Interpretable Machine Learning

Keshav Bhandari

August 11, 2018

Table of Contents

[Lift & Gain Charts on Test & Train Dataset 2](#_Toc48134609)

[Response Variable Summary Statistics 4](#_Toc48134610)

[Variable Importance 5](#_Toc48134611)

[Interpretation of PDP Plots 6](#_Toc48134612)

[duration 7](#_Toc48134613)

[euribor3m 8](#_Toc48134614)

[nr.employed 9](#_Toc48134615)

[emp.var.rate 10](#_Toc48134616)

[cons.conf.idx 11](#_Toc48134617)

[age 12](#_Toc48134618)

[pdays 13](#_Toc48134619)

[cons.price.idx 14](#_Toc48134620)

[poutcome\_success 15](#_Toc48134621)

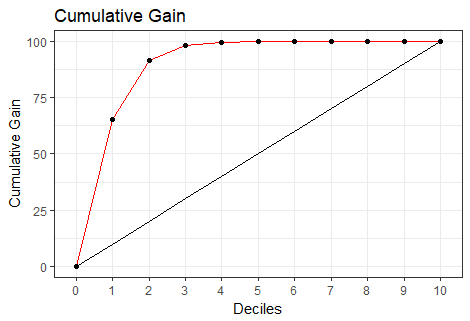
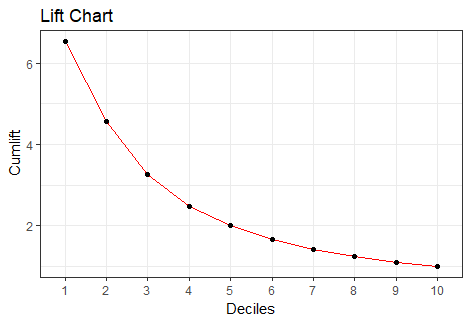
[campaign 16](#_Toc48134622)

[Frequently Asked Questions 17](#_Toc48134623)

[Methodology of Constructing PDP Plots 18](#_Toc48134624)

## Lift & Gain Charts on Test & Train Dataset

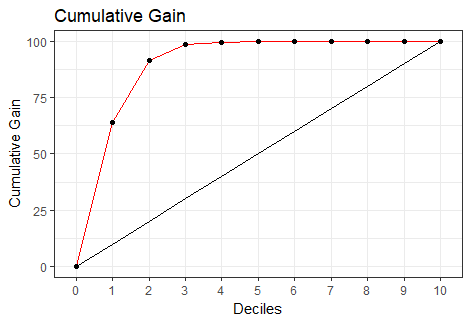
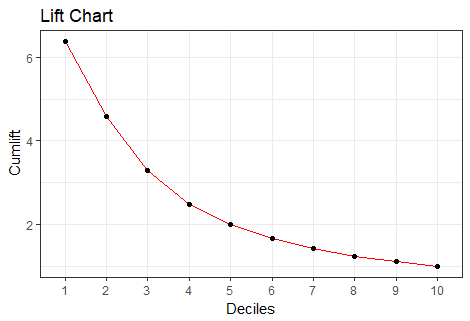
Lift & Gain Charts on Test Dataset:



Lift Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| bucket | total | totalresp | Cumresp | Gain | Cumlift | % Responders |
| 1 | 412 | 295 | 295 | 65.41 | 6.54 | 71.60 |
| 2 | 412 | 117 | 412 | 91.35 | 4.57 | 28.40 |
| 3 | 412 | 30 | 442 | 98.00 | 3.27 | 7.28 |
| 4 | 412 | 6 | 448 | 99.33 | 2.48 | 1.46 |
| 5 | 412 | 3 | 451 | 100.00 | 2.00 | 0.73 |
| 6 | 412 | 0 | 451 | 100.00 | 1.67 | 0.00 |
| 7 | 412 | 0 | 451 | 100.00 | 1.43 | 0.00 |
| 8 | 412 | 0 | 451 | 100.00 | 1.25 | 0.00 |
| 9 | 412 | 0 | 451 | 100.00 | 1.11 | 0.00 |
| 10 | 411 | 0 | 451 | 100.00 | 1.00 | 0.00 |

