Introduction to Credit Risk Modelling

Credit Scoring & Credit Control XVI
Pre-conference workshop

Dr Galina Andreeva, Galina.Andreeva@ed.ac.uk
The workshop is an introduction for complete novices, newcomers to the field of credit scoring or people with limited experience. It covers the basic principles and standard methodology of scorecard development.

1. Concepts, principles and objectives
2. What is needed
3. Stages in credit model (scorecard) development
   - Characteristics transformation
   - Modelling algorithms
     - Logistic regression
   - Measuring predictive accuracy.
4. Where to find more information.
What is credit scoring?

Decision support systems used in consumer credit
Aims at risk assessment of potential borrowers (application scoring) and existing borrowers (behavioural scoring)
Risk/ creditworthiness is usually measured by default probability
Default probability is predicted from observed borrower’s characteristics on the basis of the analysis of known performance of previous customers.
Relationship v transactional lending

40-50 years ago – relationship lending

• Judgmental assessment based on 3(5) Cs (Character, Capacity, Collateral + Capital, Conditions)
• Very slow, subjective, often inconsistent

Now – transactional lending

• Automated impersonal decisions
• Quick and consistent
• Huge volumes of unsecured lending
**Basic concepts: Probability and Odds**

**Frequentist (historic) approach to probability:**
Repeat an experiment a large # of times, \((n\) times). Count the # of times a specific outcome occurs, \((r\) times).

**Probability of an outcome** \(P = \frac{r}{n}\) - proportion of times outcome occurs.

**Odds** that an event will occur (A to B) are given by the ratio of the probability that it will occur (\(P\)) to the probability that it will not occur:
\[
A : B = P / (1 - P) \\
P = \frac{A}{A + B}.
\]

In Credit Scoring context
\(P (G)\) is the probability of being a Good customer (i.e. repaying the money borrowed), or the proportion of Goods in the portfolio or portfolio segment, \(p_G\)
\(P (B)\) is the probability of being a Bad customer, or the proportion of Bads in the portfolio or portfolio segment, \(p_B\)

\(P (G) + P (B) = 1\)

\(O (G) = P (G) : P (B) = p_G : p_B\) is ‘Good to Bad’ odds.
**Example:**

The Royal Bank of Wonderland accepts 8,000 credit applicants - everyone who applies. In a year 7,000 of them turn out to be Good, and 1,000 turn out to be Bad. The average profit from a Good account is £1,000 and the average loss for a Bad account is £10,000.

How many Goods are required to offset the losses from one Bad account? What are the break-even odds? **10:1**

What are the population odds, \( O_{Pop} = \frac{p_G}{p_B} = 7:1 \)
Example continued:

Marital Status:

|            | Good | P(x|G) | Bad | P(x|B) | Marginal Odds |
|------------|------|-------|-----|-------|---------------|
| Married    | 4900 | 0.7   | 400 | 0.4   | 49 : 4        |
| Not married| 2100 | 0.3   | 600 | 0.6   | 21 : 6        |
| Total      | 7000 | 1     | 1000| 1     |               |

Marginal Odds of Married = $0.7 : 0.4 \times 7 : 1 = 12.25$

Marginal Odds of NM = $0.3 : 0.6 \times 7 : 1 = 3.5$

*Information Odds*

NB: Marginal Odds = Information Odds $\times$ Population Odds
Let $X = (X_1, X_2, \ldots, X_m)$ be characteristics (variables) of the borrower such as age, marital status, etc.

$x = (x_1, x_2, \ldots, x_m)$ be outcomes/ attributes of characteristics.

$P(G)$ and $P(B)$ are prior probabilities.

