Enhancing modelling in a regulatory environment with machine learning

Dr. Jack Davies, CQF
True North Partners LLP
Jack.Davies@tnp.eu

Overview

- *TNP compared performance and relative development effort involved in modelling likelihood of customer contact success using two competing models*
- *We found insights from machine learning (ML) ensemble modelling can be used to enhance the logistic regression development process*

Logistic regression

- Manually screened features, applied weight of evidence (WoE) / dummy encoding and fitted stepwise logistic regression to develop candidate model:

Model Gini: 26%

Main benefits
- Understanding of data and control over binning

Main limitations
- Time-consuming, both in reducing # of features and obtaining optimal binning of features

Gradient boosting

- Applied gradient-boosted decision tree to get final candidate model
- Outperformed logistic using same variables:

Model Gini: 30%

- When introduced with original features, approach identified additional predictive features

Main benefits
- Reduced relative development effort and modelled relationships better between features

Main limitations
- Less transparent

Advanced approach

Reduced development effort of gradient-boosted decision tree, coupled with model interpretability techniques provides developers with a *robust toolkit* to enhance traditional modeling.

Feature importance ranking of boosted model provides good *starting shortlist of likely predictive variables*, allowing developers to concentrate efforts early on key features.

Partial dependence plots (PDP) provide insights into prediction levels across range of feature attributes - can *reduce feature binning effort*.

Developers can enhance traditional modelling efficiency using gradient boosting derived feature importance and variable binning.

Logistic regression

Applies an enhanced traditional scorecard development approach

Variable analysis

Variable type
- Numerical / categorical

categories
- Discrete / continuous

Missing values
- Percentage / meaning

Feature engineering
- Transformation / new features

- Removed scarce / inconsistent variables; translated others as needed

Initial screening

Assess variables univariately using two approaches, using Gini for ranking ability:

Univariate logistic regression using PROC LOGISTIC (SAS)
- No special treatment for missing values
- Continuous variables not discretised

Univariate classification tree using PROC HPSPLIT (SAS)
- Missing values grouped into most suited bucket
- Max 15 leaves per tree, i.e., continuous variables discretised

- Highest-ranked features from both approaches were investigated further and considered for WoE binning

Binning & encoding

WoE binning
- Account for non-linearity
- Account for logical bin trends
- Account for stable bin volumes

Assess binning to gauge prediction loss from variable discretisation
- Collapsing reduces granularity and may weaken predictions
- IV change guides bin selection, balancing simplicity and predictive power

Logistic regression

Model 1
- WoE version of variables as input
- Stepwise selection to obtain best model
- Further refinement, e.g., remove variables contributing least to overall regression

Model 2
- Dummy-coded version of variables as input
- Apply all other steps in same manner as Model 1

Gradient boosting decision trees

ML approach does not face same limitations as logistic regression

Offers several preferred characteristics

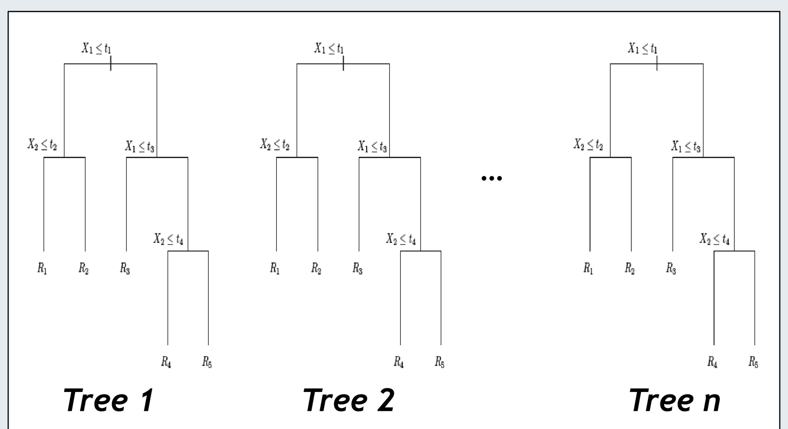
Less prone to overfitting
- Builds multiple shallow and weak trees
- Combination provides good predictive ability
- Shallower, weaker trees help prevent overfitting

