Credit Risk - Predictive Modelling

4EK614

28 April 2021
With You Today

Radek Lastovicka
Senior Manager in Credit Risk Team
Radek.Lastovicka@cz.ey.com

Jan Nusko
Senior Consultant in Credit Risk Team
Jan.Nusko@cz.ey.com

Credit Risk Team

Risk Parameters
Impairment Loss

AQR

Regulatory

Our services

Our projects

Model Development
Model Validation

Methodological Reviews
Asset Quality Reviews

Data Analysis
Data Mining

LIC©
Advanced Analytics

Stress Testing
Impairment

Regulatory Reporting
Business Intelligence

Page 2 Credit Risk - Predictive Modelling
About This Seminar

Course Structure

Day 1: Credit Risk, Underwriting Process, Predictive Modelling
Day 2: Market Risk

Classroom: Microsoft Teams
Time: 9:15 - 12:30

Course Assessment

1. Case study - you can choose market risk and/or credit risk topics:
   a) Credit Risk - Preparation of PD scorecard:
      a) Prepare development sample from portfolio of mortgage loans
      b) Model scorecard using logistic regression (or any technique you want!) and include assessment
   b) Market Risk - TBA

2. Outputs - PPT presentation or PDF, summarizing the abovementioned outputs, and scripts used.

3. Output presentation - short (10-15 minute) presentation about results of this assessment.

Prerequisites

1. Basic understanding of statistical and mathematical concepts
2. Elementary knowledge of programming (Python, R, ...)

Study materials

1. PowerPoint slides, provided after the course

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials

Study materials
Agenda

1. Credit Risk & Banking  W 09:15-09:50
2. Underwriting & Scoring W 09:50-10:25
3. Predictive Modelling W 10:30-11:15
4. Scorecard Development W 11:15-11:50
5. Model Assessment W 11:50-12:20
6. Q&A W 12:20-12:30

Operative

Don’t hesitate to ask or comment at any point
Especially since it’s just virtual 😊
We recommend teams for case study
Menti.com - 14 13 31 65
Your task is to build a PD scorecard using the provided data. The goal is to create a model that will predict a probability of default for each mortgage.

The presentation contains an overview of a proposed modelling process and some considerations to consider when developing and assessing the model.

You will be assessed on the “good modelling practice” you employ. Remember, the best model is not necessarily the one with the highest performance metric. Your goal should be to build a scorecard with enough discriminatory power, but the steps taken during the modelling process are most important.

Resources – please write an email to jan.nusko@cz.ey.com

- Mortgage_sample.csv: Modelling dataset with data about 50000 US mortgages
- Mortgage_metadata.xlsx: Data dictionary

Package suggestions:
- Python - scorecardpy
- R - scorecard
Banks

Credit Risk - Predictive Modelling
Balance sheet and off-balance sheet of a bank

<table>
<thead>
<tr>
<th>Balance sheet</th>
<th>Assets</th>
<th>Liabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cash</td>
<td>Deposits from customers</td>
</tr>
<tr>
<td></td>
<td>Deposits at central bank</td>
<td>Loans from other banks</td>
</tr>
<tr>
<td></td>
<td>Loan</td>
<td>Securities</td>
</tr>
<tr>
<td></td>
<td>Loans to other banks</td>
<td>Hybrid instruments</td>
</tr>
<tr>
<td></td>
<td>Securities</td>
<td>Other liabilities</td>
</tr>
<tr>
<td></td>
<td>Other assets</td>
<td>Equity</td>
</tr>
<tr>
<td>Off-balance sheet</td>
<td>Undrawn limits of credit lines</td>
<td>Undrawn limits of credit lines</td>
</tr>
<tr>
<td></td>
<td>Loan commitments</td>
<td>Guarantees received</td>
</tr>
<tr>
<td></td>
<td>Guarantees given</td>
<td>Derivatives</td>
</tr>
<tr>
<td></td>
<td>Derivatives</td>
<td></td>
</tr>
</tbody>
</table>
What is credit risk?