Lift & Gain Charts on Train Dataset:



Lift Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| bucket | total | totalresp | Cumresp | Gain | Cumlift | % Responders |
| 1 | 4119 | 2961 | 2961 | 63.81 | 6.38 | 71.89 |
| 2 | 4119 | 1288 | 4249 | 91.57 | 4.58 | 31.27 |
| 3 | 4119 | 325 | 4574 | 98.58 | 3.29 | 7.89 |
| 4 | 4119 | 48 | 4622 | 99.61 | 2.49 | 1.17 |
| 5 | 4118 | 15 | 4637 | 99.94 | 2.00 | 0.36 |
| 6 | 4119 | 2 | 4639 | 99.98 | 1.67 | 0.05 |
| 7 | 4119 | 0 | 4639 | 99.98 | 1.43 | 0.00 |
| 8 | 4119 | 1 | 4640 | 100.00 | 1.25 | 0.02 |
| 9 | 4119 | 0 | 4640 | 100.00 | 1.11 | 0.00 |
| 10 | 4118 | 0 | 4640 | 100.00 | 1.00 | 0.00 |

##### 

See the interpretation: [FAQ](#Frequently-Asked-Questions)

## Response Variable Summary Statistics

##### 

Train

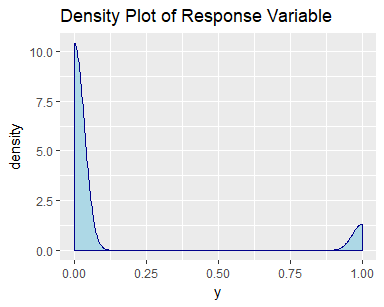
|  |  |  |
| --- | --- | --- |
| Response | Prop | Freq |
| 0 | 0.89 | 36548 |
| 1 | 0.11 | 4640 |

##### 

Test

|  |  |  |
| --- | --- | --- |
| Response | Prop | Freq |
| 0 | 0.89 | 3668 |
| 1 | 0.11 | 451 |

##### 

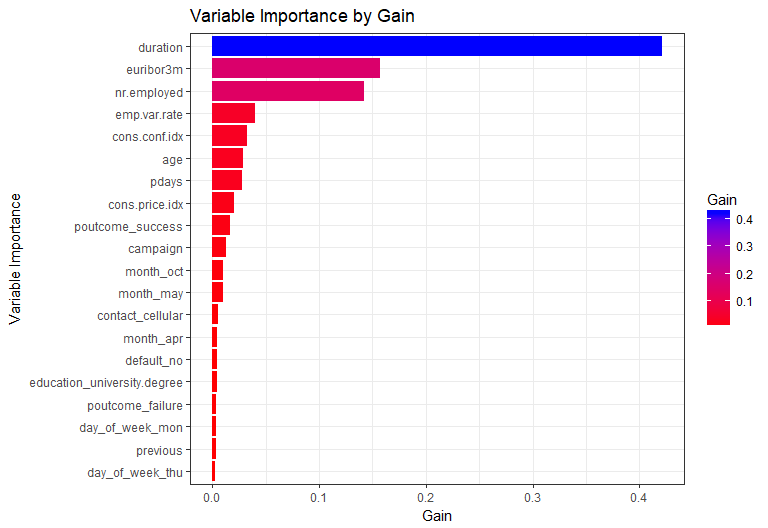


##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

## Variable Importance

The Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature’s contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating a prediction.



