Aix Marseille School of Economics

Master 2: Econometrics, Big Data and statistics

Interpretability of Machine Learning Models
Homework

Authors:
Marcel PENDA
Alexandre FAUX
Coraline BEST

November 21, 2023

Table of contents

1 Introduction

Since their development as early as the 1950s, machine learning models have found countless applications. With the increase in the availability of large datasets, open sources programming languages with rich libraries and cheaper access to large amount of computing power most academic fields and industries benefit from them. They have two main advantages. They allow for very flexible and complex relationships between the covariates and are able to deal with the curse of dimensionality allowing them to use many variables. These two aspects usually make machine learning models more accurate compared to traditional statistical models in terms of forecasting and classification accuracy.

This stems from the two different philosophies of Machine Learning and Statistics, or Econometrics in our case. The former aims at producing the best performing model in terms of an error criteria while the latter aims at identifying the data generating process behind the data. To put it in simpler words, while machine learning tries to produce the model which will give us the most precise forecast, statistics or econometrics are concerned about finding the model the closest to the true one. This is due to the fact that econometrics is aiming at finding a causal link between the dependent variable and the independent variable, whereas a machine learning model cares about this link as far as it improves the *prediction* of the dependent variable.

This difference is crucial for the interpretability of our model, because econometrics investigates causality and how our variables are related. This fact makes econometrics models most of the time interpretable, because their aim is to make sense of the world and not simply predict it. Machine Learning models often lack this property because they build complex relationships between a large number of variables, effectively making them "black boxes". This project will investigate the interpretability problem of machine learning models by presenting two methods used to make sense of complex models. In a first part, we will introduce the main challenges one faces when applying these methods to machine learning models. In a second part we will present the Partial Dependence Plot (PDP) method and illustrate it with an example. In the last part we will do the same with the Individual Conditional Expectation (ICE) method.

2 Interpretability Challenges in Machine Learning and Model Agnostic Methods

Before presenting the different methods, we need to first define what interpreability is, why it is a desirable property, and how one can apply it to Machine Learning models. To define it, interpretability is the degree to which a human can understand the decision of a model or the degree to which a human can predict the outcome of a model. Both definitions imply that having an interpretable model means having a model that we can fully understand, meaning that we can infer what it will produce next and how it arrived to a certain conclusion. At this point we have to make the distinction between interpretable and explainable. For instance, even with complex black box models such as a large neural network, we can explain the process quite easily as it boils down to linear combinations of the input variables. However, compared to a linear regression model, one can not infer the outcome value of a new observation by looking at the fitted neural network, whereas we can do it for the linear regression just by looking at the fitted coefficients. This brings us to the necessity of interpretability. As we have seen in the introduction, machine learning models do not care about causality and the explanations but only focus on correctly predicting the outcome. Interpretability is a property that is required, as most research questions do not limit themselves to prediction. Ad-

ditionally, interpretability is important for the accountability of our models. Understanding why the model estimates a specific result is a safeguard against the shortcomings of machine learning models. With interpretation we can better identify hidden biases and it also boosts social acceptance as the public will understand better why the decisions have been taken. One example of this are machine learning models applied to credit scoring. By simply looking at the data, these models could discriminate people against certain characteristics (e.g., gender or origin) which are heavily correlated to socio-economic status. With no interpretation this fact could be overlooked, perpetrated racial discrimination and cause serious reputation damages to the company and its practices. Another, more trivial example is detecting miss-specification, bugs and poorness of fit of our model. Indeed, a model producing results with weird or absurd interpretation might be wrongly specified or poorly fitted and optimized.

After having understood the meaning and the importance of interpretability, we have to understand in which shapes it comes. The first difference we can observe is between intrinsically interpretable models and black boxes of which we can make only sense of by applying an additional procedure on the fitted model. We refer to the later as post hoc interpretation methods. We focus on these for our study as they are the methods applied to black box models. Among this class of methods, we can find a few more characteristics. For instance, post hoc methods can be model specific or model agnostic, meaning they can be applied to all models. Inthis case, the interpretation can be local to one or some observations or it can be a global interpretation of the whole fitted model. These main characteristics determine the modality of how the model is interpreted. For interpretation, four main solutions are proposed to summarize the information contained in a model. The first is to focus on a metric specific to the model. Based on how a model is constructed, we can collect a statistic that indicates how this process was done. For instance, with classification and regression trees and random forests we can measure the variable importance. This metric measures how many times a variable was used to make a split. The second refers to the model's weights. Here the weights refer to the resulting parameters, after fitting the model, which will attribute to each covariate an effect on the outcome variable. For instance, regression coefficients or tree splitting thresholds are model weights. The third one is based on returning data points. These methods either work by identifying certain existing observations or by producing new data points. This way, one can see how variation in one covariate affects the outcome at some level. The final solution is based on local approximations of a complex model. To retrieve interpretation, one can locally approximate a black box model by a linear model such that we can fall back on a fully interpretable linear regression set up.

