Моделирование предрасположенности

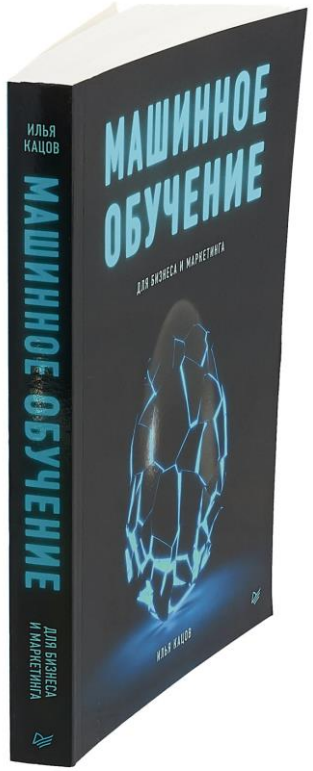
Loginom Хакатон 2020. Секции 1, 3

Введение

В секциях 1 и 3 **Loginom Хакатон 2020** необходимо построить предсказательную модель для оценки вероятности некоторого события у клиента (повторного визита или отмены заказа).

В данном руководстве мы покажем, как это сделать.

Для восприятия дальнейшего материала ознакомьтесь, пожалуйста, с **п. 3.5.4.1. Моделирование методом аналогии** книги из списка рекомендуемых – «Машинное обучение для бизнеса и маркетинга».



Постановка задачи

Мы будем решать задачу получения вероятности повторного визита клиента розничной сети методом аналогий на основе алгоритма логистической регрессии. В комплекте идет небольшой демо-набор **demo.lgd** с транзакциями.

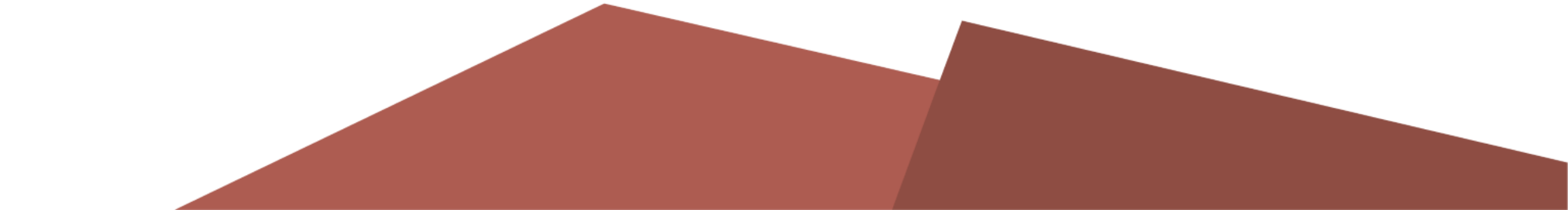
Вопросы, которые мы опустим и оставим на самостоятельное изучение:

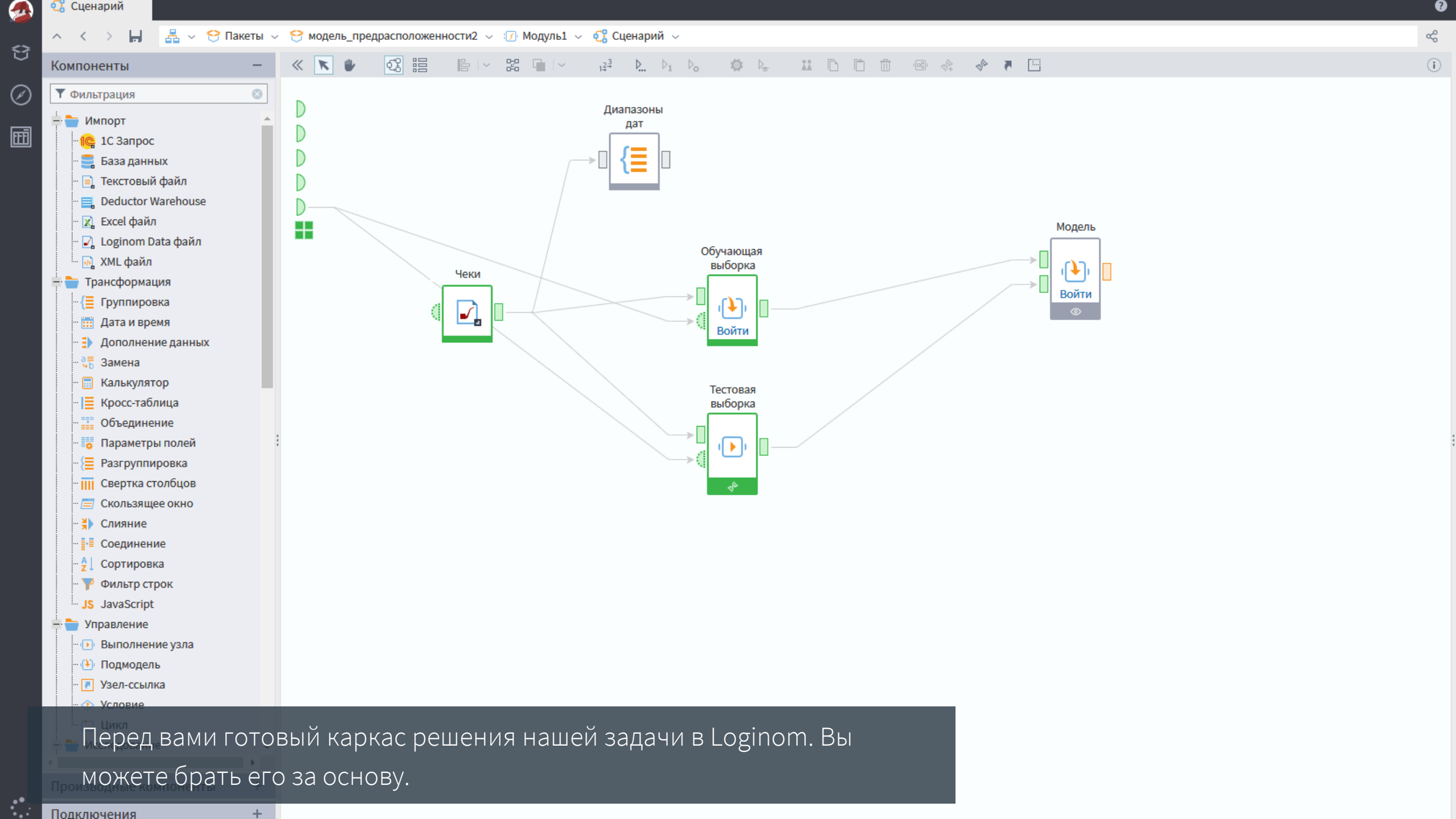
- Логрегрессия, алгоритмы Ridge и LASSO.
- Оценка качества алгоритмов машинного обучения и метрики ROC-AUC.

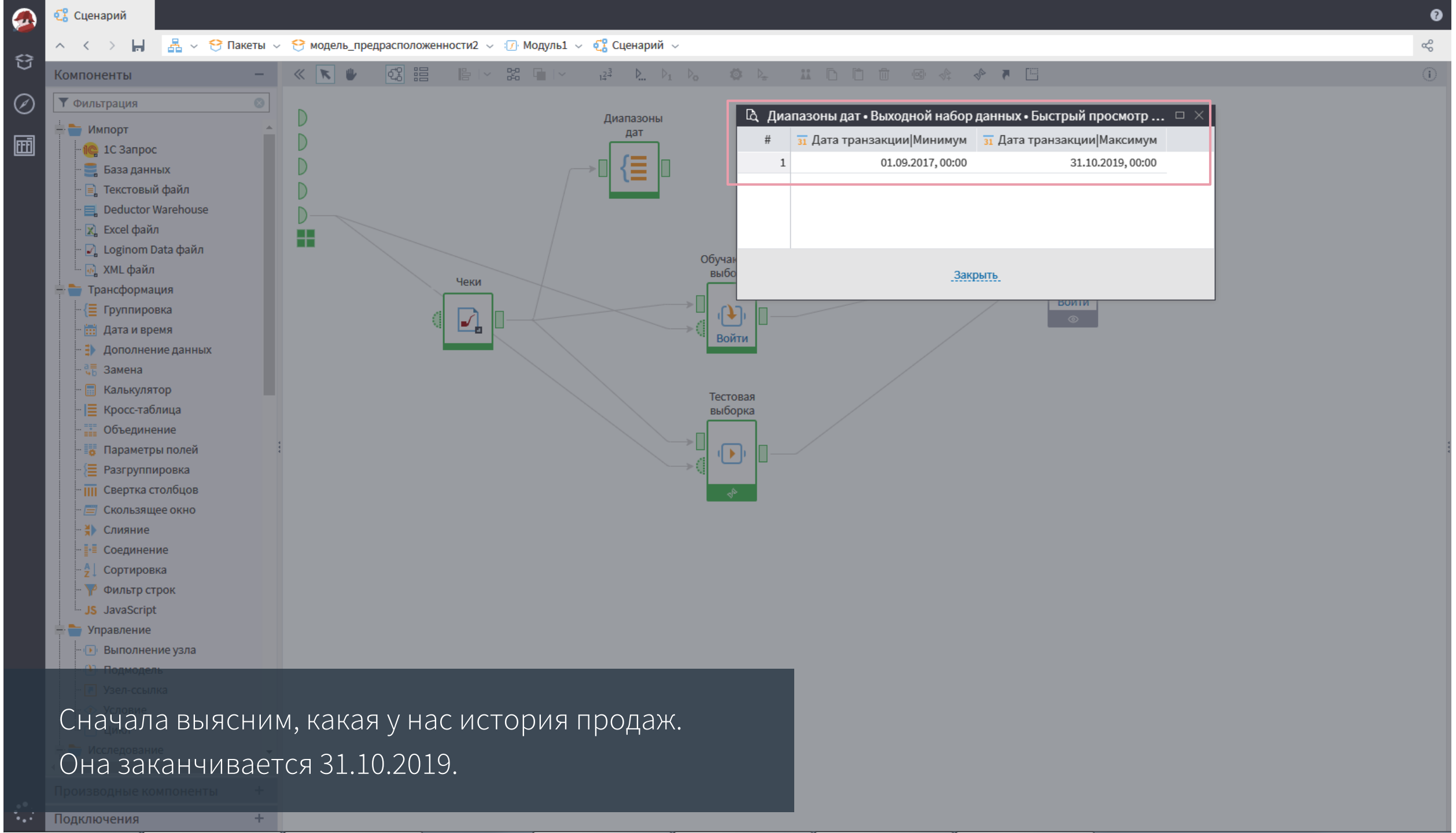
Сценарий

Для дальнейшего понимания откройте в Loginom сценарий **модель_предрасположенности**.

Работа в Loginom




Сценарий
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Deductor Warehouse
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
JavaScript
Управление
Выполнение узла
Подмодель
Узел-ссылка
Условие
Цикл
Производные компоненты
Подключения
Диапазоны дат
Чеки
Обучающая выборка
Войти
Тестовая выборка
Модель
Войти
Перед вами готовый каркас решения нашей задачи в Loginom. Вы можете брать его за основу.


Сценарий
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Deductor Warehouse
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
JavaScript
Управление
Выполнение узла
Подмодель
Узел-ссылка
Условие
Исследование
Производные компоненты
Подключения
Диапазоны дат
Чеки
Войти
Тестовая выборка
Диапазоны дат • Выходной набор данных • Быстрый просмотр ...
| Дата транзакции|Минимум | Дата транзакции|Максимум
1 | 01.09.2017, 00:00 | 31.10.2019, 00:00
Закрыть
Сначала выясним, какая у нас история продаж.
Она заканчивается 31.10.2019.

