TOWARDS OPTIMAL PARTIAL DEPENDENCY PLOTS

TECHNICAL REPORT

Kusha Sareen
McGill University
Armilla.ai

Rahm Hafiz
Armilla.ai

Dan Adamson
Armilla.ai

August 14, 2021

ABSTRACT

The predictions of machine learning models influence a number of socially important outcomes. As such, it is essential that these models are robust and trustworthy. There is a need for standards in the machine learning industry ensuring models are responsibly built and safe before they are deployed. Partial Dependency Plots (PDPs) are a tool used to visualize the relationship between a model's features and its prediction, giving insights into the inner workings of "black box" models. This report provides a methodology for selecting interesting PDPs for model validation. PDPs are often selected manually, a strenuous task for large models. This method can be used to provide a concise report of the model behaviour and suggest interesting PDPs for a validator based on several criteria.

Keywords Machine Learning · Explainability · Partial Dependency Plots

1 Introduction

The recent rise in popularity of machine learning (ML) algorithms in production has garnered discussion about the responsibilities and validation standards expected of an ML platform. Indeed, the importance of robustness, safety and explainability in ML is well-acknowledged [Sokol and Flach, 2019]. There is a need for model validation standards in industry ensuring robustness. Common algorithms used in the technology and financial sectors are often impactful on social outcomes and thus require a number of explainability tools to ensure their safe use.

The Partial Dependency Plot (PDP) is a key tool to visualize the global behaviour of these "black box" algorithms. These plots show the marginal effect one or two features have on the outcome of a machine learning model and detail the relationship between a given feature and the target prediction. The PDP provides a causal interpretation for the output of the model. Given a model f and a set of features $\mathbf{x}$, the PDP, $f_S(\mathbf{x})$, of a regression model is given by

$$f_S(\mathbf{x}) = \mathbb{E}_{\mathbf{x_C}}[f(\mathbf{x_S}, \mathbf{x_C})] = \int f(\mathbf{x_S}, \mathbf{x_C}) dP(\mathbf{x_C}), \tag{1}$$

where $\mathbf{x_C}$ is the set of features varying over their marginal distributions (in a 1D PDP, the cardinality of $\mathbf{x_C}$ is 1; for a 2D PDP, it is 2 etc.), $\mathbf{x_S}$ is the set of features held constant and thus the PDP $f_S(\mathbf{x})$ is the expected output of the model as $\mathbf{x_C}$ are varied.

In practice, $f_S(\mathbf{x})$ is approximated by a Monte Carlo method. The average model output in the dataset is computed as a single feature is varied while all others are held constant. We have

$$\hat{f}_S(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x_S}, \mathbf{x_C}^{(i)}), \tag{2}$$

where $\{\mathbf{x_C}^{(i)}\}$ is the set of finite feature values in the dataset sampled on the curve. We approximate the integral by evaluating the model at each data point in the dataset at each $\mathbf{x_C}^{(i)}$. This curve $f(\mathbf{x_S}, \mathbf{x_C}^{(i)})$ is accordingly called the Individual Conditional Expectation (ICE) curve and is a useful tool for visualizing the model behaviour locally.

