ARGUMENTA OECONOMICA
No 1 (50) 2023
ISSN 1233-5835;  e-ISSN 2720-5088

*Gero Szepannek*, Karsten Lübke***

# How much do we see? On the explainability of partial dependence plots for credit risk scoring

Risk prediction models in credit scoring have to fulfil regulatory requirements, one of which consists in the interpretability of the model. Unfortunately, many popular modern machine learning algorithms result in models that do not satisfy this business need, whereas the research activities in the field of explainable machine learning have strongly increased in recent years. Partial dependence plots denote one of the most popular methods for model-agnostic interpretation of a feature's effect on the model outcome, but in practice they are usually applied without answering the question of how much can actually be seen in such plots.

For this purpose, in this paper a methodology is presented in order to analyse to what extent arbitrary machine learning models are explainable by partial dependence plots. The proposed framework provides both a visualisation, as well as a measure to quantify the explainability of a model on an understandable scale. A corrected version of the German credit data, one of the most popular data sets of this application domain, is used to demonstrate the proposed methodology.

**Keywords:** credit scoring, interpretable machine learning (IML), partial dependence plot (PDP), explainability

**JEL Classification:** C52
**DOI:** 10.15611/aoe.2023.1.07

*Quote as:* Szepannek, G., Lübke, K. (2023). How much do we see? On the explainability of partial dependence plots for credit risk scoring. *Argumenta Oeconomica*, 1(50), 137-150.

## 1. Introduction

During the last few years several frameworks for automated machine learning (autoML, Hutter et al., 2018) have been proposed. One such framework is provided by the R package mlr3 (Lang et al., 2019). It allows to define a chain of modelling steps, including data preprocessing operations such as dimensionality reduction and imputation up to the final model evaluation using different strategies such as cross-validation, bootstrap and also holdout sets. All model specification choices can be

** Stralsund University of Applied Sciences, Germany. ORCID: 0000-0001-8456-1283.
** FOM University of Applied Sciences, Dortmund, Germany.

defined as so-called hyperparameters, and algorithms are provided in order to optimise these hyperparameters with regard to a predefined performance measure. As a consequence of the free availability of tools such as mlr3, the use of complex machine learning algorithms has been facilitated also for companies with comparatively low experience in this field. The resulting models are able to detect complex nonlinear multivariate dependencies without the need for the analyst to explicitly specify the kind of the functional relationship of the dependence. For this reason, such models are often called black box models.

In the application context of credit risk scoring, traditionally white box logistic regression models (Crook et al., 2007; Szepannek, 2022) are frequently used in business practice. Nonetheless, numerous benchmark studies have shown that properly parametrised modern machine learning algorithms, such as random forests and gradient boosting, are often of superior predictive accuracy compared to the aforementioned traditional scorecard models (for an overview cf. Louzada et al., 2014). A comprehensive benchmark study which evaluates several algorithms on a set of domain-specific data sets on a meta-level can be found in Baesens et al. (2002) and has been updated by Lessmann et al. (2015). The specific situation of unbalanced classes was addressed by Vincotti and Hand (2002) and Brown and Mues (2012), and investigated together with a systematic hyperparameter tuning for several classes of machine learning algorithms in a comprehensive benchmark study (Bischl et al., 2014). In Crook et al. (2007) and Szepannek (2017), the current challenges are discussed in a broader context, e.g. reject inference (Banasik and Crook, 2007), the Basel 2 accord (Basel Committee on Banking Supervision, BCBS, 2005), and profit scoring (Verbraken et al., 2014).

In order to prevent the concomitant lack of model understanding, the BCBS established a number of requirements on transparency from the perspective of regulation. The ”selection of certain risk drivers and rating criteria should be based not only on statistical analysis, but the relevant business experts should be consulted on the business rationale and risk contribution of the risk drivers under consideration” (European Banking Authority, 2017). This underlines the need for an appropriate methodology to understand what the models have learned, and still for their explanation.

According to Szepannek and Aschenbruck (2020), there can be different requirements to the explanation of a model depending on the context. Several authors recently applied methods of interpretable machine learning to credit scoring (Biecek et al., 2021; Bussmann et al., 2020; Dastile and Celik, 2021; Demajo et al., 2020; Torrent et al., 2020). In Bücker et al. (2021), the different requirements are linked to the corresponding methodology within a unified framework for *Transparency, Auditability and eXplainability for Credit Scoring* (*TAX4CS*). According to this, the methods can be distinguished into either *global* explainability on the model level such as variable importance (Breiman, 2001), partial dependence (PD, Friedman 2001), or accumulated local effects (ALE, Apley, 2016), or *local* explainability on the level of individual predictions such as Shapley additive explanations (SHAP, Strumbelj and Kononenko, 2014), breakdown plots (Staniak and Biecek, 2018), or local interpretable model explanations (LIME, Ribeiro et al., 2016). Many of them can be accessed via the DALEX framework (Biecek, 2018).

