ofgem
Making a positive difference
for energy consumers

Identifying customers of interest: a data-driven approach

This study focussed on a specific target group: customers who have been on a default tariff for more than 3 years. The aim was to identify attributes of long-term disengaged customers that could be used to predict long-term disengagement.

The dataset contained over 600,000 customers on default energy tariffs from five suppliers which were anonymised to Nala, Mufasa, Timon, Pumba, Simba. The dataset also contained a number of attributes relating to customers' payment and consumption behaviour

Using the open-source data science tool AutoML from h2o and the data platform built by data services in Ofgem, these data were pre-processed and run through a series of **supervised machine-learning models**. Supervised machine-learning models attempt to learn a mapping function (a formula) from input data to output data. In this case study our input data are features about a customer and our output data are whether they are long-term disengaged (on a default tariff for greater than 3 years) or not (on a default tariff for less than three years). The approach used here not only optimises each type of model in its ability to predict these output data, it compares each individual model to combinations of models leveraging new data science methods and technology to determine the most representative model of that function.

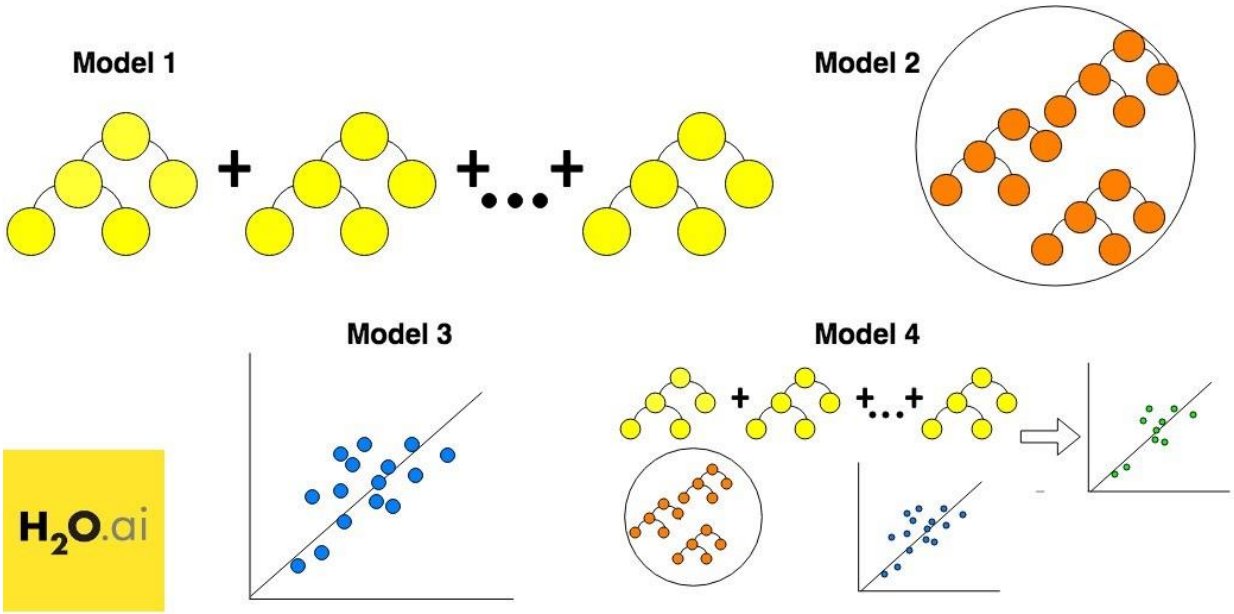

Figure 1) Different supervised learning models were tested using AutoML from h2o. Model 1 is a schematic of a sequential tree-based algorithm such as gradient-boosting or random forest algorithms. Model 2 is a schematic of a parallel tree-based algorithm such as distributed random forest. Model 3 represents an algorithm assessing a linear relationship between the target and predictors such as a generalised linear model. Model 4 represents a stacked ensemble method that takes a combination of these algorithms as base learners and then uses a meta-learner to find the optimal combination of the base learners.

The winning model was used to **rank features** within the dataset that were more likely **to predict** a long-term disengaged customer.

Once a model was selected, it was then used to assess the attributes and identify those which could help predict which customers were more likely to disengage over the long-term.

