Illuminating the Black Box: Machine Learning Model Explanation

John Oxley and Eric McVittie

Credit Scoring and Credit Control XVI
Edinburgh
28 – 30 August 2019

experian™

Introduction

Machine Learning methods hold promise for many and varied applications

But they have limitations

Additional techniques are required to make most ML Models understandable

We examined XGB and RF models using SHAP, a popular explanation technique

Does SHAP explain the model?
Inconclusive

Does SHAP reflect the Data Generating Process?
Not in general

experian™

Content

- ML predictive models and conventional statistical models
- Explain-ability in the ML context
- Levels of abstraction of information
- SHAP
- The Methodology
- Synthetic Data Generation
- Results
- Conclusions and Implications

Out of Scope

- Detailed description of Random Forest, XGBoost or Lasso methodology
- Detailed description of ML model explanation methodology
- Global ML explanation. We calculate local explanations (for each observation) and compare the order of importance with the true order according to the Data Generating Process and report rank correlations for the entire data set
- Assessment of potential detriment to consumers from erroneous decisions
- Discussion of the requirements of Regulatory bodies to achieve Compliance

Machine Learning vs. conventional methods

Leo Breiman: The Two Cultures

Artificial Intelligence Research	Classical Statistical methods
Machine Learning methods including supervised learning techniques for predictive modelling	E.g. Regression, Discriminant analysis
• Best Fit Hyperplanes • Efficient ways to partition data	Investigation of causal links
Unconstrained	Discovery of the Data Generating Process
	Assumptions about data distributions and structure

What is Explainability in the ML context?

*By "explaining a prediction", we mean presenting textual or visual artefacts that provide **qualitative** understanding of the relationship between the instance's components (e.g. application data, words in text, regions in an image) and the model's prediction. For properly opaque methods, 'ML Model Explanation' is in some sense a contradiction in terms*

What drives the demand for explanations?

- Human curiosity and learning to understand the model
- To understand the data
- Detect bias
- Increase Trust and Social acceptance
- Requirement for social interactions

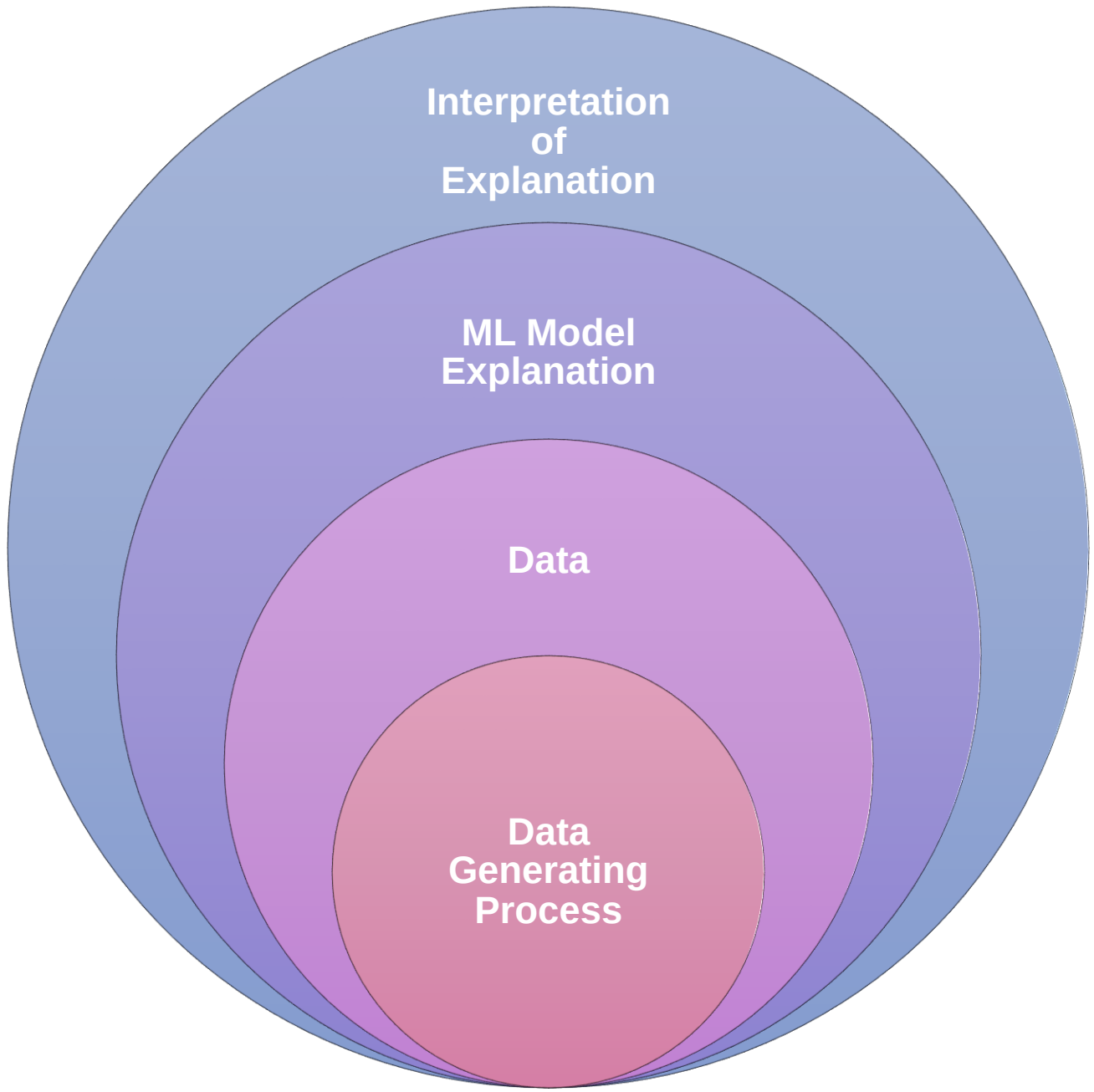
When we don't need explanations

- Model has no significant impact
- Problem is well-studied?
- When interpretation can be used to game the system?