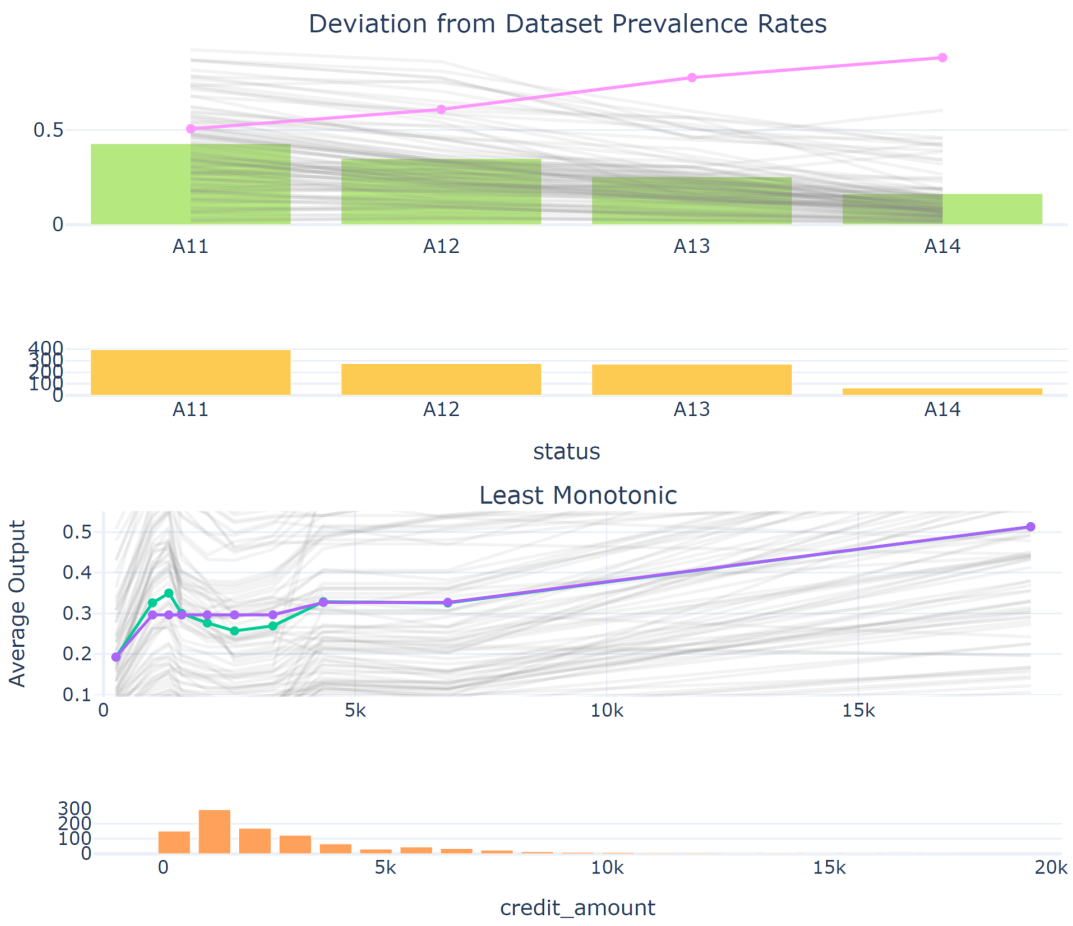

Figure 1: Partial Dependency Plots selected for a random forest model on the German Credit Risk dataset. The PDP for immigration status (green) is contrasted with the prevalence rates in the dataset (pink) using the Dataset Incidence Utility. The PDP for credit amount (green) is contrasted with its isotonic regression (purple) by the Least Monotonic Utility. ICE curves are shown in gray. Both plots are served with the underlying data distribution of the features (yellow and orange respectively)

There are some limitations in using PDPs to visualize model behaviour. The sum in Equation 2 is a poor approximation of the integral in Equation 1 when little data exists near a given $\mathbf{x_C}^{(i)}$. The average curve is also necessarily reductive and may not resemble the model behaviour at all regions in latent feature space. As such, the plots do not display potential heterogeneous effects in the model. For instance, Goldstein et al. [2014] provides such an example. As a result, plots are often presented with the underlying data distribution of the features in $\mathbf{x_C}$ and a sample of ICE curves. Additionally, Equation 1 assumes features $\mathbf{x}$ are independent and uncorrelated. If features are correlated, the average over the marginal distribution may consider data points that are unlikely to occur in the dataset.

2 The Standard Utility Function

In practice, evaluating the PDPs of large models is a strenuous task and can be prohibitive. We present a selection criteria for useful PDPs in model validation based on Inouye et al. [2019].