Finally, in addition of how these method allow us to interpret our black box models, we have to define what understanding a model means and how this understanding can be limited. The best level of understanding is global understanding. It includes how our model was fitted and how it is able to assign a prediction to a new data point. This level is actually hard to reach even with simpler methods as ordinary least squares, since imagining such models often relies on graphical visualisation. Beyond 3 parameters it is often impossible to grasp such complex relationships. Indeed, even if a linear regression uses relatively simple concepts of linear algebra, and can be imagined as a projection of an orthogonal vector to a base of two vectors, this understanding only works for 2 explanatory variables. Beyond this, it is already too complex for representation. This is why, we are more often satisfied by a partial understanding of our models, usually of how a single covariate affects our outcome. Instead of creating a complex high dimensional space in our mind, we can focus of how a small part of the model works. This is most often sufficient as we often only look for the partial effect of a variable on the outcome. Nonetheless, this approach has a main flaw. It assumes that while studying a single covariate that all others are held constant. This might lead to very

unrealistic observations (i.e. variable value combinations), and thus provides partial effects that do not happen in real life. In our example the PDP can be associated to a modular understanding of our model. The last level of understanding refers to interpretations at the level of one or a group of observations. This scope is similar to the precedent, however it is more granular. Indeed, even when looking at a single covariate, different observations can have different partial effects on the outcome. In our example the ICE would correspond to this category.

Against this background, we can better understand some of the challenges of the interpretability. First of all, not all models and interpretation methods are equal. Depending on our goal, different tools should be used. Additionally, global and total understanding is often impossible to retrieve, and thus we must settle for partial comprehension. Finally, our interpretation method should be adapted to the public that receives the explanation. Not all methods are displayed the same way and rely on the same mathematical foundations. Some methods, even if less specific or precise, can be clearer and speak to more people simply because of a graph or an image. To conclude, the challenge of interpretation encompasses the three aspects of choosing the appropriate tool depending on the research question, selecting the adequate level of interpretation, and taking into account the audience.

3 Partial Dependence Plot (PDP)

Introduced by J. H. Friedman, a partial dependence plot (PDP) is a global model-agnostic methods to examine feature effects on the final predictive values computed by black box machine learning methods. Thus, it is an a posterior evaluation method for complex machine learning models aiming at interpreting feature effects and causality.
The idea of the PDP is quite simple and can be perceived as an experiment. For a given machine learning model (e.g. random forest, CNN etc.), we examine the effect of one feature of interest X_S by systematically changing this feature's values (e.g. iteratively increasing all values) and iteratively re-fitting the model to obtain reference predictions while keeping all other features x_C fixed. More precisely, to perform a PDP, we follow the subsequent steps. For simplification, we assume an exemplary random forest model (rfm) and a sample data with a dependent variable y and a matrix of explanatory variables $\mathbf{X} = X_1,\ X_2,\ X_3$ with $X_1 = 1,\ 2,\ 3,\ 4,\ 5,\ etc.$ of $n = 100$ observations.

1. Artificially change all values of the feature of interest X_S to one specific occurrence while keeping all remaining features X_C fixed, e.g. starting with $\min(X_S)$.

   Example: Thus, in the first iteration, we assign $X_S = 0$ for all 100 observations while keeping the features X_2 and X_3 unchanged.

2. We re-estimate the given model with the modified data and obtain reference prediction values $\hat{f}_{1i}$ with $i = 1, ..., n$ for each observation based on which we calculate a referential average prediction value $\hat{f}_1$.

