МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

Кафедра цифровых технологий и
прикладной информатики

Б1.В.06 ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

Лабораторные работы. Аналитическая платформа Loginom

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к лабораторным работам и самостоятельной работе

Направление подготовки

09.03.03 Прикладная информатика

Квалификация (степень) выпускника
бакалавр

Уфа 2024

Рекомендовано к изданию методической комиссией экономического факультета (протокол № 7 от 21.03.2024 г.)

Составитель: доцент, к.ф.-м.н. Шамсутдинова Т.М.

Рецензент: ст. преподаватель Прокофьева С.В.

Ответственный за выпуск: зав. кафедрой ЦТиПИ, д.т.н.,
Беляева А.С.

г.Уфа, БГАУ, Кафедра цифровых технологий и прикладной информатики

СОДЕРЖАНИЕ

Лабораторная работа 1. Система бизнес-аналитики Loginom 5
Часть 1. Знакомство с платформой..... 5
Часть 2. Визуализаторы..... 10
Лабораторная работа 2. Кластеризация 32
Лабораторная работа 3. Нейросеть (регрессия) и линейная регрессия 43
Лабораторная работа 4. Нейросеть (классификация) и логистическая регрессия..... 55
Лабораторная работа 5. Анализ данных с использованием Loginom..... 66
Список литературы..... 67

Лабораторная работа 1. Система бизнес-аналитики Loginom

Часть 1. Знакомство с платформой

Цель работы: ознакомиться с принципами работы с ML-платформой Loginom

Содержание работы:

Аналитическая low-code-платформа Loginom [1] предоставляет инструменты для реализации всех аналитических процессов: от интеграции и подготовки данных до моделирования, развертывания и визуализации (рис. 1.1).

Рисунок 1.1 - Процессы аналитики платформы Loginom

Используя платформу Loginom, можно решать следующие бизнес-задачи [1]:

– управление рисками: кредитный конвейер, скоринг, антифрод;

– клиентская аналитика: сегментация клиентов, противодействие оттоку, кросс-продажи;

– очистка данных: очистка и удаление дублей, создание золотой записи, стандартизация НСИ (нормативно-справочная информация);

– маркетинг: директ-маркетинг, оптимизация цен, оценка эффективности рекламы;

– логистика: прогнозирование спроса, оптимизация запасов, расчет страховых запасов;

– диагностика: статистический контроль качества, оценка вероятности поломок, цифровые двойники.

Прикладные бизнес-решения на основе аналитической платформы Loginom представлены на рис. 1.2.

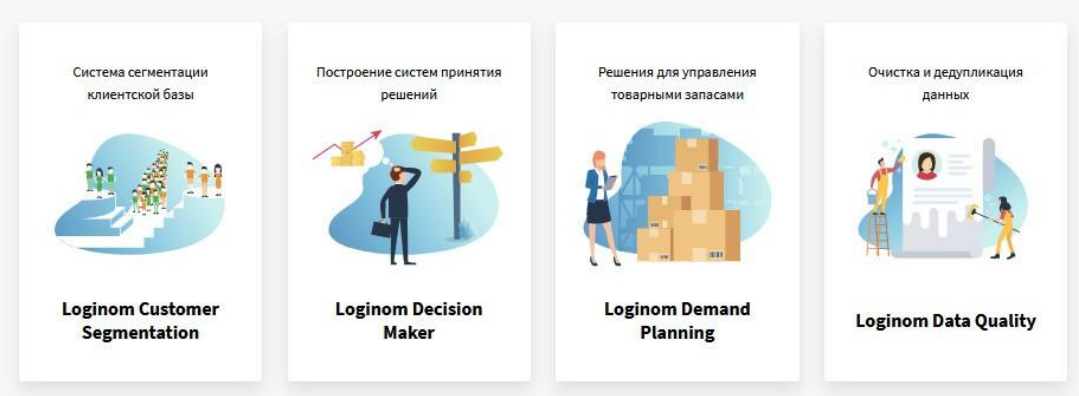


Рисунок 1.2 - Прикладные бизнес-решения

Для некоммерческого использования есть бесплатная клиентская версия Loginom Community Edition (для профессиональной бизнес-аналитики – версия Loginom Personal). Также есть три варианта серверных приложений: Team, Standard и Enterprise.

Установим бесплатную версию Loginom Community (https://loginom.ru/download) (рис. 1.3). После заполнения анкеты на указанный email придет ссылка на скачивание продукта.

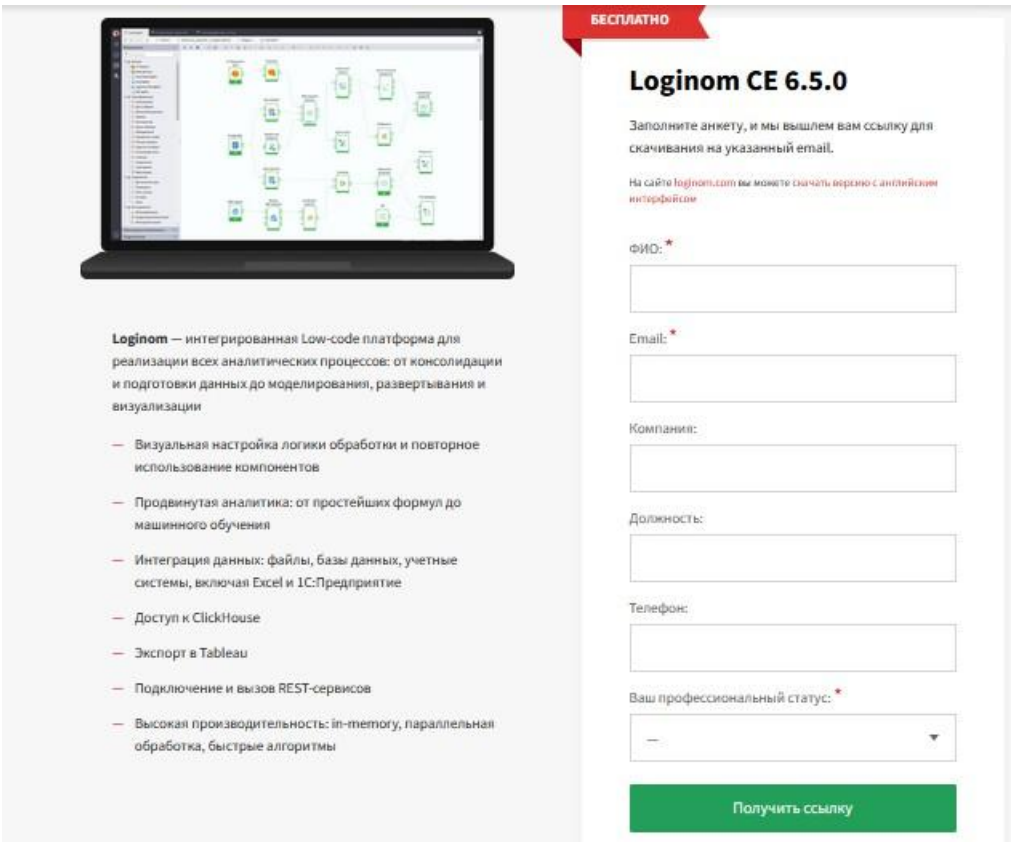


Рисунок 1.3 - Получение дистрибутива Loginom CE

В зависимости от разрядности операционной системы скачиваем по ссылке один из дистрибутивов (рис. 1.4). На момент написания пособия актуальна версия Loginom 6.5.0.

Рисунок 1.4 - Ссылка на скачивание дистрибутива Loginom CE

В начальном окне можно создать первый пакет. Назовем его Анализ_продаж (рис. 1.5).

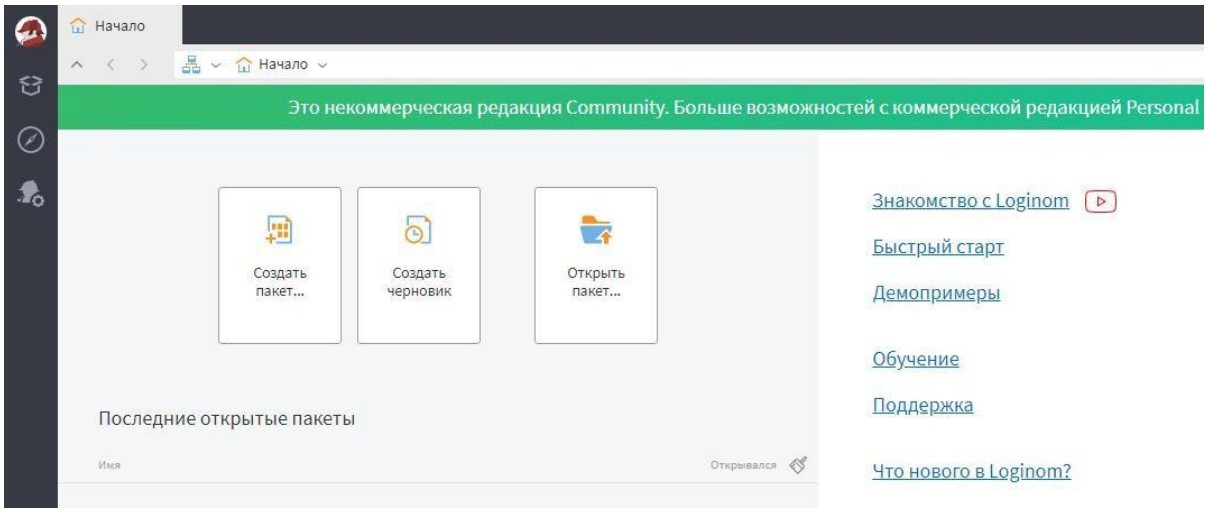


Рисунок 1.5 - Начальное окно Loginom

В первом созданном пакете по умолчанию создается Модуль1, включающий Сценарий – пока пустой (рис. 1.6).

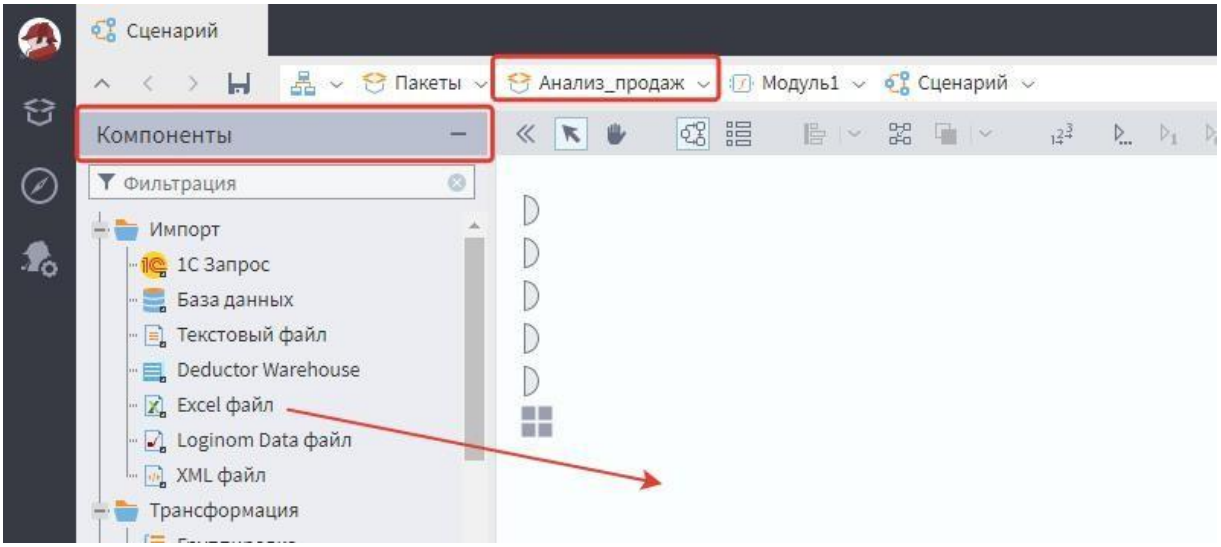


Рисунок 1.6 - Сценарий пакета Анализ_продаж

Компоненты добавляются в сценарий перетаскиванием из панели в рабочую область.

Левая колонка представлена пиктограммами меню (рис. 1.7).

Рисунок 1.7 - Пиктографическое меню

С помощью меню Навигация можно перемещаться по элементам пакета (рис. 1.8).

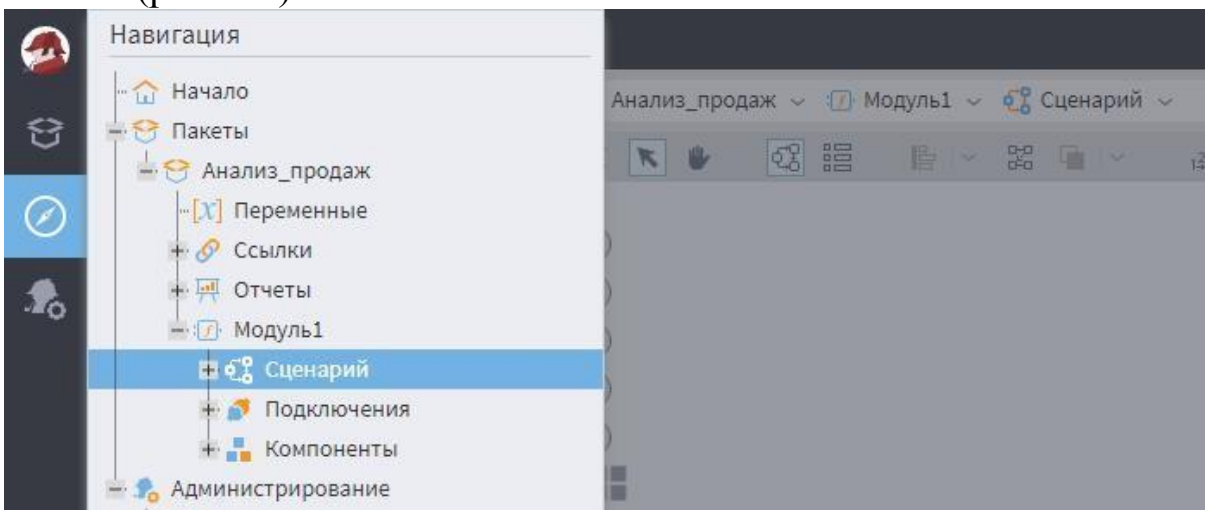


Рисунок 1.8 - Перемещение по элементам пакета

Важно: автосохранение пакетов в Loginom предусмотрено по расписанию в платных версиях, поэтому перед закрытием программы необходимо сохранить текущий пакет или все открытые во время текущей сессии пакеты.

Часть 2. Визуализаторы

Цель работы: ознакомиться с визуализаторами в Loginom.

Содержание работы:

Для составления аналитической отчетности в Loginom используются визуализаторы. Среди доступных визуализаторов: графики, таблицы, кубы.

Для построения визуализаций необходимо импортировать в пакет исходные данные. Воспользуемся набором финансовых данных от компании Microsoft (https://go.microsoft.com/fwlink/?LinkID=521962). Эти данные поставляются вместе с Power BI Desktop [2]. Скачаем набор данных Financial Sample.xlsx на диск компьютера.

В зависимости от решаемой задачи в область Сценария добавляются нужные компоненты. Они представляют собой узлы сценария. Сам сценарий – это последовательность связанных узлов. Коммуникация между узлами устанавливается с помощью портов: входные порты (располагаются слева) принимают данные, через выходные порты (располагаются справа) узел передает данные.

Перейдем в Сценарий в пакете Анализ_продаж (см. рис. 1.8). Добавим в рабочую область Сценария компонент категории Импорт: Excel файл и кликнем по нему мышкой, т.е. сделаем его активным. Станет доступно пиктографическое меню для действий с этим компонетом (рис. 2.1).

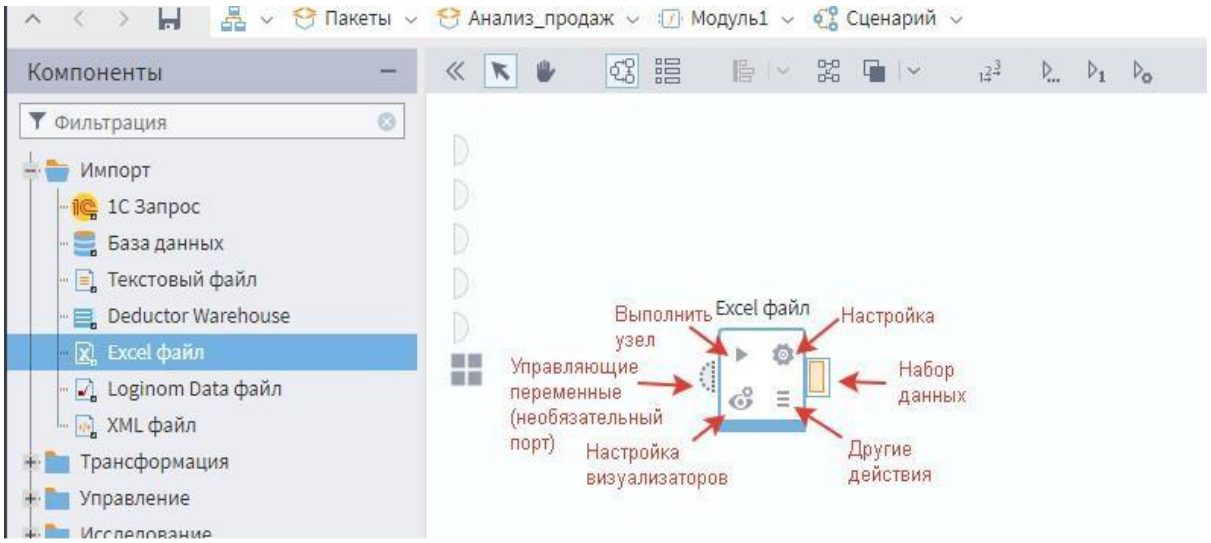


Рисунок 2.1 - Узел Excel файл

Произведем настройку узла, для этого кликнем мышкой по пиктограмме шестеренка внутри компонента. И перейдем к мастеру

заполнения настроек. В адресной строке будет отображаться путь к текущему диалоговому окну.

Настройка узла Excel файл состоит из четырех окон. В первом окне выбирается файл на локальном диске с расширением .xls или .xlsx (рис. 2.2).

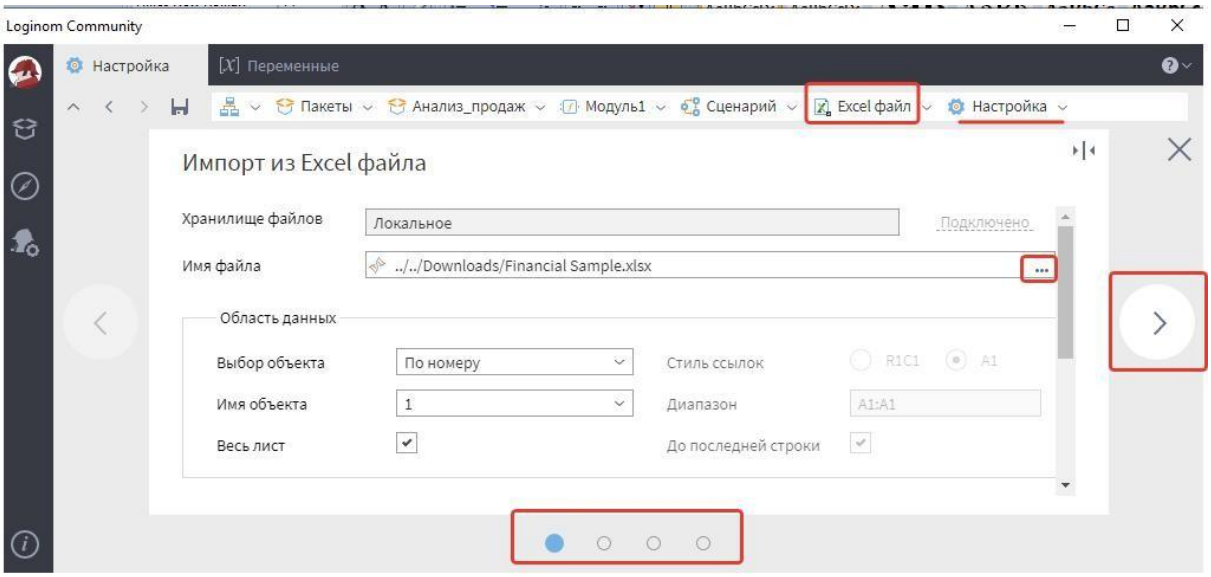


Рисунок 2.2 - Импорт данных

Во втором окне можно выбрать столбцы для импорта (с помощью галочки в строке Использовать), а также проверить корректность определения типа и вида данных (рис. 2.3).

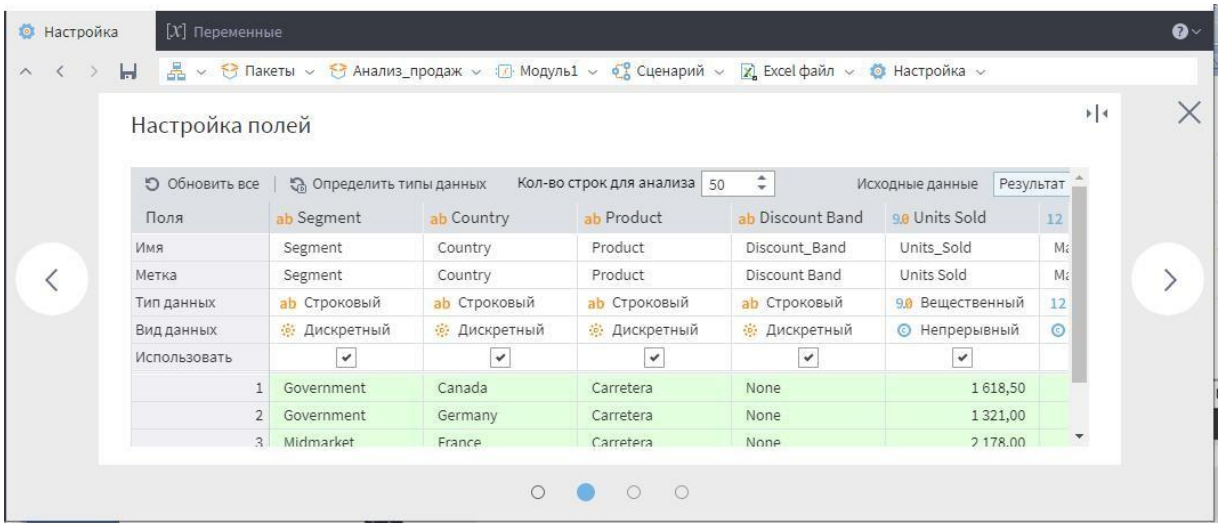


Рисунок 2.3 - Настройка полей

В третьем окне можно еще раз проверить правильность определения вида и типа данных и настроить их назначение по раскрывающейся стрелке справа для каждого столбца, если это необходимо (рис. 2.4).

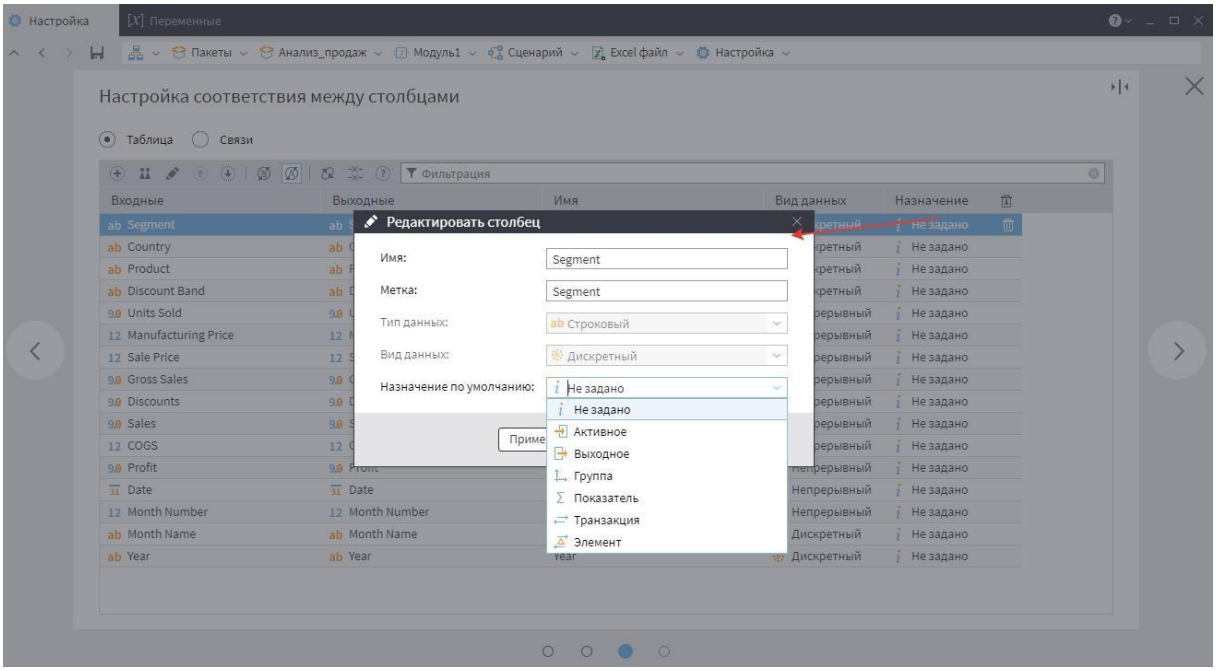


Рисунок 2.4 - Настройка назначения полей данных

В последнем окне можно изменить метку узла, создать комментарий и сохранить настройки (рис. 2.5).

