Credit Scoring

Ying Liu, Yong Shi

University of Chinese Academy of Sciences
&
People’s Bank of China

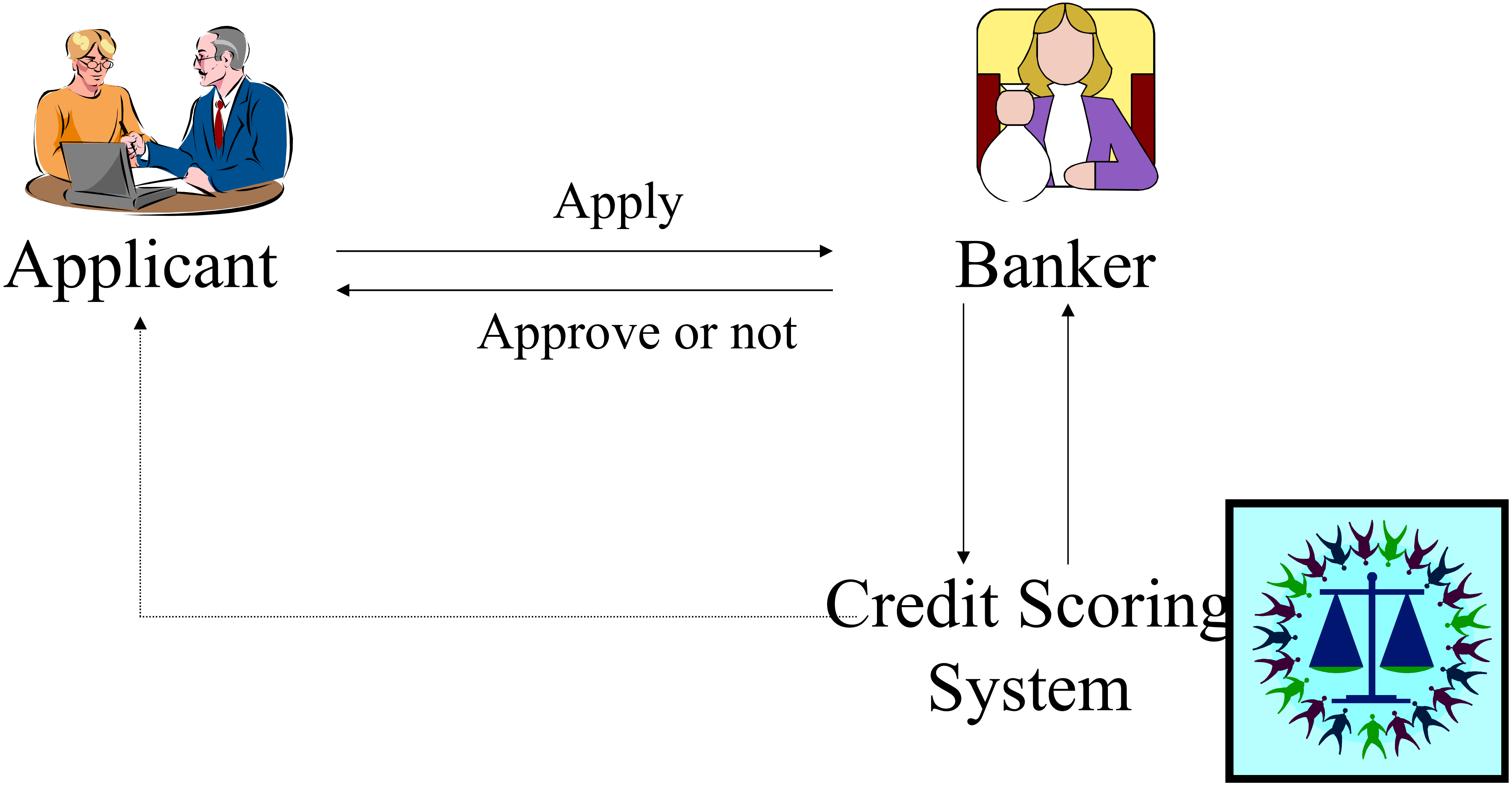
Outline

- Motivation
- Process Flow
- Methods

Motivation

- Use a customer's history of loan, mortgage, credit cards to build a classification / prediction model
- Divide customers into two groups: "good" vs. "bad"
- Assign each customer a score of risk
- A technique for financial institutions to control financial risk, reduce payment delinquency

