# Research in Predictive Modelling

**Binning, Variable Selection, Income Modelling**

Tanja Verster

Date: 2 June 2016

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOMKAMPUS

# Agenda: Research in predictive modelling

**Predictive modelling**

- Binning
- Variable selection
- Income modelling

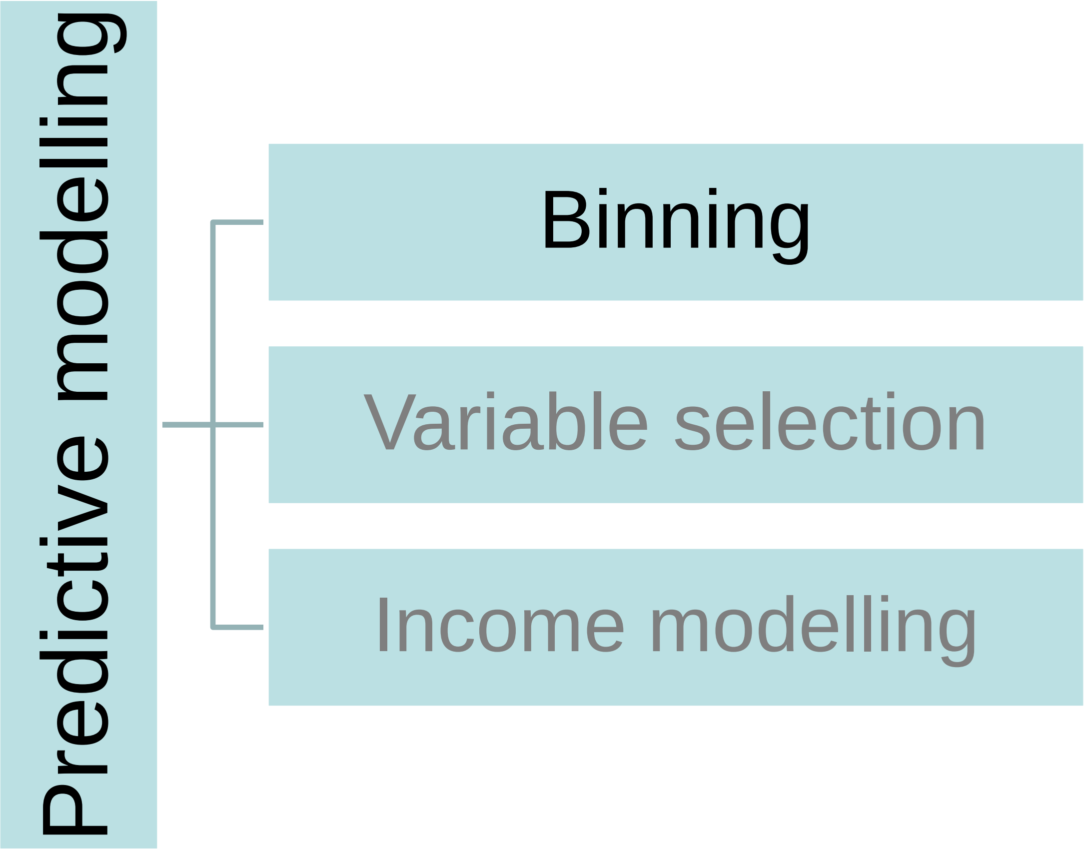
# Agenda: Research in predictive modelling

Predictive modelling

Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior.

Predictive models are widely used as analytical tools in retail credit.

# Agenda: Research in predictive modelling



Binning

Known by many names

- Classing
- Grouping
- Discretization
- Quantization

# What is binning?

**Discretization**

- Discretization is the transformation of continuous data into discrete bins (Nguyen et al, 2014)

**Binning**

- A variable transformation called "binning"—which maps variables from a large ordered set or a continuous range into a small set of bins (Oliveira et al, 2008)

# Why bin?

- Basic preparation needed to build a scorecard (Thomas, 2009)
- It is an important and general pre-processing technique, and a critical element of many data mining and data management tasks.
- The general goal is to obtain data that retains as much information in the continuous original as possible  (Nguyen et al 2014)

