Deloitte.

The application of machine learning and challenger models in IRB Credit Risk modelling

The use in model estimation

MAKING AN IMPACT THAT MATTERS since 1845

Introduction

The recent surge in data availability and storing capacity, combined with increased computing power, creates the opportunity for machine learning (hereafter: "ML") to be applied in credit risk modelling. ML models are considered better equipped than "traditional" models to keep up with the emergence of Big Data and to detect complicated patterns and dependencies, thereby potentially leading to greater risk differentiation and higher accuracy.

The main challenge is that ML models are more complex, making their results less transparent to interpret, justify and explain to management functions and supervisors. Therefore, the incorporation of ML models in the internal ratings-based (hereafter: "IRB") model landscape has been limited.

Acknowledging that ML may play an important role in shaping credit risk modelling within financial services in the future, the European Banking Authority (hereafter: "EBA") recently published a discussion paper *EBA discussion paper on machine learning for IRB models* (EBA/DP/2021/04) (hereafter: "EBA ML discussion paper"). The paper aims to provide the supervisor's expectations and recommendations for a possible and prudent use of ML models in the context of the IRB framework.

As identified in the EBA ML discussion paper, one of the use cases for ML in IRB modelling are challenger models. Challenger models are models applied in parallel to traditional models, to benchmark model performance, explore alternative modelling assumptions and identify patterns that may not be captured by traditional models.

This blog series aims to provide insights on how ML can be incorporated as challenger models in the context of IRB modelling. This blog specifically focuses on the application of ML challenger models in the process of model estimation.

Unlocked potential

According to Article 170(3)(c) of the *Capital Requirements Regulations* (hereafter: "CRR"), the structure of rating systems should ensure meaningful:

- Risk differentiation, i.e., the final model obtains adequate discriminatory power to differentiate between riskier and less riskier observations; and
- Risk quantification, i.e., the final model obtains adequate predictive ability to allow for accurate and consistent estimations.

Model estimation in IRB context denotes the procedure by which a quantitative model is developed to estimate the PD, LGD or EAD/CCF risk parameters to ensure meaningful risk differentiation and quantification. The current industry standard is to use logistic regression models to estimate these risk parameters. Benefits of regression models are the ease of use, their interpretability and their transparency.

The power of ML models is to detect complex (non-linear) patterns and dependencies (interaction effects) between risk drivers. Using ML models as challenger models can help to obtain insights into unlocked potential in the data. However, it is pivotal to understand where the unlocked potential comes from, such that these insights can be used to improve the "traditional" regression model, or potentially replace the traditional model.

Post-hoc explainers for black-box models

The generally expected improved model performance of ML models as compared to traditional methods typically comes at the cost of interpretability. This holds especially for 'black box' ML models, such as neural networks and tree-based models. However, there has been an increasing demand to gain insight in the decision-making process of these black box models, for example to prevent bias or unethical behavior (implicit discrimination based on gender, race, religion, etc.). This led to the emergence of so-called "post-hoc explainers".

Post-hoc explainers are model-agnostic methods to analyze dynamics following from the model output (after training the model and therefore post-hoc). Post-hoc explainers can provide local and global explanations:

- Local explanations focus on why the model reaches a certain prediction for a specific observation. For example, with an input set of risk drivers, how did the model come to a certain PD for obligor *XYZ*.
- Global explanations aim to provide transparency in how the model functions as a whole. Therefore, global explanations can be considered as averages or aggregations across all local predictions.

Post hoc explainers provide input into individual feature (i.e. risk driver) importance, the relationship between the risk driver and dependent variable and interaction effects. Examples of two well-known post-hoc explainers are:

- SHAP: The SHAP framework has been developed to explain and gain global and local insight on the output of any type of (ML) model. An important aspect of the SHAP framework is the SHAP value, which is the contribution of a risk driver value to the difference between the actual prediction and the mean prediction.
- Local Interpretable Model-Agnostic Explanations (LIME): Lime aims at providing local insights into why (classifier) models reaches a specific outcome for a certain prediction.