An Example

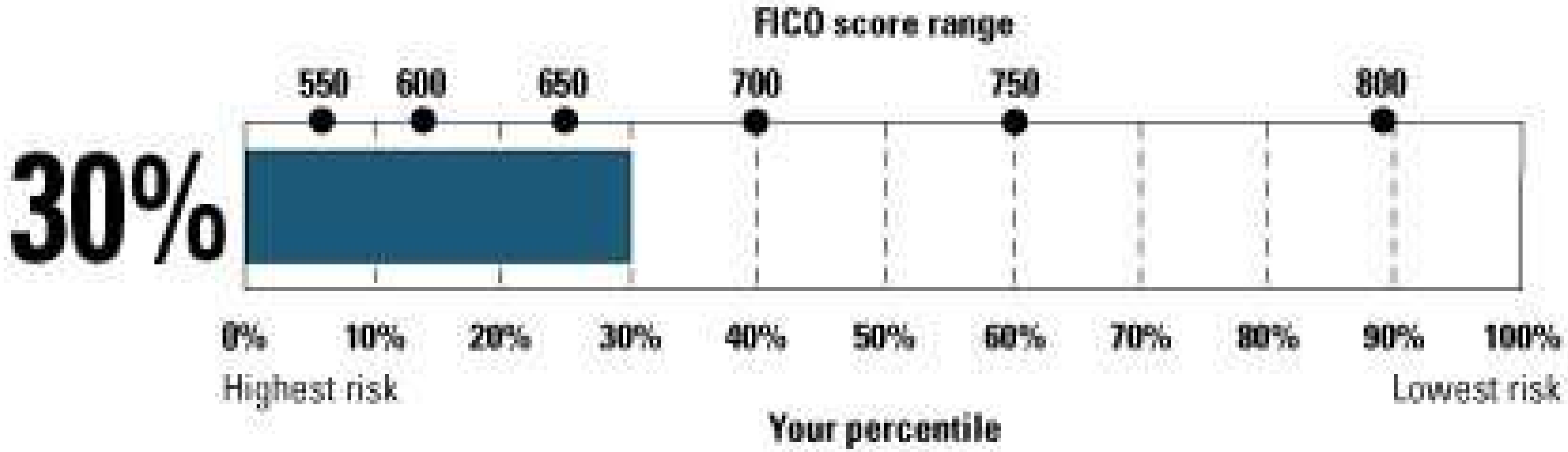

Apply
Applicant
Banker
Approve or not
Credit Scoring
System

An Example Dataset

C_id	sex	age	income	Edu	# credit cards	Payment ratio per month	# loans	Payment ratio per month	…	Good/ bad
12	0	34	50K	BS.	1	100%	1	100%	…	1
14	1	29	60K	BS.	2	20%	1	50%	…	1
135	1	46	100K	MS.	4	100%	2	100%	…	0
…	…	…	…	…	…	…	…	…	…	

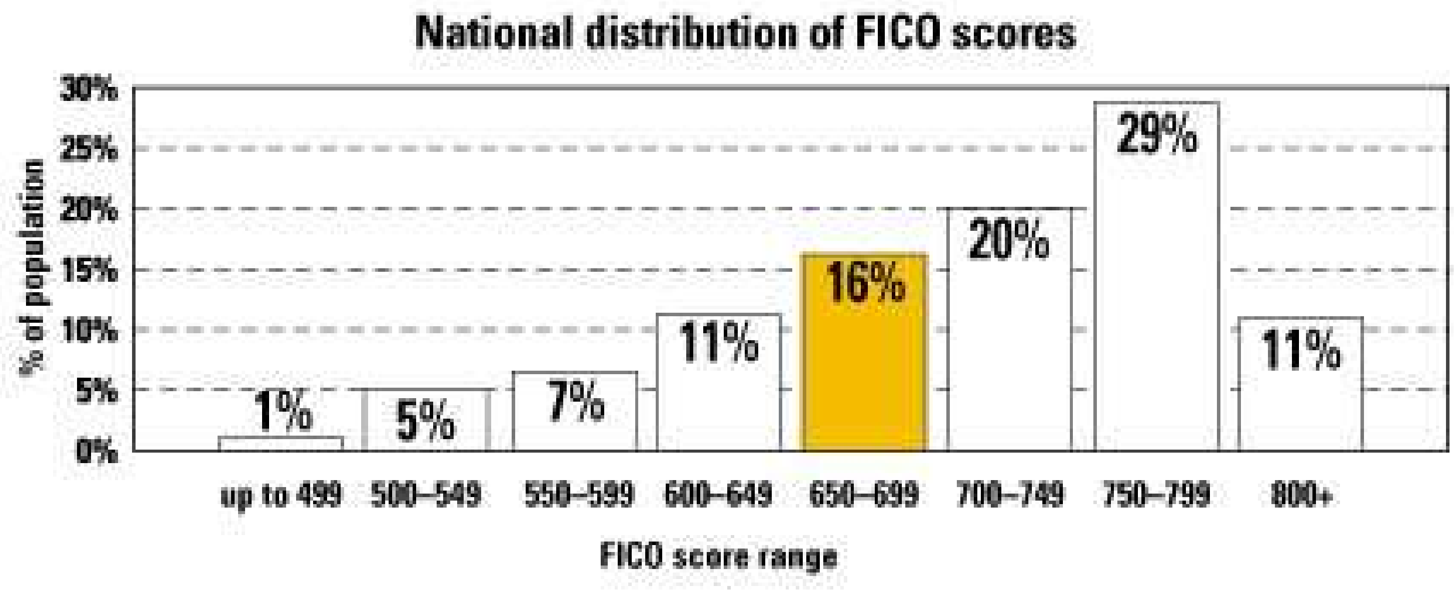
An Example

- FICO score
  - 350 – 850
  - The higher the score, the lower the risk
  - If reject customer < 670, 30% payment delinquency is circumvented