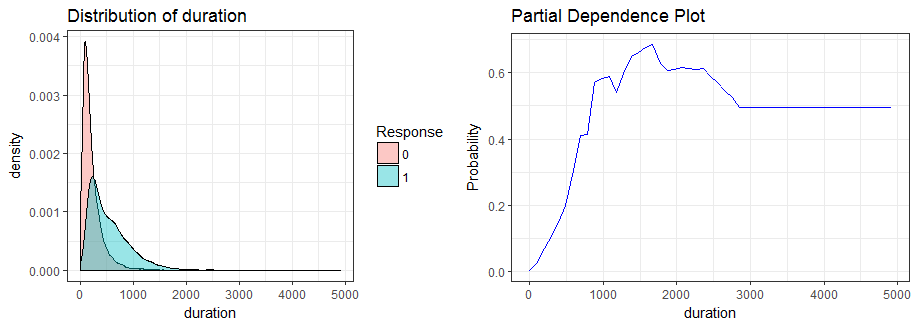
##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

## Interpretation of PDP Plots

The partial dependence plot shows the marginal effect of a feature on the predicted outcome of a previously fit model (J. H. Friedman). A partial dependence plot can show if the relationship between the target and a feature is linear, monotonic or more complex. Applied to a linear regression model, partial dependence plots will always show a linear relationship, for example. For a classification problem, an increasing trend of the partial dependence function implies a higher likelihood of the occurence of an event and vice versa.

**Example:** The following example is sourced from a well known UCL machine learning repository (bank classifying term deposit subscriptions). The response variable indicates whether the client subscribed to a term deposit (binary: 1 for ‘yes’, 0 for ‘no’). The variable of interest is ‘duration’ which is the last contact duration, in seconds (numeric) and is the most important variable in the prediction of term deposit subscriptions. An XGBoost model was fit to this dataset.



The partial dependency plot above indicates as the duration of the call goes up till around 1,600 seconds, the probability of the event occurring (client subscribing to the term deposit) is at its peak, holding all the other variables constant. This makes sense as the more the bank talks to a target client the higher the probability the target client will open a term deposit since a higher duration means a higher interest (commitment) from the potential client. Beyond 1600 seconds, the marginal effect of duration on the model’s predicted outcome starts decreasing slightly and after 2,800 seconds the trend is stagnant, indicating there is no effect of duration on term deposit subscriptions at this range. The density plot shows us the sparsity of values for the ‘duration’ variable after 2,250 seconds implying there are not many data points in that region and that we might not be able to confidently rely on the interpretation of the machine learning model predictions for values beyond this range.

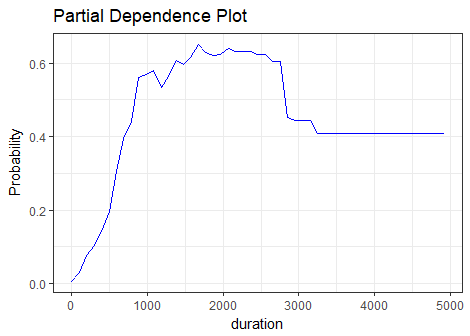
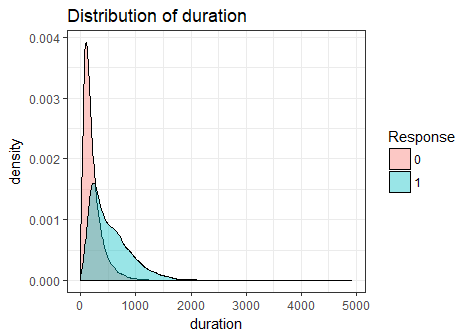
## duration

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 0 | 102 | 180 | 258.29 | 319 | 4918 | 1544 |
| validation | 0 | 103 | 181 | 256.79 | 317 | 3643 | 828 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