This paper concentrates on partial dependence, denoting a popular and well-known method for a model-agnostic assessment of a feature's effect on the model outcome. Despite its popularity and its frequent use in practice, partial dependence analysis is usually applied without addressing the corresponding question how much can actually be seen in the resulting plots. For this purpose, the methodology is presented in order to analyse to what extent arbitrary machine learning models are explainable by partial dependence plots. The proposed framework provides both a visualisation as well as a measure to quantify the explainability of a model on an understandable scale.

Molnar et al. (2020a) pointed out that the superior performance of complex machine learning models results from their ability to detect high order dependencies and nonlinearities. Such dependencies are difficult to understand for analysts, while it has to be noted that the trade-off between predictive accuracy and interpretability is not necessarily given for any data (Rudin, 2019). As a potential solution, criteria are proposed that help to quantify the interpretability of a model. Model selection can thus consist in multi-objective optimisation of both predictive performance and interpretability. The approach followed in this paper differs from this in the sense that it assumes an existing model (which may be the one with the largest predictive accuracy). Afterwards, the question addressed is ”How much can we see in the interpretation given by the partial dependence plots for a given model?”

In Section 2, partial dependence is reviewed. Based on this, a measure is presented that allows to quantify how far it explains a given model. In the case study, the methodology is applied to the real-world context of credit scoring using the South German credit data (Groemping, 2017; Szepannek and Luebke, 2021). In Section 3, an algorithm is presented that can be further used to identify a subset of variables which best serve to explain model. Finally, the research results are summarised in Section 4.

## 2. Quantifying explainability

### 2.1. Partial dependence

Referring back to Friedman (2001), partial dependence plots (PDP) are a popular tool to understand the effect of one or several features w.r.t. the output of a predictive model. One of their advantages is that they can be used for different kinds of predictive models $\hat{f}(x)$: the set of predictor variables $x = (x_s, x_c)$ is split into disjoint subsets and the partial dependence function for a subset $x_s$ is given by:

$$PD_s(X) = \int \hat{f}(X_s, X_c)\, dP(X_c) \tag{1}$$

This means that a partial dependence function computes the expected prediction given $X_s$ takes the values $x_s$. For a data set with $n$ observations, it is estimated by:

$$\widehat{PD}_s\left(x_s\right)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_s,x_{ic}\right), \tag{2}$$

where $x_{is}$ are the values that observation $i$ takes in $X_s$. Note that for $X_s = X$ it is $PD\left(X\right)=\hat{f}\left(X\right)$ corresponds to the model itself and for $s=\varnothing$ or in other words $X_c = X$, the partial dependence function ends up in:

$$PD_{\varnothing}=\int\hat{f}\left(X\right)dP\left(X\right), \tag{3}$$

which is a constant that can be estimated by $\frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_i)$.

## 2.2. Application to the South German credit data

The South German credit data is publicly available at the UCI ML benchmark repository (Dua and Graff, 2017) and has been made available by Groemping (2019; see also Szepannek and Luebke, 2021). It has 1000 observations and 21 variables where 7 predictors are numeric and 13 are categorical plus a binary target variable. The predictable event describes the default status of a loan. The overall prior default rate on the data is 0.3. For the purpose of this paper a random forest model was trained on the South German credit data using default parameters according to Liaw and Wiener (2002), which turned out to be a good choice for this purpose (Szepannek,