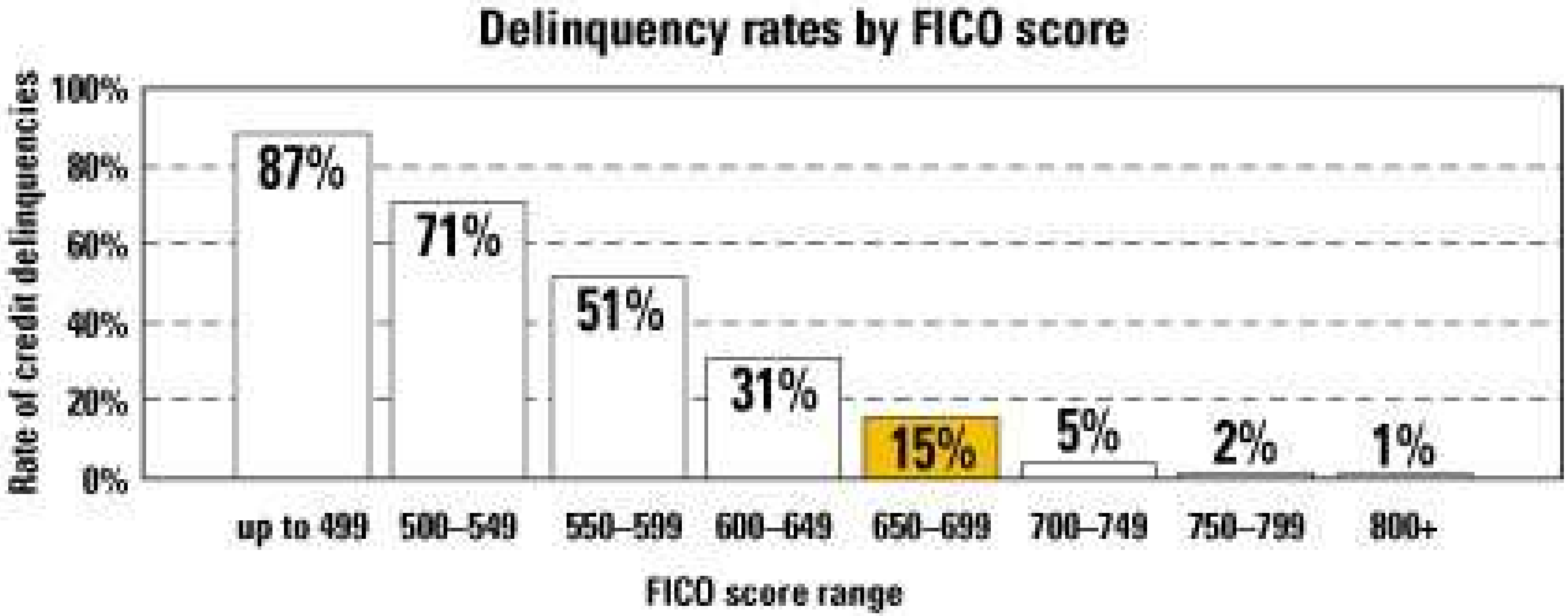


An Example

- FICO score distribution

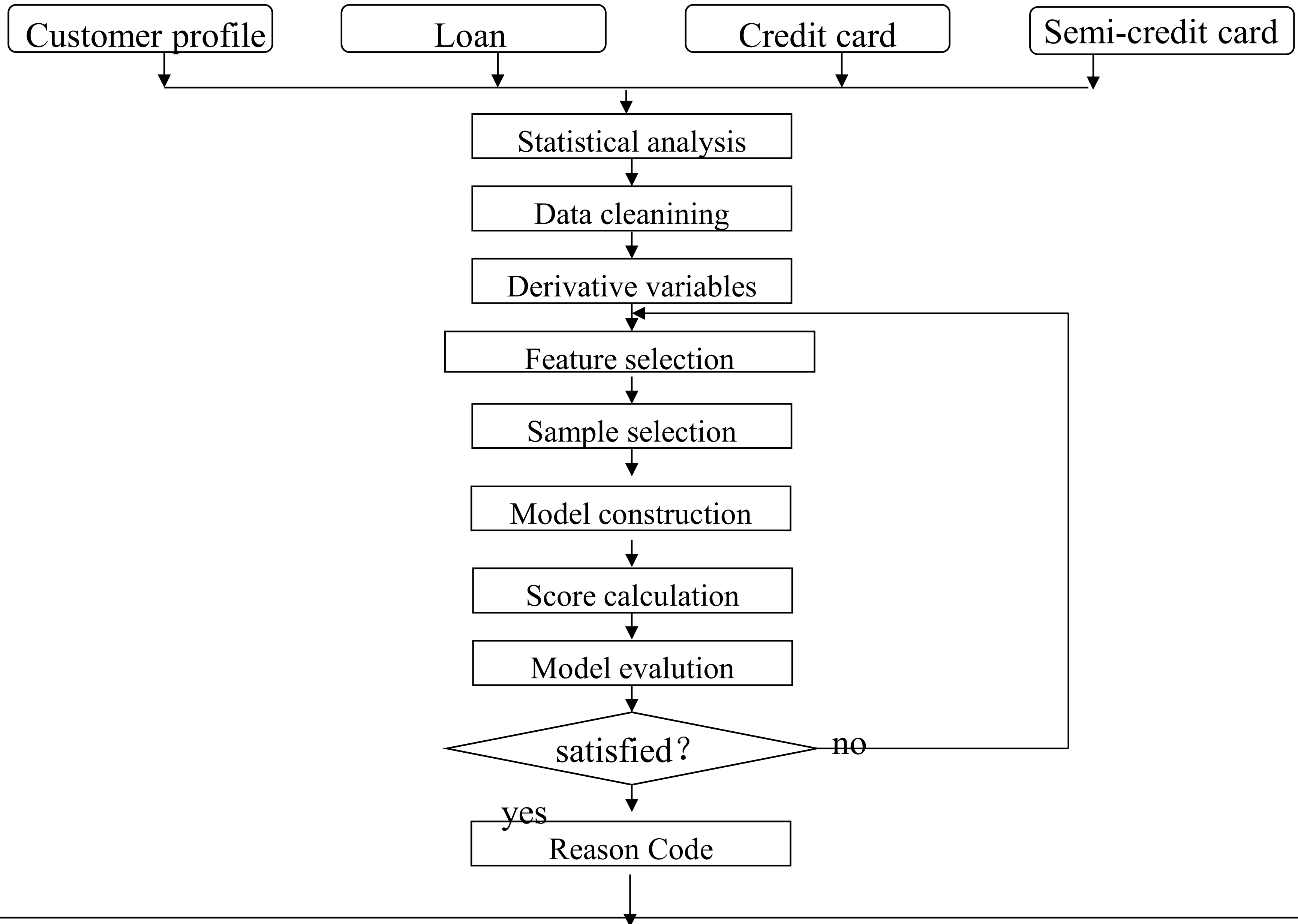


An Example

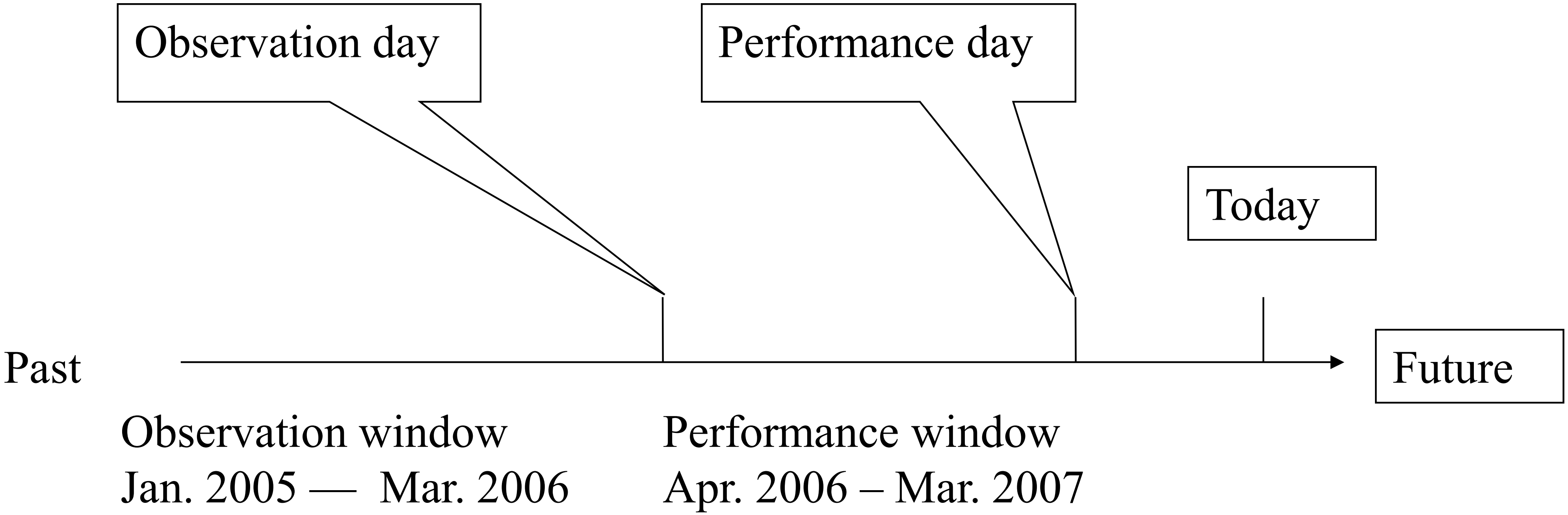


Delinquency rates by FICO score
Rate of credit delinquencies
100%
80%
60%
40%
20%
0%
87%
71%
51%
31%
15%
5%
2%
1%
up to 499
500–549
550–599
600–649
650–699
700–749
750–799
800+
FICO score range

Process Flow



Two-group Classification



- “bad” customer: consecutive payment delinquency in 3 three months in the performance window
- “good” customer: no payment delinquency in the performance window
- “grey”: in between “bad” and “good”

Statistics

- 39 attributes in the raw data
- 1999724 customers，4321357 accounts，2.16 accounts per customer
- 52461470 records
- Outliers, missing values
- Data cleaning is required

Data Cleaning

- Principle： remove any account with outlier
- 902325 customers remained， 1233300 accounts remained
- 308053 “good”， 33483 “bad”
- Fill in the missing values based on its risk tendency
- …

Derivative Variables

- Original attributes may not has strong capability in risk prediction
- Derivative variables may be more capable
- Derived from the original attributes
- Integrate background knowledge, professional experience
  - Seven categories: delinquency in history, current delinquency， debt， credit history， new account， types of loan， others
  - 459 variables in total

Derivative Variables

- 1. payment delinquency in history
  - Number of accounts without delinquency in the observation window
  - Number of accounts with delinquency in the observation window
  - Total number of delinquency in the observation window
  - The date of the most recent delinquency
  - … …

Derivative Variables

- 2. Current payment delinquency
  - Number of accounts without delinquency in the performance window