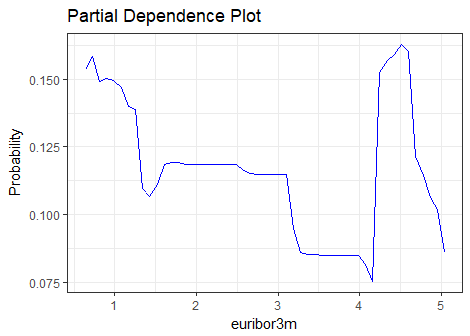
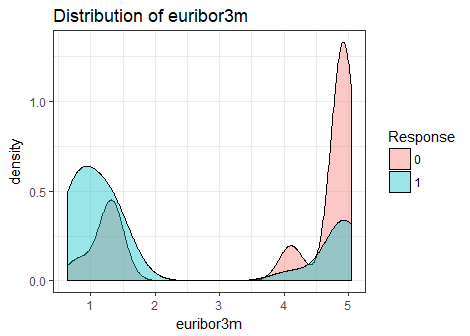
## euribor3m

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 0.63 | 1.34 | 4.86 | 3.62 | 4.96 | 5.04 | 316 |
| validation | 0.64 | 1.33 | 4.86 | 3.62 | 4.96 | 5.04 | 234 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

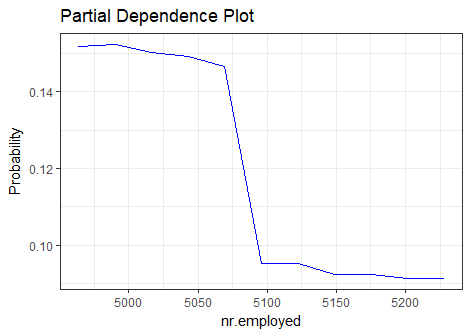
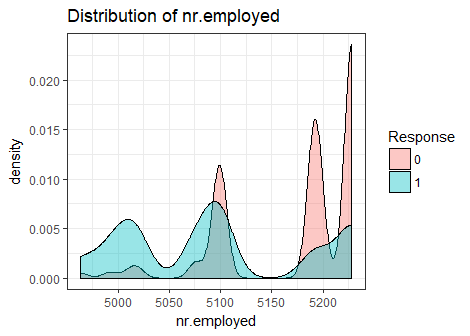
## nr.employed

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 4963.6 | 5099.1 | 5191 | 5167.04 | 5228.1 | 5228.1 | 11 |
| validation | 4963.6 | 5099.1 | 5191 | 5166.48 | 5228.1 | 5228.1 | 11 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

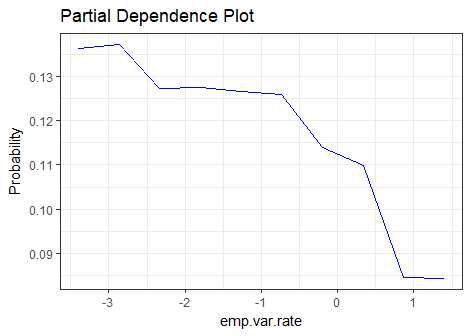
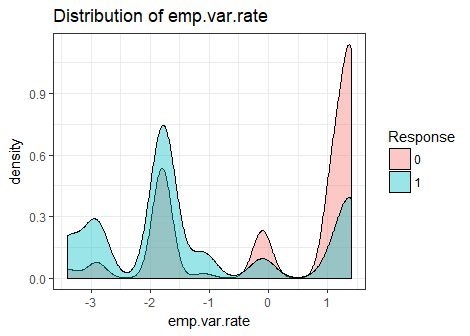
## emp.var.rate

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | -3.4 | -1.8 | 1.1 | 0.08 | 1.4 | 1.4 | 10 |
| validation | -3.4 | -1.8 | 1.1 | 0.08 | 1.4 | 1.4 | 10 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