- The risk that a counterparty fails to meet a contractual obligation

### Banking book
- Retail: mortgages, credit cards
- Corporate: Investment property financing, project financing, large corporate lending
- Wholesale: Lending to banks & sovereigns

### Trading book
- Counterparty credit risk (CCR): whenever a trade is settled in the future and/or is not “delivery versus payment” (DvP), a firm takes on credit risk

### Insurance
- Reinsurer default
- Corporate bond / ABS default / CDS
- Derivative counterparties

### Other
- Intermediary: Default on commissions receivable
- Accounts receivable: Non payment of invoice
Components of credit risk

- **PD**
  - Probability of Default: The likelihood the borrower will default on its obligation either over the life of the obligation.

- **LGD**
  - Loss Given Default: Loss that lender would incur in the event of borrower’s default. It is the exposure that cannot be recovered through bankruptcy proceedings, collateral recovery or some other form of settlement. Usually expressed as a percentage of exposure at default.

- **EAD**
  - Exposure at Default: The exposure that the borrower would have at default. Takes into account both on-balance sheet (capital) and off-balance sheet (unused lines, derivatives or repo transactions) exposures and payment schedule.

**Expected Credit Loss (ECL)** = PD x LGD x EAD
Credit risk agenda

**Governance**
- Risk management function reshaping roadmap
- Credit risk strategy and linkage to business strategy
- Risk appetite framework and statements
- Credit risk processes and segregation of duties
- Model governance framework (model request, design implementation, validation)
- Stress testing framework

**Collection services**
- Diagnostics on the effectiveness & efficiency of the collections process
- Development of a collections strategy, strategic and tactical (cost-benefit) analysis of available outsourcing options
- Design of a collections framework
- Support with collections technology requirements analysis, selection and implementation of an appropriate solution

**Application process**
- Application scoring
- Rating models
- Provisioning
- LGD models

- Business model request specification
- Application scorecard design and validation
- Design and review of the application processes
- Support with application workflow technology

- Model design / validation / internal audit reviews
- Regulatory compliance
- PD estimation
- Model usage for business purposes

- Design of impairment methodology in line with IFRS
- Effective interest rate and recognitions of fees and commissions
- Back-testing analyses
- Proprietary IT tools

- LGD estimates design and validation
- LGD (scoring) models design and validation
- LGD data warehouse specification
- Collateral valuation scenarios
Underwriting process
Underwriting process

- Underwriting (UW) process is the processing of credit application and making a decision about the final approval or decline of the application.

- Generally the UW process can end up in several different states: approval, decline, cancelation from client side, non-eligibility (for example the applicant is not meeting minimum age criteria, etc.)
Underwriting process

• Client segment is a crucial parameter to the UW process and scoring

<table>
<thead>
<tr>
<th>Private individuals</th>
<th>Entrepreneurs Freelancers</th>
<th>Small business</th>
<th>Corporates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usually automated process</td>
<td>Usually automated process with possibly manual inputs</td>
<td>Partially automated process, but mostly manual assessment</td>
<td>Typically manual assessment on yearly basis (rating process using financial, qualitative and behavioral scoring)</td>
</tr>
<tr>
<td>Scoring applications in order to assess riskiness of newly issued loans/credits</td>
<td>Scoring applications in order to assess riskiness of newly issued loans/credits</td>
<td>Scoring applications for automated products</td>
<td>Sometimes not sufficient data to use statistical approach - especially in case of project financing</td>
</tr>
<tr>
<td>Scoring client behavior on monthly basis on credit and deposit products</td>
<td>Scoring client behavior on monthly basis</td>
<td>Process for manual yearly rating (typically financial scoring, qualitative scoring and behavioral scoring)</td>
<td>Industry dependent and seasonal</td>
</tr>
<tr>
<td>Large data sets → statistical approach</td>
<td>Large data sets → statistical approach</td>
<td>Sufficient data sets for statistical approach</td>
<td>Credit registers (CRÚ, Cribis, Bisnode, etc.)</td>
</tr>
<tr>
<td>Need to verify income and over-indebtedness</td>
<td>No need to verify income and over-indebtedness</td>
<td>Credit registers (CRÚ, Cribis, Bisnode, etc.)</td>
<td></td>
</tr>
<tr>
<td>Credit registers (BRKI, NRKI, Solus)</td>
<td>Credit registers</td>
<td>Credit registers (CRÚ, Cribis, Bisnode, etc.)</td>
<td></td>
</tr>
</tbody>
</table>
Underwriting process