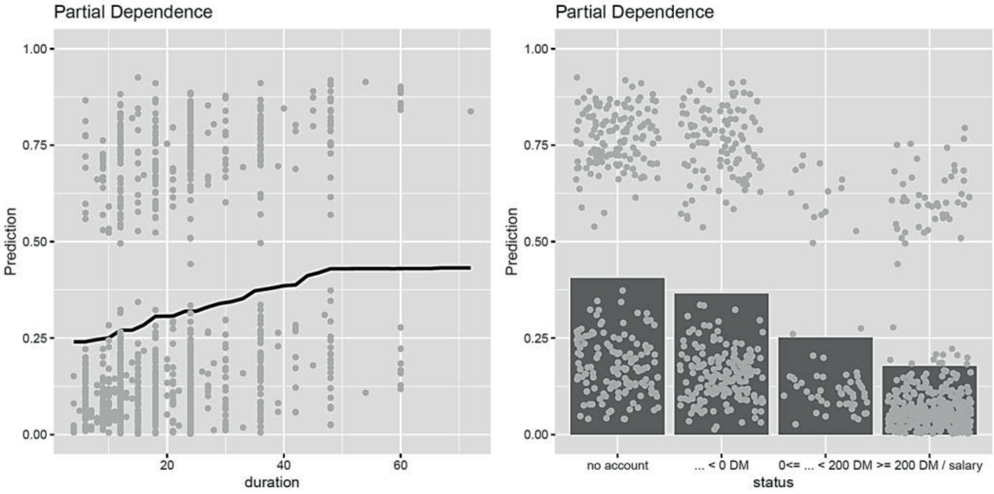

Fig. 1. Partial dependence plot for the variables duration (left, black line) and status account (right, bars), as well as the predictions on the training data (grey dots).

Source: authors' own.

2017). Usually, the data are split into training, validation and test sets in order to ensure a proper model selection and validation. As these aspects are beyond the scope of this paper, but rather the interpretability of the resulting model is of interest, no additional splits of the data were undertaken and the forest was trained on the entire data.

Figure 1 illustrates the partial dependence curves for the numeric variables duration and status account. This allows for a visual analysis of the effect of the variable on the predicted default probability and it can be easily seen that the risk (i.e. the default probability) increases for longer maturity time, whereas from roughly four years (45 months) onwards the risk stays constantly high. Analogously, it can be seen from the right plot that the risk decreases with a larger amount of money in the account. Nonetheless, when adding the predicted training data points to the graph it has to be noted that the PDP only partly explains the predictions by the models which cover a much broader range than the PDP indicates. This is obvious, as partial dependence is obtained by averaging. In turn, relying on the PD can be misleading.

### 2.3. Explainability

In the following step, a measure is derived to quantify the degree of explanation given by a partial dependence function for a model. A perfect explanation will have the same values for the partial dependence function and the predictions of the data. In this case, all points in a scatterplot of predictions vs. explanation (PX-plot) will lie on the diagonal. Such a plot is shown in Figure 2, where compared to Figure 1, the x-axis changed.

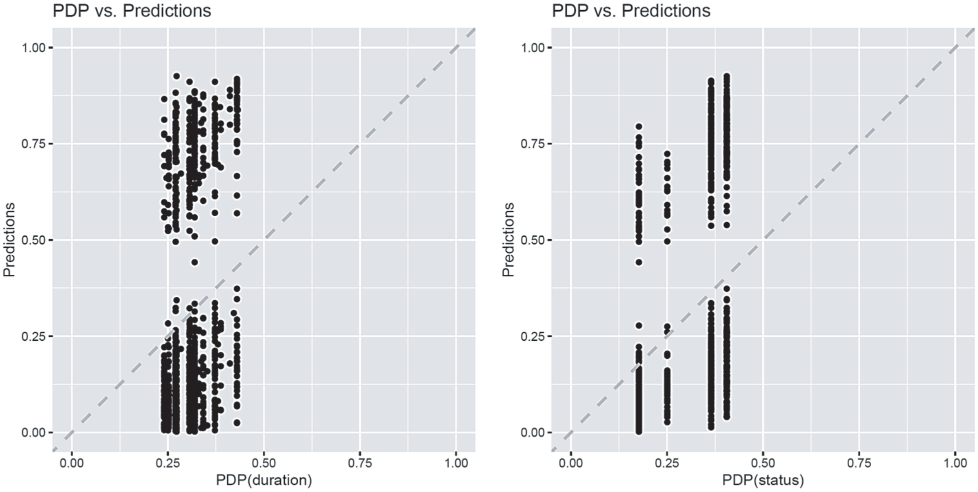

Fig. 2. Partial dependence (abscissa) vs. true predictions (ordinate) of training data for the variables duration (left) and status account (right).

Source: authors' own.

The above allows for a graphical analysis of the explainability. The more representative a PDP for a model, the closer the points to the diagonal. From this plot, it can be seen that the PDP covers a much smaller range of predicted values compared to the true model's predictions. Note that the x-axis of the right plot for the categorical variable status account takes only for distinct values – one for each category of the variable. In addition, the range of the partial dependence values is broader compared to those for the status account variable, and in particular for this variable there are only few observations with low values of the PDP $\leq 0.25$ and large predictions $> 0.75$.

