EXPLAINABLE AI IN PRACTICE

Using NDT Gen 3 with UK CRA data

Explainability is vital to consumers, lenders and regulators

ICO guidance on Automated Decision Making and Profiling

– Detailed guidance on compliance with GDPR

FCA partnering with Alan Turing Institute to explore transparency and explainability of AI in the financial sector

– July 2019

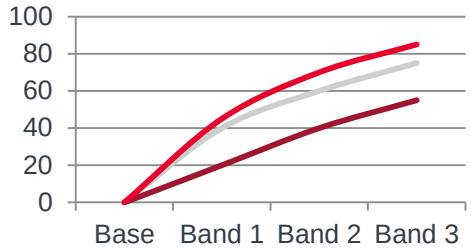
NeuroDecision Technology (NDT) is Equifax's patented Explainable AI technology. NDT produces powerful explainable models using neural networks.

Explainable scores come from consistent effects

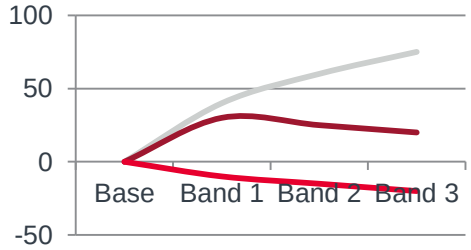
Traditional scorecards yield fixed responses

Variable	Score
Base	+0
Band 1	+40
Band 2	+60
Band 3	+75

Flexible models yield a range of responses

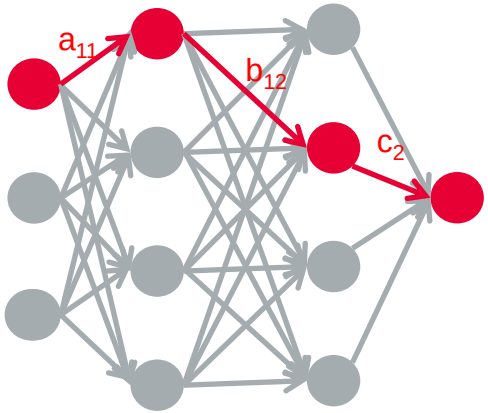


Consistent monotonic responses



Inconsistent responses

NDT enforces monotonicity by eliminating negative paths



Sign of this path
= sign(a_{11})sign(b_{12})sign(c_2)
= sign($a_{11}b_{12}c_2$)

- A path in the network is:
  - Positive if $a_{ij}b_{jk}c_k>0$
  - Zero if $a_{ij}b_{jk}c_k=0$
  - Negative if $a_{ij}b_{jk}c_k<0$
- If there are no negative paths, responses are monotonic
- NeuroDecision uses a penalty term to eliminate negative paths

UK characteristics and practice pose new challenges

Continuous variables vs. binning

- Neural networks benefit from using continuous numeric variables to better capture varying responses/interaction effects
- Equifax tests have shown NDT produces better models when using continuous variables
- Much UK practice uses binning

Bulk characteristic treatment

- UK practice often relies on ‘stepping in’ characteristics with just-in-time treatment
- Neural networks/NDT require bulk treatment & selection

