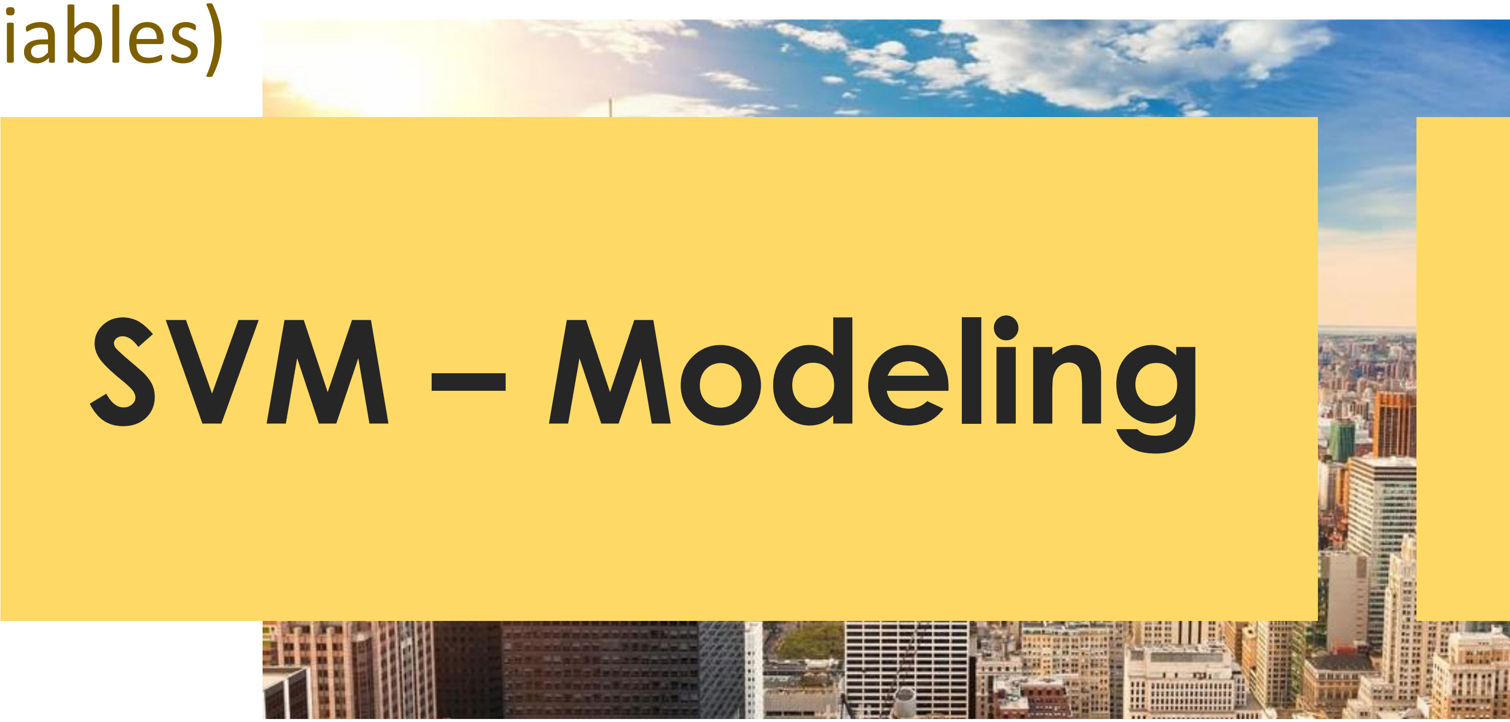
Data Set B

Basic information

PC 1

SVM – Modeling

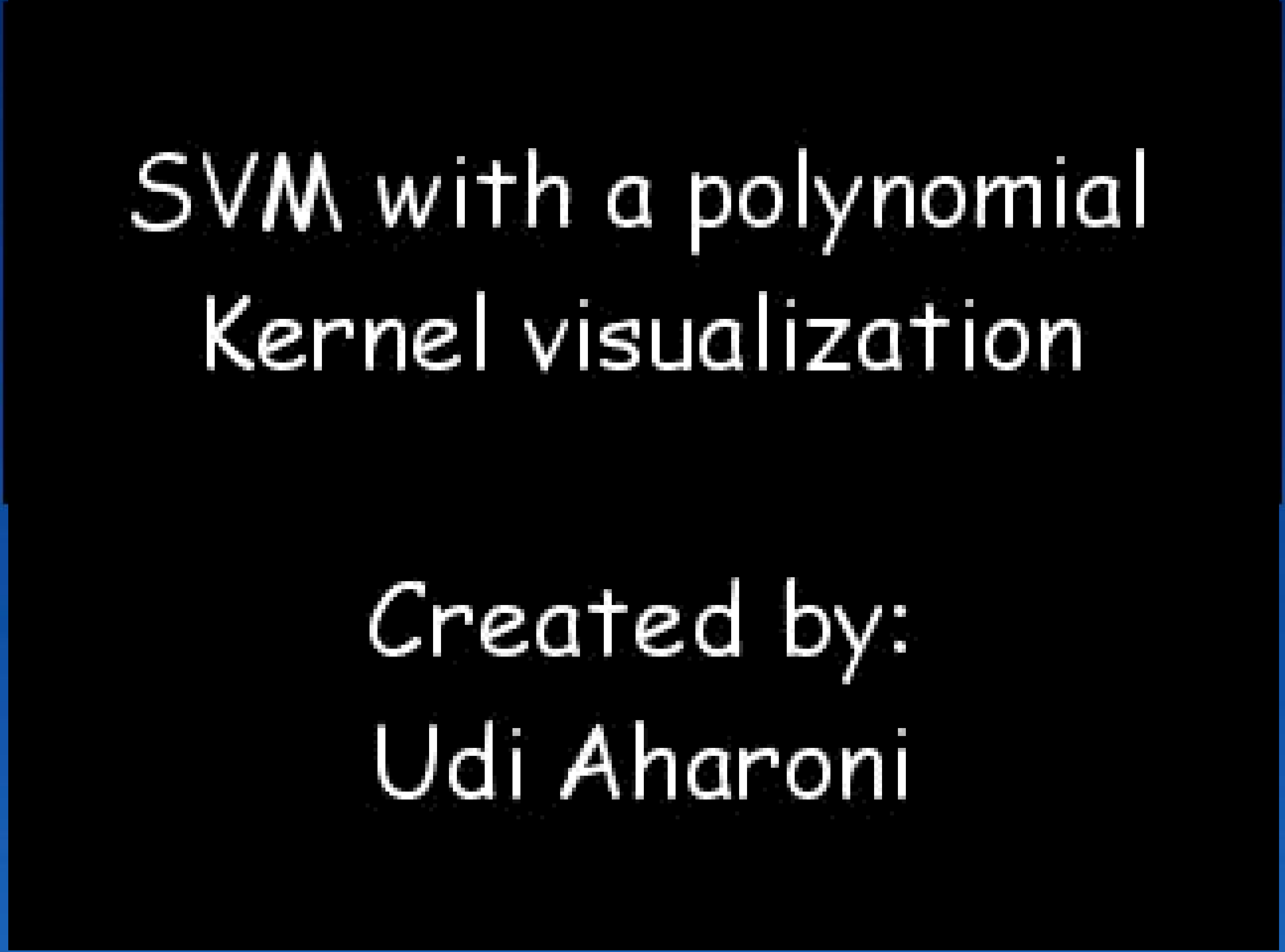
- **Compare two datasets:**
  - Data set A (recoded variables)
  - Data set B (PC 1)
- **Compare four kernels:**
  - Linear
  - Polynomial
  - Radial
  - Sigmoid
- **10-fold cross-validation accuracy rate**



SVM with a polynomial Kernel visualization

Loading......

