Risk Scorecards with Machine Learning

+44 7784 695593

Fabrizio Russo[1], Torgunn Ringsjø[1], David Smith[1], James Woodcock[1], Thomas Pile[1], Lachezara Koteva[1]

[1]4most Europe Ltd.

fabrizio.russo@4-most.co.uk

4-most.co.uk

Abstract

- Improve performance of scorecard for **risk assessment at point of application**.
- Analytical approach included:
  - Horse Race (Algorithm Comparison)
  - Sample Design
  - Model Based Inference
  - Through-the-Door (TTD) Model Build
  - Transparency through application level assessment
- Results: **8% increase in AUC using transparent XGBoost** - improved discrimination with a potential increase in **auto-decisions (+33%)** and **accept rates (+12%)** as well as decrease in **bad rates (−30%)**

Motivation

1. Develop a model that provides a better risk ranking mechanism than the scorecard used to assess new business banking applications
2. Support increased automation or simplification of risk decisions
3. Model evaluation criteria include model performance as well as transparency and auditability

Horse Race

Comparison between traditional techniques and a variety of Machine Learning algorithms. The 'best' technique is taken forward for inference and final model development.

- Traditional scorecards developed through MIV iterative approach
- ML algorithms fitted on all suitable variables with minimal feature engineering
- Non exhaustive hyper-parameters tuning

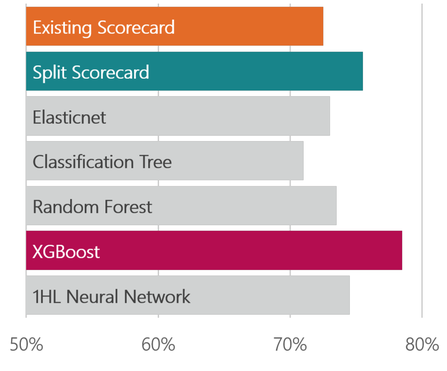

Figure 1: Algorithm Comparison - AUC on Test Data

Experiment Design

The data covered a period of two and a half years from which samples were selected to maximise stability and robustness.

- KS and PSI tests were use to ensure train, test and out-of-time (oot) sets were similar in feature and target distributions.
- Feature engineering pipeline included encoding of default, missing and ordinal values, build of financial ratios etc.
- Feature space extensively analysed to ensure no target leak

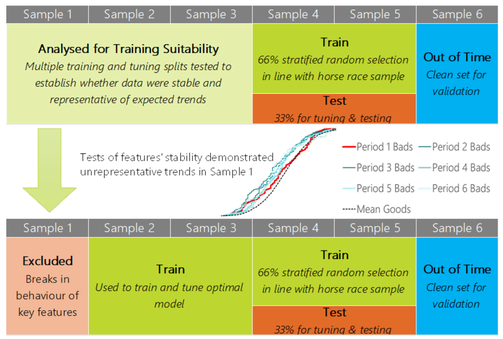

Figure 2: Sample windows selection

Model Build

Model based inference was used to obtain balanced feature importance for labelled and unlabelled observations:

- Fit a Known-Good-Bad (KGB) ranking XGBoost to taken-up population (purple)
- Fit an Accept-Reject (AR) ranking XGBoost to Through-the-Door (TTD) population (blue)
- Combine KGB and AR to infer weights to be assigned to unlabelled population
- Fit a weighted logit XGBoost to TTD population (green)

Model tuning: Grid and random search cross-validation to select hyper-parameters with highest AUC and lowest variance.

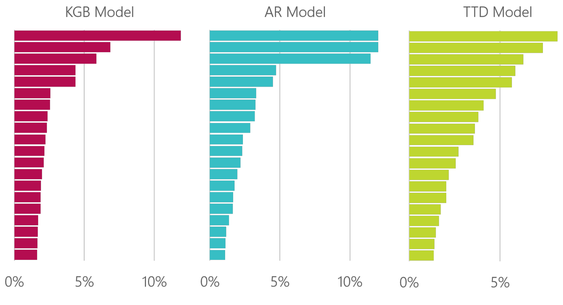

Figure 3: Comparison of feature importance distributions for KGB, AR and TTD models (top 20 - feature names omitted for confidentiality)

Performance Results

The model was tested against unseen data and results compared to existing score.

