Orange3 Explain Documentation

Biolab

Jan 24, 2022

CONTENTS

CHAPTER

ONE

WIDGETS

1.1 Feature Importance

Inspect model using the Permutation Feature Importance technique.

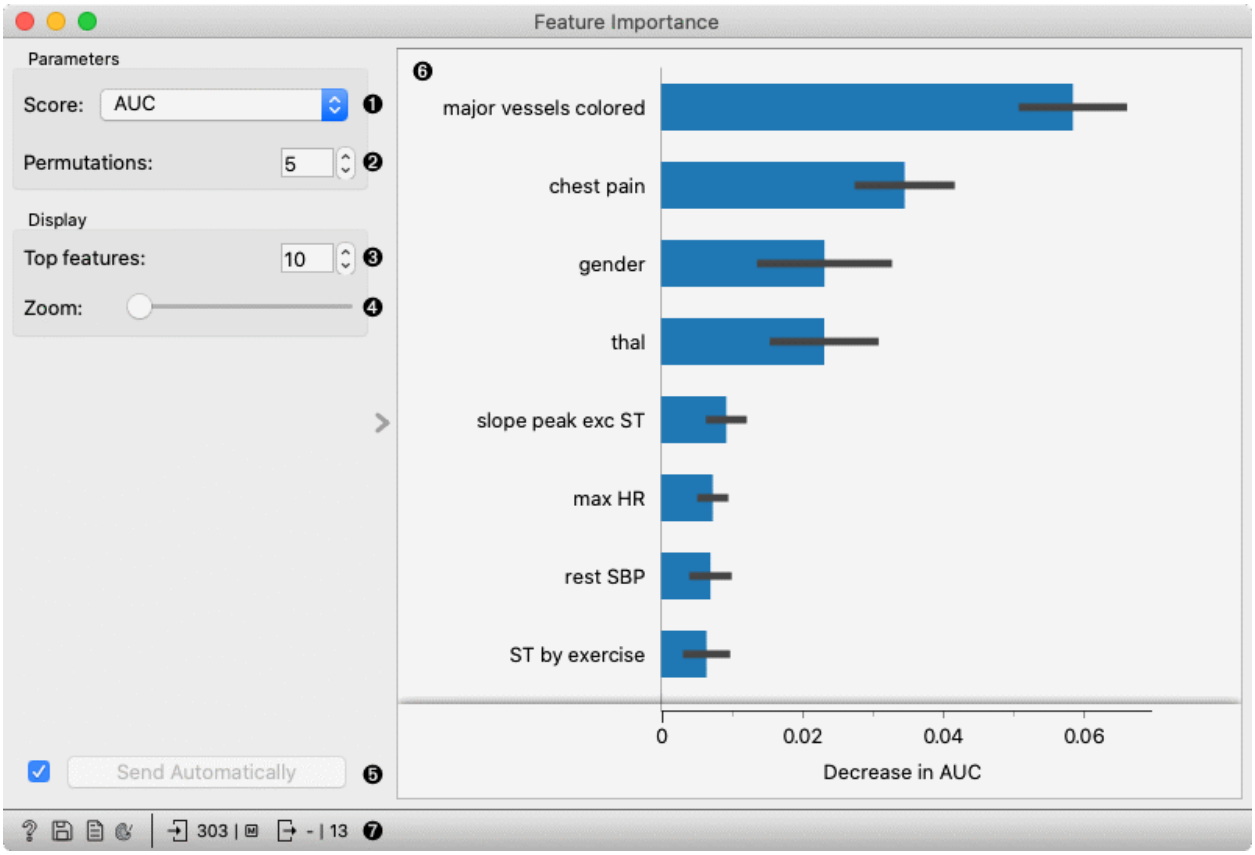
Inputs

- Data: dataset used to compute the explanations
- Model: a model which widget explains

Outputs

- Selected data: data instances that belong to selected features in the plot
- Scores: Mean and standard deviation of score for each feature.

Feature Importance widget explains classification and regression models. The widget gets a trained model and reference data on input. It uses the provided data to compute the contribution of each feature toward the prediction, by measuring the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the target.