# Many binning methods exist

**Equal-interval-width method**

**Equal-frequency-per-interval method**

**Minimal-class-entropy method**

- Example is decision trees (CART)
- RPART is procedure in R that implements CART

# New binning method introduced: Autobin

**A predictive approach towards automatic binning**

- by Hennie Venter and Tanja Verster

  - **In a nutshell:**
    - Split the training data set in two
    - Consider all possible ways to "bin" a variable, and choose the binning that minimize the loss
    - Loss is equal to
      - Prediction loss when comparing the predictions based on the first split with the observed data in the second split and the other way around
    - BUT in general the solution of this problem is hard
    - Use dynamic programming to determine solution (if variable is ordered)

# More detail on new binning method: first notation

$Y$ a zero-one response variable
$X$ is a discrete regressor
taking values in $V = \{v_1, v_2, \ldots, v_M\}$

# **More detail on new binning method:** first notation

$Y$ a zero-one response variable
$X$ is a discrete regressor
taking values in $V = \{v_1, v_2, \ldots, v_M\}$

Data: frequency table with $v_m$ occurring $f_m$ times while the corresponding numbers of 1's and 0's of $Y$ are $b_m$ and $g_m$ respectively

# **More detail on new binning method:** first notation

$Y$ a zero-one response variable
$X$ is a discrete regressor
taking values in $V = \{v_1, v_2, ..., v_M\}$

Data: frequency table with $v_m$ occurring $f_m$ times while the corresponding numbers of 1's and 0's of $Y$ are $b_m$ and $g_m$ respectively

Aim: to predict the value of *Y* given a newly observed value of *X*

# Split the training data set in two:

## Formulation from a predictive point of view

Take two random samples from the data set

$N'$ observations

Frequency tables

$f'_m \quad b'_m \quad g'_m$

$p'_m = b'_m / f'_m$

$N''$ observations

Frequency tables

$f''_m \quad b''_m \quad g''_m$

$p''_m = b''_m / f''_m$

# Formulation of loss function

The probability of getting $y_n''$=1 when $X = v_m$ is estimated from the first sample as $p_m'$

The extent of the discrepancy between prediction and observations can be measured by **a loss function**

If $y_n''$ =1 and $p_m'$ is high, small discrepancy

If $y_n''$ =1 and $p_m'$ is small, big discrepancy

$l(y, p) = -[y \log(p) + (1-y)\log(1-p)]$ for $y = 0, 1$ and $0 \leq p \leq 1$

# Formulation

Prediction loss when comparing the predictions based on the first sample with the observed data in the second sample is

$$-\sum_{m=1}^{M}\{b''_m \log(p'_m) + g''_m \log(1 - p'_m)\}$$

Prediction loss when comparing the predictions based on the second sample with the observed data in the first sample is

$-$ [redacted] $\{b'_m \log(p''_m) + g'_m \log(1 - p''_m)\}$

# Formulation of loss function

Average mutual cross validation prediction loss (CV)

$$CV(V) = -\frac{1}{N' + N''} \sum_{m=1}^{M} \left\{ \begin{matrix} b''_m \log(p'_m) + g''_m \log(1 - p'_m) + \\ b'_m \log(p''_m) + g'_m \log(1 - p''_m) \end{matrix} \right\}$$

# Formulation:

## Consider all possible ways to "bin" a variable, and choose the binning that minimize the loss

Let $\boldsymbol{C} = \{C_1, C_2,\ldots, C_K\}$ denote a partition of $V$ into $K$ bins and denote the corresponding observed data of the first sample in the bin $C_k$ by

$$f'_k(\boldsymbol{C}) = \sum_{m=1}^{M} f'_m I(v_m \epsilon C_k),\ \ b'_k(\boldsymbol{C}) = \sum_{m=1}^{M} b'_m I(v_m \epsilon C_k) \ \text{ and } \ g'_k(\boldsymbol{C}) = \sum_{m=1}^{M} g'_m I(v_m \epsilon C_k)$$

and similarly for the second sample. Denote the estimated probabilities $p'_k(\boldsymbol{C})$ and $p''_k(\boldsymbol{C})$ for bin $C_k$.

