Двумерный анализ

Loginom Хакатон 2020. Секции 1, 3

Введение

В прошлом выпуске мы показали, как строить модель предрасположенности клиента совершить определенное действие (на примере повторного визита).

Но мы подробно не остановились на важной теме – **двумерном анализе**, а по сути, **оптимальном квантовании**, которое обязательно перед применением логистической регрессии и реализовано в компоненте **Конечные классы** платформы Loginom.

Двумерный анализ

В предсказательной аналитике уделяют большое внимание анализу связей между переменными. Самым простым и типичным является случай анализа взаимосвязи (сопряженности) двух переменных, так называемый **ДВУМЕРНЫМ АНАЛИЗОМ.**

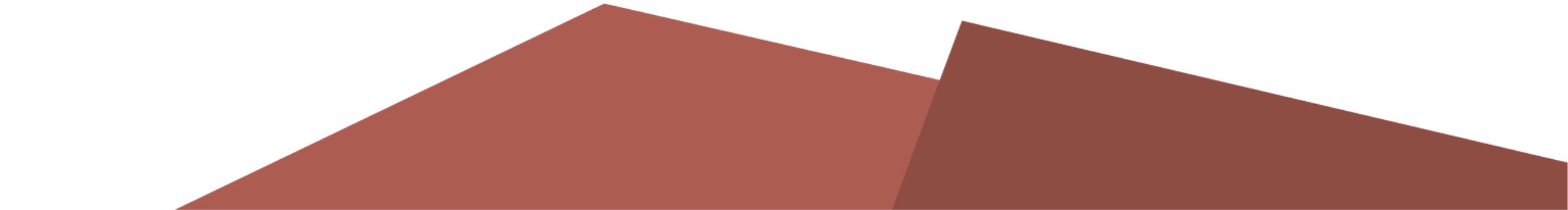
Формирование конечных классов (англ: Fine&Coarse Classing) как процесс оптимального квантования или биннинга (для непрерывных переменных) и снижения разнообразия категорий путем их объединения (для категориальных переменных) позволяет построить устойчивые модели оценки вероятности наступления события даже при использовании методов, линейных по своей сути (например, **логистическая регрессия**).

Двумерный анализ

Кроме того, процедуры формирования конечных классов позволяют получить два важных результата:

- оценить предсказательную силу отдельной переменной (анализ IV);
- оценить характер связи значения переменной с бинарной выходной переменной (анализ WoE).

Математика WoE-анализа



Первый шаг

Первый шаг в анализе взаимоотношений двух переменных является их перекрестная классификация, или построение таблицы сопряженности. Пример: реакция клиентов банка на предложение открыть кредитную карту (см. исходные данные в файле **Отклики.xls**).

Отклик	Уровень образования		
	низкий	высокий	**Всего**
Нет (0)	50	310	**360**
Да (1)	95	45	**140**
Всего	**145**	**355**	**500**

Второй шаг

Методика анализа двумерной таблицы:

- Пересчитать абсолютные частоты в проценты;
- Сравнить процентные показатели, полученные для подгрупп с разным уровнем независимой переменной, каждый раз внутри одной категории зависимой переменной.

Отклик	Уровень образования		
	низкий	высокий	**Всего**
Нет (0)	34,5%	87,3%	**72%**
Да (1)	65,5%	12,7%	**28%**
Всего	**100%**	**100%**	**100%**

WoE-анализ

WoE-анализ (Weight Of Evidence) или **совокупность доказательств** - статистический метод оценки влияния тех или иных факторов на справедливость некоторой гипотезы.

$$WoE_i = ln \frac{N_i/N}{P_i/P}$$

где i – индекс признака, для которого вычисляется показатель WoE, N_i – число не-событий в группе, N – общее число не-событий, P_i – число событий в группе, P – общее число событий.