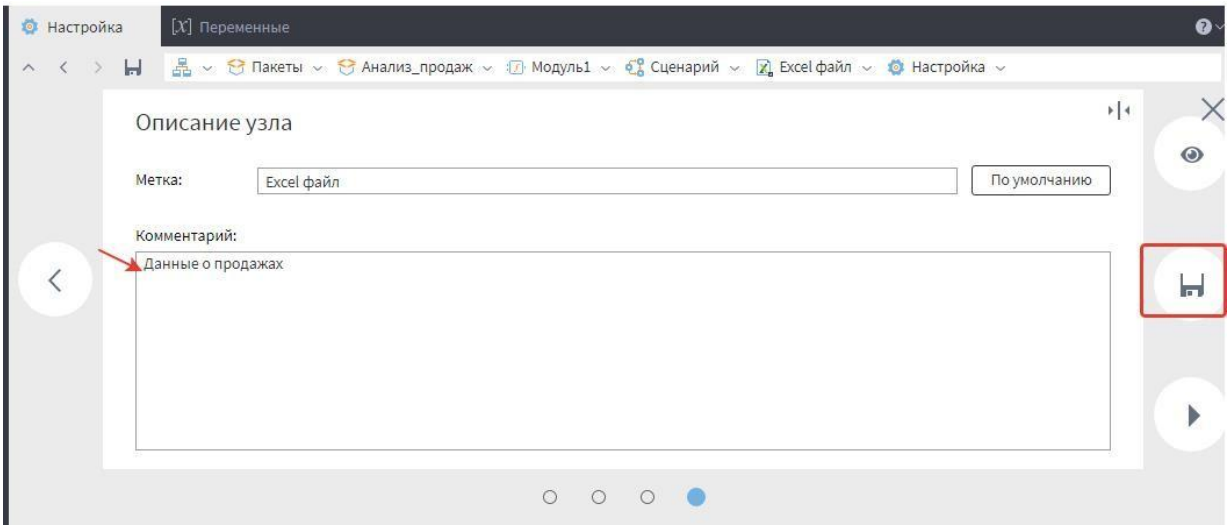


Рисунок 2.5 - Завершение работы
с мастером настройки узла Excel файл

Остановимся подробнее на типах и видах данных (рис. 2.3), с которыми работают узлы Loginom [3].

Тип данных:

- логический (значения True/False);
- дата/время;
- вещественный (числа с плавающей точкой);
- целый;
- строковый.

Вид данных: непрерывный и дискретный.

Непрерывные данные – это количественные показатели, для которых имеют смысл арифметические операции. Они потенциально могут принимать любые значения из некоторого интервала.

Дискретные данные – это обычно качественные показатели: номинальные или порядковые, например страна, сегмент, продукт, отток (ушел клиент или остался) и т.п. Тип данных у таких показателей логический или строковый, но может быть и числовым, если числа выступают метками уровней такого показателя, или число выступает в роли идентификатора: транзакции, товара и т.п. Номер телефона может быть записан как последовательность цифр без разделителей (пробелов, скобок, тире), но работа с ним как с числом не имеет смысла.

Система Loginom оперирует фактами и измерениями. Факты – это количественные показатели (тип данных целочисленный или вещественный), а также дата/время. Измерения – дискретные данные: строкового или логического типа.

В многомерной модели данных количественные и качественные признаки хранятся в различных таблицах. Количественные образуют единственную таблицу фактов. Качественные данные – таблицы измерений.

Измерения являются логической основой модели данных. Проектирование многомерной модели данных всегда начинается с измерений. В многомерной модели данных измерения играют роль индексов, служащих для идентификации конкретных значений в ячейках куба.

Измерения могут быть простыми и иерархическими. Иерархические измерения могут содержать подчиненные измерения и образовывать уровни иерархии. Например, измерение Регион может иметь иерархически подчиненное измерение Город.

Многомерная модель данных с иерархией измерений называется снежинка, а без иерархии – звезда. Автором концепции использования измерений в моделировании данных является Эдгар Кодд [4].

Факт – это показатель (признак, атрибут), количественно описывающий исследуемый процесс или объект, например, цена, количество, остаток на складе и т.д.

В многомерных структурах данных факты образуют отдельную таблицу, с которой соединены все таблицы измерений. Любой факт ассоциирован с некоторым измерением, так как количественный показатель без измерения (категориального признака) не имеет смысла. Цена – это просто столбец чисел, который приобретает смысл

благодаря связи с каким-либо измерением, например, товаром, продуктом, работой или услугой [5].

Выберем в узле Excel файл пиктограмму Визуализаторы и приступим к их построению. Начнем с добавления в рабочую область визуализатора Качество данных. Визуализатор можно добавить простым перетаскиванием нужного элемента или выделив элемент в списке и нажав на + в рабочей области. После добавления нужного визуализатора необходимо в него войти для настройки (рис. 2.6).

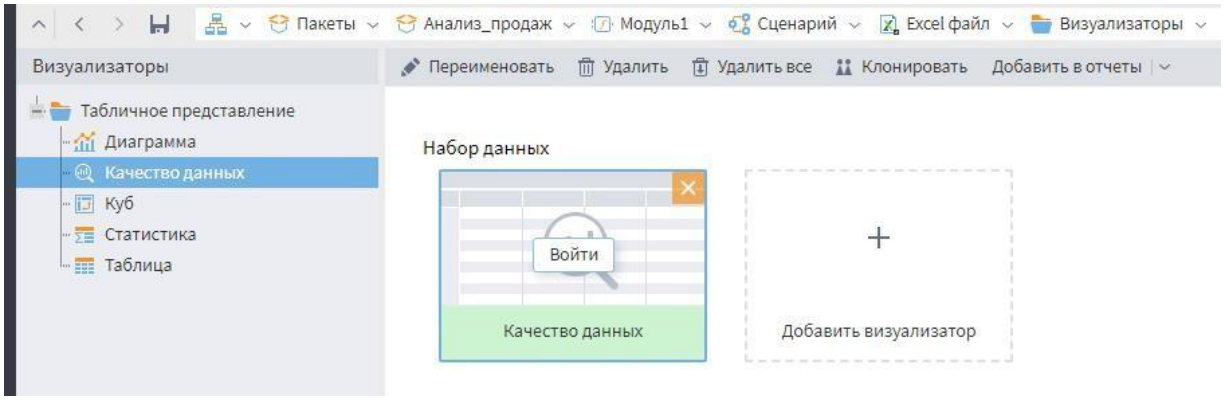


Рисунок 2.6 - Добавление визуализатора

В визуализаторе Качество данных нужно выбрать названия столбцов (рис. 2.7) для расчета статистик (по умолчанию выбираются все доступные столбцы). Итоговая сводка представлена на рис. 2.8.

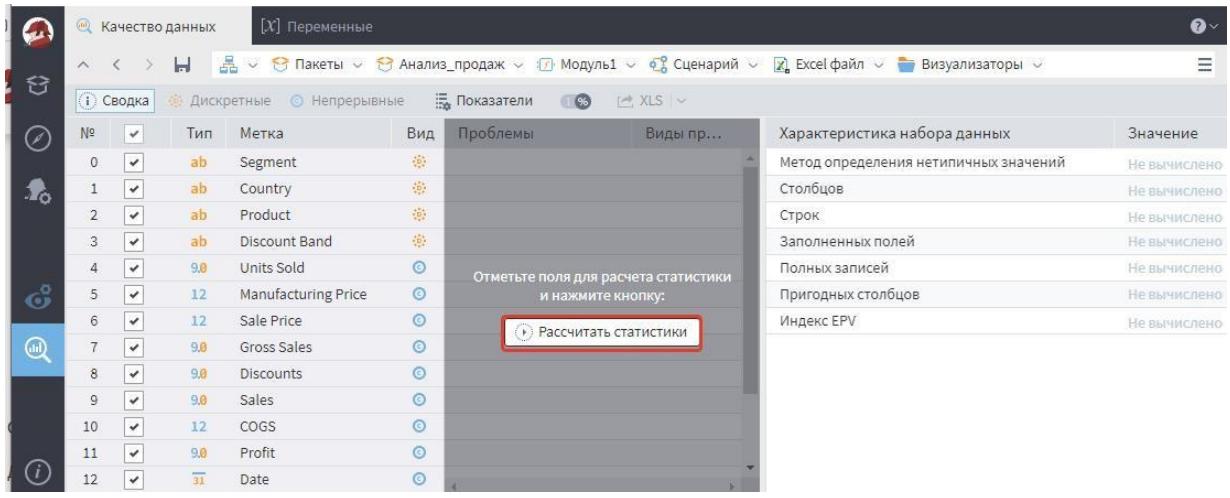


Рисунок 2.7 - Выбор столбцов/полей

Инструмент Диаграмма для текущих данных менее информативен, будем использовать построение диаграмм для визуализатора Куб.

Добавим визуализатор Куб, реализующий один из методов OnLine Analytical Processing (OLAP, OLAP-куб), организующий представление данных в виде многомерных кубов.

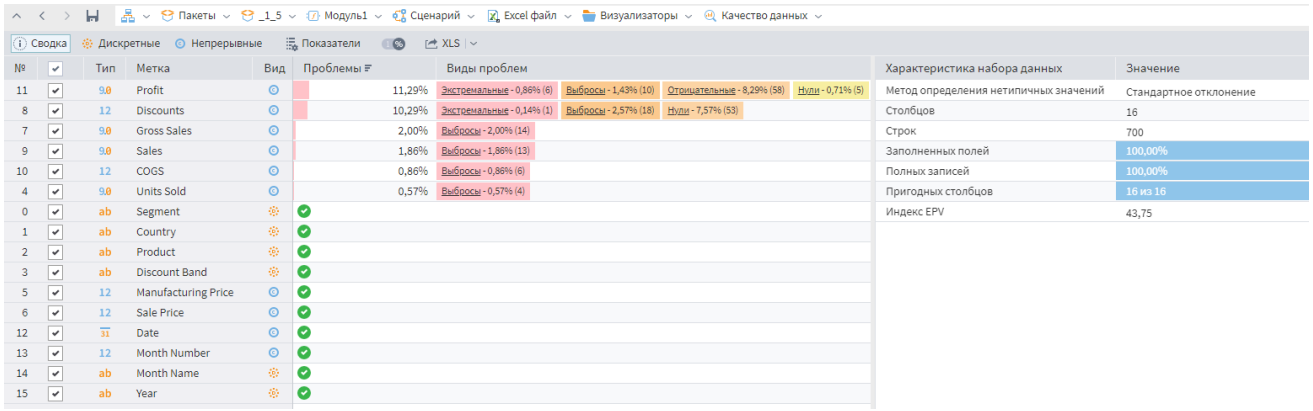


Рисунок 2.8 - Сводка по качеству данных

Войдем в визуализатор и добавим измерения в Куб перетаскиванием или кликом мышкой на кнопке +. Измерения для фильтрации добавляются в верхнюю строку рабочей области, измерения для строк добавляются слева от таблицы, измерения для столбцов добавляются перед фактами во вторую верхнюю строку рабочей области. Также выберем с помощью кнопки Факты два количественных показателя: Gross Sales и Discounts – и определим для них способ агрегации – суммирование (рис. 2.9).

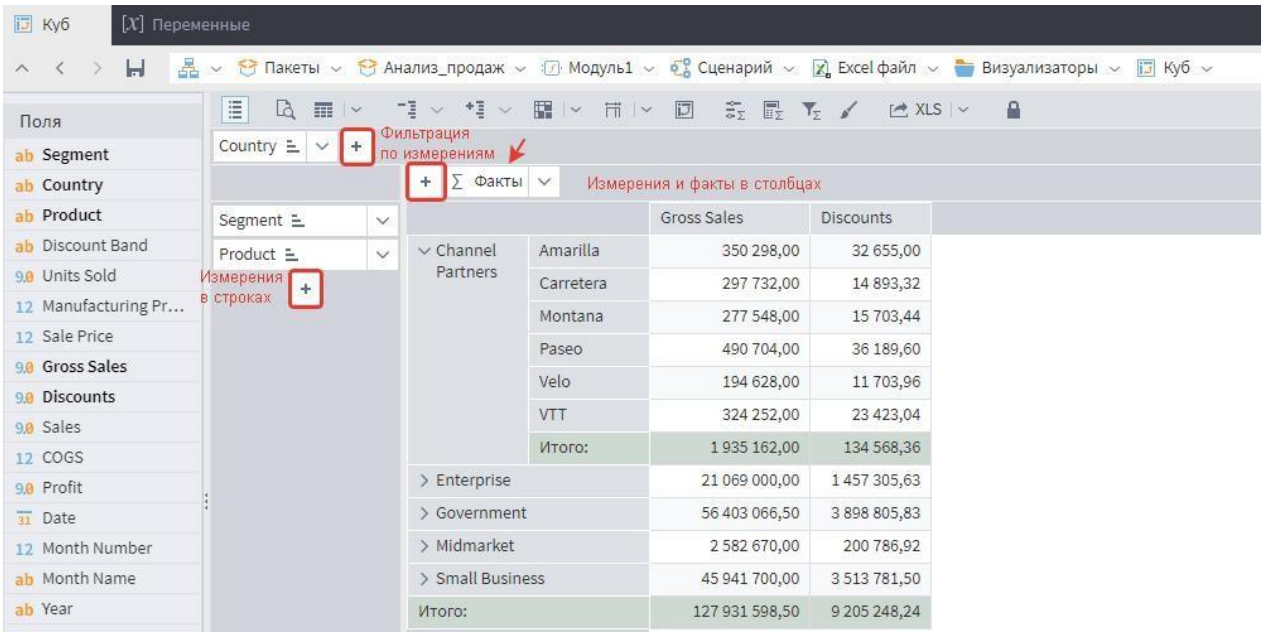
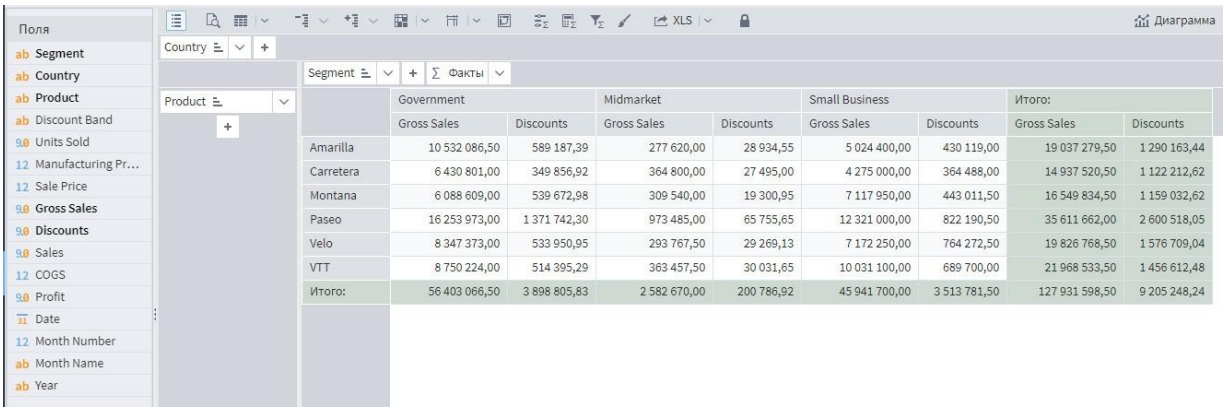


Рисунок 2.9 - Добавление фактов и измерений в Куб

Куб представляет собой сводную таблицу, в которой измерения – категориальные признаки в данных, которые будут разбивать данные

на группы, а факты – количественные показатели, значения которых необходимо вывести на экран.

Перенесем измерение Segment в измерения для столбцов. Теперь группировка в строках по продукту, а в стобцах по сегменту (рис. 2.10). Ползунок перемещен в конец сводной таблицы, так как целиком она не помещается на экране.



	Government		Midmarket		Small Business		Итого:	
	Gross Sales	Discounts	Gross Sales	Discounts	Gross Sales	Discounts	Gross Sales	Discounts
Amarilla	10 532 086,50	589 187,39	277 620,00	28 934,55	5 024 400,00	430 119,00	19 037 279,50	1 290 163,44
Carretera	6 430 801,00	349 856,92	364 800,00	27 495,00	4 275 000,00	364 488,00	14 937 520,50	1 122 212,62
Montana	6 088 609,00	539 672,98	309 540,00	19 300,95	7 117 950,00	443 011,50	16 549 834,50	1 159 032,62
Paseo	16 253 973,00	1 371 742,30	973 485,00	65 755,65	12 321 000,00	822 190,50	35 611 662,00	2 600 518,05
Velo	8 347 373,00	533 950,95	293 767,50	29 269,13	7 172 250,00	764 272,50	19 826 768,50	1 576 709,04
VTT	8 750 224,00	514 395,29	363 457,50	30 031,65	10 031 100,00	689 700,00	21 968 533,50	1 456 612,48
Итого:	56 403 066,50	3 898 805,83	2 582 670,00	200 786,92	45 941 700,00	3 513 781,50	127 931 598,50	9 205 248,24

Рисунок 2.10 - Перестроенная сводная таблица

При добавлении факта с помощью кнопки Факт выбираем нужное поле таблицы данных из списка (рис. 2.11) и перед добавлением факта определяем вид агрегации (по умолчанию – это сумма). Можно выбрать сразу несколько (рис. 2.12).

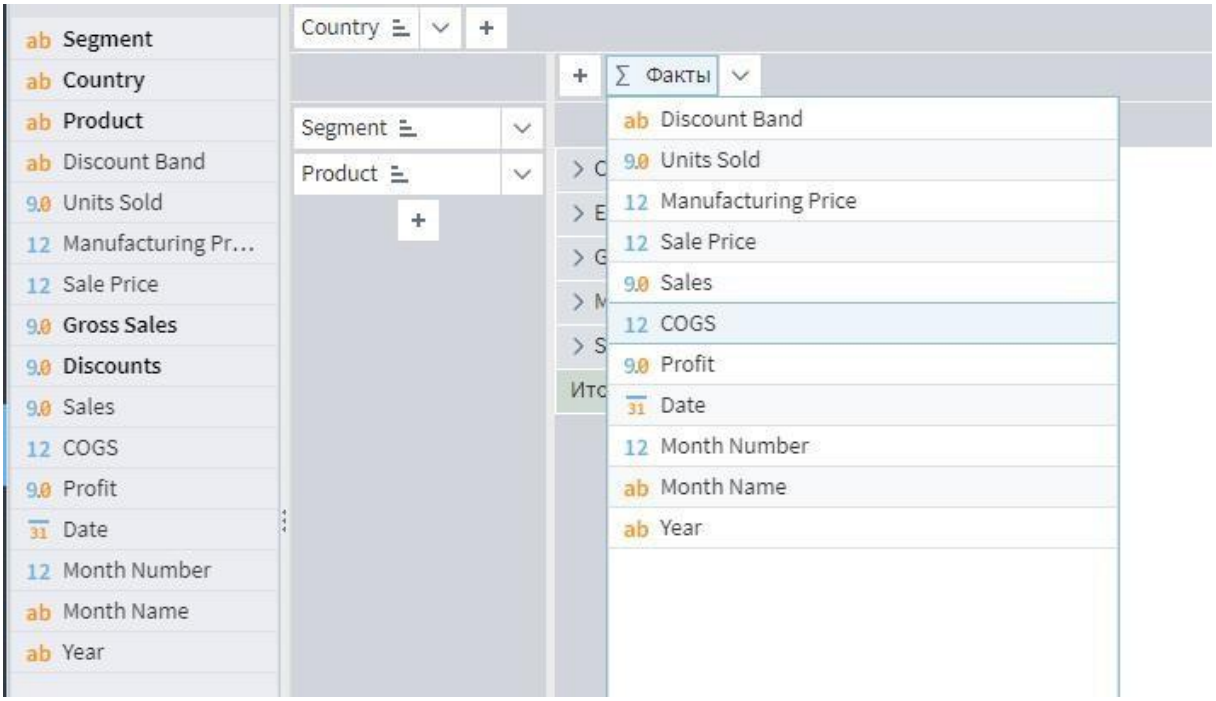


Рисунок 2.11 - Добавление факта

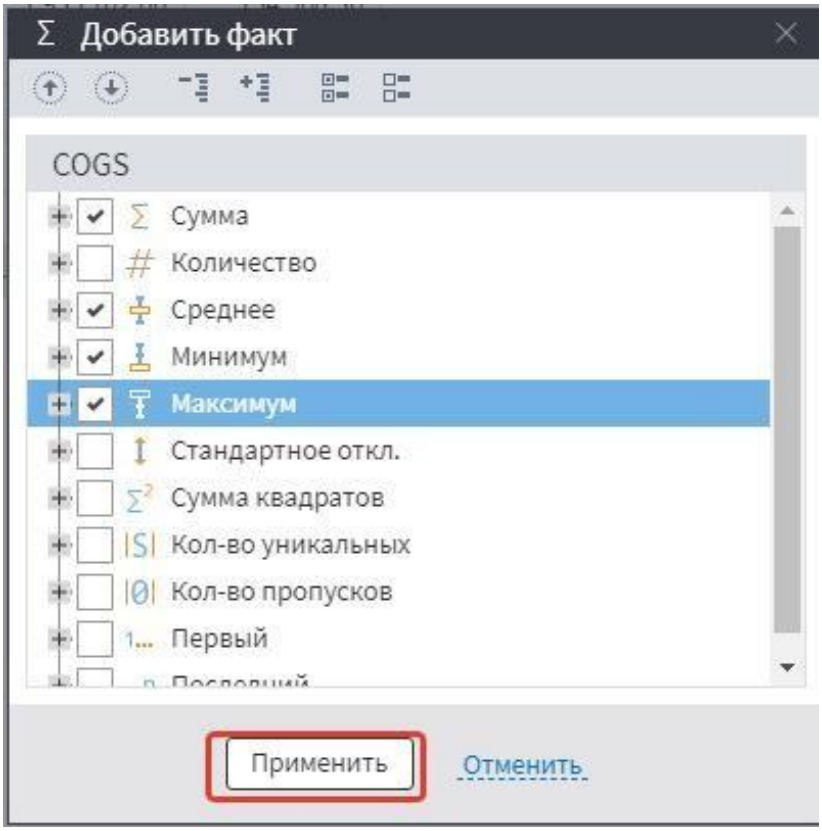


Рисунок 2.12 - Настройка вида агрегации

Если необходимо изменить вид агрегации для уже добавленного в Куб факта, то справа от кнопки Факты по стрелке раскрывают контекстное меню для работы с фактами и выбирают опцию Настроить факты (рис. 2.13).