In order to quantify the confidence in an explanation given by a partial dependence plot, one can measure the differences between the partial dependence function $PD(X_s)$ and the model's predictions. A natural way of doing this is obtained by computing the expected squared difference (ESD):

$$ESD\left(PD_s\right)=\int\left(\hat{f}\left(X\right)-PD_s\left(X\right)\right)^2 dP\left(X_s\right). \tag{4}$$

Note that in contrast to common error functions, the ESD does not measure the difference between predictions and observations, but instead between the partial dependence function $PD_s(X)$ and the model's predictions $\hat{f}\left(X\right)$.

For an easier interpretation $ESD(PD_s)$ can be benchmarked against $ESD\left(PD_\varnothing\right)$:

$$ESD\left(PD_\varnothing\right)=\int\left(\hat{f}\left(X\right)-PD_\varnothing\right)^2 dP\left(X\right). \tag{5}$$

The comparison of both $ESD(PD_s)$ and $ESD\left(PD_\varnothing\right)$ can be used to quantify the **explainability** $\Upsilon$ of model $\hat{f}\left(X\right)$ by a partial dependence function $PD_s$ via the ratio:

$$\Upsilon(PD_S) = 1 - \frac{ESD(PD_S)}{ESD(PD_\emptyset)}. \tag{6}$$

Note that $\Upsilon$ in (6) is somehow similar to the common $R^2$ as used in linear regression: $\Upsilon$ close to 1 states that a model is well represented by a PDP and the smaller it is, the fewer of the model's predictions are explained in the PDP. Real data plug-in estimates for $ESD(PD_s)$ and $ESD\left(PD_\varnothing\right)$ are obtained using $\widehat{PD}_s\left(x_s\right)$ and $\widehat{PD}_\varnothing$ as described above.

### 2.4. Application to the South German credit data

Table 1 (column $\hat{\Upsilon}$) shows the explainability of the random forest model on the South German credit data for all variables. Among all the numeric variables, duration has the highest explainability of only $\hat{\Upsilon} = 0.077$, which is nonetheless pretty far from 1 and thus reflects the visual impression as gained by considering Figures 1 and 2.

Columns $\hat{\Upsilon}_k$ of the table describe the explainability for increasing number of variables $k$ in the subset $X_s$ (cf. Section 3). It can be seen that for two subsets $X_s \subset X_{s^*}$, it is $\Upsilon(PD_s) \leq \Upsilon(PD_{s^*})$ with $\Upsilon(PD_s) = 1$ for $X_s = X$. The PX-plot in Figure 3 illustrates the fit of $PD_s$ from Table 1 with $\dim(X_s) = 9$ and $\hat{\Upsilon} = 0.8$ (which obviously cannot be visualised anymore). Compared to Figure 2, the PDP covers a broader range and the points are closer to the diagonal.

Table1

Explainability  for all variables of the South German credit data

| **Variable** | $\hat{\Upsilon}$ | **k** | $\hat{\Upsilon}_k$ | **Variable** | $\hat{\Upsilon}$ | **k** | $\hat{\Upsilon}_k$ |
|---|---|---|---|---|---|---|---|
| status.account | 0.221 | 1 | 0.221 | rate.to.income | 0.004 | 11 | 0.878 |
| duration | 0.077 | 2 | 0.304 | personal.status | 0.001 | 12 | 0.910 |
| credit.history | 0.074 | 3 | 0.366 | job | 0.000 | 13 | 0.937 |
| credit.amount | 0.054 | 4 | 0.434 | resident.since | 0.000 | 14 | 0.960 |
| purpose | 0.039 | 5 | 0.521 | housing | 0.013 | 15 | 0.977 |
| savings | 0.044 | 6 | 0.595 | other.debtors | 0.004 | 16 | 0.989 |
| age | 0.023 | 7 | 0.671 | num.credits | 0.001 | 17 | 0.995 |
| employment.since | 0.018 | 8 | 0.742 | telephone | 0.001 | 18 | 0.998 |
| property | 0.017 | 9 | 0.805 | numb.people.liable | 0.000 | 19 | 1.000 |
| other.installments | 0.017 | 10 | 0.843 | foreign.worker | 0.001 | 20 | 1.000 |

Source: authors' own.