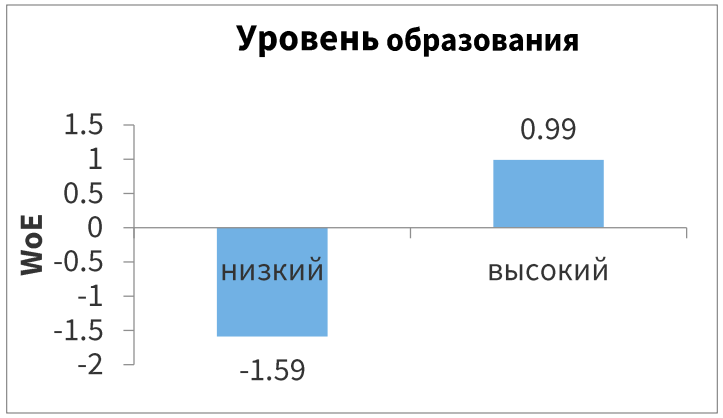
Гипотеза независимого поведения признаков: пропорция события и не-события в анализируемой подгруппе должна сохраняться такой же, как и для всей выборки в целом.

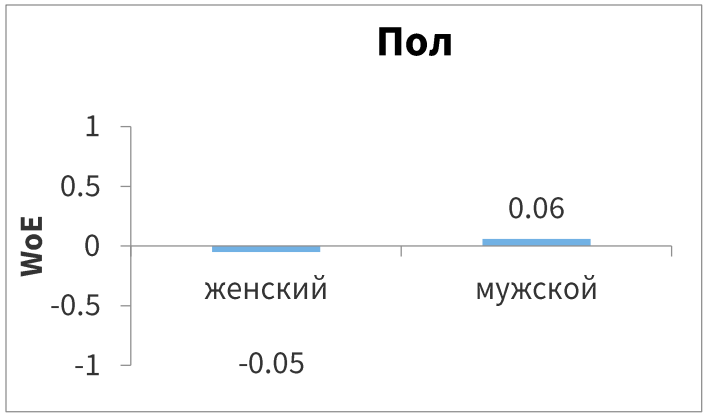
Интерпретация WoE-индексов

Если значение категории совпадает с событием большее число раз, чем с не-событием, то согласно формуле, под знаком логарифма будет значение меньше 1, что делает его отрицательным.

WoE < 0 указывает на большую вероятность появления события, а WoE > 0 - не-события.

Индекс WoE есть количественная мера предсказательной силы отдельной категории внутри переменной.





Информационный индекс

WoE является промежуточным элементом для вычисления агрегированной величины, называемой информационным **индексом IV** (Information Value):

$$IV = \sum_{i=1}^{K} \left\{ \left(\frac{N_i}{N} - \frac{P_i}{P} \right) \cdot WoE_i \right\},$$

Информационный индекс отвечает за **предсказательную способность** (силу) всей переменной.

$$IV(\text{Уровень образования}) = \left(\frac{50}{360} - \frac{95}{140} \right) \cdot (-1{,}59) + \left(\frac{310}{360} - \frac{45}{140} \right) \cdot (0{,}99)$$

$$= 1{,}39.$$

Значимость

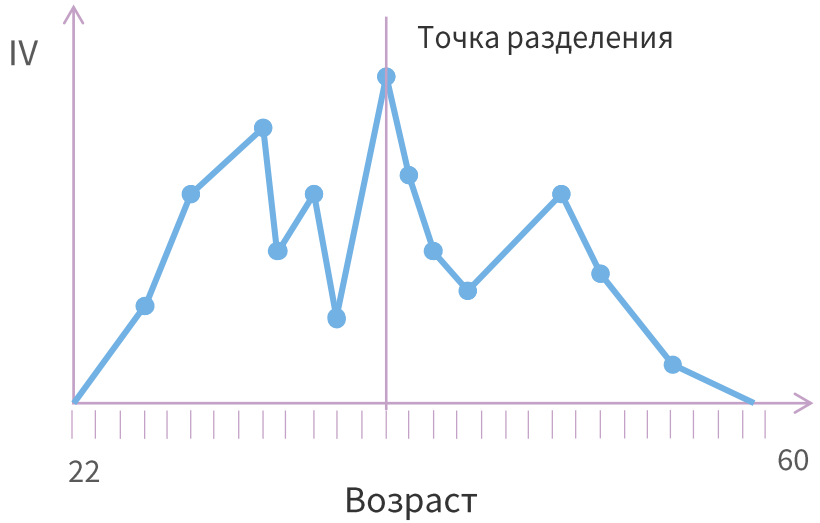
С помощью значения IV можно оценить значимость (предсказательную силу) переменной с бинарной выходной переменной:

- IV < 0,02 - отсутствует;
- 0,02 <= IV < 0,1 - низкая;
- 0,1 <= IV < 0,3 - средняя;
- IV >= 0,3 - высокая.

