Credit Risk Modelling /Scorecard

April (24-26) 2012

Presented by
Dr Mohammad Yousaf Shad
Mian Khaliq Ur Rehman
Credit scoring and the business

• What is credit scoring?
• Where is credit scoring used?
• Why is credit scoring used?
• How has credit scoring affected credit provision?
What is credit scoring?

- Credit scoring is a method to assess the risk
- Assign scores to the characteristics of debt and borrows
- Relationship lending vs Transactional lending
- Relationship lending to smaller lenders
- Transactional lending is favoured by larger lenders dealing in high-volume
Where is credit scoring used?

- Effective decision making
  - Accept/reject
  - Maximum loan value or repayment
  - Pricing/interest rate
  - Loan term (Loan duration and repayment structure, collateral)
How has credit scoring affected credit provision?

• Automate credit risk assessment
  – Data
    • Increasing amounts of information, resulting from automation, data sharing, and empowering data privacy legislation
  – Risk assessment
    • Use of credit scoring to drive transactional lending, as opposed to the relationship lending of old
  – Decision making
    • Lenders are no longer limited to the accept/reject decision, but can also use scores to set prices, and value portfolios for securitisation
  – Process automation
    • Evolving technology that has made computers—including processors, data storage, and networks—faster, smaller, and cheaper
  – Legislation
    • Fair-lending legislation and Basel II have promoted the use of credit scoring, while data privacy legislation and practices allow the sharing of data
How has credit scoring affected credit provision?

<table>
<thead>
<tr>
<th>Decision component</th>
<th>Change area</th>
<th>Credit growth drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Practices</td>
<td>Automation</td>
</tr>
<tr>
<td>Data</td>
<td>Sharing</td>
<td>Collection</td>
</tr>
<tr>
<td>Risk assessment</td>
<td>Credit scoring</td>
<td>Calculation</td>
</tr>
<tr>
<td>Decision-making</td>
<td>Risk-based pricing</td>
<td>Decision agents</td>
</tr>
<tr>
<td>Delivery</td>
<td>Securitisation</td>
<td>ATM, Internet</td>
</tr>
<tr>
<td></td>
<td>Cross-sales</td>
<td></td>
</tr>
</tbody>
</table>
## History of credit scoring

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fair Isaac (FI)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FI</td>
<td>1956</td>
<td>Founded San Francisco CA, by Bill Fair and Earl Isaac</td>
</tr>
<tr>
<td></td>
<td>1958</td>
<td>First scorecard development, for American Investments</td>
</tr>
<tr>
<td></td>
<td>1984</td>
<td>Develops first bureau score for pre-screening</td>
</tr>
<tr>
<td></td>
<td>1995</td>
<td>First use of scoring by mortgage securitisers</td>
</tr>
<tr>
<td><strong>Experian-Scorex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management Decision</td>
<td>1974</td>
<td>Founded by John Coffman and Gary Chandler</td>
</tr>
<tr>
<td>Systems (MDS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1982</td>
<td>MDS purchased by CCN</td>
</tr>
<tr>
<td>Scorex</td>
<td>1984</td>
<td>Founded in Monaco by Jean-Michel Trousse</td>
</tr>
<tr>
<td>MDS</td>
<td>1987</td>
<td>MDS develops first monthly bureau score, for bankruptcy</td>
</tr>
<tr>
<td>Experian-Scorex</td>
<td>2003</td>
<td>Created as subsidiary of Experian, after purchase of Scorex</td>
</tr>
</tbody>
</table>
## History of credit scoring

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1914</td>
<td>Western Union introduces embossed metal plate first charge card in the United States.</td>
</tr>
<tr>
<td>1920s</td>
<td>Introduction of ‘shopper’s plates’, early version of modern store cards.</td>
</tr>
<tr>
<td>1950</td>
<td>Diners Club and American Express launch first charge cards.</td>
</tr>
<tr>
<td>1951</td>
<td>Diners Club launches first credit card in New York city.</td>
</tr>
<tr>
<td>1960</td>
<td>Bank Americard established, later to become Visa.</td>
</tr>
<tr>
<td>1966</td>
<td>Master Charge established, later to become MasterCard.</td>
</tr>
<tr>
<td>1966</td>
<td>Barclaycard established in the United Kingdom.</td>
</tr>
</tbody>
</table>
The mechanics of credit scoring

• What are scorecards?
• What measures are used?
• What is the scorecard development process?
• What can affect the scorecards?
What are scorecards?

- A scorecard is actually a credit-scoring model built to evaluate risk on a unique or homogenous population.
What measures are used?

• Highly statistical and mathematical discipline
  – Process and strategy
    • Reject inference
    • Strategy

Scorecard performance
Default probability and loss severity
What is the scorecard development process?

<table>
<thead>
<tr>
<th>Task</th>
<th>Objectives</th>
</tr>
</thead>
</table>
| Data Collection        | - Identify required variables  
                        | - Define observation / performance window                                    |
| Data Cleansing         | - Eliminate duplicates / Identify exclusive accounts  
                        | - Handle missing values  
                        | - Examine / Remove outliers and abnormal values                             |
| Specification          | - Characteristic analysis  
                        | - Characteristic selection                                                   |
| Estimation             | - Determine coefficients                                                   |
| Assessment             | - Test coefficients                                                        |
| Validation             | - Measure predictive accuracy of model                                       |
| Reject Inference       | - Remove bias resulting from exclusion of rejects                           |
| Scaling                | - Create score bands                                                        |
| Cut-off Score          | - Determine accept/reject threshold                                         |
| Determination          |                                                                             |
What can affect the scorecards?

- **Economy**
  - Upturns and downturns, with changes in employment, interest rates, inflation, etc.

- **Market**
  - Changes to customer demographics, including lenders’ conscious moves up, down or across markets, as well as changes to product features that affect client behaviour

- **Operations**
  - Changes to forms, processes, systems, or calculations, used at any point, whether for new or existing business

- **Target**
  - People’s handling of debt changes over time, as do their responses to various stimuli

- **Score**
  - Driven strategies may also influence the outcomes they are meant to predict.

- **Unattributed**
  - Changes that cannot be attributed to anything, other than scorecards’ age.
Basel-I (cont’d)

- There was no clear rationale for the 8% capital requirement.
- The risk buckets were homogenous. For example, all corporate bonds faced the same capital charge regardless of (important) difference in maturity and seniority.
- The Cooke Ratio was too simple to truly evaluate solvency levels.
- Potentially risk-reducing diversification in the loan portfolio was ignored.
- Use of off-balance sheet activities to mitigate risk exposure was not recognized.
Basel-I

  - Credit Risk: \textit{Cooke ratio (capital/risk weighted assets) must exceed 8%}
  - Completely ignored market and operational risk factors
  - Faced criticism
  - Market risk factor in 1996
Baseline-II

• Baseline-II 2004
• Goal is to provide precise classifications risk levels between banks.
• Established three options
  – Standardized approach
  – Foundation/Advanced Internal rating based approaches (FIRB\AIRB)
  – Asset Securitization
Basel II (cont’d)
Minimum Requirement of IRB (SBP)

- Composition of minimum requirements
- Compliance with minimum requirements
- Rating system design
- Risk rating system operations
- Corporate governance and oversight
- Use of internal ratings
- Risk quantification
- Disclosure requirements
- Validation of Internal Estimates
Minimum Requirement of IRB (SBP) (Cont’d)

• Composition of minimum requirements
  – A meaningful assessment of borrower and transaction characteristics
  – A meaningful differentiation of risk
  – Reasonably accurate and consistent quantitative estimate of risk

• Compliance with minimum requirements
  – Must demonstrate to SBP that it meets the IRB requirements
  – Timely return to compliance in case of not complete compliance
  – SBP may ask to hold additional capital in case of non-compliance duration
Minimum Requirement of IRB (SBP) (Cont’d)

• Rating system design
  – It includes all the methods, process, control and data collection and IT system
  – Bank may have more than one models for same portfolio
  – Model requirement
    • Have a process for vetting data inputs into a statistical default
    • Data is representative of the population
    • Consider all relevant material information
    • Regular cycle of model validation
  – Time Horizon
    • Assessment horizon for PD estimate is one year (PIT)
    • Banks must use at least five years of data to estimate the PD

• Risk rating system operations
  – Coverage of ratings
  – Integrity of rating process
  – Procedure for overrides
  – Policies and procedures
  – Sound stress testing
Minimum Requirement of IRB (SBP) (Cont’d)

• Corporate governance and oversight
  – Corporate Governance (Approval of Directors and senior management)
  – Credit Risk Control
  – Audit System

• Use of internal ratings
  – Internal Ratings, default and loss estimations
  – Bank must demonstrate that it has been using its ratings system for at least three years
Minimum Requirement of IRB (SBP) (Cont’d)

• Risk quantification
  – Calculation of capital charge
• Disclosure requirements
  – Must meet certain disclosure requirements
• Validation of Internal Estimates
  – Yearly validation requirements
  – Bank may use external data for validation purpose
  – Must have standardize process of model validation
PD/Scorecard Development

• Collaboration between different departments
• Extensive team work is required.
• Work in isolation could lead to number of problems
  – Inclusion Characteristics that can not be gathered
  – Legally suspect
  – Difficult to collect operationally
Scorecard/PD Development (Cont’d)

• **Scorecard Developer**
  
  – Expertise in performing data mining and statistical analyses
  
  – An in-depth knowledge of the databases in the company

  – An in-depth understanding of statistical principles, in particular those related to predictive modeling

  – Business experience in the implementation and usage of risk models
Scorecard/PD Development (Cont’d)

- **Credit Scoring Manager**
  - Subject matter expertise in the development and implementation of risk strategies using scores
  - An in-depth understanding of corporate risk policies and procedures
  - An in-depth understanding of the risk profile of the company’s customers and applicants for products/services
  - A good understanding of the various implementation platforms for risk scoring and strategy implementation in the company
  - Knowledge of legal issues surrounding usage of particular characteristics/processes to adjudicate credit applications
  - Knowledge of credit application processing and customer management processes in the company
Scorecard/PD Development (Cont’d)

• **Product Manager(s)**
  
  – Subject matter expertise in the development and implementation of product-marketing strategies
  