Posterior probabilities:

$P(G|x)$ is the probability of being Good given certain attributes

$P(B|x)$ is the probability of being a Bad customer given certain attributes
\( P(x|G) \) and \( P(x|B) \) are likelihoods that describe how likely the attributes \( x \) are in good/bad populations

\[
P (\text{Married} \mid \text{Good}) = \frac{4900}{7000} = 0.7
\]

\[
P (\text{Married} \mid \text{Bad}) = \frac{400}{1000} = 0.4
\]

\( I(x) = \frac{P(x|G)}{P(x|B)} \) are information odds

Using Bayes’ rule

\[
O(G \mid x) = \frac{P(G \mid x)}{P(B \mid x)} = \frac{P(x \mid G) \times P(G)}{P(x \mid B) \times P(B) / P(x)} = I(x) \times O_{\text{Pop}}
\]
Example continued:

**Marital Status:**

|       | Good  | $P(x|G)$ | Bad  | $P(x|B)$ | Marginal Odds, $O(G|x)$ |
|-------|-------|----------|------|----------|-------------------------|
| Married | 4900  | 0.7      | 400  | 0.4      | 49 : 4 12.25:1          |
| Not married | 2100  | 0.3      | 600  | 0.6      | 21 : 6 3.5:1            |

**Time in Employment:**

| Time  | Count | $x$ | $P(x)$ | $y$ | $P(y)$ | Marginal Odds, $O(x|y)$ |
|-------|-------|-----|--------|-----|--------|-------------------------|
| 0     | 1050  | 0.15| 500    | 0.5 |        | 105 : 50 2.1:1          |
| up to 6 m | 1680  | 0.24| 250    | 0.25|        | 168 : 25 6.72:1         |
| 6m - 3y | 1960  | 0.28| 140    | 0.14|        | 196 : 14 14:1           |
| 3y+   | 2310  | 0.33| 110    | 0.11|        | 231 : 11 21:1           |
| Total | 7000  |     | 1000   |     |        |                         |
Assuming Independence

Assume that 2 characteristics are independent. Then can use multiplication rule: \( P(A \cap B) = P(A) \times P(B) \)

\[
O(G \mid x_1, x_2) = \frac{P(G \mid x_1, x_2)}{P(B \mid x_1, x_2)} = \frac{p_G P(x_1, x_2 \mid G)}{p_B P(x_1, x_2 \mid B)} =
\]

\[
= \frac{p_G P(x_1 \mid G) P(x_2 \mid G)}{p_B P(x_1 \mid B) P(x_2 \mid B)} = O_{pop} \times I(x_1) \times I(x_2)
\]

This result generalizes to \( n \) independent characteristics

Odds of \( n \) independent chars =

\[\text{Pop Odds} \times \text{Info Odds (Char } 1) \times \ldots \times \text{Info Odds (Char } n)\]
Assuming Independence

Assume that Marital Status and Time in Employment are independent.

Odds of Married and No Job = \( \frac{7}{1} \times \frac{0.7}{0.4} \times \frac{0.15}{0.5} = \frac{7 \times 1.75 \times 0.3}{0.3} = 3.675 \)

Odds of Not Married and 3+ years of employment = ？

\( = \frac{7}{1} \times \frac{0.3}{0.6} \times \frac{0.33}{0.11} = 7 \times 0.5 \times 3 = 10.5 \)
Odds (Married & No Job) = $7 \times 1.75 \times 0.3 = 3.675$