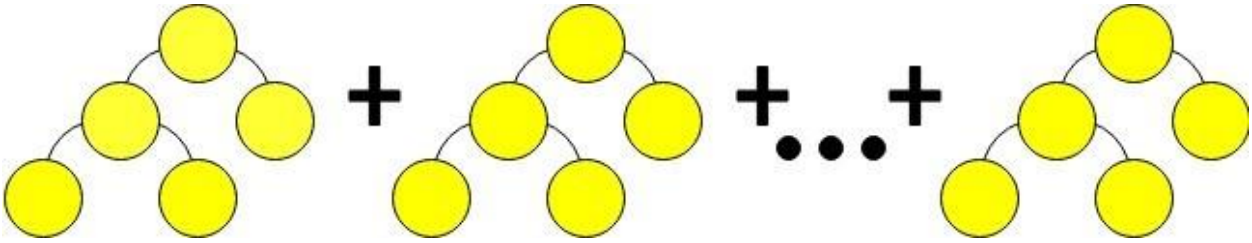

Figure 2) The winning model with the highest accuracy at predicting reserved, unseen data, was selected. This model was most accurate at mapping the function from the input data (features about energy customers) to output data (whether they were long-term disengaged).

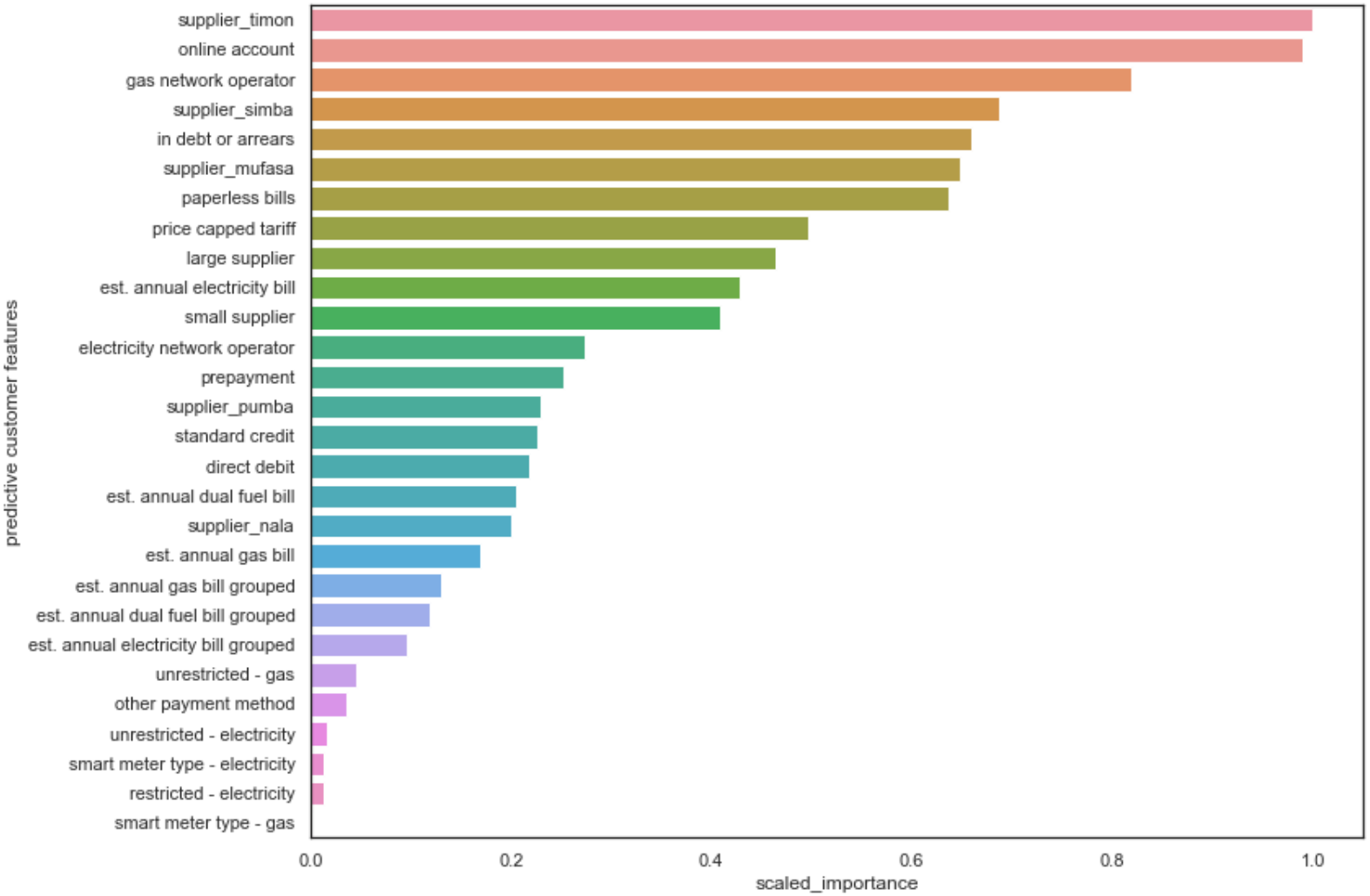

Figure 3) The attributes that identify which customers have been on a default tariff for 3 years or more.

