**Explainable Machine Learning**

Now we know who all the customers that are going to churn, the next question that the end user looks for what is reason behind this high-risk customer? why they are going to churn? Over the period of time this will be important to know the root cause for the churn.

In machine learning complex model has big issue with transparency, we don’t have any strong prove why model give that prediction and which feature are impacting the model prediction, which features are strongly contributing, and which are negative contribution for model prediction. By feature importance graph we can see which features importance by passing complete training and test dataset, but for single row of features or for any given instance it is very difficult to understand why and how model predict output. And, we will be answering to this question through the **SHAP** (Shapley Additive explanations)

**SHAP**

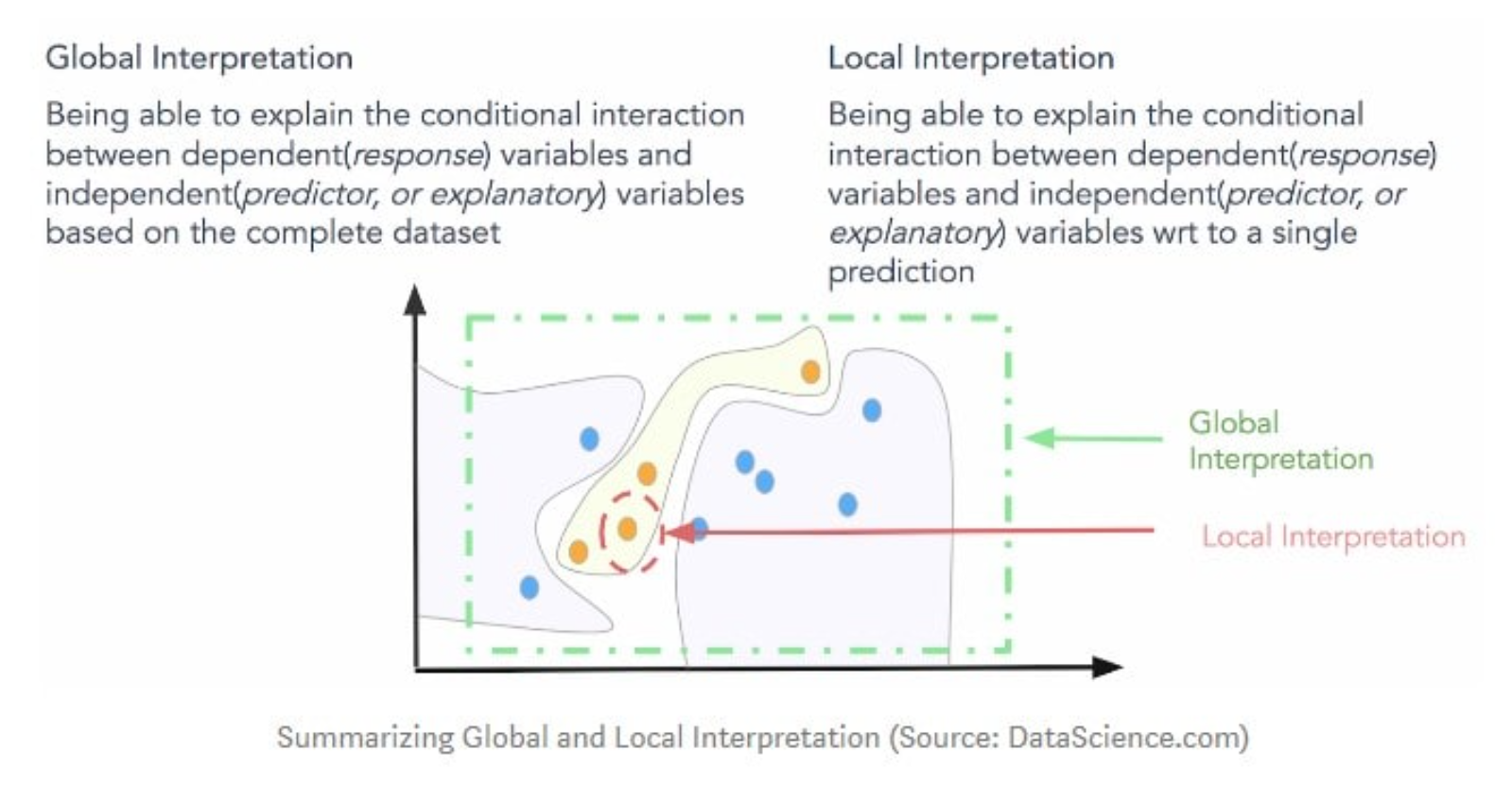
SHAP goal is to explain the prediction of a given instance X by computing the contribution of each feature to the prediction. The feature values of a data instance act as players in a coalitional game theory. SHAP prediction output is a fair distribution of all the feature Shapley values. Shapely value is actually distribution, it’s an average of model contribution made by each player(features) over all permutation of player(features). The baseline for Shapley values is the average of all predictions. In the plot, each Shapley value is an arrow that pushes to increase (positive value) or decrease (negative value) the prediction.