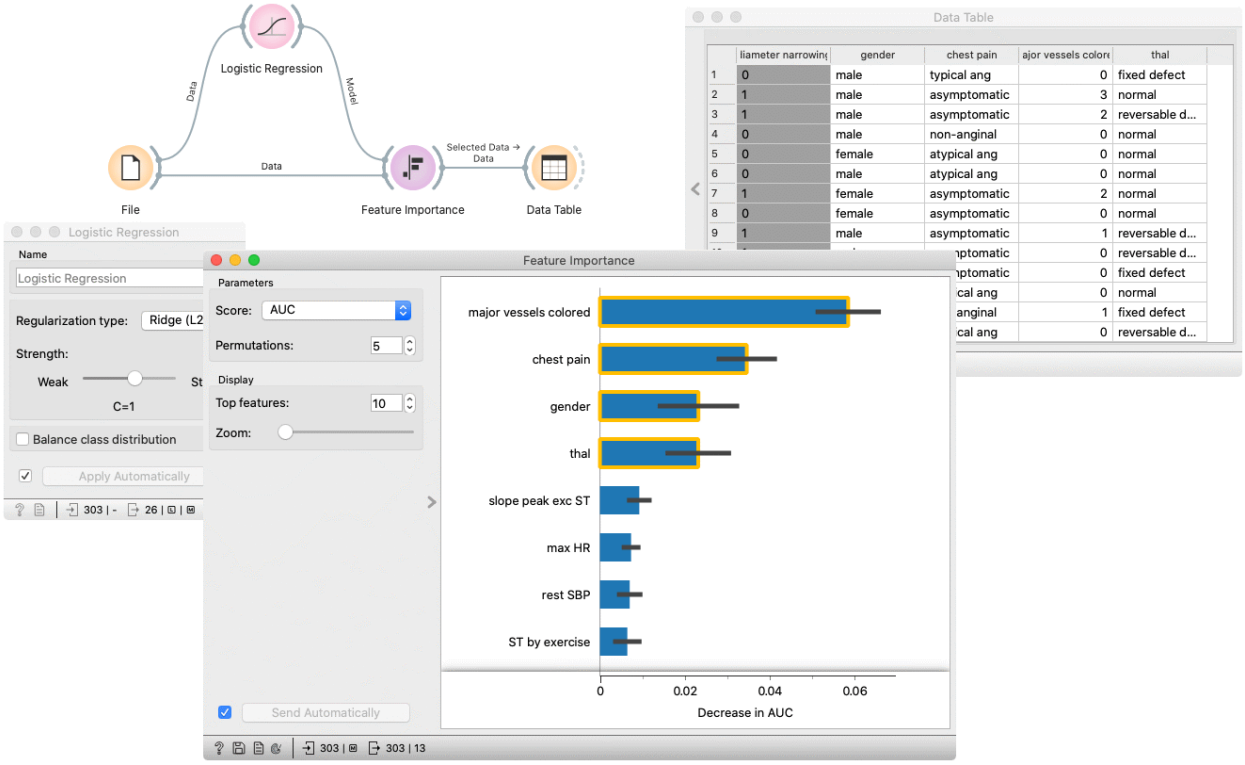
1. Select the scoring metric.
2. Select the number of times to permute a feature.
3. Select the number of the features to be shown in the plot.
4. Zoom in/out the plot.
5. Press *Apply* to commit the selection.
6. Plot which shows the selected number of features that are most important for a model.
7. Get help, save the plot, make the report, set plot properties, or observe the size of input and output data.

1.1.1 Example

In the flowing example, we use the Feature Importance widget to explain features, used in Logistic regression model. In the File widget, we open Hearth disease dataset. We connect it to Logistic regression widget, which trains the model. Feature Importance widget accepts the model and data which are used to explain the features. For an explanation, we usually use the same data than for training, but it is also possible to explain the features on different data (e.g. reference data subset).

The features in the plot are ordered by their relevance (e.g. Major vessels coloured is the most important feature).

By selecting some arbitrary features, a filtered input dataset appears on the output of the Feature Importance widget.