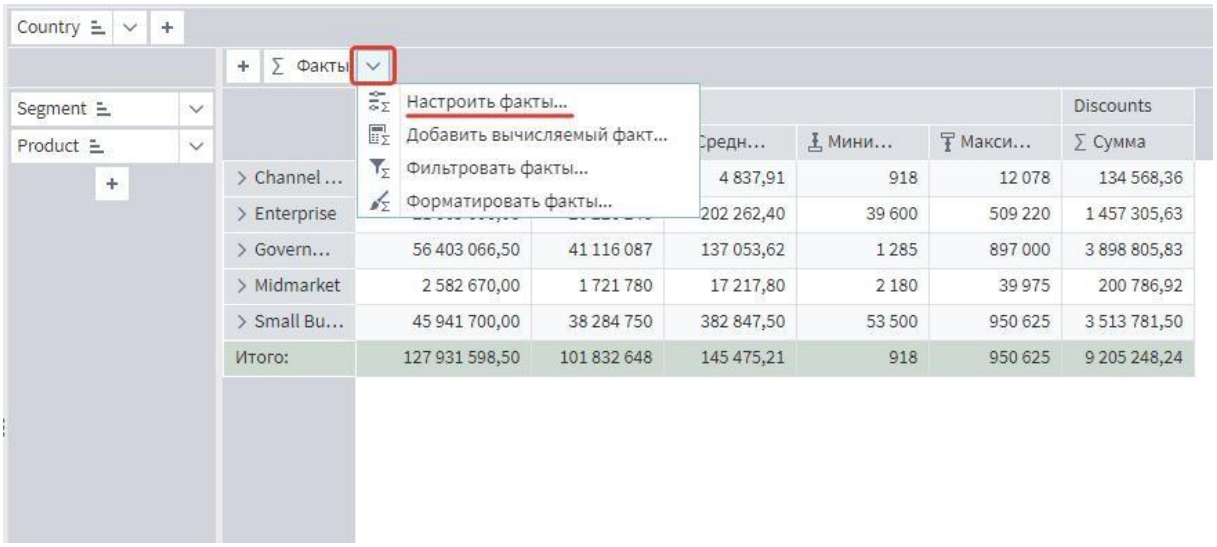


Рисунок 2.13 - Настройка фактов

Становится доступным окно настройки фактов, в котором можно убрать ранее добавленный факт или изменить для него вид агрегации (рис. 2.14).

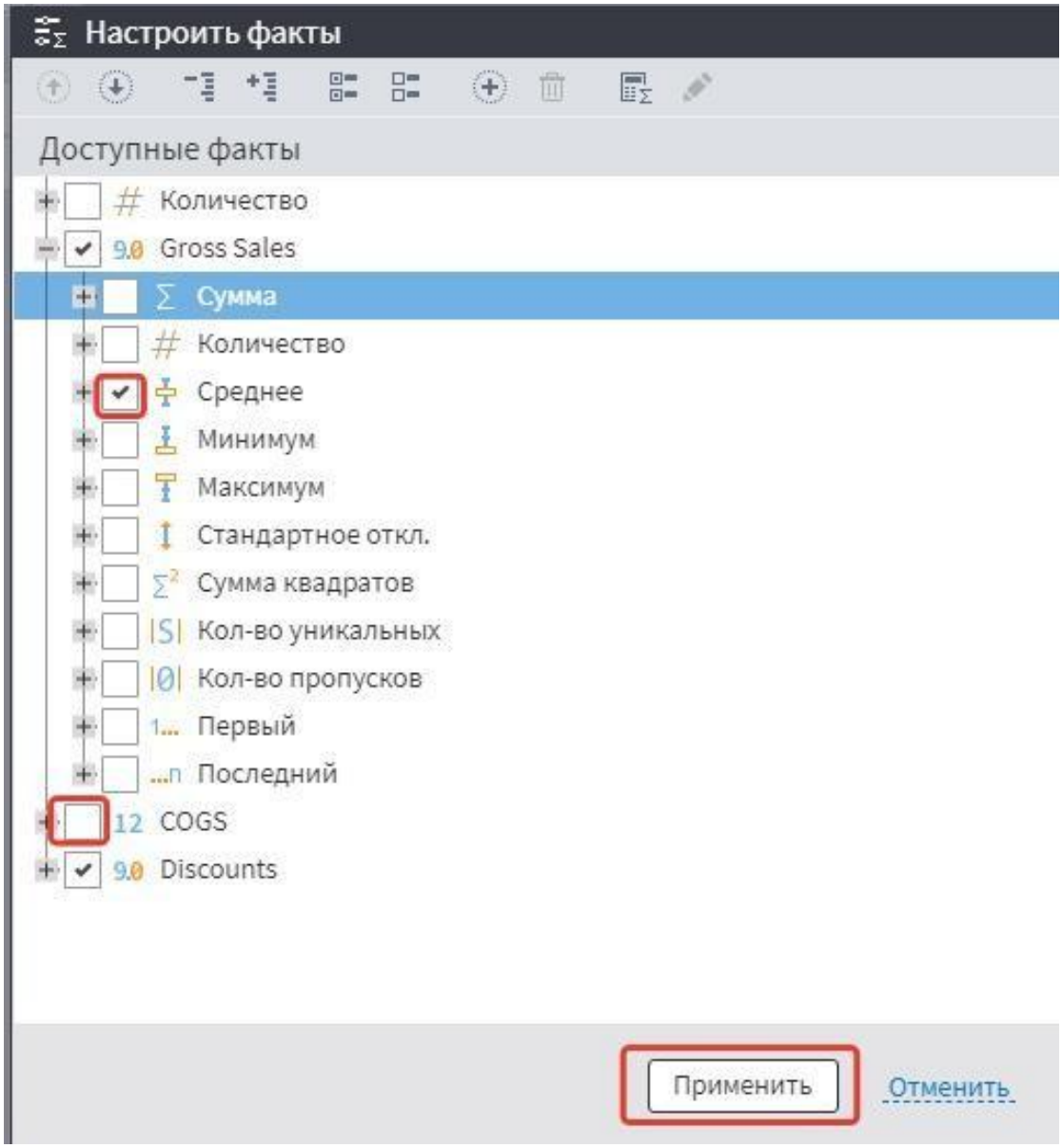


Рисунок 2.14 - Настройка фактов

Также щелчком правой кнопки мыши на одном из фактов можно вызвать контекстное меню, которое позволяет удалять факты, добавлять вычисляемые факты и производить другие манипуляции с фактами (рис. 2.15).

Добавим вычисляемый факт: Gross Sales Part, который будет определен как доля продаж каждого продукта в сегменте. С помощью контекстного меню для фактов (см. рис. 2.13) перейдем в диалоговое окно Добавить вычисляемый факт (рис. 2.16).

Обратите внимание: в имени переменной вместо пробелов используются подчеркивания. Добавим выражение для вычисления доли конкретного продукта в сегменте в общих продажах всех продуктов в сегменте. Значение переменной Gross_Sales (добавляется в поле выражения для вычисления простым щелчком на метке

соответствующего факта в левом нижнем углу) необходимо разделить на сумму значений Gross_Sales по продуктам сегмента. Чтобы добавить переменную, значение которой соответствует нужному вычислению, начинаем набирать имя переменной для суммирования и в раскрывающемся контекстном списке находим нужную переменную. Имя переменной указывает на то, какой статистический показатель она рассчитывает: Gross_Sales.Sum.Total.Product (значение которой и есть общая сумма (.Sum.Total) значений показателя Gross_Sales по продуктам (.Product)). Можно при необходимости использовать для вычисляемого выражения условия, функции и т.п. После добавления вычисляемого факта сводная таблица примет вид, как на рис. 2.17.

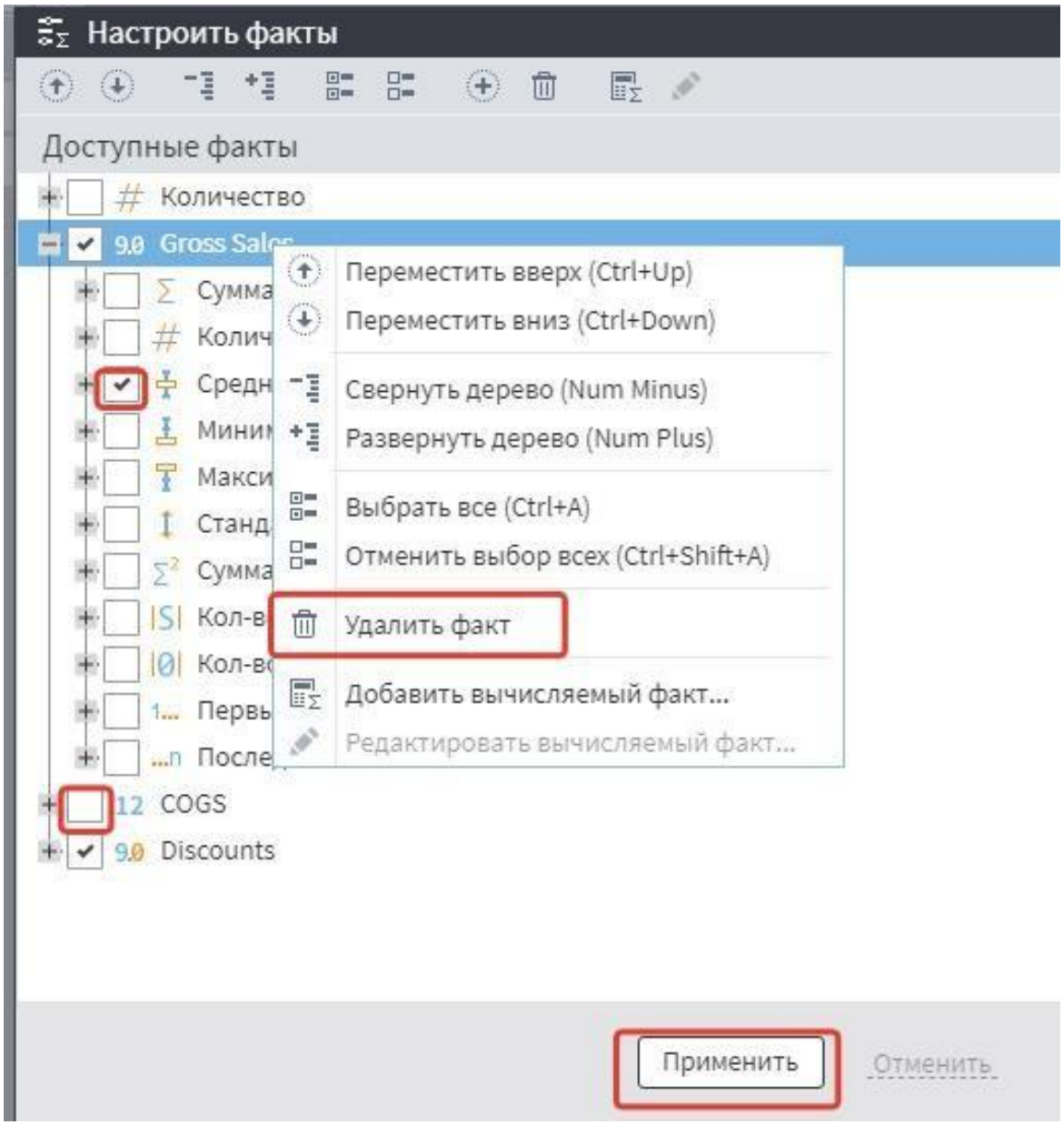


Рисунок 2.15 - Настройка фактов

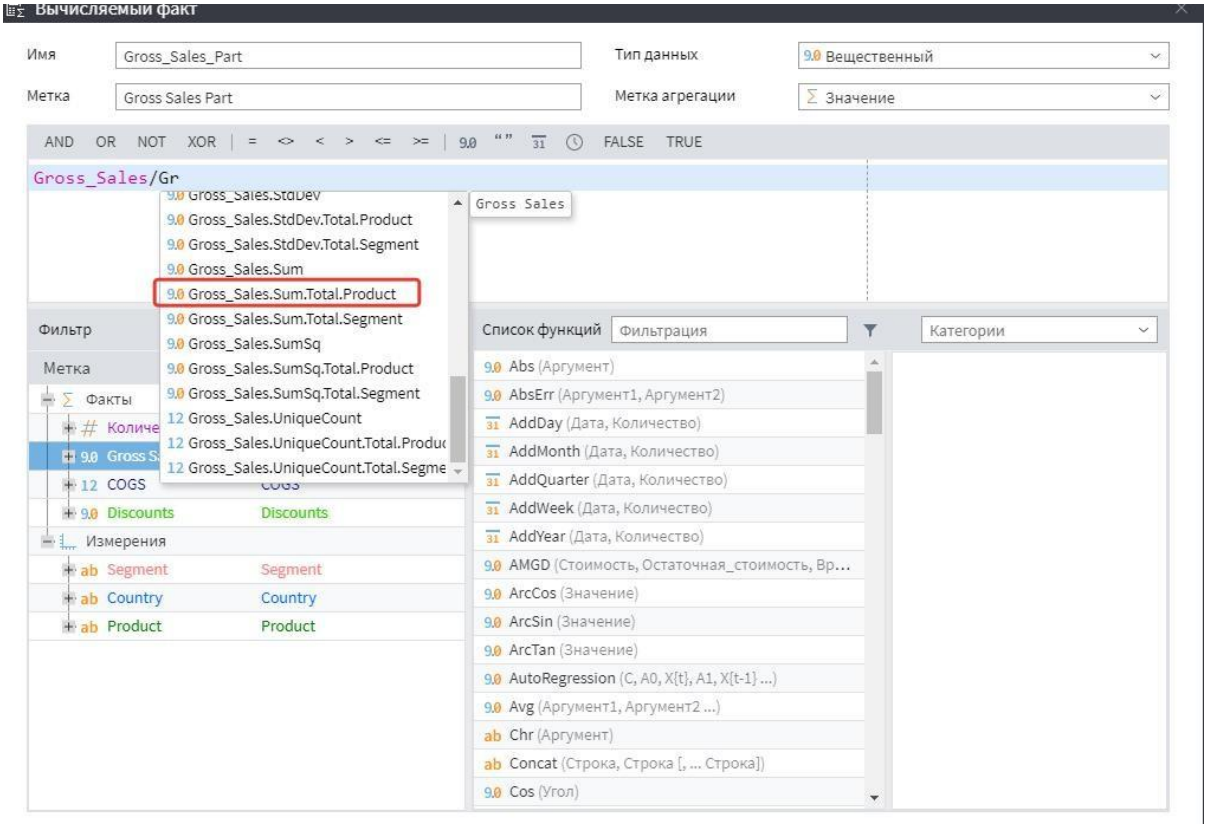
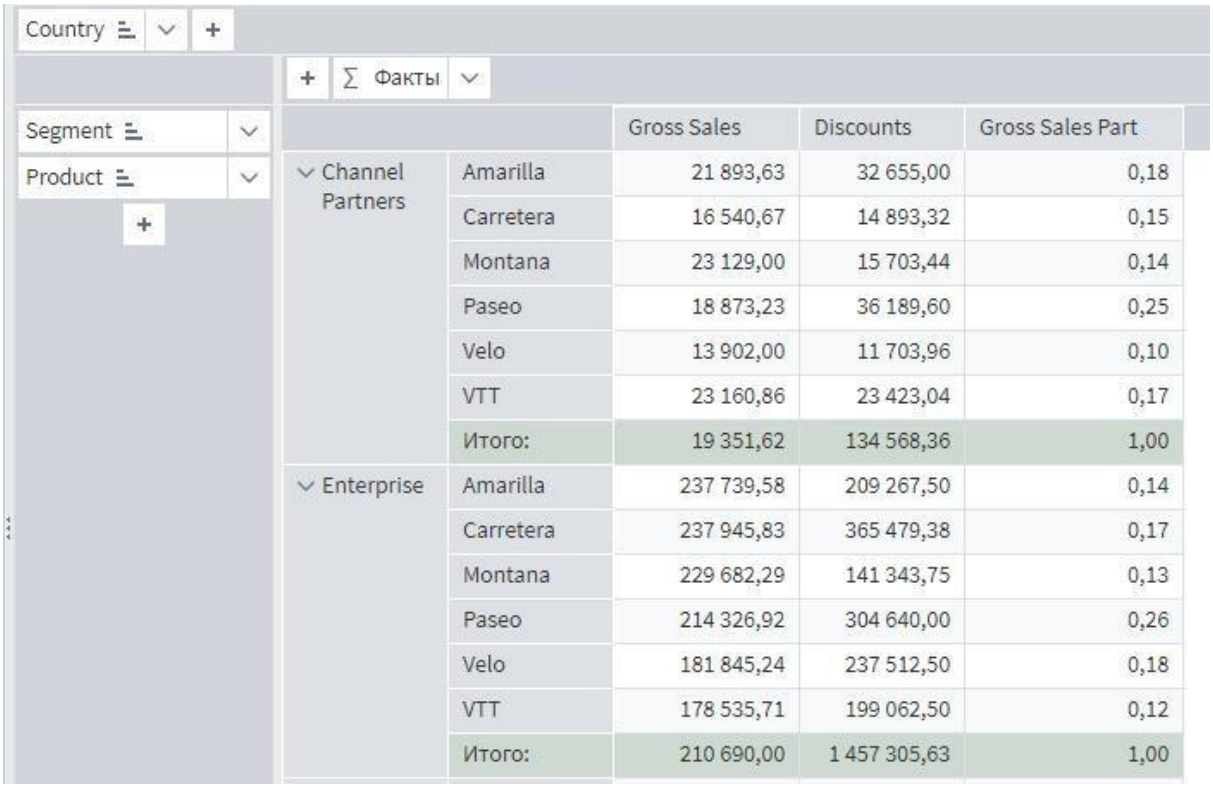


Рисунок 2.16 - Настройка Вычисляемого факта



Country

Факты

Segment

Product

		Gross Sales	Discounts	Gross Sales Part
Channel Partners	Amarilla	21 893,63	32 655,00	0,18
	Carretera	16 540,67	14 893,32	0,15
	Montana	23 129,00	15 703,44	0,14
	Paseo	18 873,23	36 189,60	0,25
	Velo	13 902,00	11 703,96	0,10
	VTT	23 160,86	23 423,04	0,17
	Итого:	19 351,62	134 568,36	1,00
Enterprise	Amarilla	237 739,58	209 267,50	0,14
	Carretera	237 945,83	365 479,38	0,17
	Montana	229 682,29	141 343,75	0,13
	Paseo	214 326,92	304 640,00	0,26
	Velo	181 845,24	237 512,50	0,18
	VTT	178 535,71	199 062,50	0,12
	Итого:	210 690,00	1 457 305,63	1,00

Рисунок 2.17 - Фрагмент сводной таблицы с вычисляемым фактом

Если необходимо отображать большее число знаков после запятой для показателей с типом данных вещественный, можно провести Форматирование фактов (см. рис. 2.13) и настроить необходимые параметры в соответствующем диалоговом окне (рис. 2.18).

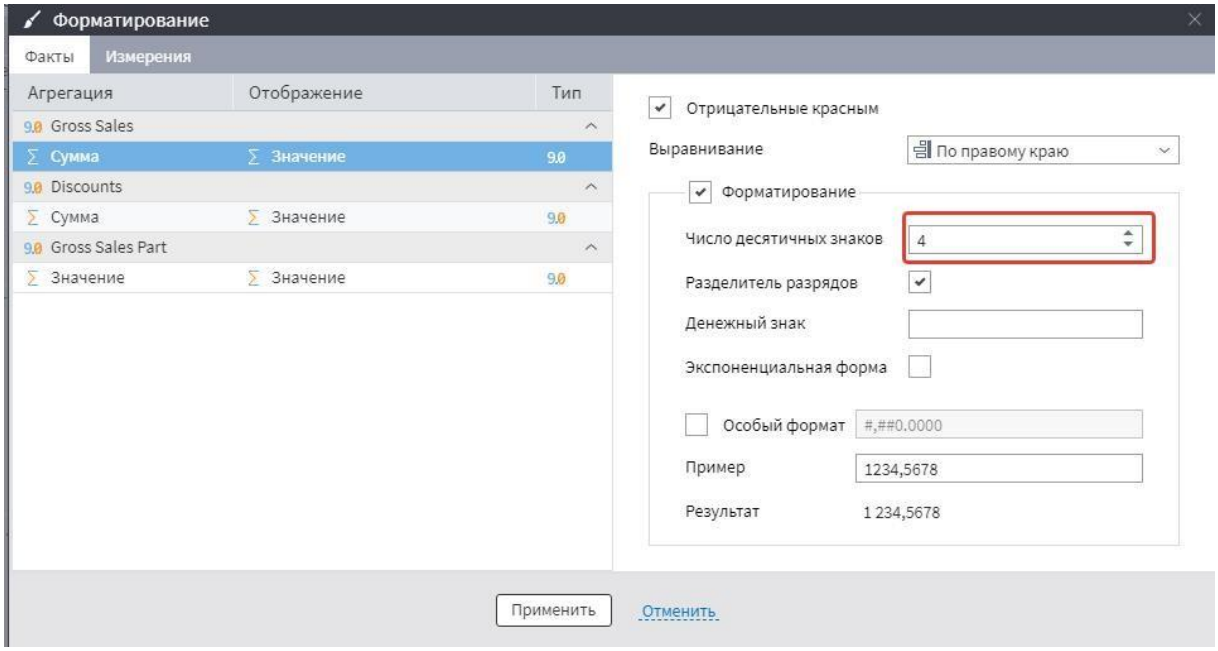


Рисунок 2.18 - Форматирование фактов

Выполним фильтрацию по измерению Country. Перейдем в контекстное меню для настройки фильтра, кликнув по этому измерению, расположенному в левом верхнем углу (рис. 2.19). Режим по умолчанию для фильтра – Множественный выбор – не изменяем. Кликнув мышкой на первой, третьей и пятой строке, делаем соответствующие значения показателя Country невидимыми при отображении (рис. 2.20).

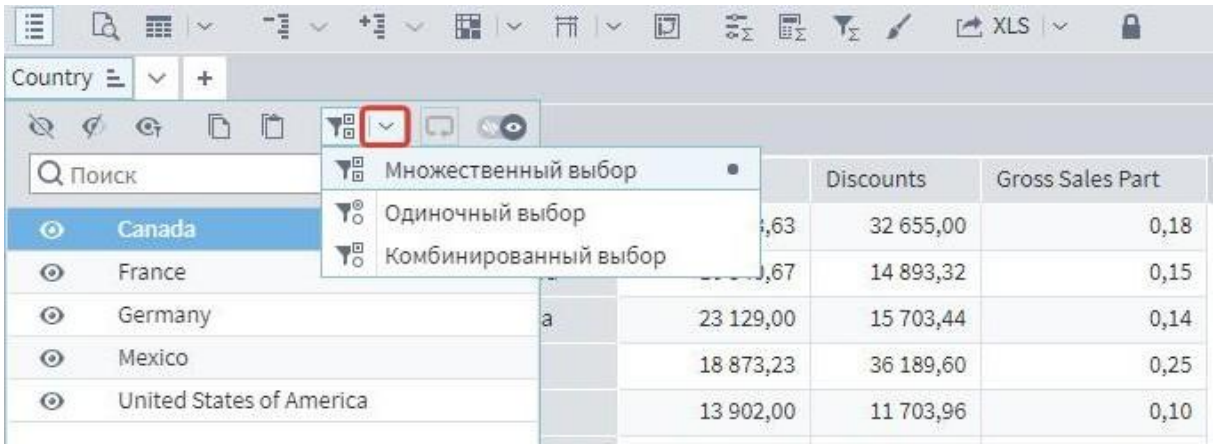


Рисунок 2.19 - Контекстное меню настройки фильтра

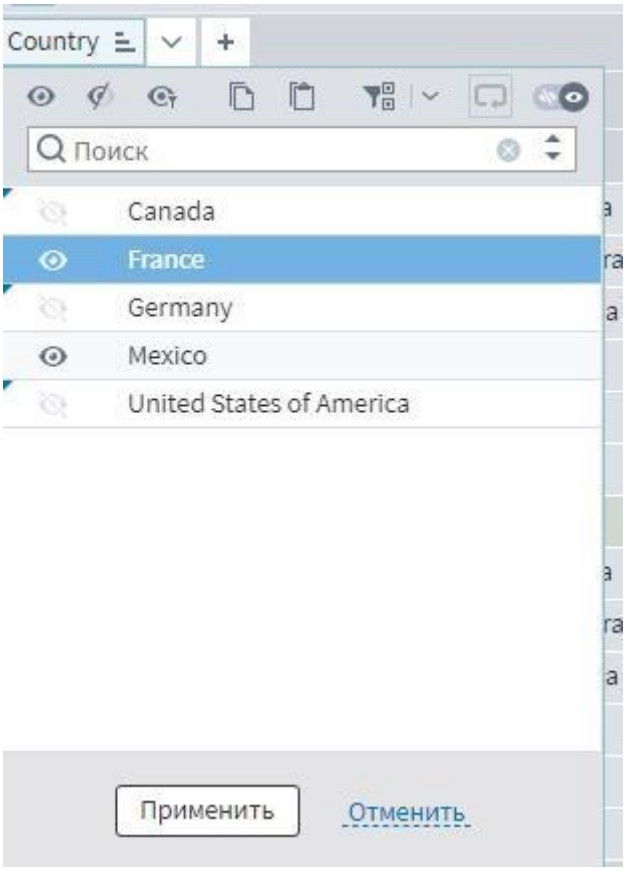
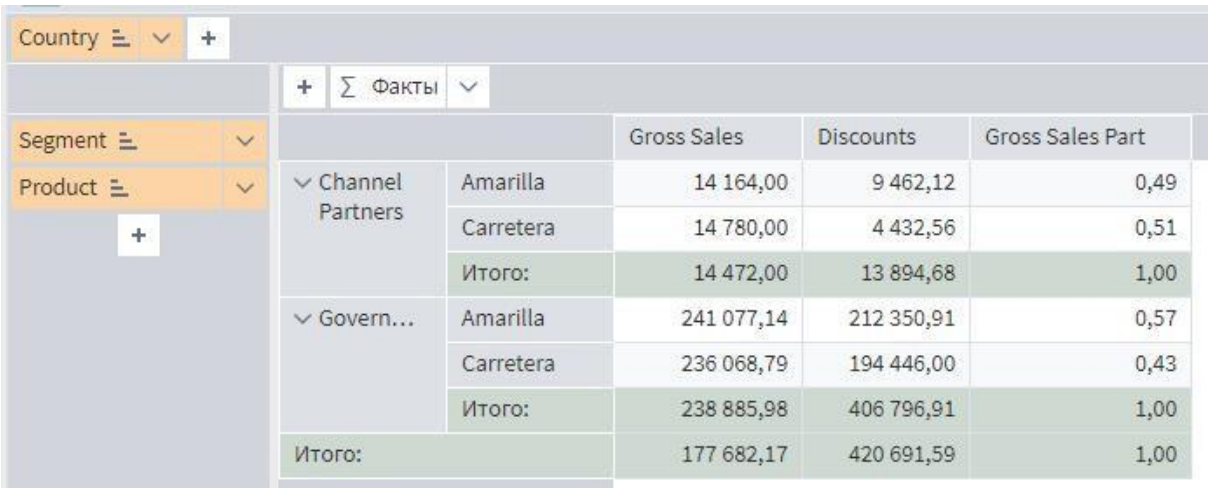


Рисунок 2.20 - Выбор значений измерения для построения среза данных

Жмем на кнопку Применить. Измерение Country будет окрашено в розовый цвет, который сигнализирует, что выбраны значения для получения среза данных.