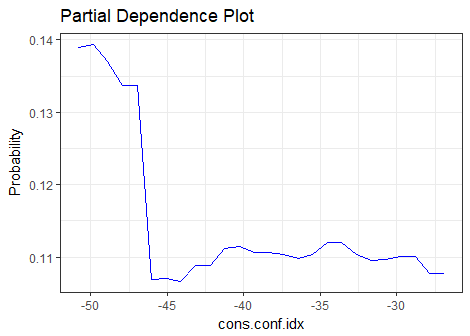
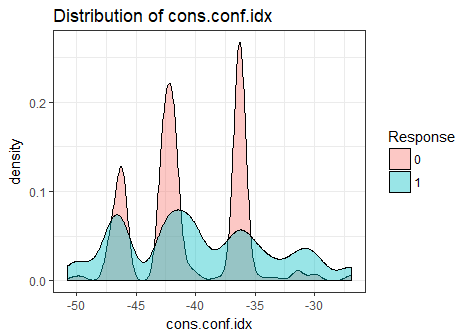
## cons.conf.idx

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | -50.8 | -42.7 | -41.8 | -40.5 | -36.4 | -26.9 | 26 |
| validation | -50.8 | -42.7 | -41.8 | -40.5 | -36.4 | -26.9 | 26 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

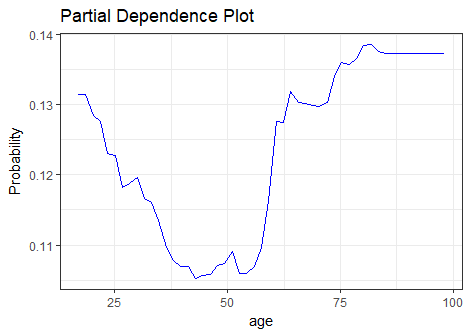
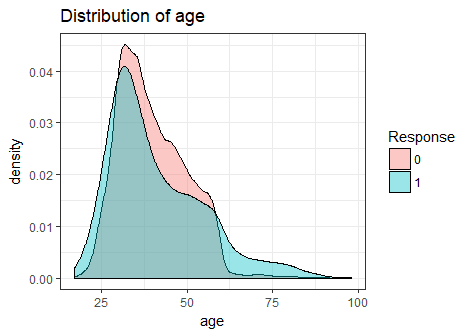
## age

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 17 | 32 | 38 | 40.02 | 47 | 98 | 78 |
| validation | 18 | 32 | 38 | 40.11 | 47 | 88 | 67 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

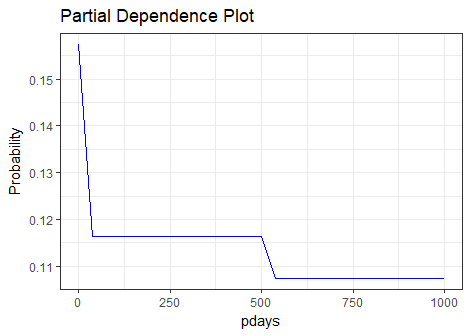
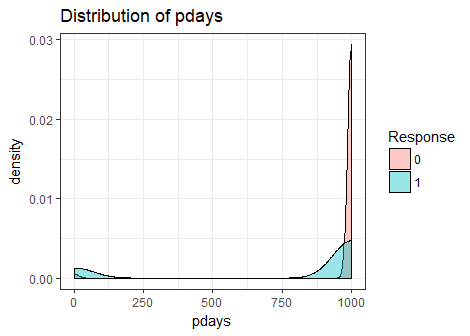
## pdays

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 0 | 999 | 999 | 962.48 | 999 | 999 | 27 |
| validation | 0 | 999 | 999 | 960.42 | 999 | 999 | 21 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

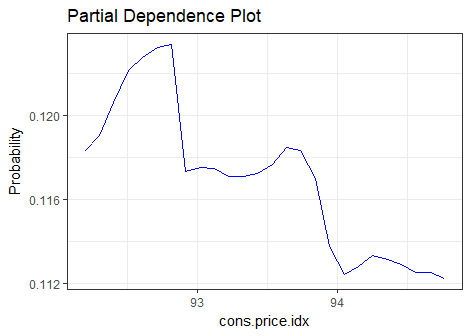
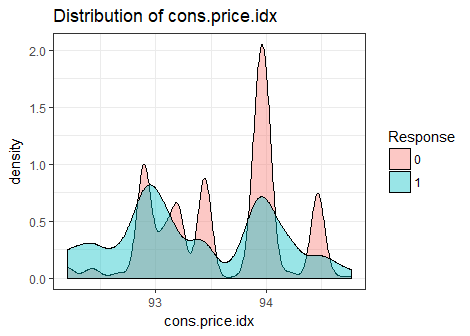
## cons.price.idx

