Credit Risk - Predictive Modelling

4EK614

09 Feb 2022

Building a better working world

With You Today

Jan Nusko

Senior Consultant in Credit Risk Team

Jan.Nusko@cz.ey.com

Our services

Credit Risk Team

- Risk Parameters
- Impairment Loss
- AQR
- Regulatory

Our projects

- Model Development
- Model Validation
- Methodological Reviews
- Asset Quality Reviews
- Data Analysis
- Data Mining
- LIC(R), EPIC, CFE
- Advanced Analytics
- Stress Testing
- Impairment
- Regulatory Reporting
- Business Intelligence

About This Seminar

Course Structure

Day 1: Credit Risk, Market Risk, Climate Risk
Day 2: Predictive Modelling in Credit Risk
Day 3: Assignment + Seminar Data Walkthrough

Time: 9:15 - 15:15

Study materials

1. PowerPoint slides, provided after the course

Prerequisites

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

Course Assessment

1. Case study – 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

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.

Coursework

Goal

- 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

- Mortgage_sample.csv: Modelling dataset with data about 50000 US mortgages
- Mortagage_metadata.xlsx: Data dictionary
- Package suggestions:
  - Python - scorecardpy
  - R - scorecard
- jan.nusko@cz.ey.com is available for a 30 min consultation, please send him an email if interested.

Agenda Day 1

1. Intro & Admin 09:15-09:45
2. Banking & Credit Risk 09:45-10:30
3. Underwriting 10:40-11:15
4. Scoring 11:15-11:50
5. Lunch 11:50-13:00
6. Market Risk 13:00-14:30
7. ESG & Taxonomy 14:35-15:30

Operative

Don’t hesitate to ask or comment at any point
We recommend teams for case study
Menti.com – 4563 7579

Agenda Day 2

1. Predictive Modelling 09:15-10:00
2. Scorecard Development 10:10-11:00
4. Model Assessment 11:10-12:00
6. Q&A 12:00-12:30

Operative

Don’t hesitate to ask or comment at any point
We recommend teams for case study
Menti.com – 4563 7579

Agenda Day 3

1.	Assignment Intro	09:15-10:00
2.	Data Walkthrough	10:10-10:50
3.	Assignment Walkthrough	11:00-12:00
5.	Lunch	12:00-13:00

Operative

Don't hesitate to ask or comment at any point

We recommend teams for case study

Banks

Balance sheet and off-balance sheet of a bank

	Assets	Liabilities
Balance sheet	Cash Deposits at central bank **Loan** Loans to other banks Securities Other assets	**Deposits from customers** Loans from other banks Securities Hybrid instruments Other liabilities **Equity**
Off-balance sheet	Undrawn limits of credit lines Loan commitments Guarantees given Derivatives	Undrawn limits of credit lines Guarantees received Derivatives

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

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

Performing portfolio

Rating models

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

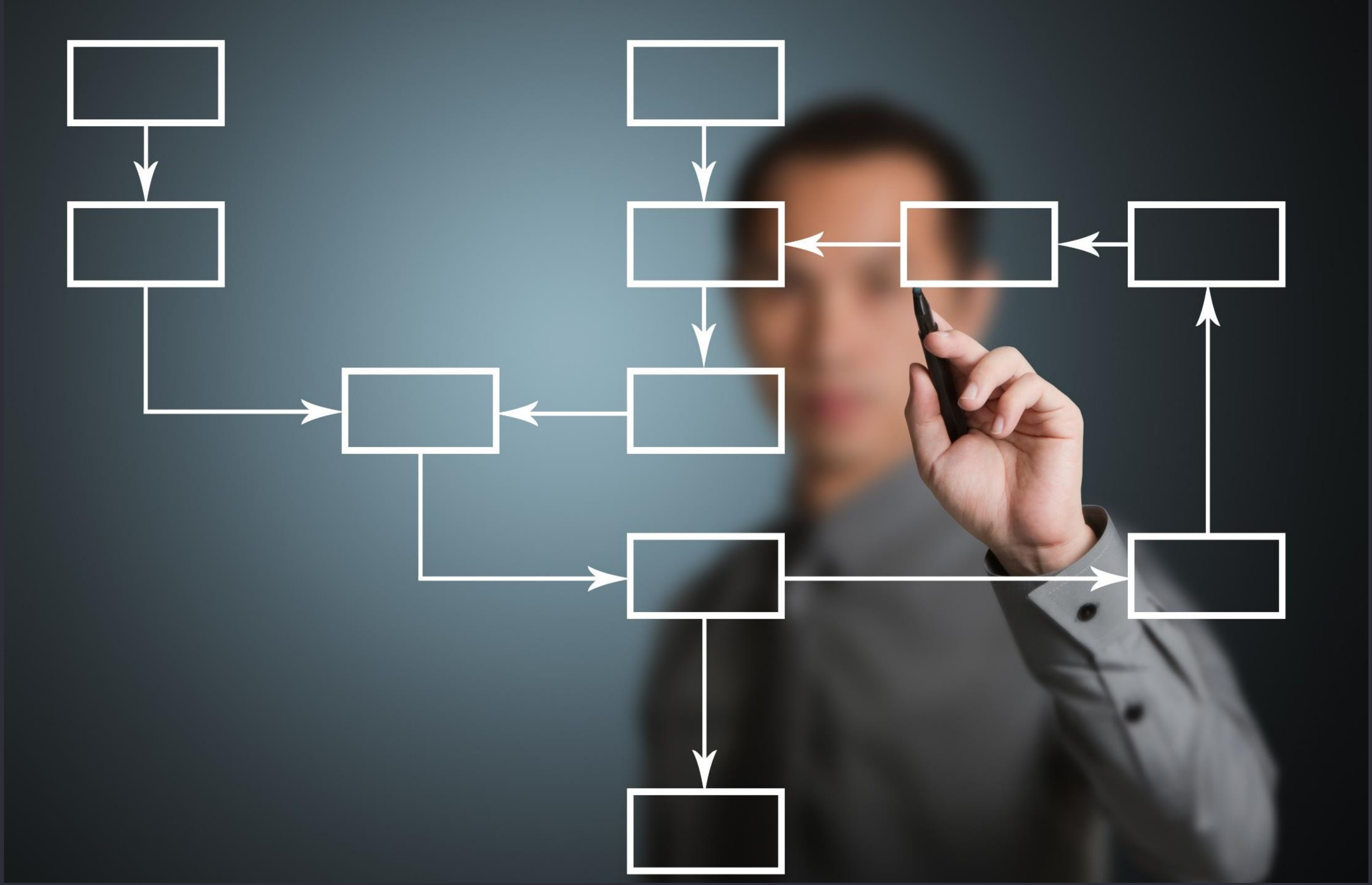
Provisioning

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

Non-performing portfolio

LGD models

- 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.)

Product parameters selection

Personal/ financial information

Client identification

Scoring and Overdebtedness verification

Decision / Contract signature

Underwriting process

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

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

Underwriting process

- Underwriting process differs significantly for different products

Mortgage	Consumer loan	Credit card, Overdraft and Revolving	Investment loan
• Financing housing needs • Subject to consumer protection • Requires real estate collateral and insurance • Large financed amount • Typically longer maturity • More thorough and detailed UW process • Partially manual assessment • Loan to value condition • Lower interest rates • Fixation periods • Co-applicants possible	• Purpose or non-purpose • Subject to consumer protection • Can have collaterals or guarantors, but usually it doesn't • Automated, easy and fast UW process • Higher interest rates • Co-applicants possible, but not that frequent as for mortgages • Medium financed amount • Medium maturity • Medium risk	• Credit limit that can be utilized, but it is not a must • Client can flexibly utilize whatever part of the limit he needs to • Grace period • High interest rates • Typically no collaterals • Lower financed amount • Maturity is not specified (contract terminates on request when fully repaid) • High risk • Credit cards come with plastic card	• Typical financing for corporate and small business segments, but also for entrepreneurs • Processed manually • Very high financed amount • Based on business and financial plan • Usually with collaterals and guarantees