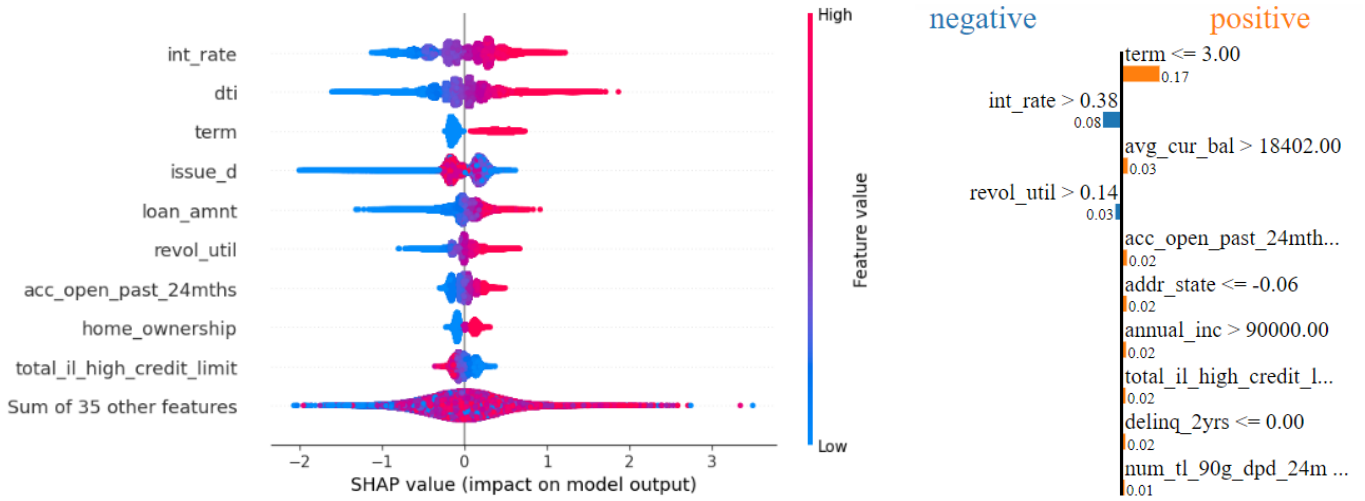

Figure 1: Visualization of SHAP global explanation (left) and LIME local explanation (right) on a CatBoost model applied to the open source Lending Club dataset.

Comparability of model outputs is crucial to draw conclusions and next steps from ML challenger models. The advantage of post-hoc explainers is that they provide the opportunity for model agnostic comparison. For example, SHAP and LIME can be run on any type of model, including black-box ML models and the "traditional" regression models, and thereby can provide insights into where the ML challenger models are outperforming the traditional models. The previous blog provided an example on using SHAP in the risk driver selection process.

However, as described in article 179(1)(a) of the CRR, the estimates following from the model should be plausible and intuitive and therefore show a clear economic link between the input and the output variables. Although post-hoc explainers provide valuable insights into black-box models, the explanations will never fully clarify the models and will always use simplifications to present the results. If the explainer would be able to fully explain the model, the explainer would be a copy of the model itself. Therefore, these post-hoc explainers only provide a partial understanding of the model and the ML model may still result in non-intuitive estimates.

Inherent interpretability from white box models

This urged researchers to explore inherently interpretable (white box) ML models. White box models are interpretable by design but still leverage advanced ML techniques in order to detect complex data patterns. Two examples of white box models are explained below.

1. Logistic model tree (LMT)

The LMT algorithm combines a decision tree with logistic regression models, where new logistic regressions are fitted at the final leaf nodes of the tree. This model is inherently interpretable because it consists of a combination of two interpretable models.

Figure 2 shows an example of how an LMT is created: by moving down the decision tree, the data is split into new leaf nodes based on slicing of risk drivers. For example, 110,169 of the 240,360 observations in leaf node 0 move to leaf node 1 based on that risk driver 21 is smaller than or equal to 0.25 for those observations. This split of data continues until the overall weighted loss of your fit to the data given a specified set of hyperparameters cannot be further decreased, often calculated using the Gini index loss; implying that the data is appropriately split into sub-sets of data.

At the final leaf nodes, indicated by light-blue in 2, logistic regressions are fitted on the sub-sets of the data, such that the overall LMT fits well on the total train data. As shown in the "Final model fit" in Figure 2, the LMT is implicitly able to capture non-linear patterns in the data.