Non-monotonic numeric variables

- E.g. credit card utilisation - lowest bad rate occurs around 10-20%

Default values

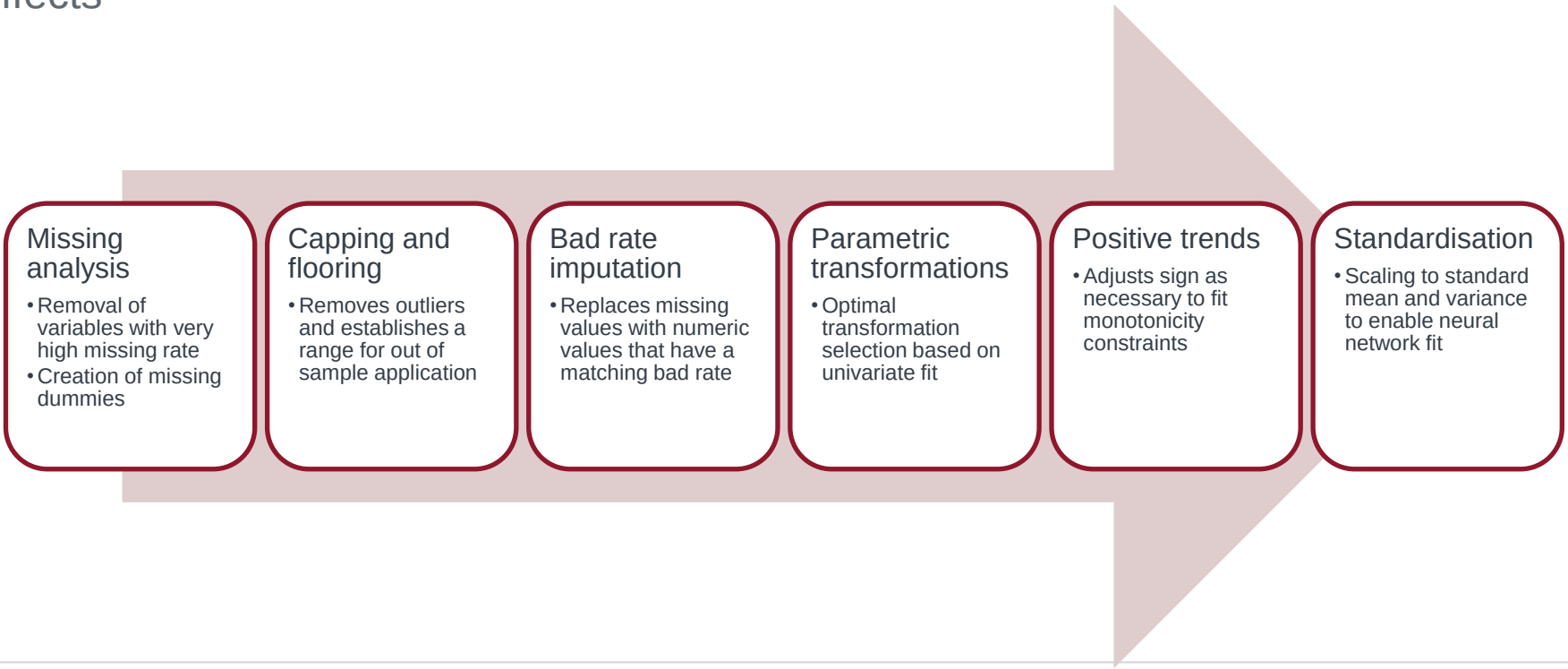
- Equifax use default values to explain why characteristics are ‘missing’. E.g.
  - C – No information qualifies for the calculation
  - M – No information in that data type
- Need recoding/imputing to treat variables as numeric

Pure categorical variables

- No natural order of effects – so what does monotonicity mean?

We apply bulk treatment for continuous variables

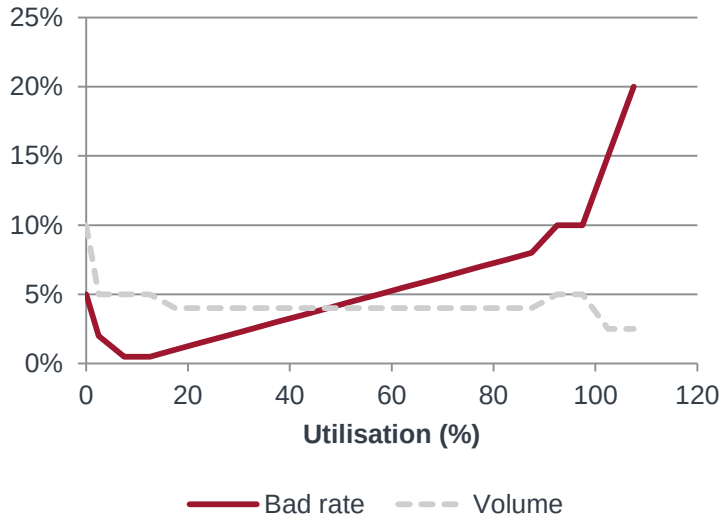
We use numeric variables in continuous form to maximise the value of interaction effects