На практике предсказание отклика по одной, двум значимым переменным, сомнительна. При решении реальных задач приходится иметь дело с большим числом категориальных переменных, имеющих большее число уникальных значений, а также с непрерывными переменными.

Обработка непрерывной переменной

Идеи конечных классов для обработки непрерывной переменной те же самые. На первом шаге все уникальные значения переменной сортируются по возрастанию, и для каждого из них рассчитывается WoE-значение, а для всей переменной – индекс IV (начальные классы).
На следующем шаге производится итерационное формирование конечных классов путем присоединения соседних начальных классов последовательно начиная с первого с новым расчетом IV.



По сути, на каждой итерации выполняется поиск оптимальной точки разделения, на основе которой будут созданы два конечных класса. Это точка, в которой значение IV будет максимальным.
Процесс выполняется итерационно до тех пор, пока не будут достигнуты ограничения на максимально допустимое число конечных классов и минимальную долю примеров в классе.

Применение

- Эффективное использование логистической регрессии.
- Переменные с отсутствующей значимостью можно не подавать на вход логистической регрессии.
- Из нескольких коррелирующих между собой переменных можно оставить одну с максимальным значением IV.

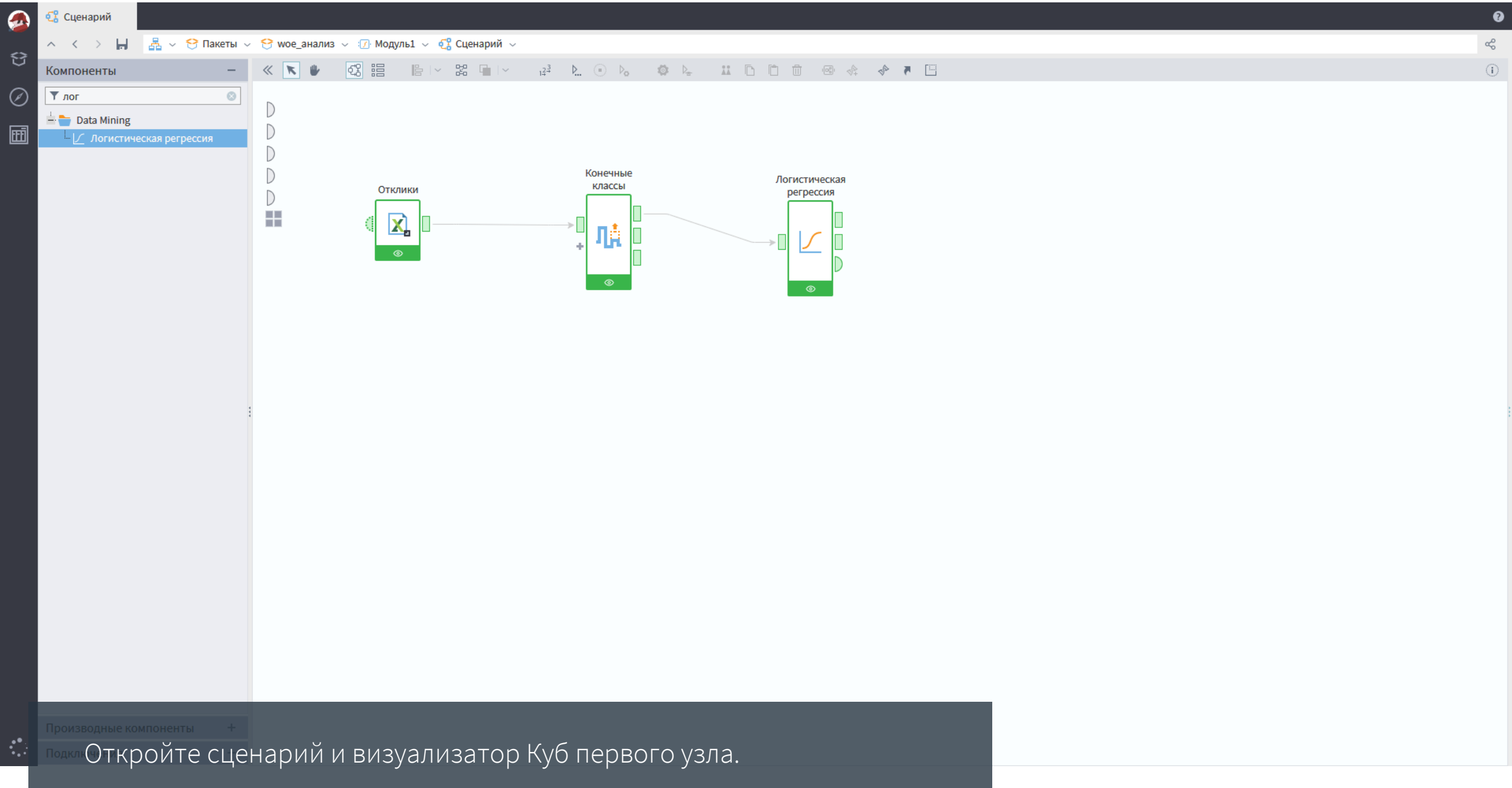
Пример

Пример расчетов можно посмотреть в файле **woe-анализ.xls**.