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 92.2 | 93.08 | 93.75 | 93.58 | 93.99 | 94.77 | 26 |
| validation | 92.2 | 93.08 | 93.75 | 93.58 | 93.99 | 94.77 | 26 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

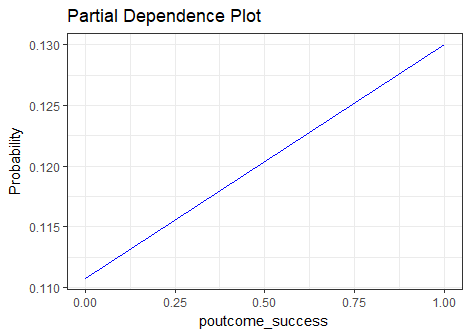
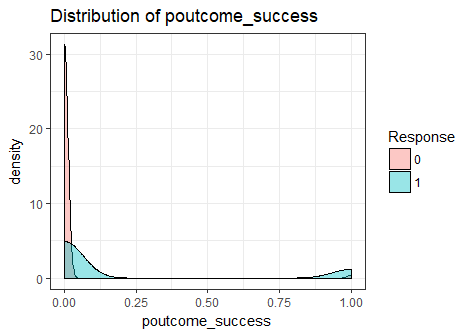
## poutcome\_success

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 0 | 0 | 0 | 0.03 | 0 | 1 | 2 |
| validation | 0 | 0 | 0 | 0.03 | 0 | 1 | 2 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

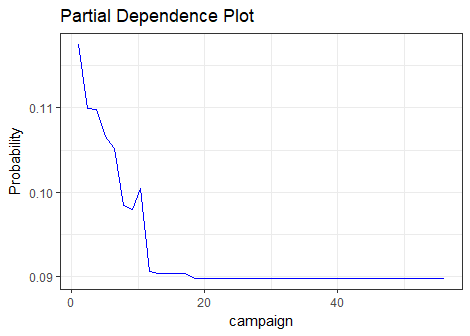
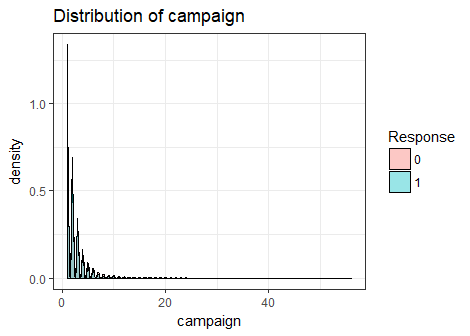
## campaign

##### 

Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Unique |
| train | 1 | 1 | 2 | 2.57 | 3 | 56 | 42 |
| validation | 1 | 1 | 2 | 2.54 | 3 | 35 | 25 |

##### 



##### 

##### 

##### 

##### 

Go back to table of contents: [Click Here](#Table_of_Contents)

## Frequently Asked Questions

**1. What do the Y-axis mean in partial dependence plots?**  
The y-axis of a partial dependence plot for classification represents the marginal impact of the independent variable to the dependent variable, holding the other variables constant. It does not represent the predicted value or relative impact of this variable to other variables.

**2. How does one interpret the density/distribution plot?**  
A density plot is a smoothed, continuous version of a histogram estimated from the data. This chart is a variation of a Histogram that uses kernel smoothing to plot values, allowing for smoother distributions by smoothing out the noise. The peaks of a Density Plot help display where values are concentrated over the interval.

