



Качество бинарной классификации

Качество бинарной классификации — в этом визуализаторе строятся диаграммы и таблицы с результатами проведенной классификации на основе логистической регрессии.

Визуализатор разделен на 3 области:

- Область настроек — в данной области происходит выбор и настройка диаграммы;
- Область диаграммы — отображает диаграмму, выбранную в области настроек;
- Область оценок классификации — содержит таблицы с оценками классификации.

Примечание: Область оценок классификации можно изменить в размерах, если потянуть за линию-разделитель левой кнопкой мыши. При двойном щелчке по линии-разделителю или одинарном посередине линии, область будет скрыта. Скрытие области возможно и для *Области настроек*.

Область настроек

Располагается в левой части визуализатора и содержит три группы параметров: *Выбор диаграммы*, *Множества* и *Порог отсечения*.

Выбор диаграммы

Группа содержит девять переключателей, каждый из которых вызывает определенную диаграмму в центре визуализатора:

- **ROC-кривая** — зависимость TPR от FPR .
- **PR-кривая** — зависимость TPR от PPV .
- **Базовые показатели** — на диаграмме строятся графики: TPR , TNR , FPR , FNR , и отображается порог отсечения.
- **Диаграмма точности** — на диаграмме строятся графики PPV , NPV , OPR , OCR , и отображается порог отсечения.
- **Диаграмма равновесия** — на диаграмме строятся графики PPV , TPR , и отображается порог отсечения.
- **% распознанных событий** — в зависимости от проставленного флажка *Кумулятивная* показывает:
 - ☒ — кумулятивный % событий от общего числа событий в зависимости от размера выборки.
 - ☐ — % событий, находящихся в диапазоне, от общего числа событий в зависимости от номера диапазона.

- **Диаграмма роста** — в зависимости от проставленного флажка *Кумулятивная* показывает:
 - ☒ — кумулятивное значение Lift (Роста) в зависимости от размера выборки.
 - ☐ — значения Lift (Роста) для диапазона в зависимости от номера диапазона.
- **Диаграмма отклика** — в зависимости от проставленного флажка *Кумулятивная* показывает:
 - ☒ — кумулятивный % событий в выборке в зависимости от размера выборки.
 - ☐ — % событий в диапазоне в зависимости от номера диапазона.
- **Диаграмма выигрыша** — показывает значение Gain (Выигрыш) в зависимости от размера выборки.

Примечание: посмотреть расшифровку сокращений можно в статье [Метрики качества моделей бинарной классификации](#).

Чекбокс *Кумулятивная* становится активным при выборе диаграмм: *% распознанных событий*, *Диаграмма роста* и *Диаграмма отклика*. По умолчанию флажок на чекбоксе проставлен, при снятии флажка становится доступным выбор диапазонов в выпадающем списке. Доступные значения:

- **10 диапазонов** — разделяет множество на 10 равных частей. Данный набор диапазонов используется по умолчанию.
- **20 диапазонов** — разделяет множество на 20 равных частей.
- **50 диапазонов** — разделяет множество на 50 равных частей.

Априорная доля событий

Возможно использование поправки на априорную вероятность, определенную в мастере настроек узла "Логистическая регрессия" до того, как будут получены результаты обработки данных с помощью классификатора. Для этого необходимо, чтобы флажок "Априорная доля событий" был проставлен. Использование поправки на априорную вероятность может изменить количество правильно классифицированных событий и не-событий, что приведёт к изменению значений в таблице с оценками классификации и в матрице ошибок. Если после проставления или снятия флажка "Априорная доля событий" никаких изменений не произошло, то необходимо проверить корректность настроек "Поправки на долю событий" в мастере настройки "Логистической регрессии". После внесения изменений в настройки узел необходимо переобучить.

Множества

Содержит два флажка:

- **Обучающее** — при проставленном флажке отображает на диаграмме график обучающего множества;
- **Тестовое** — при проставленном флажке отображает на диаграмме график тестового множества.

Возможен выбор только одного множества для диаграмм: *Базовые показатели*, *Диаграмма точности* и *Диаграмма равновесия*.

