# Credit Scoring using Random Forests

Liu Weizhi, Ji Cheng, Tian Rong

School of Management and Engineering, Nanjing University

*weizhiliu2009@gmail.com*

December 26, 2013

# Division of Tasks

- Liu Weizhi: coding
- Ji Cheng: gathering related materials, collecting data
- Tian Rong: designing beamer

# Overview

# Cause of Credit Risk Control

- Credit card brings convenient to consumers as well as huge profit to the bank.
- However, high profits usually accompany with high risk.
- Banks, as the issuers of credit cards, undertake the potential risk.

# The Development of Credit Risk Control

- The judgement is usually based on the experience of risk assessment experts.
- Find the customer with potential risk by statistical means.
- Use efficient data analysis tools and methods.

# The Background of Credit Card

- First credit card appeared in March, 1995.
- The total number of credit cards reached at 1.22 hundred millions while the total number of credit card loans reached at 6931.73 hundred millions.
- Credit card brings convenient to consumers and huge profit to the bank high accompany with high risk.

# The Background of Credit Card

**Event**

Credit card companies which earned huge profit started to lose money at the same time as the high speed development in credit cards

Two main risks banks faced in China:

- Credit Risksł Issue credit cards by the means of “No Guarantee”
- Operational Risk
  - Failure of internal control: Adopt aggressive marketing strategies because of the lack in knowledge of the credit card risk characteristics.
  - Transaction processing risk: Incomplete process and hacker attacks may easily happen.

# Means of Coping with Potential Risk

Advanced technology and means is the fundamental guarantee to the development and security of credit card business.

- Self-built credit card business system and information system in China mostly were still in the stage of beginning, which cause the operational risk of credit cards. To solve these problems, the only method is to use the advanced technology to build a impeccable system.
- For instance, establishing modern authorization exchange network system and fund settlement system is the key to solving the problem of overdraft.

# Meaning of Credit Scoring

Credit scoring is a consuming credit managing technology which is widely used in Europe and the United States.

- Based on Data Mining and Statistical Analysis.
- Building the predictive model.
- Making a comprehensive assessment of consumers' future performance with a credit score .

# Role of credit rating

Credit scoring model can provide credit managers with a large amount of highly predictive information.

- Making effective management strategy
- Realizing high profit with the help of risk control

# History of Credit Scoring

- The methods of diving the overall into several groups based on different characteristics was used by Fisher in1936 for the first time while David Durand firstly adopted this method to assess credit risk in 1941.
- Legislation which was named as “Fair Credit Law” and passed in the United States marked the fully acceptence to credit scores by the society.
- Credit scoring began to be adopted in other financial products by banks.

# Classification of Credit Scoring

- Credit Scoring of Application
  Focused on new applications for credit cards.
- Behavior Scoring
  Assessment in the probability of potential loss to Banks.
- Profit Scoring
  Assessment in the potential profit which coming from card holders to Banks .
- Repayment Scoring
  Forecasting the effect of measures when bad loans appears

# Advantages of Credit Scoring

- Objectivity: Based on huge amount of data.
- Consistency: Credit Scoring Model remain consistent during the process.
- Accuracy: Based on law of large number and statistical technology.
- Comprehensiveness: Credit Scoring Model is consisted of several predictor variables which represent all dimensions of Information .
- Efficiency: Decisions can be made within a few seconds.

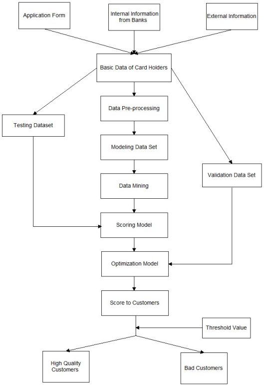
# Market development and customer maintenance

- Customer segmentation model
  Separating customers in accordance with the different research purposes.
- Customers activate model
  Solving the problems caused by sleeping cards.
- Customer leaving model
  Preventing customers from running away.

# Risk Control

- Applying Scoring Model
  Determining the line of credit.
- Behaving Scoring Model
  Assessment in the probability of happening of bad loans .
- Fraud Detecting Model
  Identifying fraudulent trading by analyzing the history of every customer.