Log Odds Score (M & NJ) = $\ln (3.675) = \ln (7) + \ln (1.75) + \ln (0.3) = 1.3$

**Naïve Bayes’ scorecard**

<table>
<thead>
<tr>
<th></th>
<th>REJECT</th>
<th>ACCEPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone</td>
<td>7:1</td>
<td></td>
</tr>
<tr>
<td>Not married</td>
<td>3.5:1</td>
<td>12.25:1</td>
</tr>
<tr>
<td>Married</td>
<td></td>
<td>36.75:1</td>
</tr>
<tr>
<td>Married, No job</td>
<td>3.675:1</td>
<td>1.05:1</td>
</tr>
<tr>
<td>Married, No job, 0 to 3 years of employment</td>
<td>3.36:1</td>
<td>11.76:1</td>
</tr>
<tr>
<td>Married, No job, 1m+ of employment</td>
<td>7:1</td>
<td>24.5:1</td>
</tr>
<tr>
<td>Married, Not married, 3+ years</td>
<td>10.5:1</td>
<td></td>
</tr>
<tr>
<td>Married, Not married, 1m+ of employment</td>
<td>21:1</td>
<td></td>
</tr>
<tr>
<td>Married, Not married, 3+ years</td>
<td>10.5:1</td>
<td></td>
</tr>
</tbody>
</table>

$S(x)$

$W(x) –$ weights of evidence
The Basic Idea

Need the data on characteristics and performance of previous borrowers
Assign points to each category of borrowers’ characteristic so that they reflect the level of risk of this category
For a new borrower, sum all the points over the relevant categories of all characteristics → get an overall score – a summary of this borrower’s creditworthiness.
### Example of a scoring table (part)

<table>
<thead>
<tr>
<th>Time at current address</th>
<th>Less than 6 months</th>
<th>6m – 2 years</th>
<th>2 – 6 years</th>
<th>6 - 10 years</th>
<th>10 + years</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>13</td>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residential Status</th>
<th>Owner</th>
<th>Tenant</th>
<th>With parents</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Banking</th>
<th>Current account</th>
<th>Saving account</th>
<th>Current and saving</th>
<th>No account</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>10</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Retired</th>
<th>Full-time</th>
<th>Part-time</th>
<th>Self-employed</th>
<th>Student</th>
<th>Other</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>21</td>
<td>16</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>18-25</th>
<th>26-31</th>
<th>32-40</th>
<th>41-54</th>
<th>55+</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

Adapted from Lewis M. (1992) *An Introduction to Credit Scoring*, FICO: San Rafael, CA
### Risk Rank Ordering

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Status</th>
<th>Employment Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>36.75:1</td>
<td>Married</td>
<td>3+ years of employment</td>
</tr>
<tr>
<td>24.5:1</td>
<td>Married</td>
<td>6m – 3y of employment</td>
</tr>
<tr>
<td>11.76:1</td>
<td>Married</td>
<td>up to 6m of employment</td>
</tr>
<tr>
<td>10.5:1</td>
<td>Not married</td>
<td>3+ years of employment</td>
</tr>
<tr>
<td>7:1</td>
<td>Not married</td>
<td>6m – 3y of employment</td>
</tr>
<tr>
<td>3.675:1</td>
<td>Married</td>
<td>No job</td>
</tr>
<tr>
<td>3.36:1</td>
<td>Not married</td>
<td>up to 6m of employment</td>
</tr>
<tr>
<td>1.05:1</td>
<td>Not married</td>
<td>No job</td>
</tr>
</tbody>
</table>
Strategy curve

![Graph showing the relationship between bad rate % and accept rate %](image-url)
Assumptions and general principles

• “Future is going to be like the past”
• Scoring models are based on associations / correlations, they are not causal models
  • It is predictive not explanatory.
  • Anything that helps predict can be used (but there is still a desire to understand risk drivers).
• It has to interact easily with organisations information system.
  • So variables that can be used must be easily and cheaply obtained and automatically updated.

Birds flying low, expect rain and a blow
## Credit risk approaches

<table>
<thead>
<tr>
<th>Lending to individuals</th>
<th>Lending to businesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively small amounts of money lent to a large number of customers</td>
<td>Large amounts of money lent to a relatively small number of businesses</td>
</tr>
<tr>
<td>- focus more on prediction, less on causality</td>
<td>- focus more on causality, less on prediction</td>
</tr>
<tr>
<td>- Management Science and Data Mining</td>
<td>- Finance and Accounting</td>
</tr>
<tr>
<td>- limited research and a few textbooks</td>
<td>- extensive research, loads of textbooks</td>
</tr>
</tbody>
</table>
What is scoring used for

New customers (application scoring):
• who is likely to default?
• who is likely to respond?
• who is likely to use the product?
• who is likely to switch to a competitor?