Аналогично настроим фильтрацию для двух других измерений Segment и Product, используемых для группировки фактов. Для Segment оставим в рассмотрении только Channel Partners и Government, а для Product – только Amarilla и Carretera (рис. 2.21).



		Gross Sales	Discounts	Gross Sales Part
Channel Partners	Amarilla	14 164,00	9 462,12	0,49
	Carretera	14 780,00	4 432,56	0,51
	Итого:	14 472,00	13 894,68	1,00
Govern...	Amarilla	241 077,14	212 350,91	0,57
	Carretera	236 068,79	194 446,00	0,43
	Итого:	238 885,98	406 796,91	1,00
Итого:		177 682,17	420 691,59	1,00

Рисунок 2.21 - Фильтрация по измерениям

Отменим действие всех фильтров. Для этого зайдем в каждое измерение. Сделаем все значения измерения видимыми и нажмем кнопку Применить.

Настроим фильтрацию на основе фактов. В контекстном меню для фактов (рис. 2.13) выберем Фильтровать факты. В диалоговом окне выберем измерение Segment, а для факта Gross Sales – агрегацию Сумма и настроим условие фильтрации по факту (рис. 2.22).

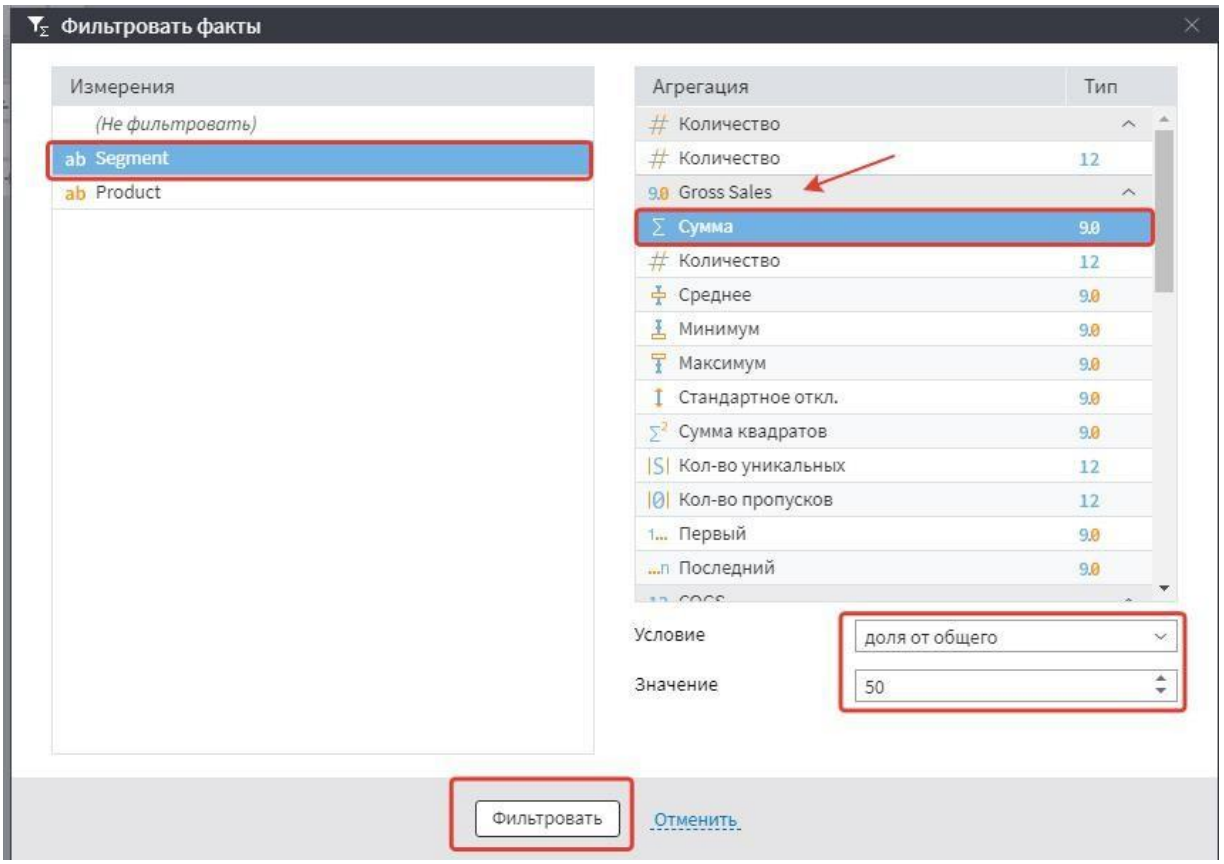


Рисунок 2.22 - Настройка условия отбора значений для факта

После выполнения фильтрации кнопка Факты окрасится в розовый цвет, а на экране останутся только те сегменты, которые удовлетворяют заданному условию, т.е. дают половину общих продаж (рис. 2.23).

Отменить фильтрацию можно выбрав уловие Не фильтровать слева вверху (рис. 2.22) и нажав на кнопку Фильтровать.

Добавим к сводной таблице диаграмму (значок расположен в строке пиктографического меню в верхнем правом углу). Проведем настройку отображения диаграммы (рис. 2.24). Уменьшим область отображения таблицы, чтобы увеличить доступное пространство экрана для диаграммы.

Country

Segment

Product

Факты

		Gross Sales	Discounts	Gross Sales Part
Govern...	Amarilla	10 532 086,50	589 187,39	0,19
	Carretera	6 430 801,00	349 856,92	0,11
	Montana	6 088 609,00	539 672,98	0,11
	Paseo	16 253 973,00	1 371 742,30	0,29
	Velo	8 347 373,00	533 950,95	0,15
	VTT	8 750 224,00	514 395,29	0,16
	Итого:	56 403 066,50	3 898 805,83	1,00
Small Business	Amarilla	5 024 400,00	430 119,00	0,11
	Carretera	4 275 000,00	364 488,00	0,09
	Montana	7 117 950,00	443 011,50	0,15
	Paseo	12 321 000,00	822 190,50	0,27
	Velo	7 172 250,00	764 272,50	0,16
	VTT	10 031 100,00	689 700,00	0,22
	Итого:	45 941 700,00	3 513 781,50	1,00
Итого:		102 344 766,50	7 412 587,33	1,00

Рисунок 2.23 - Построение среза для факта

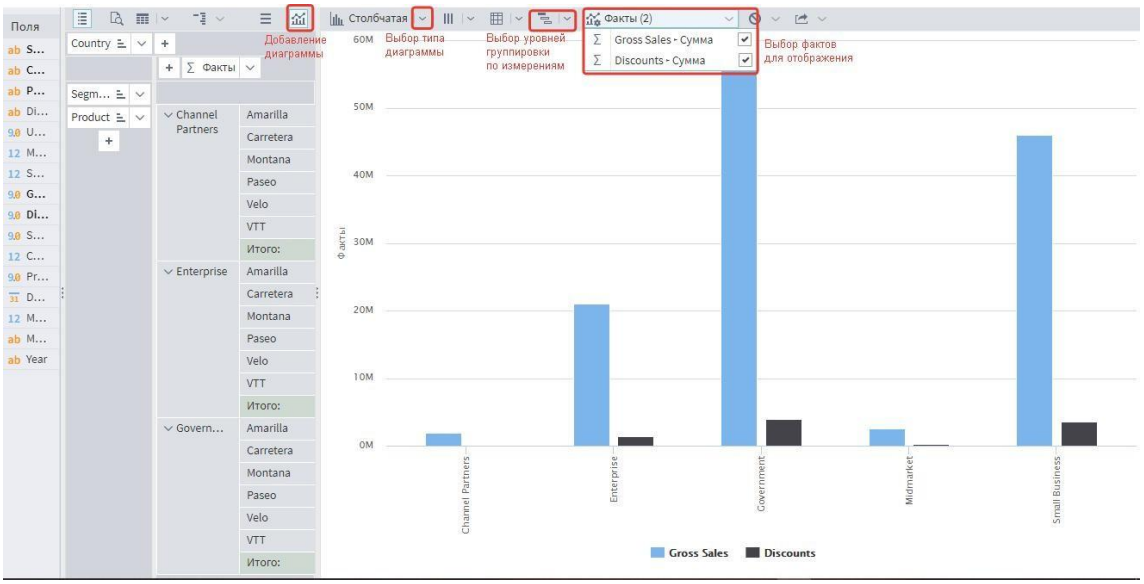


Рисунок 2.24 - Настройка диаграммы
для визуализации данных сводной таблицы

Сохраним визуализацию и перейдем на уровень вверх в рабочую область Визуализаторы (рис. 2.25).

Построенный визуализатор Куб можно добавить в отчеты (рис. 2.26).

В пакете Анализ_продаж в разделе Отчеты появится визуализатор Куб, переместиться в который можно с помощью панели навигации (рис. 2.27).

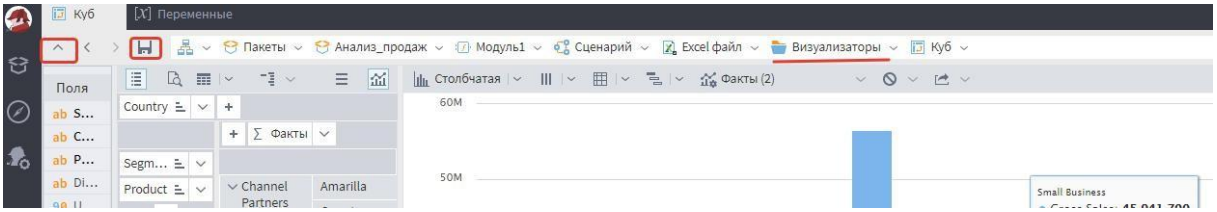


Рисунок 2.25 - Сохранение и переход в Визуализаторы

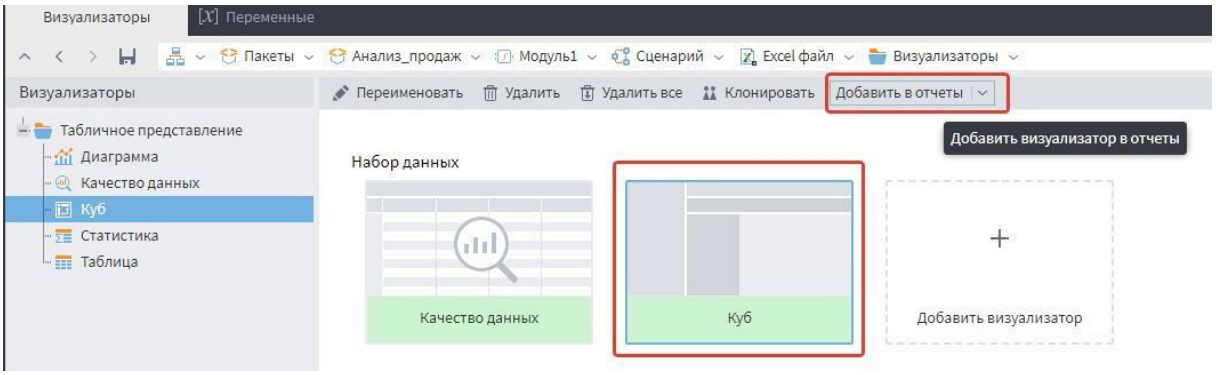


Рисунок 2.26 - Добавление визуализатора Куб в Отчеты

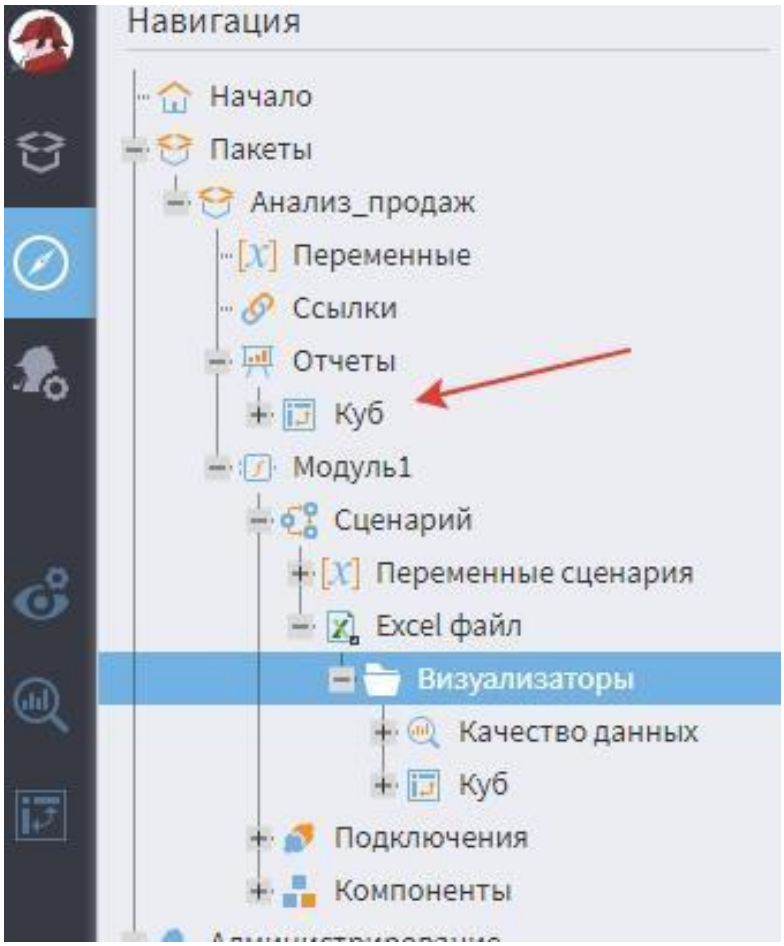


Рисунок 2.27 - Навигация по элементам пакета

Добавим в рабочую область элемента Визуализаторы компонент Статистика и войдем в него (рис. 2.28).

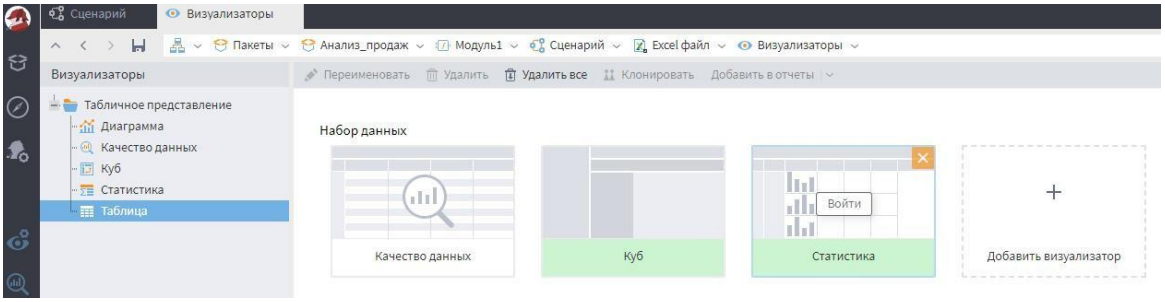


Рисунок 2.28 - Визуализатор Статистика

В этом визуализаторе настроим поля для сводки по описательным статистикам набора данных, так как по умолчанию включаются все доступные поля (рис. 2.29, 2.30).

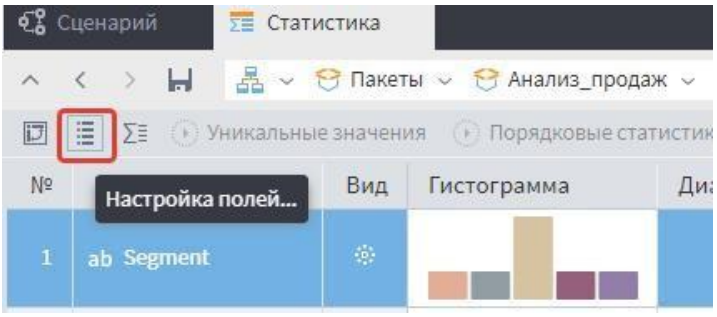


Рисунок 2.29 - Настройка полей

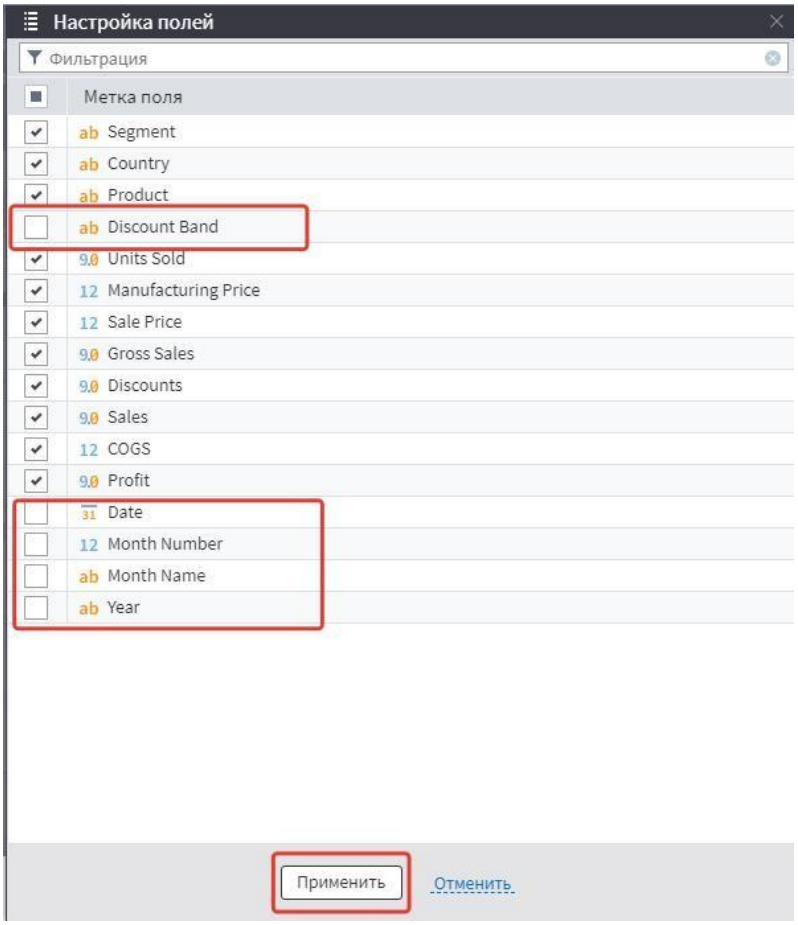


Рисунок 2.30 - Выбор полей для отображения в визуализаторе Статистика

Далее настроим отображаемые показатели в сводке по описательным статистикам (рис. 2.31, 2.32).

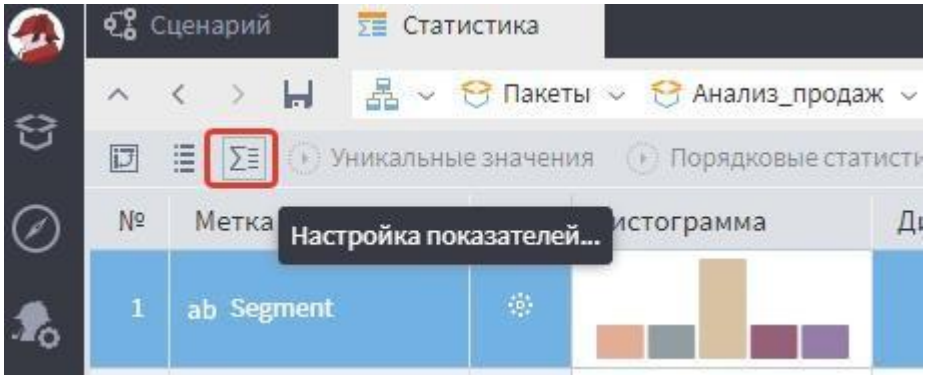


Рисунок 2.31 - Настройка показателей

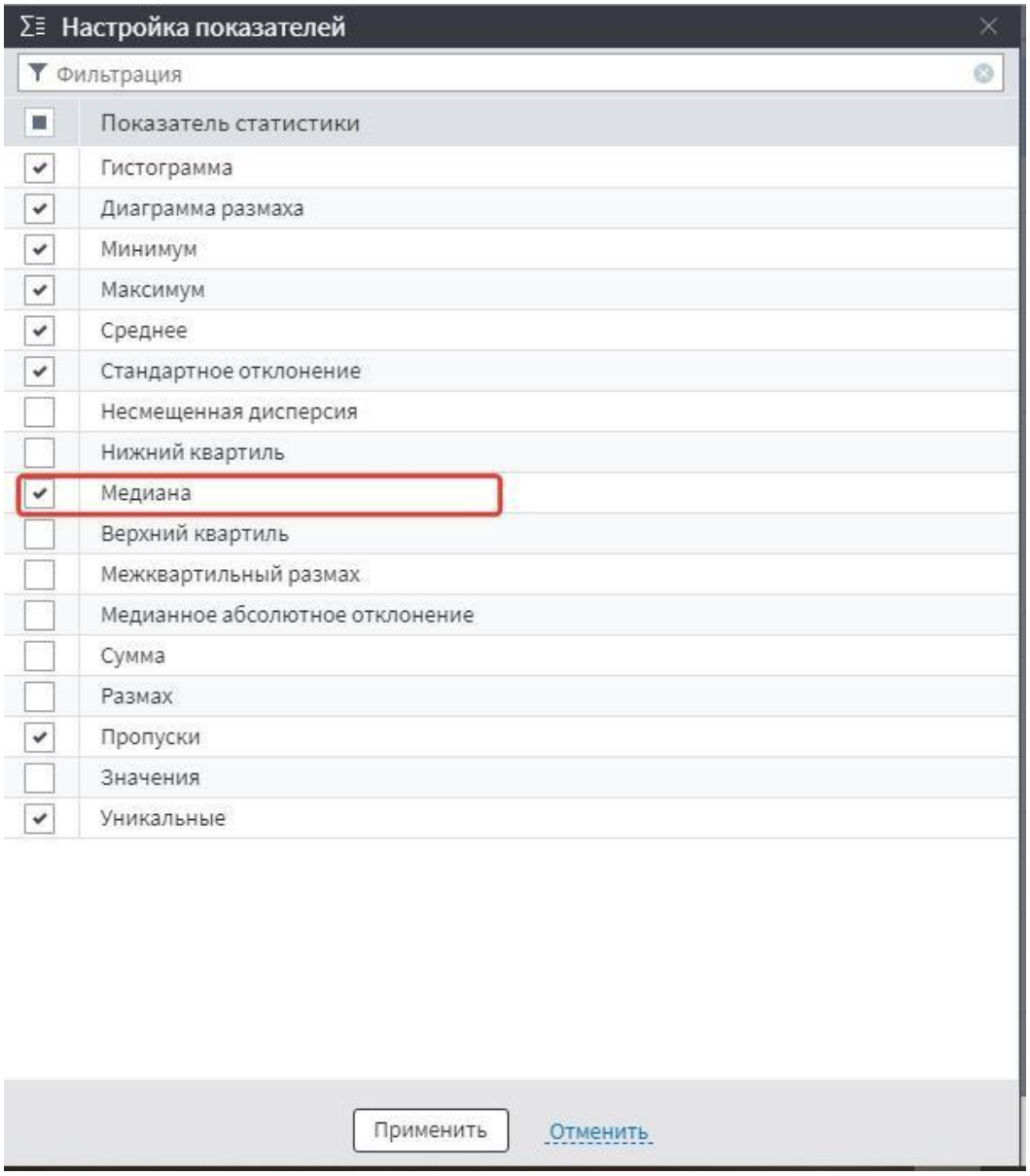
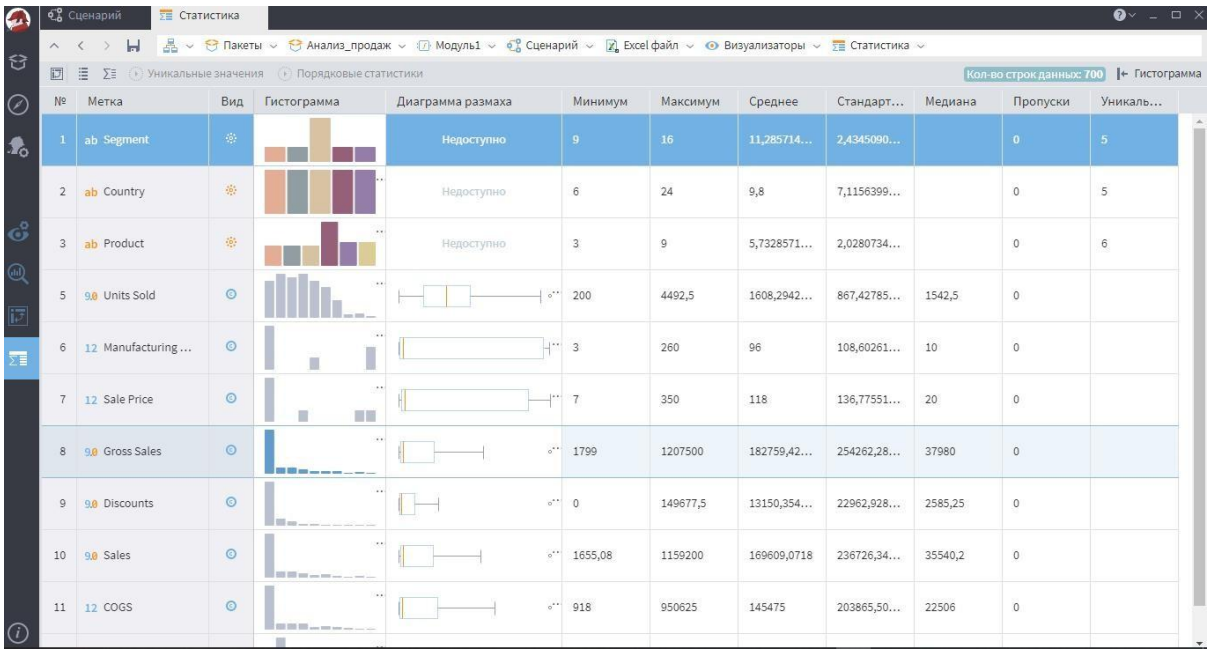