# Process of building Credit Scoring Model based on Data Mining



# Our Goal

We don't insist on calculating the specific credit score but to help the lenders to decide whether an application will turn into a bad loan in the future.
Our basic idea is using three machine leaning methods, namely logistics regression, classification tree and random forests, to predict whether borrowers will have a delinquency. The performance of each classifiers is compared using the test data set.

# Data Source

- Bad loans are defined as those loans where repayments are not being made as originally agreed (eg. specific due date).
- Two data sets were collected, one is German Credit, and another is from Kaggle.com.
- German Credit, including 1000 records whose 30% are bad loans, was retrieved from UCI machine learning repository.
- Another data set was retrieved from Kaggle's competition "Give me Some Credit" which has 250,000 records and 6.7% bad loans. (We only used the first 10,000 records due to the lack of computation ability.)

# Data Description

Take German Credit as example. German Credit consists of 21 columns whose first column indicates whether a loan was good or bad. The next 20 features are as follows: checking, duration, history, purpose, amount, savings, employ, installment, status, others, residence, property, age, otherplans, housing, cards, job, liable, tele, foreign. Some features are represented by the specific codes like ‘A34’, ‘A32’, etc.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | good_bad | checking | duration | history | purpose | amount | savings | employ | installme | status | others | residence | property | age |
| 2 | 0 | A11 | 6 | A34 | A43 | 1169 | A65 | A75 | 4 | A93 | A101 | 4 | A121 | 67 |
| 3 | 1 | A12 | 48 | A32 | A43 | 5951 | A61 | A73 | 2 | A92 | A101 | 2 | A121 | 22 |
| 4 | 0 | A14 | 12 | A34 | A46 | 2096 | A61 | A74 | 2 | A93 | A101 | 3 | A121 | 49 |
| 5 | 0 | A11 | 42 | A32 | A42 | 7882 | A61 | A74 | 2 | A93 | A103 | 4 | A122 | 45 |
| 6 | 1 | A11 | 24 | A33 | A40 | 4870 | A61 | A73 | 3 | A93 | A101 | 4 | A124 | 53 |
| 7 | 0 | A14 | 36 | A32 | A46 | 9055 | A65 | A73 | 2 | A93 | A101 | 4 | A124 | 35 |
| 8 | 0 | A14 | 24 | A32 | A42 | 2835 | A63 | A75 | 3 | A93 | A101 | 4 | A122 | 53 |
| 9 | 0 | A12 | 36 | A32 | A41 | 6948 | A61 | A73 | 2 | A93 | A101 | 2 | A123 | 35 |
| 10 | 0 | A14 | 12 | A32 | A43 | 3059 | A64 | A74 | 2 | A91 | A101 | 4 | A121 | 61 |
| 11 | 1 | A12 | 30 | A34 | A40 | 5234 | A61 | A71 | 4 | A94 | A101 | 2 | A123 | 28 |
| 12 | 1 | A12 | 12 | A32 | A40 | 1295 | A61 | A72 | 3 | A92 | A101 | 1 | A123 | 25 |
| 13 | 1 | A11 | 48 | A32 | A49 | 4308 | A61 | A72 | 3 | A92 | A101 | 4 | A122 | 24 |
| 14 | 0 | A12 | 12 | A32 | A43 | 1567 | A61 | A73 | 1 | A92 | A101 | 1 | A123 | 22 |
| 15 | 1 | A11 | 24 | A34 | A40 | 1199 | A61 | A75 | 4 | A93 | A101 | 4 | A123 | 60 |
| 16 | 0 | A11 | 15 | A32 | A40 | 1403 | A61 | A73 | 2 | A92 | A101 | 4 | A123 | 28 |
| 17 | 1 | A11 | 24 | A32 | A43 | 1282 | A62 | A73 | 4 | A92 | A101 | 2 | A123 | 32 |

# Logistic Regression Introduction

- Assuming a linear regression $y = \theta x$, where $x$ is the feature vector and $\theta$ is the corresponding parameters.
- However, the value of $y$ might range widely in the real space, which is not appropriate for the classification problems in which $y$ belongs to a discrete set, like $\{$'Positive Class', 'Negative Class'$\}$.
- Logistic Regression adopts logistic function, which can be applied to depict the probability of some events.

