FAQ
CTU
ARY

FAQCTUARY 2020-4

FEATURES WITH FLAT PARTIAL DEPENDENCE PLOTS UP TO A CERTAIN LEVEL: NOT IMPORTANT?

By Elke Gagelmans, Michel Denuit and Julien Trufin

December 2020

Detralytics

DISCLAIMER

The content of the FAQctuary's is for a pedagogical use only. Each business case is so specific that a careful analysis of the situation is needed before implementing a possible solution. Therefore, Detralytics does not accept any liability for any commercial use of the present document. Of course, the entire team remain available if the techniques presented in this FAQctuary required your attention.

Detralytics
Rue Belliard/Belliardstraat 2
1040 Brussels
www.detralytics.com
info@detralytics.eu

ABSTRACT

This paper focuses on partial dependence plots which are often used when modeling with machine learning techniques in order to better understand the effects of the features on the conditional expectation of the response variable. However, these plots must be interpreted with caution. Indeed, they can easily lead to wrong interpretations in case the analyst is not enough familiar with these plots. As noticed in a previous FAQctuary, a typical situation is the case where a feature is important because of its interactions with others while its partial dependence plot is flat. In this FAQctuary, we go one step further and we consider a very simple example with a three-way interaction effect and we show that only looking at partial dependence plots for each feature and for two features may indeed lead the analyst to wrong conclusions.

Contents

1 Introduction

Partial dependence plots are often used to determine the effect of features on the response variable when modeling with machine learning techniques. However, these plots must be interpreted with caution. Indeed, they can easily lead to wrong interpretations in case the analyst is not enough familiar with these plots. As noticed in a previous FAQctuary, a typical situation is the case where a feature is important because of its interactions with others while its partial dependence plot is flat. In such a case, an analyst who would only base his analysis on this plot could be tempted to conclude that the feature is not important to explain the conditional expectation of the response while he would be wrong. In this FAQctuary, we consider a very simple example with a three-way interaction effect and we show that only looking at partial dependence plots for each feature and for two features may indeed lead the analyst to wrong conclusions.

2 Data simulation

We consider an example in car insurance. Four features $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ are supposed to be available, that are

- X_1 = Gender: policyholder's gender (female or male);
- X_2 = Age: policyholder's age (integer values from 18 to 65);
- X_3 = Split: whether the policyholder splits its annual premium or not (yes or no);
- X_4 = Sport: whether the policyholder's car is a sports car or not (yes or no).

The variables X_1, X_2, X_3 and X_4 are assumed to be independent and distributed as follows:

$$\begin{aligned}
&\mathrm{P}\left[X_1 = female\right] = \mathrm{P}\left[X_1 = male\right] = 0.5;\\
&\mathrm{P}\left[X_2 = 18\right] = \mathrm{P}\left[X_2 = 19\right] = \ldots = \mathrm{P}\left[X_2 = 65\right] = 1/48;\\
&\mathrm{P}\left[X_3 = yes\right] = \mathrm{P}\left[X_3 = no\right] = 0.5;\\
&\mathrm{P}\left[X_4 = yes\right] = \mathrm{P}\left[X_4 = no\right] = 0.5.
\end{aligned}$$

The values taken by a feature are thus equiprobable.
The response variable Y is supposed to be the annual number of claims. Given $\boldsymbol{X} = \boldsymbol{x}$, Y is assumed to be Poisson distributed with expected claim frequency given by

$$\begin{aligned}
\lambda(\boldsymbol{x}) \;=\; & 0.1 \times \left(1 + 0.1 I_{\{x_1 = male\}}\right)\\
& \times \left(1 + \frac{1}{\sqrt{x_2 - 17}}\right)\\
& \times \left(1 + I_{\{x_4 = yes\}} \times \left(\left(0.5 I_{\{18 \leq x_2 < 35\}} - 0.5 I_{\{45 \leq x_2 < 65\}}\right) \times \left(0.7 I_{\{x_3 = yes\}} - 0.7 I_{\{x_3 = no\}}\right)\right)\right)
\end{aligned}$$