Временные периоды моделирования

Пусть мы решили прогнозировать вероятность повторного визита клиента в следующие **два месяца**. Сам выбор длины временного окна неоднозначен, и во многом связан с бизнес-процессами маркетологов компании, но для данной сети его нет смысла делать менее месяца.

Промежуточного периода у нас не будет (кстати, почему? Подумайте над этим).

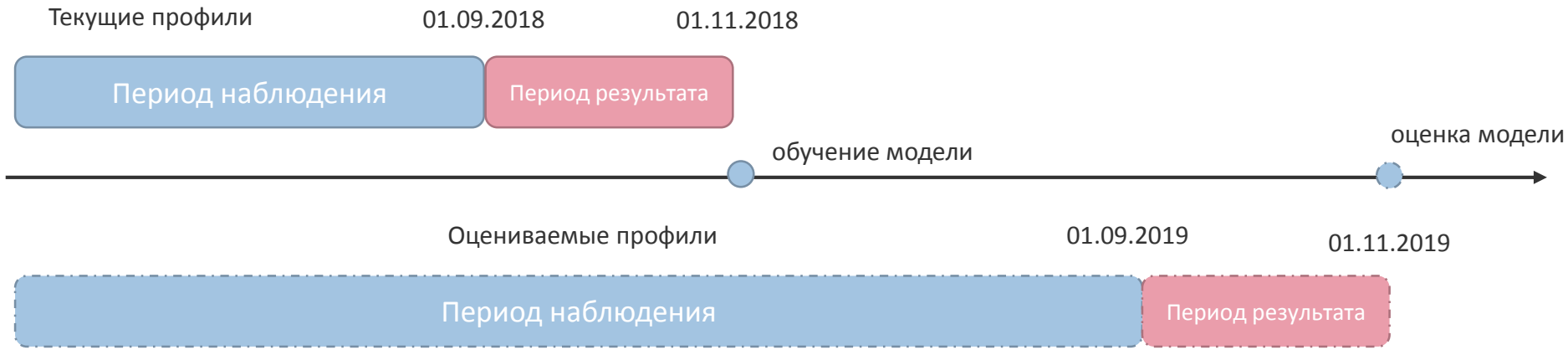
Тогда временной период, на котором мы будем обучать модель (назовем его «текущим»), будет оканчиваться 01.09.2018 (не включая эту дату).

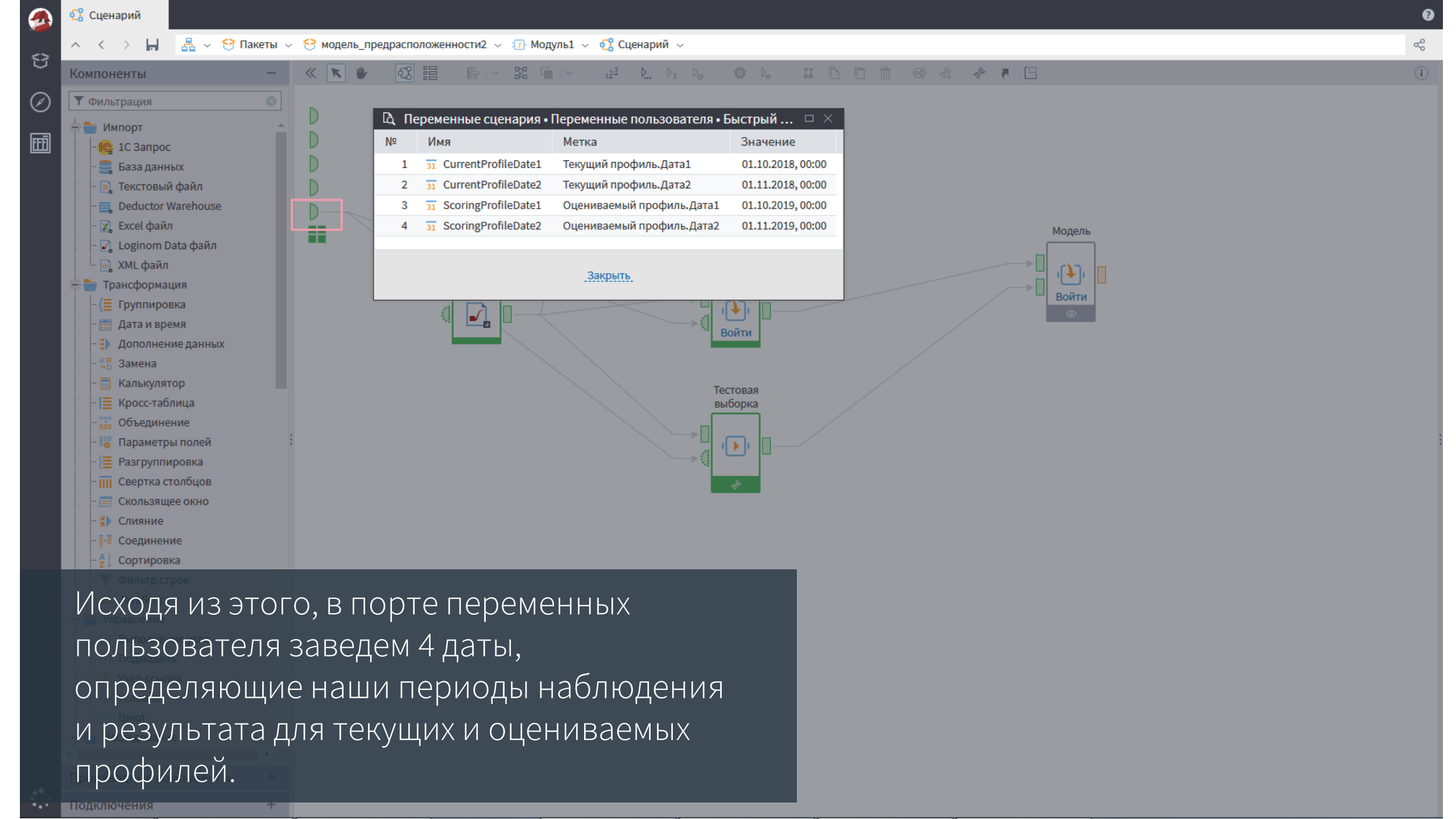
Метки событий (в данном случае факт хотя бы одного визита) будем фиксировать в следующие два месяца начиная с этой даты, то есть до 01.11.2018 (не включая эту дату).

Временные периоды моделирования

Соответственно, временной период, на котором мы будем оценивать клиентов (назовем его «оцениваемым»), а также тестировать модель, будет оканчиваться 01.09.2019. Метки событий будем аналогично фиксировать в следующие два месяца начиная с этой даты, то есть до 01.11.2019.

Рисунок показывает это схематично.



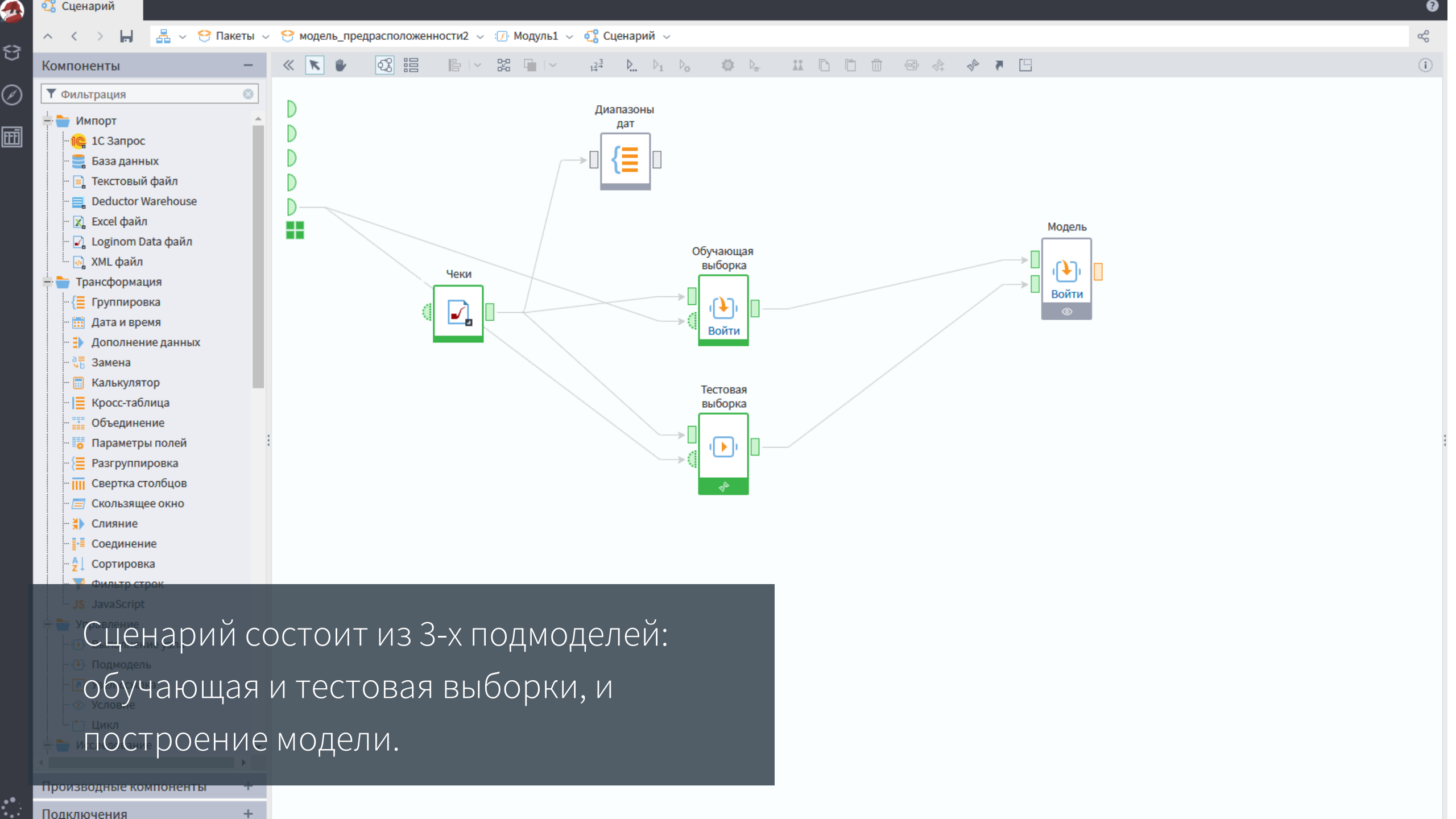

Сценарий
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Deductor Warehouse
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
Подключения
Переменные сценария • Переменные пользователя • Быстрый ...
№ Имя Метка Значение
1 CurrentProfileDate1 Текущий профиль.Дата1 01.10.2018, 00:00
2 CurrentProfileDate2 Текущий профиль.Дата2 01.11.2018, 00:00
3 ScoringProfileDate1 Оцениваемый профиль.Дата1 01.10.2019, 00:00
4 ScoringProfileDate2 Оцениваемый профиль.Дата2 01.11.2019, 00:00
Закрыть
Войти
Модель
Войти
Тестовая выборка
Исходя из этого, в порте переменных пользователя заведем 4 даты, определяющие наши периоды наблюдения и результата для текущих и оцениваемых профилей.