Where is it used in Credit Risk?

- Adverse reasons for score declines
- Behavioural model that determine treatments
- Any ML model that could cause detriment
- For extrapolations e.g. optimisation scenarios

Information levels of abstraction



SHAP: SHapley Additive exPlanations

The Shapley value - a method from cooperative game theory. It defines how to distribute fairly the 'payout' among the players in a game taking into account different coalitions of players

By analogy, predictions can be explained by configuring each feature as a 'player' in a game where the prediction is the payout

The Shapely value for a variable is the average of the marginal contributions made by that variable over all permutations of variables

Computationally very expensive as the model has to be used to score the data many times

SHAP A Unified Approach to Interpreting Model Predictions (Lundberg and Lee)

An efficient Python implementation of the Shapley Value calculation primarily for trees but there exist implementations for other ML techniques

Other cooperative Game theory methods are available although Shapley satisfies 4 important axioms

Methodology

- **Initial thoughts**
  - **Insufficient to use real world data and intuition about the order of importance of features to test explanation methods**
  - **A feature which is strong in a univariate sense should be conspicuous in the explanation … UNLESS the model finds proxies for that feature**
  - **A feature which is weak in a univariate sense but predictive in an interaction term should also be conspicuous … UNLESS the model finds proxies**
  - **Multiplicity of models / flat maximum effect**

- **Fundamentals**
  - **The fundamental thing is the Data Generating Process**
    - The structural effects, defined synthetically by the linear sum of terms
  - **Generate synthetic data with a mixture of continuous and categorical features**
  - **Include interactions between these features**
  - **Generate coefficients randomly and calculate log(odds) → Prob(Default)**
  - **Construct a 'cheat' GLM LASSO model to give a best reasonably possible benchmark**
  - **Generate RF and XGB models on the original variables**
  - **Generate SHAP explanation and compare with the known variable importance from the DGP**

experian™

One Trial

10,000 – 250,000 observations

Coefficients
Random normal vector N(0,2) for 1st order variables, N(0,0.5) for interaction terms

Data Generating Process
The ‘ground truth’ defined by the linear sum of terms

e. g.

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \boldsymbol{\beta}_3 x_3 + \boldsymbol{\beta}_4 x_1 x_2 + \boldsymbol{\beta}_5 x_2 x_3 + \cdots + e = \log(odds)$$

Features
1 to 5 continuous normal +
1 to 5 categorical (2 – 5 levels)

Interactions
Full set of quadratic interactions

Noise term
Random normal

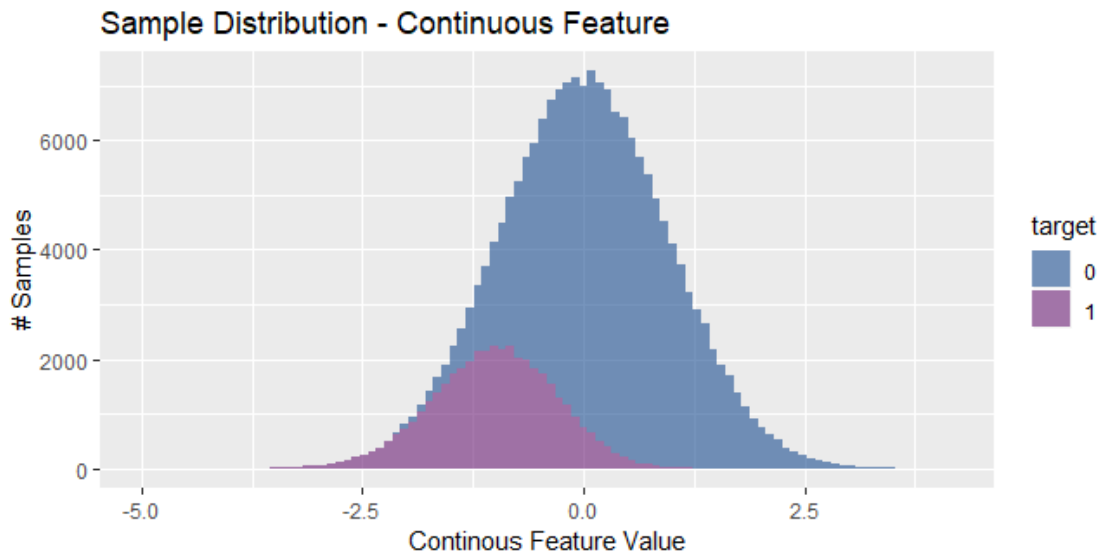
Transform log(*odds*)
Create p(Default)