1.2 Explain Model

Explains a classification or regression model. Explains which features contribute the most and how they contribute toward the prediction for a specific class.

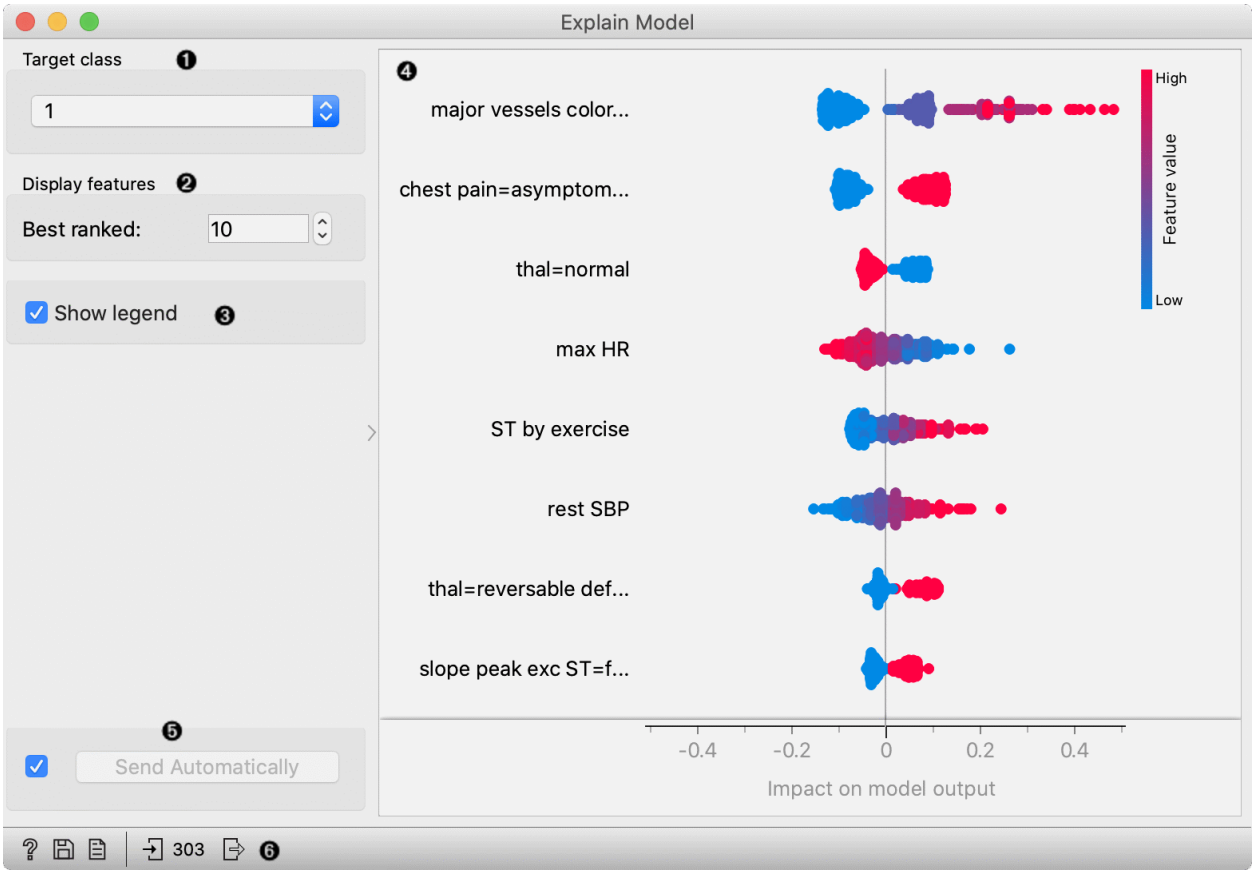
Inputs

- Data: dataset used to compute the explanations
- Model: a model which widget explains

Outputs

- Selected data: data instance that belong to selected points in the plot
- Scores: The score of each attribute. Features that contribute more toward the final prediction have higher scores.

Explain Model widget explains classification and regression models with SHAP library. The widget gets a trained model and reference data on input. It uses the provided data to compute the contribution of each feature toward the prediction for a selected class.