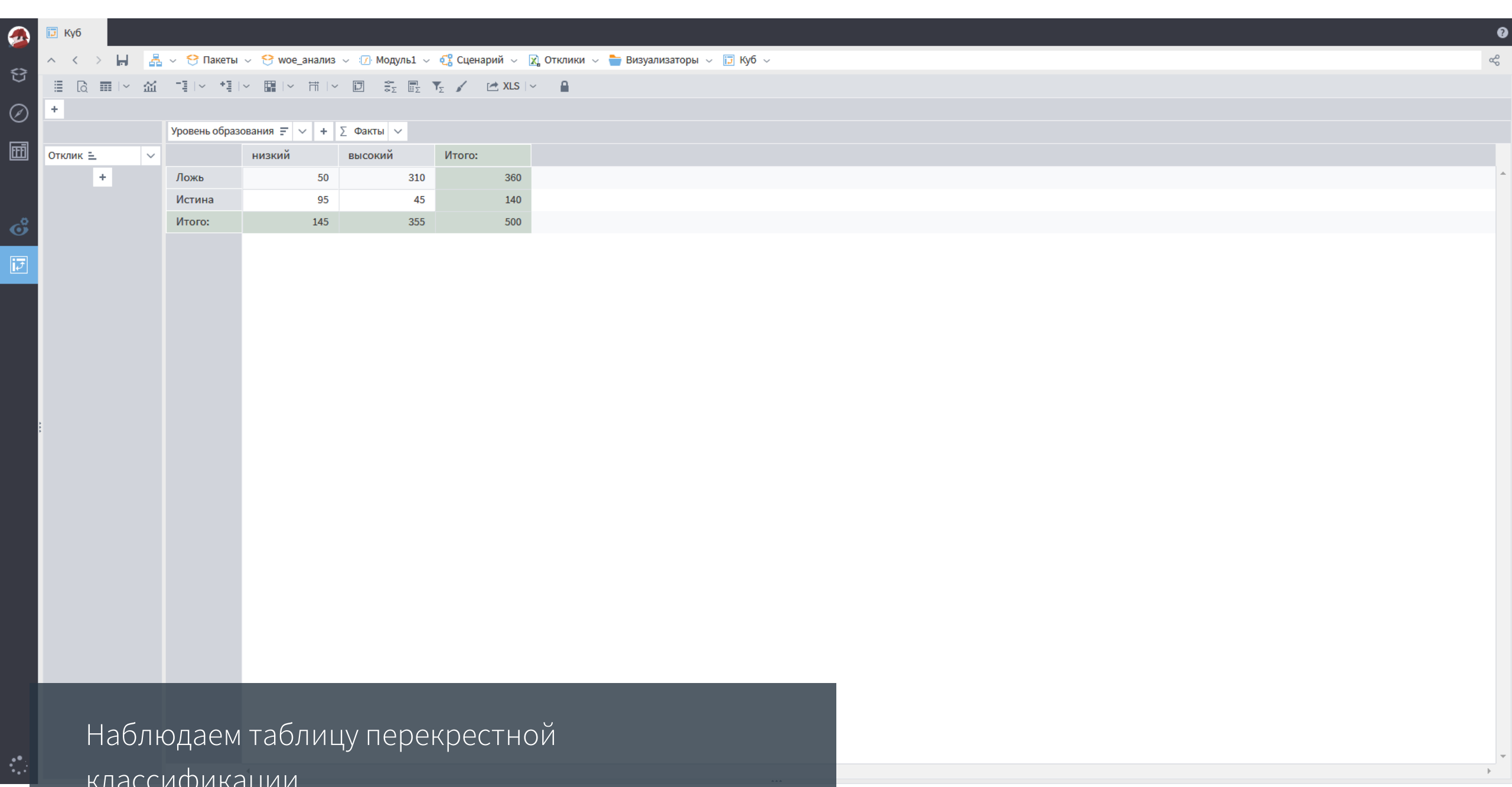
Работа в Loginom

Файлы

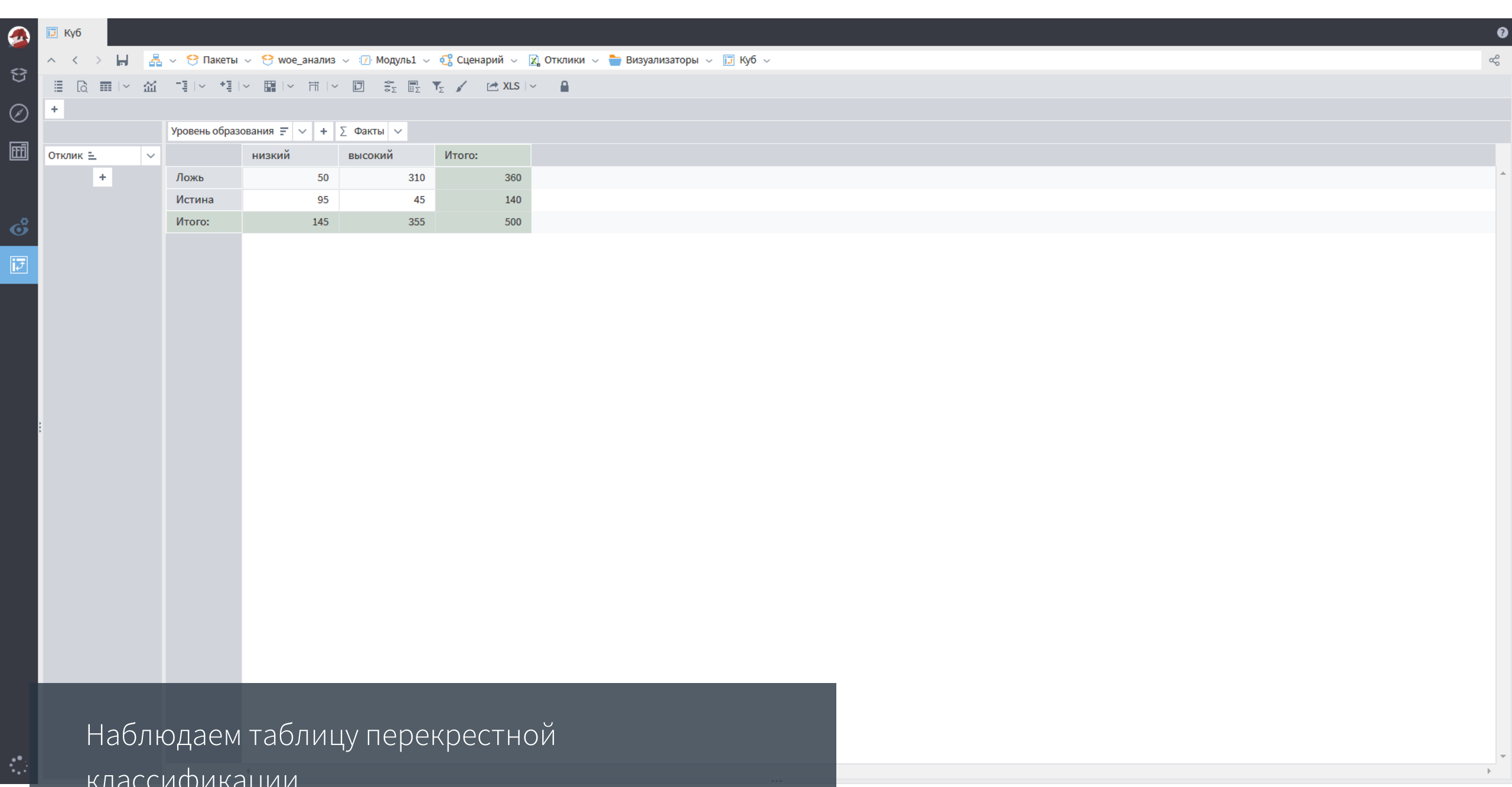
Для дальнейшего понимания откройте сценарий **woe-анализ.lgp**.