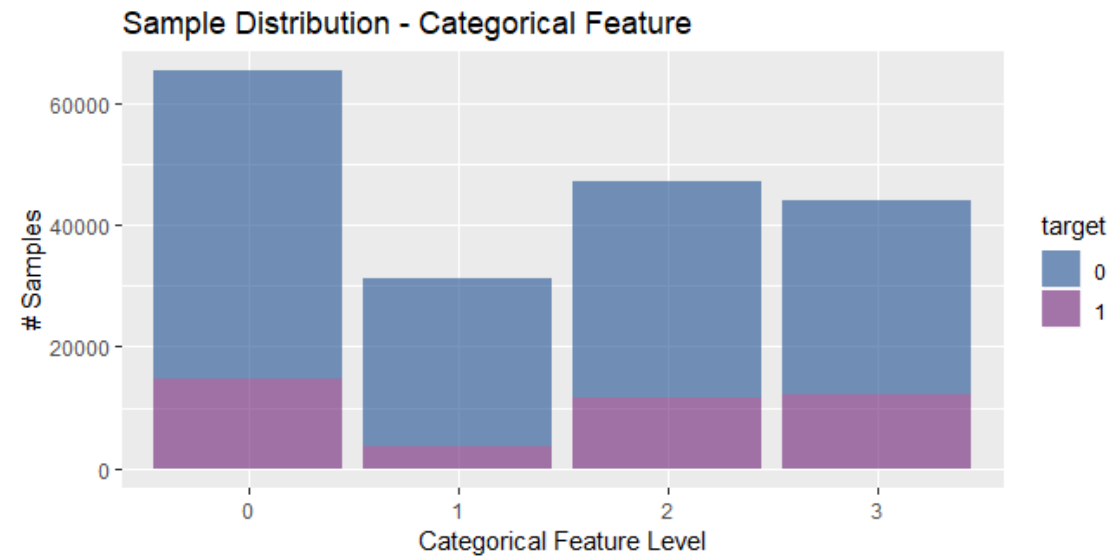
Modelling algorithms:
LASSO GLM model; RF; XGB

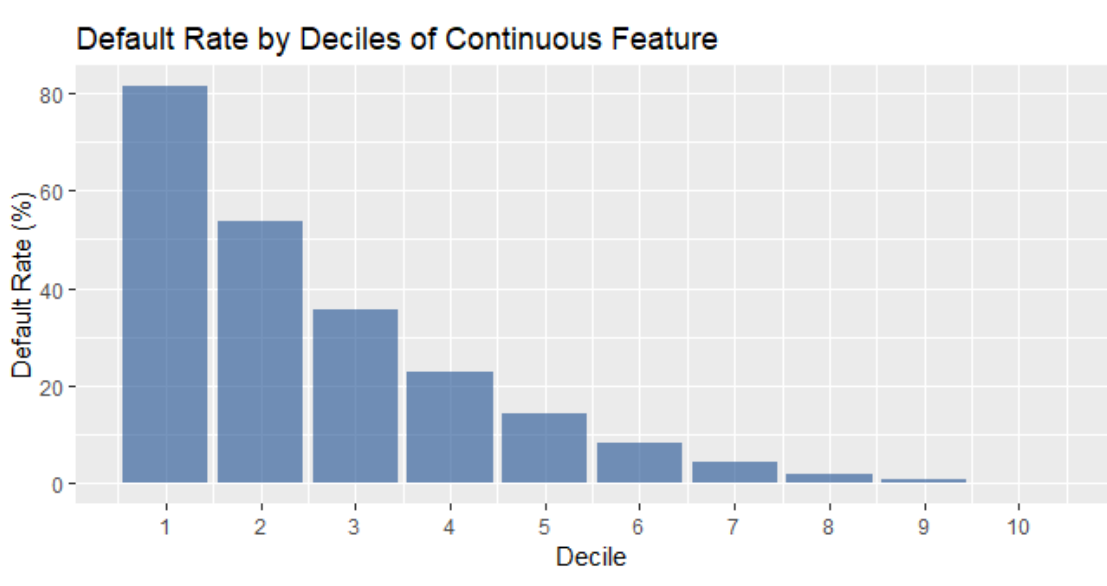
 © Experian

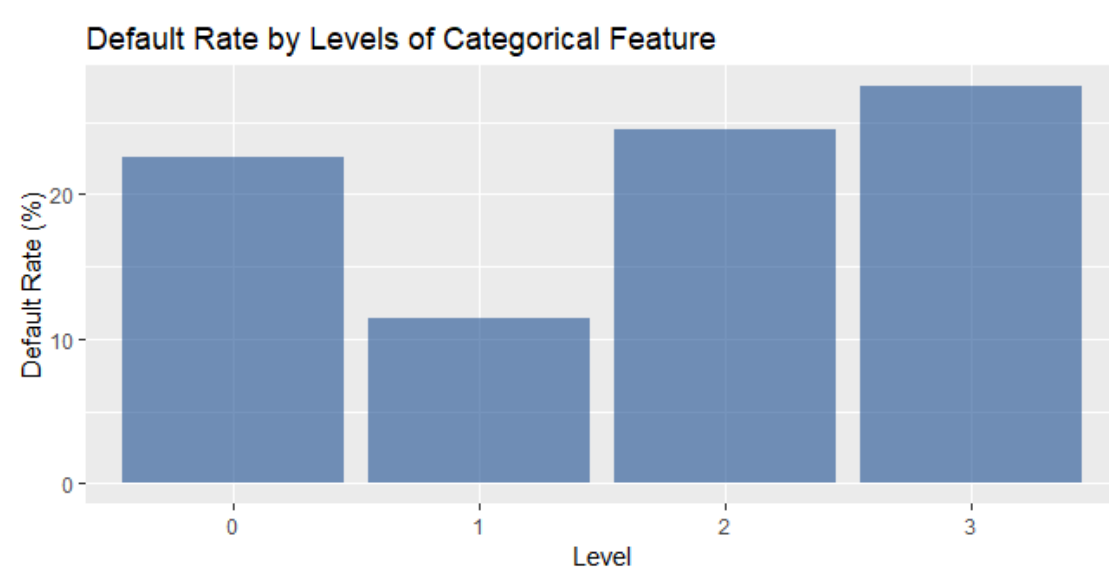
Examples of Sampled Data for a Single Trial