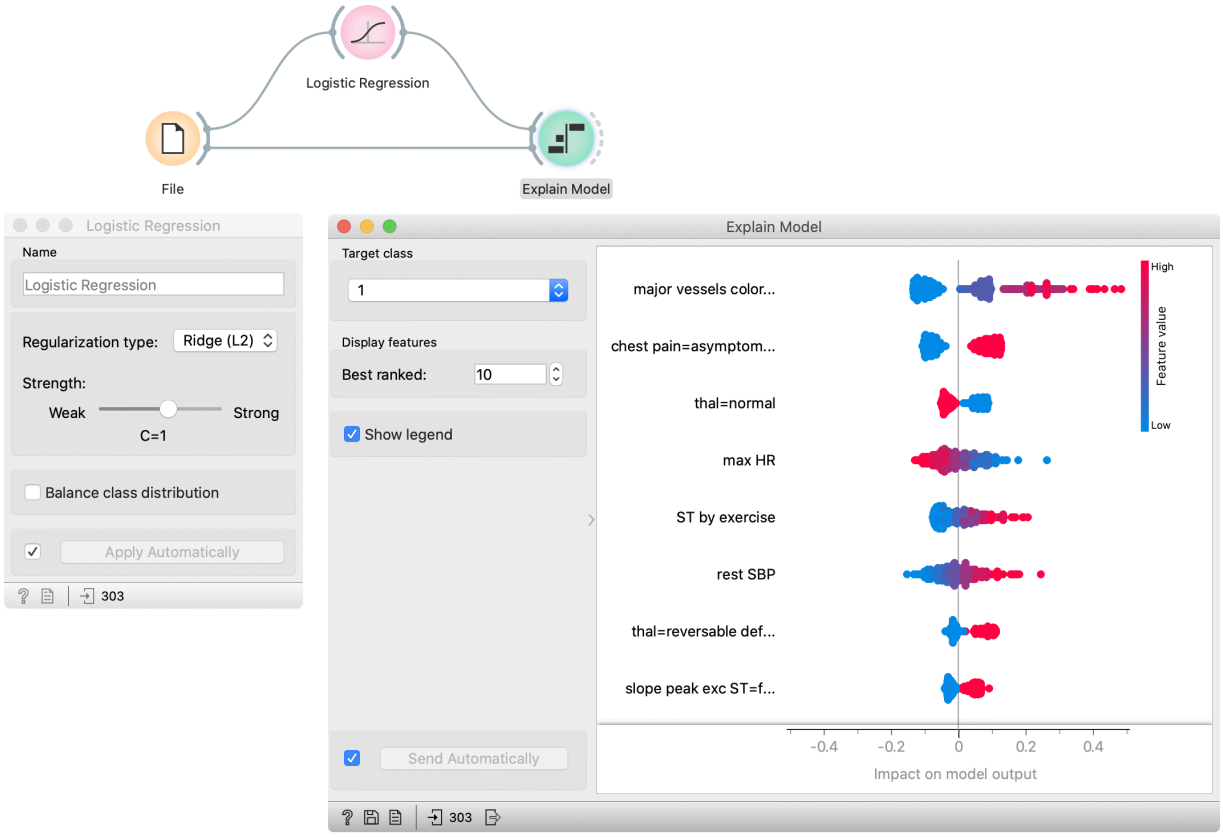


1. Select the target class – the plot will show explanations for this class.
2. Select number of the features shown in the plot.
3. Show/hide the legend.
4. Plot which shows the selected number of features that are most important for a model. For each feature, points in the graph show SHAP values (horizontal axis) for each data instance (row) in the data. SHAP value is a measure of how much each feature affect the model output. Higher SHAP value (higher deviation from the centre of the graph) means that feature value has a higher impact on the prediction for the selected class. Positive SHAP values (points right from the centre) are feature values with the impact toward the prediction for the selected class. Negative values (points left from the centre) have an impact against classification in this class. For regression, SHAP value shows how much the feature value affects the predicted value from the average prediction. Colours represent the value of each feature. Red colour represents higher feature value, while blue colour is a lower value. The colour range is defined based on all values in the dataset for a feature.
5. Press *Apply* to commit the selection.
6. Get help, save the plot, make the report, or observe the size of input and output data.

