FAQ
CTU
ARY

FAQCTUARY 2020-2

FEATURES WITH FLAT PARTIAL DEPENDENCE PLOTS: NOT IMPORTANT?

By Elke Gagelmans, Michel Denuit and Julien Trufin

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

Contents

1 Introduction

Partial dependence plots 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. 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 aim to illustrate such a situation with the help of a simulated example that is very simple.

2 Dataset

We consider a simulated dataset which acts as a portfolio in car insurance. Specifically, 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, denoted Y, is supposed to be the annual number of claims. Given $\boldsymbol{X} = \boldsymbol{x}$, Y is assumed to be Poisson distributed with expected annual claim frequency given by

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= 0.1 \times \left(1 + 0.1 I_{\{x_1 = male\}}\right)\\
&\quad \times \left(1 + \frac{1}{\sqrt{x_2 - 17}}\right)\\
&\quad \times \left(1 + 0.3 I_{\{18 \leq x_2 < 35\}} - 0.3 I_{\{45 \leq x_2 < 65\}}\right) \cdot I_{\{x_4 = yes\}},
\end{aligned}$$

where I_A is equal to one if the random event A is realized and zero otherwise.

In this example, we notice that being a male increases the expected annual claim frequency by 10%, the expected annual claim frequency decreases with the age of the policyholder, splitting its premium does not influence the expected annual claim frequency while driving a sports car increases the expected annual claim frequency with 30% for policyholders between 18 and 35 years old and decreases the expected annual claim frequency with 30% for policyholders who are between 45 and 65 years old.

The true expected annual claim frequency $\lambda(\boldsymbol{x})$ is thus known, so that we can generate observations of the random vector $(Y, \boldsymbol{X})$. Specifically, we generate $n = 500\,000$ independent realizations of $(Y, \boldsymbol{X})$, that is, we consider a dataset 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 policyholder that has been observed during a whole year.

In Table 2.1, we provide the ten first observations of the simulated dataset. 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 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 $T = 25$, $s = 5000$ and $m = 3$ were relevant in this context.

Figure 3.1 shows the corresponding partial dependence plots. These plots exhibit the marginal effects of the features on the predicted outcome. Based on these plots, the analyst could be tempted to conclude that both features X_3 (Split) and X_4 (Sport) are not important, while we actually know that X_4 (Sport) influences the expected annual claim frequency.

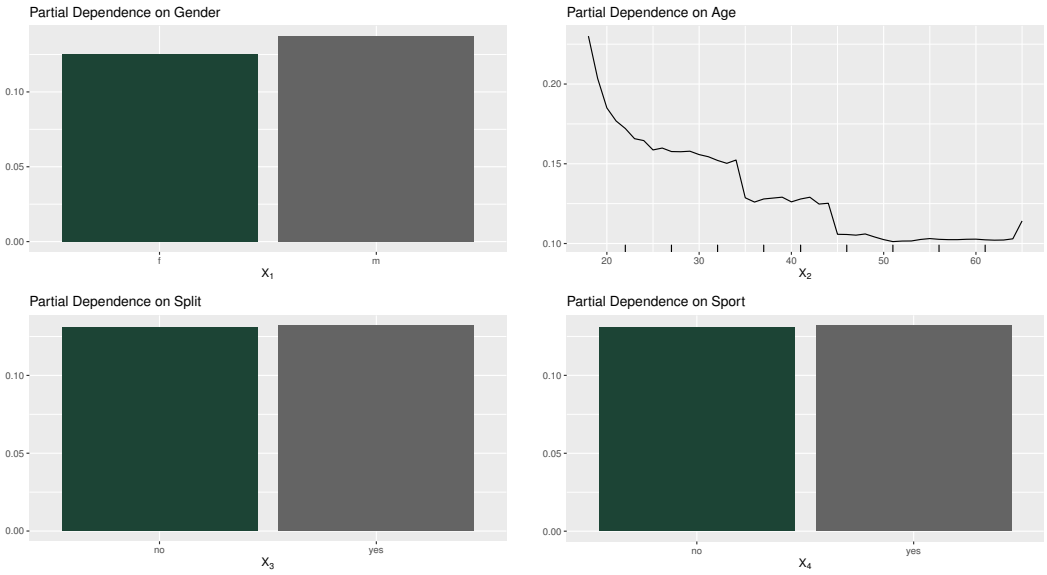

Figure 3.1: Partial dependence plots.

That is why these plots must be supplemented by other indicators such as the variable importances. In Figure 3.2, we depict the importances of the four features considered in this note. Notice that these importances are computed by permutation.

We observe that X_2 (Age) is the most important feature followed by X_4 (Sport). Adding X_4 (Sport) in the model thus leads to a significant improvement in terms of predictive accuracy, while it was not visible based on its partial dependence plot. The reason is that X_4 (Sport) is important for the expected annual claim frequency only because of its interaction with X_2 (Age), while its average marginal effect is negligible. To illustrate this observation, we show in Figure 3.3 two partial dependence plots for X_2 (Age): one for $X_4 = yes$ and the other for $X_4 = no$.

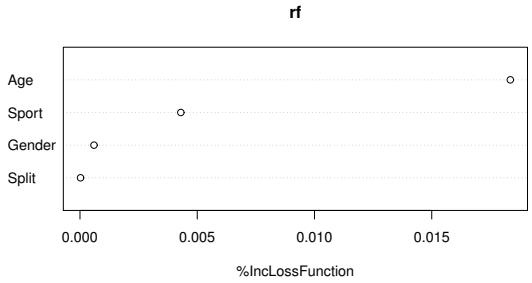

Figure 3.2: Variable importances.

We see that the effect of X_2 (Age) on the estimated expected annual claim frequency depends on the value of X_4 (Sport), indicating that there is an interaction between X_2 (Age) and X_4 (Sport), as expected. With respect to the variable X_1 (Gender), it appears to be in third position in terms of importance. X_1 (Gender) only influences the expected annual claim frequency by a marginal effect (it does not interact with other features), so that its effect was already visible on its partial dependence plot (contrary to X_4 (Sport)). Finally, as expected (since our example has been built such that X_3 (Split) does not influence the expected annual claimfrequency), the least important variable is X_3 (Split).

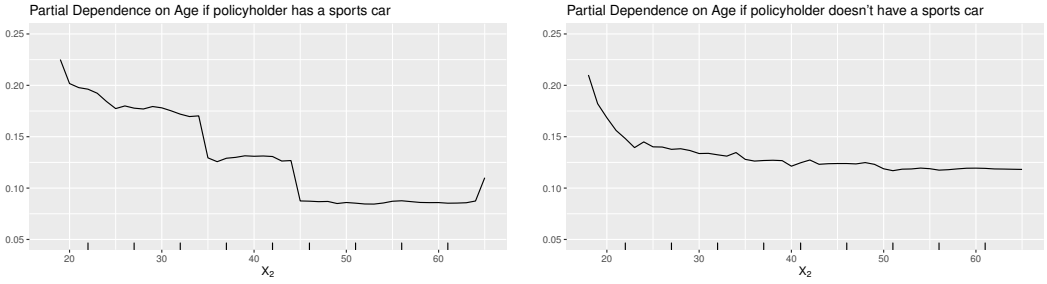

Figure 3.3: Partial dependence plots for X_2 (Age): $X_4 = yes$ (left) and $X_4 = no$ (right).

4 Conclusion

The analyst must be cautious in interpreting partial dependence plots. Indeed, interaction effects are not always visible. That is why these plots must be used in conjunction with other indicators such as the variable importances. To conclude, a feature with a flat partial dependence plot 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