  – An in-depth knowledge of the company’s typical client base and target markets
  
  – Knowledge of future product development and marketing direction

• **Operational Manager(s)**
  
  – Subject matter expertise in the implementation and execution of corporate strategies and procedures
  
  – An in-depth knowledge of customer-related issues
Scorecard/PD Development (Cont’d)

• **Project Manager**
  - Subject matter expertise in the management of projects
  - An in-depth understanding of the relevant corporate areas involved in the project

• **IT/IS Managers**
  - Subject matter expertise in the software and hardware products involved in risk management and risk scoring implementation
  - In-depth knowledge of corporate data and procedures to introduce changes to data processing
  - Knowledge of processing data from external data providers

• **Corporate Risk Management Staff**
• **Legal Staff**
Preliminaries and planning

- Create business plan
  - Reduction in bad debt/bankruptcy/claims/fraud
  - Increase in approval rates or market share in areas such as secured loans, where low delinquency presents expansion opportunities
  - Increased profitability
  - Increased operational efficiency (e.g., to better manage workflow in an adjudication environment)
  - Cost savings or faster turnaround through automation of adjudication using scorecards
  - Better predictive power (compared to existing custom or bureau scorecard)

- Identify organizational objectives and scorecard role
- Determine internal versus external development and scorecard type
- Create project plan
- Identify project risks
- Identify project team and responsibilities
Data Review & Project Parameters

• **Data Availability**
  – External data
    • More time requires to evaluate than internal data
    • Sources of external data (Credit Bureau, collaboration with another institution(s))
  – Internal data
  – Quantity
    • No. of accounts
      – Good
      – Bad
      – Declined application
  – Quality
    • Tempered data
      – Collateral value
      – CIB report
      – Financial ratios
    • Demographic data
    • Unverified applicants
No of Accounts?

- 1,500–2,000 cases of each class is a reasonable. “Credit Scoring, Response Modelling and Insurance Rating by Steven Finlay Page No (68)”

- The minimum requirements commonly quoted are 1,500 goods and 1,500 bads, with a further 1,500 rejects for selection processes. “The Credit Scoring Toolkit by Raymond Anderson (Head of Scoring at Standard Bank Africa) Page No (77)”

- As for the number in the sample, Lewis (1992) suggested that 1,500 goods and 1,500 bads may be enough. In practice, much larger samples are used, although Makuch (1999) makes the point that once one has 100,000 goods, there is no need for much more information on the goods. Thus a typical situation would be to take all the bads one can get into the sample and take 100,000+ goods. “Credit Scoring and Its Applications by Lyn C. Thomas, David B. Edelman, Jonathan N. Crook Page No (121)”
No of Accounts? (Cont’d)

• This question is common among target modelers. Unfortunately, there is no exact answer. Sample size depends on many factors. What is the expected return rate on the target group? This could be performance based such as responders, approved accounts, or activated accounts, or risk based such as defaults or claims filed. How many variables are you planning to use in the model? The more variables you have, the more data you need. The goal is to have enough records in the target group to support all levels of the explanatory variables. One way to think about this is to consider that the significance is measured on the cross-section of every level of every variable. “Data Mining Cookbook by Olivia Parr Rud Page No (44)”
No of Accounts? (Cont’d)

• **Good/Bad/Reject**
  
  – Typically about 2,000 each of goods, bads, and rejects are sufficient for scorecard development. This method is called oversampling and is widely used in the industry. Adjustments for oversampling (to be covered later in this chapter) are later applied to get realistic forecasts. An added benefit of using a large enough sample is that it reduces the impact of multicollinearity and makes the result of logistic regression statistically significant. “Credit Risk, Scorecards Developing and Implementing Intelligent Credit Scoring. Page no 63”
Rejected Accounts?

• Can we develop scorecard without rejected applications?
  – Yes
• Is it legally permissible to develop model without rejected application?
  – Yes

• If yes then how biased over scorecard model would be?
  – It depends upon how much other information was used in the decision.

“My suggestion is to develop the scorecard using what data you have, but start saving rejected applications ASAP.”

Raymond Anderson
Head of Scoring at Standard Bank Africa
Johannesburg Area, South Africa
Rejected Accounts? (Cont’d)

• Basel II asks for models without material biases. As long as the models have those removed you should be OK.

• You can adjust the bins (WOE based bins) to reduce obvious biases e.g. if bad people are performing well due to selection bias, you can artificially give them negative WOE’s

• Many banks have developed models without RI - but they tend to have lots of business input and adjustments.

Naeem Siddiqi  
Global Product Manager, Banking Analytics Solutions at SAS Institute Inc.  
Canada
Data Gathering for Definition of Project Parameters

- Definition of bad and good
- Selection of variables
  - Expected predictability power
  - Reliability and robustness
  - Ease in collection
  - Logical Interpretability
    - Parents name
    - Family background etc
  - Human intervention
  - Legal issues surrounding the usage of certain type of information
    - Some private information regarding clients that can create major problem legally
      - Religion (Shaban Azmi and Emran Hashmi incident)
      - Sexual orientation
      - Ethnicity
  - Creation of ratios based on business reasoning
    - Expense/Income of clients
    - CIB report three month status/CIB report twelve months status
  - Future availability
  - Changes in competitive environment
Proposed variables list

### Variable definitions

**Default**
- Defaulted or not defaulted client

**Socio-demographic variables**
- **Education**: The highest attained education of client, categorized variable
- **Marital status**: Status of the client, single/married, categorized variable
- **Years of employment**: The number of years in the current employment
- **Sector of employment**: The sector in which the client is employed, categorized variable
- **Sex**: Sex of the client, categorized variable
- **Date of Birth**: Date of birth of client
- **Type of employment**: Type of client’s employment, categorized variable
- **Number of employments**: The total number of employments in the last 3 years
- **Employment position**: The position of client in employment, categorized variable
- **Credit ratio 1**: Ratio of Expenditures/Income of client
- **Credit ratio 2**: Ratio of (Income-Expenditure)/Living Wage of client
- **Region**: Post Code of region of client’s address

**Bank-client relationship variables**
- **Own resources**: Declared own resources, in percentage of total amount needed
- **Amount of loan**: The total amount of loan granted
- **Purpose of loan**: The declared purpose of loan, categorized variable
- **Length of the relationship**: The length of client/bank relationship at the time of loan application
- **Date of account opening**: The year when client opened an account in the bank
- **Deposit Behavior**: The characteristics of client’s behavior with respect to her/his current account
- **Loan Protection**: The type of credit risk mitigation, categorized variable
- **Type of product**: Type of product - loan
- **Number of co-signers**: The number of co-signers for the current loan
- **Date of loan**: The year in which the loan was granted

Note: “c” denotes categorized variables.
Performance and Sample Window

• Basel II requirement is at least five years of data
• May select all the loans that were in non-default status a year ago.
Performance and Sample Window (Cont’d)

- Performance window period (at least 1 year/could be more at management’s discretion)
  - Mean period of defaulted loans
  - Weighted mean of defaulted loans
  - Industry practice (18-24 months for credit card & 3 to 5 years for mortgage portfolio)

<table>
<thead>
<tr>
<th></th>
<th>1 Mth</th>
<th>2 Mth</th>
<th>3 Mth</th>
<th>4 Mth</th>
<th>5 Mth</th>
<th>6 Mth</th>
<th>7 Mth</th>
<th>8 Mth</th>
<th>9 Mth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-03</td>
<td>0.00%</td>
<td>0.44%</td>
<td>0.87%</td>
<td>1.40%</td>
<td>2.40%</td>
<td>2.80%</td>
<td>3.20%</td>
<td>3.60%</td>
<td>4.10%</td>
</tr>
<tr>
<td>Feb-03</td>
<td>0.00%</td>
<td>0.37%</td>
<td>0.88%</td>
<td>1.70%</td>
<td>2.30%</td>
<td>2.70%</td>
<td>3.30%</td>
<td>3.50%</td>
<td></td>
</tr>
<tr>
<td>Mar-03</td>
<td>0.00%</td>
<td>0.42%</td>
<td>0.92%</td>
<td>1.86%</td>
<td>2.80%</td>
<td>3.00%</td>
<td>3.60%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr-03</td>
<td>0.00%</td>
<td>0.65%</td>
<td>1.20%</td>
<td>1.90%</td>
<td>2.85%</td>
<td>3.05%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May-03</td>
<td>0.00%</td>
<td>0.10%</td>
<td>0.80%</td>
<td>1.20%</td>
<td>2.20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun-03</td>
<td>0.00%</td>
<td>0.14%</td>
<td>0.79%</td>
<td>1.50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul-03</td>
<td>0.00%</td>
<td>0.23%</td>
<td>0.88%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aug-03</td>
<td>0.00%</td>
<td>0.16%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep-03</td>
<td>0.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NUST Collaboration with RMC Consultants
Performance and Sample Window (Cont’d)
Development of Sample specification

• Sampling
  – Scorecard development (80%)
  – Validation purpose (20%)
• Sampling would be proportion of goods, bads and rejected accounts
• Simple random sampling
• Stratified sampling
• Cluster sampling
Development of Sample specification (Cont’d)

• Appropriate/optimal size of sample
  – Proportional sampling
    • Portfolio (80% good and 20% Bad)

\[ n = \left[ \frac{z^2 p(1 - p)}{d^2} \right] \]

- \( z \) = \( z \)-statistic for the desired level of confidence (e.g., 1.96 for 95% confidence)
- \( p \) = the proportion to be achieved
- \( d \) = half the width of the desired confidence interval (e.g., within \( d \)% of the proportion you estimate from your sample)
Development of Sample specification (Cont’d)

\[ n = \left( \frac{z\sigma}{d} \right)^2 \]

\( z \) = z-statistic for the desired level of confidence (e.g., 1.96 for 95% confidence)
\( \sigma \) = the population standard deviation (usually unknown)
\( d \) = the (half) width of the desired interval
Scorecard Development Steps

1. Explore Data & Data Cleansing
2. Initial Characteristic Analysis (K - G & B)
3. Preliminary Scorecard
4. Reject Inference
5. Final Scorecard (AGB)
6. Initial Characteristic (All G & B)
7. Validation

NUST Collaboration with RMC Consultants
Explore Data

• Explore the sampling data
• Descriptive Statistics
  – Mean, Median & Mode
  – Proportion missing
  – Visual Technique
  – Interpretation of data
Data Cleansing