187,413 samples; 9 features (4 continuous, 5 categorical), target rate = 22.34%



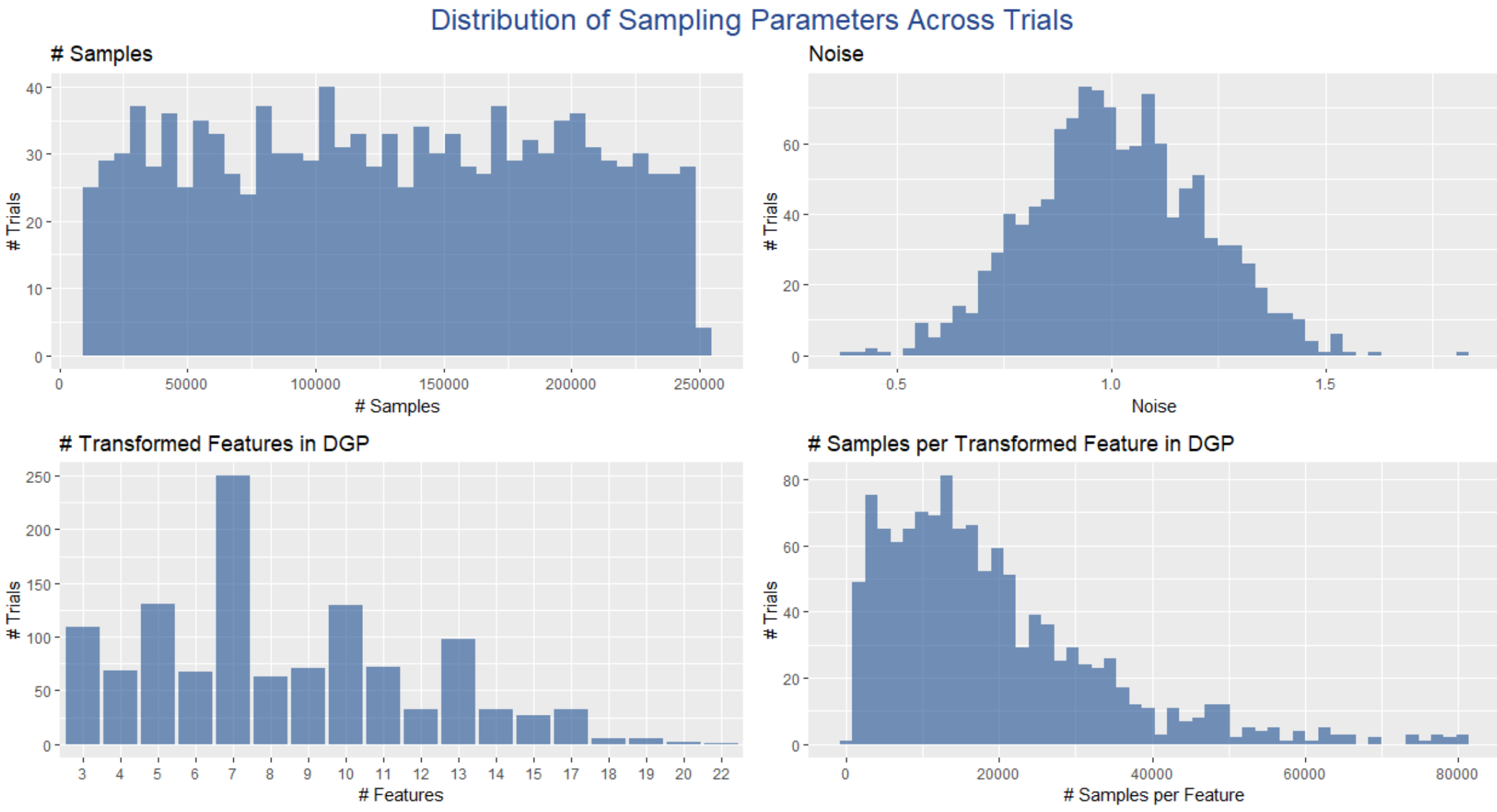






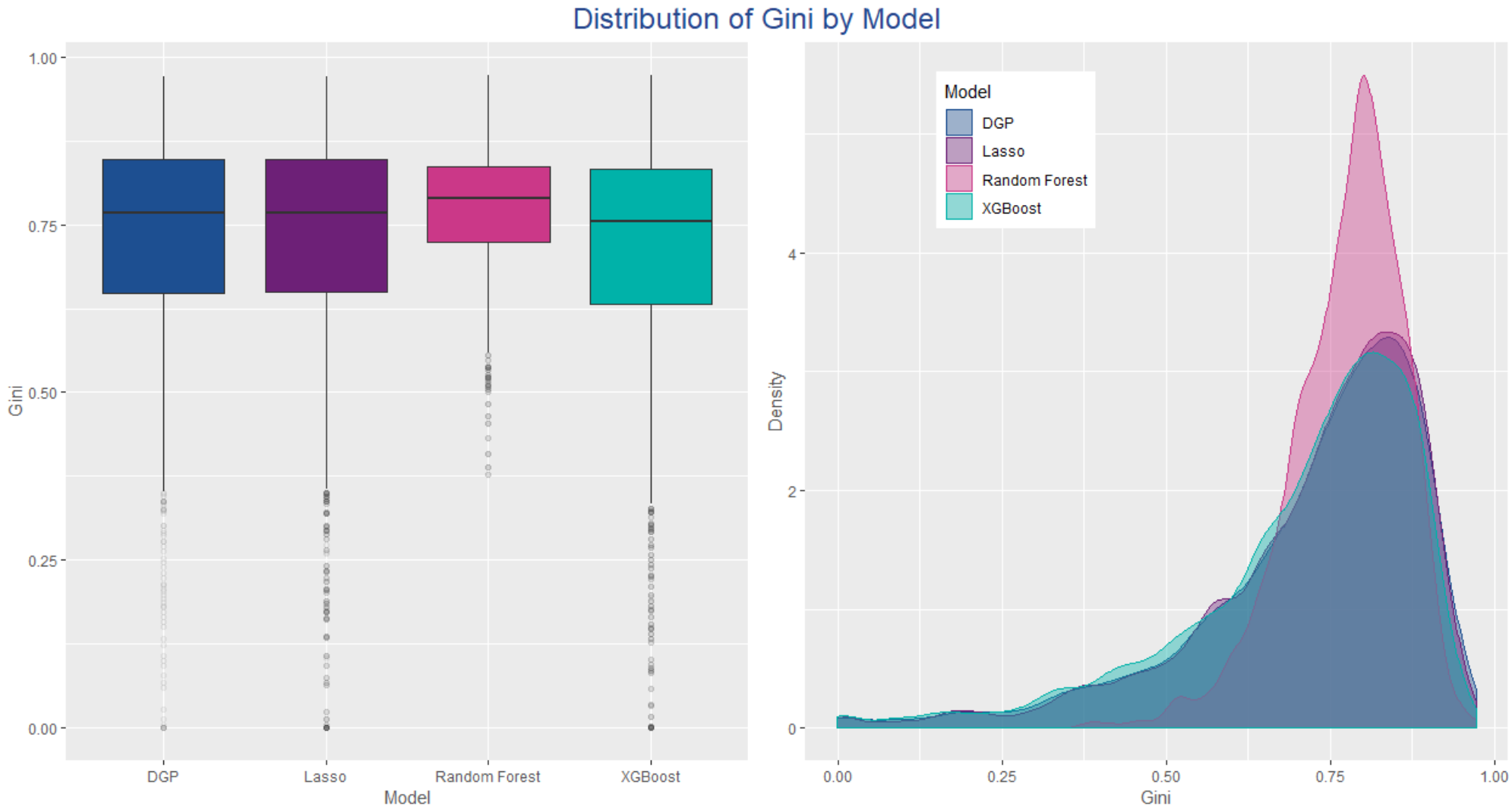
 © Experian