Профили клиентов

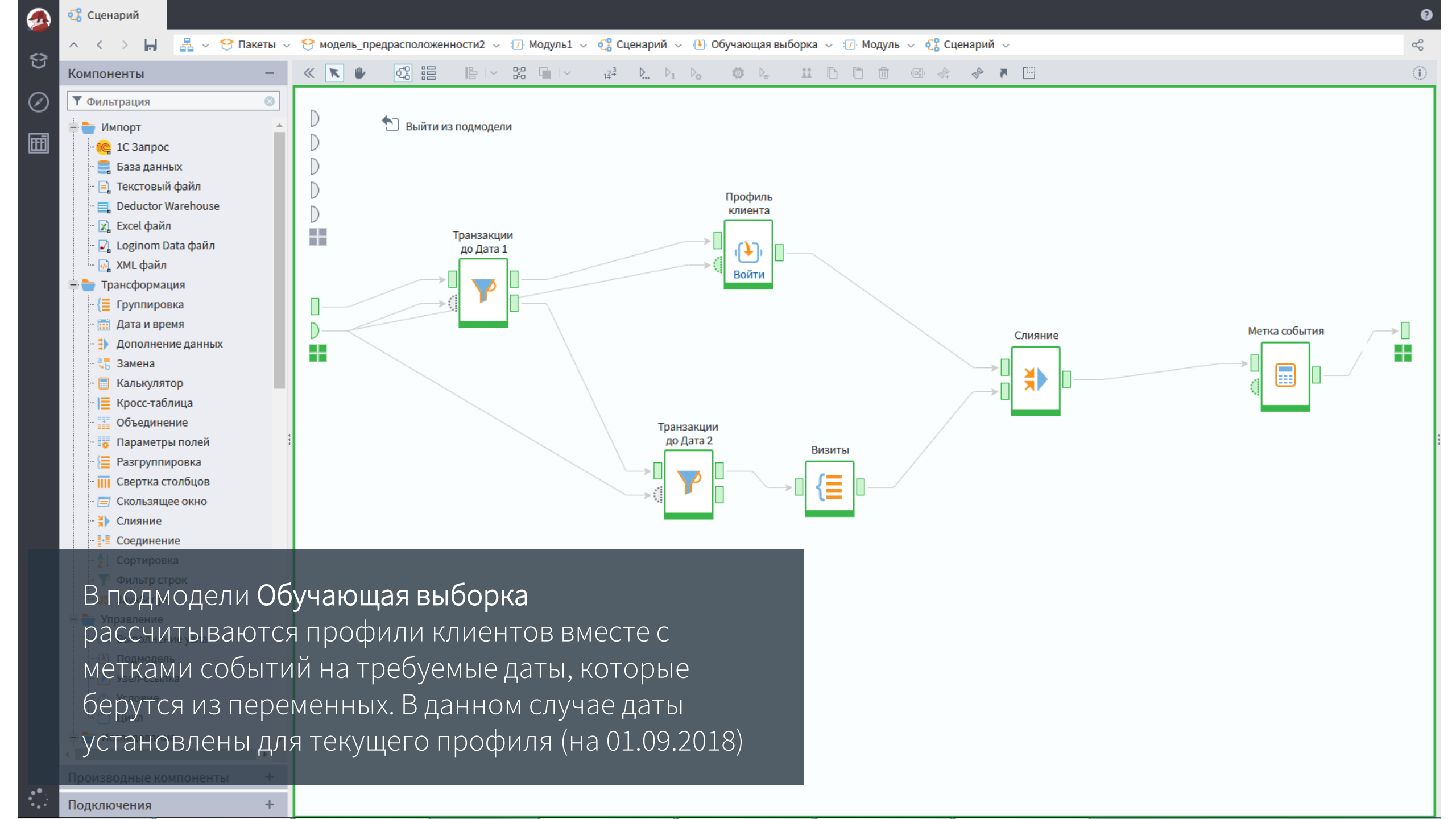
Мы построим модель на трех характеристиках клиентов, которые могут влиять на повторный визит:

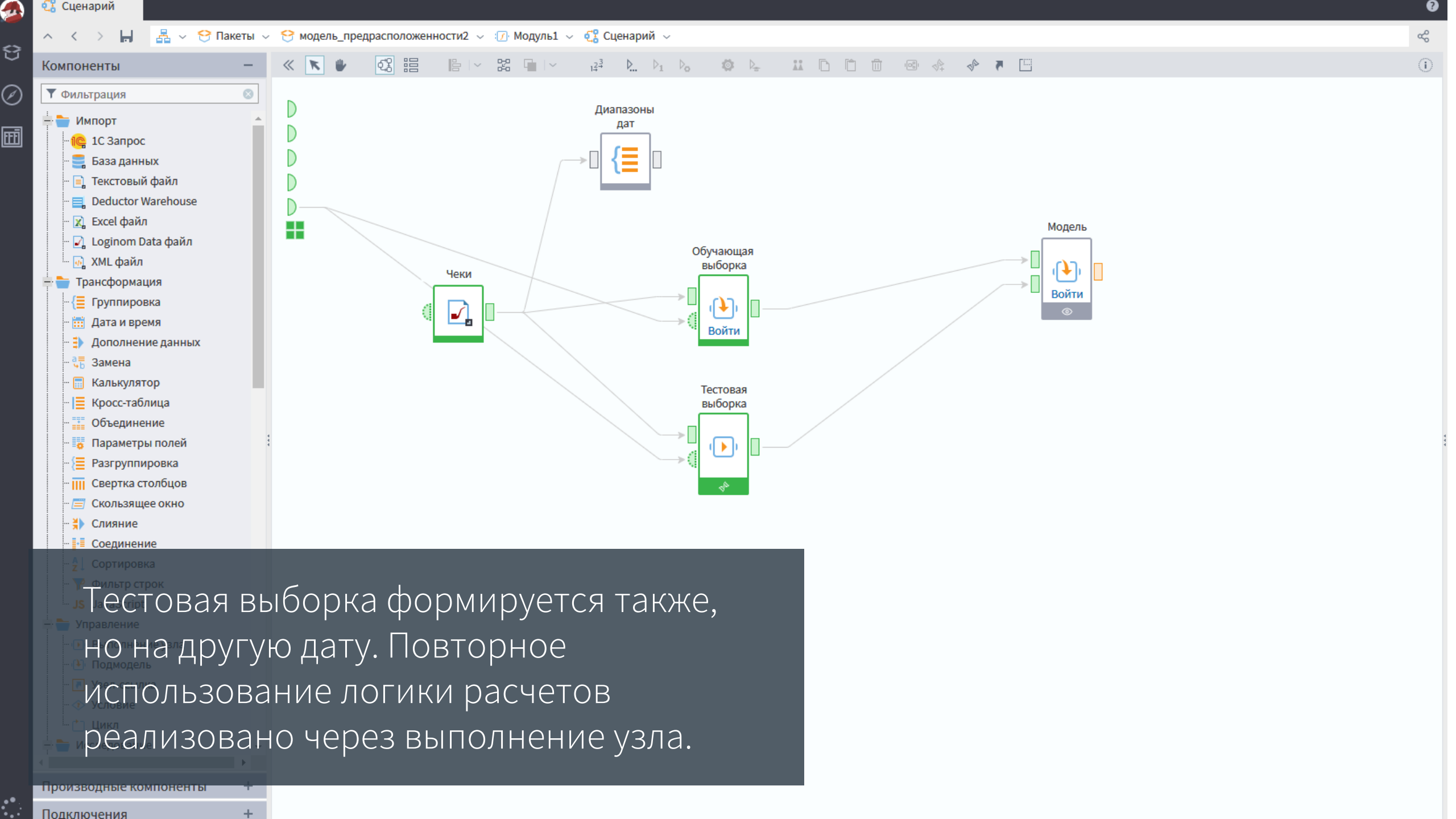
- Количество визитов – под визитом понимаем все чеки клиента в рамках одних суток;
- Время сна, дни – число дней с момента последнего визита;
- Разнообразие позиций – количество уникальных товаров, купленных клиентом на дату расчета.

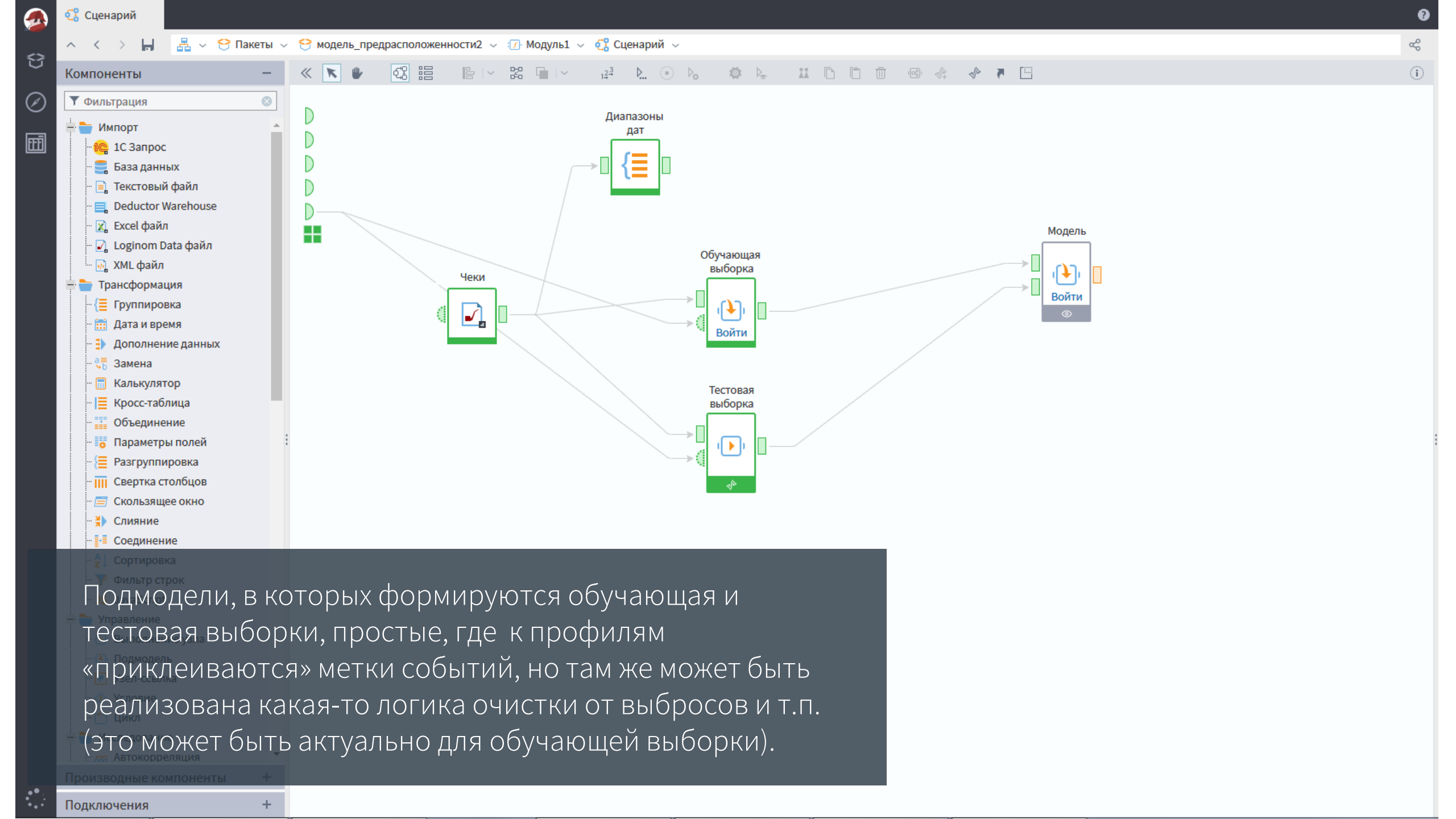
В реальности характеристик должно быть больше, десятки, но здесь мы подсказок давать не будем. Максимально используйте ту информацию, которая есть в «сырых» транзакциях.