- Underwriting process differs significantly for different products

<table>
<thead>
<tr>
<th>Product</th>
<th>Financing housing needs</th>
<th>Subject to consumer protection</th>
<th>Requires real estate collateral and insurance</th>
<th>Large financed amount</th>
<th>Typically longer maturity</th>
<th>More thorough and detailed UW process</th>
<th>Partially manual assessment</th>
<th>Loan to value condition</th>
<th>Lower interest rates</th>
<th>Fixation periods</th>
<th>Co-applicants possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortgage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer loan</td>
<td>Purpose or non-purpose</td>
<td>Subject to consumer protection</td>
<td>Can have collaterals or guarantors, but usually it doesn’t</td>
<td>Automated, easy and fast UW process</td>
<td>Higher interest rates</td>
<td>Co-applicants possible, but not that frequent as for mortgages</td>
<td>Medium financed amount</td>
<td>Medium maturity</td>
<td>Medium risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credit card, Overdraft and Revolving</td>
<td>Credit limit that can be utilized, but it is not a must</td>
<td>Client can flexibly utilize whatever part of the limit he needs to</td>
<td>Grace period</td>
<td>High interest rates</td>
<td>Typically no collaterals</td>
<td>Lower financed amount</td>
<td>Maturity is not specified (contract terminates on request when fully repaid)</td>
<td>High risk</td>
<td>Credit cards come with plastic card</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investment loan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Credit Risk – Predictive Modelling
Underwriting process

• First step in the process is the assessment of client general eligibility
  • Is the client over 18 years old?
  • Is the client eligible to sign contracts?
  • Is the client on the international sanction list?
  • Is the client a politically exposed person?
  • Has the client a tax domicile in the same country?
  • Does the client agree with all the legally required actions (credit bureau request, information protection principles, general terms and conditions, pre-contractual information, etc.)?

• Second step is the assessment of client eligibility for the given product and channel
  • Is the client below prescribed age when applying for a long term product such as mortgage?
  • Does the client have eligible income for the particular product and process?
  • Does the client have all prescribed documents (valid ID card and valid second ID document)?
  • Is the collateral for the issued loan eligible and sufficient (LTV threshold)?
Underwriting process

- There are several laws and directives that affect the underwriting process

  - Law on consumer loan
  - Consumer protection
  - Mortgage credit directive (MCD)
  - Consumer credit directive (CCD)
  - EBA guidelines
  - Basel Capital Accord
  - Anti-money laundering (AML)

  Consumer needs to be protected from dishonest and malicious practices including intentional over-indebting, but also non-intentional over-indebting - the responsibility of not over-indebting the client is now on the borrower.

  Market and economy needs to be protected against adverse economic impacts originating in the financial system.

  Society needs to be protected against criminal acts and terrorism.

  Consumer needs to be protected against loosing his money deposited in a bank by irresponsible lending and crediting banks clients.
Underwriting process

- Client authentication
- Anti-fraud module

- Expiry date check - ID not expired
- Check on validity in MPSV database
- Issue date consistency check (based on linear regression below)
- Check on issue date - not week-end or public holiday
- Check on address at MěÚ or OÚ
- Control on ID manipulation (color histogram, fonts)
- Check on consistency of bar-code and ID number
- Consistency of sex and birth number (third digit)
- Birth date divisible by 11 after 1953
- Overall control number check
- Expiry date control number check
- Birth date control number check

![Image of identification card]

![Image of another identification card]

![Graph showing linear regression]

Credit Risk - Predictive Modelling
Underwriting process

• Internal blacklists on phone numbers, ID cards, IČO of employers, ready-made companies

• Frequency checks in on-line underwriting process (applications are tracked with respect to different identificators and their combinations)

• Device fingerprint (publicly available libraries)
  - Phone number
  - Account number
  - ID card number
  - E-mail address
  - Birth number
  - IP address
  - Geolocation (via IP address and Google API) - can be used for anti-fraud as well as for scoring

• Checks on discrepancy between past applications with the same identifiers
Underwriting process

• **Individuals / Entrepreneurs:**
  - **BRKI – Banking Register of Client Information**
    - Information about applications and loan contracts shared among the banks operating in Czech Republic. Generally only banks can access it.
    - Information is stored in BRKI during the existence of credit relationship and 4 years after it terminates. If the contract with the bank has not been signed is this information in BRKI stored for one year.
  - **NRKI – Non-Banking Register of Client Information**
    - Information about applications and loan contracts shared among non-bank credit providers. Generally only those that participate on the sharing can access it.
  - **SOLUS**
    - Information about applications and loan contracts shared among participating credit providers and some other companies. Generally only those that participate on the sharing can access it. It contains both – register of negative as well as register of positive information.
    - In SOLUS participate also TELCO companies and utility providers.