# Logistic Regression Algorithm

- Logistic Function: $h(z) = \frac{1}{1+e^{-z}}$, let $z = \theta x$, then $h(\theta x) = \frac{1}{1+e^{-\theta x}} \in [0, 1]$.
- Cost Function: $cost(\theta) = -ylog(h(\theta x)) - (1-y)log(1-h(\theta x))$. Minimize this function to estimate the parameters $\theta$.

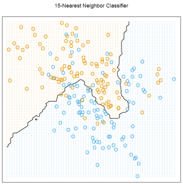



Figure: Decision Boundary

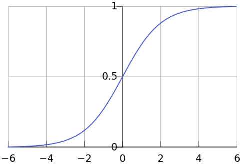



Figure: Logistic Function

# Logistic Regression Results

| | Estimate | Std. Error | z value | $Pr(\geq \|z\|)$ |
|---|---|---|---|---|
| (Intercept) | -5.48743 | 1.387014 | -3.95629 | 7.61E-05 *** |
| checking | 0.484832 | 0.089609 | 5.41055 | 6.28E-08 *** |
| duration | -0.01426 | 0.01144 | -1.24679 | 0.212473 |
| history | 0.469549 | 0.114837 | 4.088823 | 4.34E-05 *** |
| purpose | 0.068907 | 0.042475 | 1.622316 | 0.104736 |
| amount | -0.00015 | 5.07E-05 | -2.99869 | 0.002711 *** |
| savings | 0.25071 | 0.076253 | 3.287871 | 0.001009 *** |
| employed | 0.109647 | 0.095728 | 1.145409 | 0.25204 |
| installp | -0.33189 | 0.106943 | -3.10348 | 0.001913 *** |
| marital | 0.48904 | 0.155785 | 3.139197 | 0.001694 *** |
| coapp | 0.100784 | 0.225263 | 0.447406 | 0.654582 |
| resident | -0.10415 | 0.101927 | -1.02184 | 0.306856 |
| property | -0.14986 | 0.12051 | -1.24353 | 0.213674 |
| age | 0.019083 | 0.011009 | 1.73332 | 0.083039 * |
| other | 0.325128 | 0.140454 | 2.314837 | 0.020622 ** |
| housing | 0.274481 | 0.216319 | 1.268869 | 0.204488 |
| existcr | -0.39916 | 0.203087 | -1.96544 | 0.049364 ** |
| job | 0.34085 | 0.179191 | 1.902159 | 0.057150 * |
| depends | -0.24825 | 0.290759 | -0.8538 | 0.393214 |
| telephon | 0.021655 | 0.244287 | 0.088645 | 0.929364 |
| foreign | 1.699978 | 0.891746 | 1.906348 | 0.056605 * |