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

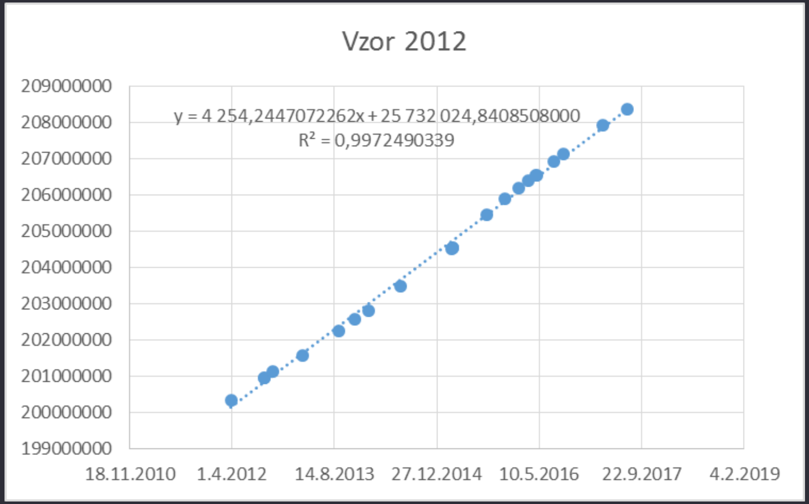
Society needs to be protected against criminal acts and terrorism

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



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)

Hardware: CPU architecture & device memory, GPU canvas, Audio stack
Software: User agent, OS version,
Storage: local storage, session storage
Display: color depth, screen size
Browser customizations: fonts, plug-ins, codecs, mime types, time zone, user language,
Miscellaneous: floating point calculations, callbacks / objects to DOM

  - 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

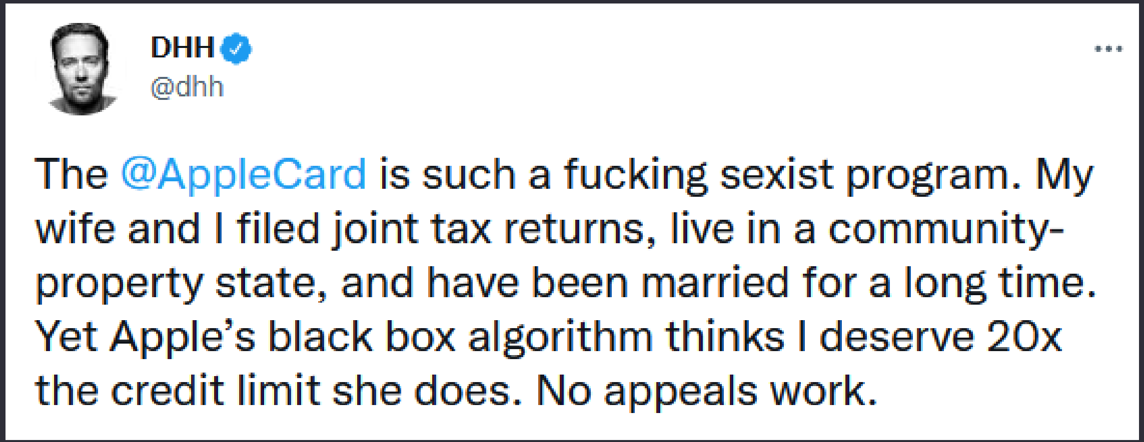
Underwriting process

- Individuals / Entrepreneurs:
  - <u>BRKI – Banking Register of Client Information</u>
    - 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.
  - <u>NRKI – Non-Banking Register of Client Information</u>
    - Information about applications and loan contracts shared among non-bank credit providers. Generally only those that participate on the sharing can access it.
  - <u>SOLUS</u>
    - 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:
  - <u>CRÚ – Kreditní Registr Úvěrů</u>
    - Information about loan contracts of entrepreneurs and companies – compulsory register operated by Czech National Bank.

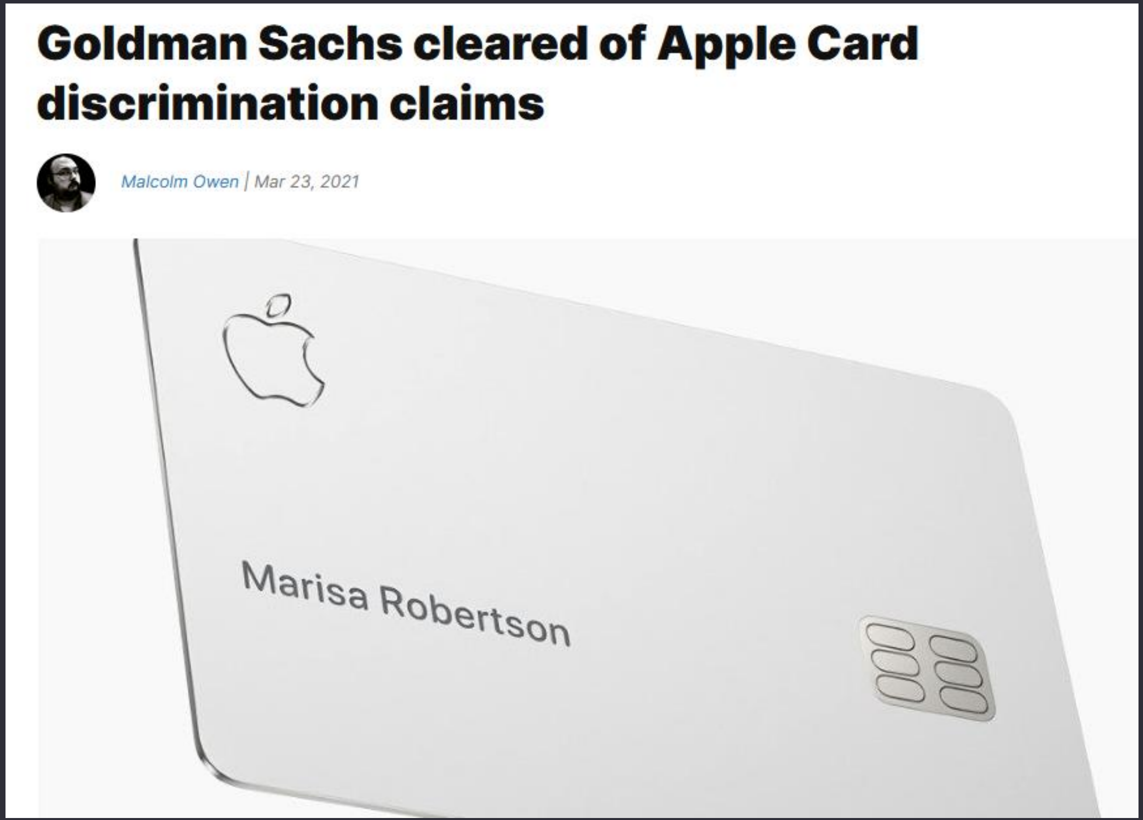
Scoring - Discrimination?