We can explore these attributes further to show exactly how they predict a long-term disengaged customer, for instance whether they are positively or negatively correlated. (See figure 5 for an example of a negatively correlated feature).

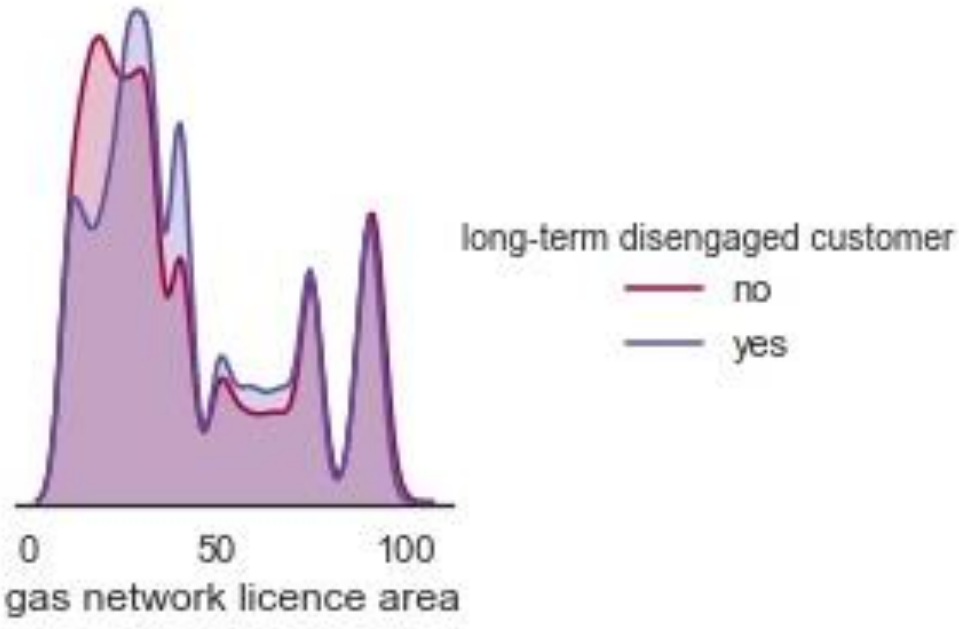

Figure 4) The volume of long-term disengaged customers ('yes') differs across the different areas compared to non-long-term disengaged customers ('no'). The numbers on the x-axis represent different regions within the UK.