We can define the standard utility U of a given PDP curve by

$$U(\hat{f_S}, g, w) = \int L(f_S\hat{(}\mathbf{x}), g(\mathbf{x})) \cdot w(\mathbf{x}) d\mathbf{x}, \tag{3}$$

where L is a particular loss function (L1 or L2, for instance), $\hat{f_S}(\mathbf{x})$ is the PDP curve, $g(\mathbf{x})$ is a comparison function depending on the type of utility, and $w(\mathbf{x})$ is a weight function depending on the local data distribution. Here $g(\mathbf{x})$ sets a prior for what we expect $\hat{f_S}(\mathbf{x})$ to look like and U measures the deviation from that prior. Numerically, we approximate

$$U(\hat{f_S}, g, w) = \frac{1}{k}\sum_{i=1}^{k} L(f_S(\hat{\mathbf{x}_\mathbf{C}}^{(i)}), g(\mathbf{x_C}^{(i)})) \cdot w(\mathbf{x_C}^{(i)}), \tag{4}$$

The weight function $w(\mathbf{x})$ for a given set of points $\{\mathbf{x_C}^{(1)}, ..., \mathbf{x_C}^{(k)}\}$ is given by.

$$w(\mathbf{x_C}^{(i)}) = \frac{|x \in D \text{ such that } (x - \mathbf{x_C}^{(i)}) = \min(x - \mathbf{x_C}^{(j)}), \forall j \in (0, k)|}{|D|}, \tag{5}$$

where D is the dataset. Since we expect $\hat{f_S}$ to be a poor approximation of the model output when little data exists nearby, we weigh the utility of a particular $\mathbf{x_C}^{(j)}$ by the fraction of data nearby. This has empirically shown to greatly improve the output and robustness of the method.

2.1 Monotonicity

It is commonly expected the relationship between certain features and the target is monotonic. For instance, if we built a model predicting an individuals height using weight as a predictor, we should expect a monotonically increasing PDP $\hat{f_S}(\mathbf{x})$. In this case, a non-monotonic PDP should elicit investigation from the model validator. As such, we define $g(\mathbf{x})$ to be the isotonic regression of $\hat{f_S}(\mathbf{x})$, attaining

$$\min sum_{i=1}^{k}(\hat{f_S}(\mathbf{x_C}^{(i)}) - g(\mathbf{x_C}^{(i)}))^2 \text{ subject to } g(\mathbf{x_C}^{(i)}) \leq g(\mathbf{x_C}^{(j)}) \text{ when } \mathbf{x_C}^{(i)} \leq \mathbf{x_C}^{(j)}. \tag{6}$$

2.2 Lipschitz-Boundedness

Similarly, it is often expected a PDP not vary greatly along a small interval in its domain. As such, it should be easy to bound its slope. It is well documented that roughness of the prediction manifold is a indicator of underfitting whereas the prediction manifold for an overfit model should vary greatly to accoadate each datapoint. Here $g(\mathbf{x})$ is a Lipschitz-bounded regression of the PDP such that we have

$$\min \sum_{i=1}^{k}(\hat{f_S}(\mathbf{x_C}^{(i)}) - g(\mathbf{x_C}^{(i)}))^2 \text{ subject to } |\frac{g(\mathbf{x_C}^{(i+1)}) - g(\mathbf{x_C}^{(i)})}{\mathbf{x_C}^{(i+1)} - \mathbf{x_C}^{(i)}}| < l, \forall i \in (0, k-1), \tag{7}$$

where l is an arbitrary Lipschitz constant.

2.3 Similarity to Dataset Incidence

It is expected the PDP be similar to prevalence rates in the dataset. For example, a model trained on data that shows individuals with greater bank balance are less likely to default on their credit payments should, on average, be more likely to predict an individual with a large balance will not default on their credit payment. Though this utility does not take into account data bias, it can reveal model bias. For a classifier and numeric $\mathbf{x_C}$, we take

$$g(\mathbf{x_C}^{(i)}) = \frac{|x \in D^* \text{ such that } (x - \mathbf{x_C}^{(i)}) = \min(x - \mathbf{x_C}^{(j)}), \forall j \in (0, k)|}{|x \in D \text{ such that } (x - \mathbf{x_C}^{(i)}) = \min(x - \mathbf{x_C}^{(j)}), \forall j \in (0, k)|}, \tag{8}$$

where D is the dataset and $D^* \subseteq D$ is the subset of data with an affirmative target. Similarly, for a regressor given a categorical $\mathbf{x_C}$ we have

$$g(\mathbf{x_C}^{(i)}) = \frac{1}{|D_{\mathbf{x_C}^{(i)}}|}\sum\nolimits_{x \in D_{\mathbf{x_C}^{(i)}}} \hat{f_S}(x), \tag{9}$$

where $D_{\mathbf{x_C}^{(i)}} \subseteq D$ is the subset of data with the category $\mathbf{x_C}^{(i)}$.