$$\times\left(1+I_{\{x_4=no\}}\times\left(\left(0.3I_{\{25\leq x_2<50\}}-0.3I_{\{x_2\leq 25\ \vee\ x_2>50\}}\right)\times\left(0.7I_{\{x_3=yes\}}-0.7I_{\{x_3=no\}}\right)\right)\right),$$

where I_A is equal to one if the random event A is realized and zero otherwise.
In this example, being a male increases the expected annual claim frequency by 10% and the expected annual claim frequency decreases with the age of the policyholder. The feature Age interacts with features Sport and Split. A policyholder with a sports car and who is between 18 and 35 years old, sees its premium increase (resp. decrease) by $50\%\times 70\% = 35\%$ if the policyholder splits (resp. does not split) its premium. A policyholder with a sports car and who is between 45 and 65 years old, sees its premium decrease (resp. increase) by 35% if the policyholder splits (resp. does not split) its premium. When considering policyholders with no sports cars, we notice that the effect of Age and Split on the expected annual claim frequency is different. For people aged from 25 to 50 years old, the premium increases (resp. decreases) by $30\% \times 70\% = 21\%$ when the policyholder splits (resp. does not split) its premium. For policyholders with another age, the premium decreases (resp. increases) by 21% when the policyholder splits (resp. does not split) its premium.
In this example, the true model $\lambda(\boldsymbol{x})$ is known and we can simulate realizations of the random vector $(Y, \boldsymbol{X})$. Specifically, we generate $n = 500\,000$ independent realizations of $(Y, \boldsymbol{X})$, that is, we consider a learning set made of 500 000 observations $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_{500\,000}, \boldsymbol{x}_{500\,000})$. An observation represents a policy that has been observed during a whole year.
In Table 2.1, we provide the ten first observations of the learning set. While the nine first policies made no claim over the past year, the tenth policyholder, who is a 49 years old man without a sports car and splitting his premium, experienced one claim.

	Y	X_1 (Gender)	X_2 (Age)	X_3 (Split)	X_4 (Sport)
1	0	male	27	no	yes
2	0	female	23	no	no
3	0	male	23	no	yes
4	0	female	49	yes	no
5	0	male	43	no	no
6	0	female	65	yes	yes
7	0	female	21	no	yes
8	0	female	55	no	yes
9	0	female	32	no	yes
10	1	male	49	yes	no

Table 2.1: Ten first observations of the simulated dataset.

In this simulated dataset, the proportion of males is approximately 50%, so are the proportions of sports cars and policyholders splitting their premiums. For each age 18,19,...,65, there are between 10 199 and 10 614 policyholders.

3 Random forests

Based on the simulated dataset composed of 500 000 observations described above, we want to model the expected annual claim frequency using all the features available. In that goal, we fit a random forest.
A random forest depends on several parameters that need to be fine-tuned. Among these parameters, we can quote

- The number of trees T composing the random forest;
- The size of the trees s, here controlled by the minimum number of observations required in each terminal node;
- The number of features m that are selected at random as candidates for splitting at each node.

In order to fine-tune the random forest, we split the dataset into a training set (80% of the observations) and a validation set (20% of the observations). The training set is used to build the random forest while the validation set aims to fine-tune its parameters. After having conducted an extensive analysis, we found that the following parameters $T = 50$, $s = 5000$ and $m = 3$ were relevant in this context. For instance, Figure 3.1 displays the out-of-bag error with respect to the number of trees. One can see that the out-of-bag error stabilizes from approximately 50 trees.