Порог отсечения

Представляет собой выпадающий список, содержащий следующие значения:

- **Из настроек узла** — используется порог, заданный узлом *Логистическая регрессия*.
- **Заданный** — дает возможность задать свой порог отсечения. Порог вводится, или в поле *Значение порога*, или задается перемещением ползунка под полем.
- **Баланс ($TPR = TNR$)** — задает порог отсечения, при котором TPR и TNR равны.
- **Максимум ($TPR + TNR$)** — задает порог отсечения, при котором сумма TPR и TNR имеет максимальное значение.
- **Точка равновесия ($TPR = PPV$)** — задает порог отсечения, при котором TPR и PPV равны.
- **Максимум общей точности** — задает порог отсечения, при котором значение OCR максимально.
- **Максимум $F1$ Score** — задает порог отсечения, при котором значение $F1$ Score максимально.
- **Коэффициент Мэтьюса (MCC)** — задает порог отсечения, при котором значение MCC максимально.

Область диаграммы

Область содержит следующие элементы:

- **Заголовок** — выводит название отображаемой диаграммы.
- **Событие** — выводит метку выходного поля и значение, являющееся событием.
- **Диаграмма** — выводит в центр диаграмму, которая всегда расположена в центре области, и имеет одинаковую высоту и ширину.
- **Легенда** — содержит названия серий, выводимых на диаграмме. При нажатии на серию в легенде можно скрыть или показать ее на диаграмме.

Примечание: в зависимости от свободного пространства области легенда будет располагаться под диаграммой или справа от нее.

Области оценок классификации

Располагается в правой стороне экрана и содержит таблицы, описывающие результаты классификации. Данные в этих таблицах обновляются с изменением *Порога отсечения*.

Оценки классификации

Таблица содержит столбцы:

- **Показатель** — отображает название рассчитанного элемента.
- **Множества** — группа из двух столбцов:
 - **Обучающее** — в столбце выводятся значения оценки показателей, рассчитанных для обучающего множества.

- **Тестовое** — в столбце выводятся значения оценки показателей, рассчитанных для тестового множества.

Таблица разделена на две части и имеет следующие строки:

- **Оценки классификации** — эта часть содержит оценки, которые рассчитываются для всей модели и не зависят от *Порога отсечения*:
 - **AUC ROC** — выводит площадь под ROC-кривой.
 - **AUC PR** — выводит площадь под PR-кривой. Определяется аналогично ROC-кривой, только по осям откладываются не FPR и TPR, а точность (Precision) и полнота (Recall).
 - **Коэффициент Джини** — коэффициент Джини.
 - **KS** — статистика критерия Колмогорова-Смирнова.
- **Порог отсечения** — в этой части расположены оценки, зависящие от *Порога отсечения*. Название порога выводится после двоеточия, например, *Порог отсечения: Заданный*:
 - **Значение** — выводит используемое значение *Порога отсечения*.
 - **TPR (Чувствительность)** — выводит значение *TPR* при используемом пороге.
 - **TNR (Специфичность)** — выводит значение *TNR* при используемом пороге.
 - **FPR (1-Специфичность)** — выводит значение *FPR* при используемом пороге.
 - **PPV** — выводит значение *PPV* при используемом пороге.
 - **F1 Score** — выводит значение *F1 Score* при используемом пороге.
 - **MCC** — выводит значение коэффициента корреляции Мэтьюса при используемом пороге.

Примечание: если множество данных отсутствует, то соответствующие ему ячейки таблиц будут серыми, и в них ничего не будет выводиться.

Матрицы ошибок

Содержит матрицы ошибок для обучающего и тестового множества и имеет следующий вид:

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Множество	P	N	
Событие	TP	FP	TP+FP
Не-событие	FN	TN	FN+TN

Примечание: над верхним правым углом таблицы расположен переключатель  **1 % / 1 %**
Абсолютные или относительные значения, при переключении которого данные в таблице будут представлены в процентах или в количестве записей, попавших в категорию.

Распознано

В данной таблице показано отношение правильно распознанных событий к общему количеству событий во множествах. Таблица состоит из двух полей:

- **Множество** — в этом столбце отображается тип множества.
- **Распознано** — количество распознанных событий от общего числа событий. Определяется по формуле $\frac{TP+TN}{TP+TN+FP+FN}$

Примечание: над верхним правым углом таблицы расположен переключатель 1 % / 1 %

Абсолютные или относительные значения, при переключении которого данные в таблице будут представлены в процентах или в количестве записей, попавших в категорию.