2.4 Similarity Among Models

Finally, when a validator is comparing multiple models, it can be useful to know how the models differ. Differences in models trained on the same dataset can reveal model bias and exhibit which relationships have been learned by a particular model in reference to its peers. Additionally, in ensembles of specialists, it can be useful to see how models vary in specialist categories. We simply take

$$g(\mathbf{x}) = \hat{f}_{S,j}(\mathbf{x}), \tag{10}$$

where $\hat{f_{S,j}}$ is the PDP of the jth comparison model.

3 Other Utility Measures

Additional characteristics of interest can also be used to select PDPs.

3.1 Slope and Derivative-Based Measures

Similar to Subsection 2.2, it can be useful to find the PDP with the greatest slope over it's normalized range. Here,

$$U(\hat{f_S}, g, w) = \sup \frac{\partial \hat{f_S}}{\partial \mathbf{x_C}}. \tag{11}$$

3.2 Variance

Finally, the variance of the ICE curves of a PDP is an interesting metric. When averaging over the marginal distribution, if the model output does not greatly vary, this indicates the predictors $\mathbf{x_C}$ are significant contributors to the model's prediction. As such, we expect a feature with small ICE curve variance to be an important predictor. We can select for high or low variance with

$$U(\hat{f_S}, g, w) = \pm \frac{1}{k} \sum_{i=1}^{k} \text{Var}\left(f(\mathbf{x_S}, \mathbf{x_C}^{(i)})\right). \tag{12}$$

4 Selecting 2D PDPs

A two dimensional PDP is identical mathematically to the 1D plot execpt that $|\mathbf{x_C}| = 2$ and it is often served on a meshgrid. However, for the interpreter, higher dimensional plots often are less concerned with the feature-target relationship but rather sensitive feature 'regimes' which have a particular output and interactions between features.

We have included 3 criteria for selecting 2D PDPs accordingly.

4.1 Feature Correlation

Selecting 2D PDPs according to the greatest feature correlation can point to areas where the PDP breaks down - regimes where little data exists that may be prone to adverserial examples - in addition to areas where there is significant feature interaction leading to potentially non-linear effects in the 2D PDP.

4.2 Feature Importance

Additionally, selecting the two most important features allows for visualization of the general behaviour of the model. This is the best possible low resolution picture of how the model makes a prediction.

4.3 Individual Utility Intersection

Finally, there is flexibility to select 2D PDPs using the utility functions defined in Sections 2 and 3. These may prove useful for selection of features regimes that match a certain characteristic. For instance, using the Lipschitz-Boundedness utility in Section 2.2 can select regimes of the model that are difficult to bound, regions that are potentially underfit or overfit.