[illegible]%



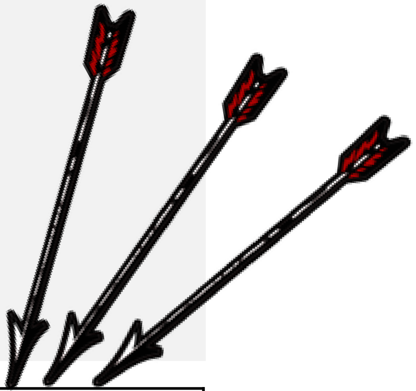
SVM with a polynomial Kernel visualization

Created by:
Udi Aharoni

SVM – Comparisons

10-fold cross-validation accuracy rate

- Data set A is better than Data set B.
- Sigmoid performed worst.



	Linear	Polynomial	Radial	Sigmoid
Data set A	80.37%	80.34%	80.39%	71.73%
Data set B	77.88%	77.88%	77.88%	65.69%

SVM – Comparisons

Running time

- Linear spent the least time.

	Linear	Polynomial	Radial	Sigmoid
Data set A	8.29 mins	10.67 mins	29.89 mins	10.11 mins
Data set B	3.44 mins	**1.39 hours**	25.14 mins	12.89 mins

PART ONE

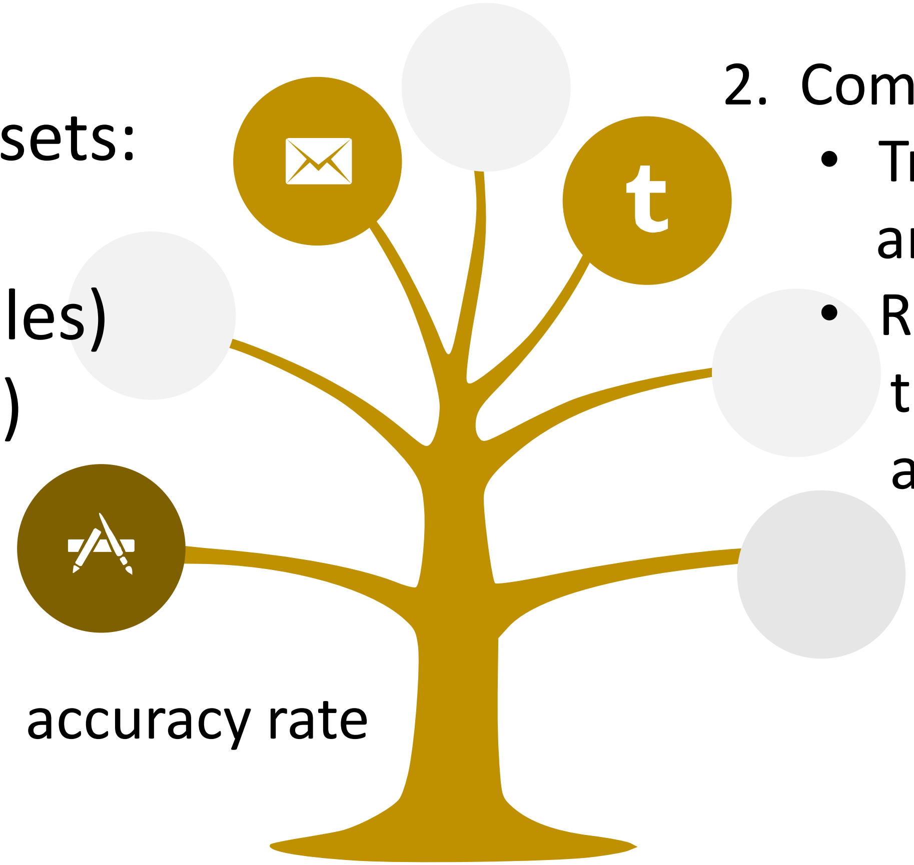
SVM

SVM – Results

Based on the 10-fold cross-validation accuracy rate and running time, **SVM with linear kernel** applying to **Data set A** may be preferable.