- Eliminate duplicates / Identify exclusive accounts (VIP)

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
<td>Sex</td>
<td>Age</td>
<td>Address</td>
<td>Time</td>
<td>Marital Status</td>
<td>Occupation</td>
<td>Home Ownership</td>
<td>Checking</td>
<td>Savings</td>
<td>Good/Bad Mark</td>
</tr>
<tr>
<td>2</td>
<td>John Smith</td>
<td>male</td>
<td>45</td>
<td>8.5 Married</td>
<td>Professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Mary White</td>
<td>female</td>
<td>19.7</td>
<td>0.21 Married</td>
<td>Blue Collar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Bill Gates</td>
<td>male</td>
<td>65</td>
<td>11 Married</td>
<td>Business Owner</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Sam Williams</td>
<td>male</td>
<td>41.4</td>
<td>5 Married</td>
<td>Professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Paul Gardner</td>
<td>male</td>
<td>17.8</td>
<td>11 Married</td>
<td>Professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Cleansing

• Handle missing values
  – Complete case
  – With average method
  – K-nearest approach
  – Regression modelling
• Examine / Remove outliers and abnormal values
Handle missing values

• Complete case
• With average method
• K-nearest approach
• Regression modelling
Examine / Remove outliers & abnormal values

• Central tendency and dispersion
  – Point Estimate
• Calculation of Standard Error
  \[ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \]
• Test statistics
  – T or Z statistics
• Formulation of Confidence Interval
  \[ \text{point estimate} \pm (\text{reliability factor} \times \text{standard error}) \]
Correlation / Multicollinearity

- Pearson Correlation
- Test Statistics

\[ t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}} \]

- Complete Principle Components Analysis (PCA) to handle correlation in data which includes:
  - Calculation of Eigenvalues
  - Calculation of Eigenvectors
  - Formulation of artificial variables

- Variance Inflation Factor (VIF)

\[ \text{VIF} = \frac{1}{1 - R^2} \]
Correlation

• **Covariance.** The magnitude of the covariance depends on the magnitude of the individual variables’ standard deviations and the relationship between their co-movements. The covariance is an absolute measure and is measured in return units squared.

\[
\text{Cov}_{1,2} = \rho_{1,2} \sigma_1 \sigma_2
\]

• Covariance can be standardized by dividing by the product of the standard deviations of the two variables being compared. This standardized measure of co-movement is called correlation and is computed as:

\[
\rho_{1,2} = \frac{\text{Cov}_{1,2}}{\sigma_1 \sigma_2}
\]

• **Pearson Correlation** product-moment correlation coefficient, also known as r, R, or Pearson's r, a measure of the strength of the linear relationship between two variables that is defined in terms of the (sample) covariance of the variables divided by their (sample) standard deviations.
The term $\rho_{1,2}$ is called the correlation coefficient between variables. The correlation coefficient has no units. It is a pure measure of the co-movement of the two variables’ and is bounded by $-1$ and $+1$.

A correlation coefficient of $+1$ means that variables always change proportionally in the same direction. They are perfectly positively correlated.

A correlation coefficient of $-1$ means that variables always move proportionally in the opposite directions. They are perfectly negatively correlated.

A correlation coefficient of zero means that there is no linear relationship between the two variables’. They are uncorrelated. One way to interpret a correlation (or covariance) of zero is that, in any period, knowing the actual value of one variable tells you nothing about the other.
Test Statistics

• As indicated earlier, the closer the correlation coefficient is to plus or minus one, the stronger the correlation.
• As indicated earlier, the closer the correlation coefficient is to plus or minus one, the stronger the correlation. With the exception of these extremes (i.e., $r = \pm 1.0$), we cannot really speak of the strength of the relationship indicated by the correlation coefficient without a statistical test of significance.
• We test whether the correlation between the population of two variables is equal to zero.
• The appropriate null and alternative hypotheses can be structured as a two-tailed test as follows:

$$H_0: \rho = 0 \text{ versus } H_a: \rho \neq 0$$

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

Reject $H_0$ if $+t_{critical} < t$, or $t < -t_{critical}$
Multicollinearity

- **Collinearity** is a linear relationship between *two* explanatory variables. Two variables are perfectly collinear if there is an exact linear relationship between the two. For example, $X_1$ and $X_2$ are perfectly collinear if there exist parameters $\lambda_0$ and $\lambda_1$ such that, for all observations $i$, we have

$$X_{2i} = \lambda_0 + \lambda_1 X_{1i}.$$ 

- **Multicollinearity** refers to the condition when two or more of the independent variables, or linear combinations of the independent variables, in a multiple regression are highly correlated with each other.

- This condition distorts the standard error of estimate and the coefficient standard errors, leading to problems when conducting T-tests for statistical significant of parameters.
Multicollinearity (Cont’d)

• As a result of multicollinearity, there is a greater probability that we will incorrectly conclude that a variable is not statistically significant.

• Multicollinearity is likely to be present to some extent in most economic models.

• The issue is whether the multicollinearity has a significant effect on the regression results.
Principal Components Analysis

- What is Principal Component Analysis
- Computing the components in PCA
- Dimensionality Reduction using PCA
- A 2D example in PCA
- Applications of PCA in computer vision
- Importance of PCA in analyzing data in higher dimensions
Principal Components Analysis (Cont’d)

Principal components analysis (PCA)

• A method of dimensionality reduction without (much) sacrificing the accuracy.
• Dimensions are no of independent variables.
• Principal Component Analysis aims to:
  • Summarize data with many independent variables to a smaller set of derived variables.
  • In such a way, that first component has maximum variance, followed by second, followed by third and so on.
  • The covariance of any of the components with any other components is zero.

In a way PCA, redistributes total variance in such a way, that first K components explains as much as possible of the total variance.

Total variance in case of N independent variables= variance of variable 1 + variance of variable 2 + variance of variable 3 +.................................+ Variance of variable N.

Why it is required?

• In many data analysis/mining scenario, independent variables are highly correlated, which affects models accuracy and reliability.
• It increases cost as once you deploy a model, you need to capture all superfluous variables.
Principal Components Analysis (Cont’d)

How to calculate Principal Components?

It involves four steps

- Get covariance/correlation matrix
- Get Eigen Values
- Get Eigen Vectors, which are the direction of principal components
- Fine co-ordinates of each data point in the directions of principal components
Principal Components Analysis
(Cont’d)

• Get the covariance matrix/correlation matrix of the independent variables. This is also called characteristic matrix.

• If the variables are on same scale (say all the variables are measured on meter scale), It is better use covariance matrix.

• If the variables are measured on quite different scales, it is advisable to first standardized the variables and then take covariance matrix (which is same as correlation matrix).

• If you don’t standardized variables in such cases, those variables which has huge numeric values because of scale will dominate the total variance hence all further analysis.

• Solve determinant \(|A-\lambda I|=0\) to get values of \(\lambda\)
  – It will give you \(N\) values of \(\lambda\), In case of \(N\) independent variables.
  – The biggest value of \(\lambda\) is called first Eigen Value and Second next value of \(\lambda\) is called second Eigen Value and so on.
Principal Components Analysis
(Cont’d)

• Put the value of biggest Eigen value in matrix $[A - \lambda I]]=0$

• Note $[A - \lambda I]*[X]=0$ by definition

• Hence once you get value of $[X]$, for which it is true, it is called first Eigen Vector (Direction of PCA).

• Repeat above three steps to get second Eigen Vector and so on.
Principal Components Analysis
(Cont’d)

- Step 4: is to obtain coordinates of data point in the direction of Eigen Vectors
- We obtain this by multiplying centered data matrix to the Eigen vector matrix.

\[
\begin{pmatrix}
\mathbf{v}_1 & \mathbf{v}_2 & \mathbf{w}_1 & \mathbf{w}_2
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 3 & 4
5 & 6 & 7 & 8
9 & 10 & 11 & 12
\end{pmatrix}
\]

Projection on the line of first principle component
Projection on the line of 2nd principle component

Variance of projections in the line of principal components

\[
\text{var}(AC_1, AC_2) = \text{var}(AD_1, AD_2)
\]
Variance Inflation Factor (VIF)

• Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

• It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

\[
VIF = \frac{1}{1 - R^2_i}
\]

• A common rule of thumb is that if \( VIF > 5 \) then multicollinearity is high.
Variable Selection Methods

- Chi Square
- Weight of Evidence & Information Value
- What is Bin?
Chi-Squared Statistics

- The chi-squared test is used to decide if there is sufficient evidence to conclude that two distributions, classed into a number of discrete intervals, are the same or different.
- Make the hypothesis that the good:bad ratio is the same in each of the bins or bands and then uses the CS statistic to check this hypothesis.
- Usually one hopes the value is low so that the hypothesis is true but here we look for binnings with large values of the CS statistic since in that case there will be significant differences in the good:bad ratio between different bins.
- The hypothesis that the distribution of goods (and bads) in each bin is the same as that in the whole population odds pG (pB) and would suggest the expected number of goods (and bads) in the bin is nkpG (nkpB).
- The CS statistic for the goodness of this fit is the sum of the squares of the differences in the forecast and observed numbers, normalized by dividing by the theoretical variance, and summed over all the bins.
Chi-Squared Statistics (Cont’d)

• Number of goods in a bin has a binomial distribution \( B(nk, pG) \) with mean \( nk \cdot pG \) and variance \( nk \cdot pG(1 - pG) \).

• The CS test then calculates for each cell the square of the difference between the actual and expected numbers in that cell divided by the expected numbers.

\[
CS = \sum_{k=1}^{K} \frac{(\text{expected number of goods in interval } k - \text{observed number of goods in interval } k)^2}{\text{expected number of goods in interval } k} + \sum_{k=1}^{K} \frac{(\text{expected number of bads in interval } k - \text{observed number of bads in interval } k)^2}{\text{expected number of bads in interval } k}.
\]

\[
CS = \sum_{k=1}^{K} \frac{(nk \cdot pG - g_k)^2}{nk \cdot pG(1 - pG)}.
\]
Chi-Squared Statistics (Cont’d)

- Chi-squared values can range from zero to infinity.
- Values greater than about 30 generally indicate a moderate relationship, with values greater than 150 indicating a strong relationship.