  - Number of accounts with delinquency in the performance window
  - Total number of delinquency in the performance window
  - Total balance in the accounts with delinquency
  - Total account limit in the accounts with delinquency
  - … …

Derivative Variables

- 3. Debt
  - Average balance in credit cards
  - Max balance in credit cards
  - Current balance in credit cards
  - Monthly credit card payment
  - Monthly mortgage payment
  - Average amount of utilization of credit cards
  - Max amount of utilization of credit cards
  - Proportion of outstanding loans
  - … …

Derivative Variables

- 4. Credit history
  - Average length of accounts
  - Max length of accounts
  - Min length of accounts

Derivative Variables

- 5. New account
  - Number of credit score inquiries
  - Number of new accounts

Derivative Variables

- 6. Types of loan
  - Account types
  - Number of accounts that have been paid off
  - Number of accounts not paid off

Derivative Variables

- 7. Temporal data
  - Compress the monthly records, e.g. average for the last 12 months, average for the most recent 6 months, average for the most recent 3 months, …

Feature Selection

- T-test
  - For each variable in normal distribution, test if the “good” samples and the “bad” samples are distinguishable
- Non-parametric test
  - For each variable in non-normal distribution, test if the “good” samples and the “bad” samples are distinguishable
- log(odds)
  - Partition the range of a variable into bins
  - Calculate the rate of the “good” and “bad” for each bin
  - Test if the log(odds) is almost the same, the variable is weak in distinguishing

Feature Selection

- Partition the remaining variables into 7 categories
- Compute the correlation coefficients among variables within each category
- Keep the variables with least correlation in each category