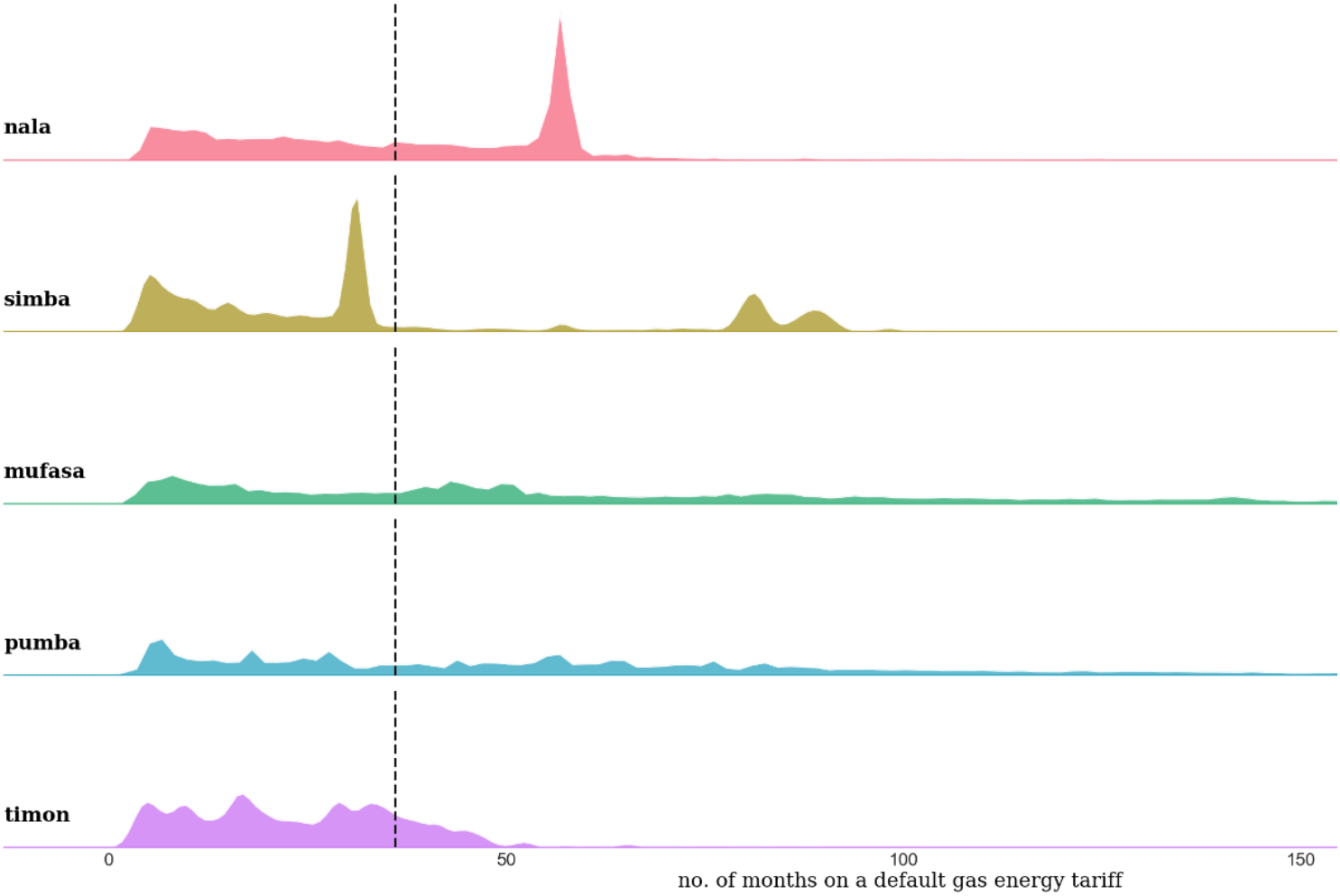

Figure 5) The relative number of long-term disengaged customers differs between suppliers. To the right of the dashed line, customers are considered long-term disengaged customers. Timon had the smallest relative number of long-term disengaged customers compared to the other suppliers. As a strong predictor it negatively predicts long-term disengagement.

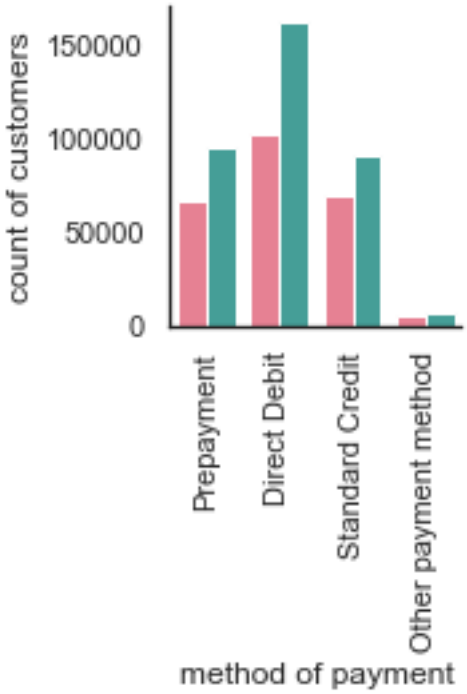

Figure 6) More direct debit customers are long-term disengaged customers ('yes') than non-long-term customers ('no').

These attributes of long-term disengaged customers can be used to identify specific customers who may in future be likely to disengage over the long-term.