1.2.1 Example

In the flowing example, we use the Explain Model widget to explain Logistic regression model. In the File widget, we open Hearth disease dataset. We connect it to Logistic regression widget, which trains the model. Explain Model widget accepts the model and data which are used to explain the model. For an explanation, we usually use the same data than for training, but it is also possible to explain the model on different data. In the Explain model widget, we set the target class on the class to 1 – it means that we observe features that contribute the most to the prediction of a patient with diagnosed heart disease.

Features in the plot are ordered by their relevance to the prediction. Major vessels coloured is the most important for the prediction in class 1. Instances with higher values of this feature (red colour) have higher SHAP values which mean they contribute toward the prediction of class 1. Lower values of this attribute (blue) contribute against the prediction of this class. The second most important attribute is chest pain (categorical attribute) with value asymptomatic. The presence of this category for the patient (red colour) contributes toward the prediction of class 1, while the absence of this category contributes against class 1.



1.3 Explain Prediction

Explains which features contribute the most to the prediction for a single instance based on the model and how they contribute.

Inputs

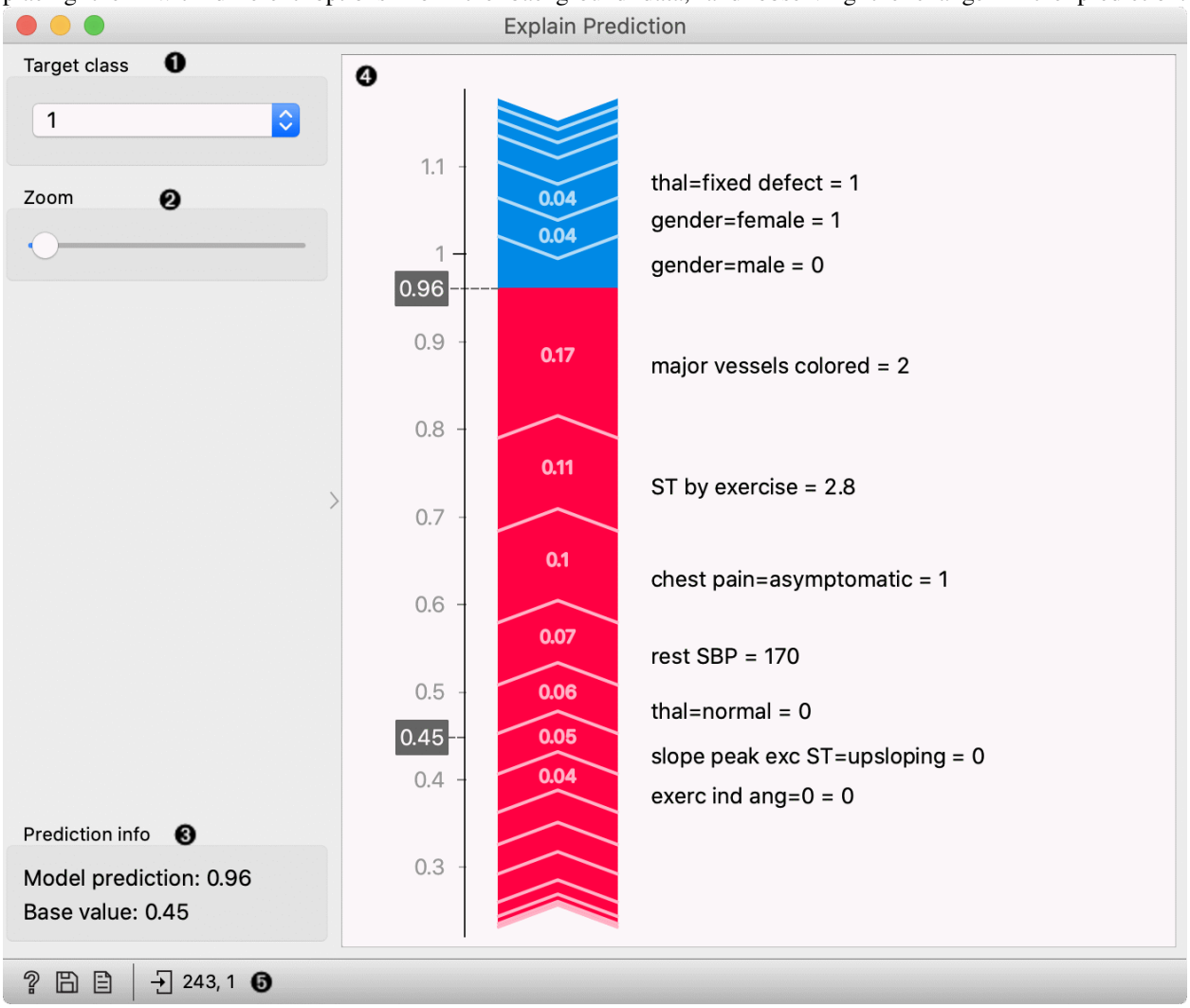
- Model: a model whose predictions are explained by the widget
- Background data: data needed to compute explanations