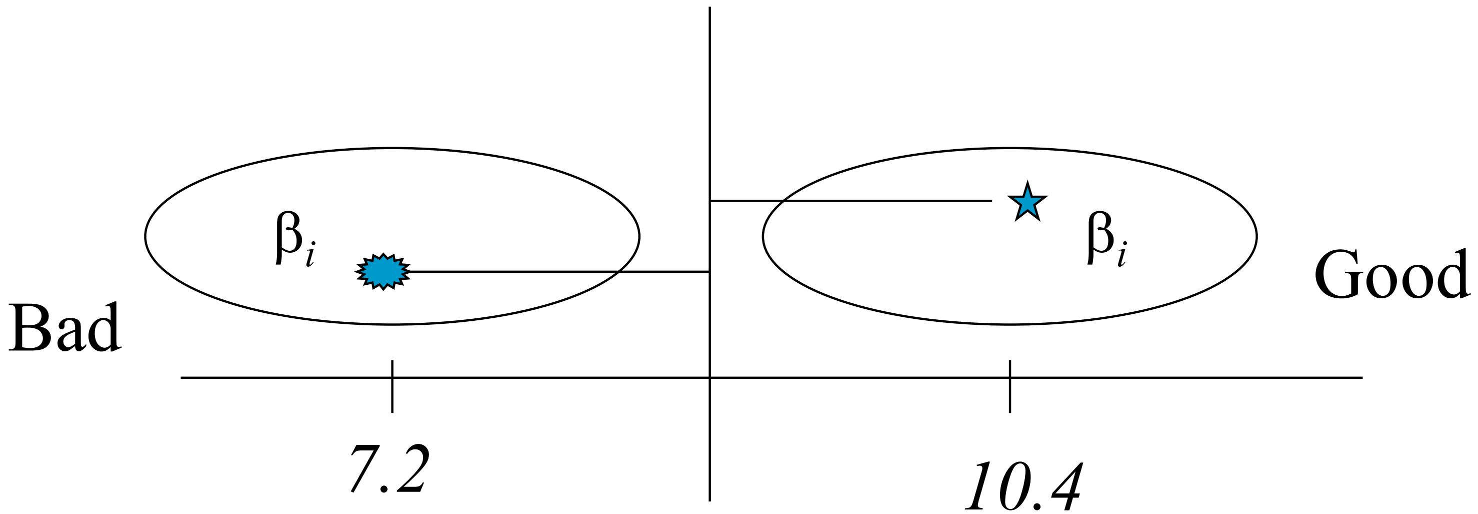
Sample Selection

- Stratified sampling
  - Use disproportional allocation
- Training dataset
  - 5000 “good” vs. 5000 “bad”

Classification Models

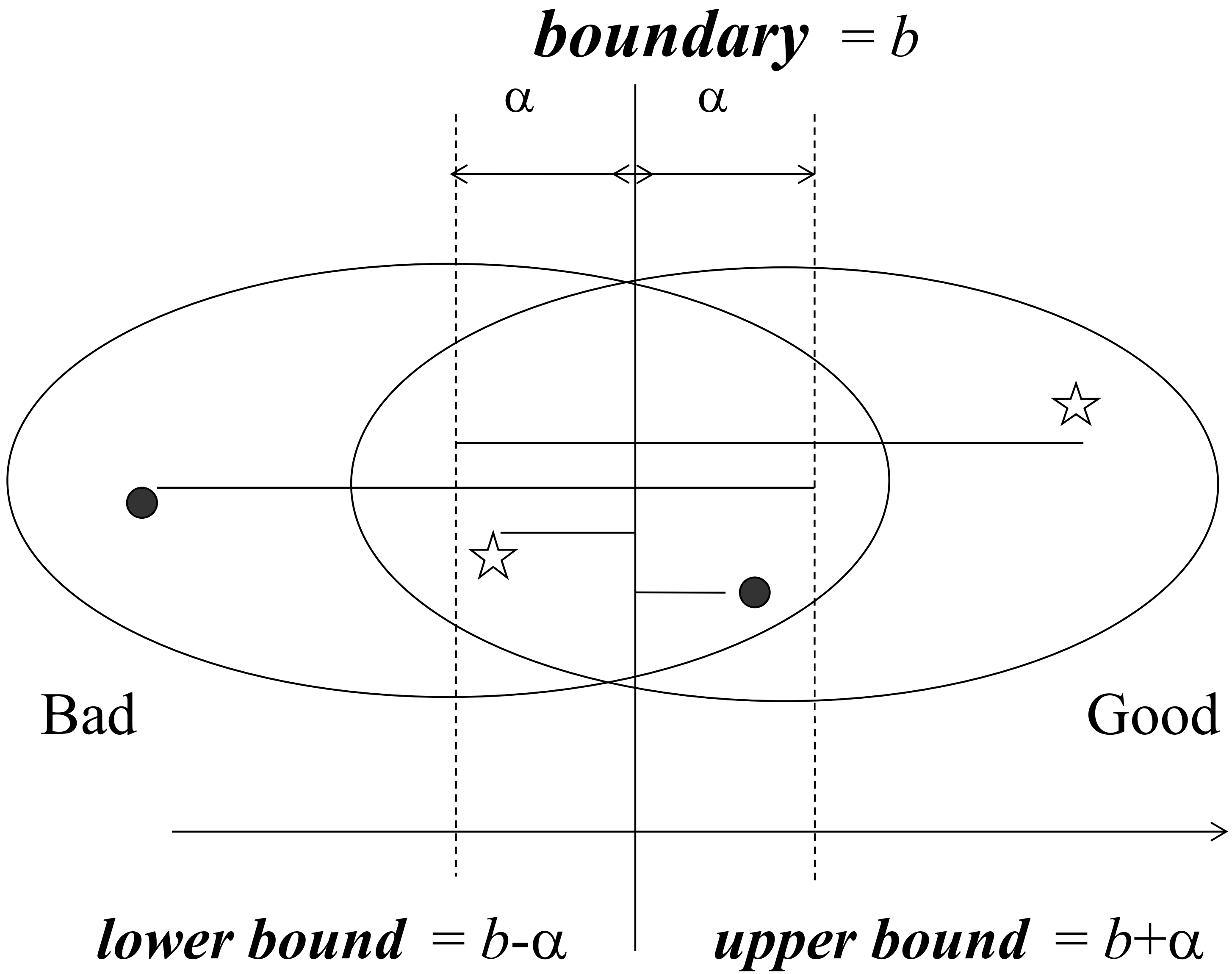
- Logistic regression
- SVM
- MCLP
- MCQP
- Neural Networks
- Decision tree

MCLP-based Classifier



- ***boundary*** = 10 (overlapping = 0)

MCLP-based Classifier



MCLP-based Classifier

- Simple Models (Freed and Glover 1981):
  - Minimize $\Sigma_i h_i \alpha_i$
  - Subject to

$$\boldsymbol{A}_i \boldsymbol{X} \leq b + \alpha_i, \ \boldsymbol{A}_i \in \text{Bad},$$

$$\boldsymbol{A}_i \boldsymbol{X} \geq b - \alpha_i, \ \boldsymbol{A}_i \in \text{Good},$$

where $\boldsymbol{A}_i$ are given, $\boldsymbol{X}$ and b are unrestricted， and $\alpha_i \geq 0$.