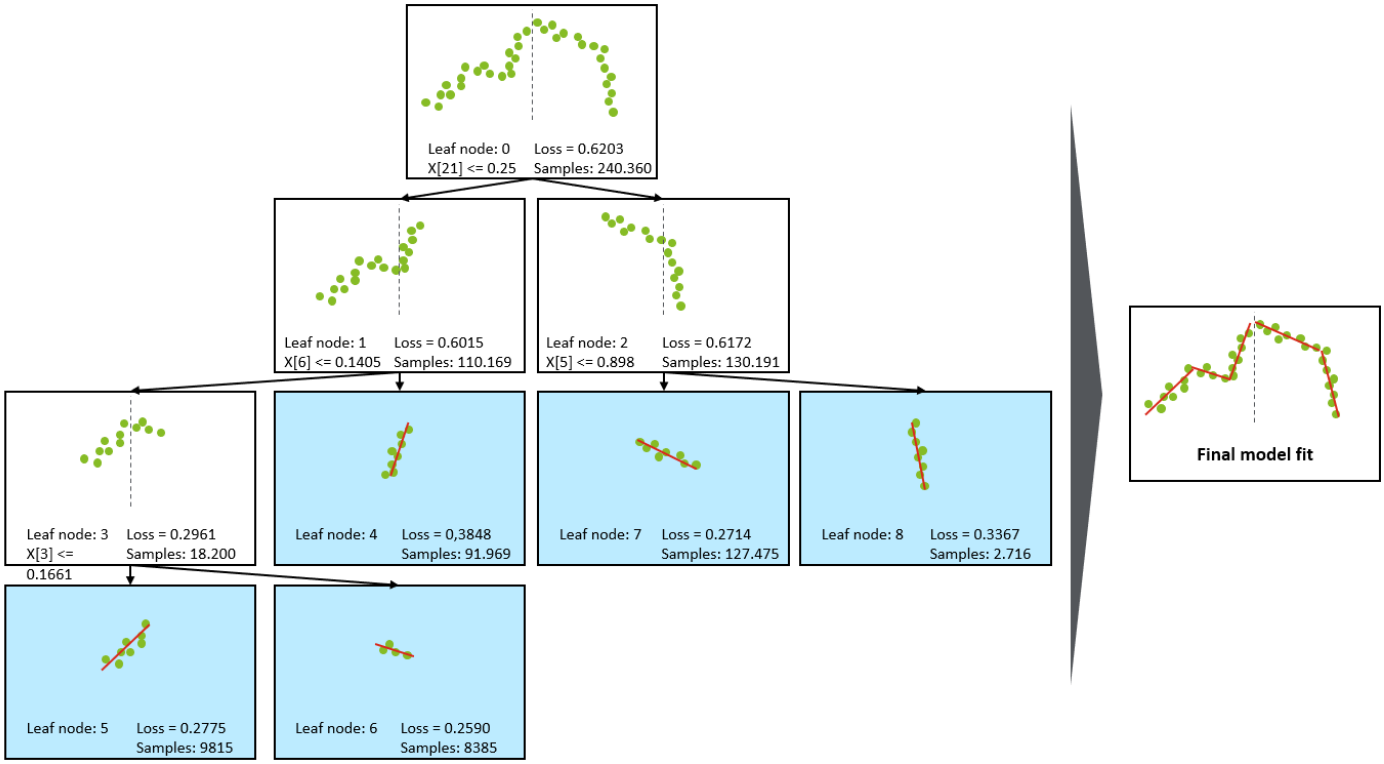

Figure 2: Visualization of an LMT with the number of samples and average loss at each leaf node.

The non-linear patterns can be further detected in visualisations like Figure 3, which shows the coefficients of selected risk drivers for the different logistic regressions fitted at the final leaf nodes. It can be seen that the coefficient values of a single risk driver can vary significantly across the leaf nodes. For example, the coefficients for risk driver "*loan_amnt*" varies between approximately 0.2 for leaf node 8 and 1.75 for leaf node 6, indicates a possible non-linear relationship to the dependent variable. In contrast, risk driver "*mo_sin_rcnt_tl*" shows its coefficients are concentrated around 0.1, indicating a possible linear relationship to the dependent variable.

A big advantage of the LMT compared to "black-box" models, is that the LMT can be used to directly compare the coefficients from your "traditional" single logistic regression model to the coefficients calculated by the LMT. If the coefficient of your traditional model does not compare well to LMT coefficients, the decision tree can provide further insights into the root cause of the difference.
Finally, it is important to note that similarly as tree-based models, such as random forests, the LMT is prone to overfitting and it is always recommended to test the model both out-of-sample and out-of-time, as also described in Article 175(4)(b) of the CRR.