The problem is to find the partition $\boldsymbol{C}$ that minimizes

$$CV(\boldsymbol{C}) = -\frac{1}{N' + N''} \sum_{k=1}^{K} \begin{Bmatrix} b''_k(\boldsymbol{C}) \log\big(p'_k(\boldsymbol{C})\big) + g''_k(\boldsymbol{C}) \log\big(1 - p'_k(\boldsymbol{C})\big) + \\ b'_k(\boldsymbol{C}) \log\big(p''_k(\boldsymbol{C})\big) + g'_k(\boldsymbol{C}) \log\big(1 - p''_k(\boldsymbol{C})\big) \end{Bmatrix}$$

# **Formulation:**

Consider all possible ways to "bin" a variable, and choose the binning that minimize the loss

Let $\boldsymbol{C} = \{C_1, C_2,\dots, C_K\}$ denote a partition of $V$ into $K$ bins and denote the corresponding observed data of the first sample in the bin $C_k$ by

$$f_k'(\boldsymbol{C}) = \sum_{m=1}^{M} f_m' I(v_m \epsilon C_k),\ \ b_k'(\boldsymbol{C}) = \sum_{m=1}^{M} b_m' I(v_m \epsilon C_k) \text{ and } g_k'(\boldsymbol{C}) = \sum_{m=1}^{M} g_m' I(v_m \epsilon C_k)$$

and similarly for the second sample. Denote the estimated probabilities $p_k'(\boldsymbol{C})$ and $p_k''(\boldsymbol{C})$ for bin $C_k$.

The problem is to find the partition $\boldsymbol{C}$ that minimizes

$$CV(\boldsymbol{C}) = -\frac{1}{N' + N''}\sum_{k=1}^{K}\begin{Bmatrix} b_k''(\boldsymbol{C})\log\big(p_k'(\boldsymbol{C})\big) + g_k''(\boldsymbol{C})\log\big(1 - p_k'(\boldsymbol{C})\big) + \\ b_k'(\boldsymbol{C})\log\big(p_k''(\boldsymbol{C})\big) + g_k'(\boldsymbol{C})\log\big(1 - p_k''(\boldsymbol{C})\big) \end{Bmatrix}$$

In general the solution of this problem is hard.

If the value set $V = \{v_1, v_2, \dots, , v_M\}$ is ordered an efficient dynamic programming solution is possible.

# Summary: Autobin

**A predictive approach towards automatic binning**

- by Hennie Venter and Tanja Verster
    - In a nutshell:
        - Split the training data set in two
        - Consider all possible ways to "bin" a variable, and choose the binning that minimize the loss
        - Loss is equal to
            - Prediction loss when comparing the predictions based on the first split with the observed data in the second split and the other way around
        - BUT in general the solution of this problem is hard
        - Use dynamic programming to determine solution (if variable is ordered)

**Tuning parameter 1:** minimum number of observations in a bin

# Compare Autobin with decision tree (RPART)

- **Recursive Partitioning**
  - In a nutshell
    - RPART is a procedure in R.
    - Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 "*Classification and Regression Trees*" CART book by Breiman, Friedman, Olshen and Stone.
    - Run automatically from SAS using PROC R

**Tuning parameter 1:** minimum number of observations in a bin

**Tuning parameter 2:** cp (complexity parameter)

# Comparing binning methods

| | RPART | RPART | AUTOBIN | AUTOBIN |
|---|---|---|---|---|
| Minimum in bin | 5% | 5% | 5% | 2% |
| Parameters | cp=0.001 | cp=0.0001 | | |
| Average IV | 0.0771 | 0.0778 | 0.0725 | 0.0779 |
| Average # of bins | 4.325 | 4.858 | 4.514 | 5.935 |

## Dataset used:

- 300 000 observations, 2000 variables

# Summary of new binning method:

The approach has benefits on two fronts:

Automatic

- Time saving

Prediction

- Often in credit scoring, the binning is required for more accurately predicting the default probabilities

# Agenda: Research in predictive modelling

Predictive modelling

- Binning
- Variable selection
- Income modelling

# What is variable selection?

Variable selection

The process of finding the best subset of features from the original set of features in a given dataset (Cios, et al. 1998)