MCLP-based Classifier

$$F(x)= 6.8205*x112 - 0.6076*x474 - 7.0563 * x155 + 0.7789 * x492 - 1.2858 * x498 + 0.0004 * x366 + 0.3890 * x505 - 1.3190 * x305 - 0.0702 * x611 + 0.3974 * x312) * 1000000$$

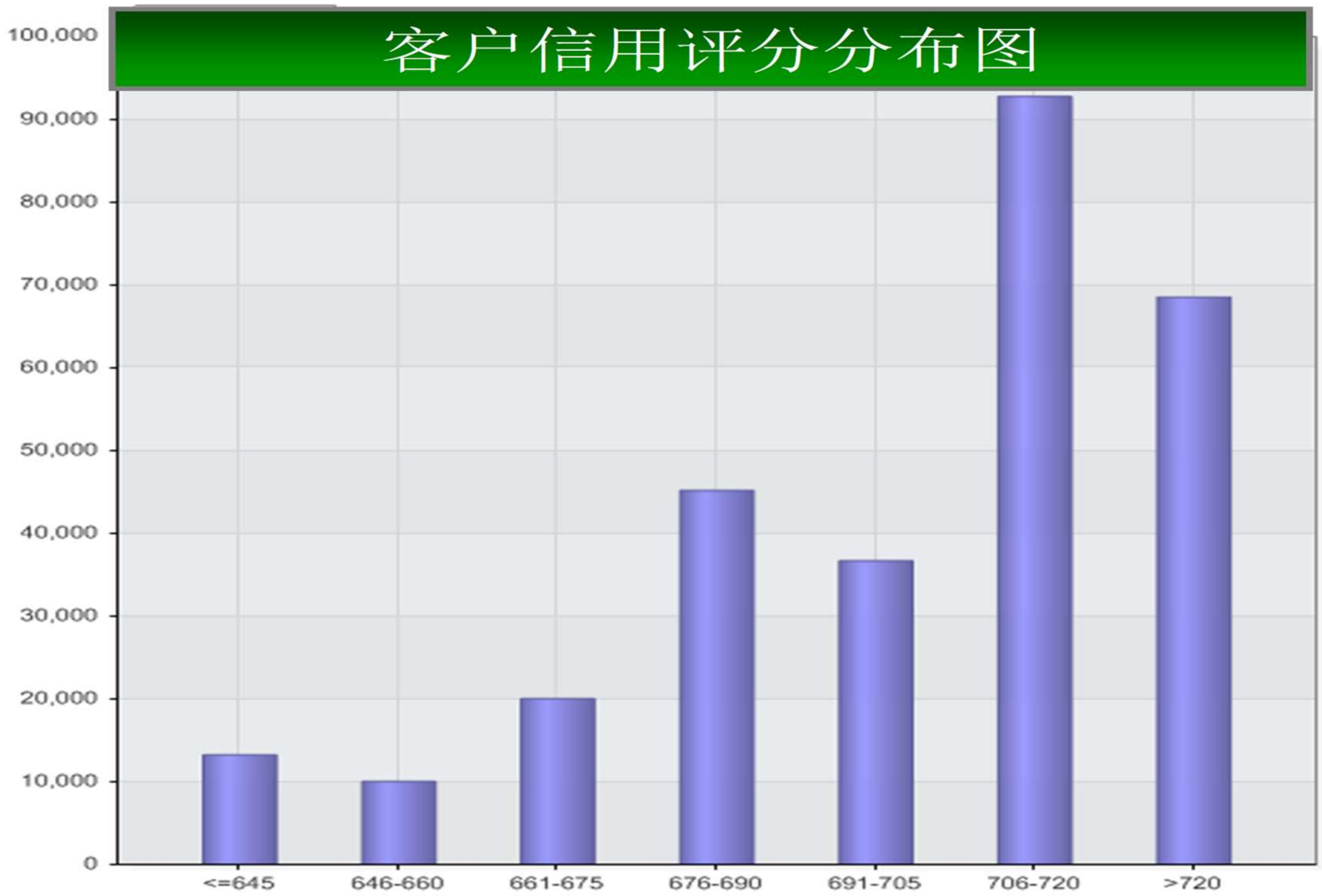
$$p = 1 / (1 + \exp(-(-0.02048780266712+F(x) * (-0.00000033965946))))$$

Score Calculation

- Probability $\longrightarrow$ score
- Linear transformation
- Formula: score=log(odds)*factor + offset
- Score range：300-850
- The odd at 600 is 1:1
- Odds doubles for every 15 points
  - Factor = 15/log(2)
  - Offset = 600