• **Companies / Entrepreneurs:**
  - **CRÚ – Kreditní Registr Úvěrů**
    - Information about loan contracts of entrepreneurs and companies – compulsory register operated by Czech National Bank.
Scoring is one of the tools to measure the creditworthiness of a business or person. It is the result of scoring, where different scales are given different weight. This procedure results in a credit score.
Predictive modelling
Predictive Modelling - Model life-cycle

Model lifecycle steps

1. Model request
   - Business line Management

2. Model development
   - Management
   - Risk

3. Model implementation & usage
   - IT & data
   - Loan approval
   - Capital requirement calculation

4. Model monitoring & review
   - Monitoring
   - Independent review
   - Internal audit

Credit Risk - Predictive Modelling
Problem to be solved:

\[ T(f(\text{age}_i, \text{status}_i, \text{education}_i, \text{housing}_i, \ldots) - 1[i \text{ defaulting in } 1 \text{ year from snapshot }]) \]

is in some sense minimized

the value \( f(.) \) we call probability of default (PD)

Can be numeric, ordinal, nominal or even missing

Attains only values 0, or 1
Predictive Modelling - Discrimination
Predictive Modelling - Workflow

• 1) Data exclusions
• 2) Missing values analysis
• 3) Outlier treatment
• 4) Variable transformation (feature engineering)
• 5) Univariate analysis
• 6) Correlation analysis
• 7) Modelling
  ▪ Selection of shortlist of variables
  ▪ Estimation of coefficients based

Name | Age | Status | Education | Housing | Target
---|---|---|---|---|---
Adam | 29 | single | high school | rent | 0
Annie | 27 | single | elementary | with parents | 1
Jane | 31 | single | high school | own house | 0
John | 30 | married | university | mortgage | 0
Since we will be using a regressive approach, we need to keep in mind that we cannot have dependent observations.

To avoid this, a cohort approach is used:

- Flexible cohort - fixed number of snapshots after first observation
- Fixed cohort - fixed snapshot date (e.g. from every September)

For our target, we define a “performance” window - usually 12 months

No balancing needed ;)
- Unless we’re talking about LDP portfolios
Predictive Modelling - Linear regression

- Problem to be solved:

We choose linear function

\[ f(\vec{x}) = \alpha + \sum_{j=1}^{k} \beta_j x_j \]

Historical data with already known target value

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Status</th>
<th>Education</th>
<th>Housing</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>29</td>
<td>single</td>
<td>high school</td>
<td>rent</td>
<td>0</td>
</tr>
<tr>
<td>Annie</td>
<td>27</td>
<td>single</td>
<td>elementary</td>
<td>with parents</td>
<td>1</td>
</tr>
<tr>
<td>Jane</td>
<td>31</td>
<td>single</td>
<td>high school</td>
<td>own house</td>
<td>0</td>
</tr>
<tr>
<td>John</td>
<td>30</td>
<td>married</td>
<td>university</td>
<td>mortgage</td>
<td>0</td>
</tr>
</tbody>
</table>

29 years

single

own house

university degree

Credit Risk - Predictive Modelling
Predictive Modelling - Logistic regression

Problem to be solved:

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Status</th>
<th>Education</th>
<th>Housing</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>29</td>
<td>single</td>
<td>high school</td>
<td>rent</td>
<td>0</td>
</tr>
<tr>
<td>Annie</td>
<td>27</td>
<td>single</td>
<td>elementary</td>
<td>with parents</td>
<td>1</td>
</tr>
<tr>
<td>Jane</td>
<td>31</td>
<td>single</td>
<td>high school</td>
<td>own house</td>
<td>0</td>
</tr>
<tr>
<td>John</td>
<td>30</td>
<td>married</td>
<td>university</td>
<td>mortgage</td>
<td>0</td>
</tr>
</tbody>
</table>

29 years

Historical data with already known target value

We choose logistic function

\[ f(\mathbf{x}) := \frac{1}{1 + e^{-\alpha - \sum_{j=1}^{k} \beta_j x_j}} \]
Predictive Modelling - Logit

- We can choose other functions, but market standard is to use the logit link function

- Using linear function is not proper as it can give estimates above 1 or below 0, which is not convenient for estimating probability of default

- Selection of the link function if it preserves the output between 0 and 1

- The reason for choosing logit function instead of others is mainly interpretational - the log-odds ratio defined below is a linear combination of the predictors

\[
\text{Log - odds ratio} = \ln \left( \frac{PD}{1 - PD} \right) = f^{-1}(PD)
\]

- By central limit theorem under very general conditions the log-odds ratio distribution converges in distribution to a normal distribution
Let’s say we have processed our data (deduplication, formatting, primary keys, consistency checks...)