Variation in sampling across 1200 trials



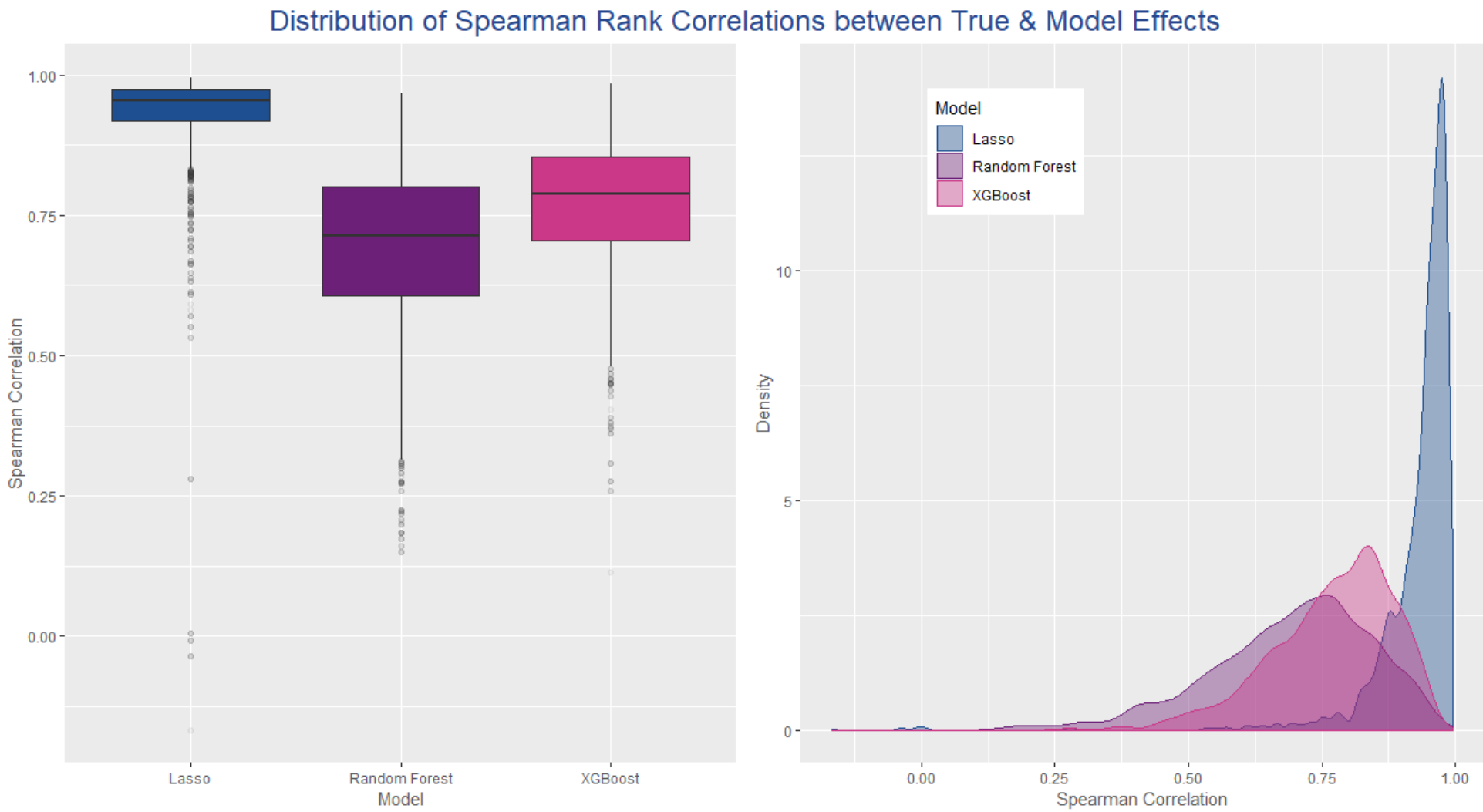
 © Experian

Distribution of Model Discrimination (Gini) in hold out samples across 1200 trials