- Data: Single data instance whose prediction is explained by the widget

Outputs

- Scores: The SHAP value of each features value. Features that contribute more to prediction have higher score deviation from the 0.

Explain Prediction widget explains classification or regression model's prediction for the provided data instance. The widget shows what features affect the prediction of selected class the most and how they contribute (towards or against the prediction). The explanation is computed with removing features, replacing them with different options from the background data, and observing the change in the prediction.



1. Select the target class – the plot will show explanations for this class.
2. Zoom in/out the plot.
3. Observe the prediction probability for a class and base value – an average probability in the dataset.
4. Plot which shows features that affect the prediction the most (features with longer tape length) and how they affect it. Red features increase the probability for a selected class while blue features decrease the probability. On the right from the tape, you can see the feature name and its value* for the selected instance. The length of the tape segment (and number on the tape) represent the SHAP value for feature contribution – it is how much the feature affects the probability for the selected class. Numbers in the gray boxes indicate the prediction probability for the selected class is (0.6) and the baseline probability (0.45) (the average probability in the data).

5. Get help, save the plot, make the report, or observe the size of input and output data.

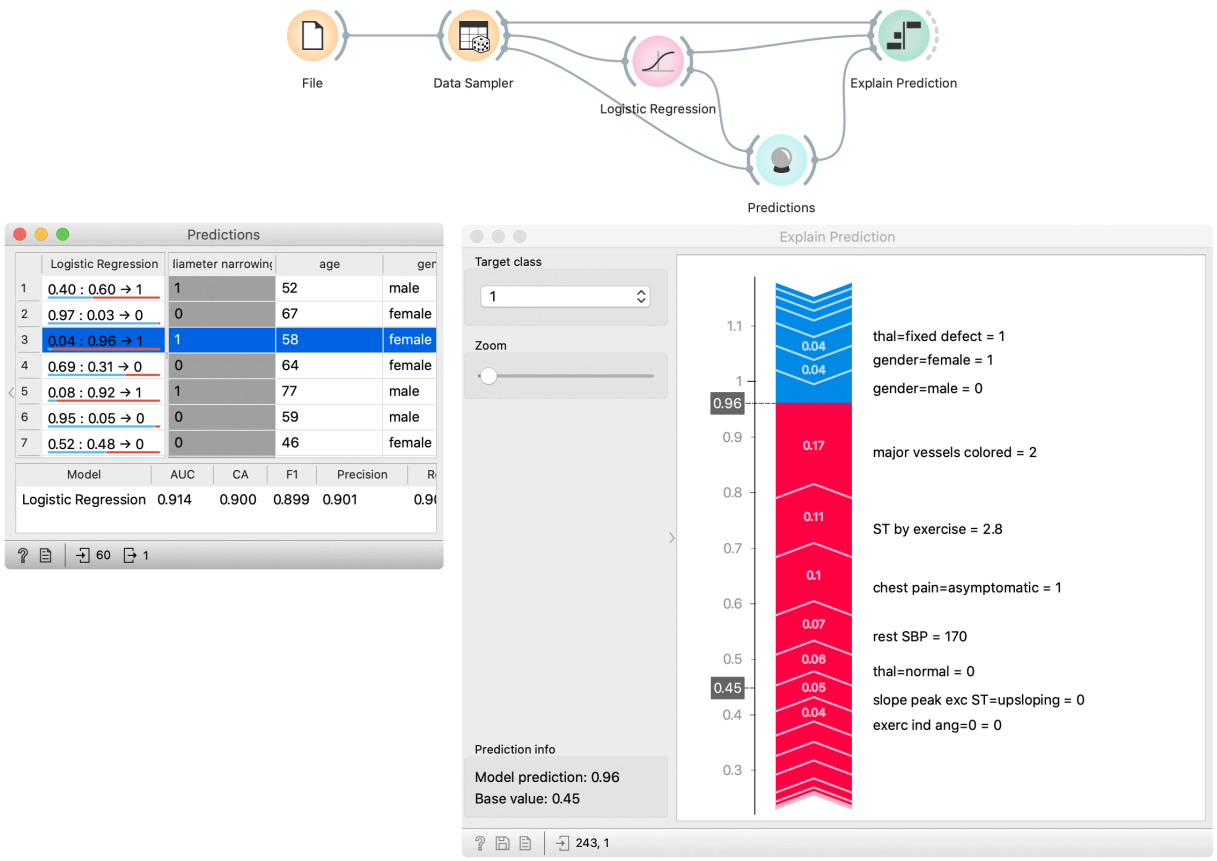
* Some models (including logistic regression) extend the categorical feature to more features with the technique named one-hot encoding. It means each value in the feature gets a new column which has value 0 (the instance does not have this feature value) or 1 (the instance has this feature value) for each instance. In those cases categorical features will be labeled with the format `feature-name=feature-value = 0/1` – e.g. `chest pain=asymptomatic = 1`. It means that the feature chest pain has value asymptomatic. Model, in this case, made more columns for feature chest pain, one of them was asymptomatic, and it was the case for the selected data instance.

1.3.1 Example