**Global vs Local Interpretation:**

Here we will understand the model at both Global and Local level.

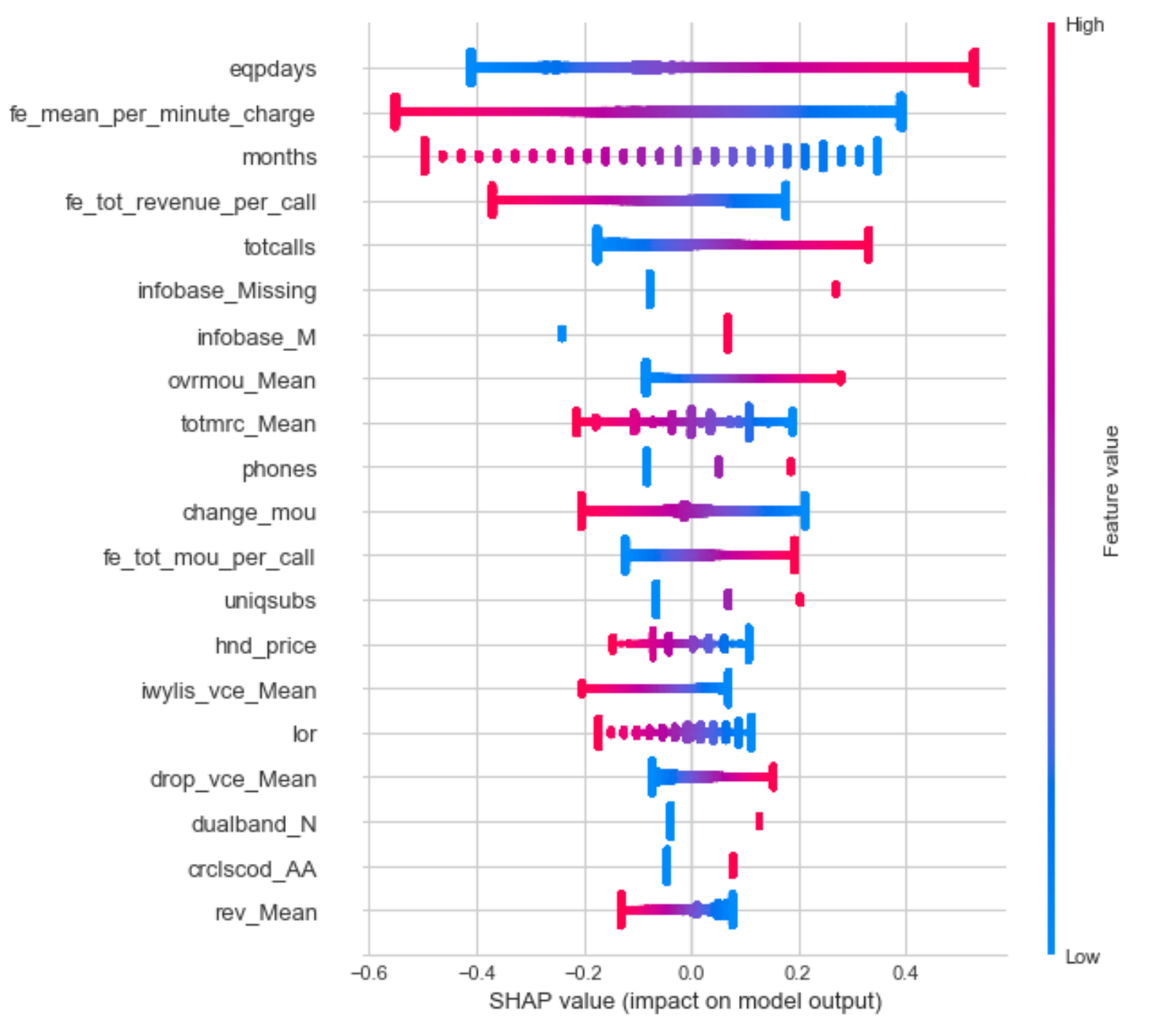