<table>
<thead>
<tr>
<th>Age bands</th>
<th>Actual no. goods</th>
<th>Expected no. goods</th>
<th>Actual no. bads</th>
<th>Expected no. bads</th>
<th>((\text{Act} - \text{Exp})^2/\text{Exp goods})</th>
<th>((\text{Act} - \text{Exp})^2/\text{Exp bads})</th>
<th>CS = sum of terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 30</td>
<td>175</td>
<td>180.9</td>
<td>26</td>
<td>20.1</td>
<td>0.19</td>
<td>1.73</td>
<td>2.41</td>
</tr>
<tr>
<td>Over 30</td>
<td>725</td>
<td>719.1</td>
<td>74</td>
<td>79.9</td>
<td>0.05</td>
<td>0.44</td>
<td>2.44</td>
</tr>
<tr>
<td>Under 40</td>
<td>350</td>
<td>365.4</td>
<td>56</td>
<td>40.6</td>
<td>0.65</td>
<td>5.84</td>
<td>10.93</td>
</tr>
<tr>
<td>Over 40</td>
<td>550</td>
<td>534.6</td>
<td>44</td>
<td>59.4</td>
<td>0.44</td>
<td>3.99</td>
<td>10.93</td>
</tr>
<tr>
<td>Under 50</td>
<td>525</td>
<td>536.4</td>
<td>71</td>
<td>59.6</td>
<td>0.24</td>
<td>2.18</td>
<td>6.00</td>
</tr>
<tr>
<td>Over 50</td>
<td>375</td>
<td>363.6</td>
<td>29</td>
<td>40.4</td>
<td>0.36</td>
<td>3.22</td>
<td>6.00</td>
</tr>
</tbody>
</table>
Weight of Evidence

- Weight of Evidence
  - The principle underpinning the weight of evidence transformation is simply to replace the value of a predictor variable with the associated weight of evidence.
  - The WoE is a standardized way of comparing the interval good:bad odds with the average good:bad odds of the sample.
  - WOE is used to assess the relative risk of different attributes for a characteristic, to get an indication of which are most likely to feature within a scorecard.
  - Any interval that has a higher proportion of goods than average will have a positive weight of evidence, and any interval that has a higher than average proportion of bads will have a negative WoE.
  - A WoE of 1 means the good:bad odds are $e^{2.718}$ times average, a value of 2 $e^{7.389}$ times average and so on. Likewise, a value of –1 means the odds are 2.718 times smaller than average.
  - Unfortunately, the WoE does not consider the proportion of accounts with that attribute, only the relative risk.
Weight of Evidence (Cont’d)

• WoE is mostly engaged in data prep. At one end is variable reduction. On the other end, variable derivation. In terms of variable quantity, these two goals seem to go against each other.

• WoE is able to take care of missing values in numeric variables. So no imputation will be needed.

• WoE is able to take care of outliers on both tails of numeric variables. So no Winsorization is needed. “Winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.”

• WoE with monotonic constraint is less likely to generate models with over-fitting, which is a major problem in many others.

• Implementation of WoE is very scalable in large-scale model deployment with thousands of predictors and provides the possibility to automate the modeling process and reduce the time to market.

• WoE is computational cheap for large data (tens of Gs), which is likely to be ignored by research-oriented statisticians.

• A useful property of WoE is that it always displays a linear relationship with the natural log of the good:bad odds (Ln(odds)).

\[
\text{WOE} = \left[ \ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right) \right] \times 100
\]
Weight of Evidence (Cont’d)

Predictive power of each attribute. The weight of evidence (WOE) measure is used for this purpose.

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
<th>Tot Distr</th>
<th>Goods</th>
<th>Distr Good</th>
<th>Bads</th>
<th>Distr Bad</th>
<th>Bad Rate</th>
<th>WOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td>1,000</td>
<td>2.50%</td>
<td>860</td>
<td>2.38%</td>
<td>140</td>
<td>3.65%</td>
<td>14.00%</td>
<td>-42.719</td>
</tr>
<tr>
<td>18-22</td>
<td>4,000</td>
<td>10.00%</td>
<td>3,040</td>
<td>8.41%</td>
<td>960</td>
<td>25.00%</td>
<td>24.00%</td>
<td>-108.980</td>
</tr>
<tr>
<td>23-26</td>
<td>6,000</td>
<td>15.00%</td>
<td>4,920</td>
<td>13.61%</td>
<td>1,080</td>
<td>28.13%</td>
<td>18.00%</td>
<td>-72.613</td>
</tr>
<tr>
<td>27-29</td>
<td>9,000</td>
<td>22.50%</td>
<td>8,100</td>
<td>22.40%</td>
<td>900</td>
<td>23.44%</td>
<td>10.00%</td>
<td>-4.526</td>
</tr>
<tr>
<td>30-35</td>
<td>10,000</td>
<td>25.00%</td>
<td>9,500</td>
<td>26.27%</td>
<td>500</td>
<td>13.02%</td>
<td>5.00%</td>
<td>70.196</td>
</tr>
<tr>
<td>35-44</td>
<td>7,000</td>
<td>17.50%</td>
<td>6,800</td>
<td>18.81%</td>
<td>200</td>
<td>5.21%</td>
<td>2.86%</td>
<td>128.388</td>
</tr>
<tr>
<td>44+</td>
<td>3,000</td>
<td>7.50%</td>
<td>2,940</td>
<td>8.13%</td>
<td>60</td>
<td>1.56%</td>
<td>2.00%</td>
<td>164.934</td>
</tr>
</tbody>
</table>

Total   | 40,000| 100%      | 36,160| 100%       | 3,840| 100%      | 9.60%    |
Weight of Evidence (Cont’d)

• Experts Opinion
  – A general “minimum 5% in each bucket” rule could be applied to enable meaningful analysis.

  – The bad rate and WOE should sufficiently be different from one group to the next (i.e., the grouping has to be done in a way to maximize differentiation between goods and bads). “Naeem Siddqui”

  – One need to partition your data sample into discrete sub-groups of factors and have sufficient observations within each bucket of each factor. So overall size of sample is less important than having sufficient observations within each sub cell. Paul Waterhouse (Head of The Analytical Corp)

  – You have to make sure that the distinct classes you offer to the modelling algorithm are sufficiently populated (with both goods and bads). The smaller the number of bads, the smaller will the number of characteristics be that you can use in the traditional model. Manfred Puckhaber (Head of Pricing at Hoist AG)
Weight of Evidence (Cont’d)

Illogical Trend

[Graph showing illogical WOE trend for age]
Weight of Evidence (Cont’d)

Logical Trend

LOGICAL WOE TREND FOR AGE

Predictive Strength

Weight

Missing 18–22 23–26 27–29 30–35 35–44 44 +

Age

NUST Collaboration with RMC Consultants
Weight of Evidence (Cont’d)

- WoE gives a measure of the magnitude of difference between the interval bad rate and the average bad rate.
- The Z-statistic is used to compare the proportion of bads in each interval with the rest of the sample.
- Z-statistic tells you whether or not the difference is statistically significant.
- The Z-statistic (sometimes referred to as a Z-score) provides a measure of how likely it is that the bad rates of two populations are different.

where:
- \( BR_{\neq i} \) is the bad rate for observations in all intervals except interval \( i \).
- \( BR_i \) is the bad rate for interval \( i \).
- \( BR_{all} \) is the overall bad rate for the entire sample.
- \( n_i \) is the number of observations in interval \( i \).
- \( n_{\neq i} \) is the number of observations in all intervals except interval \( i \).

\[
Z\text{-statistic} = \frac{BR_{\neq i} - BR_i}{\sqrt{\frac{BR_{all} (1 - BR_{all})}{n_{\neq i}} + \frac{BR_{all} (1 - BR_{all})}{n_i}}}
\]
Z-Statistic

• The Z-statistic is used to compare the proportion of bads in each interval with the rest of the sample.
• The larger the absolute value of the Z-statistic, then the more confident one can be that there is a genuine difference between the bad rates of the two groups.
• Any given value of the Z-statistic translates to a level of confidence that the observed difference is genuine and not some random feature of the process used to produce the sample.
• A Z-statistic with an absolute value greater than 3.58 can be taken to mean a very strong (>99.9%) level of confidence that the bad rate of the interval is different from the bad rate of the rest of the sample.
• Absolute values between 1.96 and 3.58 indicate a moderate degree of confidence (somewhere between 95 percent and 99.9 percent) that the bad rate within the interval is significantly different.
Z-Statistic (Cont’d)