We could take advantage of models with some sort of elimination

E.g. - Lasso regression

- Least absolute shrinkage and selection operator
- Performs both variable selection and regularization

Out of 649 parameters, 192 were set exactly to zero and the obtained lasso model has 457 parameters. Among these parameters, 139 have estimated values in absolute value greater or equal to 0.1. We will present 20 coefficients with highest absolute values.

Is this a good model?
Scorecard points (score)

The motivation is to derive a scale such that:
- It's a linear combination of log-odds ratio
- More score points means lower PD
- Double odds ratio corresponds to a prescribed number of score points $A$:

$$A = \alpha + \beta \ln \left( \frac{2 \cdot PD_i}{1 - PD_i} \right) - \alpha - \beta \ln \left( \frac{PD_i}{1 - PD_i} \right) = \beta \ln 2 \Rightarrow \beta = \frac{A}{\ln 2}$$

$B$ score points corresponds to a prescribed PD value $x$:

$$B = \alpha + \frac{A}{\ln 2} \ln \left( \frac{x}{1 - x} \right) \Rightarrow \alpha = B - \frac{A}{\ln 2} \ln \left( \frac{x}{1 - x} \right)$$

$A$ is usually set to be 50 score points
$B$ is usually set to 500 score points and the corresponding $x = \frac{1}{51}$
Another standardly used technique is binning of predictors and WoE transformation:

- Weight of evidence for i-th bin: $WoE_i = \ln \left( \frac{GOODS_i}{BADS_i} \right)$

---

<table>
<thead>
<tr>
<th>BIN</th>
<th>GOODS</th>
<th>BADS</th>
<th>DR</th>
<th>WoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[−inf, 33)</td>
<td>69</td>
<td>52</td>
<td>0.429752</td>
</tr>
<tr>
<td>2</td>
<td>[33, 37)</td>
<td>63</td>
<td>45</td>
<td>0.416667</td>
</tr>
<tr>
<td>3</td>
<td>[37, 40)</td>
<td>72</td>
<td>47</td>
<td>0.394958</td>
</tr>
<tr>
<td>4</td>
<td>[40, 46)</td>
<td>172</td>
<td>89</td>
<td>0.340996</td>
</tr>
<tr>
<td>5</td>
<td>[46, 48)</td>
<td>59</td>
<td>25</td>
<td>0.297619</td>
</tr>
<tr>
<td>6</td>
<td>[48, 51)</td>
<td>99</td>
<td>41</td>
<td>0.292857</td>
</tr>
<tr>
<td>7</td>
<td>[51, 58)</td>
<td>157</td>
<td>62</td>
<td>0.283105</td>
</tr>
<tr>
<td>8</td>
<td>[58, inf)</td>
<td>93</td>
<td>25</td>
<td>0.211864</td>
</tr>
<tr>
<td>9</td>
<td>MISSING</td>
<td>19</td>
<td>11</td>
<td>0.366667</td>
</tr>
</tbody>
</table>

- Why binning? solves leverage points, solves informative missings, solves non-numerical (either ordinal or multinomial) variables, assesses robustness
- Why WoE transformation? normalizes predictors values, enables easy interpretation (under reasonable conditions always attains negative and positive values, zero value represents portfolio default rate)
Predictive Modelling - WoE

WoE is the new value of binned predictor

Coefficient is the estimated parameter from logistic regression corresponding to the variable or to the absolute term (intercept)

In case number in some bin is zero, we need to compensate:

\[
WoE = \ln \left( \frac{(BADS + 0.5)/(GOODS + 0.5)}{BADS/GOODS} \right)
\]

Missing category can be treated

Scorecard points serve as a standardized linear transformation of log-odds so that certain criteria are met - it is motivated mainly by interpretation