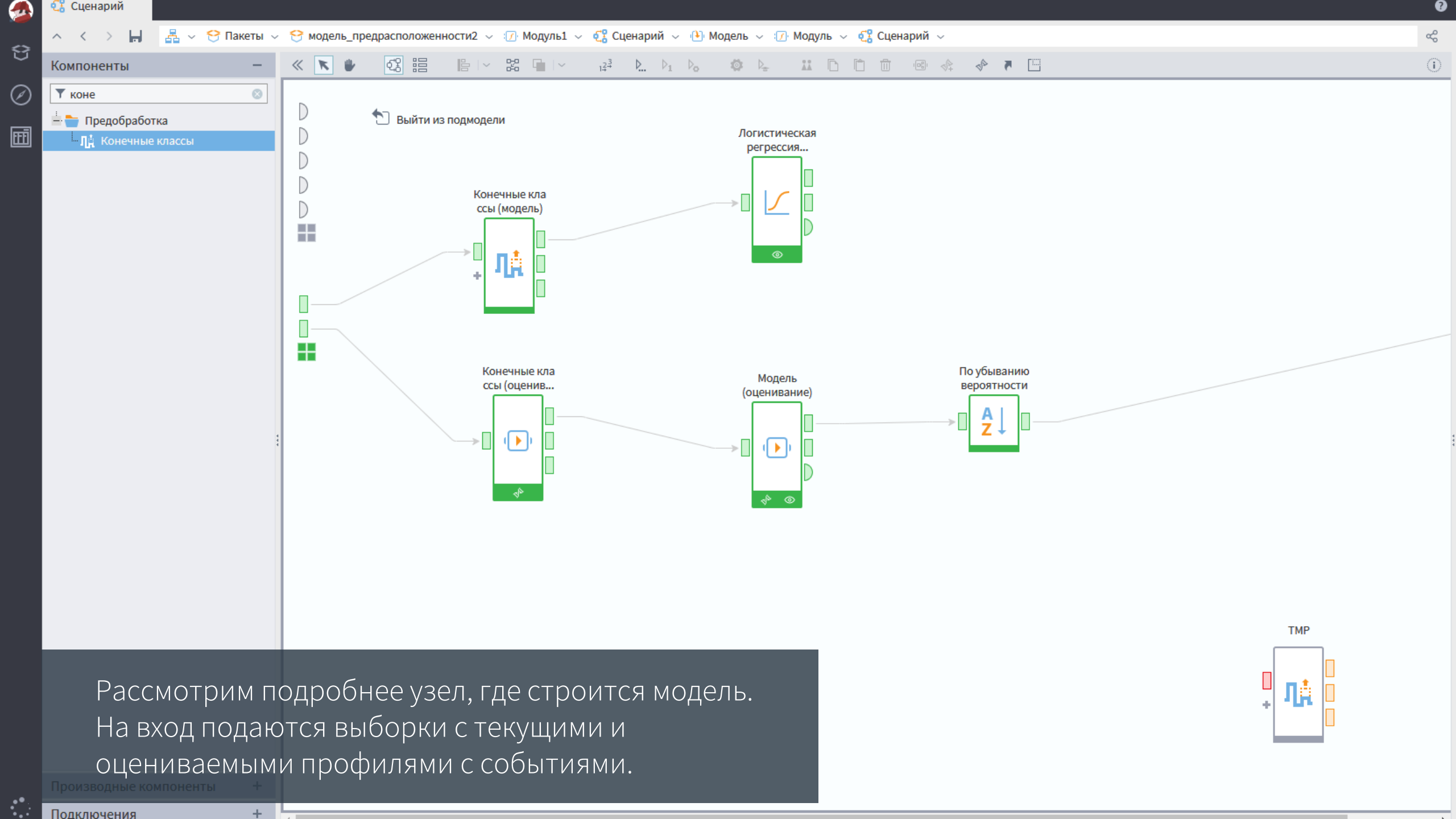
Сценарий состоит из 3-х подмоделей: обучающая и тестовая выборки, и построение модели.


Сценарий
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Обучающая выборка
Модуль
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Deductor Warehouse
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
Управление
Подмодель
Производные компоненты
Подключения
Выйти из подмодели
Транзакции до Дата 1
Профиль клиента
Войти
Слияние
Метка события
Транзакции до Дата 2
Визиты
В подмодели Обучающая выборка рассчитываются профили клиентов вместе с метками событий на требуемые даты, которые берутся из переменных. В данном случае даты установлены для текущего профиля (на 01.09.2018)

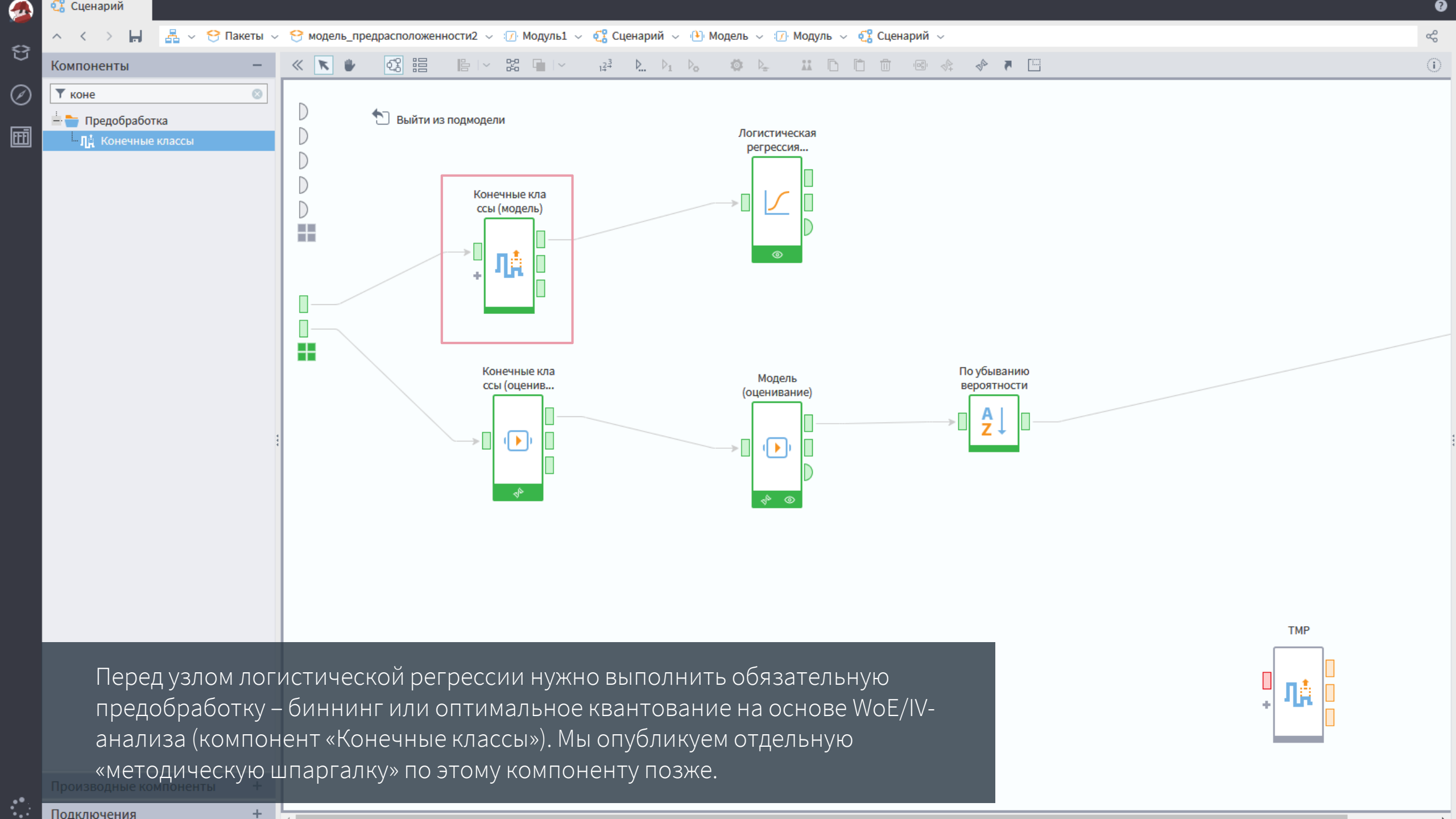

Сценарий
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Компоненты
Фильтрация
Импорт
1С Запрос
База данных
Текстовый файл
Deductor Warehouse
Excel файл
Loginom Data файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
Управление
Подмодель
Условие
Цикл
Производные компоненты
Подключения
Диапазоны дат
Чеки
Обучающая выборка
Войти
Тестовая выборка
Модель
Войти
Тестовая выборка формируется также, но на другую дату. Повторное использование логики расчетов реализовано через выполнение узла.



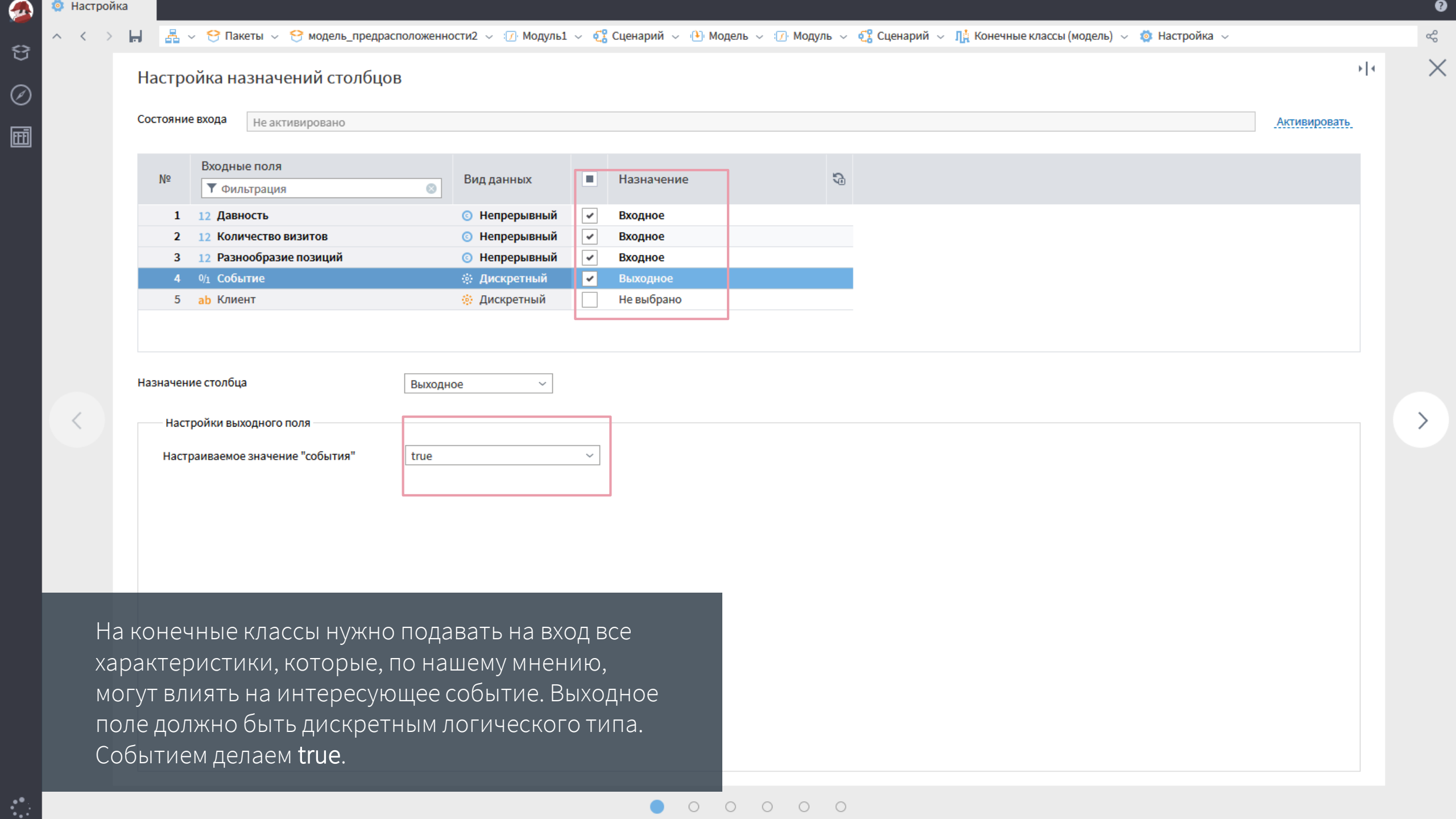
Подмодели, в которых формируются обучающая и тестовая выборки, простые, где к профилям «приклеиваются» метки событий, но там же может быть реализована какая-то логика очистки от выбросов и т.п. (это может быть актуально для обучающей выборки).