Рисунок 2.32 - Добавление к показателям, рассчитываемым по умолчанию, Медианы

В результате применения настроек визуализатор Статистика примет вид, как на рис. 2.33.



№	Метка	Вид	Гистограмма	Диаграмма размаха	Минимум	Максимум	Среднее	Стандарт...	Медиана	Пропуски	Уникаль...
1	ab Segment			Недоступно	9	16	11,285714...	2,4345090...		0	5
2	ab Country			Недоступно	6	24	9,8	7,1156399...		0	5
3	ab Product			Недоступно	3	9	5,7328571...	2,0280734...		0	6
5	9.0 Units Sold				200	4492,5	1608,2942...	867,42785...	1542,5	0	
6	12 Manufacturing ...				3	260	96	108,60261...	10	0	
7	12 Sale Price				7	350	118	136,77551...	20	0	
8	9.0 Gross Sales				1799	1207500	182759,42...	254262,28...	37980	0	
9	9.0 Discounts				0	149677,5	13150,354...	22962,928...	2585,25	0	
10	9.0 Sales				1655,08	1159200	169609,0718	236726,34...	35540,2	0	
11	12 COGS				918	950625	145475	203865,50...	22506	0	

Рисунок 2.33 - Визуализатор Статистика

Вернемся в Сценарий с помощью навигации. Добавим в рабочую область узел Параметры полей (рис. 2.34).

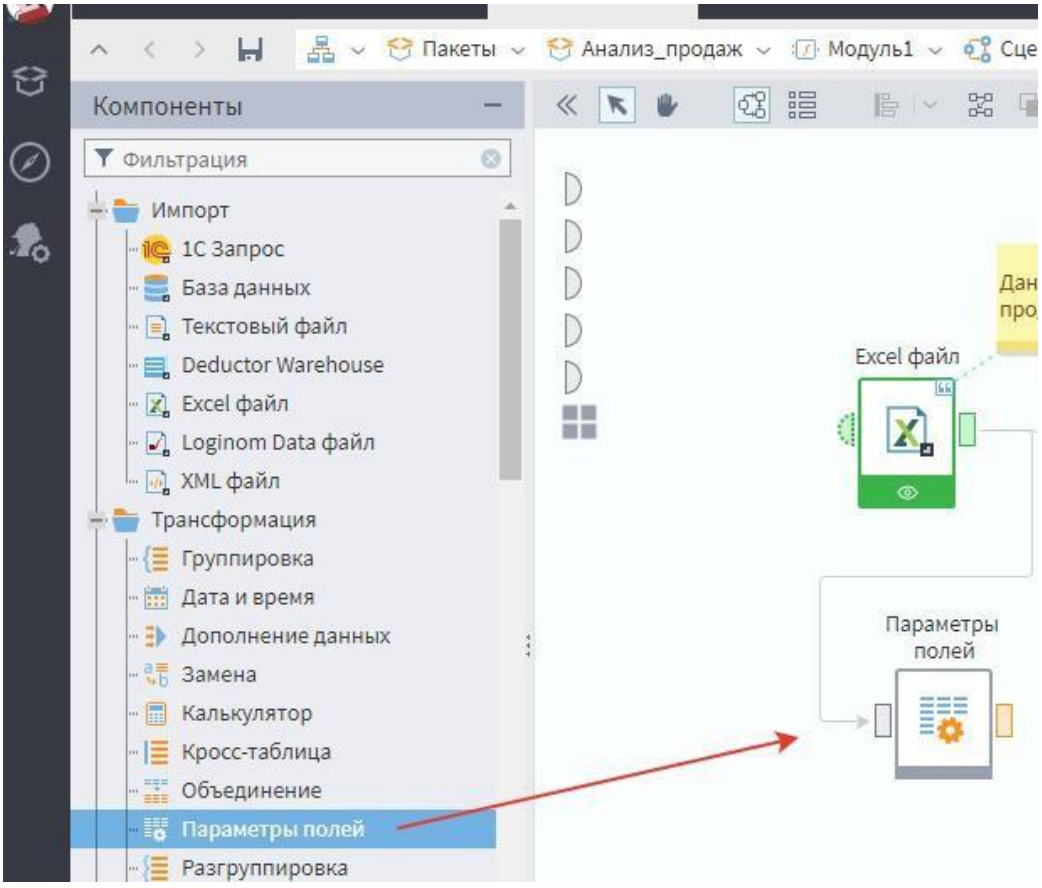


Рисунок 2.34 - Добавление узла Параметры полей

Перейдем в настройки этого узла (шестеренка внутри узла) и создадим набор данных, в котором останется только одно поле Date, содержащее все нужные составляющие момента времени осуществления продажи (рис. 2.35).

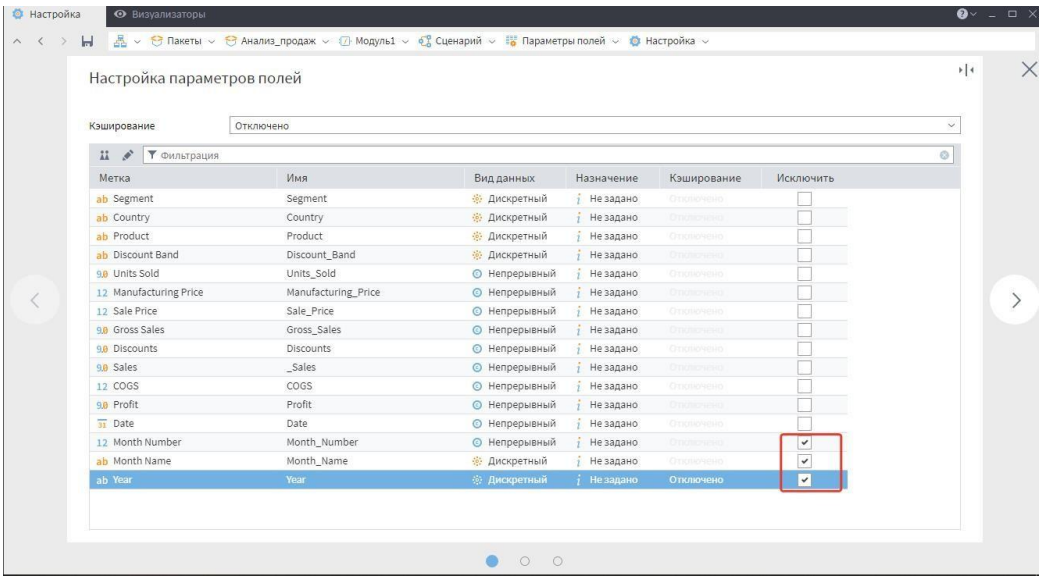


Рисунок 2.35 - Исключение полей

Сохраняем настройки узла и добавляем в рабочую область Сценария новый узел Дата и время (рис. 2.36).

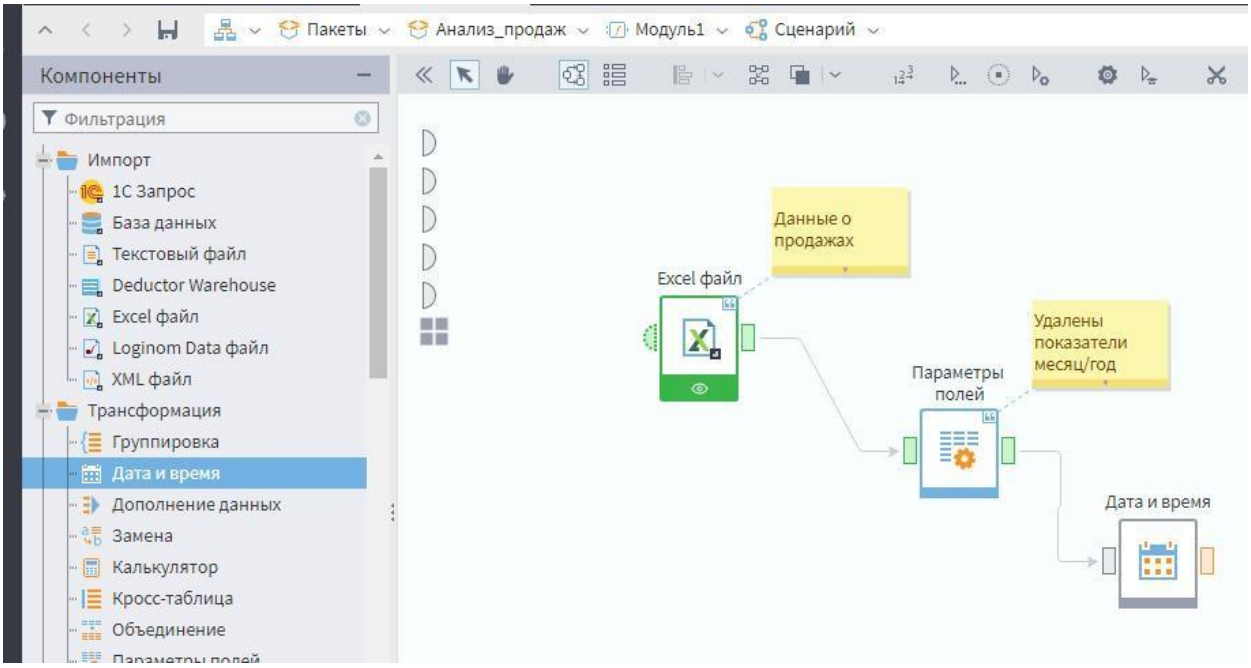


Рисунок 2.36 - Добавление узла Дата и время

Переходим к настройкам этого узла и выбираем в диалоговом окне Преобразование даты/времени представление в виде Год+Месяц (рис. 2.37).

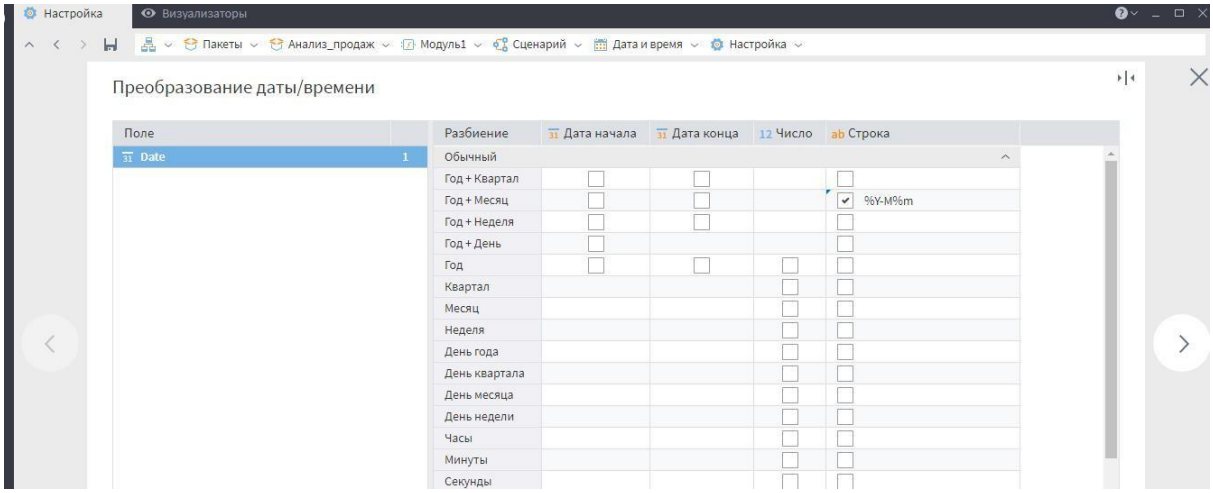


Рисунок 2.37 - Преобразование даты/времени

Сохраняем настройки и переходим в визуализаторы этого узла. Добавляем визуализатор Куб, настройки для которого представлены на рис. 2.38.

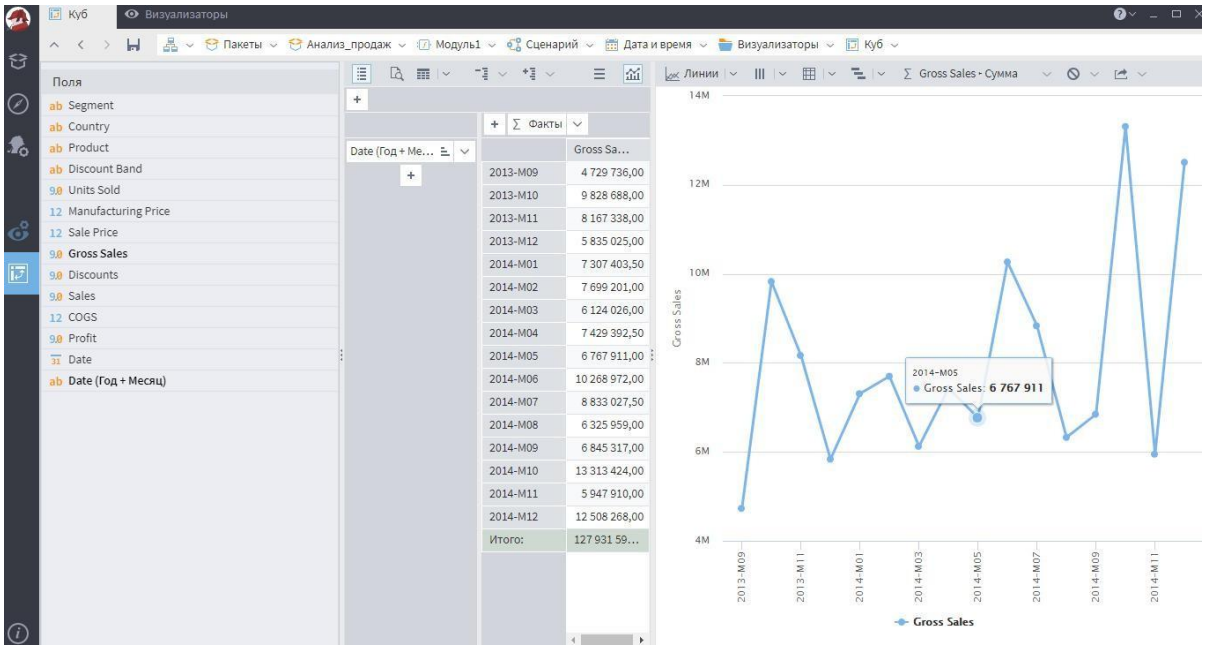


Рисунок 2.38 - Визуализатор Куб для узла Дата и время

Получили динамический ряд ежемесячного объема продаж. Сохраняем визуализатор и возвращаемся в Сценарий (рис. 2.39).

Итоговый сценарий для пакета Анализ_продаж состоит из трех узлов, для двух из которых настроены визуализаторы (нижняя полоска в прямоугольниках этих узлов содержит стереотип визуализатора).

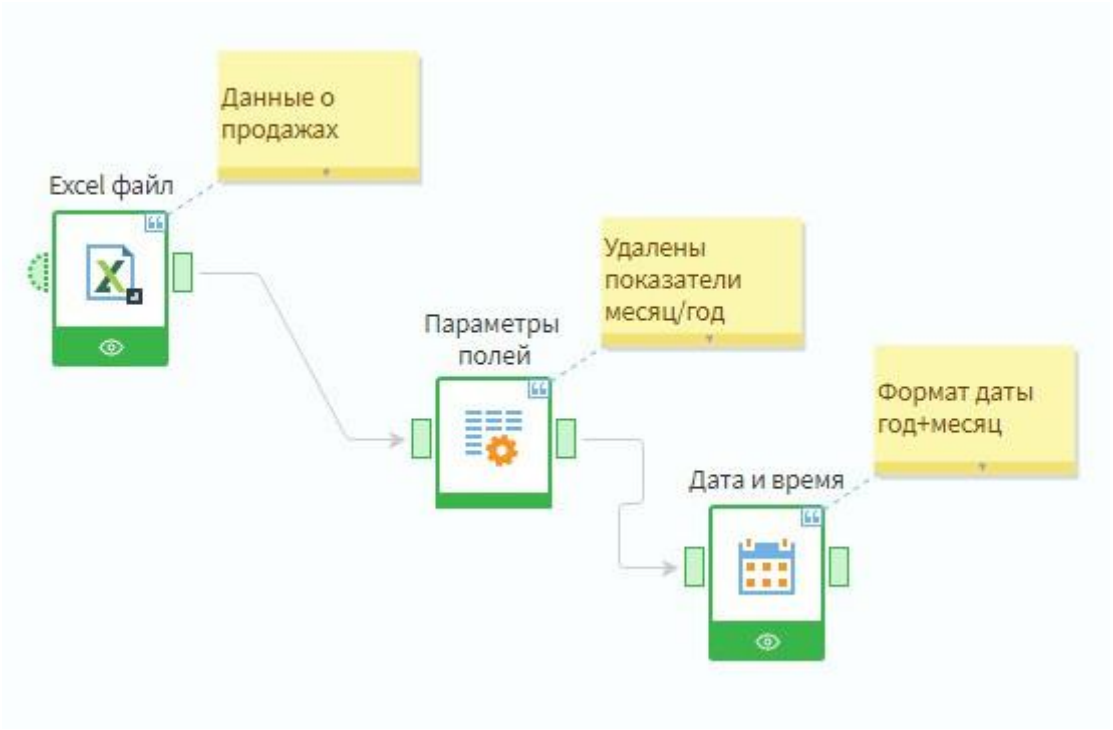


Рисунок 2.39 - Сценарий пакета Анализ_продаж

Контрольные вопросы

1. Охарактеризуйте понятие бизнес-аналитики.
2. Какие бизнес-задачи можно решать, используя платформу Loginom, приведите примеры.
3. Как установить бесплатную версию Loginom Community.
4. Какие визуализаторы используются в Loginom?

Лабораторная работа 2. Кластеризация

Цель работы: ознакомиться с возможностями Loginom по кластеризации данных.

Содержание работы:

Кластеризация (сегментация) – это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от других, которые вошли в другие кластеры.

В Loginom обработчик производит кластеризацию объектов на основе алгоритмов k-means и g-means. Если количество кластеров известно, то применяется алгоритм k-means, в противном случае – g-means, который определяет это количество автоматически в рамках заданного интервала [6].

Создадим новый пакет Медицинское_страхование. В Сценарий из категории Импорт добавим первый узел Текстовый файл (рис. 3.1).

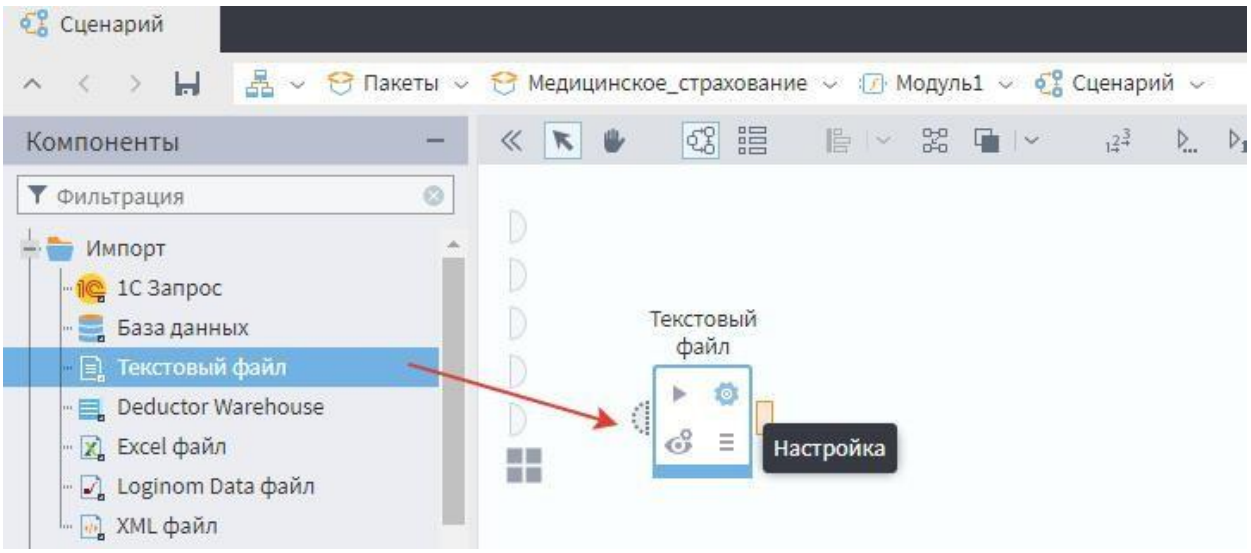


Рисунок 3.1 - Узел Текстовый файл

Прежде чем приступить к настройке этого узла, скачаем на локальный диск данные. Это набор данных о расходах на медицинское обслуживание тех, кто имеет медицинскую страховку: https://www.kaggle.com/mirichoi0218/insurance.

На ресурсе kaggle требуется регистрация для скачивания файлов данных (рис. 3.2). Зарегистрироваться можно с учетной записью ресурсов Google @gmail.com.

Файл скачается как архив, из которого его нужно извлечь перед импортом в Loginom.