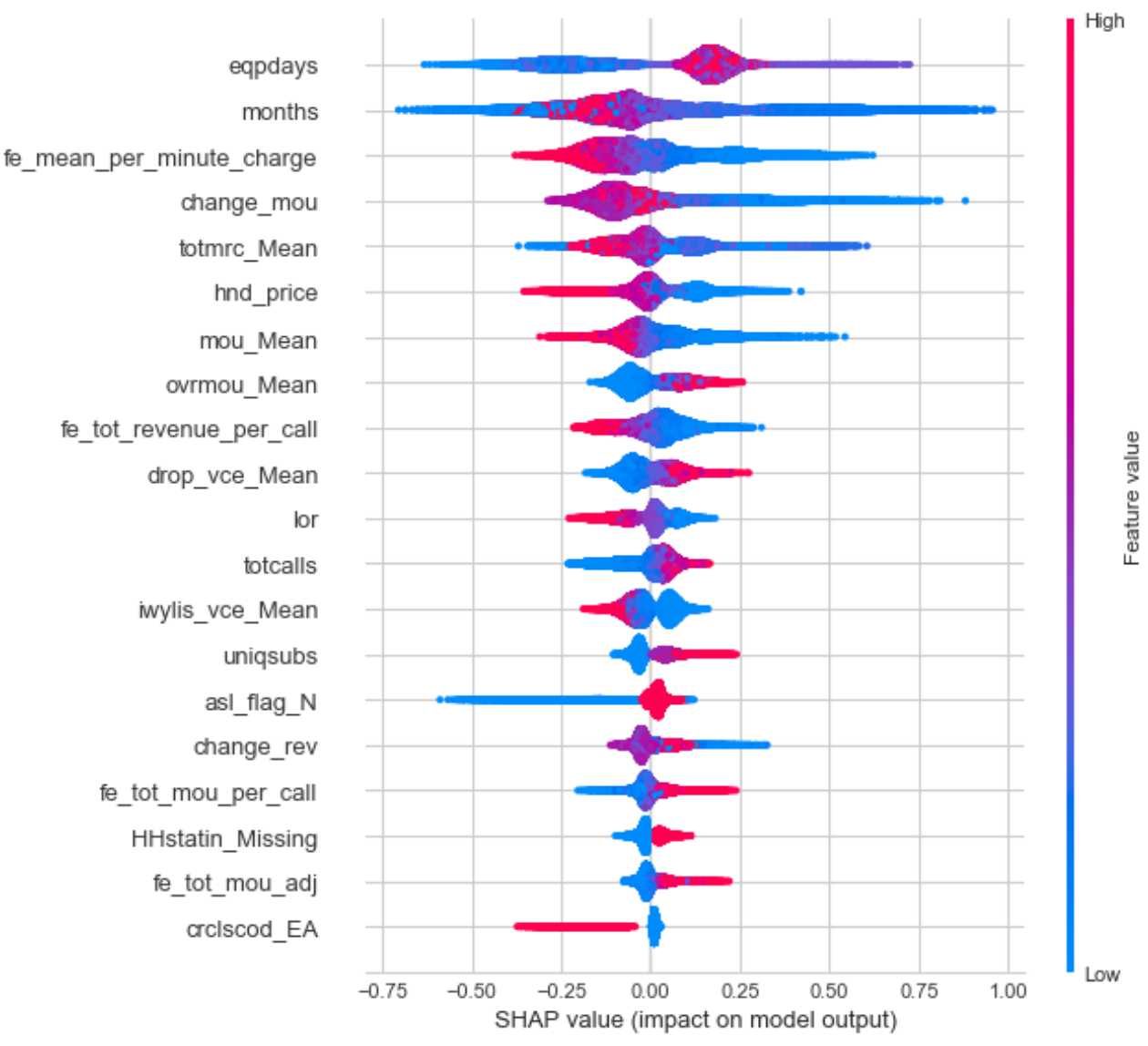
At Global level we are explaining the interaction of our churn variable with Independent variabels on the complete dataset.