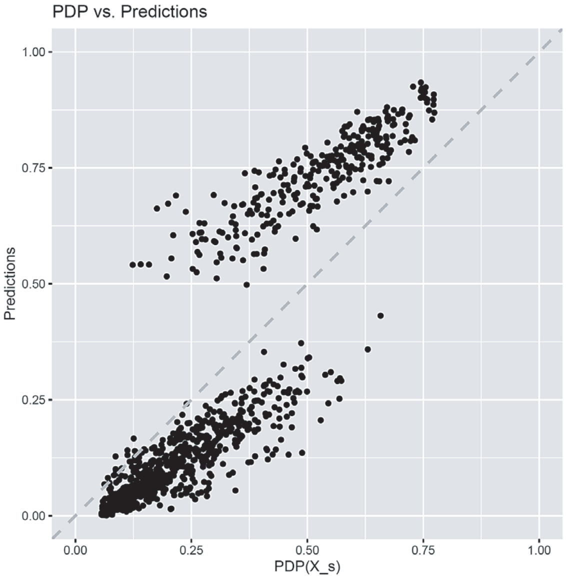

Fig. 3. Partial dependence vs. true predictions of training data for dim($X_s$) = 9 and  = 0.8.

Source: authors' own.

## 2.5. Connection to the existing methodology

Note that the proposed measure of explainability $\Upsilon$ reflects the difference between the PDP and the model's prediction, which is an important but not the only aspect of interest with regard to explainability. A well-known limitation of partial dependence curves is that they might be misleading in the case of correlated predictor variables

(Hooker and Mentch, 2019). In Friedman and Popescu (2008), an $H^2$ statistic is proposed that can be used to identify the existence of interactions between predictor variables. For correlated predictor variables, accumulated local effect plots (ALE, Apley, 2016) have been shown to be more appropriate than partial dependence plots. ALE plots are beyond the scope of this paper, but the extension of $\Upsilon$ for ALE plots may be a subject for future research.

A popular visual tool to analyse hidden variability behind a partial dependence curve are individual conditional expectation (ICE) curves (Goldstein et al., 2015), where instead of averaging over all the observations, a separate PD curve is drawn for each observation $x_i$:

$$\widehat{ICE}_s\left(x_i\right)=\hat{f}\left(x_s,x_{ic}\right). \tag{7}$$

The resulting plot of the ICE curves enables to understand the heterogeneity of the PD as a function of $x_s$ (cf. Figure 4 (left) for the variable duration). In particular, ICE plots can be used for a visual analysis of whether the individual curves show the same trend. Yet, to the best of the author‘s knowledge, this can only be analysed visually, but no objective measure has been proposed in order to quantify this. In contrast, explainability $\Upsilon$ quantifies the observed variation hidden behind a partial dependence function into one single and interpretable value that is close to 1 (for small variation) and close to 0 (for strong variation) by integrating over the distribution $P(X_s)$.

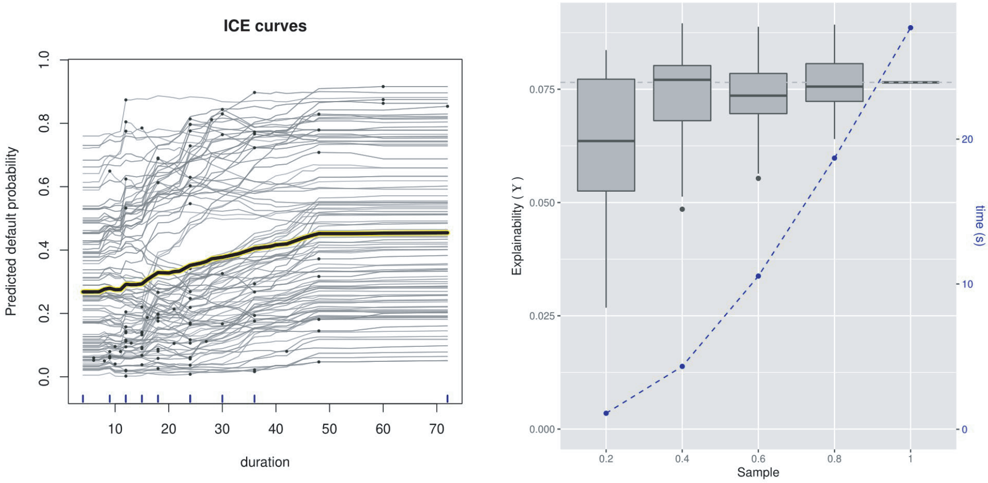

Fig. 4. ICE curves for the variable duration (left) and simulation results for the computation time based on subsamples of different size, as well as the resulting distribution of $\hat{\Upsilon}$ for the variable duration (right).

Source: authors' own.

Another issue of PDPs is their extrapolation to areas where little or no training data is available (Hooker and Mentch, 2019, Molnar et al., 2020b). Note that explainability as a global measure reflects the distribution of the training data w.r.t. the predictor variables, i.e. a large value of $\Upsilon$ does not prevent from misinterpreting extrapolations of the model outside the range of the training data.