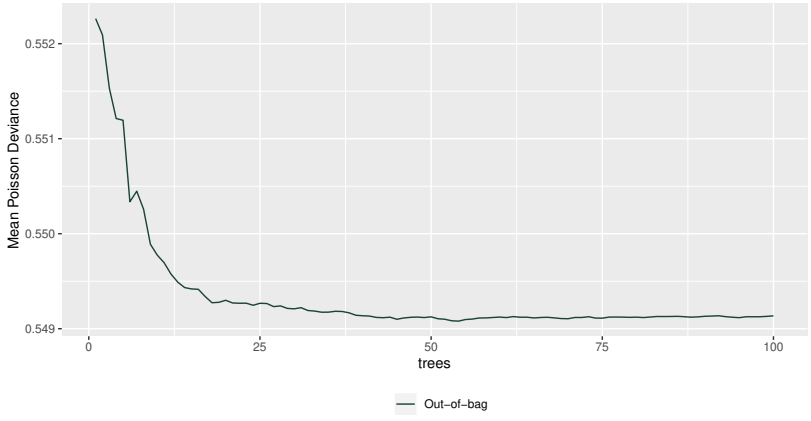

Figure 3.1: Out-of-bag error with respect to the number of trees.

Figure 3.2 depicts the partial dependence plots for each of the four features considered in this example. These plots aim to capture the marginal effects of the features on the predicted outcome [**Friedman**]. Figure 3.2 shows that there is a noticeable difference between males and females. The expected annual claim frequency decreases with the age of the policyholder. There seems to be almost no difference between policyholders who split their premium and the ones who do not, and between policyholders who have sports cars and the ones who do not. Therefore, features X_3

and X_4 may seem not to be important to predict the number of claims. However, we know that these two features are actually important to explain the expected annual claim frequency.

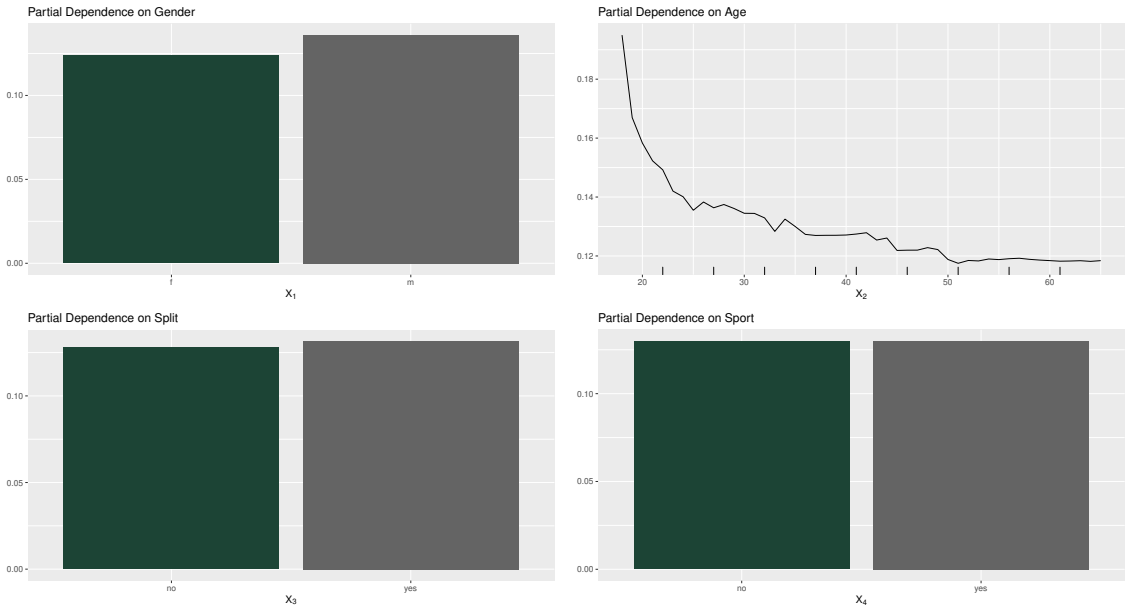

Figure 3.2: Partial dependence plots.