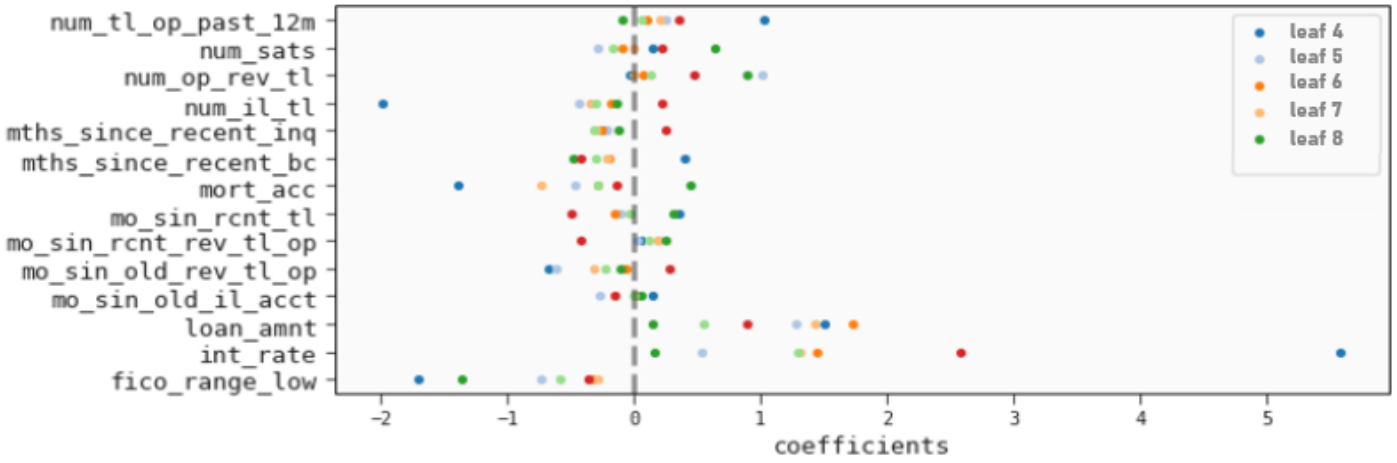

Figure 3: Plot of the coefficients as calculated by each leaf node for a selection of features.

2. Generalized additive model with structured interactions (GAMI-Net)

Multi-layered (deep) neural networks are often described as the most difficult kind of black box to interpret. The GAMI-Net is built with the intention to create explainable neural networks. The GAMI-Net consists of sub-sets of neural networks, where each network is designed to:

- capture one single main effect, i.e., relationship of one single risk driver to the dependent variable; or
- one pairwise interaction effect, i.e., the relationship of two interacting risk drivers to the dependent variable. Note that the GAMI-Net only checks interactions between two risk drivers and therefore does not consider higher-order (more than two) interactions.

The main effects are first trained in the model, after which the pairwise interactions are fitted to remaining unexplained data patterns (the residuals). The output of the neural networks are (non-linear) shape functions, which represent the predicted relationship of each single neural network to the dependent variable.

Although the shape functions in the GAMI-Net are constructed by a deep neural network, which are not intrinsically interpretable, the final model is. That is because the neural networks are only used to compute the shape functions and will be disregarded from the model afterwards.

The shape functions can be visualised like in Figure 4, where 1D line plots are used for numerical risk drivers, bar charts for categorical and partial dependence heatmap plots for pairwise interaction effects. The final model is a combination of the mentioned elements. These graphs clearly show that the GAMI-Net is able to capture non-linear relationships and interaction effects.