2017


DHH
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

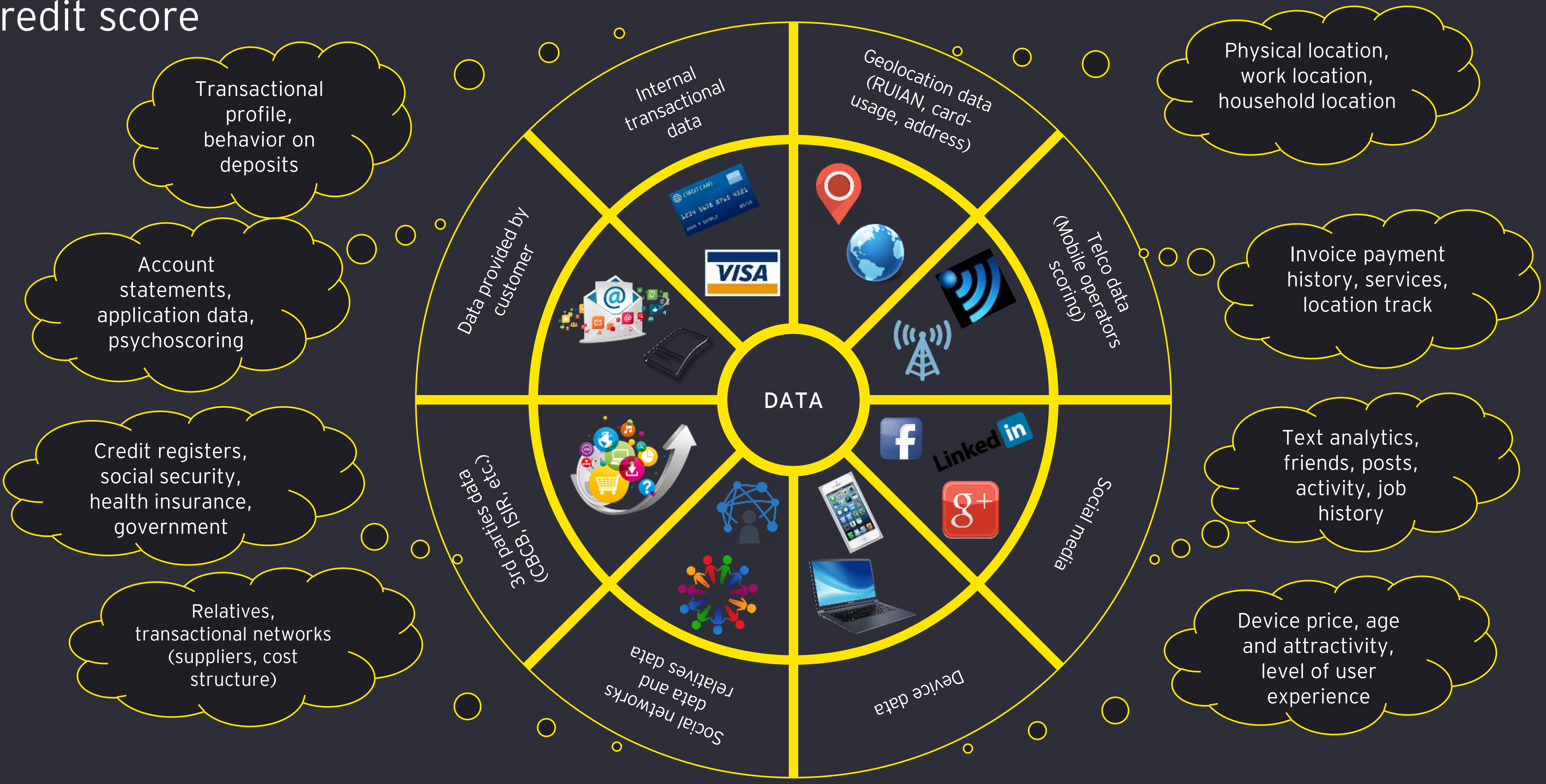
2021


Goldman Sachs cleared of Apple Card discrimination claims

Malcolm Owen | Mar 23, 2021

Underwriting process

- 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



Scoring - Goal

- Problem to be solved:

single

own house

university degree

29 years

i=1 i=2 i=3 i=4 .....

predictors

target

find a function f such that

$$\boxed{T(f(age_i, status_i, education_i, housing_i, \dots) - \mathrm{I}[i \text{ defaulting in 1 year from snapshot }])}$$

is in some sense minimized

Can be numeric, ordinal, nominal or even missing

Attains only values 0, or 1

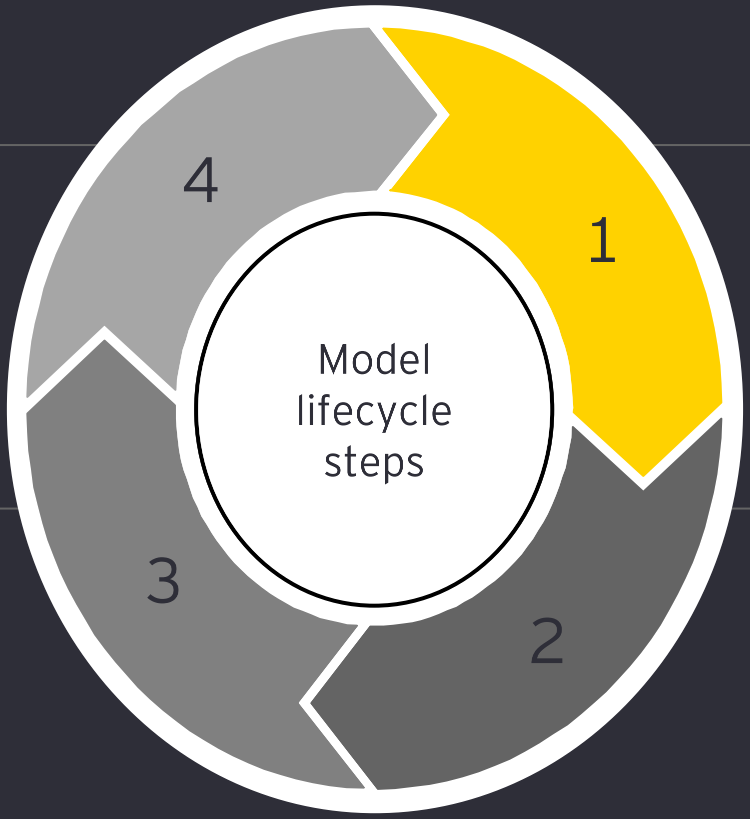
the value f(.) we call probability of default (PD)

Predictive modelling

Predictive Modelling - Model life-cycle

Model monitoring & review
Monitoring
Independent review
Internal audit

Model request
Business line
Management



Model implementation & usage
IT & data
Loan approval
Capital requirement calculation

Model development
Management
Risk

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

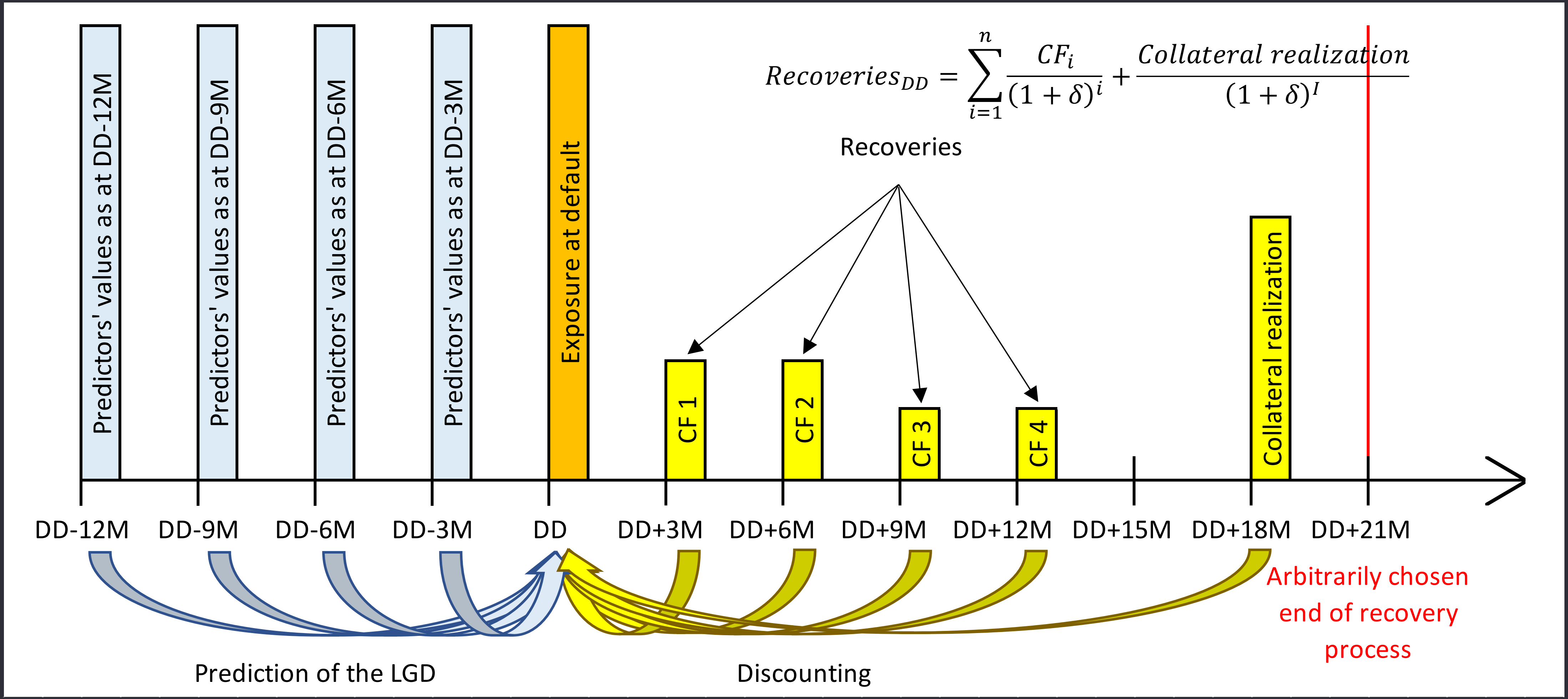
- Loss Given Default: Loss that lender would incur in the event of borrower's default. It is the exposure that cannot be recovered through bankruptcy proceedings, collateral recovery or some other form of settlement. Usually expressed as a percentage of exposure at default.