**3. What does the Y-axis of density plots represent?**  
The y-axis in a density plot is the probability density function for the kernel density estimation. However, we need to be careful to specify this is a probability density and not a probability. The difference is the probability density is the probability per unit on the x-axis. To convert to an actual probability, we need to find the area under the curve for a specific interval on the x-axis. Somewhat confusingly, because this is a probability density and not a probability, the y-axis can take values greater than one. The only requirement of the density plot is that the total area under the curve integrates to one. One can think of the y-axis on a density plot as a value only for relative comparisons between different categories.

**4. How do the density plots add value to partial dependence plots?**  
The density plots enhance the interpretation of partial dependence plots by showing us the distribution of an independent variable with respect to the dependent variable. Sparsely populated intervals of the density plot indicate low confidence for the corresponding X-axis range in a partial dependence plot. In contrast, highly populated areas of the density plot add greater trust to the interpretation of the partial dependence plot for the corresponding X-axis interval of the variable of interest.

In addition, it helps Validate general business interpretations for the variable of interest by comparing its actual distribution broken out by the response variable to what is predicted by the model. To see the interpretation of PDP plots with an example: [Click Here](#Interpretation-of-PDP-Plots)

**5. How does one interpret the gain and lift charts?**  
Gain and Lift charts are used to evaluate performance of classification model. They measure how much better one can expect to do with the predictive model than without it.

Gain at a given decile level is the ratio of cumulative number of targets (events) up to that decile to the total number of targets (events) in the entire data set.

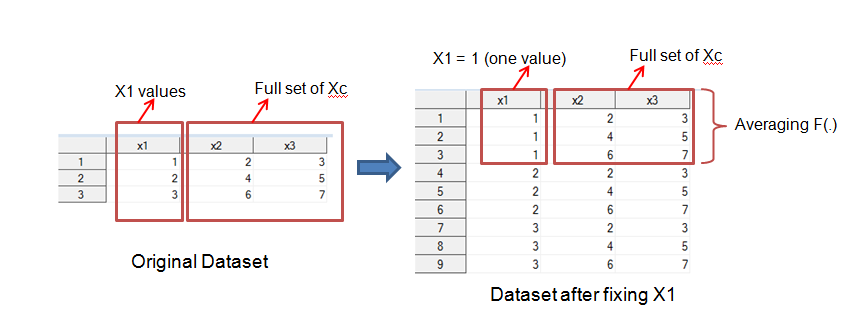
Lift measures how much better one can expect to do with the predictive model comparing without a model. It is the ratio of gain % to the random expectation % at a given decile level. The random expectation at the 5th decile is 50%.

Go back to table of contents: [Click Here](#Table_of_Contents)

## Methodology of Constructing PDP Plots

Partial dependence plot is plotting the independent variable x1 vs the model outcome y\_bar, after considering the average effect of other independent variables in the model. Notice that partial dependence plot is NOT plotting relationship between independent variables and target variable.

Suppose we have the model equation F(.). One independent variable is X1 and all other independent variables are Xc. Suppose X1 has multiple values (X11,X12,X13….X1n). For each value (X1i), we can have a partial dependence function: avg(F(X1=X1i, Xc=Xc\_full)). Here we fix X1=X1i, while other variables Xc\_full is the full set of data. Let’s use a graph to further explain:



From the above explanation, we can find that we are plotting the “marginal effect” of X1 on the predict outcome F(.) by averaging the effect of other variables on F(.). In other words, we are averaging the predicted probabilities obtained from the XGBoost model when X1 = 1, when X1 = 2 and when X1 = 3. The 3 average values obtained from the table on the top right side would then be used to construct the partial dependence plot. This plot is particular useful when we want to figure out the “marginal effect” of one (or several) particular variable(s) on the predicted outcome when using ensemble data mining methods (like boost or bagging). Because of its complex nature, we cannot identify the marginal effect by simply reviewing the equation.

Go back to table of contents: [Click Here](#Table_of_Contents)