Henceforth, we make a focus on feature X_4. Regarding the partial dependence plots, there seems to be no difference in the expected annual claim frequency between drivers who own a sports car and the ones who do not. Nevertheless X_4 could interact with another feature, so that its impact on the expected annual claim frequency could be hided. We need to look at conditioned partial dependence plots to reveal interaction effects. Figure 3.3 shows the partial dependence plot of X_4, conditioned on X_1. For both male and female drivers, there is no difference between policyholders who own a sports car and the ones who do not. Hence there is no interaction between X_1 and X_4. Figure 3.4 shows the partial dependence plot for X_2 conditioned on feature X_4. The effect of the age of the policyholder is similar for policyholders who drive with a sports car and who do not. This indicates that there is no interaction between X_2 and X_4. Figure 3.5 displays the partial dependence plot for X_4 conditioned on X_3. For both policyholders who split their premium and who do not, the expected annual claim frequency is not impacted by the variable Sport. Hence there is no interaction between X_3 and X_4. From these graphs, we conclude that X_4 is not interacting with any other feature.

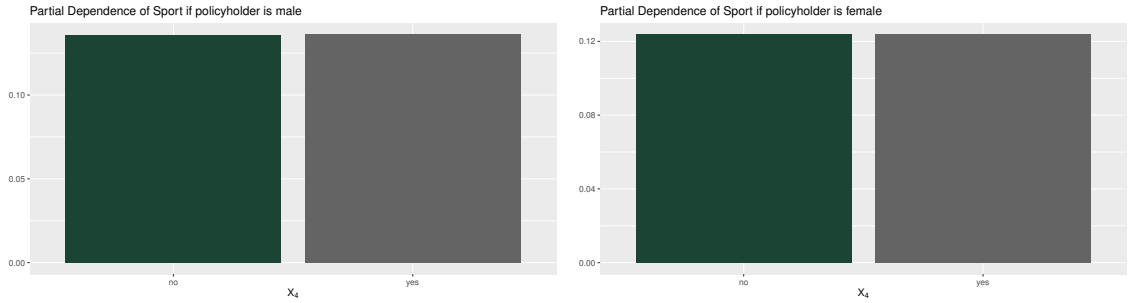

Figure 3.3: Partial dependence plot for X_4 (Sport): $X_1 = male$ (left) and $X_1 = female$ (right).

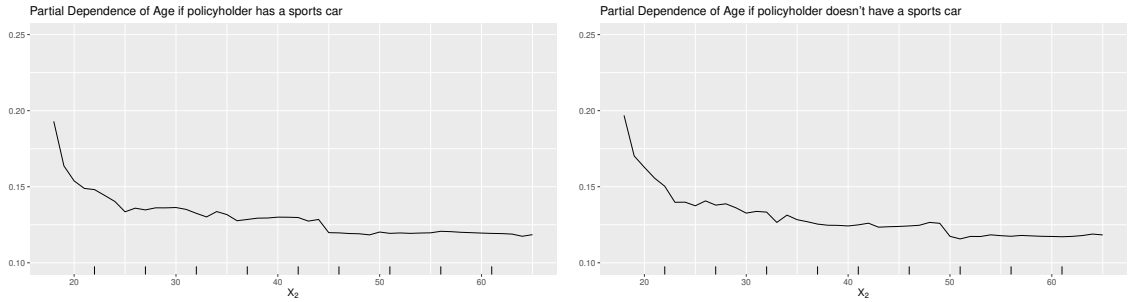

Figure 3.4: Partial dependence plot for X_2 (Age): X_4 (Sport) $= yes$ (left) and X_4 (Sport) $= no$ (right).

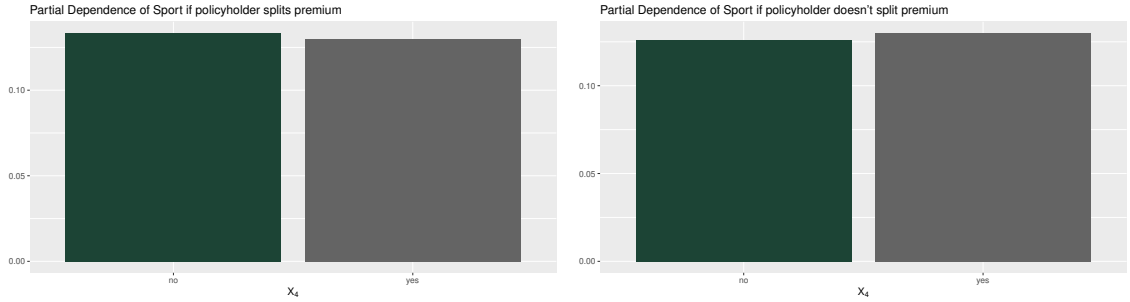

Figure 3.5: Partial dependence plot for X_4 (Sport): X_3 (Split) $= yes$ (left) and X_3 (Split) $= no$ (right).