First, we open heart disease dataset in File widget. With the Data Sampler widget, we split the dataset on the training and test set. The training set is used to train the logistic regression model with the Logistic Regression widget. We compute predictions for the test part of the dataset (remaining data from Data Sampler) with the Predictions widget. In the Predictions widget (the left window) we select the data instance whose prediction we would like to explain – we select the third row in the data which is predicted to belong to class 1 (diagnosed heart disease).

Explain Prediction widget accept three inputs. First is the model from the Logistic Regression widget, background data from the Data Sampler (we usually use model's training data as background data), and the data instance whose prediction we want to explain with the Explain Prediction widget. In the widget we select class 1 as a target class, it means we are explaining what features and how they affect the prediction probability for the selected class 1. Numbers in the gray boxes in the plot indicate that the prediction probability for the selected class is 0.6 (border between red and yellow tape) and the baseline probability is 0.45 (the average probability in the data).

Features marked red on the tape push probabilities from the baseline probability toward probability 1.0 (prediction of the selected class), and blue features push against the prediction of the selected class. Numbers on the tape are SHAP values for each feature – this is how much the feature (and its value) changes the probability toward or against the selected class. We can see that the highest impact on the prediction has the feature *major vessels coloured* with the value 2 and *ST by exercise* with the value 2.8. Two important features that push against the prediction of class 1 are *gender=male* with value 0 (which means that the patient is not male) and *gender=female* with the value 1 (patient is female - actually another feature with the same meaning that



previous).

1.4 Explain Predictions

Explains which features contribute the most to the predictions for the selected instances based on the model and how they contribute.

Inputs

- Model: model whose predictions are explained by the widget
- Background data: dataset needed to compute explanations
- Data: dataset whose predictions are explained by the widget

Outputs

- Selected Data: instances selected from the plot
- Data: original dataset with an additional column showing whether the instance is selected
- Scores: SHAP values for each feature. Features that contribute more to prediction have a higher score deviation from 0.

Explain Predictions widget explains classification or regression model's predictions for the provided data instances.

CHAPTER

TWO

INDICES AND TABLES

- genindex
- modindex
- search