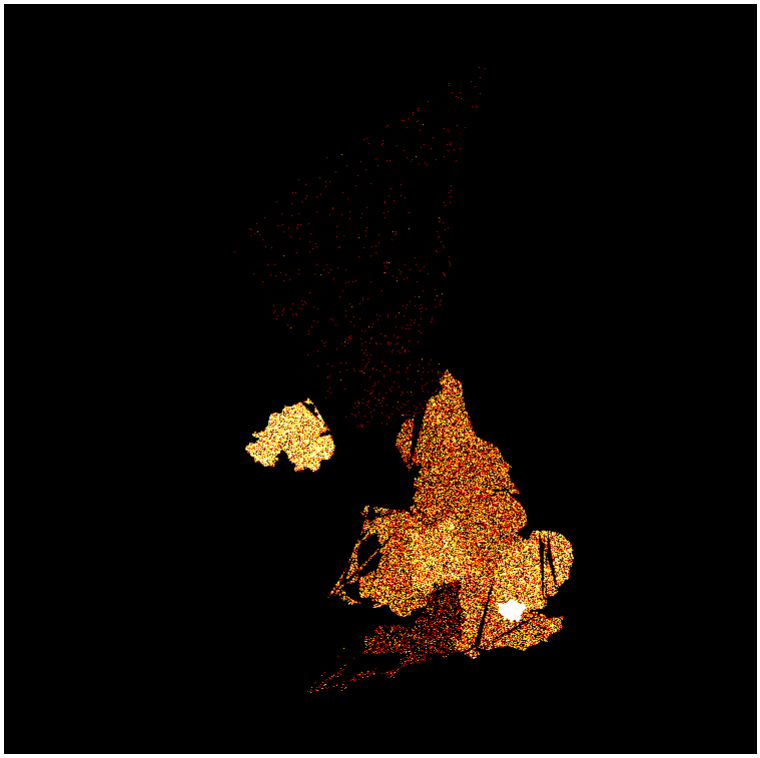

Figure 7A) Each point represents a single ***customer that has been on a default tariff for under three years*** (178,985 customers).

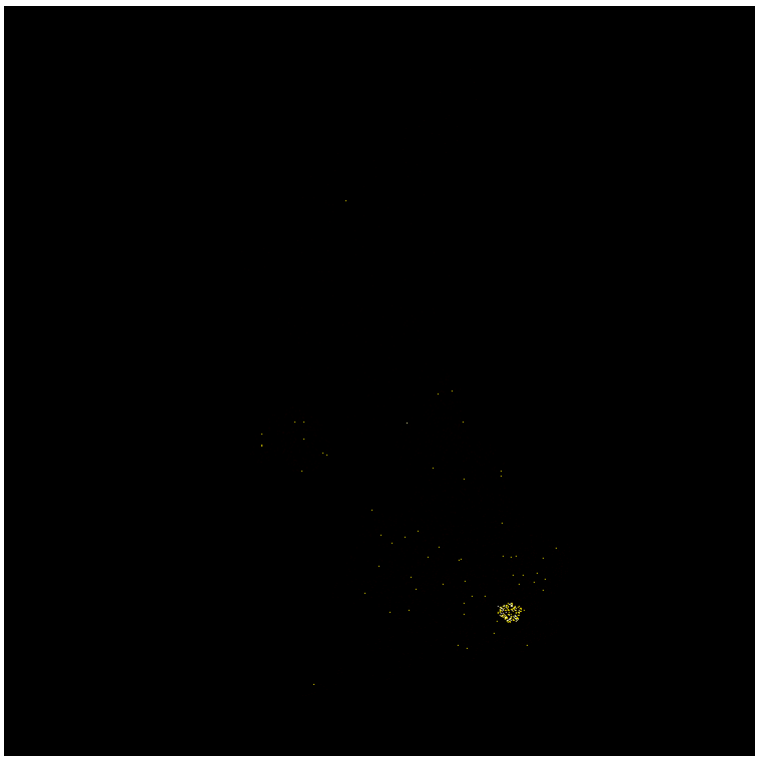

Figure 7B) Each point represents a single customer that has been on a default tariff for less than three years and some of the features that predict they will be long-term disengaged i.e. not with supplier Timon, not using online account management, on a price-capped tariff, in debt/arrears and with a large supplier. (2482 customers)

In Figure 7B each point represents a customer on a default tariff for under three years and has features that strongly predict long term disengagement. There is an opportunity to engage with this sub-group to encourage switching, lessening the likelihood of future disengagement.

Data: shaping policy & driving engagement
This approach of using data science to identify customer groups of interest can be applied to large datasets with a multitude of features to find the most accurate ways of predicting consumer behavior. This in turn means we can shape policy more effectively and drive increases in consumer engagement.