Univariate analysis report for age

<table>
<thead>
<tr>
<th>Interval</th>
<th>Attribute</th>
<th>Number of goods</th>
<th>Number of bads</th>
<th>Total number</th>
<th>% of total</th>
<th>Good rate (response rate)</th>
<th>Bad rate (non-response rate)</th>
<th>Good:bad odds</th>
<th>Weight of evidence</th>
<th>Z-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;=21</td>
<td>72</td>
<td>8,928</td>
<td>9,000</td>
<td>5.00</td>
<td>0.80%</td>
<td>99.20%</td>
<td>0.008</td>
<td>-0.831</td>
<td>-7.414</td>
</tr>
<tr>
<td>2</td>
<td>22–24</td>
<td>85</td>
<td>8,915</td>
<td>9,000</td>
<td>5.00</td>
<td>0.94%</td>
<td>99.06%</td>
<td>0.010</td>
<td>-0.667</td>
<td>-6.385</td>
</tr>
<tr>
<td>3</td>
<td>25–27</td>
<td>99</td>
<td>8,901</td>
<td>9,000</td>
<td>5.00</td>
<td>1.10%</td>
<td>98.90%</td>
<td>0.011</td>
<td>-0.509</td>
<td>-5.228</td>
</tr>
<tr>
<td>4</td>
<td>28–30</td>
<td>117</td>
<td>8,883</td>
<td>9,000</td>
<td>5.00</td>
<td>1.30%</td>
<td>98.70%</td>
<td>0.013</td>
<td>-0.340</td>
<td>-3.770</td>
</tr>
<tr>
<td>5</td>
<td>31–32</td>
<td>122</td>
<td>8,879</td>
<td>9,000</td>
<td>5.00</td>
<td>1.35%</td>
<td>98.65%</td>
<td>0.014</td>
<td>-0.302</td>
<td>-3.406</td>
</tr>
<tr>
<td>6</td>
<td>33–34</td>
<td>128</td>
<td>8,872</td>
<td>9,000</td>
<td>5.00</td>
<td>1.42%</td>
<td>98.58%</td>
<td>0.014</td>
<td>-0.251</td>
<td>-2.896</td>
</tr>
<tr>
<td>7</td>
<td>35–36</td>
<td>131</td>
<td>8,870</td>
<td>9,000</td>
<td>5.00</td>
<td>1.45%</td>
<td>98.55%</td>
<td>0.015</td>
<td>-0.230</td>
<td>-2.677</td>
</tr>
<tr>
<td>8</td>
<td>37–38</td>
<td>142</td>
<td>8,858</td>
<td>9,000</td>
<td>5.00</td>
<td>1.58%</td>
<td>98.42%</td>
<td>0.016</td>
<td>-0.141</td>
<td>-1.708</td>
</tr>
<tr>
<td>9</td>
<td>39–40</td>
<td>150</td>
<td>8,850</td>
<td>9,000</td>
<td>5.00</td>
<td>1.67%</td>
<td>98.33%</td>
<td>0.017</td>
<td>-0.086</td>
<td>-1.074</td>
</tr>
<tr>
<td>10</td>
<td>41–42</td>
<td>161</td>
<td>8,839</td>
<td>9,000</td>
<td>5.00</td>
<td>1.79%</td>
<td>98.21%</td>
<td>0.018</td>
<td>-0.016</td>
<td>-0.200</td>
</tr>
<tr>
<td>11</td>
<td>43–44</td>
<td>166</td>
<td>8,834</td>
<td>9,000</td>
<td>5.00</td>
<td>1.84%</td>
<td>98.16%</td>
<td>0.019</td>
<td>0.013</td>
<td>0.164</td>
</tr>
<tr>
<td>12</td>
<td>45–46</td>
<td>176</td>
<td>8,824</td>
<td>9,000</td>
<td>5.00</td>
<td>1.96%</td>
<td>98.05%</td>
<td>0.020</td>
<td>0.074</td>
<td>1.002</td>
</tr>
<tr>
<td>13</td>
<td>47–48</td>
<td>189</td>
<td>8,811</td>
<td>9,000</td>
<td>5.00</td>
<td>2.10%</td>
<td>97.90%</td>
<td>0.021</td>
<td>0.147</td>
<td>2.059</td>
</tr>
<tr>
<td>14</td>
<td>49–50</td>
<td>198</td>
<td>8,802</td>
<td>9,000</td>
<td>5.00</td>
<td>2.20%</td>
<td>97.80%</td>
<td>0.022</td>
<td>0.195</td>
<td>2.787</td>
</tr>
<tr>
<td>15</td>
<td>51–52</td>
<td>207</td>
<td>8,793</td>
<td>9,000</td>
<td>5.00</td>
<td>2.30%</td>
<td>97.70%</td>
<td>0.024</td>
<td>0.240</td>
<td>3.516</td>
</tr>
<tr>
<td>16</td>
<td>53–55</td>
<td>212</td>
<td>8,789</td>
<td>9,000</td>
<td>5.00</td>
<td>2.35%</td>
<td>97.65%</td>
<td>0.024</td>
<td>0.262</td>
<td>3.880</td>
</tr>
<tr>
<td>17</td>
<td>56–59</td>
<td>221</td>
<td>8,780</td>
<td>9,000</td>
<td>5.00</td>
<td>2.45%</td>
<td>97.55%</td>
<td>0.025</td>
<td>0.305</td>
<td>4.609</td>
</tr>
<tr>
<td>18</td>
<td>60–62</td>
<td>225</td>
<td>8,775</td>
<td>9,000</td>
<td>5.00</td>
<td>2.50%</td>
<td>97.50%</td>
<td>0.026</td>
<td>0.326</td>
<td>4.973</td>
</tr>
<tr>
<td>19</td>
<td>63–66</td>
<td>230</td>
<td>8,771</td>
<td>9,000</td>
<td>5.00</td>
<td>2.55%</td>
<td>97.45%</td>
<td>0.026</td>
<td>0.346</td>
<td>5.338</td>
</tr>
<tr>
<td>20</td>
<td>&gt;=67</td>
<td>243</td>
<td>8,757</td>
<td>9,000</td>
<td>5.00</td>
<td>2.70%</td>
<td>97.30%</td>
<td>0.028</td>
<td>0.405</td>
<td>6.431</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3,271</td>
<td>176,729</td>
<td>180,000</td>
<td>100.00</td>
<td>1.82%</td>
<td>98.18%</td>
<td>0.019</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>
Measure of Association

• Weights of Evidence and the Z-statistic can be useful when looking at individual attributes within a variable, but what is often of greater interest are combined measures of association that take into account all of the individual attributes that a predictor variable contains.

• Information value Probably the most popular measure of association used for classification problems.

• Information value is used to measure the predictive power of a characteristic. It is most commonly used measure to assess the uni-variate predictive power of characteristics in retail credit scoring.

• Information values can range from zero to infinity, but values in the range 0–1 are most common.
Calculation of Weight of Evidence & Information Value

\[
\text{WOE} = \left[ \ln \left( \frac{\text{Distr Good}}{\text{Distr Bad}} \right) \right] \times 100
\]

\[
\text{Information Value} = \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) \times \ln \left( \frac{\text{Distr Good}_i}{\text{Distr Bad}_i} \right)
\]

<table>
<thead>
<tr>
<th>Value of IV</th>
<th>Statistical strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than 0.02</td>
<td>a very weak statistical relation</td>
</tr>
<tr>
<td>0.02 – 0.1</td>
<td>a weak statistical relation</td>
</tr>
<tr>
<td>0.1 – 0.3</td>
<td>an average statistical relation</td>
</tr>
<tr>
<td>0.3 – 0.5</td>
<td>a strong statistical relation</td>
</tr>
<tr>
<td>greater than 0.5</td>
<td>an extremely strong statistical relation</td>
</tr>
</tbody>
</table>
Calculating information value for employment status

<table>
<thead>
<tr>
<th>Interval</th>
<th>Classing</th>
<th>Number of goods</th>
<th>Number of bads</th>
<th>Total number</th>
<th>% of total</th>
<th>Bad rate (non-response rate)</th>
<th>Good:bad odds</th>
<th>Weight of evidence</th>
<th>Z-statistic</th>
<th>Information value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Full-time employed</td>
<td>1,954</td>
<td>111,582</td>
<td>113,536</td>
<td>63.08</td>
<td>98.28%</td>
<td>0.018</td>
<td>-0.055</td>
<td>-4.002</td>
<td>0.002</td>
</tr>
<tr>
<td>2</td>
<td>Part-time employed</td>
<td>228</td>
<td>15,870</td>
<td>16,098</td>
<td>8.94</td>
<td>98.58%</td>
<td>0.014</td>
<td>-0.253</td>
<td>-3.993</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>Self-employed</td>
<td>515</td>
<td>8,482</td>
<td>8,997</td>
<td>5.00</td>
<td>94.28%</td>
<td>0.061</td>
<td>1.188</td>
<td>28.460</td>
<td>0.130</td>
</tr>
<tr>
<td>4</td>
<td>Student</td>
<td>170</td>
<td>2,886</td>
<td>3,056</td>
<td>1.70</td>
<td>94.44%</td>
<td>0.059</td>
<td>1.158</td>
<td>15.633</td>
<td>0.041</td>
</tr>
<tr>
<td>5</td>
<td>Homemaker</td>
<td>168</td>
<td>12,815</td>
<td>12,983</td>
<td>7.21</td>
<td>98.71%</td>
<td>0.013</td>
<td>-0.345</td>
<td>-4.635</td>
<td>0.007</td>
</tr>
<tr>
<td>6</td>
<td>Retired</td>
<td>117</td>
<td>13,744</td>
<td>13,861</td>
<td>7.70</td>
<td>99.16%</td>
<td>0.009</td>
<td>-0.777</td>
<td>-8.929</td>
<td>0.033</td>
</tr>
<tr>
<td>7</td>
<td>Unemployed</td>
<td>119</td>
<td>11,350</td>
<td>11,469</td>
<td>6.37</td>
<td>98.96%</td>
<td>0.011</td>
<td>-0.565</td>
<td>-6.431</td>
<td>0.016</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>3,271</strong></td>
<td><strong>176,729</strong></td>
<td><strong>180,000</strong></td>
<td><strong>100.00</strong></td>
<td><strong>98.18%</strong></td>
<td><strong>0.019</strong></td>
<td><strong>0.00</strong></td>
<td><strong>0.234</strong></td>
<td></td>
</tr>
</tbody>
</table>
# Final List of Variables

## Variable definitions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Default</strong></td>
<td>Defaulted or not defaulted client</td>
</tr>
<tr>
<td><strong>Socio-demographic variables</strong></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>The highest attained education of client, categorized variable</td>
</tr>
<tr>
<td>Marital status</td>
<td>Status of the client, single/married, categorized variable</td>
</tr>
<tr>
<td>Years of employment</td>
<td>The number of years in the current employment</td>
</tr>
<tr>
<td>Sector of employment</td>
<td>The sector in which the client is employed, categorized variable</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex of the client, categorized variable</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>Date of birth of client</td>
</tr>
<tr>
<td>Type of employment</td>
<td>Type of client’s employment, categorized variable</td>
</tr>
<tr>
<td>Number of employments</td>
<td>The total number of employments in the last 3 years</td>
</tr>
<tr>
<td>Employment position</td>
<td>The position of client in employment, categorized variable</td>
</tr>
<tr>
<td>Credit ratio 1</td>
<td>Ratio of Expenditures/Income of client</td>
</tr>
<tr>
<td>Credit ratio 2</td>
<td>Ratio of (Income-Expenditure)/Living Wage of client</td>
</tr>
<tr>
<td>Region</td>
<td>Post Code of region of client’s address</td>
</tr>
<tr>
<td><strong>Bank-client relationship variables</strong></td>
<td></td>
</tr>
<tr>
<td>Own resources</td>
<td>Declared own resources, in percentage of total amount needed</td>
</tr>
<tr>
<td>Amount of loan</td>
<td>The total amount of loan granted</td>
</tr>
<tr>
<td>Purpose of loan</td>
<td>The declared purpose of loan, categorized variable</td>
</tr>
<tr>
<td>Length of the Relationship</td>
<td>The length of client/bank relationship at the time of loan application</td>
</tr>
<tr>
<td>Date of account opening</td>
<td>The year when client opened an account in the bank</td>
</tr>
<tr>
<td>Deposit Behavior</td>
<td>The characteristics of client’s behavior with respect to her/his current account</td>
</tr>
<tr>
<td>Loan Protection</td>
<td>The type of credit risk mitigation, categorized variable</td>
</tr>
<tr>
<td>Type of product</td>
<td>Type of product - loan</td>
</tr>
<tr>
<td>Number of co-signers</td>
<td>The number of co-signers for the current loan</td>
</tr>
<tr>
<td>Date of loan</td>
<td>The year in which the loan was granted</td>
</tr>
</tbody>
</table>

Note: “c” denotes categorized variables.
Regression Analysis
Regression Analysis

• The purpose of simple linear regression is to explain the variation in a dependent variable in terms of the variation in a single independent variable. Here, the term “variation” is interpreted as the degree to which a variable differs from its mean value. Don’t confuse variation with variance—they are related but are not the same.