# Many variable selection methods exist

**Subset selection**

- Forward, Backward, Stepwise

**Regularization**

- Elastic nets

**New method is introduced**

- $\lambda$-good subsets

# New variable selection method introduced:

- **$\lambda$-good subsets**
  - New technique developed by Prof Hennie
  - Basic idea (in a nutshell):
    - Adding new variable to current subset of variables, only if the criterion of fit improved by at least $\lambda$
    - Removing variable from current subset of variables, only if the criterion of fit reduced by at least $\lambda$

**Tuning parameter: $\lambda$**

# More detail: $\lambda$-good subsets

$Y$ a zero-one response variable

Possible criterion of fit

Log likelihood — maximize

AUC — maximize

# More detail: $\lambda$-good subsets

$Y$ a zero-one response variable

- Possible criterion of fit
  - Log likelihood — maximize
  - AUC — maximize

$l(y,p) = -[y\log(p) + (1-y)\log(1-p)]$ for $y = 0, 1$ and $0 \le p \le 1$

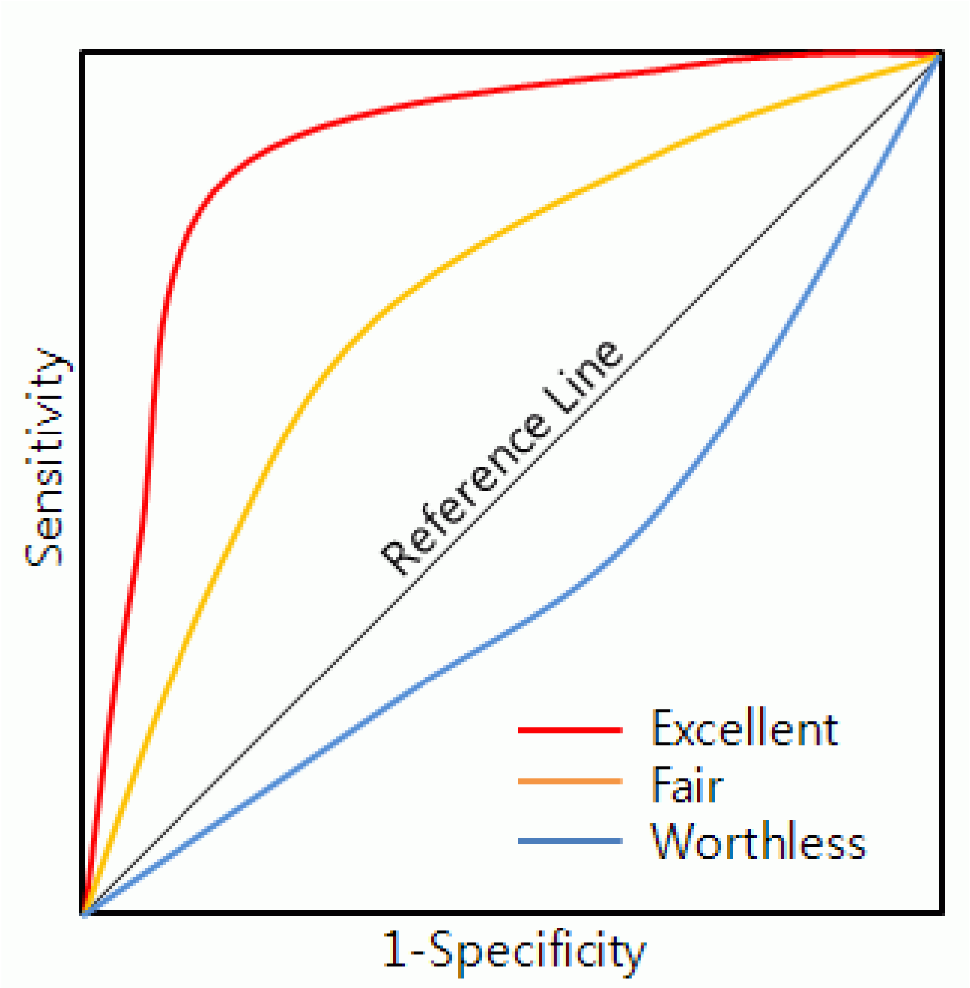
# AUC (Area under the ROC curve)