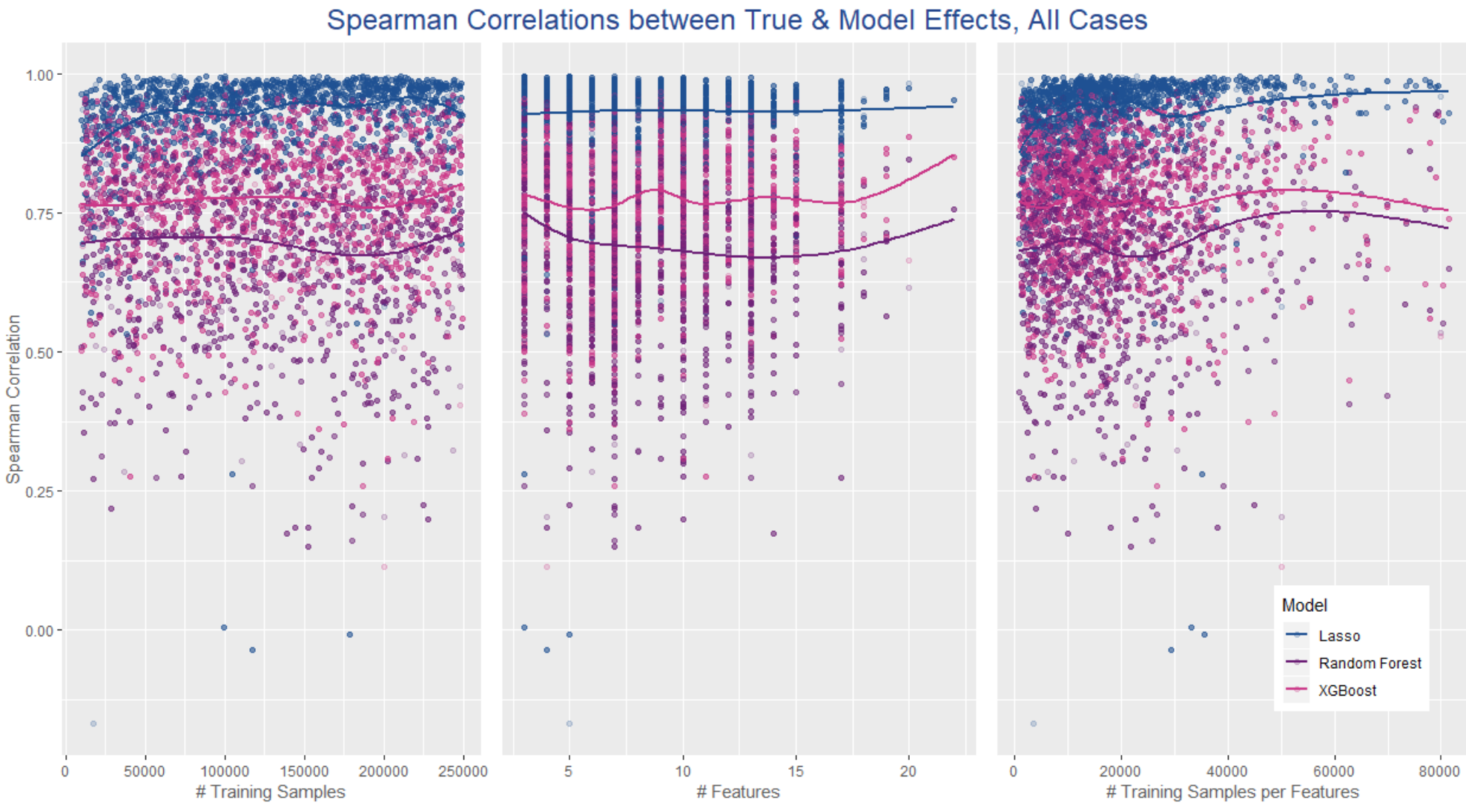
 © Experian 07/08/2019

Correspondence between true and model effects across 1200 trials based on all samples in each trial