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.

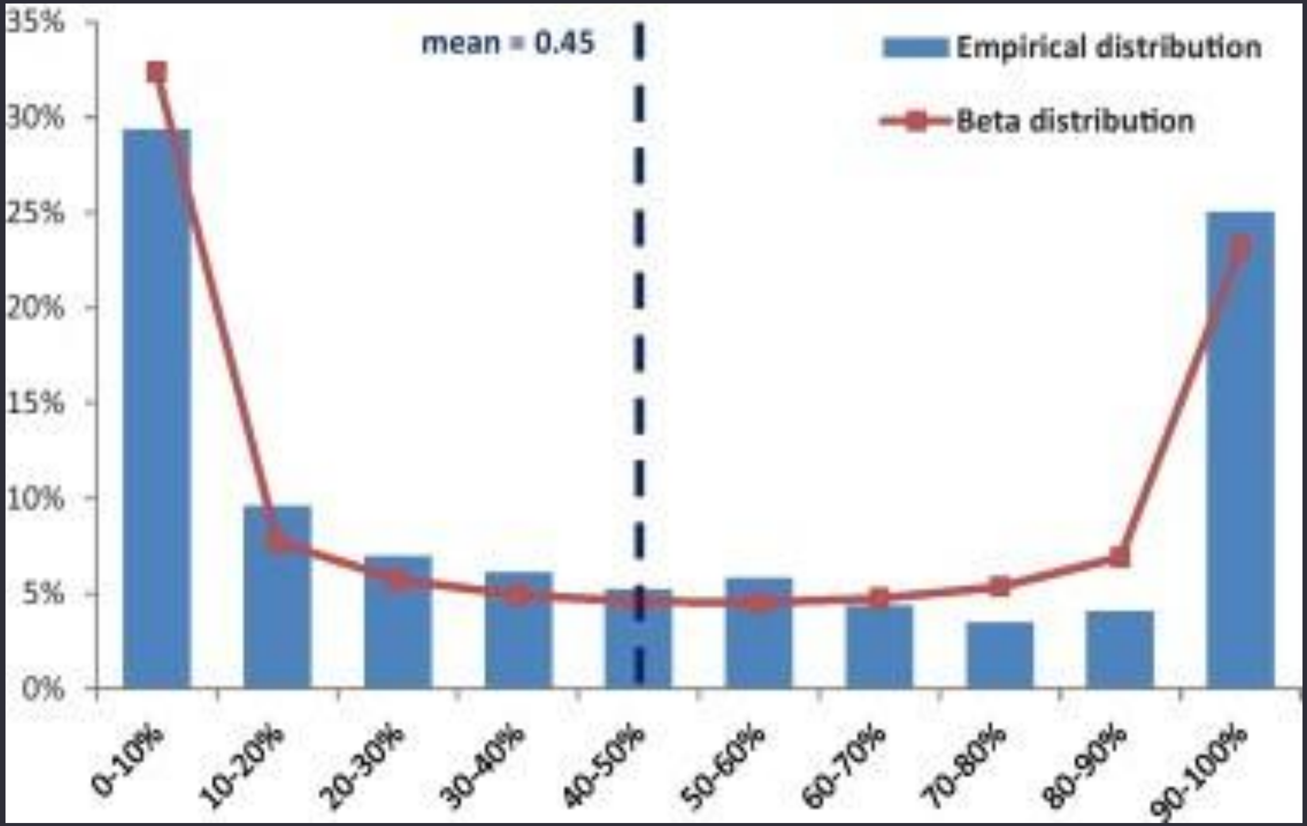
Expected Credit Loss (ECL) = PD x LGD x EAD

Loss given default



LGD models

- **"U-shape"**
- **It does not make sense to use average LGD** = 45% for these clients
- Real LGD is lower then 10% for the best 1/3 of the clients and higher then 90% for the worst 1/3 of the clients



Predictive Modelling - Goal

- Problem to be solved:

predictors

i=1 i=2 i=3 i=4 .....

target

find a function f such that

$$\boxed{T(f(age_i, status_i, education_i, housing_i, \ldots) - \mathrm{I}[i \text{ defaulting in 1 year from snapshot }])}$$

is in some sense minimized

Can be numeric, ordinal, nominal or even missing

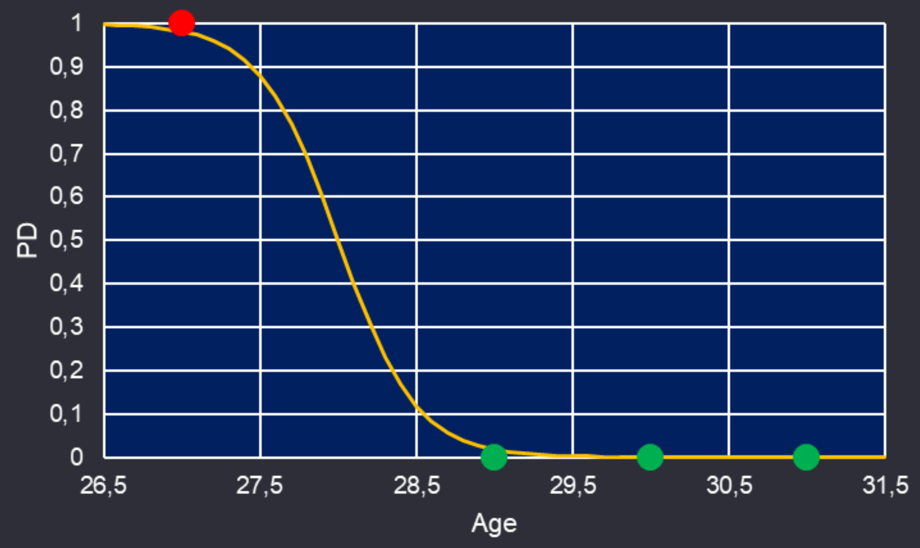
Attains only values 0, or 1

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

Predictive Modelling - Workflow

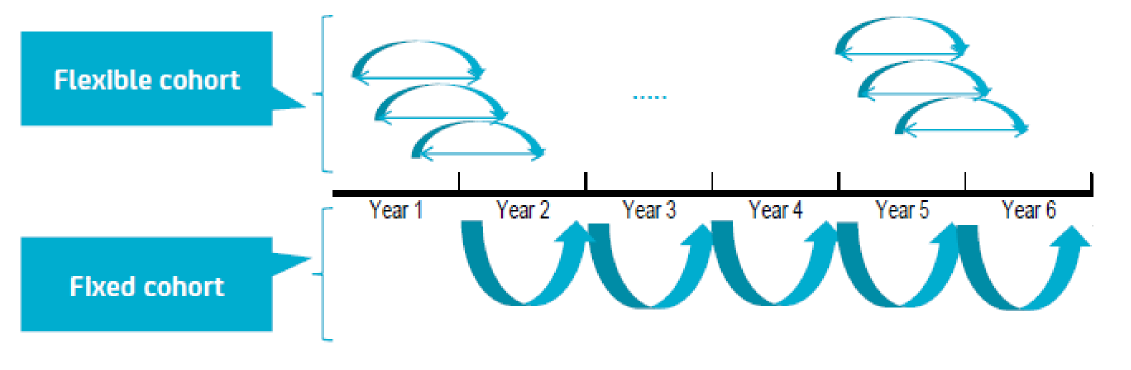
- 1) Data exclusions
- 2) Missing values analysis
- 3) Outlier treatment
- 4) Variable transformation (feature engineering)
- 5) Univariate analysis
- 6) Correlation analysis
- 7) Modelling
  - Selection of shortlist of variables
  - Estimation of coefficients based