Coefficients should be negative when using WoE

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Group</th>
<th>Scorecard Points</th>
<th>WoE</th>
<th>DR</th>
<th>Percentage of population</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>25</td>
<td></td>
<td>-3.6578</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;25</td>
<td>0</td>
<td>-1.0109</td>
<td>18.1%</td>
<td>16.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;35</td>
<td>27</td>
<td>-0.4535</td>
<td>11.2%</td>
<td>23.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;55</td>
<td>83</td>
<td>0.7352</td>
<td>3.7%</td>
<td>33.9%</td>
<td></td>
<td>-0.6572</td>
</tr>
<tr>
<td>&gt;=55</td>
<td>111</td>
<td>1.3272</td>
<td>2.1%</td>
<td>25.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>50</td>
<td>0.0420</td>
<td>7.1%</td>
<td>1.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary</td>
<td>0</td>
<td>-1.0188</td>
<td>18.2%</td>
<td>6.9%</td>
<td></td>
<td>-0.8213</td>
</tr>
<tr>
<td>Vocational</td>
<td>18</td>
<td>-0.7154</td>
<td>14.1%</td>
<td>15.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>82</td>
<td>0.3762</td>
<td>5.2%</td>
<td>38.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>83</td>
<td>0.3877</td>
<td>5.2%</td>
<td>38.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>85</td>
<td>0.4215</td>
<td>0.0%</td>
<td>0.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Housing type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.7765</td>
</tr>
<tr>
<td>With parents</td>
<td>0</td>
<td>-0.8902</td>
<td>16.3%</td>
<td>10.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rent</td>
<td>2</td>
<td>-0.8489</td>
<td>15.8%</td>
<td>13.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooperative</td>
<td>44</td>
<td>-0.1144</td>
<td>8.3%</td>
<td>21.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortgage</td>
<td>105</td>
<td>0.9786</td>
<td>2.9%</td>
<td>42.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Own</td>
<td>82</td>
<td>0.5794</td>
<td>4.3%</td>
<td>9.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>79</td>
<td>0.5216</td>
<td>0.0%</td>
<td>1.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.5903</td>
</tr>
<tr>
<td>Single</td>
<td>0</td>
<td>-0.8118</td>
<td>15.3%</td>
<td>29.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>85</td>
<td>1.1710</td>
<td>2.4%</td>
<td>21.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divorced</td>
<td>47</td>
<td>0.2865</td>
<td>5.7%</td>
<td>36.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Widowed</td>
<td>76</td>
<td>0.9634</td>
<td>3.0%</td>
<td>10.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>76</td>
<td>0.8736</td>
<td>0.0%</td>
<td>1.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total scorecard points = Intercept scorecard points + \( \sum_{i=1}^{m} \text{Variable scorecard points} \)
Model performance - GINI

- ROC (Receiver Operation Characteristics) curve, GINI
- Measuring discriminatory power - only ordering matters, not the actual score values

**True positive rate**

If we sort the clients increasingly by the score value, the true positive rate can be calculated for the i-th observation as the number of observations with target=1 and index lower or equal to i divided by the total number of observations with target=1.

**False positive rate**

If we sort the clients increasingly by the score value, the false positive rate can be calculated for the i-th observation as the number of observations with target=0 and index lower or equal to i divided by the total number of observations with target=0.

### Example Table

<table>
<thead>
<tr>
<th>Client</th>
<th>Event</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annie</td>
<td>1</td>
<td>325</td>
</tr>
<tr>
<td>Paul</td>
<td>0</td>
<td>398</td>
</tr>
<tr>
<td>Lisa</td>
<td>1</td>
<td>415</td>
</tr>
<tr>
<td>Jane</td>
<td>0</td>
<td>463</td>
</tr>
<tr>
<td>Jack</td>
<td>1</td>
<td>499</td>
</tr>
<tr>
<td>Adam</td>
<td>0</td>
<td>520</td>
</tr>
<tr>
<td>John</td>
<td>0</td>
<td>611</td>
</tr>
<tr>
<td>Mary</td>
<td>0</td>
<td>672</td>
</tr>
</tbody>
</table>

### ROC Curve

- **AUC (Area Under Curve)** = \[
\frac{1}{3} \times \frac{1}{5} + \frac{2}{3} \times \frac{1}{5} + \frac{3}{3} \times \frac{3}{5} = \frac{4}{5} = 0.8
\]