# Top 3 Reasons for Denial

| | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | age | other | housing | existcr | job | depends | telephon | foreign | good_bad | score1 | score2 | score3 | topk | |
| 2 | 67 | 3 | 2 | 2 | 3 | 1 | 2 | 1 | good | 0.947141 | 0.785714 | 0.846 | savings;history;age | |
| 3 | 22 | 3 | 2 | 1 | 3 | 1 | 1 | 1 | bad | 0.446593 | 0 | 0.326 | installp;property;existcr | |
| 4 | 53 | 3 | 2 | 1 | 3 | 1 | 1 | 1 | good | 0.931717 | 0.963235 | 0.982 | checking;age;savings | |
| 5 | 61 | 3 | 2 | 1 | 2 | 1 | 1 | 1 | good | 0.906865 | 0.963235 | 0.926 | checking;age;savings | |
| 6 | 28 | 3 | 2 | 2 | 4 | 1 | 1 | 1 | bad | 0.67592 | 0.666667 | 0.674 | history;marital;job | |
| 7 | 24 | 3 | 1 | 1 | 3 | 1 | 1 | 1 | bad | 0.242119 | 0 | 0.272 | purpose;existcr;other | |
| 8 | 28 | 3 | 1 | 1 | 3 | 1 | 1 | 1 | good | 0.398376 | 0.3 | 0.266 | installp;amount;existcr | |
| 9 | 25 | 1 | 2 | 3 | 3 | 1 | 1 | 1 | good | 0.145249 | 0.266667 | 0.466 | savings;purpose;installp | |
| 10 | 44 | 3 | 3 | 1 | 4 | 1 | 2 | 1 | bad | 0.298514 | 0.785714 | 0.276 | job;housing;employed | |
| 11 | 44 | 3 | 1 | 1 | 3 | 2 | 1 | 1 | good | 0.765885 | 0.7 | 0.814 | installp;savings;duration | |
| 12 | 44 | 3 | 2 | 1 | 3 | 1 | 1 | 1 | good | 0.889959 | 0.7 | 0.868 | history;amount;age | |
| 13 | 39 | 3 | 2 | 1 | 2 | 1 | 1 | 1 | good | 0.779685 | 0.963235 | 0.766 | checking;marital;amount | |
| 14 | 63 | 3 | 2 | 2 | 3 | 1 | 2 | 1 | bad | 0.483161 | 0.2 | 0.378 | age;purpose;history | |
| 15 | 30 | 3 | 2 | 2 | 3 | 1 | 2 | 1 | good | 0.558439 | 0.3 | 0.612 | installp;marital;other | |
| 16 | 25 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | bad | 0.45848 | 0 | 0.398 | history;marital;other | |
| 17 | 30 | 1 | 2 | 1 | 4 | 1 | 1 | 1 | good | 0.839057 | 0.782609 | 0.858 | checking;job;savings | |
| 18 | 26 | 3 | 2 | 1 | 3 | 1 | 2 | 1 | good | 0.647998 | 0.7 | 0.672 | amount;savings;resident | |

# Classification Tree Introduction

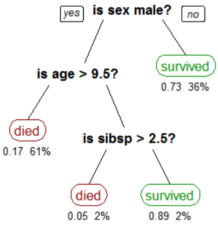



Figure: Titanic Survival Tree

- Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.
- In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.
- The core of classification tree is to split father node to make the children node's subset more pure which can be depicted by **entropy**.

# Classification Tree ID3 Algorithm

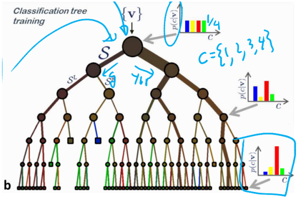
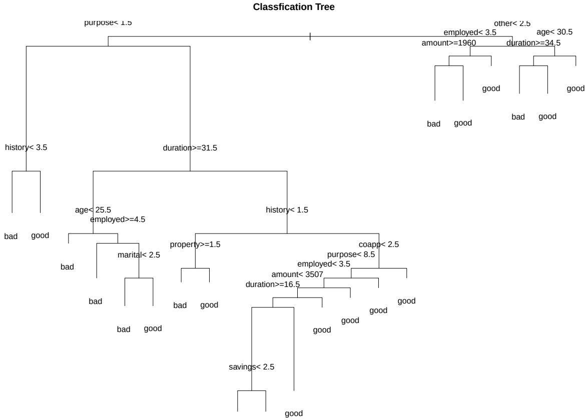



Figure: Classification Tree Entropy

- Entropy: The purity of subset $S$ can be calculated by $H(S) = -\sum_i p(i)log(p(i))$, where $p_i$ is the frequency of class $i$ in $S$. If there is only one class in the set $S$, then the entropy is 0.
- Information Gain: IG is the measure of difference in entropy between the original set $S$ and sets $T$ splited on attribute $\mathscr{Y}$. $IG(\mathscr{Y}) = H(S) - \sum_{i \in T} p(i)log(p(i))$
- Select the splitting attribute which shares the highest information gain.

# Classification Tree Results



# Random Forests Introduction

- Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.
- The most powerful aspect about random forests is variable importance ranking which estimates the predictive value of variables by scrambling the variable and seeing how much the model performance drops.
- Kinect has used the random forests to detect humans' body movement.