- **ROC: Receiving operating curve**
  - Plot Sensitivity i.e. true positive rate
  - against (1 – Specificity) i.e. false positive rate

| **AUC** | **Gini** |
|---|---|
| AUROC | Wilcoxon Mann-Whitney U statistic |
| Area under the ROC curve | Summary statistic under CAP curve |
| C-statisci | Somer's D |

$$Gini = 2(AUC) - 1$$

# AUC (Area under the ROC curve)



# $\lambda$-good subsets

Criterion of fit

Adding new variable to current subset of variables, only if the criterion of fit improved by at least $\lambda$

Removing variable from current subset of variables, only if the criterion of fit reduced by at least $\lambda$

# Compare these binning techniques in combination with variable selection techniques

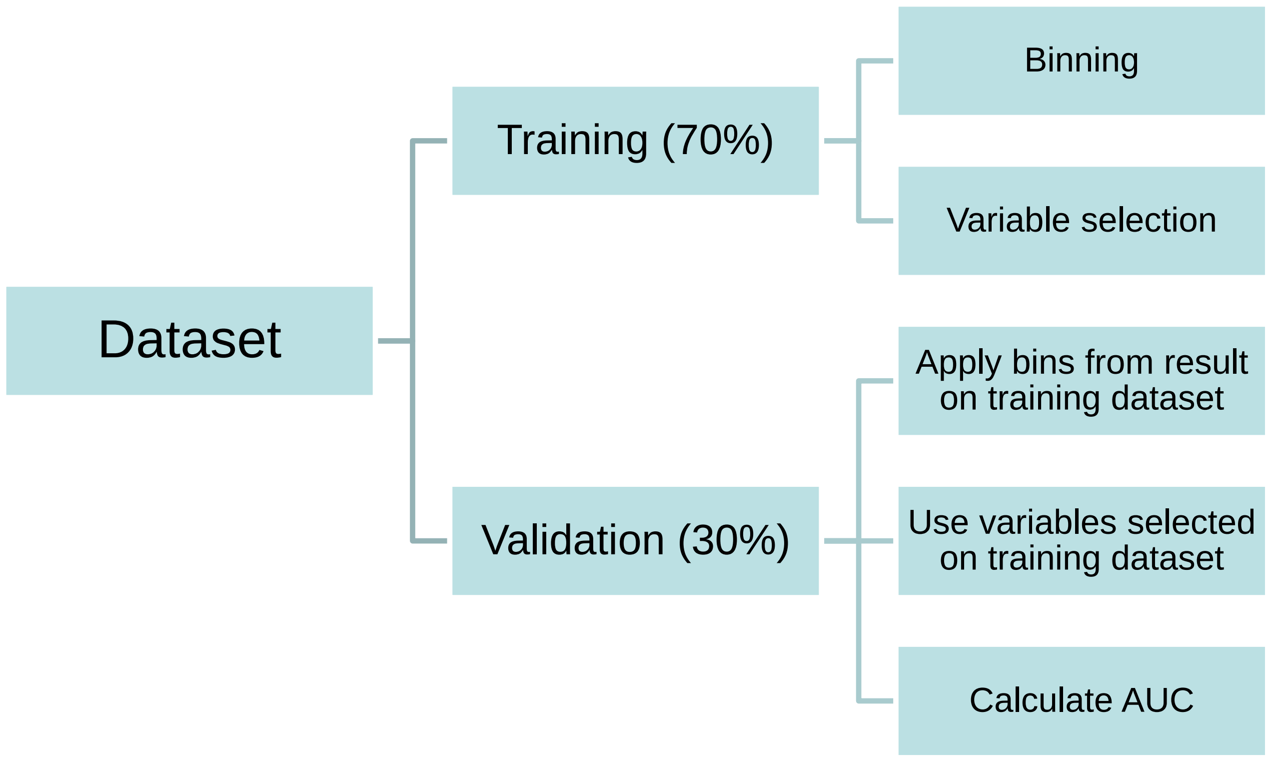
- **Elastic nets**
  - Elastic nets using package ‘glmnet’ in R
    - Friedman, J., Hastie, T. and Tibshirani, R. (2008) *Regularization Paths for Generalized Linear Models via Coordinate Descent*
    - Extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression, logistic and multinomial regression models, poisson regression and the Cox model.
  - **Tuning parameter 1**: $\alpha$
  - **Tuning parameter 2:** After initial analysis, the number of variables chosen was very high, and a new restriction was added: maximum number of variables that may be chosen pmax=30