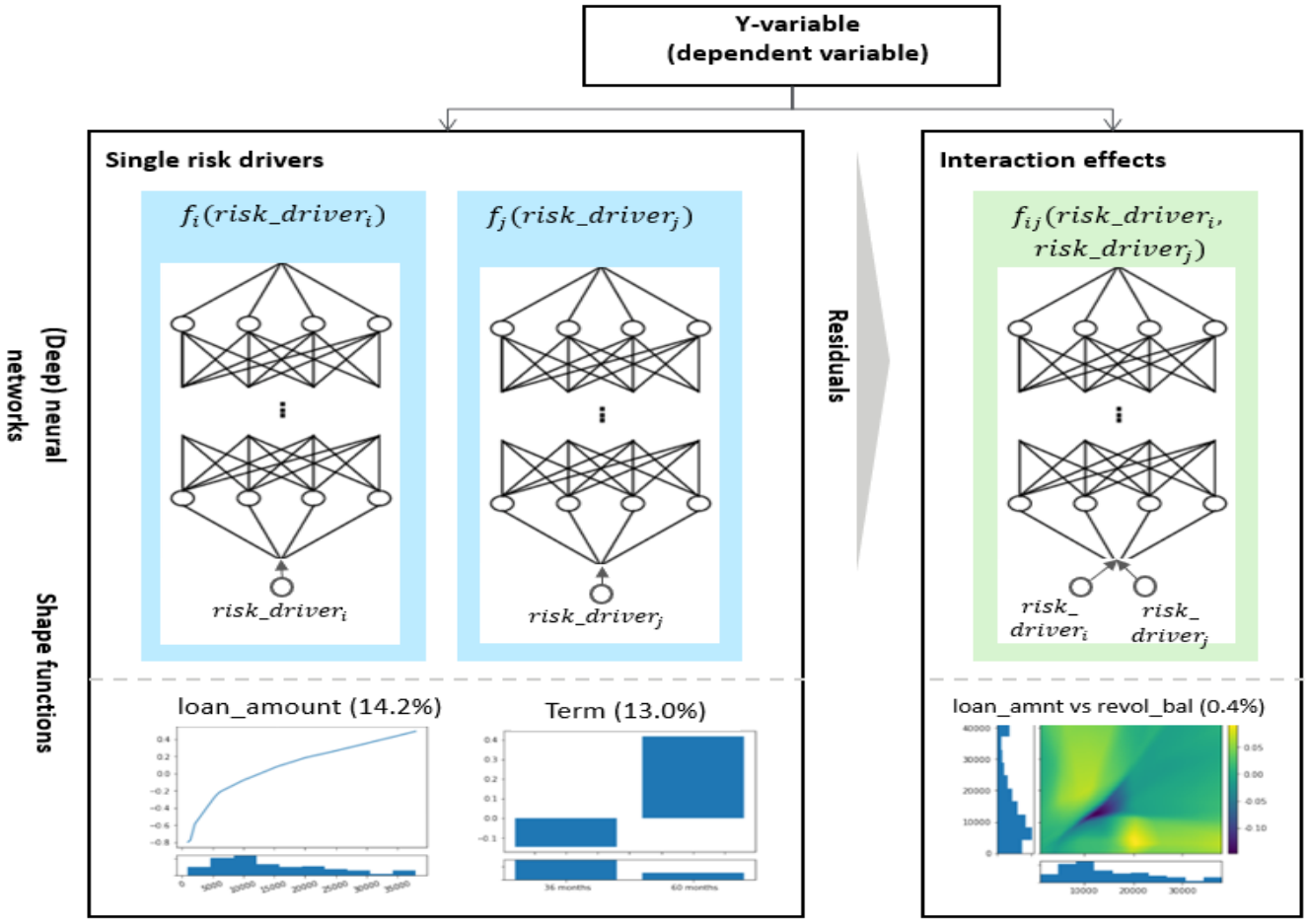

Figure 4: GAMI-Net neural networks and its translation to main effects and interaction effects.

The percentages shown in Figure 5 on top of the shape function plots represent the amount of variance explained by the risk driver, which can be interpreted as the importance of the risk driver. The values correspond to the ones shown in Figure 5, where the importance of each risk driver and pairwise interactions in the model is ranked.

Figure 5: Feature importance of all main effects and interaction effects in the model.

Conclusion

Currently, “traditional” regression models are most commonly used for the estimation of IRB credit risk components. Traditional models have several advantages, including the fast training time, ease of understanding and its simplicity. However, in terms of model performance, there are limitations as non-linear dynamics and interaction effects are not implicitly taken into account.

This blog described how unlocked data potential can be extracted and visualised from ML models to challenge and improve the traditional models. Firstly, insights from so-called “black box” models, like the random forest and neural network, can be extracted with post-hoc explainers; such as LIME and SHAP for, respectively, local and global explanations. Another advantage is that post-hoc explainers provide the opportunity for model agnostic comparison and can be run on both black-box ML models and the “traditional” regression models. A disadvantage is that post-hoc explainers only provide a partial understanding of the model and the ML model can still produce non-intuitive estimates.

Secondly, “white box” models were discussed, which are inherently interpretable but still leverage advanced ML techniques in order to detect complex data patterns. The Logistic Model Tree can be used to challenge non-linear effects that are not captured by model coefficients in the “traditional” regression approach. The GAMI-Net model can additionally be used to detect non-linear effects via the estimated shape functions, but also to provide an intuitive solution to detect and challenge interaction effects via among others the feature importance plots.

Authors

Marco Folpmers

Reinout Kool

Arwyn Goos

Charlotte Prins

Wouter Hottenhuis

Ime Meulenbelt