Multivariate variable screening
- Tree building considers variable-target relationships in multivariate setup
- Trees used to create combined ranking of variable importance, giving multivariate view

Inherently handles non-linearity
- Multiple trees are inherently able to handle non-linearity due to their non-parametric nature

Less data preparation intensive
- Trees discretise variables inherently
- SAS allows for missing value handling by algorithm, so no upfront missing value treatment is required

Gradient-boosted decision tree



- Slower learner building multiple shallow trees
- Trees developed sequentially, i.e. subsequent trees developed using outcome of previous trees
- Sequential resamples data, weighting observations with residuals from previous trees
- Next tree aims to improve on errors of previous
- Approach is non-parametric

Limitation: Model interpretability

Note this work was originally performed by Annemie Badenhorst (Annemie.Badenhorst@tnp.eu)

Considerations

Two separate boosted models with different favourable characteristics were developed to obtain insights for improving logistic regression model

Model 1: Improve logistic model fit

Use same variables as in logistic regression and fit gradient boosted decision tree

Assesses benefit of non-parametric vs. parametric
- Potential for improved modelling of non-linearity

Assess differences in bin granularity
- Potential for improved binning

More accurately capture multivariate relationships
- Potential identification of variable interaction(s)

Model 2: Re-screen all variables

Re-screen all variables used in logistic model

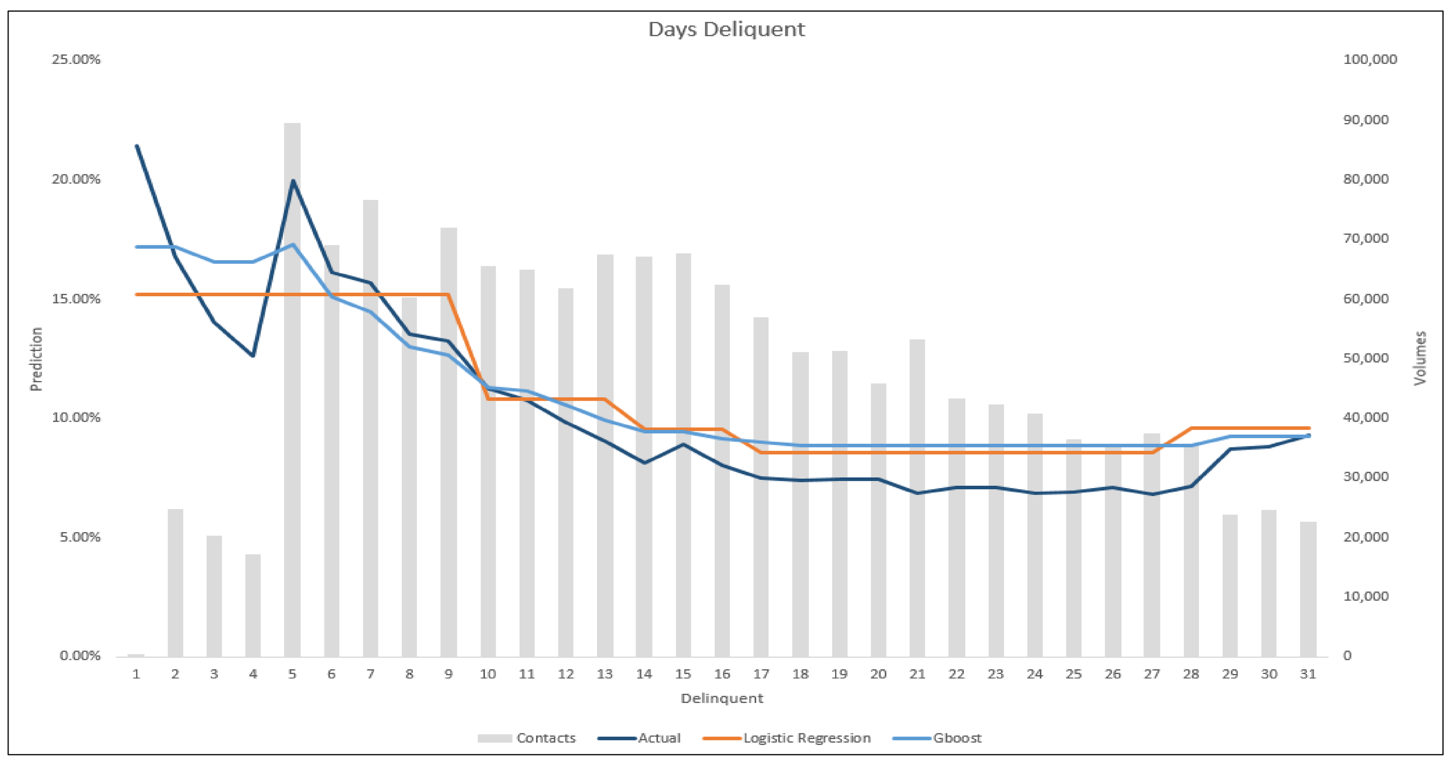
Assess if variables were erroneously eliminated through univariate screening for logistic regression
- Identify variables predictive only in multivariate setup
- Identify variables wrongly excluded in univariate screening due to strict criteria
- Identify wrongly excluded features for logistic model