Note that for the individual curves in an ICE plot, the values of $x_s$ are varied regardless of how likely they are to occur, conditional on $x_{ic}$, which might be misleading. $\Upsilon$, in addition, takes into account the joint distribution of variables from $X_s$ and $X_c$, as from each curve only the observed points $x_{is}$ (dots in the graph) are used.

### 2.6. Computational considerations

For the common implementations of partial dependence plots, e.g. those in Greenwell (2017), Biecek (2018) and Molnar et al.,(2018), the scope consists in visuali-sation of the PD curve and it is sufficient to restrict on computing $\widehat{PD}_s$ for a subset grid of the data. In contrast, $\widehat{\Upsilon}$ accounts for distribution $P(X_s)$, and thus requires computation of the partial dependence $\widehat{PD}_s\left(x_i\right)$ for all observations.

Computation of $\widehat{PD}_s\left(x_i\right)$ requires the Cartesian product $x_s \otimes x_c$ of the two variable subsets $x_s$ and $x_c$ of the data, therefore the calculation of for given data is $O(n^2)$ in the number of observations $n$ with regard to both computation time and memory usage. In order to circumvent this issue arising with large sample sizes, an alternative consists in its computation on a random subsample of the $x_{is}$, $i = 1, ..., n$. Note that a similar approach was proposed to reduce the computation cost for Shapley additive explanations (Strumbelj and Kononenko, 2014), where random subsets of variables are used in order to avoid enumerating all possible permutations of variable subsets. Naturally, this trades off with the variance of the estimate. Figure 4 (right) illustrates both the reduction in (average) computation time (dashed line) as well as the increasing variability of the estimates (box plots) for 50 random samples of the $x_s$ using an INTEL Xeon CPU E3-1505M v5 2.8Ghz 8 core with 32GB RAM.

## 3. Maximising explainability

### 3.1. Based variable selection

According to Table 1 (column $\widehat{\Upsilon}$), $\Upsilon$ can be used to compare different variables with regard to their ability to explain a model (using a PDP). Consequently, a forward variable selection can be carried out to maximise the explainability of a model with as few variables as possible (cf. Algorithm 1). Note that, as opposed to traditional variable selection or variable importance, the variables here are not selected with regard to the model's performance but rather with regard to the degree of explanation that they provide for an existing model.

**Algorithm 1** $\hat{\Upsilon}$ based forward variable selection in order to maximize explainability.

```
Initialize $X_s = \emptyset$ and $X_c = X$.
repeat
    for all variables $X_j \in X_c$ do
        $X_s^{candidate} = X_s \cup X_j$
        Compute $\hat{\Upsilon}(X_s^{candidate})$
    Determine $X_{j^*}$ that maximizes $\hat{\Upsilon}(X_s^{candidate})$.
    Set new $X_s = X_s \cup X_{j^*}$ and $X_c = X_c \setminus X_{j^*}$
until $X_c \neq \emptyset$
```

Source: authors' own.

### 3.2. Application to the South German credit data

Table 1 (column $\widehat{\Upsilon}_k$) provides an example of variable selection based on $\Upsilon$ to maximise explainability (the step number is indicated in column *k*): a PDP of only two variables already provides an explainability of 0.304 and for dim ($X_s$) = 5 (/9/12) an explainability $\widehat{\Upsilon}_{\dim(X_s)}$ = 0.5 (/0.8 /0.9) is obtained. Figure 5 shows a trellis visualisation (Cleveland, 1993) of a two-dimensional PDP (as implemented in e.g. Greenwell, 2017) for the two variables: status account and duration, with the highest explainability. It reveals the same trend of increasing risk with longer maturity times for all status levels of the account, but an observable interaction exists for existing accounts with a low or negative balance where the increase in risk is stronger.

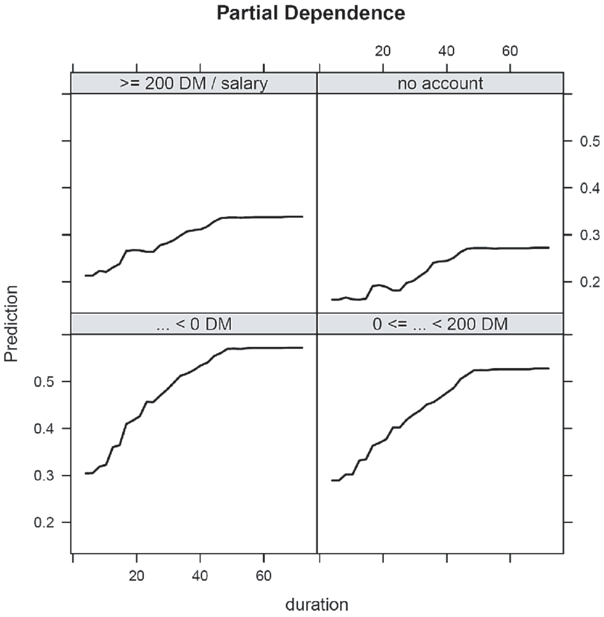

Fig. 5. 2D PDP for the variables status account and duration.