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



Predictive Modelling – Sample definition

- 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 - Workflow

- 1) Data exclusions
- 2) Missing values analysis (> 50%?)
- 3) Outlier treatment (< 5^{th} Q/> 95^{th} Q?)
- 4) Variable transformation (feature engineering) (Binning)
- 5) Univariate analysis (GINI below .2?)
- 6) Correlation analysis (Spearman >.5?)
- 7) Modelling
  - Selection of shortlist of variables
  - Estimation of coefficients based

Predictive Modelling – Linear regression

- Problem to be solved:

single

own house

university degree

29 years

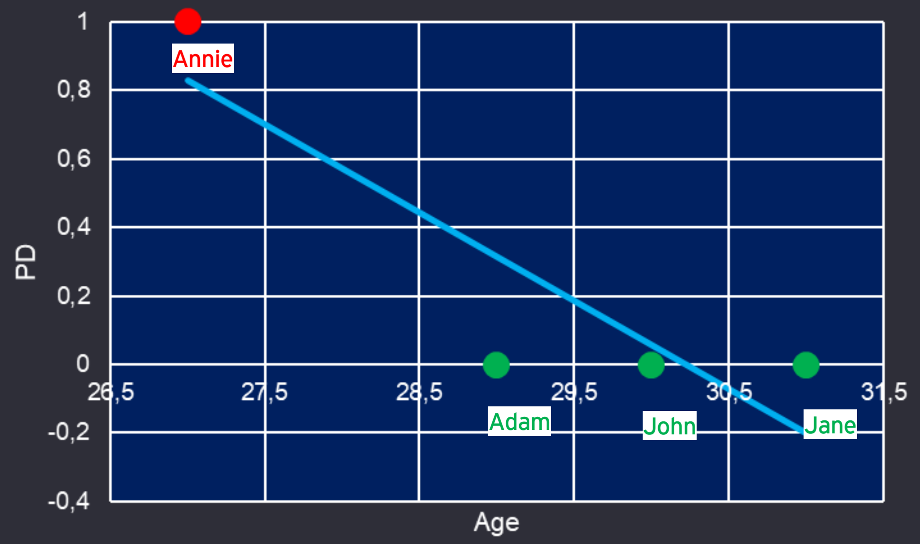
i=1 i=2 i=3 i=4 .....

Historical data with already known target value

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

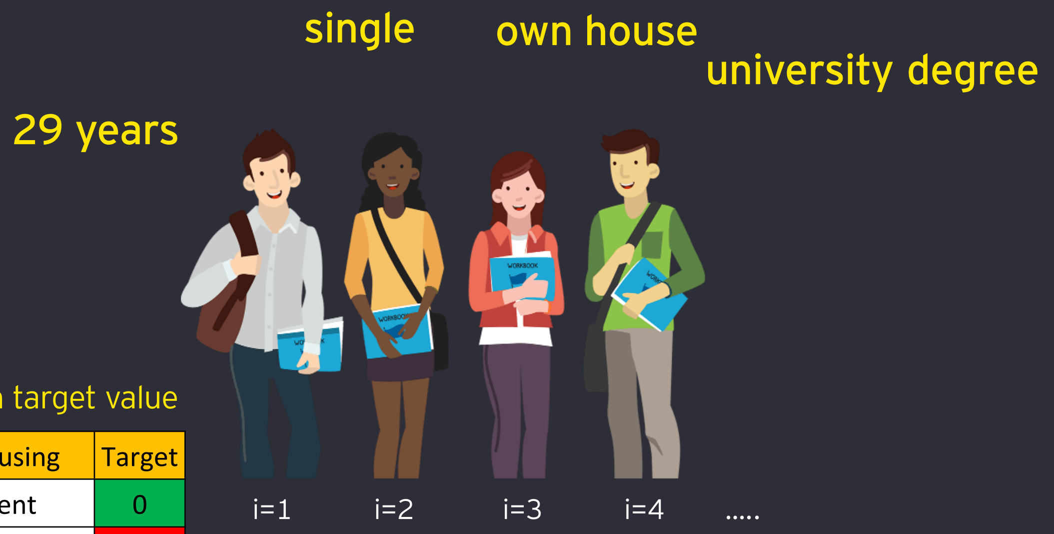
We choose linear function

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



Predictive Modelling – Logistic regression

- Problem to be solved:

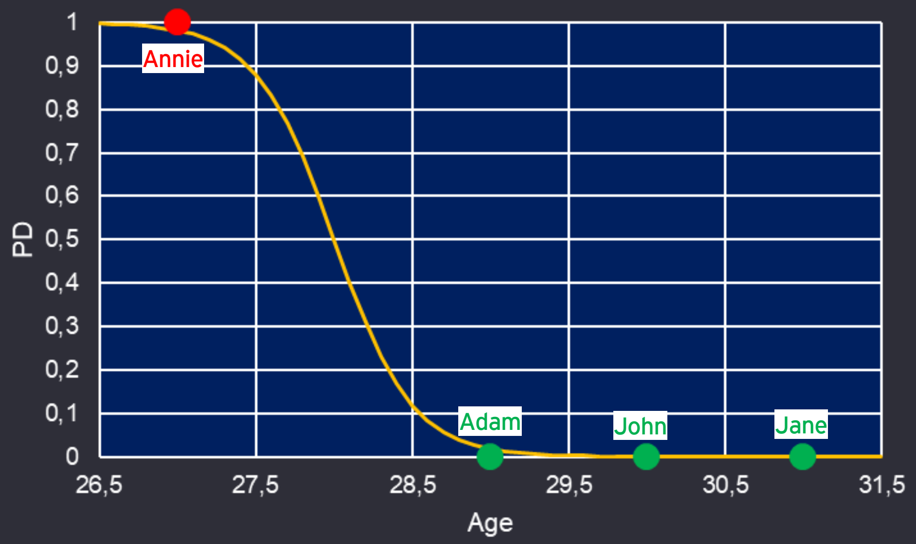


Historical data with already known target value

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

We choose logistic function

$$f(\vec{x}) := \frac{1}{1 + e^{-\alpha - \sum_{j=1}^{k} \beta_j x_j}}$$



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

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

- By central limit theorem under very general conditions the log-odds ratio distribution converges in distribution to a normal distribution

Predictive Modelling - Prediction

- 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?

Comparison of different modelling techniques

Logistic Regression

- Logistic regression with variables grouping, WOE transformation
  - Quasi-maximum likelihood method for not independent observations (especially autocorrelation on client level)
  - Multinomial logistic regression for multinomial target

PROS	CONS
Fully under control Robust Easy to interpret	Many assumptions (predictors uncorrelated) Variable selection process

Decision Trees

- Decision trees
  - Regression trees for real target
  - Classification trees for multinomial target
  - Can use a combination of these two

PROS	CONS
No assumptions (dependent predictors) Easy to interpret	Lower predictive power Lack of sensitivity

Gradient boosting

- Gradient boosting
  - Regression as well as tree version
  - Based on iterative algorithm boosting the performance power by fitting the residuals

PROS	CONS
Higher prediction power than trees	Not easy to interpret Can be overfitted Implementation, running time

Association Analysis

- Association analysis is a data mining technique
  - Searches for small parts of the predictors space and finds irregularities in terms of certain target (default rate, approval rate, etc.)