• Multiple regression is regression analysis with more than one independent variable. It is used to quantify the influence of two or more independent variables on a dependent variable.

• The dependent variable is the variable whose variation is explained by the independent variable. We are interested in answering the question, “What explains fluctuations in the dependent variable?” The dependent variable is also referred to as the explained variable, the endogenous variable, or the predicted variable.

• The independent variable is the variable used to explain the variation of the dependent variable. The independent variable is also referred to as the explanatory variable, the exogenous variable, or the predicting variable.
Regression Analysis (Cont’d)

• Assumptions underlying linear regression.

Linear regression requires a number of assumptions. As indicated in the following list, most of the major assumptions pertain to the regression model’s residual term ($\varepsilon$).

• A linear relationship exists between the dependent and the independent variable.
• The independent variable is uncorrelated with the residuals.
• The expected value of the residual term is zero [$E(\varepsilon) = 0$].
• The variance of the residual term is constant for all observations [$E(\varepsilon_i^2) = \sigma^2_\varepsilon$].
• The residual term is independently distributed; that is, the residual for one observation is not correlated with that of another observation [$E(\varepsilon_i \varepsilon_j) = 0, j \neq i$].
• The residual term is normally distributed.
Regression Analysis (Cont’d)

• The following linear regression model is used to describe the relationship between two variables, $X$ and $Y$:

$$Y_i = b_0 + b_1X_i + \varepsilon_i, \ i = 1, \ldots, n$$

– $Y_i$ = $i$th observation of the dependent variable, $Y$
– $X_i$ = $i$th observation of the independent variable, $X$
– $b_0$ = regression intercept term
– $b_1$ = regression slope coefficient
– $\varepsilon_i$ = residual for the $i$th observation (also referred to as the disturbance term or error term)

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i$$
Regression Analysis (Cont’d)

• Ordinary Least Squares Regression

  – Ordinary least squares (OLS) estimation is a process that estimates the population parameters.

  – The estimated slope coefficient ($\hat{b}_1$) for the regression line describes the change in $Y$ for a one unit change in $X$. It can be positive, negative, or zero, depending on the relationship between the regression variables. The slope term is calculated as:

    \[ \hat{b}_1 = \frac{\text{cov}_{XY}}{\sigma_X^2} \]

  – The intercept term ($\hat{b}_0$) is the line’s intersection with the Y-axis at $X = 0$. It can be positive, negative, or zero. A property of the least squares method is that the intercept term may be expressed as:

    \[ \hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} \]
Regression Analysis (Cont’d)

• Standard errors of coefficients
  – The standard error of estimate (SEE) measures the degree of variability of the actual Y-values relative to the estimated Y-values from a regression equation. The SEE gauges the “fit” of the regression line. The smaller the standard error, the better the fit.
  – The SEE is the standard deviation of the error terms in the regression. As such, SEE is also referred to as the standard error of the residual, or standard error of the regression.
Regression Analysis (Cont’d)

• Hypothesis Testing

  – Hypothesis testing for a regression coefficient may use the confidence interval for the coefficient being tested.

  – We check whether an estimated slope coefficient is statistically different from zero.

  – Null hypothesis is $H_0: b_1 = 0$ and the alternative hypothesis is $H_a: b_1 \neq 0$. 
Regression Analysis (Cont’d)

• A t-test may also be used to test the hypothesis that the true slope coefficient, $b_1$, is equal to some hypothesized value. Letting $\hat{b}_1$ be the point estimate for $b_1$, the appropriate test statistic with $n - 2$ degrees of freedom is:

$$t_{b_1} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

Reject $H_0$ if $t > +t_{critical}$ or $t < -t_{critical}$

• Rejection of the null means that the slope coefficient is different from the hypothesized value of $b_1$. 
Regression Analysis (Cont’d)

• Interpreting p-Values

  – The p-value is the smallest level of significance for which the null hypothesis can be rejected. An alternative method of doing hypothesis testing of the coefficients is to compare the p-value to the significance level:

  – If the p-value is less than significance level, the null hypothesis can be rejected.

  – If the p-value is greater than the significance level, the null hypothesis cannot be rejected.
Regression Analysis (Cont’d)

• **Total sum of squares (SST)** measures the total variation in the dependent variable. SST is equal to the sum of the squared differences between the actual Y-values and the mean of Y:

\[
SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2
\]

• **Regression sum of squares (RSS)** measures the variation in the dependent variable that is explained by the independent variable. RSS is the sum of the squared distances between the predicted Y-values and the mean of Y.

\[
RSS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2
\]

• **Sum of squared errors (SSE)** measures the unexplained variation in the dependent variable. It’s also known as the sum of squared residuals or the residual sum of squares. SSE is the sum of the squared vertical distances between the actual Y-values and the predicted Y-values on the regression line.

\[
SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]
Regression Analysis (Cont’d)

• Thus, total variation = explained variation + unexplained variation, or:
  – SST = RSS + SSE
Regression Analysis (Cont’d)

• Coefficient of Determination ($R^2$):
  – $R^2$ is the percentage of the total variation in the dependent variable explained by the independent variable:

\[ R^2 = \frac{\text{total variation (SST)} - \text{unexplained variation (SSE)}}{\text{total variation (SST)}} \]

\[ = \frac{\text{explained variation (RSS)}}{\text{total variation (SST)}} \]

  – Regression value can range from Minus Infinity to plus infinity.
Regression Analysis (Cont’d)

• The F-Statistic
  – An F-test assesses how well a set of independent variables, as a group, explains the variation in the dependent variable. In multiple regression, the F-statistic is used to test whether at least one independent variable in a set of independent variables explains a significant portion of the variation of the dependent variable.

• The *F-statistic, which is always a one-tailed test, is calculated as:*

  \[ F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-k-1)} \]

  – RSS = regression sum of squares
  – SSE = sum of squared errors
  – MSR = mean regression sum of squares
  – MSE = mean squared error

  \[ H_0: b_1 = b_2 = b_3 = b_4 = 0 \text{ versus } H_a: \text{ at least one } b_j \neq 0 \]
Regression Analysis (Cont’d)

• What is Heteroskedasticity?
  – **Heteroskedasticity** occurs when the variance of the residuals is not the same across all observations in the sample. This happens when there are subsamples that are more spread out than the rest of the sample.
  – **Unconditional heteroskedasticity** occurs when the heteroskedasticity is not related to the level of the independent variables, which means that it doesn’t systematically increase or decrease with changes in the value of the independent variable(s). While this is a violation of the equal variance assumption, it usually causes no major problems with the regression.
  – **Conditional heteroskedasticity** is heteroskedasticity that is related to the level of (i.e., conditional on) the independent variables. For example, conditional heteroskedasticity exists if the variance of the residual term increases as the value of the independent variable increases.
Regression Analysis (Cont’d)

• Effect of Heteroskedasticity on Regression Analysis
  – There are four effects of heteroskedasticity you need to be aware of:
    • The standard errors are usually unreliable estimates.
    • The coefficient estimates \( \hat{b}_j \) aren’t affected.
    • If the standard errors are too small, but the coefficient estimates themselves are not affected, the t-statistics will be too large and the null hypothesis of no statistical significance is rejected too often. The opposite will be true if the standard errors are too large.
    • The F-test is also unreliable.
Regression Analysis (Cont’d)

• Detecting Heteroskedasticity

  – There are two methods to detect heteroskedasticity:
    • Examining scatter plots of the residuals and using the Breusch-Pagan chi-square test. A scatter plot of the residuals versus one or more of the independent variables can reveal patterns among observations.
    • The more common way to detect conditional heteroskedasticity is the Breusch-Pagan test, which calls for the regression of the squared residuals on the independent variables. If conditional heteroskedasticity is present, the independent variables will significantly contribute to the explanation of the squared residuals. The test statistic for the Breusch-Pagan test, which has a chi-square distribution, is calculated as:

  \[
  \text{BP chi-square test} = n \times R^2_{\text{resid}} \quad \text{with k degrees of freedom}
  \]

  – where:
    - \( n \) = the number of observations
    - \( R^2_{\text{resid}} \) = \( R^2 \) from a second regression of the squared residuals from the first regression on the independent variables
    - \( k \) = the number of independent variables

  • This is a one-tailed test because heteroskedasticity is only a problem if the \( R^2 \) and the BP test statistic are too large.
Regression Analysis (Cont’d)

• Limitations of regression analysis include the following:

  – Linear relationships can change over time. This means that the estimation equation based on data from a specific time period may not be relevant for forecasts or predictions in another time period. This is referred to as parameter instability.

  – Even if the regression model accurately reflects the historical relationship between the two variables, its usefulness in investment analysis will be limited if other market participants are also aware of and act on this evidence.

  – If the assumptions underlying regression analysis do not hold, the interpretation and tests of hypotheses may not be valid. For example, if the data is heteroskedastic (nonconstant variance of the error terms) or exhibits autocorrelation (error terms are not independent), regression results may be invalid.
Logistic Regression
Logistic Regression

• Regression analysis often calls for the use of a model that has a qualitative dependent variable, a dummy variable that takes on a value of either zero or one.

• Dependent variable may take on a value of one in the event of default and zero in the event of no default.

• An ordinary regression model is not appropriate for situations that require a qualitative dependent variable.