Описание переменных набора:

– age: возраст основного бенефициара;

– sex: пол застрахованного;

– bmi: индекс массы тела;

– children: число детей, охваченных медицинским страхованием / число иждивенцев;

– smoker: курит ли застрахованный;

– region: жилой район получателя в США, Северо-Восток, Юго-Восток, Юго-Запад, Северо-Запад;

– charges: индивидуальные медицинские расходы, оплачиваемые страховкой.

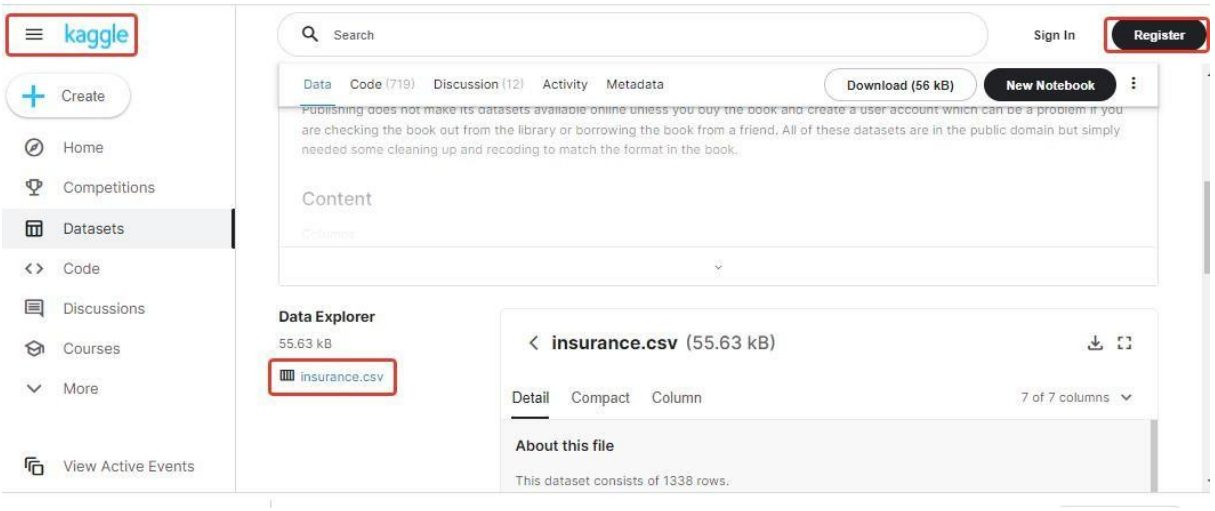


Рисунок 3.2 - Ресурс kaggle

Проведем настройку узла Текстовый файл: выполним импорт (рис. 3.3).

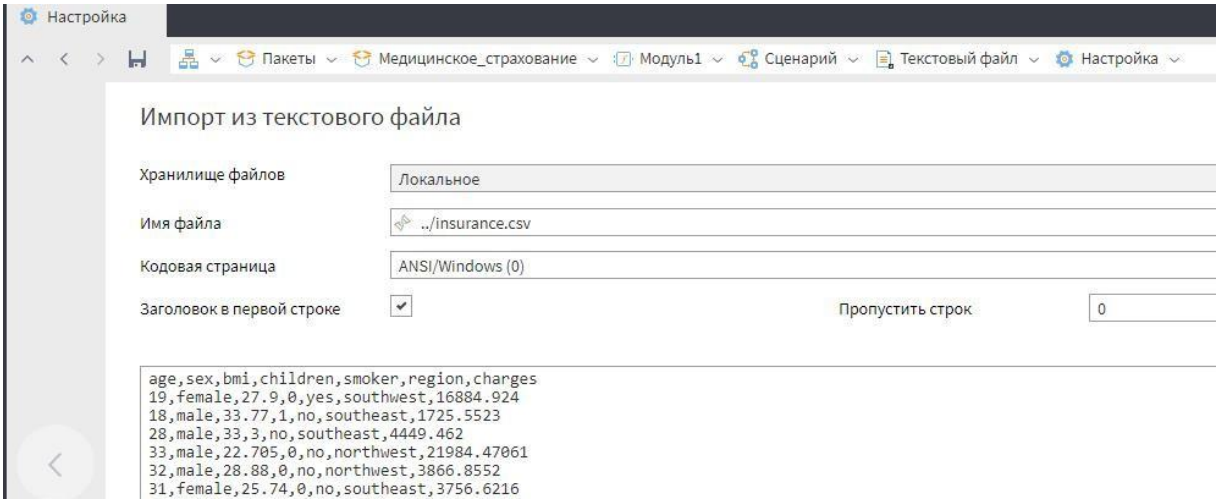


Рисунок 3.3 - Импорт

При настройке форматов импорта два поля bmi и charges определены с ошибочным типом данных (рис. 3.4). Это связано с десятичным разделителем, который в импортируемых данных точка, а по умолчанию мастер настройки ожидает запятую.

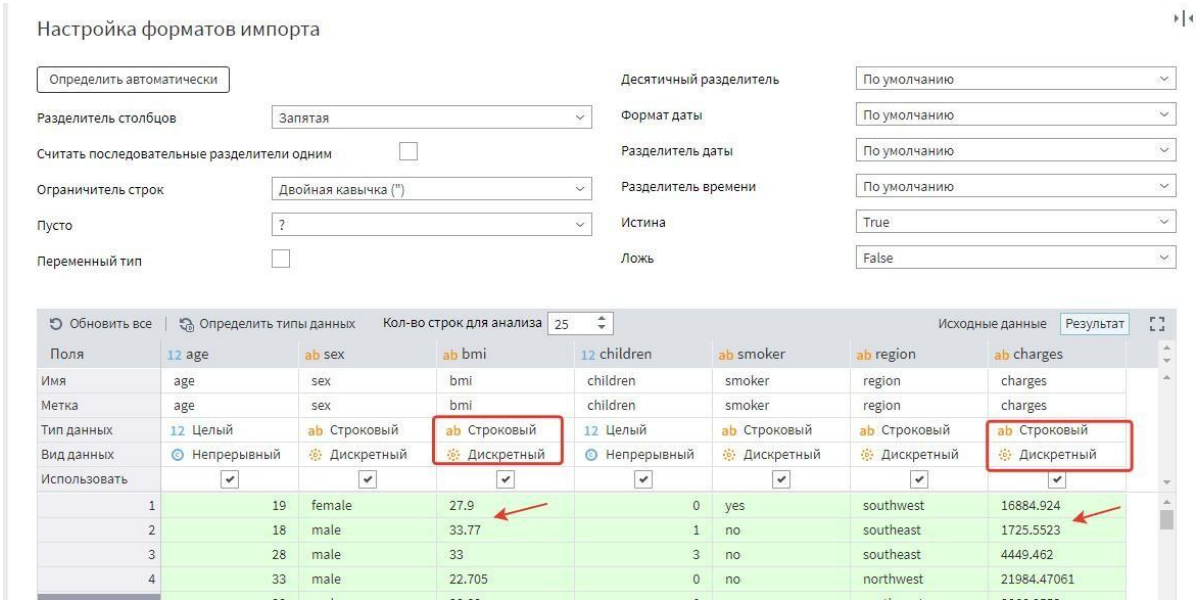


Рисунок 3.4 - Настройка форматов импорта

Выберем в качестве десятичного разделителя точку и кликнем по кнопке Определять автоматически. Теперь тип данных для этих полей будет настроен верно (рис. 3.5).

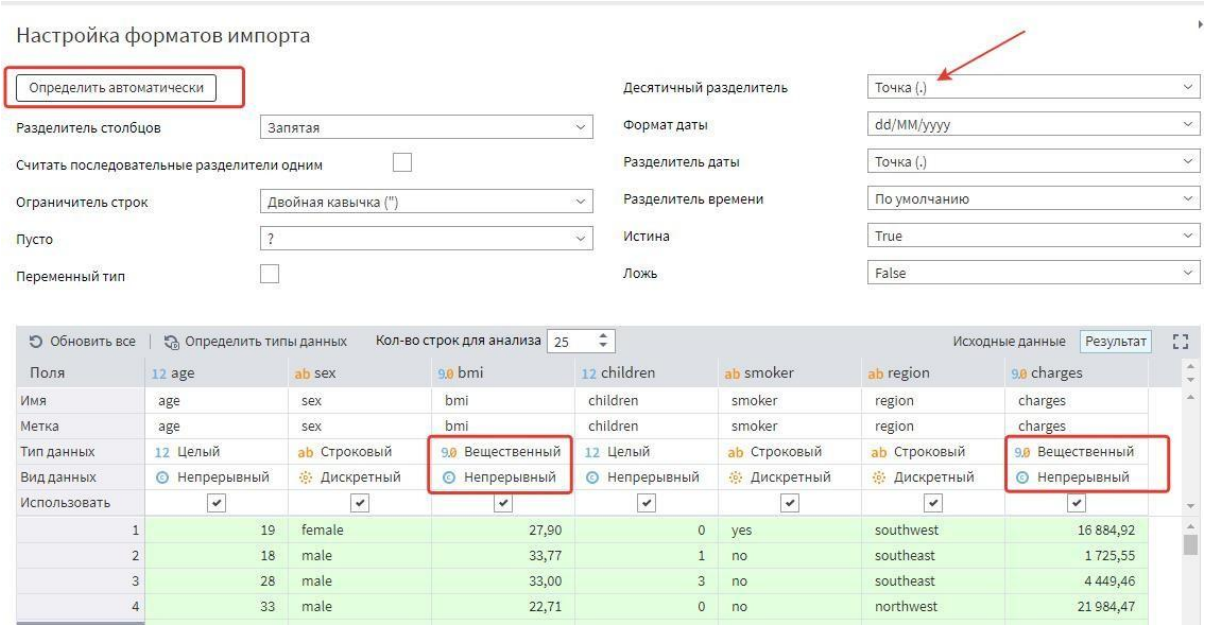


Рисунок 3.5 - Настройка форматов импорта

Завершим настройку, сохраним ее и добавим в рабочую область Сценария узел Кластеризация (рис. 3.6).

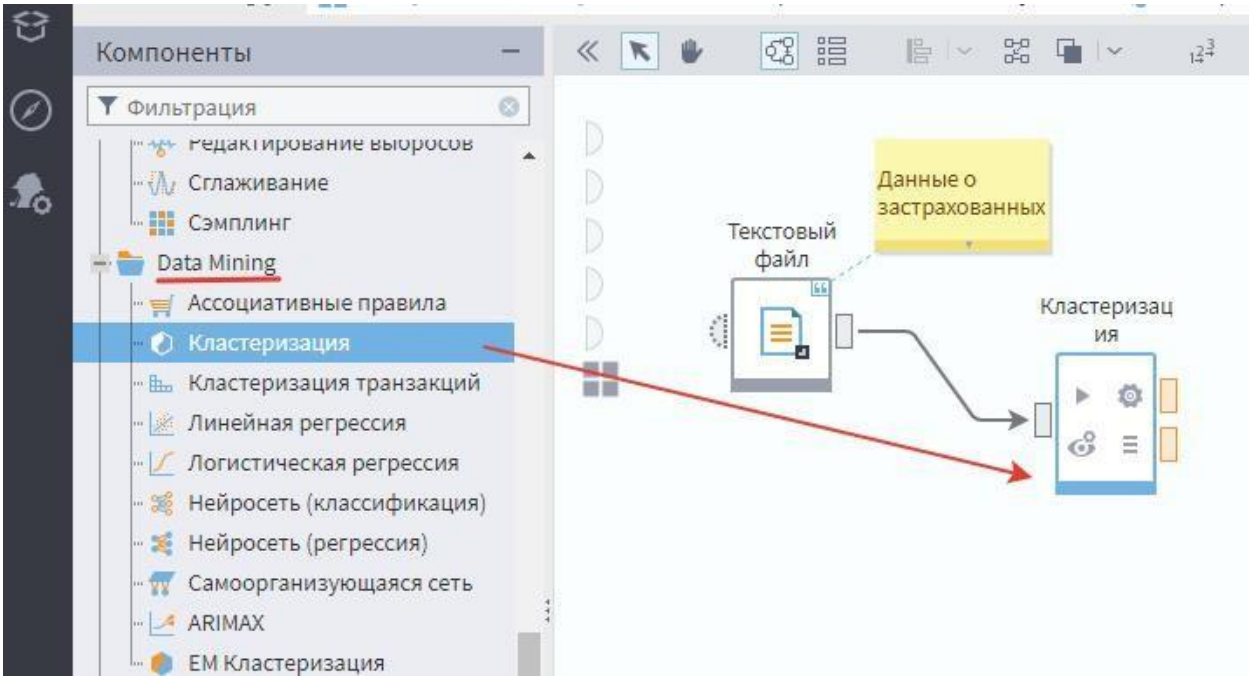


Рисунок 3.6 - Добавление узла Кластеризация

Перейдем к настройке узла Кластеризация. В первом окне необходимо произвести настройку входных столбцов (рис. 3.7). Кликнув мышкой на ячейке Не задано в столбце Назначение для тех полей, по значениям которых должна быть произведена кластеризация, изменим ее значение на Используемое (рис. 3.8).

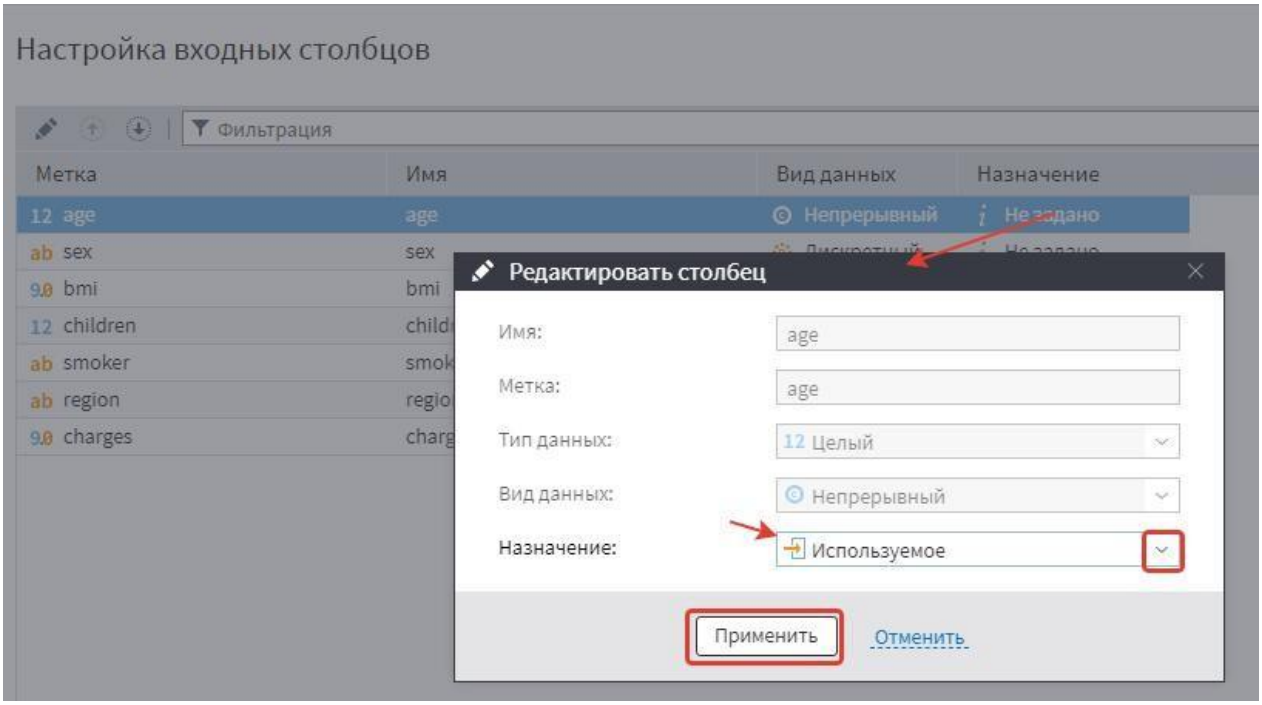


Рисунок 3.7 - Отбор столбцов для проведения кластеризации

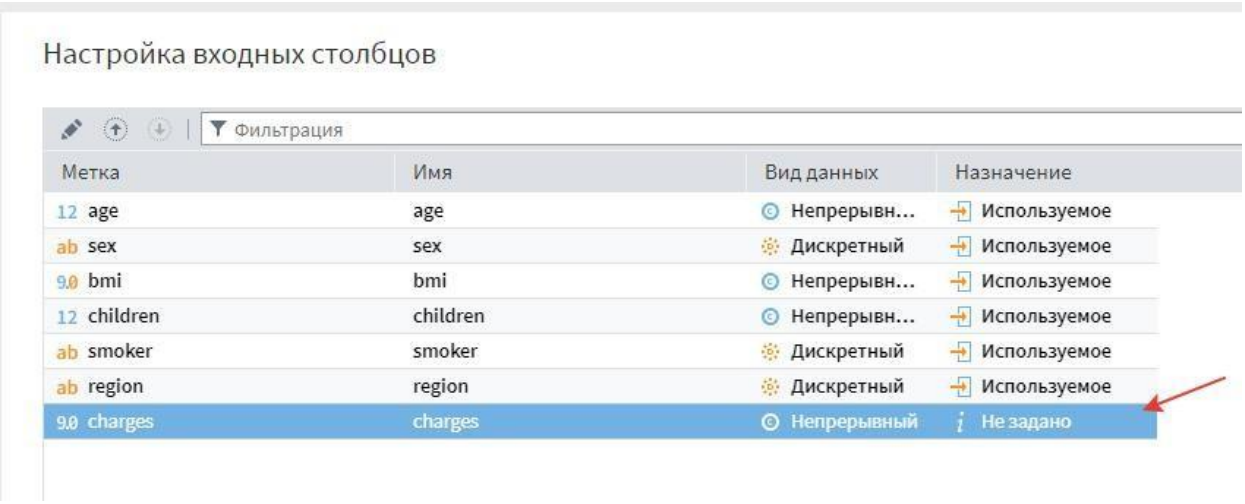


Рисунок 3.8 - Итог настройки входных столбцов

Только столбец charges не используется для кластеризации. Настройку нормализации пропустим. Для параметров кластеризации применим следующие настройки: уберем галочку с автоопределения числа кластеров, число кластеров определим равным 3 (рис. 3.9). В кластеризации будет реализован алгоритм k-means. Для алгоритма g-means нужно было оставить настройки в этом диалоговом окне по умолчанию, т.е. автоопределение числа кластеров.

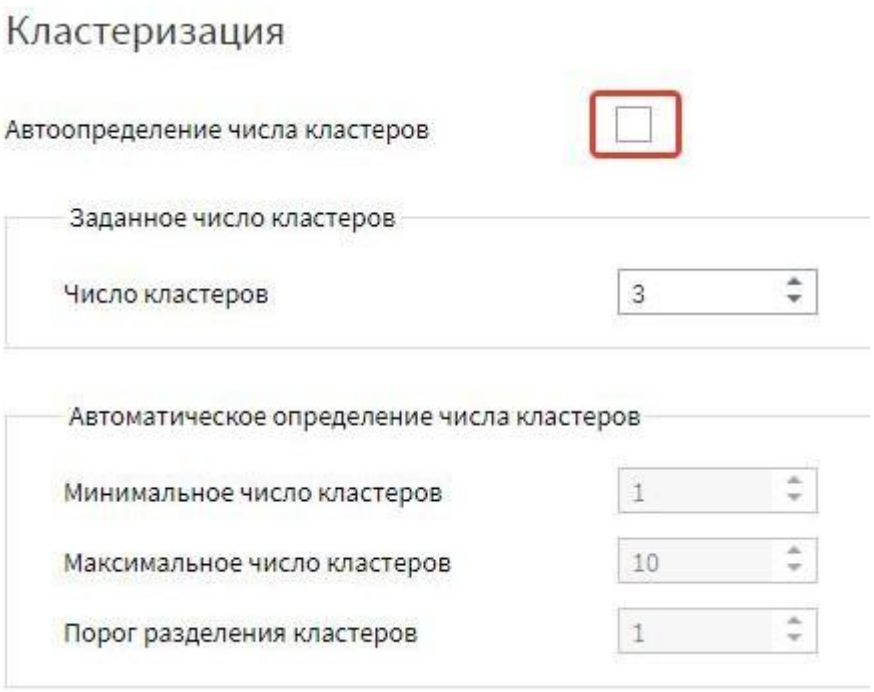


Рисунок 3.9 - Настройка параметров кластеризации

Сохраним настройки для кластеризации и в рабочей области вызовем контекстное меню, кликнув правой кнопкой мышки на узле Кластеризация (рис. 3.10).

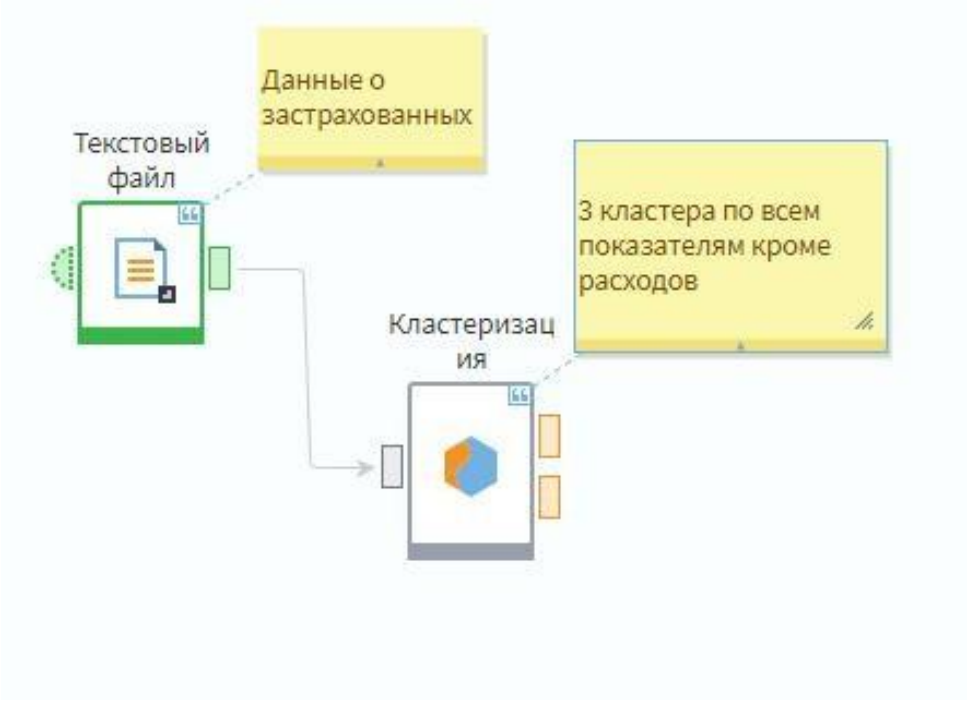


Рисунок 3.10 - Рабочая область Сценария

В контекстном меню выберем опцию Переобучить узел (рис. 3.11).