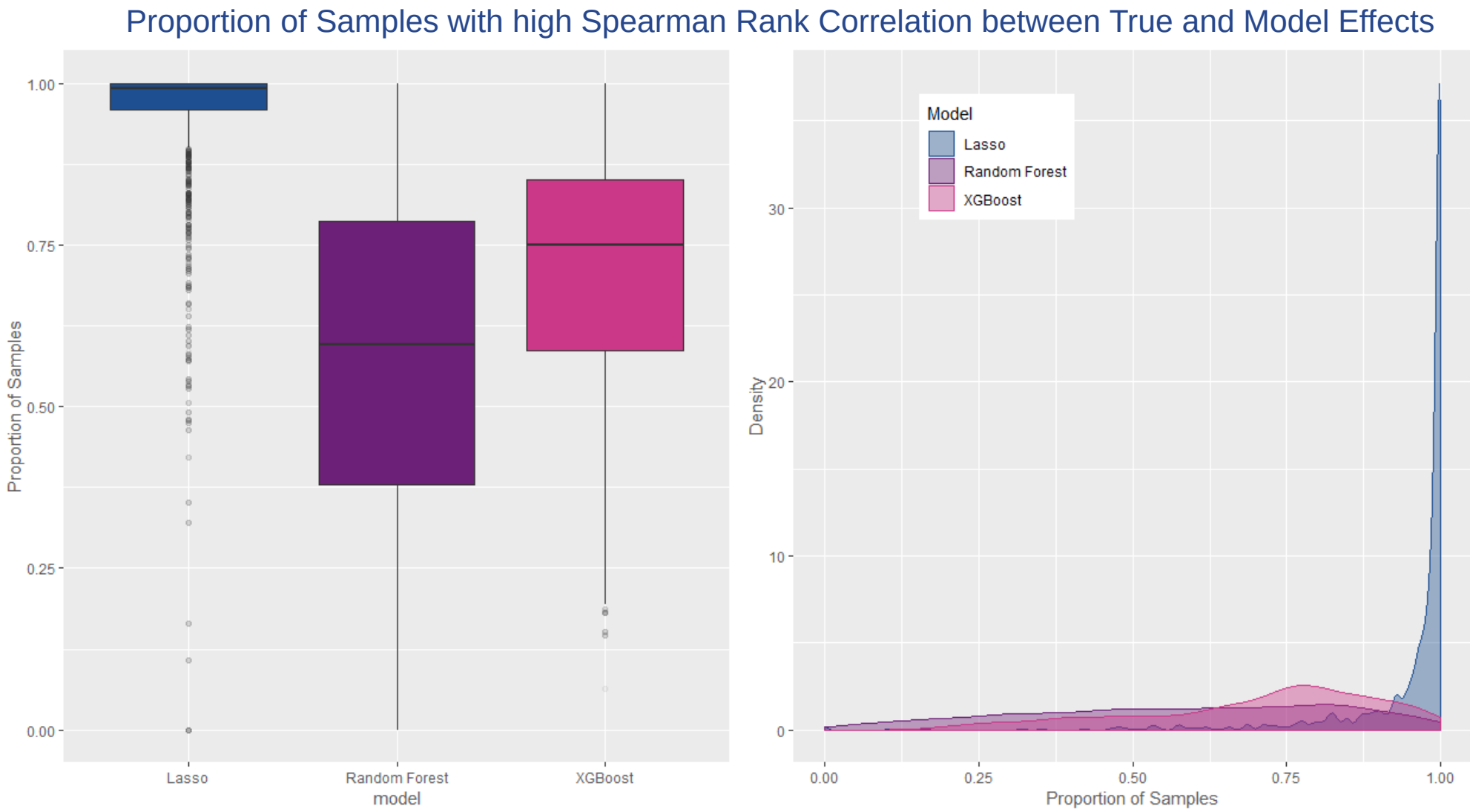
 © Experian 07/08/2019

Correspondence between true and model effects across 1200 trials



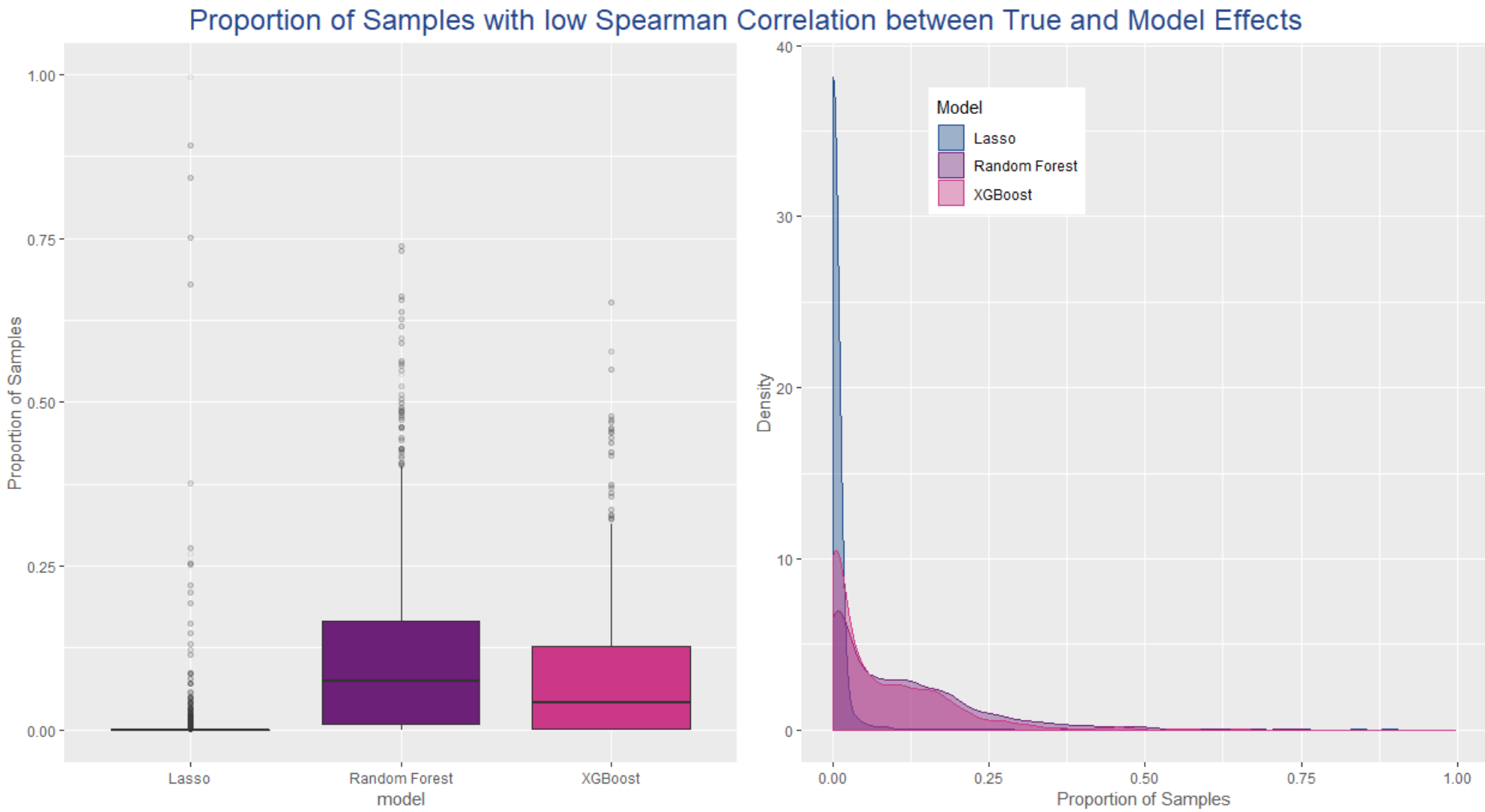
 © Experian 07/08/2019

Correspondence between true and model effects across 1200 trials



 © Experian

Correspondence between true and model effects across 1200 trials



 © Experian 07/08/2019

Conclusions

- We've used synthetic data to compare ML model explanations according to SHAP with the data's ground truth
- The performance (Gini) of the RF and XGB models is comparable with the LASSO model which incorporated the exact DGP (plus random noise) which is impressive and we infer that RF and XGB have produced effective predictive models
- However, we cannot confirm that they are in any way efficient / parsimonious – the main effects and interactions may have been modelled with proxies using correlations within the data so that predictions for new observations may not be intuitive
- Feature importance according to SHAP may reflect the ML model's interpretation of the DGP but we are unable to verify this

- *For a relatively simple DGP popular combinations of ML + explainer (RF or XGB + SHAP) have variable ability to replicate the 'true' effects in the data – and sometimes the results only weakly correspond, at best, to the way the data were generated.*
- *You can't rely on ML + explainer to reveal 'facts about the world' – if that's essential (or even just desirable) for your application, then you need to impose the required structure when defining in the model prior to training – e.g. through imposing constraints on model weights or using a Bayesian methodology.*