Figure 4: Comparison of good and bad distributions for oot

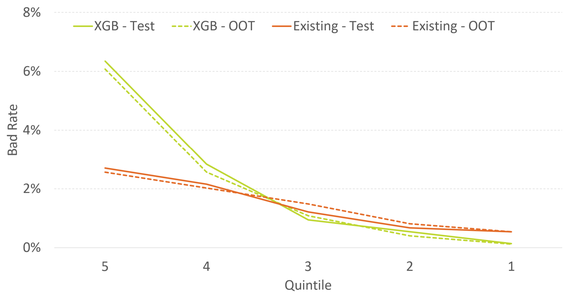

Figure 5: Bad rate by quintile (5=High PD - 1=Low PD)

As shown in Figure 4, XGBoost separates goods and bads significantly better than traditional scorecards; in Figure 5 is illustrated the potential for reduced bad rates and increased auto-decisioning through a strategy tailored to risk grades (quintile).

The features' values contribution to the final score was analysed through partial dependence plots (PDP) in combination with observed accept and bad rates.

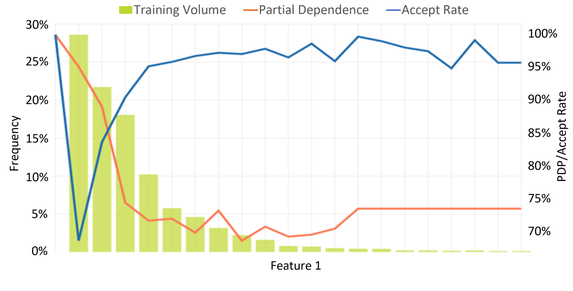

Figure 6: Feature distribution, accept rate and partial dependence

PDPs were reviewed with stakeholders applying monotonicity constraints, features binning or exclusion as applicable; this ensured model alignment with business logic and fairness requirements.

Interpretability Results

XGB explainer was used to break down the contributions of key features to individual applications ensuring transparency and auditability.

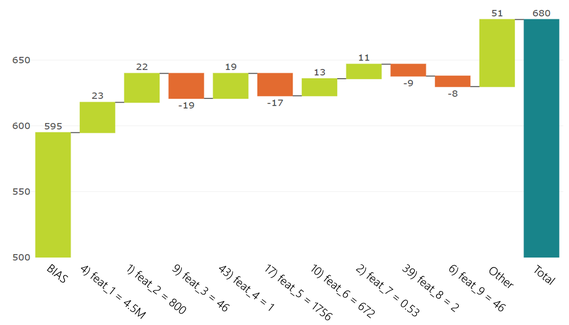

Figure 7: Score waterfall (local feature importance) for application X

Stakeholders' review of sample applications was instrumental to model sign-off, informed the cut-off strategy and the underwriter guidelines. It ensured that:

- the implemented model was calibrated in line with the business' risk appetite
- the business adopted the model's recommendations into their strategy with the confidence that they would be able to explain the decisions to both regulators and customers in line with GDPR

Conclusion & Future Work

This Machine Learning application demonstrated that ML can improve discrimination and trust in risk assessment and ultimately increase accuracy and automation of risk decisions.

1. XGBoost proved to **significantly outperform traditional scorecards'** ability to predict risk at point of application with more than **8% increase in AUC**
2. Improved discrimination meant a potential increase in **auto-decisions (+33%)** and **accept rates (+12%)** as well as decrease in **bad rates (−30%)**
3. It was demonstrated that an ML model does not have to be a black box. Both variable and observation level analysis made the score **transparency comparable to a traditional scorecard**
4. The implementation of this ML Risk Score represents a **step towards acceptance of advanced analytics in a heavily regulated environment**
5. Future work will investigate
   - How **feature interactions impact changes in contributions** for single application explanations
   - The use of more extensive yet **explainable feature engineering**