PROS	CONS
Can run parallelly to classical scoring Fraud detection ability	Running time Can affect scoring

SVM* and NN*

- Are powerful, but can be easily overfitted and can have high impact to reject inference (should be used as challengers)
  - Support vector machines (SVM)
  - Neural networks

PROS	CONS
Higher prediction power than other methods	Overfitting Not interpretable Not deterministic optimization

Bagging Ensemble Methods*

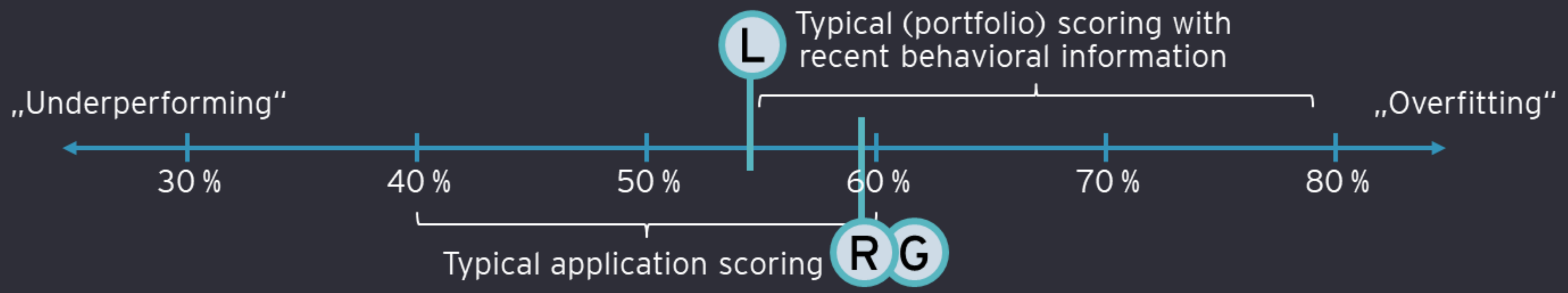
- Are based on developing many models on random subsamples or with different predictors and putting them together by ensemble rule (random forests, etc.)

PROS	CONS
Higher prediction power than standard linear methods Sometimes higher stability	Overfitting Not interpretable Not sufficient track record

* issue of selection of hyperparameters

Comparison of different modelling techniques

- We found that the predictive power of the logistic regression model and more advanced approaches is in the same league

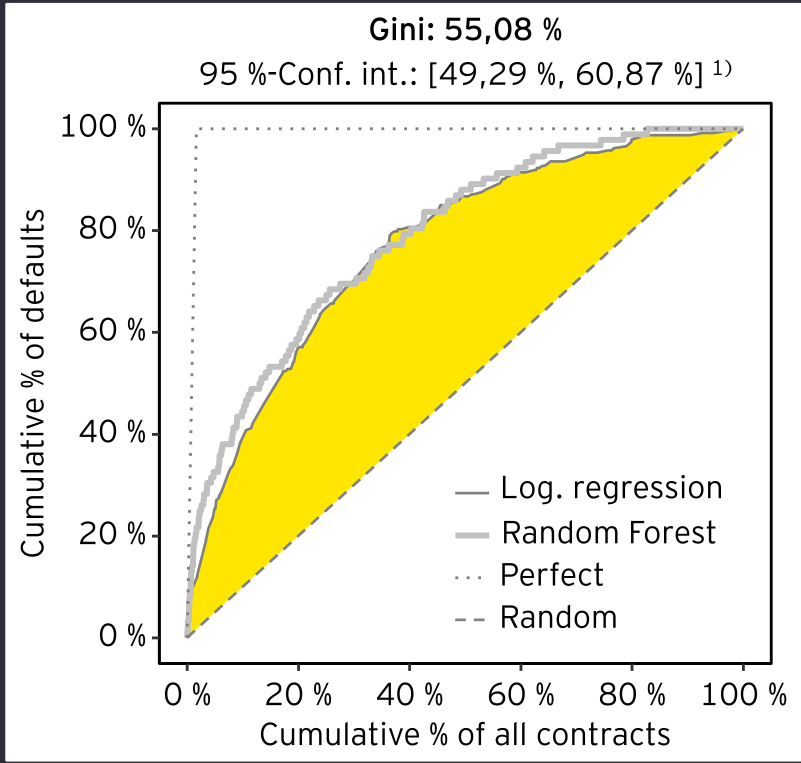


Modelling approach	Predictive power (GINI)	Variable selection
Logistic Regression Model (L)	55,08 %	Stepwise regression
Random Forest (R)	59,11 %	Default settings in h2o
Gradient Boosting Machine (G)	59,22 %	Default settings in h2o

A potential increase in predictive power with Random Forests is highly subject to information in the data (nonlinearity etc.)

Comparison of different modelling techniques

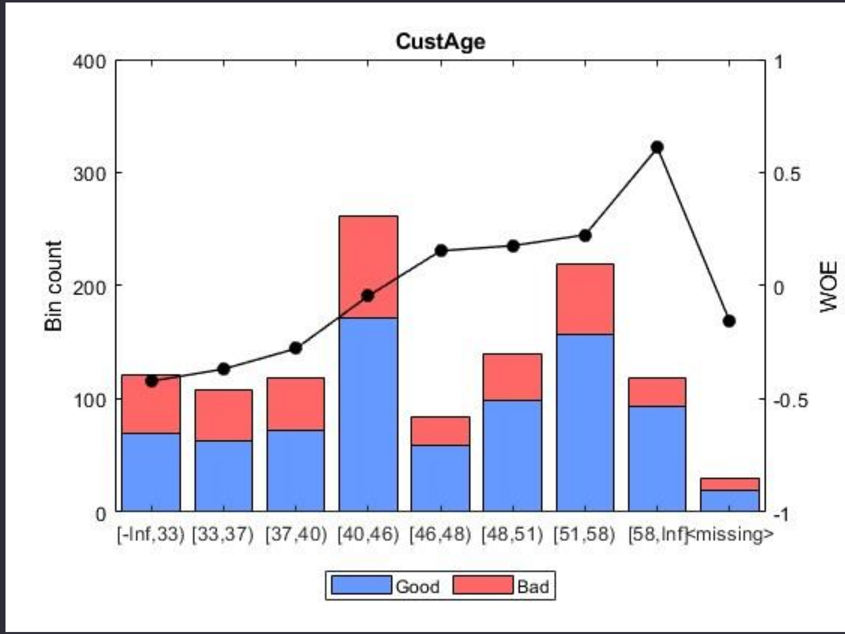
- All methods were applied after data cleansing, feature extraction and the categorization of all features in the short list.
- As only a few categories were allowed for each feature, nonlinear characteristics in the data may have been reduced by the loss of information from categorization.
- Hence, the advantage of Random Forests to cover also nonlinearity in the model is only of minor importance.
- The features for the logistic model were selected by stepwise regression.
- We further used a Random Forest implementation in VBA in order to validate the result which we obtained using the H2O algorithm in R.