# Comparing different variable selection techniques

| | | |
|---|---|---|
| Elastic nets | | |
| $\alpha = 0$ | $\alpha = 1$ | $0 < \alpha < 1$ |
| Ridge Regression | Lasso Regression | Elastic nets |

# Results:
# **Dataset** 300 000 observations, 2000 variables

2 different binning techniques (RPART and Autobin)
3 variables selection techniques (LR, $\lambda$-good subsets, elastic nets)



# Results:
# Dataset 300 000 observations, 2000 variables

2 different binning techniques (RPART and Autobin)
3 variables selection techniques (LR, $\lambda$-good subsets, elastic nets)

| Binning method | RPART used to bin 5% (cp = 0.001) | | RPART used to bin 5% (cp = 0.00001) | | AUTOBIN used to bin 5% | | AUTOBIN used to bin 2% | |
|---|---|---|---|---|---|---|---|---|
| | # Var | AUC (Valid) | # Var | AUC (Valid) | # Var | AUC (Valid) | # Var | AUC (Valid) |
| Stepwise | 63 | 0.8791 | 60 | 0.8813 | 82 | 0.8726 | 74 | 0.8867 |
| Backward fast | 74 | 0.8794 | 77 | 0.8818 | 72 | 0.8723 | 84 | 0.8867 |
| Forward | 65 | 0.8789 | 69 | 0.8814 | 66 | 0.8722 | 74 | 0.8866 |
| λ-good subsets | 35 | 0.8778 | 23 | 0.8743 | 29 | 0.8708 | 32 | 0.8784 |
| Ridge (α = 0) | 252 | 0.8722 | 258 | 0.8746 | 191 | 0.8647 | 257 | 0.8778 |
| Elasticnet α=0.3 pmax = 30 | 19 | 0.8520 | 18 | 0.8527 | 25 | 0.8439 | 21 | 0.8546 |
| Elasticnet α=0.5 pmax = 30 | 26 | 0.8565 | 24 | 0.8571 | 22 | 0.8486 | 19 | 0.8584 |
| Elasticnet α=0.7 pmax = 30 | 25 | 0.8587 | 23 | 0.8589 | 18 | 0.8507 | 19 | 0.8624 |
| Lasso (α=1 pmax = 30) | 23 | 0.8606 | 23 | 0.8631 | 20 | 0.8570 | 21 | 0.8669 |

# Results

Which binning performs best per variable selection technique?

Each time: Only marginally

| Binning method | RPART used to bin 5% (cp = 0.001) | | RPART used to bin 5% (cp = 0.00001) | | AUTOBIN used to bin 5% | | AUTOBIN used to bin 2% | |
|---|---|---|---|---|---|---|---|---|
| | # Var | AUC (Valid) | # Var | AUC (Valid) | # Var | AUC (Valid) | # Var | AUC (Valid) |
| Stepwise | 63 | 0.8791 | 60 | 0.8813 | 82 | 0.8726 | 74 | 0.8867 |
| Backward fast | 74 | 0.8794 | 77 | 0.8818 | 72 | 0.8723 | 84 | 0.8867 |
| Forward | 65 | 0.8789 | 69 | 0.8814 | 66 | 0.8722 | 74 | 0.8866 |
| λ-good subsets | 35 | 0.8778 | 23 | 0.8743 | 29 | 0.8708 | 32 | 0.8784 |
| Elasticnet α=0.3 pmax = 30 | 19 | 0.8520 | 18 | 0.8527 | 25 | 0.8439 | 21 | 0.8546 |
| Elasticnet α=0.5 pmax = 30 | 26 | 0.8565 | 24 | 0.8571 | 22 | 0.8486 | 19 | 0.8584 |
| Elasticnet α=0.7 pmax = 30 | 25 | 0.8587 | 23 | 0.8589 | 18 | 0.8507 | 19 | 0.8624 |
| Lasso (α=1 pmax = 30) | 23 | 0.8606 | 23 | 0.8631 | 20 | 0.8570 | 21 | 0.8669 |