From the discussed graphs, we would conclude that X_4 is not an important feature to predict the expected annual claim frequency. But there are also other tools available to consider the effect and importance of the different features. Figure 3.6 depicts the variable importances of the four features. The measure is computed from permuting out-of-bag observations [**rfCount**]. Age (X_2) is the most important feature according to this graph, the effect of the age was also clearly visible on the partial dependence plot. The second one is Split (X_3). Adding X_3 leads to a significant improvement. This effect is not so clear on the partial dependence plot, there it seems that the expected annual claim frequency is almost the same for persons who split and who do not split their premium. The third one is Sport (X_4), adding this feature also leads to an improvement

which was not noticeable on Figure 3.2. There was also no interaction effect visible between X_4 and any other feature on the conditioned partial dependence plots. The least important one is Gender (X_1). The effect of this feature was visible on the partial dependence plot, since it was only added as a marginal effect. Hence X_1 has an impact on the expected annual claim frequency. Since it is the least important feature according to Figure 3.6, all features are important here to estimate the expected annual claim frequency.

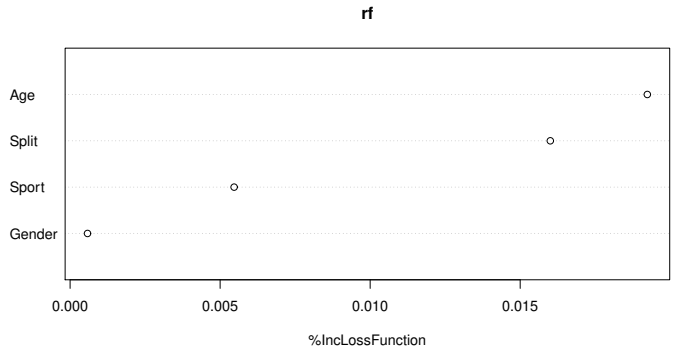

Figure 3.6: Variable importance.

We have seen that X_4 is an important feature. However, we did not highlight marginal effects nor two-way interaction effects. Let us then check whether X_4 interacts with two other features at the same time. Again, conditioned partial dependence plots are used. Figure 3.7 shows the partial dependence plot for X_2 conditioned on X_3 and X_4. All the subfigures depict different trends. Hence, the age of the policyholder has a different impact on the expected annual claim frequency for the 4 subclasses, meaning that there is a three-way interaction between X_2, X_3 and X_4. Therefore, we can now understand the reason why X_4 is actually an important feature to predict the expected annual claim frequency.

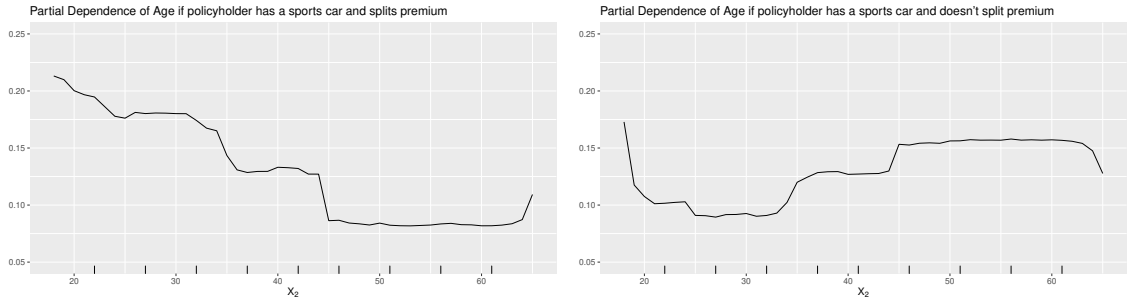

Figure 3.7: Partial dependence plot for X_2 (Age), accordingly to if the policyholder splits premium (right) or not (left) and has a sports car (top) or not (bottom).