Existing customers (behavioural):
• shall we increase a credit limit?
• can we cross-sell other products?

Problem customers (collections):
• shall we sell the debt off?
• if not, what method of contacting to be used?

Newer Uses:
• Risk-based pricing
• Securitisation
• Profit Scoring
• Customer Scoring
• Capital Adequacy (Basel New Accord)
• IFRS9
What is needed

Need a sample of past customers- their application forms and subsequent credit history.
Managers need to define which history is "good" and which is "bad".
Sample should have

- Enough goods and bads (at least 1500 goods and 1500 bads - but often 20,000-50,000 goods and all bads they can find)
- be a random sample from application population
- long enough credit histories to decide between good and bad
- as current as is possible.
What is needed for a scoring model (1)

Acceptance/ Sample period

6 -12 months

Outcome/ Performance period

9 – 24 months

• Acceptance period should compensate for seasonal fluctuations

• The length of outcome period depends on the type of credit
• Definition of a Good/Bad client

Usually a Bad account is an account that misses 3 consecutive payments within the observation period, otherwise – Good, but …

May have a third category – Indeterminate.

Basel II definition of Default for retail credit: 90 days in arrears.
Characteristics of a client that come from

- Application form (age, marital status, time at address, etc.)
- Credit bureau (previous credit history)
- Internal bank records (in case of existing customers)
- Some new sources
  - open banking/ financial transactions
  - social media, web browsing, mobile usage, text and messages, geo-location
  - rental and utilities data
  - psychometrics.

(NB: ethical issues!)
Process of scorecard development

Data preparation → Sample selection → Characteristics analysis → Fitting Good/Bad model

Cut-off selection → Evaluation of model performance → Reject inference
Characteristics analysis: coarse - classification

Categorical variables often have too many small categories (attributes), need to combine them for a robust model.

NB: Missing (No answer) can be modelled as a separate category.

<table>
<thead>
<tr>
<th>Residential Status</th>
<th>Goods</th>
<th>Bads</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owner</td>
<td>6000</td>
<td>300</td>
<td>20 : 1</td>
</tr>
<tr>
<td>Rent unfurnished</td>
<td>1600</td>
<td>400</td>
<td>4 : 1</td>
</tr>
<tr>
<td>Rent furnished</td>
<td>350</td>
<td>140</td>
<td>2.5:1</td>
</tr>
<tr>
<td>With parents</td>
<td>950</td>
<td>100</td>
<td>9.5 : 1</td>
</tr>
<tr>
<td>Other</td>
<td>90</td>
<td>50</td>
<td>1.8 : 1</td>
</tr>
<tr>
<td>No answer</td>
<td>10</td>
<td>10</td>
<td>1 : 1</td>
</tr>
</tbody>
</table>

Example (from Thomas et al.(2002), p.132)
Coarse – classification of numeric characteristics

For numeric characteristics coarse-classification helps to preserve non-linear and non-monotonic patterns

Good Rate by Applicant’s Age
Split a characteristic into 10-20 approximately equal groups (fine classes)
Compute Good:Bad Odds or (Good Rate or Bad Rate) for each fine class
Band the adjacent fine classes together into coarse classes

Coarse – classification of numeric characteristics

Good rate by Applicant’s Age

Good rate by Spouse’s Age
Before modelling Good/Bad: variable transformation

After coarse-classifying characteristics, they need to be converted into a form suitable for further analysis.