# Random Forests - Random Features

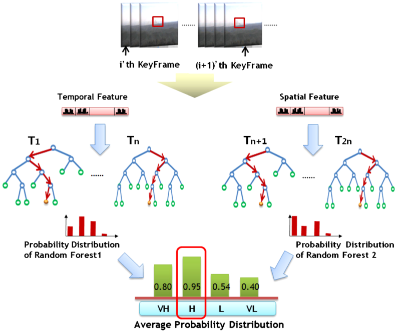
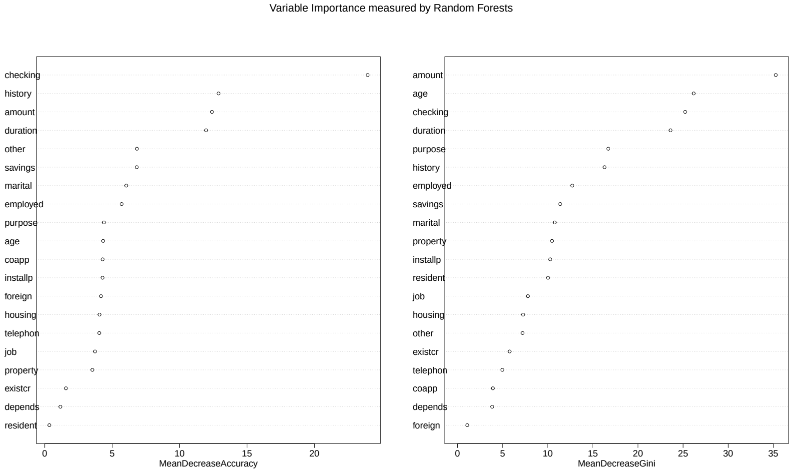



Figure: Random Forests Algorithm

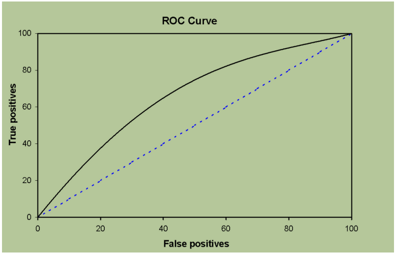
- Random Forests: The forests consist of many trees whose data set is the random bootstrap resampling of original data.
- Random Features: Each tree shares a random subset feature of original feature set.

# Variable Importance



Classifiers are build upon training set and their performance is calculated by the test set. The main performance indicator includes:

- Accuracy: # of correct predictive results divided by the total # of data set.
- K-S statistic: Mainly used in credit industry.
- AUC: Area under an ROC curve.




| | actual | |
|---|---|---|
| predictive | TP | FP |
| | FN | TN |

# Performance Results based on German Credit

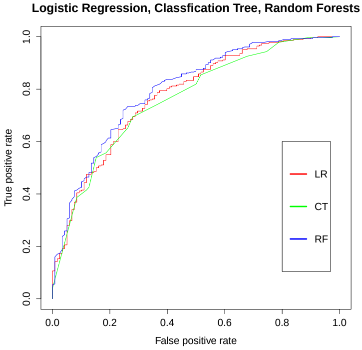



Table: German Credit

| Model | KS | AUC | Accuracy | Cutoff |
|---|---|---|---|---|
| LR | 0.425 | 0.776 | 0.773 | 0.421 |
| CT | 0.415 | 0.762 | 0.753 | 0.250 |
| RF | 0.474 | 0.798 | 0.780 | 0.472 |

# Performance Results based on Kaggle

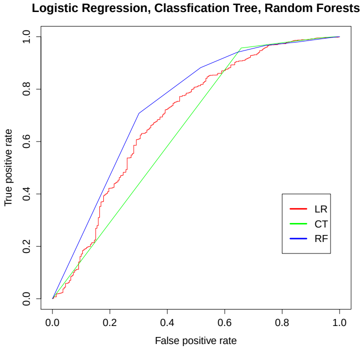



Table: Kaggle

| Model | KS | AUC | Accuracy | Cutoff |
|---|---|---|---|---|
| LR | 0.329 | 0.695 | 0.933 | 0.727 |
| CT | 0.300 | 0.651 | 0.933 | 0.308 |
| RF | 0.407 | 0.741 | 0.932 | 0.300 |

# Conclusions

### #1

In the light of ROC curve, we can find that random forests have the most extraordinary performance followed by logistic regression, and classification tree.

### #2

In comparison with classification tree, random forests illustrates the efficiency and accuracy of ensemble learning.

### #3

Random forests will reach a higher performance with the help of logistic regression.

# Thanks!