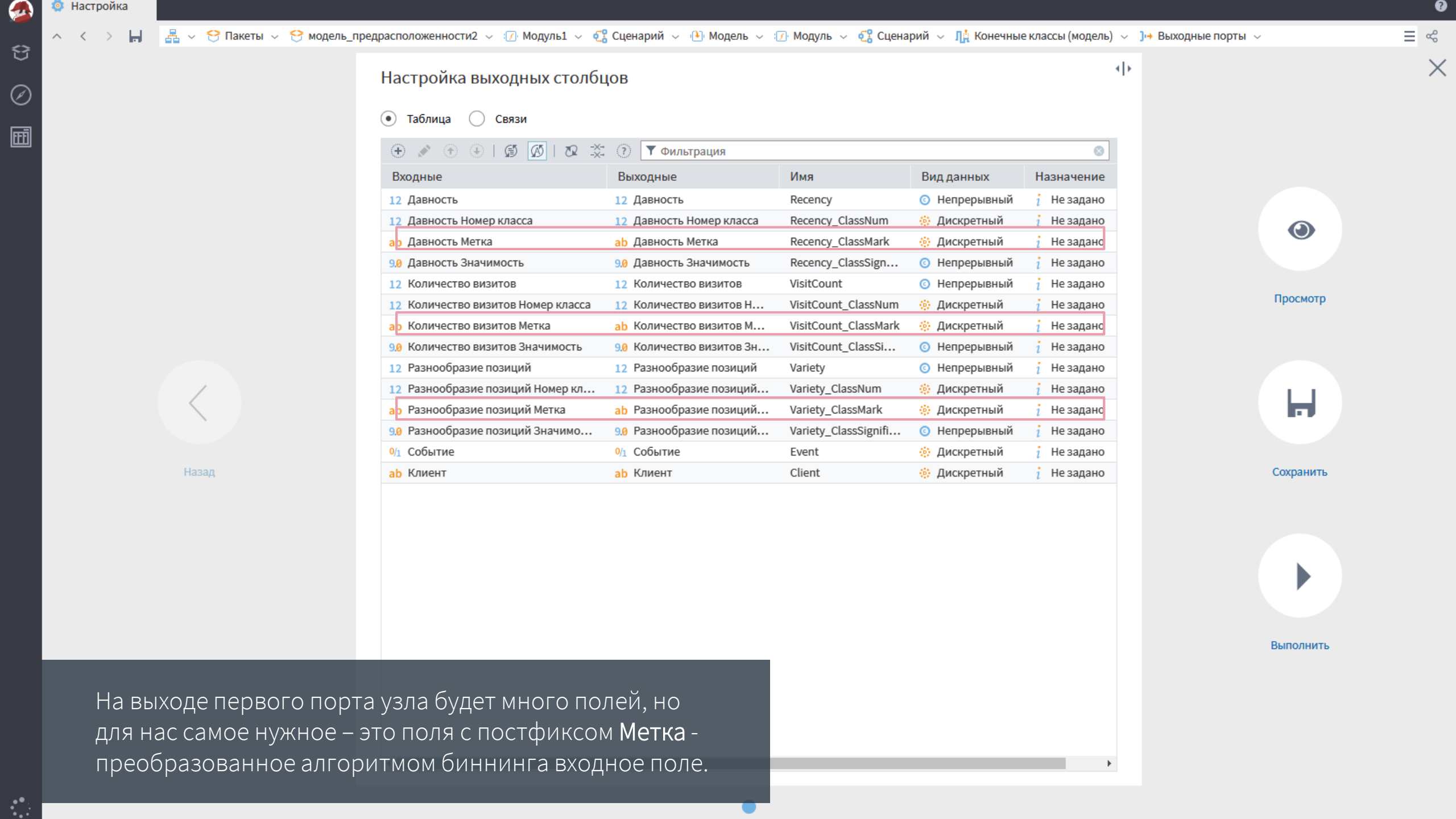


Рассмотрим подробнее узел, где строится модель.
На вход подаются выборки с текущими и оцениваемыми профилями с событиями.