At local level we are explaining the interaction of churn variable on each single prediction.



**Global Interpretation**:

At a global level, the below graphs are summarizing the effects of all the explanatory variables on the model output, colour coded to show the direction of the impact (red means an increase, while blue shows a decrease), with SHAP value more far away from zero meaning a bigger impact. It is also visually easy to see which variables have the strongest relationship with the target variable. In this way, SHAP can also be used as a tool for variable selection.

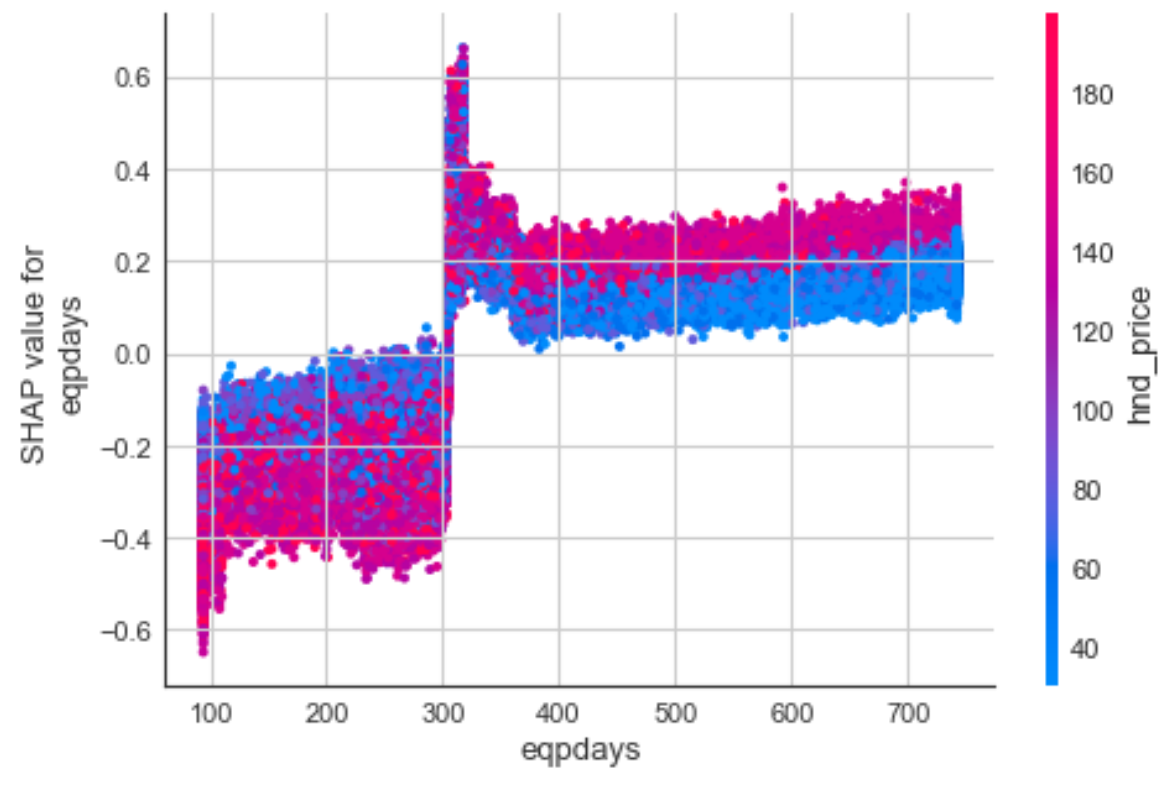
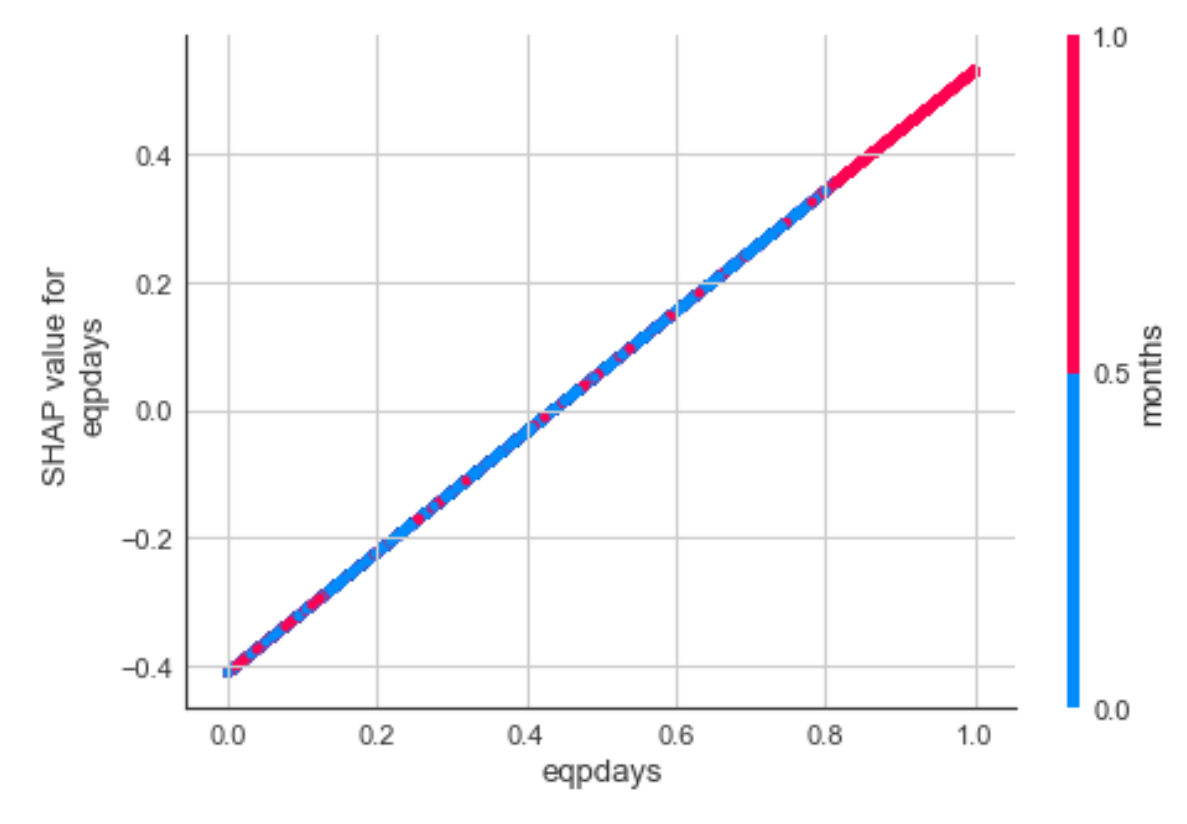
 