CART – Modeling

1. Compare two datasets:
   - Data set A (recoded variables)
   - Data set B (PC 1)



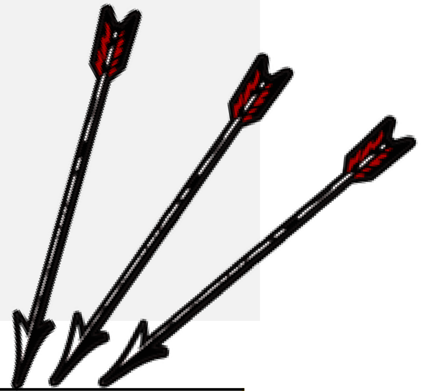
2. Compare tree with random forest:
   - Tree: minbucket = 100 and maxdepth = 5
   - Random forest: number of tree = 500 and number of variable = 3

3. 10-fold cross-validation accuracy rate

CART – Comparisons

10-fold cross-validation accuracy rate

- Data set A is better than Data set B.
- Tree is better than random forest.



	Tree	Random forest
Data set A	80.46%	79.98%
Data set B	77.88%	76.35%

CART – Comparisons

Running time

- Tree computed faster than random forest.

	Tree	Random forest
Data set A	5.33 secs	3.47 mins
Data set B	1.38 secs	3.14 mins

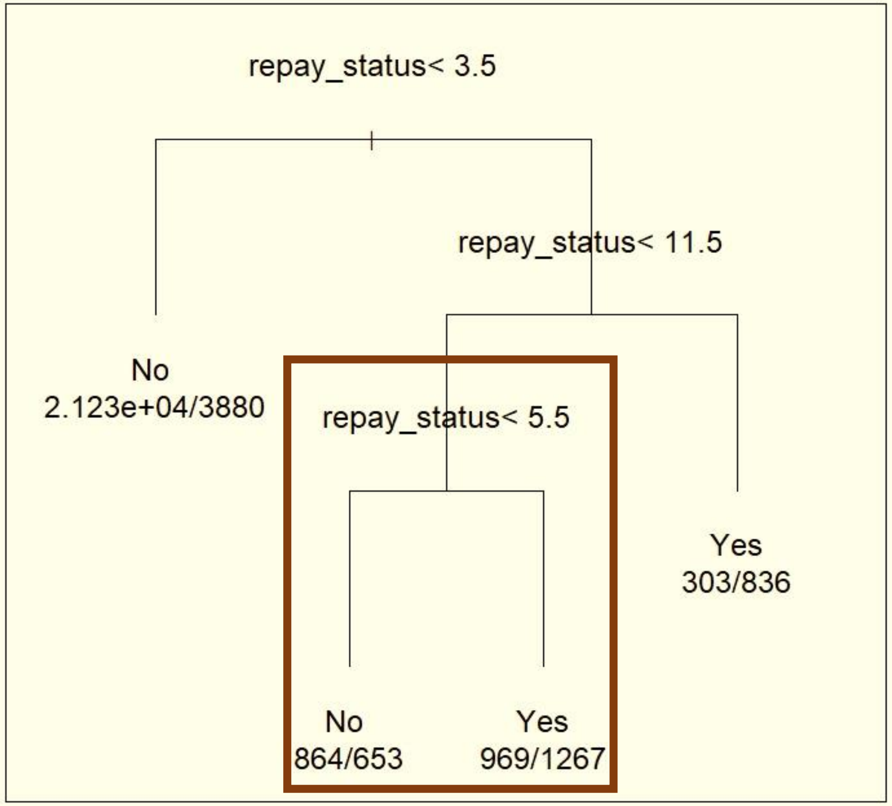
PART TWO

CART

CART – Results

- The repayment status (recoded) is informative.
- The third split performed poor.

CART model with Data set A



Conclusion

Origin data: **77.88% (No) versus 22.12 (Yes)**
SVM with linear kernel for Data set A: **80.37%**
CART for Data set A: **80.46%**

Depends on user!

- *Data quality*
- *Data typing error*
- *Lack of other features*
- *Another classification approach*

Final Discussion