© Experian

Implications

- Since SHAP does not consistently represent the DGP, its ranking of feature importance cannot be used to make inferences about it
- Like ML models, conventional scorecards are ‘best fit lines’ but crafted to include structural effects explicitly:
  - **https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/03/Design-of-software-tools-for-continuous-characteristic-analysis-Gayler.pdf**
  - Counterfactuals / scenario extrapolation are thereby made less prone to errors
- If SHAP is used to derive adverse reasons for an unconstrained ML model, any remedial action by the declined applicant may improve their score but it may not improve their credit risk
- Using an ML model for scenario simulation e.g. in optimisation may not have the expected effect
- Structural effects could be forced into the ML model e.g. with monotonicity constraints but this contradicts the reason for using an ML model in the first place

- Next steps
  - Can we determine when SHAP is most and least successful?
  - Attempt to determine “Does SHAP explain the model?”

 © Experian

A few references

The Mythos of Model Interpretability, Zachary C. Lipton

- https://arxiv.org/pdf/1606.03490.pdf

Model based Machine Learning - Chris Bishop

- http://mbmlbook.com/toc.html

DARPA

- https://www.darpa.mil/program/explainable-artificial-intelligence

ML ↔ Statistical model comparison

- http://www.fharrell.com/post/stat-ml/ & https://www.fharrell.com/post/medml/

© Experian 07/08/2019 This presentation represents the opinions of the authors and not necessarily Experian