Evaluation

- Score distribution of the population

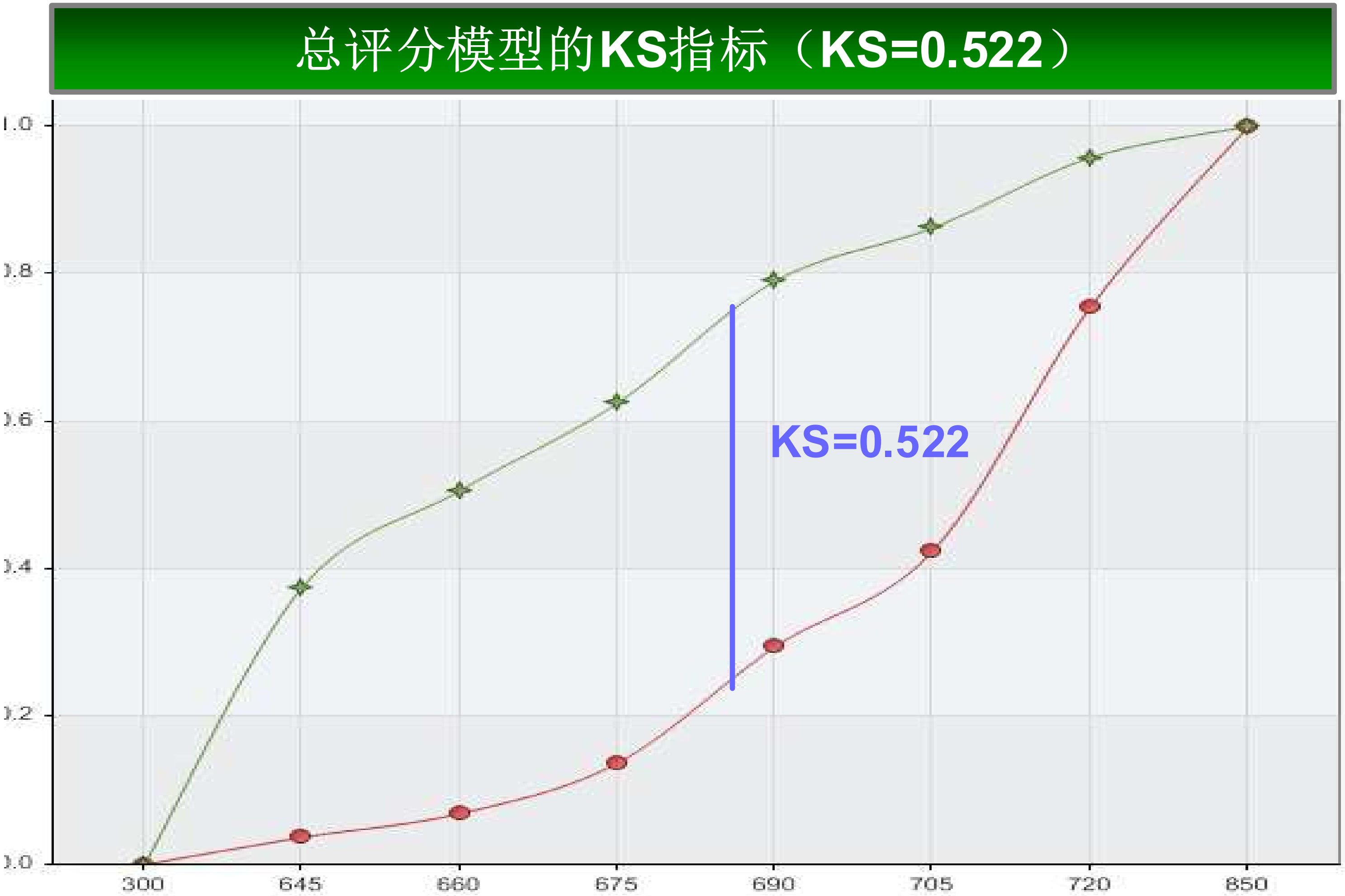


Evaluation

Score range	Accumulative rate of "good"	Accumulative rate of "bad"	Odd (good/bad)
<=645	3.67%	37.54%	3.3628
646-660	6.90%	50.65%	8.4802
661-675	13.75%	62.54%	19.8172
676-690	29.50%	79.03%	32.8525
691-705	42.47%	86.27%	61.6741
706-720	75.51%	95.76%	119.6866
>720	100.00%	100.00%	198.9504