Source: authors' own.

Although in general, partial dependence functions are not restricted with regard to the dimension of $X_s$, their visualisation is limited to one or two dimensions. For more than two variables one can create scatterplot matrices (Cleveland, 1993), but this still does not allow to visualise higher order interactions between variables. This should be kept in mind when partial dependence plots are used to explain black box machine learning models. For the random forest model on the South German credit data, the most explainable two-dimensional PDP from Figure 5 only explains 30% of the variation of the model's predictions.

## Conclusion

In recent years, several failures of AI applications have occurred. As a result, regulatory requirements for business applications of machine learning and the ongoing hype around the methodology for explainable AI (*XAI*) have emerged, but a unified framework on up to what extent the explanations by *XAI* can be misleading is still missing. Hence, a methodology was presented that allows to analyse to what degree predictive black box machine learning models can be explained by partial dependence plots. The framework provides both a graphical analysis of the mismatch between the PD curve and the predictions by the model in terms of PX-plots, as well as a measure ($\Upsilon$) to quantify explainability of a model by a PDP on an interpretable scale. An algorithm was presented to maximise explainability with a low-dimensional PDP.

The proposed methodology was applied in this study to the publicly available South German credit data using a random forest model. It appears that a reasonable and well-interpretable partial dependence curve as it is observed for the variable duration, can still deviate noticeably from the predictions of the model – which has to be taken into account when explaining it. The proposed measure of explainability $\Upsilon$ can help to support business decisions by validating the model's interpretability. A PDP of the two most explainable variables, i.e. status account and duration, is more appropriate in order to understand how the model behaves.

In general, the explainability of the model becomes better when an increasing number of variables are taken into account, but for >2D PDPs can no longer be visualised and thus an analyst will not be able to understand any high-order dependencies that impact on the model's predictions. An R package with implementations of the described methodology is available on Github under https://github.com/g-rho/xPDPy.

Note that the proposed measure of explainability $\Upsilon$ only reflects the difference between the partial dependence curve and the predictions by the model under investigation, which is an important but not the only aspect of interest with regard to explainability. For example, individual conditional expectation (ICE) plots allow for a visual analysis whether the individual curves for all the observations show the same trend. Yet, currently there is no objective measure to quantify this – which could be a scope of future research.

There is also a need for ongoing research to develop methodology to understand high order interactions, e.g. based on the ideas presented in Britton (2019), and Gosiewska and Biecek (2019). Psychology provides a reasonable number of dimensions that can be simultaneously assessed by humans, there seems to be somewhere around seven (Miller, 1956), while naturally there might be differences depending on the experience of the analyst. However, it is questionable to what degree humans will ever be able to understand nonlinear high-order interactions. For this reason, the proposed measure can be considered as a tool to quantify the degree of explainability of a black box machine learning model.

Molnar et al. (2020a) suggested an approach to simultaneously optimise a trade-off between predictive accuracy and interpretability. In contrast, other authors claim to rather use interpretable models (Rudin, 2019) which may trade off with predictive power (but not always, cf. e.g. Buecker et al., 2021). To conclude, the benefits of more complex but uninterpretable models over interpretable ones should be carefully analysed during model selection.

An important challenge consists in the development of fair scoring models (Kusner and Loftus, 2020; Szepannek and Luebke, 2021) and future research on this topic will be based on causal inference (cf. e.g. Luebke et al., 2020 for some examples). According to the results of Zhao and Hastie (2019), partial dependence curves can be used for this purpose. This makes the suggested measure of explainability also an important concept on the road towards developing fair scores.

## References

Apley, D. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv:1612.08468.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen J. (2002). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, 54(6), 627-635.

Banasik, J. and Crook, J. (2007). Reject inference, augmentation and sample selection. *European Journal of Operational Research*, 183, 1582-1594.

Basel Committee on Banking Supervision. International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Bank for International Settlements, 2005.