- **GINI** = \[2 \times (\text{AUC} - 0.5)\] = 0.6

- GINI attains values between -1 and 1, but relevant are only values between 0 and 1.
- GINI=0 stands for theoretical random model (no predictive power).
- GINI=1 stands for perfectly discriminating model.
Somers’ D is a measure of association between two variables X and Y calculated as follows:

$$\text{Somers’ D}(X, Y) = \frac{N_C(X, Y) - N_D(X, Y)}{N_C(X, X) - N_D(X, X)}$$

- $N_C(X, Y)$ is the number of concordant pairs $(x_i, y_i)$ and $(x_j, y_j)$, i.e. either $x_i < x_j$ and $y_i < y_j$, or $x_i > x_j$ and $y_i > y_j$.
- $N_D(X, Y)$ is the number of dis-concordant pairs $(x_i, y_i)$ and $(x_j, y_j)$, i.e. either $x_i < x_j$ and $y_i > y_j$, or $x_i > x_j$ and $y_i < y_j$.

- Somers’ D is more general measure than GINI as it can be used on other than binary targets.
- For binary target and no ties in the independent variable it equals GINI.
- Clearly: $N_D(X, X) = 0$
Model performance - Somers’ D

Illustrative example (X = grade, Y = time spent studying):

<table>
<thead>
<tr>
<th>Time spend studying</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Bad</td>
</tr>
<tr>
<td>Good</td>
</tr>
</tbody>
</table>

Using the formula:

\[
\text{Somers' D}(X,Y) = \frac{N_C(X,Y) - N_D(X,Y)}{N_C(X,X) - N_D(X,X)}
\]

\[
\text{Somers' D}(X,Y) = \frac{200 - 30}{400 - 0} = 0.4250
\]

\[
\text{Somers' D}(Y,X) = \frac{200 - 30}{390 - 0} = 0.4359
\]

Somers' D differentiates between independent and dependent variable.
Representativeness/Stability - PSI

- PSI (Population Stability Index) is a measure of difference between two discrete distributions.
- It is typically used in order to assess representativity - i.e. assess whether distribution of a binned variable differs in two different data samples which are typically from two different time periods (threshold of 0.2 is frequently used).

\[ PSI = \sum_{i=1}^{n} (\text{Actual}\%_i - \text{Expected}\%_i) \times \ln\left(\frac{\text{Actual}\%_i}{\text{Expected}\%_i}\right) \]

where \( n \) is the number of bins.

<table>
<thead>
<tr>
<th>Score bands</th>
<th>Actual %</th>
<th>Expected %</th>
<th>Ac-Ex</th>
<th>ln(Ac/Ex)</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 251</td>
<td>5%</td>
<td>8%</td>
<td>-3%</td>
<td>-0.470</td>
<td>0.014</td>
</tr>
<tr>
<td>251–290</td>
<td>6%</td>
<td>9%</td>
<td>-3%</td>
<td>-0.410</td>
<td>0.012</td>
</tr>
<tr>
<td>291–320</td>
<td>6%</td>
<td>10%</td>
<td>-4%</td>
<td>-0.510</td>
<td>0.020</td>
</tr>
<tr>
<td>321–350</td>
<td>8%</td>
<td>13%</td>
<td>-5%</td>
<td>-0.490</td>
<td>0.024</td>
</tr>
<tr>
<td>351–380</td>
<td>10%</td>
<td>12%</td>
<td>-2%</td>
<td>-0.180</td>
<td>0.004</td>
</tr>
<tr>
<td>381–410</td>
<td>12%</td>
<td>11%</td>
<td>1%</td>
<td>0.090</td>
<td>0.001</td>
</tr>
<tr>
<td>411–440</td>
<td>14%</td>
<td>10%</td>
<td>4%</td>
<td>0.340</td>
<td>0.013</td>
</tr>
<tr>
<td>441–470</td>
<td>14%</td>
<td>9%</td>
<td>5%</td>
<td>0.440</td>
<td>0.022</td>
</tr>
<tr>
<td>471–520</td>
<td>13%</td>
<td>9%</td>
<td>4%</td>
<td>0.370</td>
<td>0.015</td>
</tr>
<tr>
<td>520 &lt;</td>
<td>9%</td>
<td>8%</td>
<td>1%</td>
<td>0.120</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Population Stability Index (PSI) = 0.1269