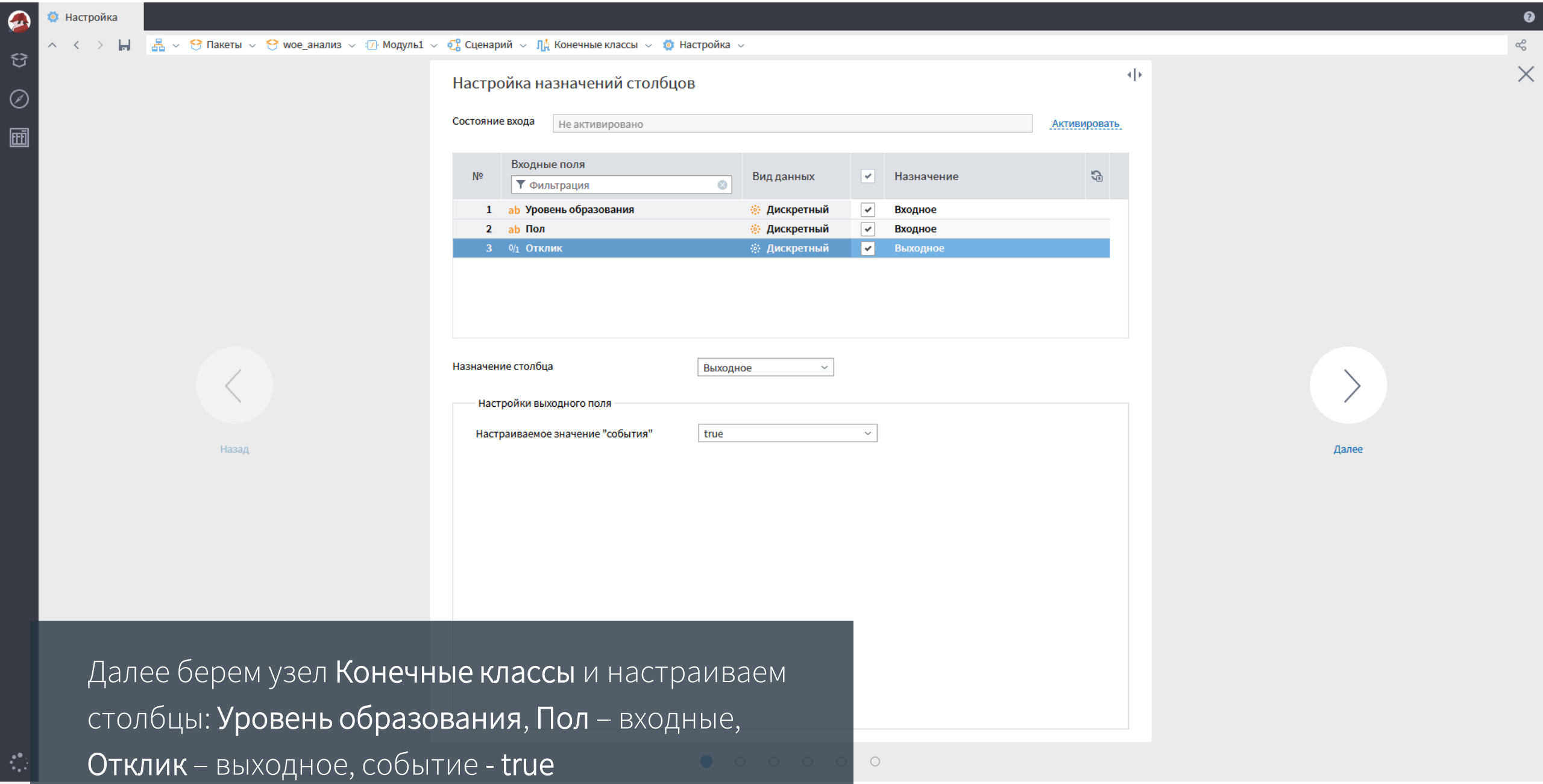
Откройте сценарий и визуализатор Куб первого узла.