We split non-monotonic continuous variables

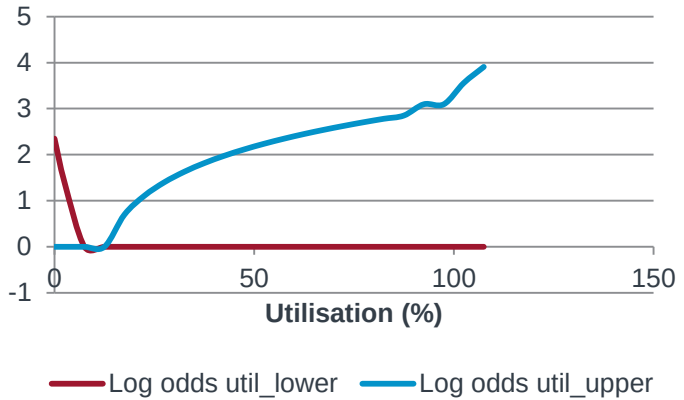
Example: credit card utilisation

U-shaped response with minimum ~ 10%



We split into two variables with monotonic responses:

- util_lower = min(util,10)
- util_higher = max(util,10)



We apply a bank of expert rules to treating default values

- Equifax use default values to represent levels of missing information in our characteristics
- Treating default values in bulk requires expert rules
  - Some missing values can be replaced by zero
    - E.g. LSC350 ‘number of insight accounts’
  - Others represent ratios that cannot be calculated
    - E.g. Average credit card utilisation
  - Some ‘zeroes’ capture important information
    - E.g. ‘no searches’ (M) vs ‘no searches L3M’ (C)
- Remaining ‘missing’ values are treated with bad rate imputation
- We also apply manual caps, e.g. utilisation capped at 120%

Value	Meaning
_	No information supplied (for ‘supplied addresses’ chars)
T	Address not found
M	No information in that data type
C	No information qualifies for the calculation
H	Characteristic cannot be calculated (zero denominator)
F	Characteristic is falsely positive (negative over negative)
G	Characteristic is negative
K	Denominator is negative

Example: SSC4 (Number of credit searches last 12 months, subject, current address)

- M → zero - no searches
- C → missing - to be imputed by bad rate

C indicates non-credit searches, i.e. debt collection

What does *monotonicity* mean for categorical variables?

- True categorical (not ordinal) variables have no natural order of effects
  - E.g. residential status:
    - Private tenant / Social tenant/ Living with parents
    - Mortgage/ Outright owner / Other
- We insist the order of effects be the same for all inputs
  - Gives consistent explanations
  - Ensures monotonicity of other variables is not compromised through interactions
- First, find the average marginal order of effects from a logistic regression with dummy encoding
  - Need not match univariate analysis
- Then, use ordinal/cumulative encoding so that monotonicity constraints have the desired effect

Char value	LR dummy estimate	Difference from previous
B	-1.2	
R	0	1.2
O	0.8	0.8
G	1.5	0.7

Char value	X_RvB	X_OvR	X_GvO	Total effect
B	0	0	0	Int
R	1	0	0	Int + X_RvB
O	1	1	0	Int + X_RvB + X_OvR
G	1	1	1	Int + X_RvB + X_OvR + X_GvO
Expected value	1.2	0.8	0.7	