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models. *R. Journal of Machine Learning Research*, 19(84), 1-5.

Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., Sztachelski, J. and Wojewnik, P. (2021). Enabling Machine Learning Algorithms for Credit Scoring – Explainable Artificial Intelligence (XAI) methods for clear understanding complex predictive models. arXiv:2104.06735.

Bischl, B., Kühn, T. and Szepannek, G. (2014). On Class Imbalance Correction for Classification Algorithms in Credit Scoring. In Operations Research Proceedings 2014. *Selected Papers of the Annual International Conference of the German Operations Research Society (GOR)*, 37-43, 2016.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Britton, M. (2019). VINE: Visualizing Statistical Interactions in Black Box Models. arXiv:1904.00561.

Brown, I. and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453.

Bücker, M., Szepannek, G., Gosiewska, A. and Biecek, P. (2021). Transparency, auditability and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1), 70-90.

Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, 3.

Cleveland, W. (1993). Visualizing Data. Hobart Press.

Crook, J., Edelman, D., and Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183, 1447-1465.

Dastile, X., and Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. *IEEE Access*, 9.

Demajo, L., Vella, V. and Dingli, A. (2020). Explainable AI for interpretable credit scoring. arXiv:2012.03749.

Dua, D., and Graff, C. (2017). UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/ml/index.php, 2017.

European Banking Authority. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189-1232.

Friedman, J. and Popescu, B. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3), 916-954.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44-65.

Gosiewska, A. and Biecek, P. (2019). iBreakDown: Uncertainty of model explanations for non-additive predictive models. arXiv:1903.11420.

Greenwell, B. (2017). An R Package for constructing partial dependence plots. *The R Journal*, 9(1), 421-436.

Groemping, U. (2019). South German credit data: Correcting a widely used data set. Department II, Beuth University of Applied Sciences Berlin.

Hooker, G. and Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. arXiv:1905.03151.

Hutter, F., Kotthoff, L. and Vanschoren, J. (2018). Automated Machine Learning: Methods, Systems, Challenges. Springer.

Kusner, K. and Loftus, J. (2020). The long road to fairer algorithms. *Nature*, 534, 34-36.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L. and Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework. *R. Journal of Open Source Software*.

Lessmann, S., Baesens, B., Seow, H.V., and Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136.

Liaw, A. and Wiener, M. (2002). Classification and regression by random Forest. *R News*, 2(3), 18-22.

Louzada, F., Ara, A. and Fernandes, G. (2016). Classification methods applied to credit scoring: A systematic review and overall comparison. *Surveys in OR and Management Science*, 21(2), 117-134.

Luebke, K., Gehrke, M. Horst, J. and Szepannek, G. (2020). Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, 28(2), 133-139.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2).

Molnar, C., Bischl, B and Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, 3(26).

Molnar, C., Casalicchio, G. and Bischl, B. (2020a). Quantifying model complexity via functional decomposition for better post-hoc interpretability. In Machine Learning and Knowledge Discovery in Databases (pp. 193-204). Springer International Publishing.

Molnar, C, König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C., Casalicchio, G., Grosse-Wentrup, M. and Bischl, B. (2020b). Pitfalls to avoid when interpreting machine learning models. arXiv:2007.04131.

Ribeiro, M., Singh, S. and Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206-215.

Staniak, M. and Biecek, P. (2018). Explanations of model predictions with live and break Down Packages. *The R Journal*, 10(2), 395-409.

Strumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647-665.

Szepannek, G. (2017). On the practical relevance of modern machine learning algorithms for credit scoring applications. *WIAS Report Series*, 29, 88-96.

Szepannek, G. (2022). An Overview on the landscape of R packages for open source scorecard modelling. *Risks*, 10(3), 1-33.

Szepannek G. and Luebke, K. (2021). Facing the challenges of developing fair risk scoring models. *Frontiers in Artificial Intelligence*, 4.

Szepannek G. and Aschenbruck, R. (2020). Predicting eBay prices: selecting and interpreting machine learning models – Results of the AG DANK 2018 Data Science Competition. *Archives of Data Science* A, 7(1), 1-17.

Torrent, N., Visani, G. and Enrico Bagli, E. (2020). PSD2 Explainable AI model for credit scoring. arXiv:2011.10367.

Verbraken, T., Bravo, C., Weber Richard, and Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2), 505-513.

Vincotti, V. and Hand, D. (2002). Scorecard construction with unbalanced class sizes. *Journal of the Iranian Statistical Society*, 2, 189-205.

Zhao, Q. and Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business and Economic Statistics*, 2019.

*Received: June 2022, revised: October 2022*