Predictive Modelling - Binning

- Another standardly used technique is binning of predictors and WoE transformation:
  - Weight of evidence for i-th bin: $WoE_i := \ln\left(\frac{GOODS_i/BADS_i}{GOODS/BADS}\right)$

	BIN	GOODS	BADS	DR	WoE
1	[-inf,33)	69	52	0,429752	-0,42156
2	[33,37)	63	45	0,416667	-0,36795
3	[37,40)	72	47	0,394958	-0,27790
4	[40,46)	172	89	0,340996	-0,04556
5	[46,48)	59	25	0,297619	0,15424
6	[48,51)	99	41	0,292857	0,17712
7	[51,58)	157	62	0,283105	0,22469
8	[58,inf)	93	25	0,211864	0,60930
9	MISSING	19	11	0,366667	-0,15787

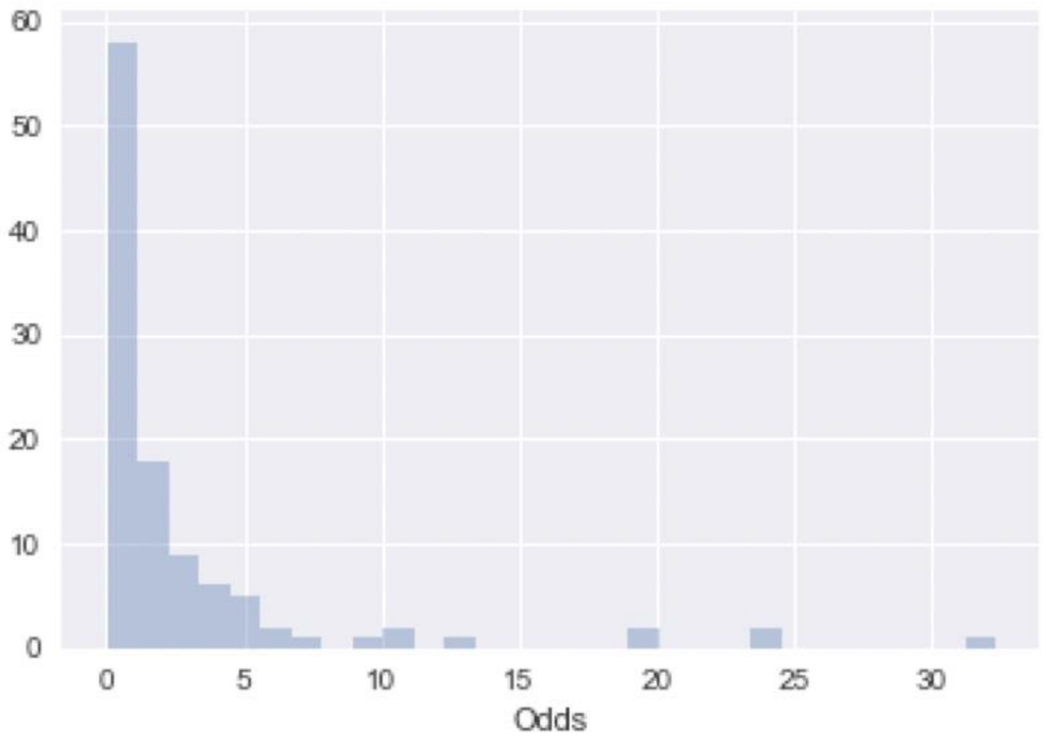


- Why binning? solves leverage points, solves informative missings, solves non-numerical (either ordinal or multinomial) variables, assesses robustness
- Why WoE transformation? normalizes predictors values, enables easy interpretation (under reasonable conditions always attains negative and positive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

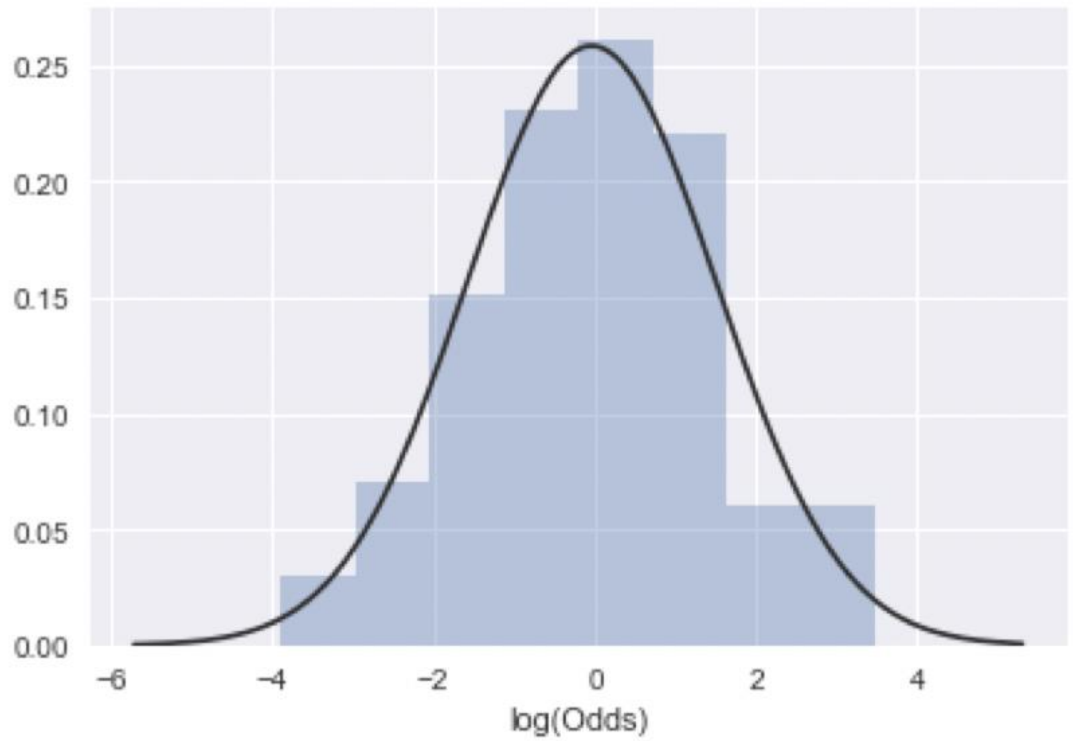
Log - odds

Take odds of random numbers which add up to 100

Ex: 20/80, 70/30, etc.



Taking log(odds) of the same



Predictive Modelling - WoE

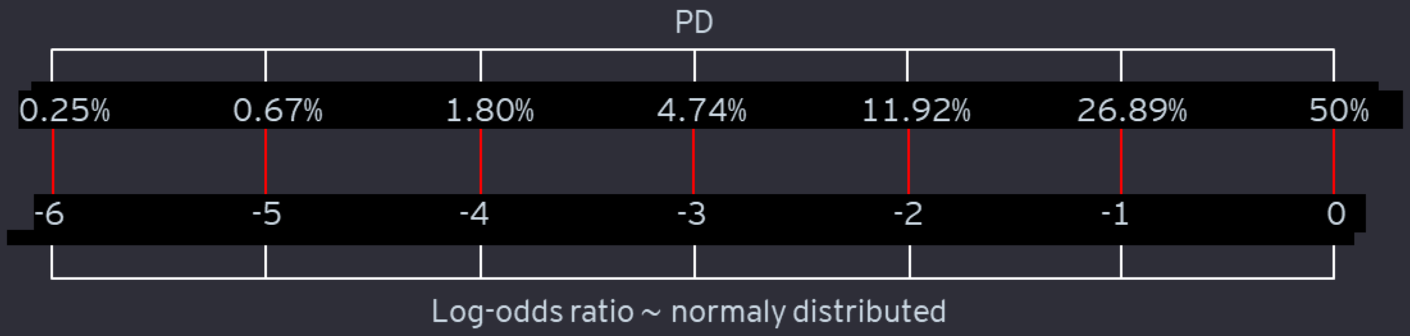
Predictor	Group	Scorecard Points	WoE	DR	Percentage of population	Coefficient
Intercept		25				-3,6578
Age	<25	0	-1,0109	18,1%	16,2%	-0,6572
	<35	27	-0,4535	11,2%	23,3%	
	<55	83	0,7352	3,7%	33,9%	
	>=55	111	1,3272	2,1%	25,1%	
	Missing	50	0,0420	7,1%	1,5%	
Education	Elementary	0	-1,0188	18,2%	6,9%	-0,8213
	Vocational	18	-0,7154	14,1%	15,6%	
	High school	82	0,3762	5,2%	38,1%	
	University	83	0,3877	5,2%	38,5%	
	Missing	85	0,4215	0,0%	0,9%	
Housing type	With parents	0	-0,8902	16,3%	10,9%	-0,7765
	Rent	2	-0,8489	15,8%	13,9%	
	Cooperative	44	-0,1144	8,3%	21,5%	
	Mortgage	105	0,9786	2,9%	42,9%	
	Own	82	0,5794	4,3%	9,7%	
	Missing	79	0,5216	0,0%	1,0%	
Marital status	Single	0	-0,8118	15,3%	29,4%	-0,5903
	Married	85	1,1710	2,4%	21,5%	
	Divorced	47	0,2865	5,7%	36,8%	
	Widowed	76	0,9634	3,0%	10,6%	
	Missing	76	0,9736	0,0%	1,7%	

$$Total\ scorecard\ points = Intercept\ scorecard\ points + \sum_{k=1}^{m} Variable_i\ scorecard\ points$$