Two main approaches:

1. Each coarse-class of a characteristic is converted into a binary dummy variable (taking 0/1 values). That leads to \( k-1 \) dummy variables, where \( k \) is the number of coarse classes for a characteristic

   - leads to a large number of variables, but does not impose any dependencies between characteristics and Good:Bad odds
Before modelling Good/Bad: variable transformation

2. Weights of evidence (WOE) transformation is applied to each coarse-class of a characteristic $X$

$$W(x) = \ln(I(x)) = \ln\left(\frac{P(x | G)}{P(x | B)}\right) = \ln\left(\frac{P(G | x)/P(B | x)}{P(G)/P(B)}\right)$$

- cuts down the number of variables but imposes dependency with Good:Bad odds and does not allow for effects of other characteristics.
Classification methods

**Statistical**
- Discriminant analysis (DA or LDA),
- Linear regression (LR),
- Logistic regression (LogR),
- Nearest-neighbours approach (k-NN),
- Classification trees.

**Machine-Learning**
- Support Vector Machines (SVM),
- Neural networks (NN),
- Genetic algorithms,
- Random Forests,
the list is not exhaustive…
Logistic Regression

- Most commonly used method

- Produces log-odds score which is a weighted sum of attribute values

\[ S(x) = \ln \left( \frac{P(G | x)}{P(B | x)} \right) = \ln \left( \frac{p(x)}{1 - p(x)} \right) = \mathbf{c} \cdot \mathbf{x} = c_0 + c_1 x_1 + \cdots + c_m x_m \]

- The observed dependent variable is binary (0/1)
- The expected value of Y is the probability of Y=1

\[ E(Y) = p(y = 1 \mid x_1, x_2, \ldots x_m) \]
Continuous v binary dependent variable
Logit function

Probability of being Good

Log-Odds Score, S(x)
Logistic predictions
Obtaining probabilities

\[
\ln\left(\frac{p(x)}{1 - p(x)}\right) = c \cdot x = c_0 + c_1 x_1 + \ldots + c_m x_m
\]

\[
p(x) = e^{c \cdot x} (1 - p(x)) = e^{c \cdot x} - p(x)e^{c \cdot x}
\]

\[
p(x) + p(x)e^{c \cdot x} = e^{c \cdot x}
\]

\[
p(x)(1 + e^{c \cdot x}) = e^{c \cdot x}
\]

\[
p(x) = \frac{e^{c \cdot x}}{1 + e^{c \cdot x}} = \frac{1}{1 + e^{-c \cdot x}}
\]
Contingency table revisited

<table>
<thead>
<tr>
<th>FLAG BINARY</th>
<th>phone1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Bad</td>
<td>102</td>
<td>645</td>
</tr>
<tr>
<td>Good</td>
<td>265</td>
<td>3988</td>
</tr>
<tr>
<td>Total</td>
<td>367</td>
<td>4633</td>
</tr>
</tbody>
</table>

\[
\ln \left( \frac{p(x)}{1 - p(x)} \right) = c_0 + c_1 x_1
\]

\(x_1\): Phone1 = 1 if home phone given; 0 otherwise
Marginal (Posterior) Odds of being Good if phone1=1
\[ p(G| x_1 =1) : p(B| x_1 =1) = \]

Marginal (Posterior) Odds of being Good if \( x_1 =0 \)
\[ p(G| x_1 = 0) : p(B| x_1 =0) = \]

Relative Odds (Odds Ratio) = \[ \frac{p(G| x_1 =1) : p(B| x_1 =1)}{p(G| x_1 = 0) : p(B| x_1 =0)} \]
Fitting Logistic Regression

\[ s(x) = c_0 \quad \text{- baseline model} \]

\[
 s(x) = \ln \left( \frac{p(G \mid x)}{p(B \mid x)} \right) = \ln \left( \frac{p(x)}{1 - p(x)} \right) = c \cdot x = c_0 + c_1 x_1
\]
Interpreting results