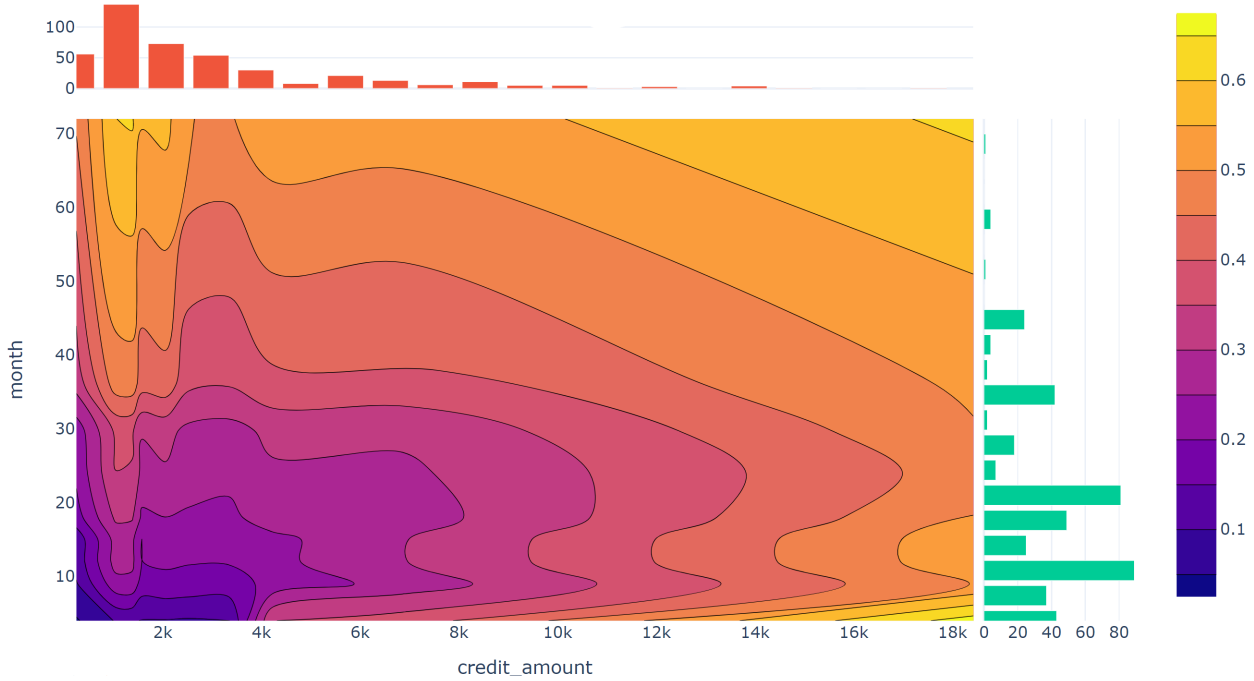

Figure 2: A selected 2D PDP for a random forest model on the German Credit Risk dataset. This plots the interaction between credit amount and months since data collection initiated. This PDP was selected based on all three criteria: feature importance, feature correlation, and the Least Lipschitz Utility.

5 Explainable Models, Importance and PDP selection

Additionally, this methodology for selecting PDPs can also be adapted to match different feature importance measures to better understand the typical behaviour of the model. A type of feature importance that is highly relevant is that which can be calculated by transfer learning a large complex model into an explainable analog, denoted model fingerprinting. This can point to the explainability of each feature.

6 Experiment: German Credit Risk Dataset

Here, we show a sample experiment of the use of automated PDPs in the German Credit Risk Dataset. Figure 1 shows the plots selected by the Least Monotonic and Dataset Incidence Utilities. As we see in the immigration status plot (top), the model has learned that individuals with immigration status A11 and A12 are more likely to receive credit than those with A13 and A14 whereas the dataset prevalence rates seem to show the opposite relationship. This raises questions as to why these associations were learned. Is the model underfit? Are there correlations in the data that result in this trend? The model validator is forced to address questions like these before the model can be used responsibly.

Additionally, the credit amount plot (bottom) shows the model has learned an nonmonotonic relationship between the amount of credit an individual has and how likely they are to receive new credit. This poses a challenge to validators to explain why such a nonlinear relationship should exist and whether it is founded in the data. This plot is also selected by the Least Lipschitz-Bounded Utility. Here, we see a large deviation in model output from 1k to 4k credit and we must consider whether such a deviation is reasonable.

In Figure 2, the relationship between model output and a mix of credit amount and month is shown in a 2D PDP. This PDP is selected on the basis of feature importance, feature correlation, and the Least Lipschitz Utility. Again, this plot may be concerning since regularized models often have smooth prediction manifolds. This raises a question of whether the model is potentially underfit or overfit.

ICE curves along with feature distribution help confirm the conclusions we draw from these PDPs are well-founded in the rest of latent feature space. Clearly, this approach selects PDPs of interest for validators.

7 Conclusion

AutoPDP provides a methodology for selecting interesting Partial Dependency Plots for model validation. This task, though usually prohibitive for large models, gives insights into the inner workings of black-box models. It can be used to provide a concise report of the model behaviour and suggest interesting PDPs for a validator based on several criteria.

References

Kacper Sokol and Peter A. Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. *CoRR*, abs/1912.05100, 2019. URL `http://arxiv.org/abs/1912.05100`.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, 2014.

David I. Inouye, Liu Leqi, Joon Sik Kim, Bryon Aragam, and Pradeep Ravikumar. Diagnostic curves for black box models. *CoRR*, abs/1912.01108, 2019. URL `http://arxiv.org/abs/1912.01108`.