• Logistic Regression
  – Binary Logistic Regression
  – Multinomial / Ordinal Logistic Regression
Probit vs Logit Model

- Probit and logit models. A probit model is based on the normal distribution, while a logit model is based on the logistic distribution. Application of these models results in estimates of the probability that the event occurs (e.g., probability of default). The maximum likelihood methodology is used to estimate coefficients for probit and logit models. These coefficients relate the independent variables to the likelihood of an event occurring, such as a merger, bankruptcy, or default.

- Logistic and linear regression are both based on many of the same assumptions and theory. While convenient in many ways this presents a minor problem with regard to the outcome. Since the outcome is dichotomous, predicting unit change has little or no meaning. As an alternative to modeling the value of the outcome, logistic regression focuses instead upon the relative probability (odds) of obtaining a given result category. As it turns out the natural logarithm of the odds is linear across most of its range, allowing us to continue using many of the methods developed for linear models.
Logistic Regression (Cont’d)

- Logistic Regression value can range between 0 and 1.

• There are two defining properties of probability.
  - The probability of occurrence of any event \((E)\) is between 0 and 1 (i.e., \(0 \leq P(E) \leq 1\)).
  - If a set of events, \(E_1, E_2, \ldots, E_n\), is mutually exclusive and exhaustive, the probabilities of those events sum to 1 (i.e., \(\sum P(E_i) = 1\)).
Logistic Regression (Cont’d)
Logistic Regression (Cont’d)

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
<th>Duration in month</th>
<th>Credit history</th>
<th>Purpose</th>
<th>Credit amount</th>
<th>Savings account/bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>6 critical account/ other credits existing</td>
<td>radio/television</td>
<td>1169 unknown/ no savings account</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td></td>
<td>48 existing credits paid back duly till now</td>
<td>radio/television</td>
<td>5951 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>12 critical account/ other credits existing</td>
<td>education</td>
<td>2096 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>42 existing credits paid back duly till now</td>
<td>furniture/equipment</td>
<td>7882 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td></td>
<td>24 delay in paying off in the past</td>
<td>car (new)</td>
<td>4870 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>36 existing credits paid back duly till now</td>
<td>education</td>
<td>9055 unknown/ no savings account</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>24 existing credits paid back duly till now</td>
<td>furniture/equipment</td>
<td>2835,500 ≤ ... &lt; 1000 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>36 existing credits paid back duly till now</td>
<td>car (used)</td>
<td>6948 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>12 existing credits paid back duly till now</td>
<td>radio/television</td>
<td>3059 &lt; 1000 DM</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td></td>
<td>30 critical account/ other credits existing</td>
<td>car (new)</td>
<td>5234 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td></td>
<td>12 existing credits paid back duly till now</td>
<td>car (new)</td>
<td>1295 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td></td>
<td>48 existing credits paid back duly till now</td>
<td>business</td>
<td>4308 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>12 existing credits paid back duly till now</td>
<td>radio/television</td>
<td>1596 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>24 critical account/ other credits existing</td>
<td>car (new)</td>
<td>1449 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>15 existing credits paid back duly till now</td>
<td>car (used)</td>
<td>1403 &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td></td>
<td>24 existing credits paid back duly till now</td>
<td>radio/television</td>
<td>1282,100 ≤ ... &lt; 500 DM</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>24 critical account/ other credits existing</td>
<td>radio/television</td>
<td>2424 unknown/ no savings account</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td>30 no credits taken/ all credits paid back duly</td>
<td>business</td>
<td>8072 unknown/ no savings account</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>goodbad</th>
<th>durmonth</th>
<th>credhist</th>
<th>purpose</th>
<th>credamo</th>
<th>savac</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1169</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>48</td>
<td>2</td>
<td>1</td>
<td>5951</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>2096</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>42</td>
<td>2</td>
<td>3</td>
<td>7882</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>3</td>
<td>4</td>
<td>4870</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>36</td>
<td>2</td>
<td>2</td>
<td>9055</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>24</td>
<td>2</td>
<td>3</td>
<td>2835</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>36</td>
<td>2</td>
<td>5</td>
<td>6948</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>3059</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>1</td>
<td>4</td>
<td>5234</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>2</td>
<td>4</td>
<td>1295</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>48</td>
<td>2</td>
<td>6</td>
<td>4308</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>1567</td>
<td>2</td>
</tr>
</tbody>
</table>
Logistic Regression (Cont’d)

\[ \log \left( \frac{p}{1-p} \right) = \text{const} + \beta_1 \text{Age} (\leq 25) + \beta_2 \text{Age} (\text{from} 25 - 30) + \beta_3 \text{Age} (\geq 30) + \beta_4 \text{Salary} (\leq 1500) + \beta_5 \text{Salary} (1500 - 2300) + \beta_6 \text{Salary} (2300 +) + \cdots \]
# Probability of default Calculation

Probability of Default

\[ f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} \]

<table>
<thead>
<tr>
<th>$z$</th>
<th>Exponential $z$</th>
<th>$1 + \text{Exponential } z$</th>
<th>Final PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6304</td>
<td>1.878361774</td>
<td>2.878361774</td>
<td>65.26%</td>
</tr>
</tbody>
</table>

NUST Collaboration with RMC Consultants
PD Risk Bucketing & Pooling

• Risk Buckets
  – Internal buckets
    • Calinski–Harabasz Statistic
  – External mapping
    • Benchmarking

• Pooled PD
  – Historical default experience approach
  – Statistical model approach
  – External mapping approach
Calinski–Harabasz Statistic

• An algorithm usually associated with determining the optimal number of clusters, which can also be used to compare different grouping options.

• 1974, clustering technique provided by Calinski and Harabasz.

• It is best possible means of determining the optimal number of groups.

• The goal is to define clusters that maximise within group similarities, and between group differences, using their Variances.
Calinski–Harabasz Statistic (Cont’d)

Calinski–Harabasz Statistic

\[
CH(g) = \frac{BSS/(g-1)}{WSS/(n-g)} = \frac{\sum_{k=1}^{g} n_k (p_k - \bar{p})^2 / (g-1)}{\sum_{k=1}^{g} \sum_{i=1}^{n_k} n_k (P_{i,k} - p_k)^2 / (n-g)}
\]

- \( g \) - is the total number of groups.
- \( n \) - is the number of observations.
- \( p \) - is an observed probability.
- \( P \) - is a default 0/1 indicator for each record.
- \( i \) and \( k \) are indices for each record and group respectively, and BSS and WSS are the between and within group sums of squares.
Benchmarking

• In many cases, lenders want to map PDs onto a set of risk grades.
• The grades may be internal or external benchmarks.
• A good example is where PDs, used to rate companies, are mapped to an equivalent Moody, S&P, or Fitch letter grade (Local companies could be PACRA & JCR-VIS).
Benchmarking (Cont’d)

Benchmark breakpoints

\[
\min \sum_{s_1, \ldots, s_{k-1}} n_k \left( \ln \left( \frac{1 - p_k^b}{p_k^b} \right) - \ln \left( \frac{1 - p_k}{p_k} \right) \right)^2
\]

• Determine the number of classes required.
• Determine the score breaks that provide groups of almost equal size.
• Calculate the sum of the squared differences between the benchmark and banded natural log odds for each.
• For each break, calculate the sum of squares for a shift up or down one point.
• Choose the modification that provides the greatest reduction.
• Repeat from (iv), until no further reductions can be achieved.
Pooled PD

• Under a historical default experience approach, the pooled PD for a risk bucket is estimated using historical data on the frequency of observed defaults among obligors assigned to that bucket.

• Default Frequency (DF)
  – The default frequency (DF) for a risk bucket is defined as the observed default rate for the bucket over a fixed assessment horizon (usually one year).

  \[ DF_t = \frac{D_t}{N_t} \]

• Long-run default frequency
  – The long-run default frequency (LRDF) for a risk bucket is simply the average of that bucket’s annual default rates taken over a number of years.

  \[ LRDF = \frac{1}{T} \sum_{t=1}^{T} DF_t \]
Pooled PD (Cont’d)

• Statistical models
  – The statistical models approach relies on an underlying empirical default prediction model.

  – Pooled PD is derived by taking the average of the estimated obligor-specific PDs for the obligors currently assigned to the risk bucket.

  – It is important to recognize that the statistical models approach is only as accurate as the underlying default prediction model.

  – The practical challenge for bank supervisors and risk managers will lie in verifying that this model produces accurate estimates
Pooled PD (Cont’d)

• **External Mapping**
  – A bank simply establishes a mapping between its internal rating system and an external scale such as that of Moody’s or S&P.

  – Calculates a pooled PD for each external grade using an external reference dataset, and then assigns the pooled PD for the external grade to its internal grade by means of the mapping.

  – Despite its apparent simplicity, this approach poses some difficult validation challenges for supervisors and risk managers.

  – To validate the accuracy of a bank’s pooled PDs, supervisors and risk managers must first confirm the accuracy of the pooled PDs 20 Studies on the Validation of Internal Rating Systems associated with the external rating scale.

  – They must then validate the accuracy of the bank’s mapping between internal and external grades.
Where do you use PD?