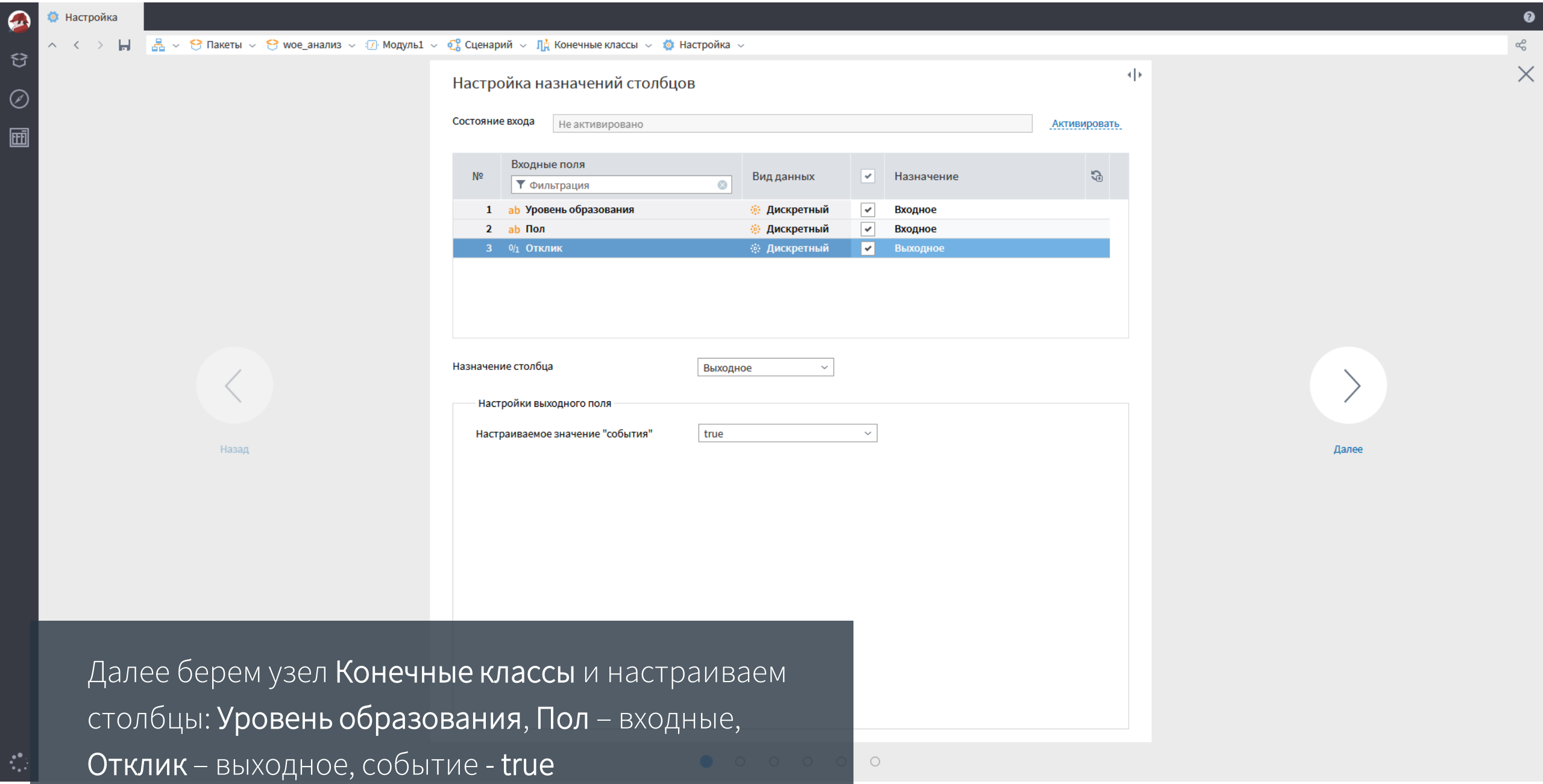
Наблюдаем таблицу перекрестной классификации.