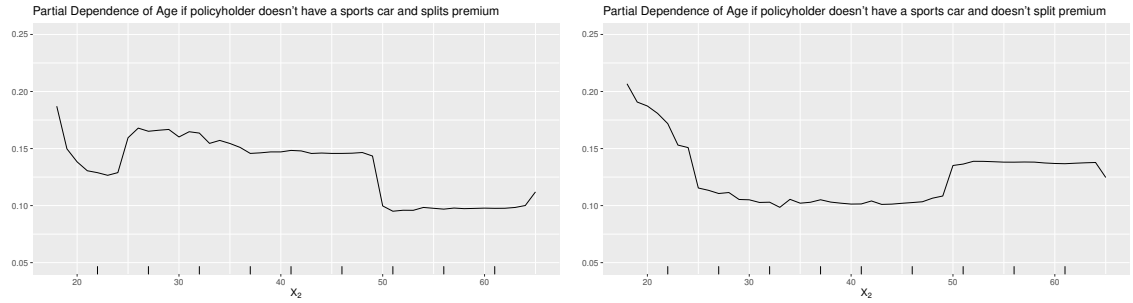

Figure 3.7: Partial dependence plot for X_2 (Age), accordingly to if the policyholder splits premium (right) or not (left) and has a sports car (top) or not (bottom).

4 Conclusion

We need to be careful with the interpretation of partial dependence plots. Conditioned partial dependence plots should also be used to reveal interaction effects. However, because variables do not always interact with just one other variable, checking all these potential interactions is not an easy task to achieve in practice. That is why these plots must be used in conjunction with other indicators such as the variable importances. To conclude, a feature X_j with flat partial dependence plots up to a certain level may still be important!

5 About the serie and the authors...

5.1 The FAQctuary's

The FAQctuary's are a series of educational papers dedicated to the insurance sector. Each issue addresses a specific actuarial topic, expressed as a question asked by market players. FAQctuary's are published by members of the Detralytics team and written in a clear and accessible language. The team combines academic expertise and business knowledge. Detralytics was founded to support companies in the advancement of actuarial science and the solving of the profession's future challenges. It is within the scope of this mission that we make our work available through our Detra Notes and FAQctuary's series.

5.2 Authors' biographies

Elke Gagelmans

Elke is part of the Talent Accelerator Program (TAP) at Detralytics. Prior to joining Detralytics, Elke did an internship at EY, where she worked on a project about micro-reserving with machine learning techniques and on a project about reporting in powerBI. She holds a Master's degree in Actuarial and Financial Engineering and a Master's degree in Mathematics, both from KU Leuven. Her thesis focused on micro-reserving and the use of GLMs to model the reserve.

Michel Denuit

Michel is Scientific Director at Detralytics, as well as a Professor in Actuarial Science at the Université Catholique de Louvain. Michel has established an international career for some two decades and has promoted many technical projects in collaboration with different actuarial market participants. He has written and co-written various books and publications. A full list of his publications is available at : `https://uclouvain.be/en/directories/michel.denuit` .

Julien Trufin

Julien is Scientific Director at Detralytics, as well as a Professor in Actuarial Science at the department of mathematics of the Université Libre de Bruxelles. Julien is a qualified actuary ofthe Instute of Actuaries in Belgium (IA|BE) and has experience as a consultant, as well as a compelling academic background developed in prominent universities such as Université Laval (Canada), UCL and ULB (Belgium).He has written and co-written various books and publications. A full list of his publications is available at : `http://homepages.ulb.ac.be/~jtrufin/`.

Expertise and innovation at the service of your future

www.detralytics.com - info@detralytics.eu