   Example: $X_{1i} = 1 \,\forall\, i \ \rightarrow\ rfm \ \rightarrow\ \hat{f}_{1i} \ \rightarrow\ \hat{f}_1 = 1/n * \sum_{=1}^{n} \hat{f}_{1i}$

3. We repeat steps 1. and 2. as many times as all occurrences were assigned. Put differently, we systematically change the feature's values to the very same occurrence from the feature range $X_1 = 1, 2, 3, 4, 5, etc.$, starting with $\min(X_1)$ until all values from the feature's set were assigned and all reference predictive values were estimated. Note, this step makes the PDP a Monte Carlo Experiment: We "randomly" draw a subsample without replacement based on which we

compute the average of a statistic of interest. We obtain a distribution of these average values, here $\hat{f}_1$

Example: We plot the data points (1, $\hat{f}_1$), (2, $\hat{f}_2$), (3, $\hat{f}_3$), (4, $\hat{f}_4$), and (5, $\hat{f}_5$).

4. Plotting all these referential average prediction values (y-axis) against the feature's set of numbers (x-axis), i.e. all possible values in the sample set the feature can take on, we obtain a step function displaying the change in the prediction value for the respective change in the feature's value while keeping all other feature's fixed. In other words, we obtain the feature's average effect on the prediction value, and thus the interpretation of the direction and magnitude of the feature's effect.

   Example: Let's assume an increase in feature X_1 from $X_{1i} = 1 \,\forall\, i$ to $X_{1i} = 100 \,\forall\, i$ leads to an increase of 5% in the prediction value y, resulting in an increasing step function. Thus, we can interpret that an increase in feature X_1 leads to an increase in the prediction value, and thus may conclude that feature X_1 contributes important information to obtain accurate predictions from our random forest model.

However, the actual impact of one feature can only be assessed after plotting a PDP for each of the explanatory variables and comparing their respective contribution to estimating the prediction values. That said, it is possible to computed the so called relative feature importance using the PDP approach.

More generally, the PDP takes the following mathematical form:

$$\hat{f}_S(X_S) = E_{X_C}[\hat{f}(X_S, X_C)] = \int \hat{f}(X_S, X_C) dP(X_C)$$

We usually focus on one or two features of interest, called S. The remaining variables (C) are treated as random variables. As stated in the formula, together X_S and X_C represent the total feature space. This allows us to observe the relationship between the feature of interest and the predicted outcome. It's like asking, "What if only my income changed, and nothing else?".

The partial function $\hat{f}_S$ can be obtained by repeatedly averaging predictions from our training data. Thus we get a reliable estimate of how our chosen features impact the predictions (Monte Carlo experiment).

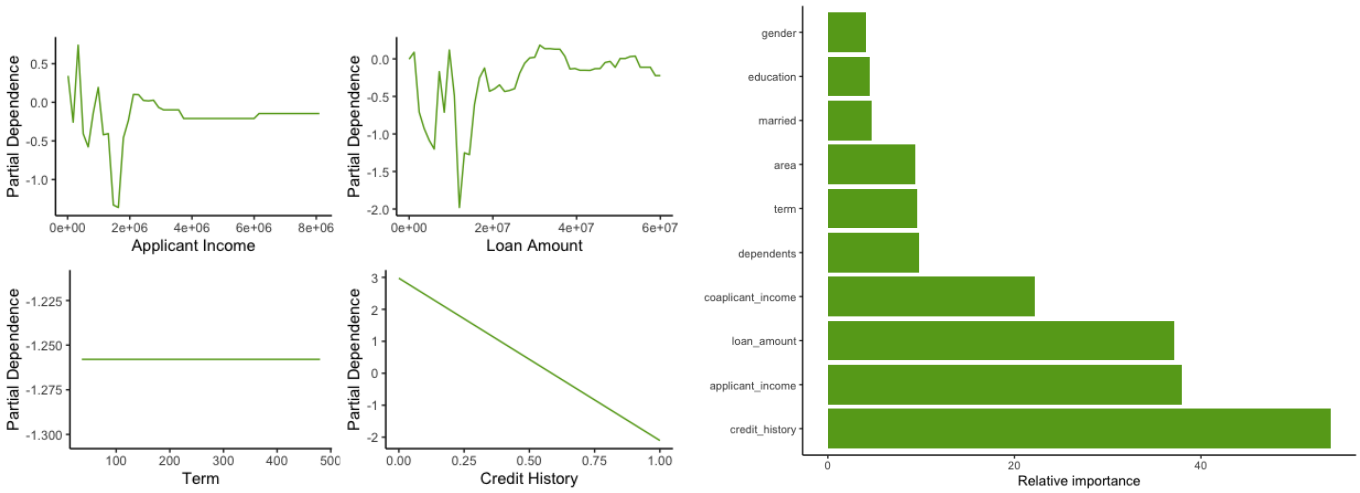
$$\hat{f}_S(X_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(X_S, X_{Ci})$$