An adapted NDT workflow for UK data

Variable Treatment

- Dummy encoding for categorical variables
- Expert treatment for default values & caps
- Missing analysis
- Capping & flooring
- Imputation
- Non-monotonic treatment
- Parametric transformations
- Positive trends

Variable Selection

- Automated selection using GBM
- Review treatment for selected variables
- Correlation analysis and manual reduction
- Fit logistic regression model
- Ordinal encoding for categorical variables

Model Building (NDT process)

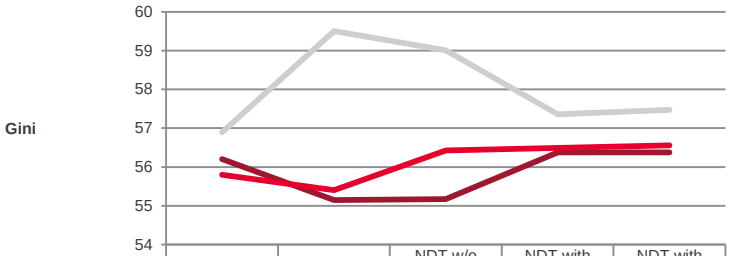
- Fit LR model
- Choose NN architecture
  - Loose fit unconstrained NN models
  - Tune penalty terms to achieve monotonicity
- Fit chosen NDT network to convergence
- Tune regularisation parameters
- (Remove redundant variables)

We tested NDT against credit application scorecards

- Comparison of NDT to traditional scorecards on existing model development samples
- Two populations
  - A – prime. 88k records
  - B – subprime. 61k records
- Circa 1700 EFX characteristics and 3 external variables
- Existing models are production grade credit application scorecards using logistic regression
  - 19 chars for pop A
  - 13 chars for pop B
- No new data

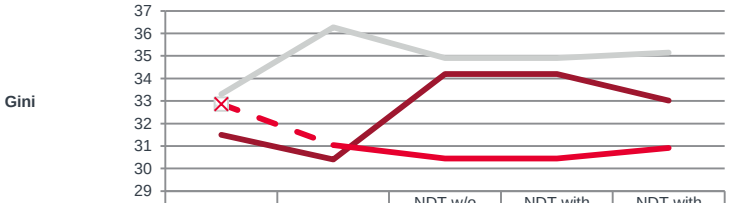
Results show the value of NDT

Pop A – prime (46 chars)



	Original scorecard	Unconstrained NN	NDT w/o regularisation	NDT with regularisation	NDT with variable tuning
Dev	56.90	59.50	59.01	57.36	57.47
Val (OOS)	56.20	55.15	55.17	56.38	56.38
Test (OOT)	55.80	55.40	56.43	56.49	56.56

Pop B – subprime (48 chars)



	Original scorecard	Unconstrained NN	NDT w/o regularisation	NDT with regularisation	NDT with variable tuning
Dev	33.30	36.27	34.91	34.91	35.15
Val (OOS)	31.50	30.41	34.19	34.19	33.02
Test (OOT)	32.86	31.05	30.45	30.45	30.92

Unconstrained NNs show high Gini in development sample but poor generalisation

NDT constraints reduce development sample fit, but generalisation is stronger

NDT with regularisation shows better (pop A) or equal (pop B) generalisation

Variable tuning (removing redundant variables) is of questionable value

Summary – Proving NDT in the UK

- NeuroDecision Technology (NDT) uses constrained neural networks to build explainable models through monotonicity constraints
- Neural networks require a change in variable treatment approach from just-in-time binning to bulk numeric transformation and selection
- Using NDT with UK CRA data requires additional steps to handle default values, non-monotonic response and categorical variables
- Equifax have established a workflow to build NDT models with UK data
- NDT produces a small Gini uplift OOS and OOT on our prime sample, and a substantial OOS uplift (8.5%) on our subprime sample
  - Subprime OOT results may be unreliable
- We are extending our internal tests to other model samples and discussing POCs with several clients