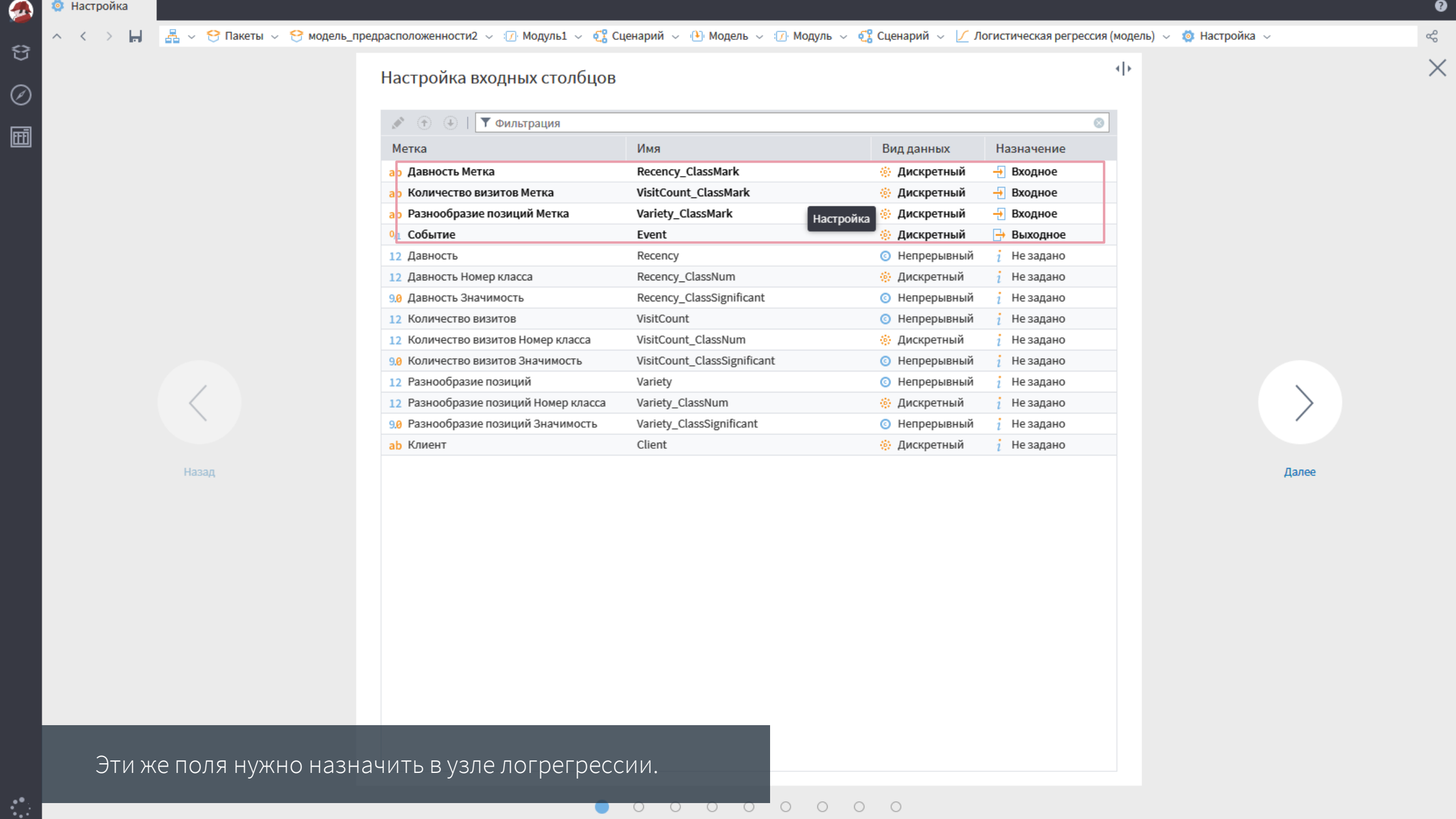


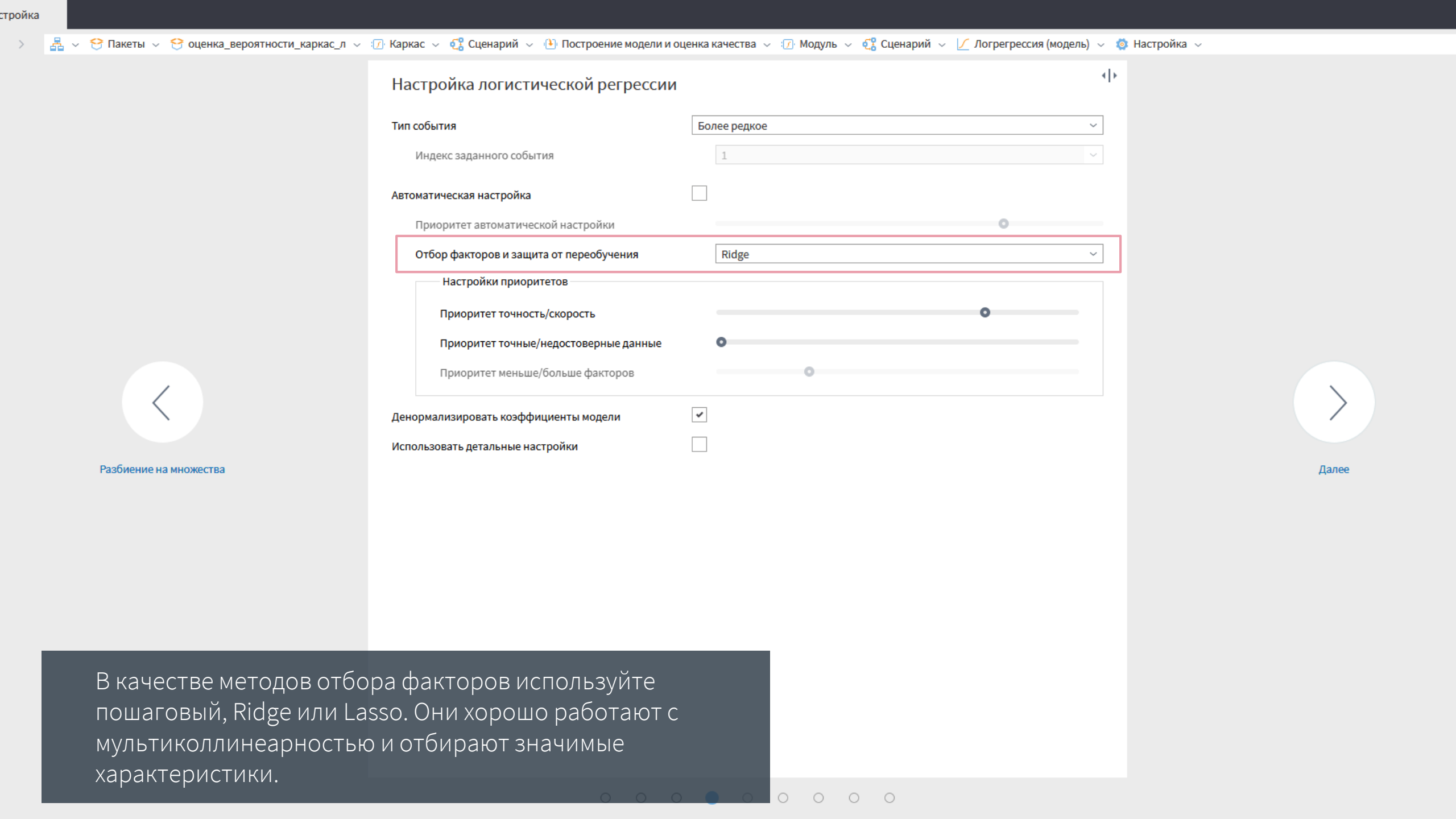
Перед узлом логистической регрессии нужно выполнить обязательную предобработку – биннинг или оптимальное квантование на основе WoE/IV-анализа (компонент «Конечные классы»). Мы опубликуем отдельную «методическую шпаргалку» по этому компоненту позже.

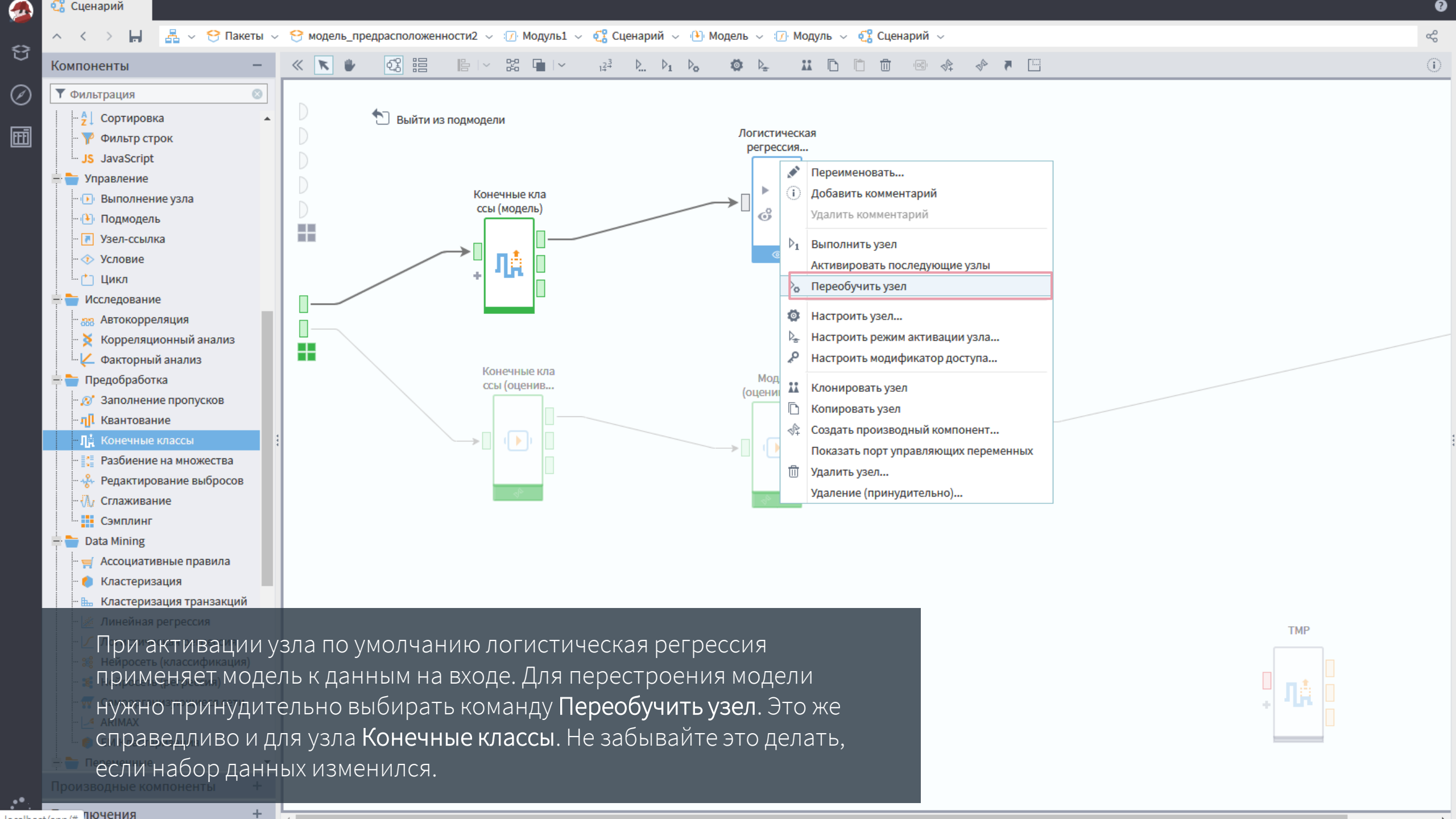

Настройка
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Модель
Модуль
Сценарий
Конечные классы (модель)
Настройка
Настройка назначений столбцов
Состояние входа
Не активировано
Активировать
№
Входные поля
Фильтрация
Вид данных
Назначение
1 12 Давность Непрерывный Входное
2 12 Количество визитов Непрерывный Входное
3 12 Разнообразие позиций Непрерывный Входное
4 0/1 Событие Дискретный Выходное
5 ab Клиент Дискретный Не выбрано
Назначение столбца
Выходное
Настройки выходного поля
Настраиваемое значение "события"
true
На конечные классы нужно подавать на вход все характеристики, которые, по нашему мнению, могут влиять на интересующее событие. Выходное поле должно быть дискретным логического типа. Событием делаем true.



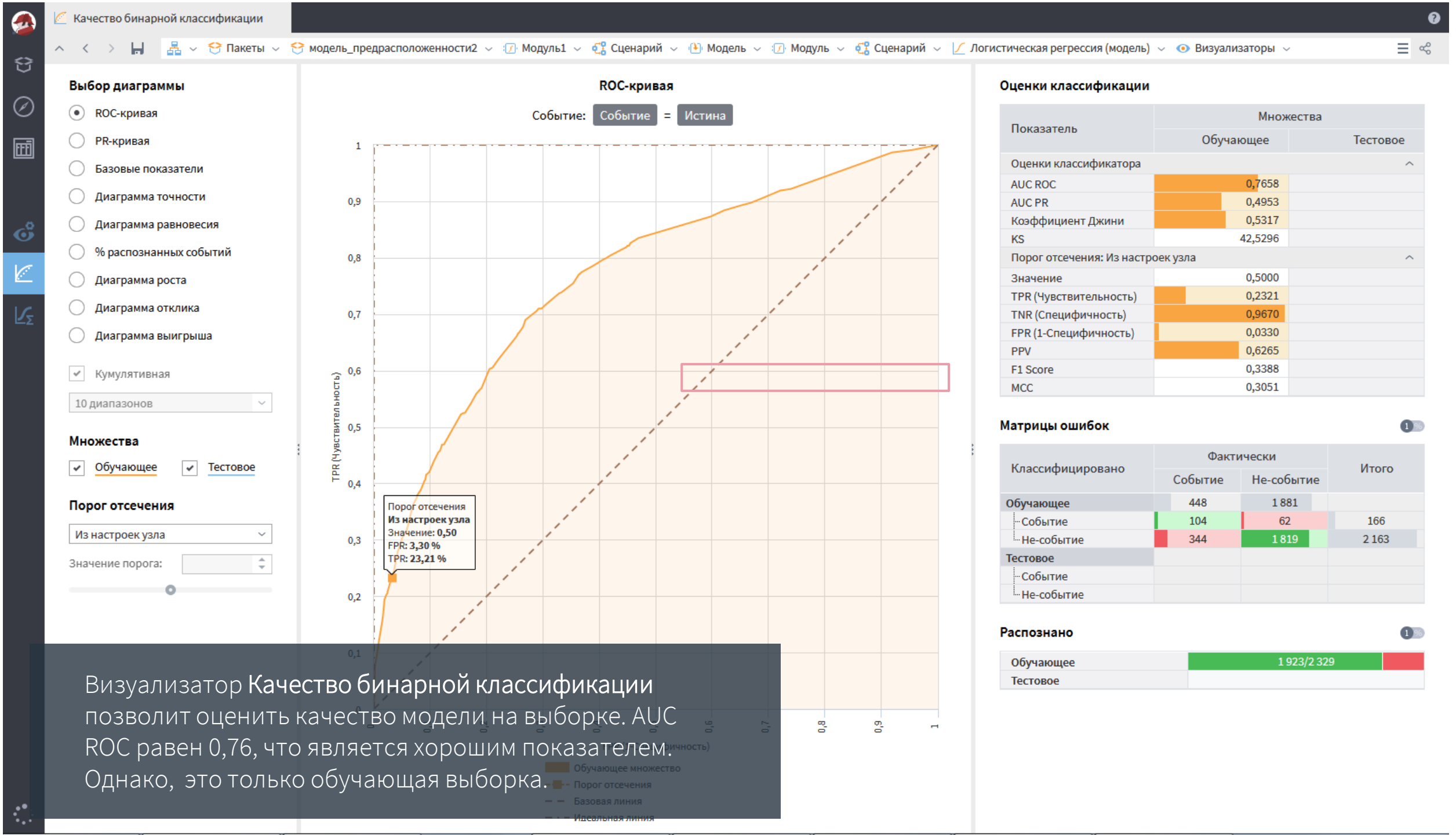
На выходе первого порта узла будет много полей, но для нас самое нужное – это поля с постфиксом **Метка** - преобразованное алгоритмом биннинга входное поле.


Настройка
Пакеты
модель_предрасположенности2
Модуль1
Сценарий
Модель
Модуль
Сценарий
Логистическая регрессия (модель)
Настройка
Настройка входных столбцов
Фильтрация
Метка
Имя
Вид данных
Назначение
Давность Метка
Recency_ClassMark
Дискретный
Входное
Количество визитов Метка
VisitCount_ClassMark
Дискретный
Входное
Разнообразие позиций Метка
Variety_ClassMark
Настройка
Дискретный
Входное
Событие
Event
Дискретный
Выходное
Давность
Recency
Непрерывный
Не задано
Давность Номер класса
Recency_ClassNum
Дискретный
Не задано
Давность Значимость
Recency_ClassSignificant
Непрерывный
Не задано
Количество визитов
VisitCount
Непрерывный
Не задано
Количество визитов Номер класса
VisitCount_ClassNum
Дискретный
Не задано
Количество визитов Значимость
VisitCount_ClassSignificant
Непрерывный
Не задано
Разнообразие позиций
Variety
Непрерывный
Не задано
Разнообразие позиций Номер класса
Variety_ClassNum
Дискретный
Не задано
Разнообразие позиций Значимость
Variety_ClassSignificant
Непрерывный
Не задано
Клиент
Client
Дискретный
Не задано
Назад
Далее
Эти же поля нужно назначить в узле логрегрессии.