K-S Curve

- K-S index = 0.522

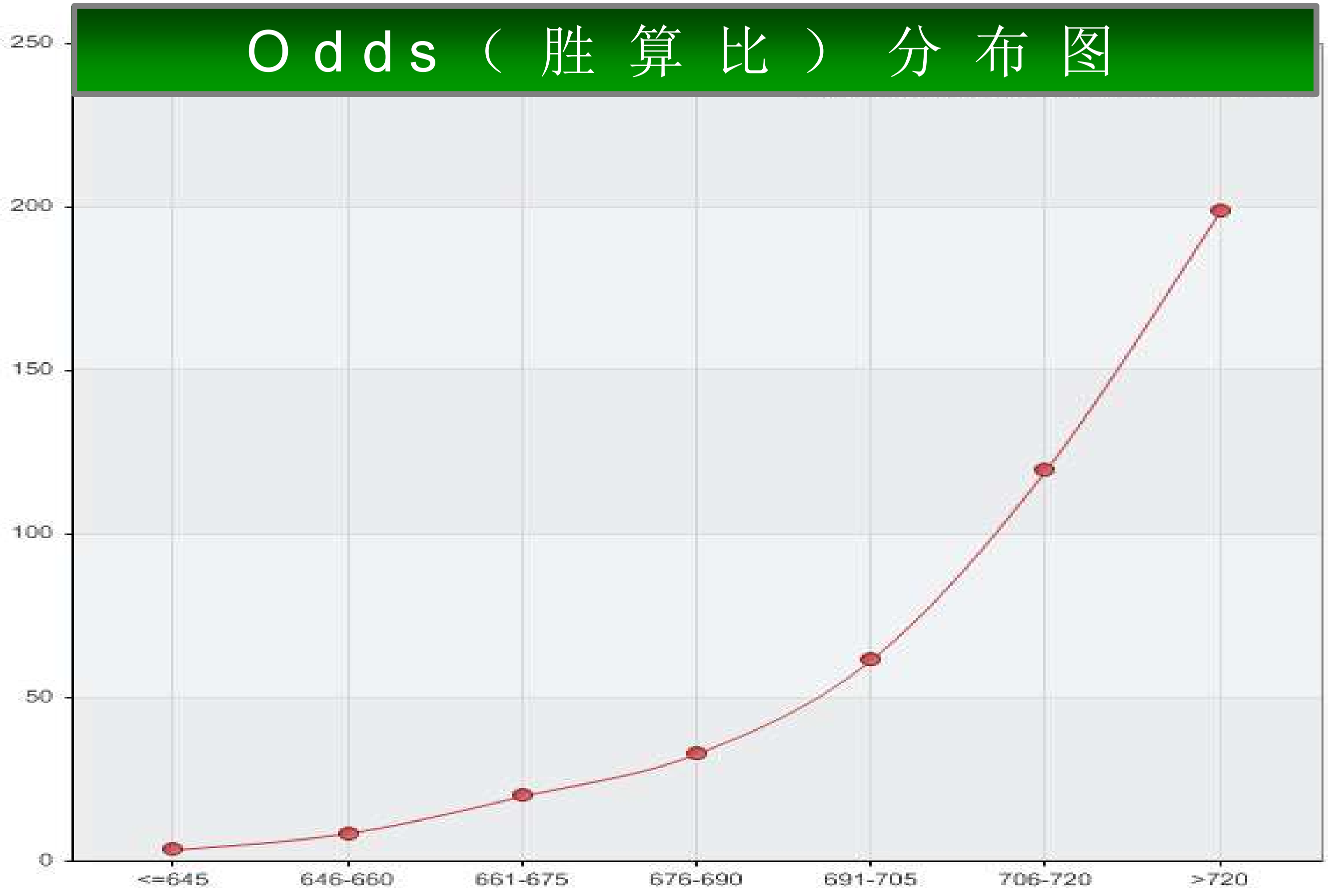


K-S Curve

- For example
  - If we turn down the applications of customers with 682 or less, we will turn down 73% potential “bad” customers, while lose 20.8% “good” customers

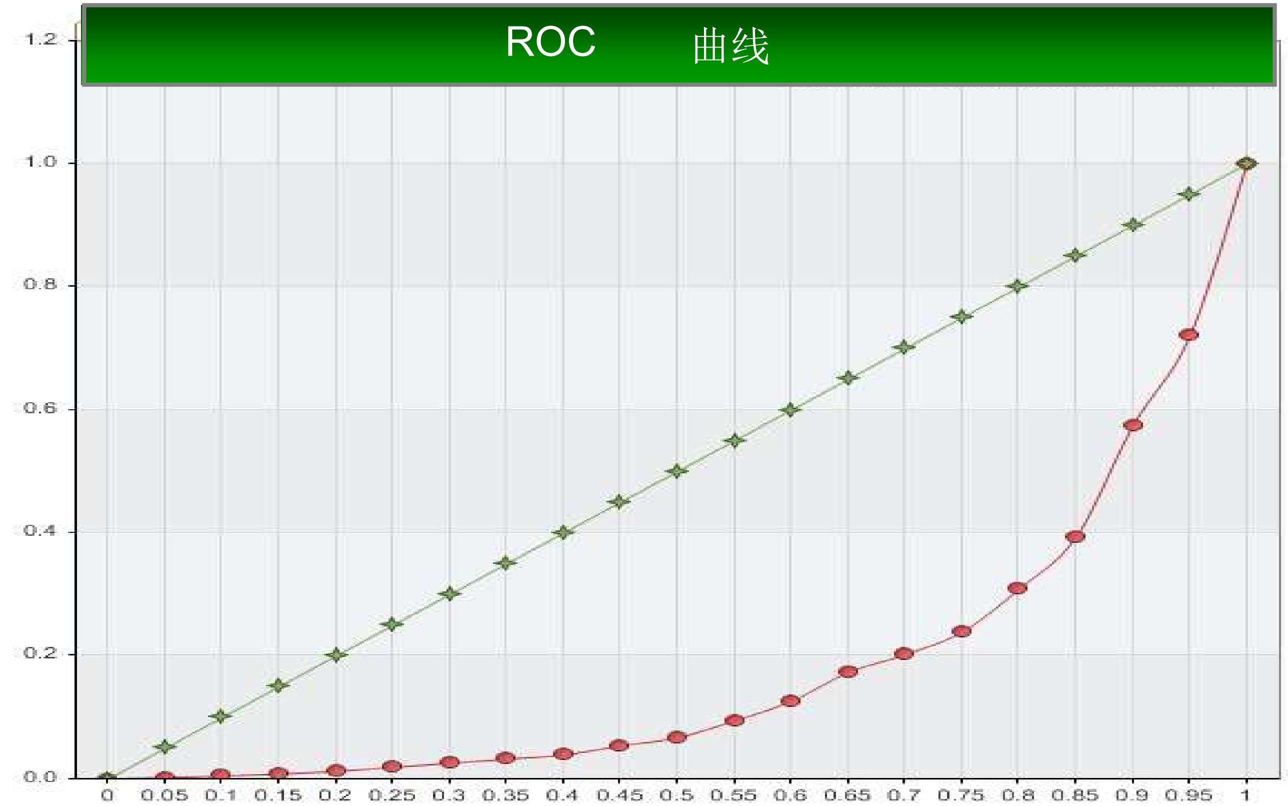
Distribution of Odds

- Odd, # good/ # bad
- Odd increasing linearly with the score



ROC Curve

- Y axis: accumulative rate of “good”
- X axis: accumulative rate of “bad”
- The area under the diagonal denotes the prediction capability



Reason Code

- Give the top 5 factors that result in the score
- Help the bank clerks to explain to the customers
- Help customers to improve their qualifications

Reason Code

- Independent variable, $X = (x_1, x_2, \cdots, x_n)$

  respondent variable, $Y=\{0, 1\}$
- Score model $S = f(x_1, x_2, \cdots, x_n)$
- Mean of every variable $\mu_1, \mu_2, \cdots, \mu_n$
- Average score of the overall population

$$S_\mu = f(\mu_1, \mu_2, \cdots, \mu_n)$$

- Average score on a given variable

$$S_1 = f(x_1, \mu_2, \cdots, \mu_n)$$

$$S_2 = f(\mu_1, x_2, \cdots, \mu_n)$$

$$S_n = f(\mu_1, \mu_2, \cdots, x_n)$$

Reason Code

- Gap between Sμ and S_i

$$C_1 = |S_1 - S_\mu|$$

$$C_2 = |S_2 - S_\mu|$$

$$C_n = |S_n - S_\mu|$$

- Sort C_i in descending order
- Pick the top 5 factors which impact the score most

Reason Code

- Use natural language to express the top *m* factors impacting the credit

Reason Code	Description	Variable ID
A	The number of loan accounts with no delinquency in the past 3 months	Xs
B	The number of delinquency in debit cards in the past 6 months	Xt
…	…	…