From SPSS:

\[ s(x) = c_0 + c_1 x_1 = 0.955 + 0.867 x_1 \]

\[ \exp(c_0) = e^{0.955} \approx 2.6 \]

\[ \exp(c_1) = e^{0.867} \approx 2.4 \]
**Variables in the Equation**

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone1</td>
<td>.942</td>
<td>.129</td>
<td>53.516</td>
<td>1</td>
<td>.000</td>
<td>2.564</td>
</tr>
<tr>
<td>AGE</td>
<td>.054</td>
<td>.004</td>
<td>173.924</td>
<td>1</td>
<td>.000</td>
<td>1.056</td>
</tr>
<tr>
<td>Constant</td>
<td>-.948</td>
<td>.182</td>
<td>27.113</td>
<td>1</td>
<td>.000</td>
<td>.387</td>
</tr>
</tbody>
</table>

a. Variable(s) entered on step 1: phone1, AGE.
Interpreting coefficients (1)

• Difficult to interpret effects on the probability, can only be evaluated for a given value, therefore interpretation involves log-odds and odds

• A one-unit change in the independent variable produces the change in the predicted log-odds given by estimated coefficient

\[
\ln \left( \frac{p}{1-p} \right) = -0.948 + 0.942 \times \text{Phone1} + 0.054 \times \text{Age}
\]

where \( p \) is the probability of being Good
Interpreting coefficients (2)

• Exponentiating both sides provides interpretation for Odds

Odds of being Good =

\[ e^{-0.948 + 0.942 \times \text{Phone1} + 0.054 \times \text{Age}} \]

\[ = e^{-0.948} \times e^{0.942 \times \text{Phone1}} \times e^{0.054 \times \text{Age}} \]

• Change in Odds due to a one-unit change in the independent variable

When Phone1 = 1 \[ \Rightarrow e^{0.942} = 2.564\text{-times increase in Odds of being Good} \]

When Age increases by 1 year \[ \Rightarrow e^{0.054} = 1.056\text{-times increase in Odds of being Good} \]
Interpreting coefficients - WOE

- With WOE coding it is easier to use log-linear equation
- should be interpreted keeping in mind WOE values of each class

Log-Odds of being Good =
= 1.741 + 1.093 PhoneWOE + 1.018 AgeWOE

<table>
<thead>
<tr>
<th>WOE values</th>
<th>PhoneWOE</th>
<th>AgeWOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.79</td>
<td>-0.93</td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td>-0.69</td>
<td></td>
</tr>
<tr>
<td>-0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predictive accuracy and hold-out sample

• If predictive accuracy is measured on the development/training dataset, the results are going to be biased
• Hence, the need for an independent sample (hold-out or test)
• When drawing a random sample from the population, the standard practice is to split it into 2 parts (70% and 30%). Use 70% as a development/ training sample, and reserve the remaining one as a hold-out for model performance measurement.
• Additional validation is required on ‘out-of-time’ sample, before the launch of the scorecard
• After the scorecard is implemented, regular monitoring is needed.
Measures of predictive accuracy

- Kolmogorov-Smirnov (K-S) statistic
- Area under the ROC curve (AUROC)
  - Gini coefficient
- Error rate
- Bad rate.
Kolmogorov-Smirnov (K-S) statistic

\[ KS = \max_s |F_b(s) - F_g(s)|, \]

where \( F_b(s) \) and \( F_g(s) \) are the cumulative distributions of Bads and Goods with and below a score \( s \) respectively.
ROC curve – Thomas et al. (2002)

or Lorenz diagram
ROC curve - alternative representation

\[ F_b(s) \]

\[ F_g(s) \]

% Rejected Goods

% Rejected Bads

CRC Credit Research Centre
Gini coefficient (G)

\[ G = 2 \times (\text{AUROC} - 0.5) \]

\[ G = 2\left[ \sum_{i=1}^{L} (F_g(s_i) - F_g(s_{i-1}))(F_b(s_i) + F_b(s_{i-1}))/2 - \frac{1}{2} \right] = \]

\[ \sum_{i=1}^{L} (F_g(s_i) - F_g(s_{i-1}))(F_b(s_i) + F_b(s_{i-1})) - 1 \]
Measures with cut-off

- K-S statistic and Gini coefficient do not require any cut-off, they provide a summary of predictive accuracy across the whole range of possible cut-offs.