Light GBM

Logistic Regression

In both models, the eqpdays (Number of days (age) of current equipment) variable has the most predictive power. The order of importance, further on, is slightly different, given that the regression is constrained in fitting the relationship to a linear one, while the LightGBM can use non-linear components to describe it. This is also apparent in the single variable graphs, which, in addition to showing the positive/negative relationships to the target, are also showing the form of the relationship.

The plots below represent the change in propensity score of churn as the most predictive variable (eqpdays) changes. Vertical dispersion at a single value of eqpdays (more visible on the LightGBM plot) represents interaction effects with other features. To help reveal these interactions, the plot automatically selects another feature for colouring (see right vertical axis).



Light GBM

Logistic Regression

The LightGBM model reveals a more complex relationship being captured, and it better explains interactions between variables.

**Local Interpretation**:

Another great aspect of SHAP is that it determines a separate set of values for each observation in the dataset. This feature can have multiple usages:

* It allows to explain why model output takes given value for each observation (in case of case, each risk/non risk customers can be explained).
* It can determine the observations where a certain variable or a set of variables are more/less predictive, and thus it aids in segmentation.
* It can help in optimizing the model by removing outliers (observations where SHAP values are low for a big number/ all variables)
* It can help in explaining interdependencies between variables at a local/ segment level
* It can help with model exclusions, as missing features have no attributed impact to the model parameters.

For a local view which makes it clearer which way each variable is ‘pushing’ the model output towards, the following plots can be used, selecting the row/observation we want to show:

Logistic Regression output