3.1 Empirical Example

In this empirical example, our objective is to predict the likelihood of loan repayment based on diverse features like gender, credit history, education etc. To do so, we used a random forest model which is often considered as a black-box model due to its challenging interpretability. To gain insights into the average impact of features such as applicant income, loan amount, term, and credit history on the prediction, we generated Partial Dependence Plots (PDPs) (Figure a).

Notably, the PDP for the "term" feature reveals a negligible impact on predicting loan repayment. Meanwhile, when examining the PDPs for applicant income and loan amount, noteworthy spikes are observed, primarily attributable to outliers. Upon closer inspection of the data, it becomes apparent

Figure 1: Partial dependence plots



(a) Partial dependence plot

(b) Relative importance of the predictors

that only three individuals had an annual income lower than 100,000 dollars, explaining the very high variability observed. Beyond an annual income and loan amount of 200,000 dollars, there is a minimal effect on predicting loan reimbursement. Additionally, credit history exhibits a negative linear relationship with the prediction. Figure b illustrates the relative importance of the predictors. It indicates a weak partial dependence of education on the probability of loan reimbursement, which appears somewhat inconsistent. This implies that the weak main effect of education may be concealing more robust interaction effects with other variables. Given the extensive research by economists on the correlation between income and education, it is reasonable to assume a strong association between education and income levels. Consequently, income and education are likely to be correlated.

3.2 Advantages

PDP offers a **clear interpretation** when the features are uncorrelated. As a result, they precisely describe how the features affect the prediction on average. This direct interpretation provides insights into how the average prediction changes when a specific feature changes.

PDP are **easy to implement**.

PDP allow for a **causal interpretation**. By analyzing how a feature affect a given prediction, we can describe the causal relationship between the two. Consequently, the relationship can be considered causal for the model, as the outcome is modeled as a function of the features.

3.3 Disadvantages

The primary challenge associated with PDP is the strong assumption of independence. This assumption relies on the fact that the feature, or features, for which the PDP is computed, are not correlated with the other features. In our empirical example, we aim to predict loan reimbursement based on various characteristics such as gender, credit history, and education. If we take the partial dependence plot of the feature of loan amount, we need to assume that it is not correlated

with other variables, which is obviously not the case. Indeed, the amount of the loan is more likely to be correlated with the individual's income.

Realistically, partial dependence functions are designed to handle at most two features. Graphical representation and interpretation become increasingly complex as the dimensionality surpasses three.

PDPs might mask heterogeneous impacts, as they primarily illustrate the average marginal effects. Imagine a situation where half of the data points demonstrate a positive correlation between a feature and the prediction, while the other half display a negative correlation. This could lead to a PD plot resembling a horizontal line, erroneously suggesting a lack of influence from the feature on the prediction. Individual conditional expectation (ICE) can address this issue by plotting individual conditional expectation curves instead of a consolidated line, and unveil heterogeneous effects.

4 Individual Conditional Expectation (ICE)

Different to the PDP, individual conditional expectation (ICE) plots are local model-agnostic methods, i.e. they explain how a feature of interest affects the predicted value of a specific observation. More precisely, the ICE plots a curve for *every single* instance, and thus illustrates how the predicted value of each single observation changes with a change in the feature of interest (while keeping all other features fixed). Put differently, the ICE decomposes the PDP (average feature effect on the predicted value) into n conditional relationships between each observation i and its predicted value $\hat{f_{Si}}$ for a change in the feature of interest X_{Si}, and thus the PDP represents the average of the ICE curves. More formally, for each observation contained in $\{(X_{Si}, X_{Ci})\}_{i=1}^{N}$, the ICE displays the predicted values $\hat{f_{Si}}$ (y-axis) against the change in the feature of interest X_{Si} (x-axis), keeping X_{Ci} fixed.