• Credit risk capital charge
• Acceptance/Rejection
• Assigning Credit Limit
• Credit Pricing
Points to double the odds (PDOs)

• What is PDO
• How it works
• Odds at a certain score
• Points to double the odds

\[ pdo = \text{Factor} \times \ln (2) \]
What is Factor

• What is Factor
• Variable increment/Decrement Factor

\[ \text{Factor} = \frac{pd0}{\ln (2)} \]
What is Offset

• What is Offset
• Score at Odd ratio is 1:1
• It’s a Constant

\[
\text{Offset} = \text{Score} - \{\text{Factor} \times \ln (\text{Odds})\}
\]
PDOs, Factor & Offset

• Over time Scorecard degrade somewhat
• Effectiveness is measured by slope of the Odd vs Score curve
• Low PDO indicates that the scorecard identifies risk very effectively and vice versa
• Factor and Offset remain constant
Scaling Of Scorecard

\[ \text{Score} = \text{Offset} + \text{Factor} \times \ln (\text{odds}) \]

- WOE = weight of evidence for each grouped attribute
- \( \hat{a} \) = regression coefficient for each characteristic
- \( a \) = intercept term from logistic regression
- \( n \) = number of characteristics
- \( k \) = number of groups (of attributes) in each characteristic
Scaling Of Scorecard

\[ -\left( \sum_{j, i=1}^{k, n} (\text{woe}_j \times \beta_i) + a \right) \times \text{factor} + \text{offset} \]

- **WOE** = weight of evidence for each grouped attribute
- \( \hat{a} \) = regression coefficient for each characteristic
- **a** = intercept term from logistic regression
- **n** = number of characteristics
- **k** = number of groups (of attributes) in each characteristic
Scaling Of Scorecard

\[- \left( \sum_{j, i=1}^{k, n} \left( \text{woe}_j \times \beta_i + \frac{a}{n} \right) \right) \times \text{factor} + \text{offset} \]

- WOE = weight of evidence for each grouped attribute
- \( \hat{a} \) = regression coefficient for each characteristic
- a = intercept term from logistic regression
- n = number of characteristics
- k = number of groups (of attributes) in each characteristic
Scaling Of Scorecard

\[ \sum_{j=1}^{k} \sum_{i=1}^{n} \left( - \left( woe_j \cdot \beta_i + \frac{a}{n} \right) \cdot \text{factor} + \frac{\text{offset}}{n} \right) \]

- WOE = weight of evidence for each grouped attribute
- \( \hat{\alpha} \) = regression coefficient for each characteristic
- a = intercept term from logistic regression
- n = number of characteristics
- k = number of groups (of attributes) in each characteristic

NUST Collaboration with RMC Consultants
Scaling

**Sample Scorecard**

<table>
<thead>
<tr>
<th>Characteristic Name</th>
<th>Attribute</th>
<th>Scorecard Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>. -&gt; 23</td>
<td>63</td>
</tr>
<tr>
<td>Age</td>
<td>23 -&gt; 25</td>
<td>76</td>
</tr>
<tr>
<td>Age</td>
<td>25 -&gt; 28</td>
<td>79</td>
</tr>
<tr>
<td>Age</td>
<td>28 -&gt; 34</td>
<td>85</td>
</tr>
<tr>
<td>Age</td>
<td>34 -&gt; 46</td>
<td>94</td>
</tr>
<tr>
<td>Age</td>
<td>46 -&gt; 51</td>
<td>103</td>
</tr>
<tr>
<td>Age</td>
<td>51 -&gt; .</td>
<td>105</td>
</tr>
<tr>
<td>Cards</td>
<td><em>American Express,</em> <em>Visa Others,</em> <em>Visa MyBank,</em> <em>No Credit Cards</em></td>
<td>80</td>
</tr>
<tr>
<td>Cards</td>
<td><em>Cheque Card,</em> <em>MasterCard/Eurocard,</em> <em>Other Credit Card</em></td>
<td>90</td>
</tr>
<tr>
<td>EC_card</td>
<td>0</td>
<td>88</td>
</tr>
<tr>
<td>EC_card</td>
<td>1</td>
<td>83</td>
</tr>
<tr>
<td>Income</td>
<td>. -&gt; 500</td>
<td>93</td>
</tr>
<tr>
<td>Income</td>
<td>500 -&gt; 1,550</td>
<td>81</td>
</tr>
<tr>
<td>Income</td>
<td>1,550 -&gt; 1,850</td>
<td>75</td>
</tr>
<tr>
<td>Income</td>
<td>1,850 -&gt; 2,550</td>
<td>90</td>
</tr>
<tr>
<td>Income</td>
<td>2,550 -&gt; .</td>
<td>88</td>
</tr>
<tr>
<td>Status</td>
<td><em>Income Type</em></td>
<td>79</td>
</tr>
</tbody>
</table>

NUST Collaboration with RMC Consultants
Reject Inference

• What is Reject Inference

• Methods to adjust rejects applicants
  – Assign all rejects to Bads
  – Ignore the rejects altogether
  – Approve all applications
Reject Inference (Cont’d)

- Augmentation Techniques
  - Simple augmentation
  - Parcelling
  - Fuzzy augmentation
Simple Augmentation
Simple Augmentation (Cont’d)
Fuzzy Augmentation
Fuzzy Augmentation (Cont’d)
### Rejection Inference Using Parceling

<table>
<thead>
<tr>
<th>Score</th>
<th># Bad</th>
<th># Good</th>
<th>% Bad</th>
<th>% Good</th>
<th>Reject</th>
<th>Rej - Bad</th>
<th>Rej - Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–169</td>
<td>290</td>
<td>971</td>
<td>23.0%</td>
<td>77.0%</td>
<td>1,646</td>
<td>379</td>
<td>1,267</td>
</tr>
<tr>
<td>170–179</td>
<td>530</td>
<td>2,414</td>
<td>18.0%</td>
<td>82.0%</td>
<td>1,732</td>
<td>312</td>
<td>1,420</td>
</tr>
<tr>
<td>180–189</td>
<td>365</td>
<td>2,242</td>
<td>14.0%</td>
<td>86.0%</td>
<td>3,719</td>
<td>521</td>
<td>3,198</td>
</tr>
<tr>
<td>190–199</td>
<td>131</td>
<td>1,179</td>
<td>10.0%</td>
<td>90.0%</td>
<td>7,334</td>
<td>733</td>
<td>6,601</td>
</tr>
<tr>
<td>200–209</td>
<td>211</td>
<td>2,427</td>
<td>8.0%</td>
<td>92.0%</td>
<td>1,176</td>
<td>94</td>
<td>1,082</td>
</tr>
<tr>
<td>210–219</td>
<td>213</td>
<td>4,047</td>
<td>5.0%</td>
<td>95.0%</td>
<td>3,518</td>
<td>176</td>
<td>3,342</td>
</tr>
<tr>
<td>220–229</td>
<td>122</td>
<td>2,928</td>
<td>4.0%</td>
<td>96.0%</td>
<td>7,211</td>
<td>288</td>
<td>6,923</td>
</tr>
<tr>
<td>230–239</td>
<td>139</td>
<td>6,811</td>
<td>2.0%</td>
<td>98.0%</td>
<td>3,871</td>
<td>77</td>
<td>3,794</td>
</tr>
<tr>
<td>240–249</td>
<td>88</td>
<td>10,912</td>
<td>0.8%</td>
<td>99.2%</td>
<td>4,773</td>
<td>38</td>
<td>4,735</td>
</tr>
<tr>
<td>250+</td>
<td>94</td>
<td>18,706</td>
<td>0.5%</td>
<td>99.5%</td>
<td>8,982</td>
<td>45</td>
<td>8,937</td>
</tr>
</tbody>
</table>

NUST Collaboration with RMC Consultants
Adverse Code

- What is adverse code
- Adverse Scoring
- WOE is zero

\[- \left( \frac{a}{n} \right) \times \text{factor} + \frac{\text{offset}}{n} \]

**Reasons for Decline with Neutral Score**

<table>
<thead>
<tr>
<th>Scorecard</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>56</td>
</tr>
<tr>
<td>Time at Address</td>
<td>43</td>
</tr>
<tr>
<td>Postal Code</td>
<td>22</td>
</tr>
<tr>
<td>Inquiries 3 Mths</td>
<td>20</td>
</tr>
<tr>
<td>% Trades Delinquent</td>
<td>43</td>
</tr>
<tr>
<td>Oldest Trade</td>
<td>68</td>
</tr>
<tr>
<td>Debt Service Ratio</td>
<td>42</td>
</tr>
<tr>
<td>Utilization</td>
<td>25</td>
</tr>
<tr>
<td>Worst Rating</td>
<td>30</td>
</tr>
<tr>
<td>Neutral Score</td>
<td>31</td>
</tr>
</tbody>
</table>
Neutral Score Using Weighted Average Approach

• What are neutral and weighted average scores
• Calculation of weighted average scores
• Ranking of adverse codes

\[ \sum_{i=1}^{n} (\text{Distribution}_i \times \text{score}_i) \]

<table>
<thead>
<tr>
<th>Time at Res</th>
<th>Distribution</th>
<th>Score</th>
<th>(D \times S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–6</td>
<td>18%</td>
<td>12</td>
<td>2.16</td>
</tr>
<tr>
<td>7–18</td>
<td>32%</td>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>19–36</td>
<td>26%</td>
<td>28</td>
<td>7.28</td>
</tr>
<tr>
<td>37+</td>
<td>24%</td>
<td>40</td>
<td>9.6</td>
</tr>
<tr>
<td>Weighted Average</td>
<td></td>
<td></td>
<td>27.04</td>
</tr>
</tbody>
</table>

NUST Collaboration with RMC Consultants
Logical Distribution of Points Allocation

<table>
<thead>
<tr>
<th>Age</th>
<th>Weight</th>
<th>Scorecard 1</th>
<th>Scorecard 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td>-55.50</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>18-22</td>
<td>-108.41</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>23-26</td>
<td>-72.04</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>27-29</td>
<td>-3.95</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>30-35</td>
<td>70.77</td>
<td>35</td>
<td>38</td>
</tr>
<tr>
<td>35-44</td>
<td>122.04</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>44+</td>
<td>165.51</td>
<td>51</td>
<td>52</td>
</tr>
</tbody>
</table>

NUST Collaboration with RMC Consultants
Choosing a Scorecard

- Development of more than one scorecard
- Which scorecard is best?
- How good is the scorecard?

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Actual</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>True Positive</td>
</tr>
<tr>
<td>Bad</td>
<td>False Positive</td>
</tr>
</tbody>
</table>
## Final scorecard with AGB

### Application scorecard example

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Attributes</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Years @ address</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;3 years</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>3–6 years</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>&gt;6 years</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Blank</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td><strong>Years @ employer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;2 years</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>2–8 years</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>9–20 years</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>&gt;20 years</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Blank</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td><strong>Home phone</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Not Given</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td><strong>Accom. status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Own</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Rent</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Parents</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td><strong>Bankers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Us</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Them</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Blank</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td><strong>Credit card</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bank or travel</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>Retail or garage</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Blank</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td><strong>Judgments on bureau</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>–16</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>–30</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>–54</td>
<td></td>
</tr>
<tr>
<td><strong>Past experience</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>New</td>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>Up-to-date</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Arrears</td>
<td>3</td>
<td>Reject</td>
</tr>
<tr>
<td>Write-off</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Reject</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final score</strong></td>
<td>267</td>
<td></td>
</tr>
</tbody>
</table>
Questions