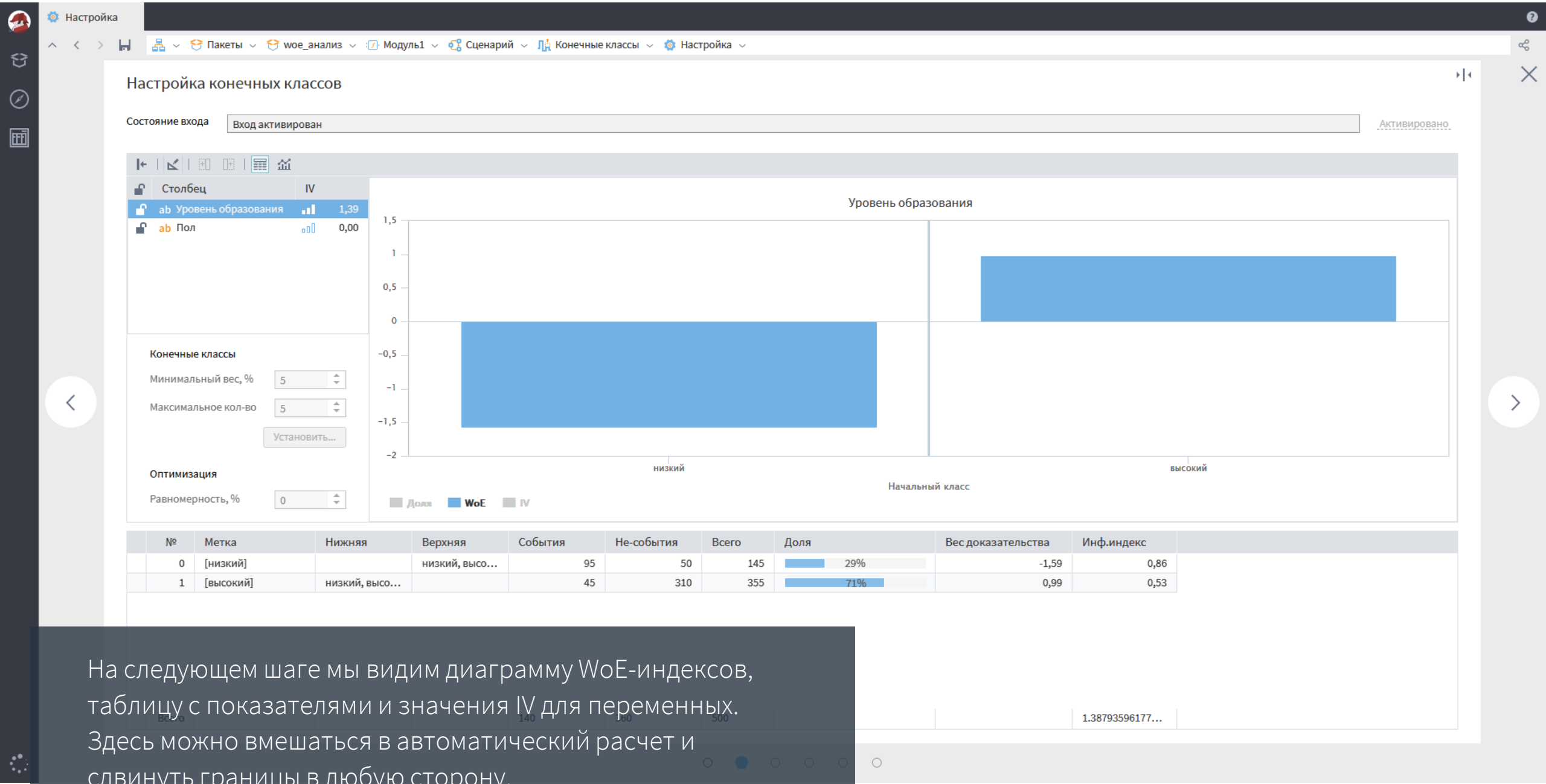
Наблюдаем таблицу перекрестной классификации.



Далее берем узел **Конечные классы** и настраиваем столбцы: **Уровень образования**, **Пол** – входные, **Отклик** – выходное, событие - **true**



Далее берем узел **Конечные классы** и настраиваем столбцы: **Уровень образования**, **Пол** – входные, **Отклик** – выходное, событие - **true**



На следующем шаге мы видим диаграмму WoE-индексов, таблицу с показателями и значения IV для переменных. Здесь можно вмешаться в автоматический расчет и сдвинуть границы в любую сторону.