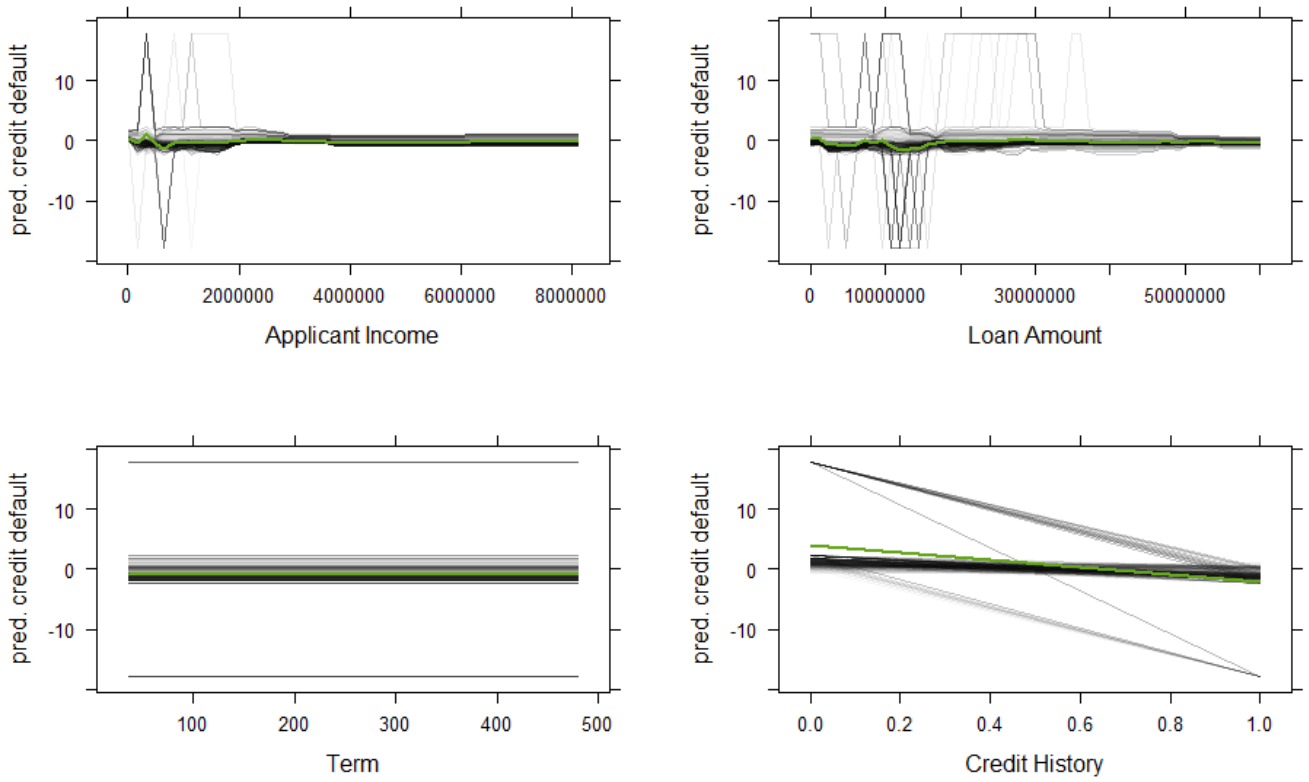
4.1 Empirical Example

Considering our empirical example, we can now examine the existence of possible heterogeneity between individuals for each feature, and thus assess whether the PDP is an appropriate tool for the interpretation of feature effects.

Figure 2 shows the ICE plots for the applicant income, loan amount, term of the loan and the credit history. In each plot, the black lines display the ICE curves while the green lines are the respective PDP, i.e. the average of the ICE curves. Note that compared to the PDPs (Figure 1), the scale of the y-axis changes. One reason for this might be a different "initial" predicted repayment probability, i.e. the observation's characteristics defined by the fixed features X_{Ci} that already "predetermine" the predicted reimbursement probability (e.g., credit history ICE plot). Second, some of the ICE curves, i.e. some single observations, show a much stronger effect on the predicted value for the same change in the feature compared to the ICE average (e.g., applicant loan and loan amount).