- It is also necessary to measure the predictive accuracy of a model at the chosen cut-off level.

- This is achieved by measures based on ‘Confusion matrix’
  - Error Rate
  - Bad Rate among Accepts
### General confusion matrix / Classification Table

<table>
<thead>
<tr>
<th>True/Observed Class</th>
<th>( G )</th>
<th>( B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G )</td>
<td>( a )</td>
<td>( b )</td>
</tr>
<tr>
<td>( B )</td>
<td>( c )</td>
<td>( d )</td>
</tr>
</tbody>
</table>

- \( a \) – true Goods predicted as Goods
- \( b \) – true Goods predicted as Bads
- \( c \) – true Bads predicted as Goods
- \( d \) – true Bads predicted as Bads
Error Rate and Bad Rate among Accepts

There are two types of error one can make in credit scoring:

1. classify an application as Good and therefore accept it for credit, and it turns out to be Bad → direct loss
2. classify an application as Bad and therefore reject it, whereas it is Good → lost opportunity

Error Rate (ER) measures the proportion of both errors in a sample:

$$ER = \frac{b + c}{a + b + c + d}$$

Normally the first type of error is perceived to be more serious, so Bad Rate among Accepts (BR) focuses on this particular error:

$$BR = \frac{c}{a + c}$$
Predictive accuracy of different methods, % of correctly classified

<table>
<thead>
<tr>
<th>Authors</th>
<th>Linear Reg / Discriminant Analysis</th>
<th>Logistic Reg-n</th>
<th>Classification trees</th>
<th>Linear Prog</th>
<th>Neural Nets</th>
<th>Nearest Neighbours</th>
<th>Support Vector Machines</th>
<th>Genetic Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henley (1995)</td>
<td>43.4</td>
<td>43.3</td>
<td>43.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Boyle et al. (1992)</td>
<td>77.5</td>
<td>-</td>
<td>75</td>
<td>74.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Srinivasan and Kim (1987)</td>
<td>87.5</td>
<td>89.3</td>
<td>93.2</td>
<td>86.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yobas et al. (1997)</td>
<td>68.4</td>
<td>-</td>
<td>62.3</td>
<td>-</td>
<td>62.4</td>
<td>-</td>
<td>-</td>
<td>64.5</td>
</tr>
<tr>
<td>Desai et al. (1997)</td>
<td>66.5</td>
<td>67.3</td>
<td>67.3</td>
<td>-</td>
<td>64.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Baesens (2003)</td>
<td>74.4</td>
<td>74.4</td>
<td>74.8</td>
<td>74.8</td>
<td>75.0</td>
<td>74.8</td>
<td>74.8</td>
<td>-</td>
</tr>
</tbody>
</table>
## More Results

<table>
<thead>
<tr>
<th>Classification algorithm</th>
<th>Hold-out AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogR</td>
<td>0.779</td>
</tr>
<tr>
<td>LogR with interaction variables</td>
<td>0.777</td>
</tr>
<tr>
<td>SVM: linear</td>
<td>0.783</td>
</tr>
<tr>
<td>SVM: polynomial</td>
<td>0.755</td>
</tr>
<tr>
<td>SVM: Gaussian RBF</td>
<td>0.783</td>
</tr>
<tr>
<td>LDA</td>
<td>0.781</td>
</tr>
<tr>
<td>kNN</td>
<td>0.756</td>
</tr>
</tbody>
</table>

**Support vector machines for credit scoring and discovery of significant features**

Tony Bellotti and Jonathan Crook

Books on credit scoring


References