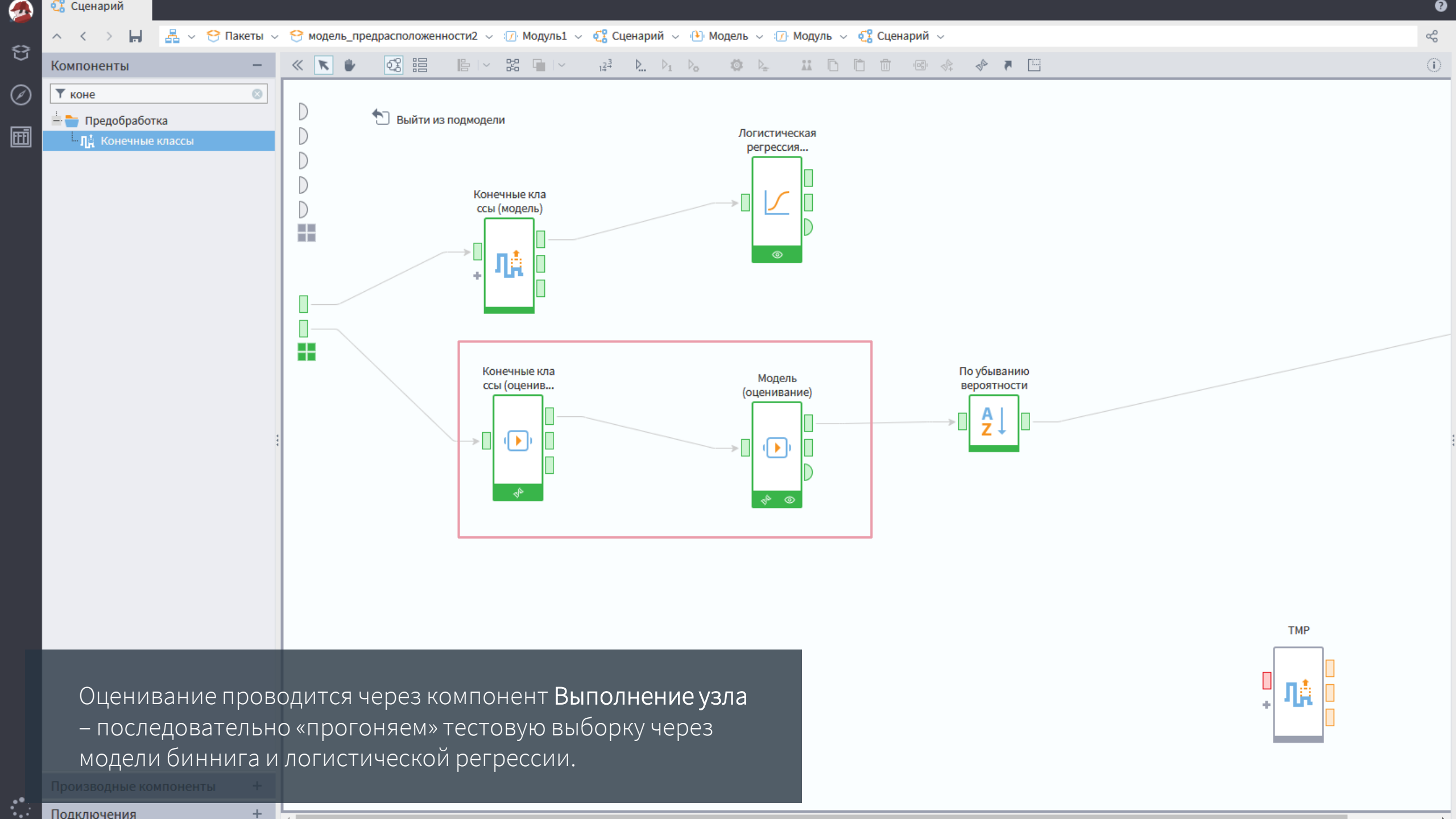

стройка
Пакеты
оценка_вероятности_каркас_л
Каркас
Сценарий
Построение модели и оценка качества
Модуль
Сценарий
Логрегрессия (модель)
Настройка
Настройка логистической регрессии
Тип события
Более редкое
Индекс заданного события
1
Автоматическая настройка
Приоритет автоматической настройки
Отбор факторов и защита от переобучения
Ridge
Настройки приоритетов
Приоритет точность/скорость
Приоритет точные/недостоверные данные
Приоритет меньше/больше факторов
Денормализировать коэффициенты модели
Использовать детальные настройки
Разбиение на множества
Далее
В качестве методов отбора факторов используйте пошаговый, Ridge или Lasso. Они хорошо работают с мультиколлинеарностью и отбирают значимые характеристики.



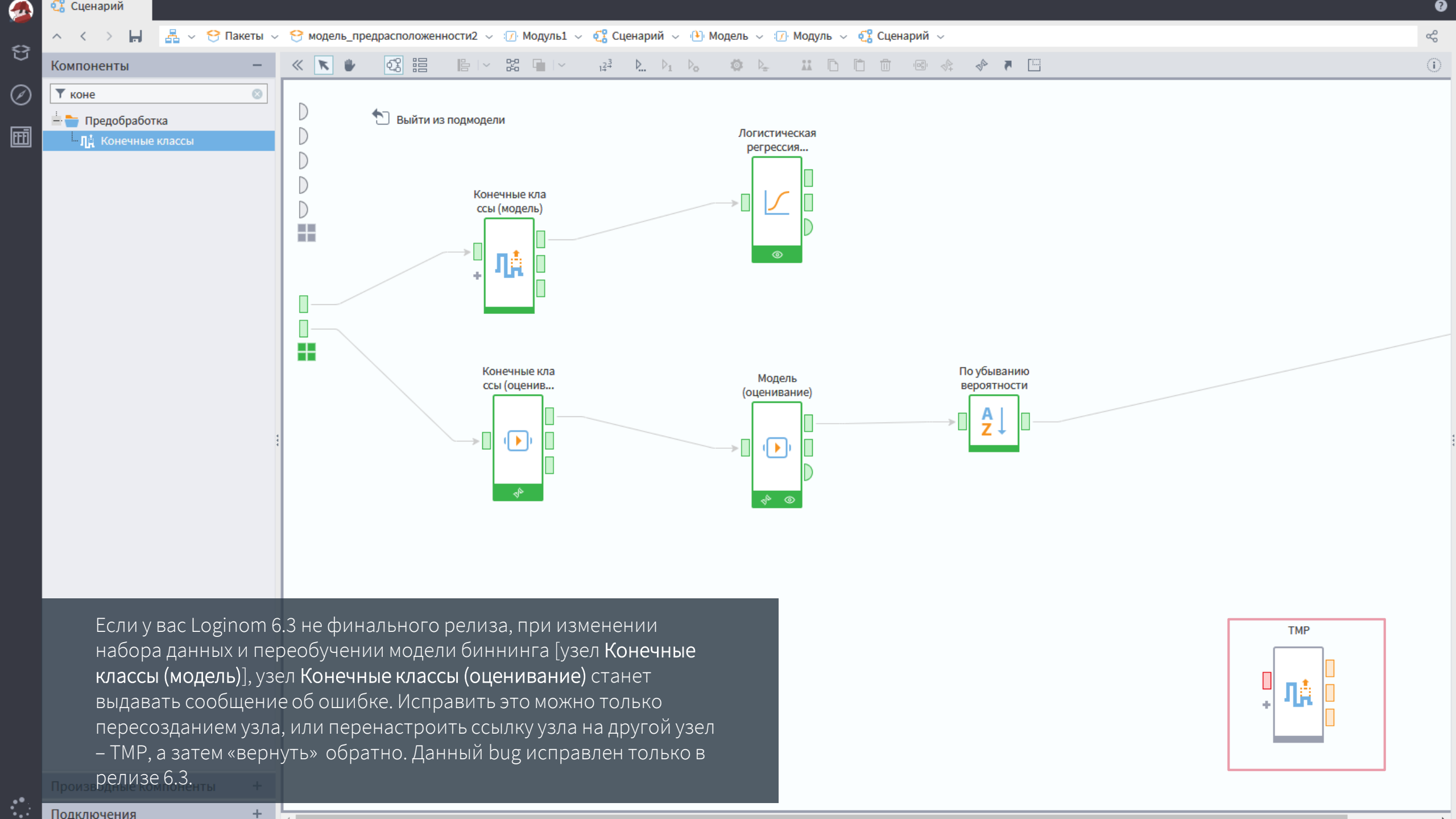
При активации узла по умолчанию логистическая регрессия применяет модель к данным на входе. Для перестроения модели нужно принудительно выбирать команду **Переобучить узел**. Это же справедливо и для узла **Конечные классы**. Не забывайте это делать, если набор данных изменился.



Визуализатор **Качество бинарной классификации** позволит оценить качество модели на выборке. AUC ROC равен 0,76, что является хорошим показателем. Однако, это только обучающая выборка.



Оценивание проводится через компонент **Выполнение узла** – последовательно «прогоняем» тестовую выборку через модели биннига и логистической регрессии.



Если у вас Loginom 6.3 не финального релиза, при изменении набора данных и переобучении модели биннинга [узел **Конечные классы (модель)**], узел **Конечные классы (оценивание)** станет выдавать сообщение об ошибке. Исправить это можно только пересозданием узла, или перенастроить ссылку узла на другой узел – TMP, а затем «вернуть» обратно. Данный bug исправлен только в релизе 6.3.