- 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: $WoE = \ln\left(\frac{(BADS_i + 0.5)/(GOODS_i + 0.5)}{BADS/GOODS}\right)$
- 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

Predictive Modelling - Scorecard

- 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 :

$$\boxed{Score\ points_i = \alpha + \beta * \ln\left(\frac{PD_i}{1 - PD_i}\right)}$$

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

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

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

A is usually set to be 50 score points

B is usually set to 500 score points and the corresponding $x = \frac{1}{51}$

Scorecard points ~ normaly distributed

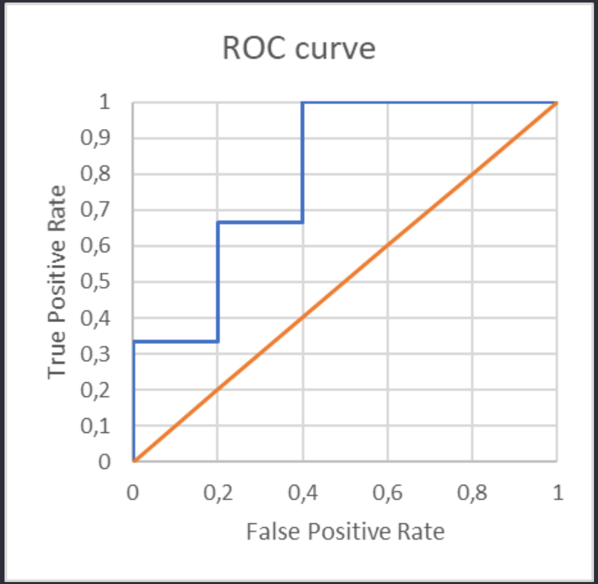
Model performance - GINI

- 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 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 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

Client	Event	Score
Annie	1	325
Paul	0	398
Lisa	1	415
Jane	0	463
Jack	1	499
Adam	0	520
John	0	611
Mary	0	672

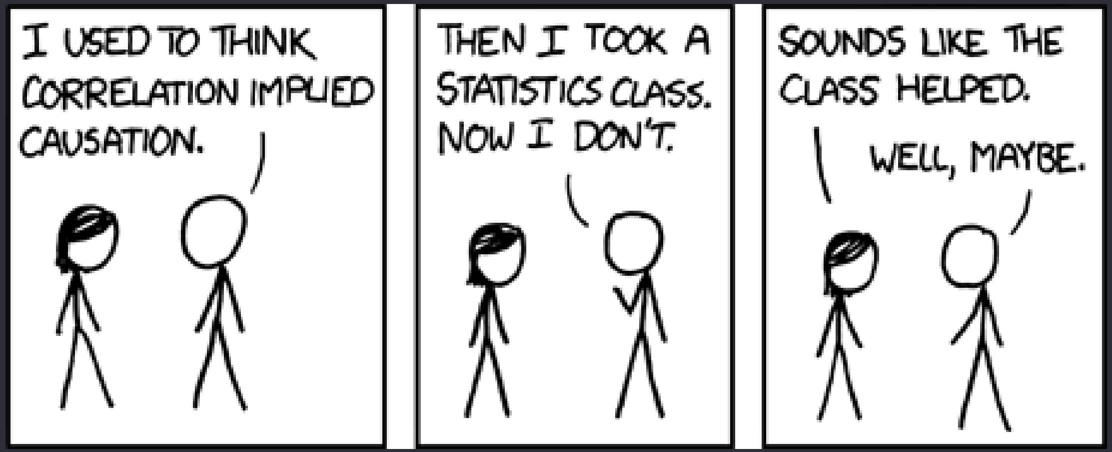


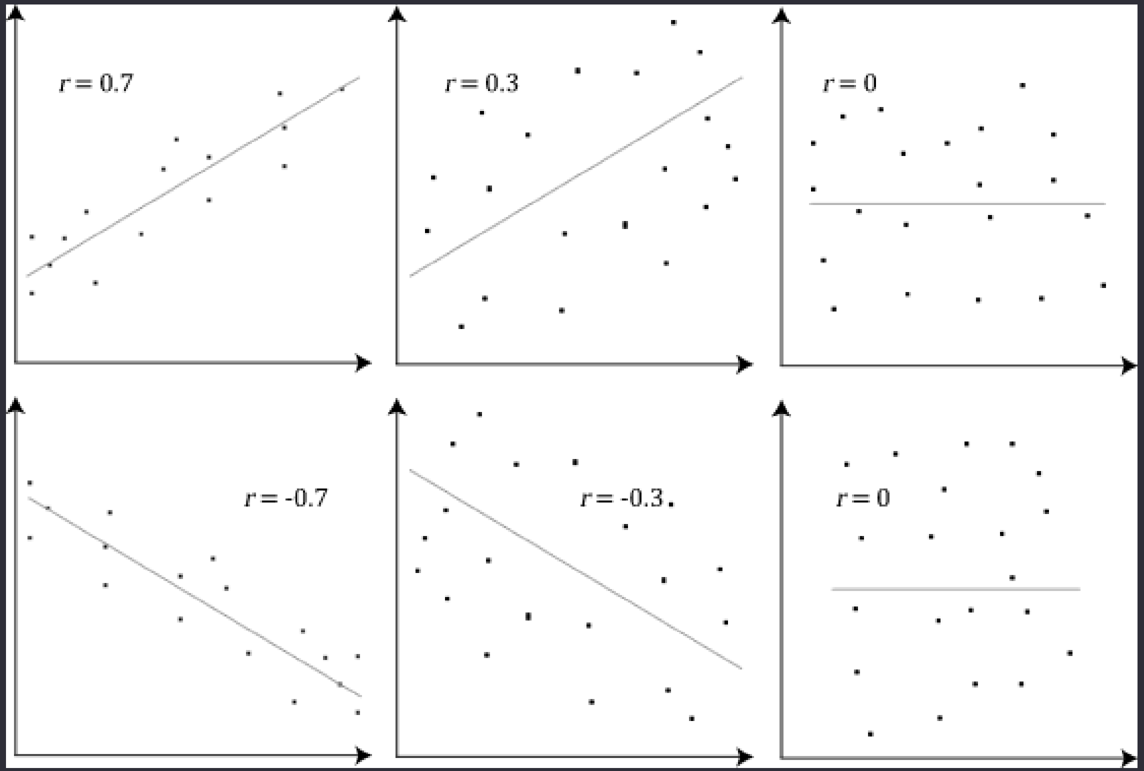
AUC (Area Under Curve) = 1/3*1/5+2/3*1/5+3/3*3/5= 4/5=0.8

GINI = 2*(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

Correlation





- Degree of statistical association between two random variables
- Pearson correlation coefficient
  - Sensitive to linear relationships
- Spearman correlation coefficient
  - More robust, sensitive to nonlinearity

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)

$$PSI = \sum_{i=1}^{n} (Actual\%_i - Expected\%_i) * \ln\left(\frac{Actual\%_i}{Expected\%_i}\right)$$

where n is number of bins

Score bands	Actual %	Expected %	Ac-Ex	ln(Ac/Ex)	Index
< 251	5%	8%	-3%	-0,470	**0,014**
251–290	6%	9%	-3%	-0,410	**0,012**
291–320	6%	10%	-4%	-0,510	**0,020**
321–350	8%	13%	-5%	-0,490	**0,024**
351–380	10%	12%	-2%	-0,180	**0,004**
381–410	12%	11%	1%	0,090	**0,001**
411–440	14%	10%	4%	0,340	**0,013**
441–470	14%	9%	5%	0,440	**0,022**
471–520	13%	9%	4%	0,370	**0,015**
520 <	9%	8%	1%	0,120	**0,001**
Population Stability Index (PSI) =					**0,1269**