# Results

Overall winner: Difficult?
Take into account AUC (on Validation dataset) as well as number of variables

| Binning method | RPART used to bin 5% (cp = 0.001) | | RPART used to bin 5% (cp = 0.00001) | | AUTOBIN used to bin 5% | | AUTOBIN used to bin 2% | |
|---|---|---|---|---|---|---|---|---|
| | # Var | AUC | # Var | AUC | # Var | AUC | # Var | AUC |
| Stepwise | 63 | 0.8791 | 60 | 0.8813 | 82 | 0.8726 | 74 | 0.8867 |
| Backward fast | 74 | 0.8794 | 77 | 0.8818 | 72 | 0.8723 | 84 | 0.8867 |
| Forward | 65 | 0.8789 | 69 | 0.8814 | 66 | 0.8722 | 74 | 0.8866 |
| λ-good subsets | 35 | 0.8778 | 23 | 0.8743 | 29 | 0.8708 | 32 | 0.8784 |
| Elasticnet α=0.3 pmax = 30 | 19 | 0.8520 | 18 | 0.8527 | 25 | 0.8439 | 21 | 0.8546 |
| Elasticnet α=0.5 pmax = 30 | 26 | 0.8565 | 24 | 0.8571 | 22 | 0.8486 | 19 | 0.8584 |
| Elasticnet α=0.7 pmax = 30 | 25 | 0.8587 | 23 | 0.8589 | 18 | 0.8507 | 19 | 0.8624 |
| Lasso (α=1 pmax = 30) | 23 | 0.8606 | 23 | 0.8631 | 20 | 0.8570 | 21 | 0.8669 |

# Results

Overall winner: Difficult?
Take into account AUC on VALID as well as number of variables

| Binning method | RPART used to bin 5% (cp = 0.001) | | RPART used to bin 5% (cp = 0.00001) | | AUTOBIN used to bin 5% | | AUTOBIN used to bin 2% | |
|---|---|---|---|---|---|---|---|---|
| | # Var | AUC | # Var | AUC | # Var | AUC | # Var | AUC |
| Stepwise | 63 | 0.8791 | 60 | 0.8813 | 82 | 0.8726 | 74 | 0.8867 |
| Backward fast | 74 | 0.8794 | 77 | 0.8818 | 72 | 0.8723 | 84 | 0.8867 |
| Forward | 65 | 0.8789 | 69 | 0.8814 | 66 | 0.8722 | 74 | 0.8866 |
| λ-good subsets | 35 | 0.8778 | 23 | 0.8743 | 29 | 0.8708 | 32 | 0.8784 |
| Elasticnet α=0.3 pmax = 30 | 19 | 0.8520 | 18 | 0.8527 | 25 | 0.8439 | 21 | 0.8546 |
| Elasticnet α=0.5 pmax = 30 | 26 | 0.8565 | 24 | 0.8571 | 22 | 0.8486 | 19 | 0.8584 |
| Elasticnet α=0.7 pmax = 30 | 25 | 0.8587 | 23 | 0.8589 | 18 | 0.8507 | 19 | 0.8624 |
| Lasso (α=1 pmax = 30) | 23 | 0.8606 | 23 | 0.8631 | 20 | 0.8570 | 21 | 0.8669 |

Promising results: using $\lambda$-good subsets with Autobin

Future research

# Agenda: Research in predictive modelling

- Predictive modelling
  - Binning
  - Variable selection
  - Income modelling

# Income modelling

- In the field of retail credit, most predictive models involve extending credit to consumers. The National Credit Act requires that credit providers must properly test affordability before extending credit. An income model can be used to accurately predict the gross monthly income of individuals in order to support these affordability requirements. Various methods to model income need to be compared.