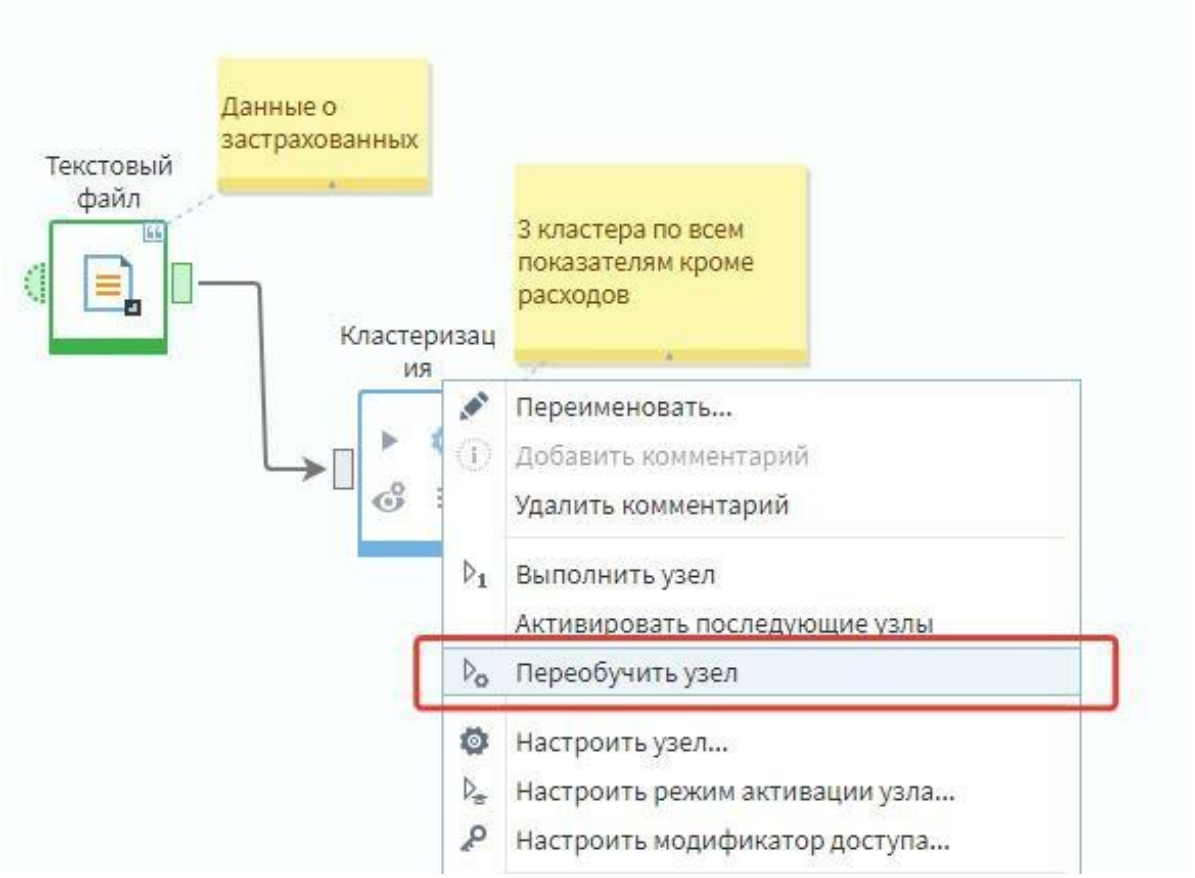


Рисунок 3.11 - Переобучить узел

После выполнения этой операции станут доступны выходные данные узла (рис. 3.12, 3.13).

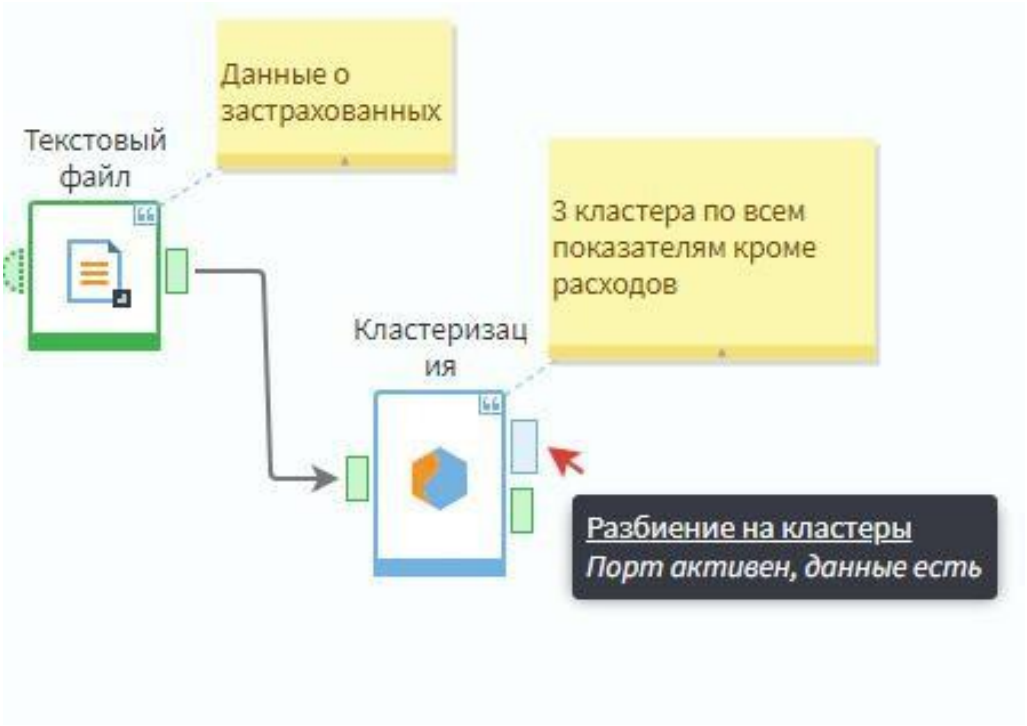


Рисунок 3.12 - Выходной порт с данными разбиения на кластеры

Рисунок 3.13 - Выходной порт с данными центров кластеров

Выходные данные при клике мышки на соответствующем выходном порте отображаются в нижней части рабочей области. Но более информативными будут визуализаторы, которые можно настроить для узла Кластеризация (рис. 3.14).

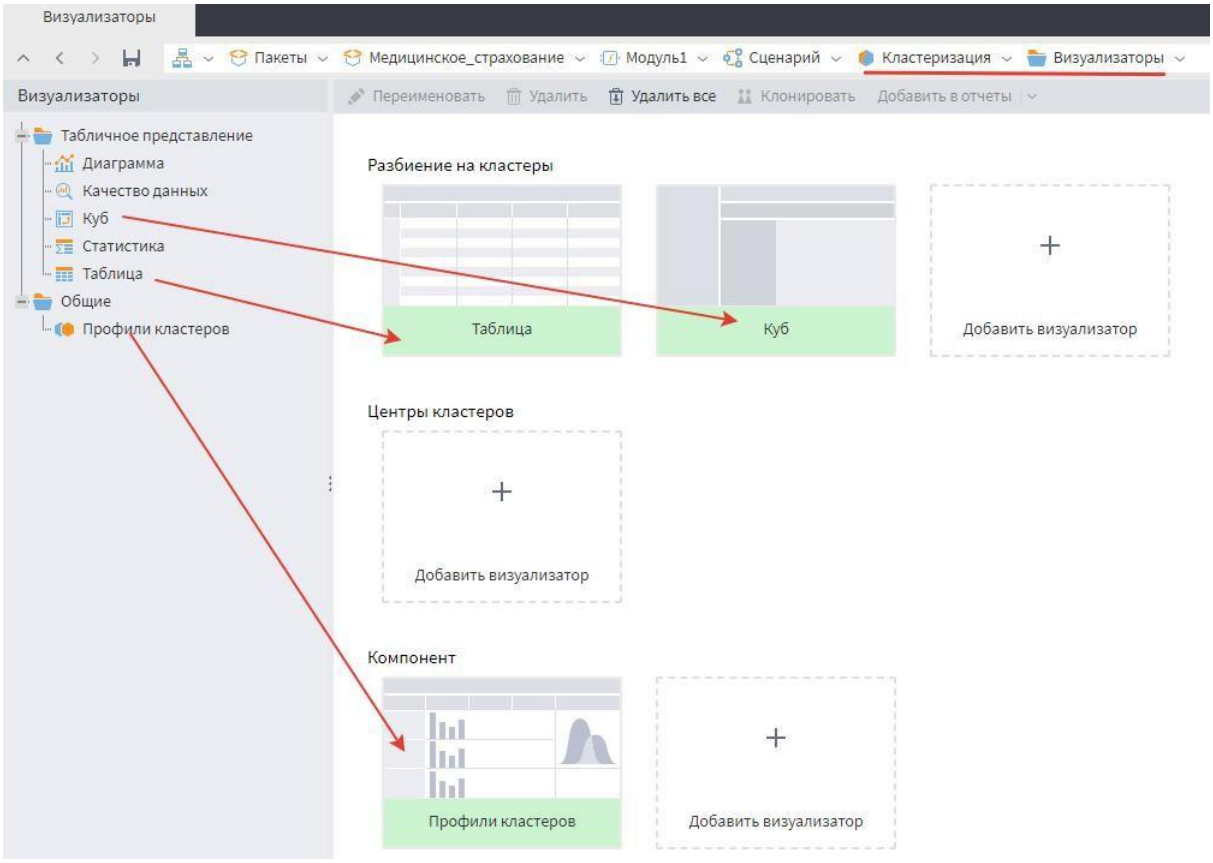


Рисунок 3.14 - Добавление визуализаторов

Настроим первый визуализатор Таблица. Проведем сортировку с использованием нескольких полей, установив иерархию: номер кластера, затем упорядочиваем внутри каждого кластера по убыванию возраста, а затем среди застрахованных одного года рождения по убыванию расходов на медицинское обслуживание (рис. 3.15).

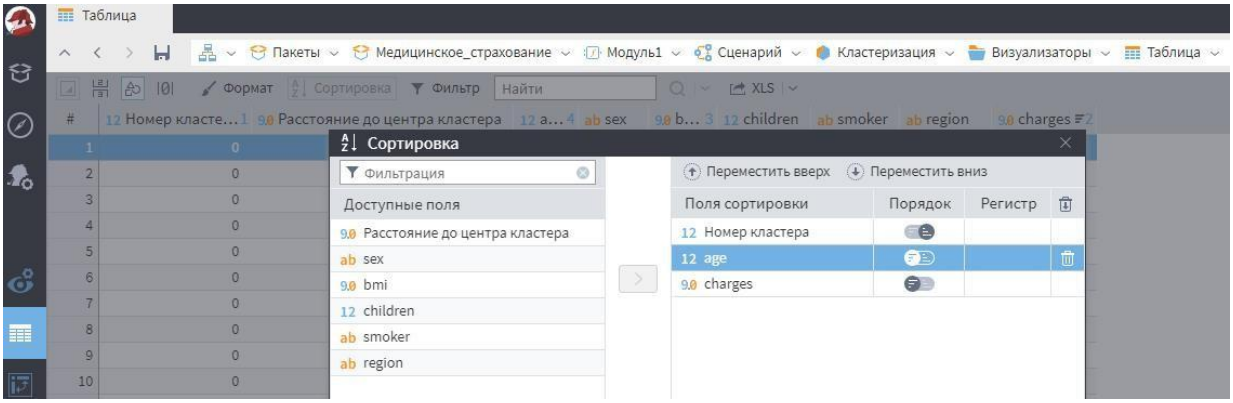


Рисунок 3.15 - Настройка сортировки по нескольким полям

Если необходимо убрать с экрана какой-либо столбец таблицы, то вызываем контекстное меню по стрелке рядом с именем столбца и убираем галочку в списке столбцов для отображения (рис. 3.16).

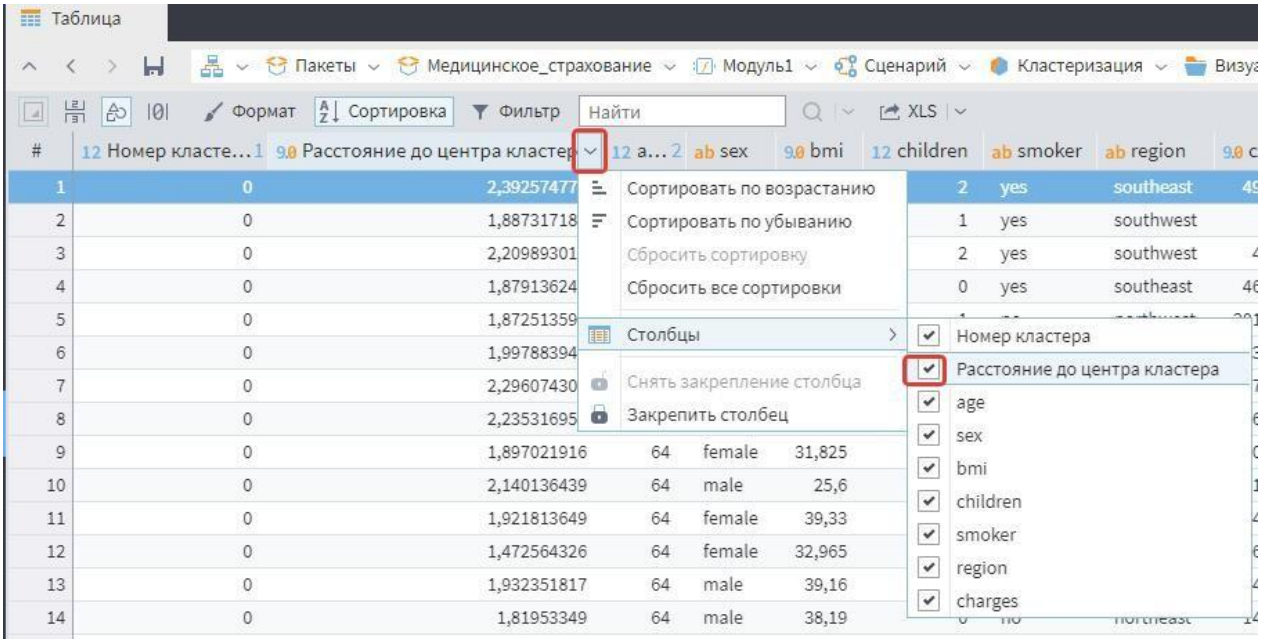
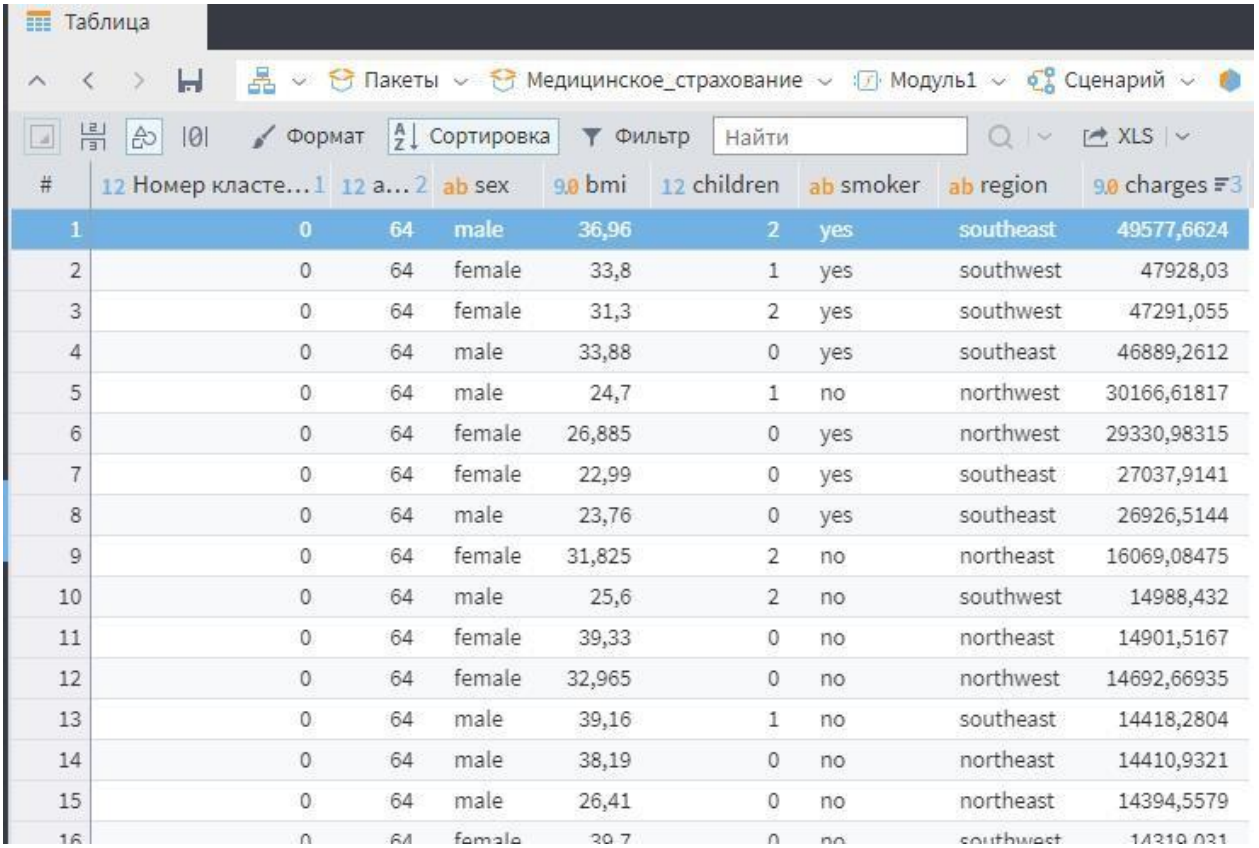


Рисунок 3.16 - Выбор столбцов для отображения

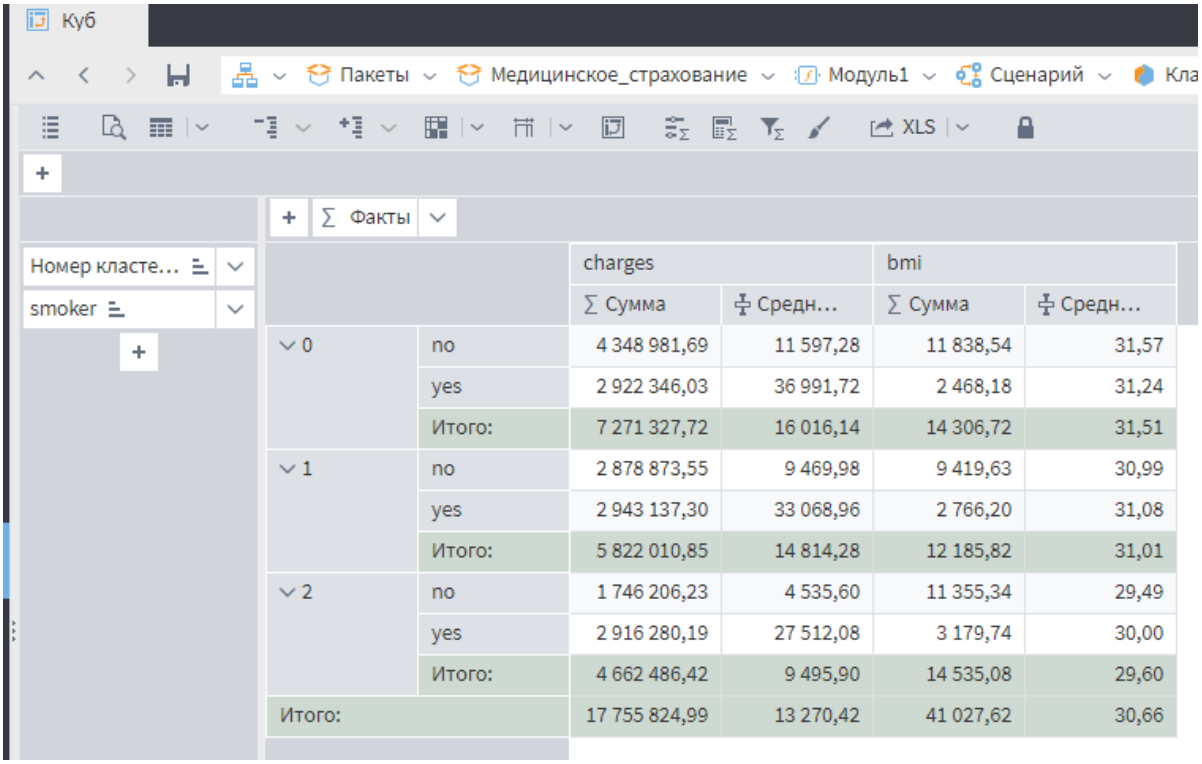
Итоговой результат настройки визуализатора Таблица приведен на рис. 3.17.



#	Номер кластера	age	sex	bmi	children	smoker	region	charges
1	0	64	male	36,96	2	yes	southeast	49577,6624
2	0	64	female	33,8	1	yes	southwest	47928,03
3	0	64	female	31,3	2	yes	southwest	47291,055
4	0	64	male	33,88	0	yes	southeast	46889,2612
5	0	64	male	24,7	1	no	northwest	30166,61817
6	0	64	female	26,885	0	yes	northwest	29330,98315
7	0	64	female	22,99	0	yes	southeast	27037,9141
8	0	64	male	23,76	0	yes	southeast	26926,5144
9	0	64	female	31,825	2	no	northeast	16069,08475
10	0	64	male	25,6	2	no	southwest	14988,432
11	0	64	female	39,33	0	no	northeast	14901,5167
12	0	64	female	32,965	0	no	northwest	14692,66935
13	0	64	male	39,16	1	no	southeast	14418,2804
14	0	64	male	38,19	0	no	northeast	14410,9321
15	0	64	male	26,41	0	no	northeast	14394,5579

Рисунок 3.17 - Визуализатор Таблица для узла Кластеризация

Сохраним визуализатор Таблица, перейдем в настройки визуализатора Куб. Добавим измерения и факты для Куба (рис. 3.18).



		charges		bmi	
		∑ Сумма	Средн...	∑ Сумма	Средн...
0	no	4 348 981,69	11 597,28	11 838,54	31,57
	yes	2 922 346,03	36 991,72	2 468,18	31,24
	Итого:	7 271 327,72	16 016,14	14 306,72	31,51
1	no	2 878 873,55	9 469,98	9 419,63	30,99
	yes	2 943 137,30	33 068,96	2 766,20	31,08
	Итого:	5 822 010,85	14 814,28	12 185,82	31,01
2	no	1 746 206,23	4 535,60	11 355,34	29,49
	yes	2 916 280,19	27 512,08	3 179,74	30,00
	Итого:	4 662 486,42	9 495,90	14 535,08	29,60
Итого:		17 755 824,99	13 270,42	41 027,62	30,66

Рисунок 3.18 - Настройки визуализатора Куб

Сохраним визуализатор Куб, перейдем в настройки визуализатора Профили кластеров (рис. 3.19).

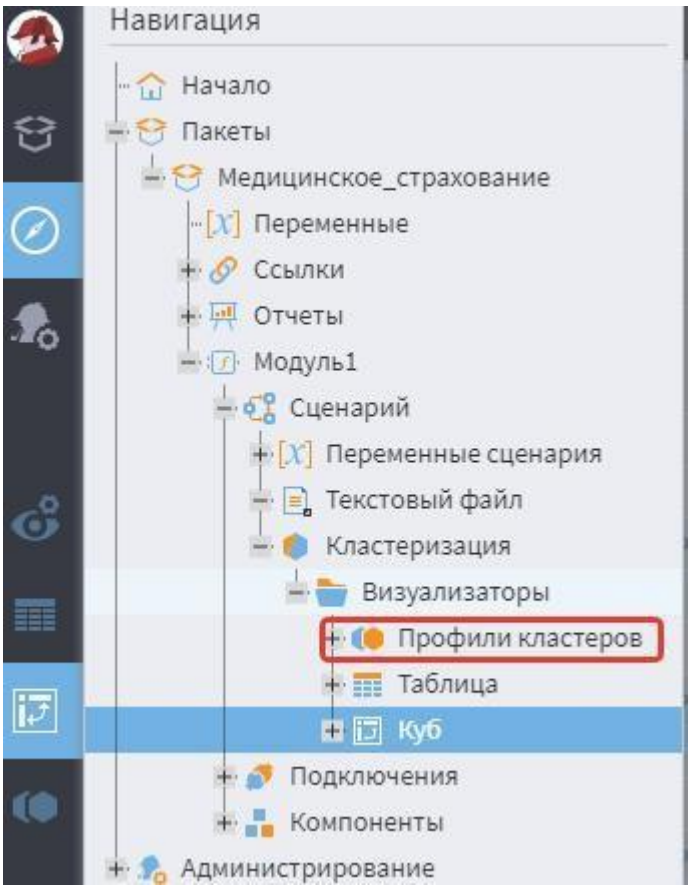


Рисунок 3.19 - Переход в настройки визуализатора Профили кластеров

Транспонируем таблицу, расположив кластеры в столбцах, если необходимо упорядочим столбцы по номеру кластера (простым перетаскиванием), выделим три ячейки в строке age (возраст) для сравнения (выделение нескольких ячеек одновременно возможно при нажатой клавише Ctrl). Справа от таблицы будет проведено сравнение кластеров по показателю age (рис. 3.20).