Качество бинарной классификации

Пакеты › модель_предрасположенности2 › Модуль1 › Сценарий › Модель › Модуль › Сценарий › Модель (оценивание) › Визуализаторы

Выбор диаграммы

- (•) ROC-кривая
- ( ) PR-кривая
- ( ) Базовые показатели
- ( ) Диаграмма точности
- ( ) Диаграмма равновесия
- ( ) % распознанных событий
- ( ) Диаграмма роста
- ( ) Диаграмма отклика
- ( ) Диаграмма выигрыша

☑ Кумулятивная

10 диапазонов

Множества

☑ Обучающее ☑ Тестовое

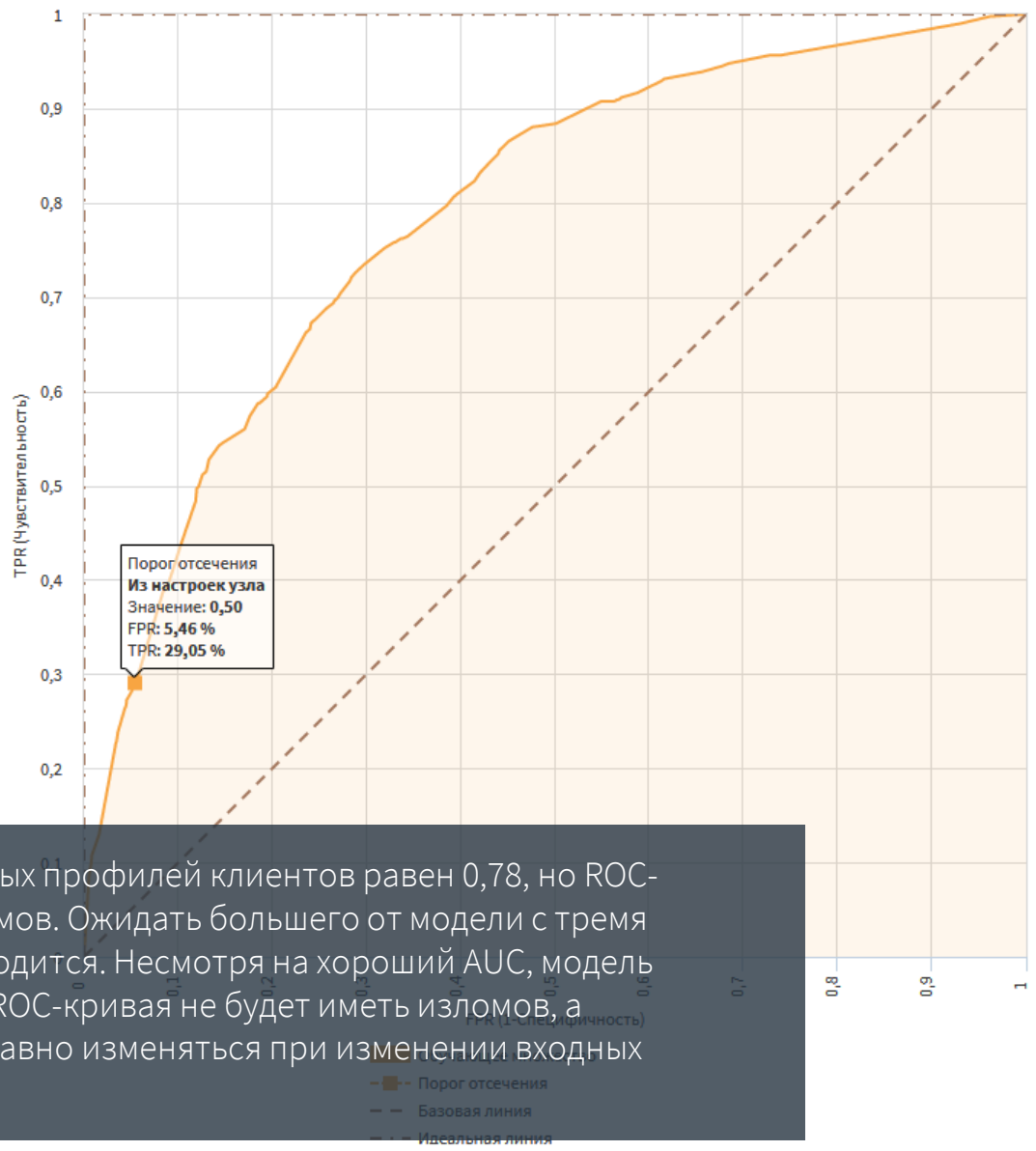
Порог отсечения

Из настроек узла

Значение порога:

ROC-кривая

Событие: Событие = Истина



Оценки классификации

Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,7876	
AUC PR	0,4906	
Коэффициент Джини	0,5753	
KS	44,1419	
Порог отсечения: Из настроек узла		
Значение	0,5000	
TPR (Чувствительность)	0,2905	
TNR (Специфичность)	0,9454	
FPR (1-Специфичность)	0,0546	
PPV	0,5628	
F1 Score	0,3832	
MCC	0,3106	

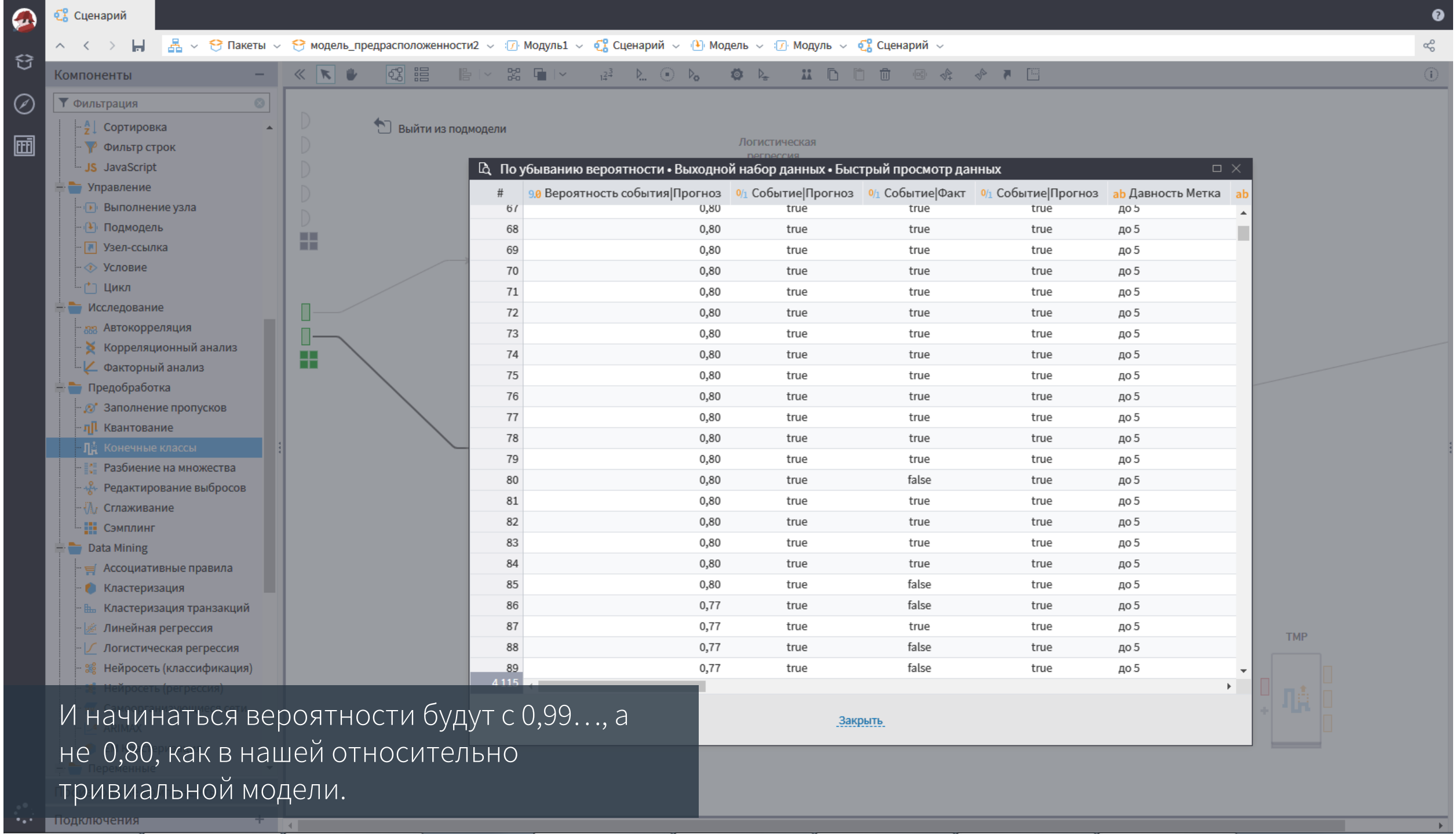
Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	802	3 313	
Событие	233	181	414
Не-событие	569	3 132	3 701
Тестовое			
Событие			
Не-событие			

Распознано

Обучающее	3 365/4 115
Тестовое	

Индекс AUC для оцениваемых профилей клиентов равен 0,78, но ROC-кривая имеет больше изломов. Ожидать большего от модели с тремя характеристиками не приходится. Несмотря на хороший AUC, модель грубая. У хорошей модели ROC-кривая не будет иметь изломов, а вероятности клиентов – плавно изменяться при изменении входных признаков.



#	Вероятность события|Прогноз	Событие|Прогноз	Событие|Факт	Событие|Прогноз	Давность Метка
67	0,80	true	true	true	до 5
68	0,80	true	true	true	до 5
69	0,80	true	true	true	до 5
70	0,80	true	true	true	до 5
71	0,80	true	true	true	до 5
72	0,80	true	true	true	до 5
73	0,80	true	true	true	до 5
74	0,80	true	true	true	до 5
75	0,80	true	true	true	до 5
76	0,80	true	true	true	до 5
77	0,80	true	true	true	до 5
78	0,80	true	true	true	до 5
79	0,80	true	true	true	до 5
80	0,80	true	false	true	до 5
81	0,80	true	true	true	до 5
82	0,80	true	true	true	до 5
83	0,80	true	true	true	до 5
84	0,80	true	true	true	до 5
85	0,80	true	false	true	до 5
86	0,77	true	false	true	до 5
87	0,77	true	true	true	до 5
88	0,77	true	false	true	до 5
89	0,77	true	false	true	до 5

4 115

И начинаться вероятности будут с 0,99…, а не 0,80, как в нашей относительно тривиальной модели.

Качество модели

Хорошая модель, конечно, должна включать больше характеристик, чем три. Хотя при прогнозировании повторного визита **Давность** и **Количество визитов** почти всегда имеют очень сильную значимость и могут определять до 90% индекса ROC-AUC. Но даже в таком виде модель способна лучше прогнозировать будущее, чем методы сегментации типа RF- и RFM-анализа, поскольку последние опираются только на прошлое, а модель обучается по меткам будущих событий.

Возможные вопросы

Является ли подмодель с расчетом профилей кандидатом на включение в библиотеку компонентов нашего командного проекта (для секции 1)?

Ответ: Определенно да, это будет плюсом, но мы не настаиваем.

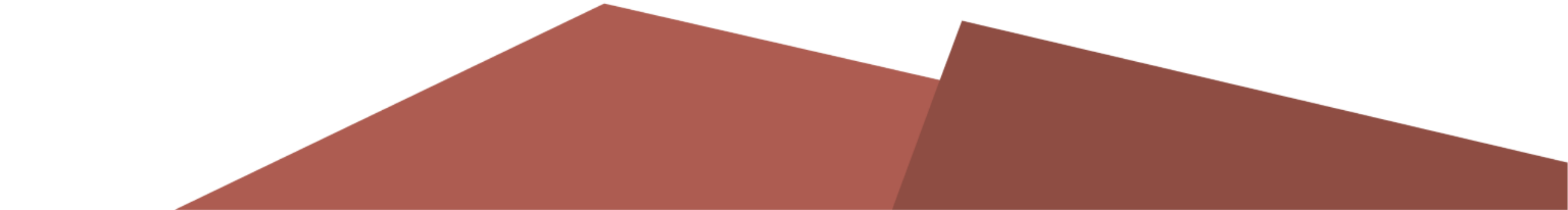
Сколько характеристик должно поступать на вход модели?

Ответ: Ограничений нет, но легко придумать 2-3 десятка.

В конечных классах оценивается значимость характеристики. Можно ли в логрегрессию не подавать характеристики с отсутствующей значимостью?

Ответ: Можно, но процедуры Ridge/Lasso и пошаговый отбор сами могут отсечь незначимые и коррелирующие характеристики.

Дополнительно для секции 3



Важно

В задаче секции № 3 построение характеристик клиентов более сложное, чем в рассмотренном нами примере. Их нужно считать не на определенную дату, а на каждую дату перед новым заказом клиента.

Внимательно следите за логикой расчетов: большой риск при формировании характеристик заглянуть «в будущее», то есть опираться на данные, которые не известны на момент расчетов. Эта ошибка аналитиков известна как Data Leakage.