- **Masters degree student (Z Kruth) project in 2014 on "**Estimation of household income and affordability for the purpose of credit decisions**" at XDS**

# Summary: Research in predictive modelling

- Predictive modelling
  - Binning
    - Status: in process
    - Advantage: saves time (automatic binning)
  - Variable selection
    - Status: in process
    - Advantage: preliminary results looks promising
  - Income modelling
    - Status: in process
    - Advantages: Student project

# Future research

- **Binning: Autobin**
  - Investigate stability of random samples chosen i.e. is results consistent if training dataset is split using a different random seed?
  - Investigate optimal value of tuning parameter (minimum size of bin)
- **Variable selection: $\lambda$-good subsets**
  - Investigate the optimal results
    - how to balance the AUC and the number of variables
    - Investigate optimal value of tuning parameter, $\lambda$
- **Income modelling**
  - Various methods to model income need to be compared.
  - Research methods to increase accuracy of
    - Income Estimation
    - Disposable Income Calculation
    - The Affordability Index
  - Including longitude and latitude data

# Thank You

# Summary slide

## Title: Research in predictive modelling - Binning, Variable Selection, Income modelling

**(Client: Dries de Wet (XDS) ; BMI: Tanja Verster)**

**Objective:**

The following are important aspects in predictive modelling that requires further research:

1) Binning: More effective and more automated ways to bin variables need to be researched.
2) Variable selection: Several variable selection techniques need to be compared including stepwise regression, factor analysis, variable clustering, partial least squares and other state-of-the-art techniques.
3) Income modelling: Investigate various methodologies that can effectively be used in income modelling.

**Progress**

1) A new technique to automatically bin variables has been developed and is being finalised. A research paper is planned for 2017.
2) A new technique to select variables has been developed and is being refined (ongoing research). A research paper is planned for 2017.
3) The new techniques have been compared with current techniques on data from XDS and a presentation of these results were given to XDS on 15 January 2016.
4) Different methods of income modelling research is on-going .

**Status:** In process

**Highlights:**

1) This research increased XDS's insight into binning and variable selection techniques.

**Issue:** None.

# References

ANDERSON, R. (2007). The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford: Oxford University Press. 723p.

CHMIELEWSKI, MR; GRZYMALA-BUSSE, JW. (1996) Global discretization of continuous attributes as preprocessing for machine learning. INTERNATIONAL JOURNAL OF APPROXIMATE REASONING. November 1996. 15(4):319-331.

ENGELMAN, B. and RAUHMEIER, R (Eds). (2011) The Basel II Risk Parameters: Estimation, Validation, and Stress Testing. Berlin, Springer, 2011 (second edition).

FAYYAD, U. M., and IRANI, K. B., (1992) On the handling of continuous-valued attributes in decision tree generation, Mach. Learning 8, 87-102, 1992.

HAMERLE A, RAUHMEIER R, ROSCH D (2003), Uses and Misuses of Measures for Credit Rating Accuracy, Working Paper

MCNAB, H. & WYNN, A. (2000). Principles and Practice of Consumer Credit Risk Management. Canterbury: Financial World Publishing. 222 p

NGUYEN H-V, MüLLER E, VREEKEN J, BöHM K. (2014). Unsupervised interaction-preserving discretization of multivariate data. Data Min Knowl Discov (2014), 28:1366–1397.

OLIVEIRA, I., CHARI, M., and HALLER S. (2008). SAS/OR®: Rigorous Constrained Optimized Binning for Credit Scoring. SAS Global Rofurm 2008: Paper 153-2008, SAS Institute Inc., Cary NC

QUINLAN, J. R., (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, Calif., 1993.

SAS Institute Inc. (2010). SAS Institute Course Notes. *Predictive Modeling Using Logistic Regression.* NC: SAS Institute Inc. 200 p

SIDDIQI, N. (2006). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. New Jersey: John Wiley & Sons. 196p

SIDDIQI, N. (2006). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. New Jersey: John Wiley & Sons. 196p

THOMAS L.C. (2009). Consumer credit models: Pricing, Profit and Portfolios. Oxford University Press. 385 p.

# CV

- Tanja Verster associate professor at the Centre for Business Mathematics and Informatics (BMI) at North-West University, South Africa, began her career with a Masters degree in Quantitative Risk Management (at BMI). After working at First National Bank as a quantitative analyst, she continued her career at the North-West University at Centre for BMI as a lecturer in May 2003. She completed her PhD in Risk Analysis (focussing on credit scoring) in the Centre for BMI in 2007 and currently lectures post-graduates in courses ranging from credit scoring and predictive modelling to data mining. She is also involved in applied research projects (focussing on predictive modelling in the credit scoring environment).

*Tanja.Verster@nwu.ac.za*