Model 1

Observations
- Improvement in ranking ability can be attributed to:
  - Improved feature usage: Boosted algorithm may leverage features better and bin more effectively
  - Capturing of interactions: Boosted algorithm inherently captures variable interactions
  - Non-parametric nature: Boosted algorithm captures multivariate non-linearity better than logistic regression

Outcome
- Techniques for model interpretability can be employed to extract insights from boosted model
- PDP plots can be used to understand variable binning
- Assessments of variable importance can be done to understand differences in use of variables, etc.
- Insights obtained can be used to enhance logistic regression model, ultimately aiming to achieve comparable performance with boosted model

PDP of features provides insight that helps refine binned variables for logistic regression



Granular binning
- Observation: Boosted model inherently creates more granular bins where data volumes are sufficient, better capturing nuances in trends
- Outcome: Refinements can be made to bins for logistic regression, which in turn may improve logistic model accuracy and ranking ability

Feature ranking
- Observation: Boosted model ranks type of vehicle as second most important variable, while it ranks this feature eighth in logistic model
- Outcome: Type of vehicle is likely not optimally used by logistic regression model - re-binning and variable interactions can be investigated

Model 2

Observations & outcomes
- Observation: Boosted model applies less bins than other screening approaches - since it uses an ensemble of weaker learners, this prevented overfitting, in contrast with other screening approach where too many bins were retained
- Outcome: ML can be useful for variable screening, and can consider selection of variables without overfitting

Conclusion

- *ML approaches do not need to replace traditional modelling*
- *Can serve as an additional tool to increase development efficiency and enhance traditional modelling for compliance in a regulatory space*

Development step	Tools applied	Benefits	Applications
Overall	Develop models directly using ML	Model relationships via non-linearity and feature interactions	Challenger models; Performance benchmarks
Feature selection and variable importance ranking	Permutation feature importance; Missing value assessments	Reduced effort during initial screening; Do not need to directly treat missing values; Considers non-linearity	Obtain view of important variables with less data manipulation; Assist developers to focus on more predictive variables sooner
Feature engineering through model interpretability	Partial dependence plots	Obtain view of how variables in black box ML models influence prediction levels; Translate insights to traditional model binning	Binning with less initial expert input, less manual input and consideration to overfitting

TRUE NORTH PARTNERS
FINANCE | RISK | STRATEGY