Indeed, the ICE plots indicate some heterogeneity within the explanatory variables. For the applicant income ICE curves (upper left-hand plot), for instance, we can observe that most of the curves follow a similar course. However, we can also identify some applicants for whom an increase in income alone has no relevant impact on the credit being repaid (rather flat ICE lines above the green

Figure 2: ICE Plots



PDP). These applicants seem to already have a relatively high chance for credit reimbursement due to other features contained in X_C "predetermining" the reimbursement outcome. Moreover, we can observe some applicants that seem to have a significant increase (decrease) in the predicted approval status (peaks below and above the green PDP line), suggesting the presence of heterogeneity between individuals. This could be explained by two reasons. One might be the poor fit of our model on individuals with low income due to missing values. The second reason might be interactions with other observed or omitted variables.

With regards to the applicant income, possible interaction features might be the coapplicant income or the applicant's credit history. For example, potential creditors applying with a coapplicant who has a rather high income might drastically decrease the chances of credit default by reaching a certain total income threshold. Put differently, if for a given (high) coapplicant income, the applicant's income increases and thus the applicant can now show the necessary (personal) financial means, the banks risk of credit default might decrease greatly due to surpassing the minimum applicant income threshold. Similarly, if an applicant has a low credit history score, indicating excellent creditworthiness, but not the required income, an increase in the income, and thus surpassing a certain threshold, might decrease the credit default risk considerably. The ICE curves indicating that an increase in the applicant's income leads to a drastic decrease in the predicted value is probably due to a poor fit of our random forest model. The low number of observations for small loans and incomes probably the causes the prediction of the model to have a very high variance, causing these spikes.

For the credit history ICE plot, we can observe that all ICE curves show the same trend. No matter the applicant (i.e. for any combination of fixed features X_C), an increase in the credit history score, indicating poor creditworthiness, has a negative effect on the predicted credit approval. However, for some applicants this effect is stronger which might be due to downwards pushing interactions (some of which are outlined above). For example, with an increase in the credit history score, indicating poor creditworthiness (e.g. due to past repayment difficulties), an overall promising credit applicant will experience a constantly declining chance for credit approval, no matter how promising the applicants other fixed features X_C. This interpretation is captured by the decreasing linear line above the green PDP line. In general, however, the aggregated credit history PDP line captures the individual ICE curves quite adequately.

The lower left-hand plot suggests that any applicant neither increases nor reduces their chance to obtain a credit by changing their repayment term (no matter the fixed remaining features X_C).

4.2 Advantages

Amplifying the PDP, the interpretability of the ICE curves is even more intuitive, since the feature's effect on the respective prediction value is displayed for each of the n observations separately.

Furthermore, compensating the disadvantage of the PDP of not capturing heterogeneous feature effects, the ICE plot is very practical. The ICE curves reveal heterogeneous relationships by decomposing the PDP into the n ICE curves, allowing to interpret some observations separately and to investigate possible problems of fit, feature interaction or omitted variables.

4.3 Disadvantages

As for the PDP, one disadvantage to keep in mind is that some points may be unrealistic or meaningless due to collinearity between the feature of interest X_S and (an)other feature(s) from X_C . In our empirical example such a case is less likely. Even though, we might assume a correlation between the loan amount and an applicant's income, any combination of these two features, i.e. low loan amount and high income and vice versa, is not unrealistic.

Also similar to the PDP, an ICE plot is best suited for visualizing the effect of one feature. Thus, we need to plot separate graphs for each feature respectively. Unfortunately, an ICE plot can quickly become difficult to interpret when plotting a large number of observations. This might be resolved by depicting only a subsample of observations or using transparent lines.

5 Conclusion and Limitations

This exercise aimed at presenting the main idea behind the PDP and ICE. After having introduced the importance of interpretability in the context of machine learning models, we provided a theoretical explanation of the PDP and ICE methods. While the ICE is a local model-agnostic method, the PDP is the aggregated average of the ICE curves, and thus a global model-agnostic method. By use of an empirical example, we applied these methods to a random forest model for credit approval prediction. Based on the PDPs, we were able to globally interpret the average feature effects, and thus the credit approvals predicted by a *block box* algorithm. The ICE plots provided a more in depth interpretation, indicating some potential feature interactions and pitfalls of our model.

Nevertheless, we have to mention some limitations to fully present the PDP and ICE methods. Since our data set has a rather small number of observations, we refrained from removing outliers which would lead to an even smaller sample. Hence, the plotted PDPs and ICE curves are stretched on the x-axis, making the identification of interactions and effect directions rather complicated. These limitations refers to some of the disadvantages listed in sections 3.3. and 4.3. If this were not the case, we could have made more precise interpretations for specific feature intervals.