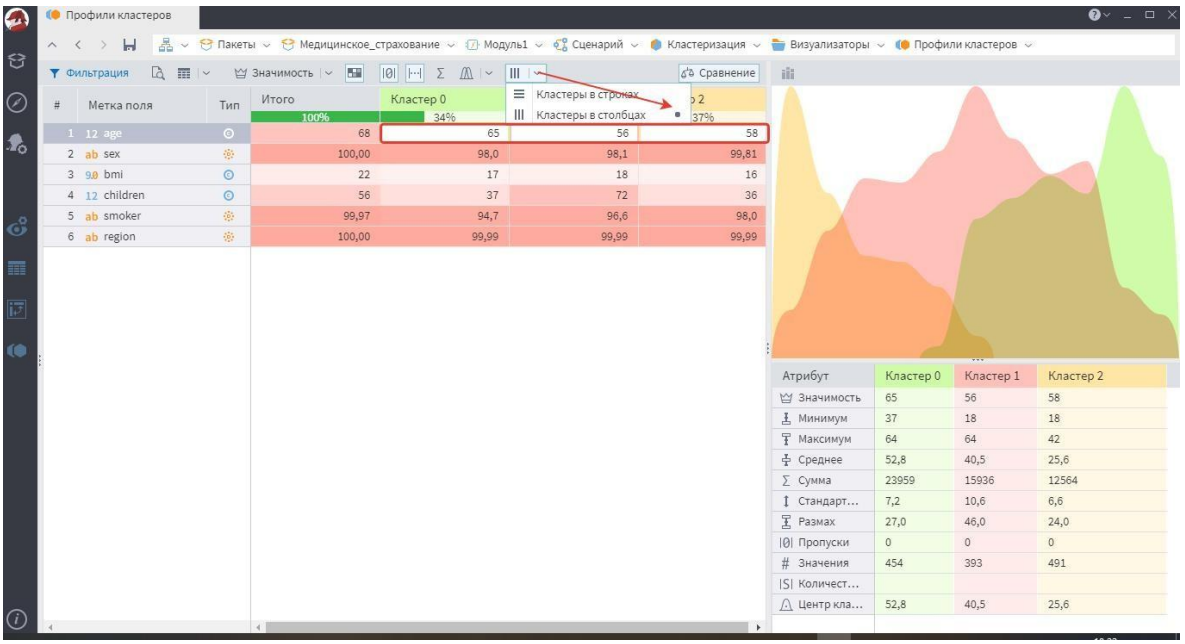


Рисунок 3.20 - Сравнение кластеров по показателю age

Мощность кластера показывает количество строк исходного набора, попавших в кластер. Задает диапазон от 0 до числа строк в исходном наборе данных.

Значимость поля – мера влияния поля на попадание поля в некоторый кластер. Задает диапазон от 0 до 100 %.

Значимость ячейки – мера влияния ячейки на попадание ячейки в некоторый кластер. Задает диапазон от 0 до 100 %.

В отчете по лабораторной работе опишите построенные кластеры, чем они характеризуются. Также опишите, как взаимосвязаны расходы, оплаченные медицинской страховкой, (charges) и характеристики (age, bmi, smoker и др.) застрахованных.

Контрольные вопросы

1. Охарактеризуйте термин кластеризация.
2. На основе каких алгоритмов кластеризации в Loginom обработчик производит кластеризацию объектов
3. Как выполнить настройку узла Кластеризация
4. Что показывает Мощность кластера?

Лабораторная работа 3. Нейросеть (регрессия) и линейная регрессия

Цель работы: ознакомиться принципами построения нейросетей (регрессия) в Loginom.

Содержание работы:

Линейная регрессия представляет собой модель зависимости между входными и выходными переменными с линейной функцией связи. Это один из наиболее часто используемых алгоритмов в машинном обучении.

В случае двух переменных: входная – x, выходная – y, графически построение линейной регрессии можно представить как приближение (аппроксимацию) связи между значениями этих переменных прямой линией (рис. 4.1).

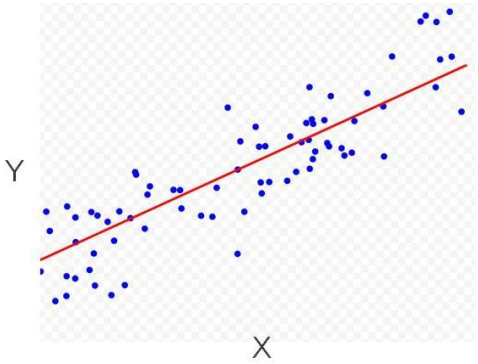


Рисунок 4.1 - Графическое представление линейной регрессии

Выходная переменная y линейной регрессии – количественная. Входными переменными могут быть любые и категориальные, и количественные показатели. Линейная регрессия оценивает, как входные показатели влияют на выход. По модели линейной регрессии для известных значений набора входных показателей можно предсказать значение выходной переменной.

Модель нейронной сети включает (рис. 4.2):

- входной слой;
- скрытые (вычислительные) слои;
- выходной слой.

Технически обучение нейронной сети заключается в нахождении весов, т.е. коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между

входными параметрами и выходными, а также выполнять обобщение. В выходном наборе нейросеть выдаст прогнозируемое значение переменной, зависимое от множества входных параметров [8].

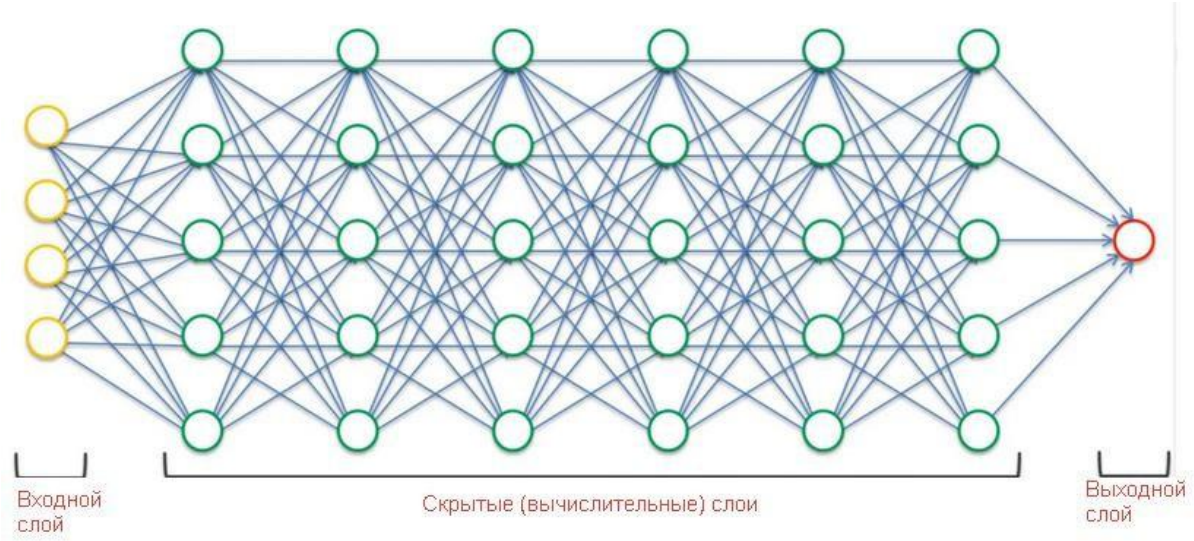


Рисунок 4.2 - Модель нейронной сети

На том же наборе данных (файл insurance.csv) построим модель линейной регрессии, а также создадим нейросеть (регрессия).

Для этого в пакете Медицинское_страхование добавим Модуль2 (рис. 4.3).

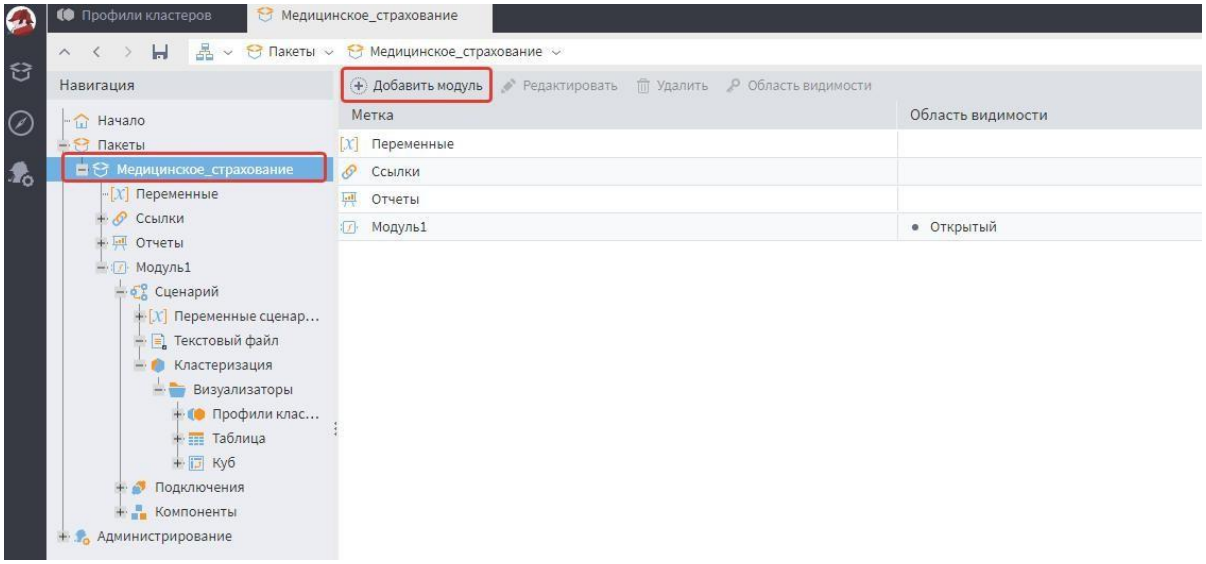


Рисунок 4.3 - Добавление модуля в пакет

В Сценарий Модуль2 из Сценария Модуль1 скопируем узел Текстовый файл (с помощью контекстного меню для узла: опции Копировать и Вставить). В рабочей области Сценария Модуль2 разместим два узла: Линейная регрессия и Нейросеть (регрессия) (рис. 4.4).

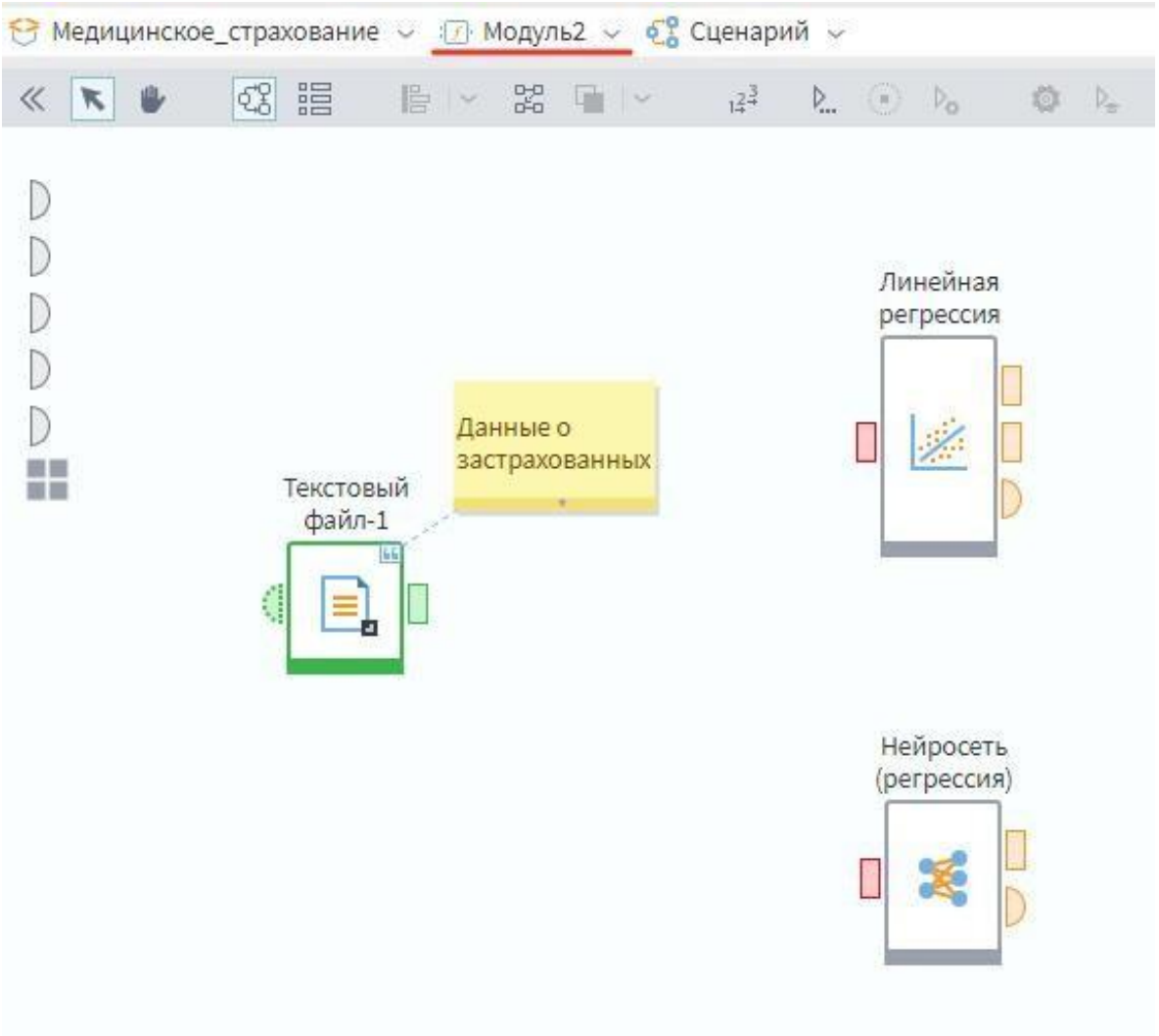


Рисунок 4.4 - Добавление узлов в сценарий

Проведем настройку узла Линейная регрессия (шестеренка внутри узла). Все показатели, кроме charges (медицинских расходов), являются входными (рис. 4.5). При редактировании значений столбца Назначение для дискретных переменных возможен выбор только варианта входной (по умолчанию у всех переменных значение не задан, т.е. не используется в регрессионной модели), для непрерывных переменных оба варианта (и входной, и выходной) доступны для выбора. Это связано с тем, что выходом в линейной регрессии может быть только количественный непрерывный показатель.

Настройки нормализации пропустим. В разбиении на множества могли бы использовать все данные для обучения модели, но разделим набор данных случайным образом в отношении 8/2 (рис. 4.6):

– 80 % наблюдений будут использоваться для построения модели, т.е. для оценки параметров линейной регрессии;

– 20 % – для тестирования ее качества, т.е. определения того, насколько сильно отклоняются для наблюдений тестового множества

прогноз выходного параметра и его значение в наборе данных для соответствующего наблюдения.

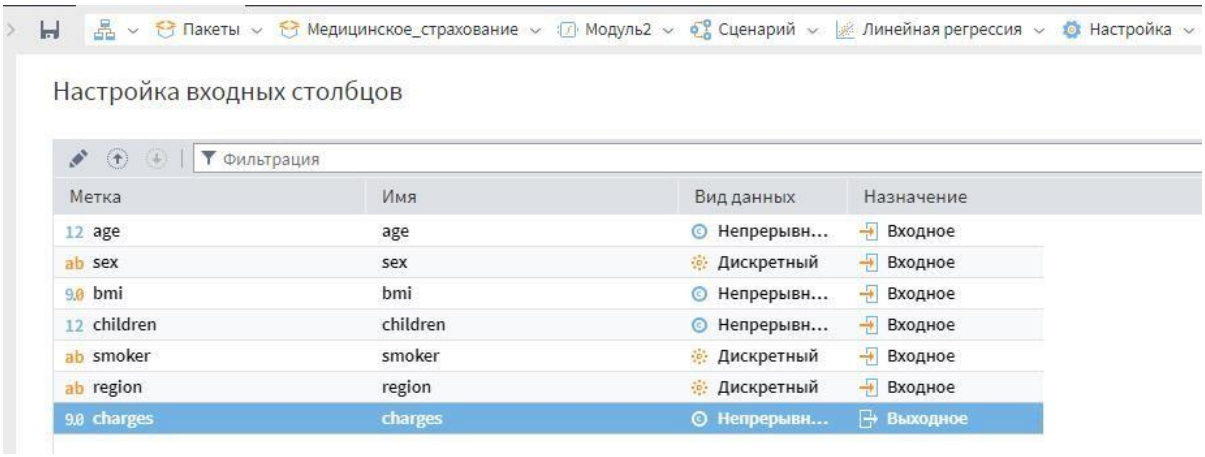


Рисунок 4.5 - Настройка входных столбцов

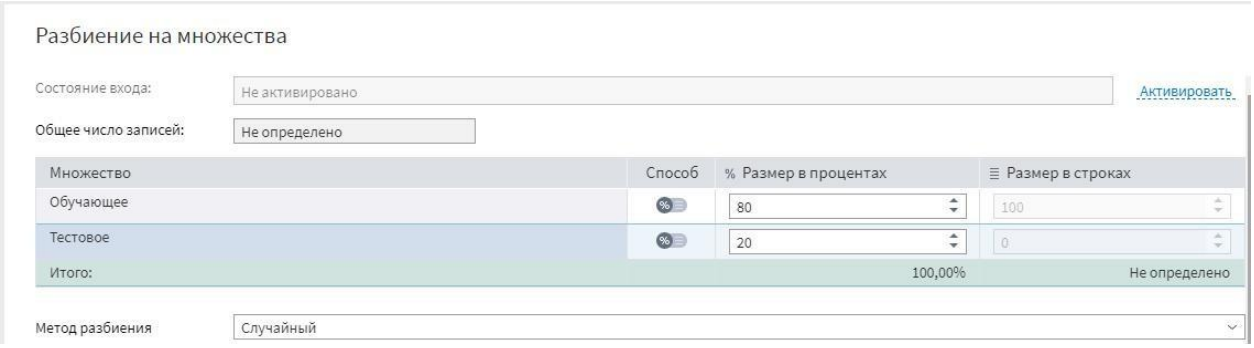


Рисунок 4.6 - Разбиение на множества

Настройки линейной регрессии оставим такие, какие есть в варианте по умолчанию. Описание дополнительных параметров настройки узла Линейная регрессия можно найти в соответствующем источнике [7].

Сохраним настройки. Вызовем контекстное меню для узла Линейная регрессия и переобучим узел (рис. 4.7).

После обучения модели будут доступны выходные порты (рис. 4.8).

Кликнув дважды на любом выходе мышкой, получим дополнительное окно с тремя вкладками в рабочей области сценария. Первая вкладка – выход линейной регрессии (рис. 4.9). В таблицу исходных данных набора добавлен столбец с выходом регрессии (т.е. рассчитанным по регрессионной модели значением выходного параметра или предсказанным/прогнозным значением выходного показателя).

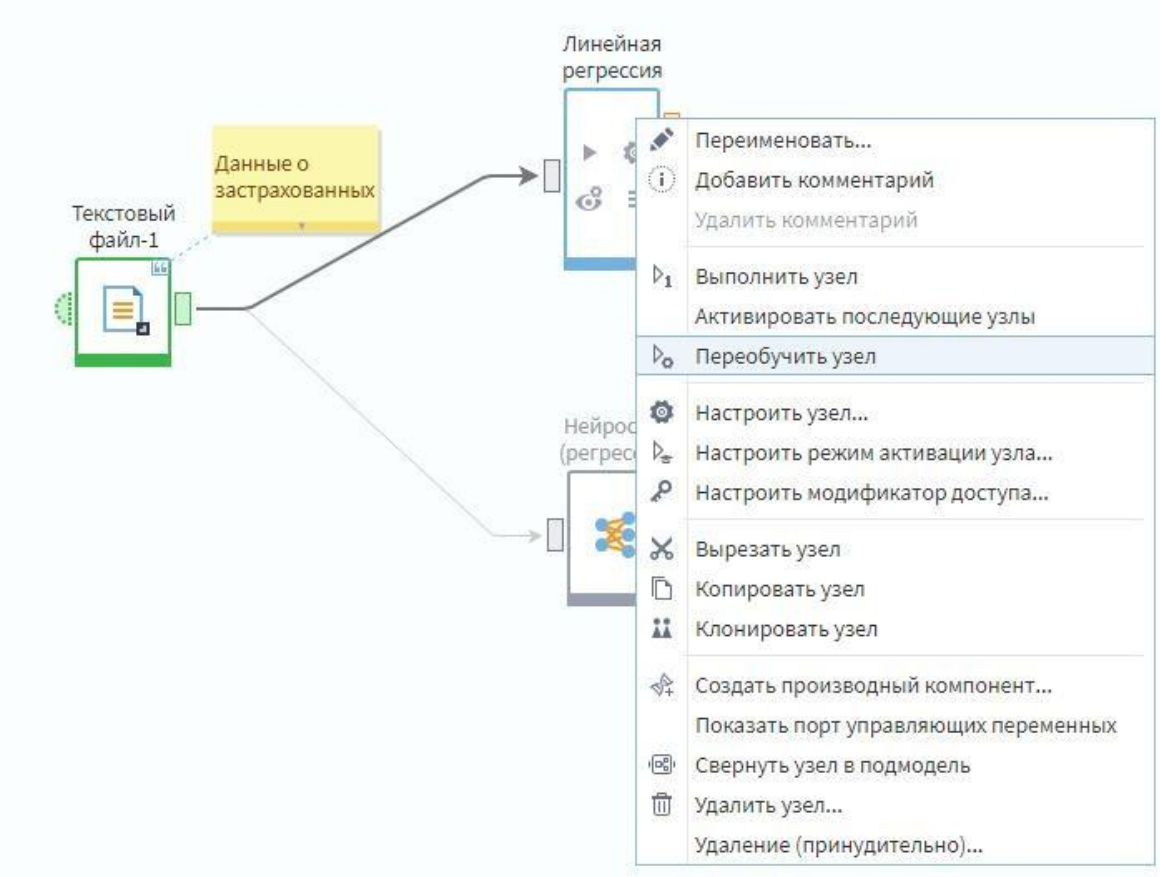


Рисунок 4.7 - Переобучение узла

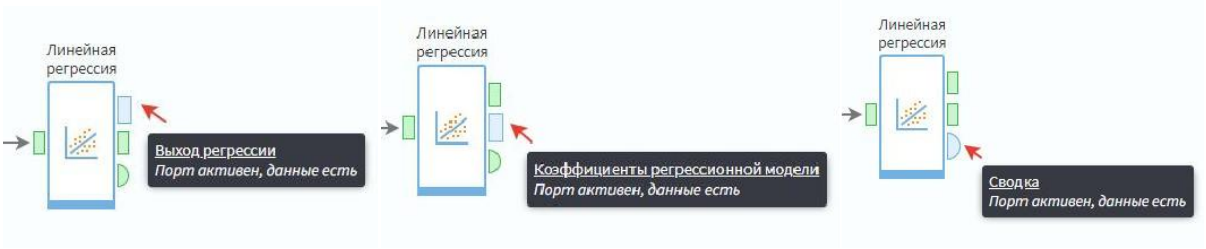


Рисунок 4.8 - Выходные порты линейной регрессии

Выход регрессии | Коэффициенты регрессионной модели | Сводка | Линейная регрессия

#	12 age	ab sex	9.0 bmi	12 children	ab smoker	ab region	9.0 charges|Регрессия	9.0 charges
1	19	female	27,90	0	yes	southwest	24 937,45	16 884,92
2	18	male	33,77	1	no	southeast	3 327,56	1 725,55
3	28	male	33,00	3	no	southeast	6 539,63	4 449,46
4	33	male	22,71	0	no	northwest	3 675,25	21 984,47
5	32	male	28,88	0	no	northwest	5 510,25	3 866,86
6	31	female	25,74	0	no	southeast	3 791,89	3 756,62
7	46	female	33,44	1	no	southeast	10 735,16	8 240,59
8	37	female	27,74	3	no	northwest	7 992,33	7 281,51
9	37	male	29,83	2	no	northeast	8 479,93	6 406,41
10	60	female	25,84	0	no	northwest	11 998,64	28 923,14
11	25	male	26,22	0	no	northeast	3 263,24	2 721,32

Рисунок 4.9 - Выход регрессии

На вкладке Коэффициенты регрессионной модели можем увидеть оценку параметров линейной регрессии (рис. 4.10), оценить значимость включенных в модель показателей.

Обратите внимание, что для дискретных показателей происходит их замена на dummy-переменные. Это переменные, принимающие два значения 0/1. Их количество на единицу меньше числа уровней категориального показателя. Так, если sex имеет два уникальных значения, то переменная-заместитель вводится только одна, причем значение female (рис. 4.10) заменяется на метку 1, а значение male – на 0. Для показателя smoker также вводится одна замещающая переменная и значение yes заменяется на 1, а no – на 0. У переменной region замещающих переменных три (так как уникальных значений четыре): в первой значение northeast имеет метку 1, все остальные заменяются на 0, во второй northwest заменяется на 1, все остальные заменяются на 0, в третьей значение southwest заменяется на 1, все остальные заменяются на 0.

Выход регрессии | Коэффициенты регрессионной модели | Сводка | Линейная регрессия

#	Имена входных полей	Метки входных полей	Уникальные значения	charges|Коэффициенты	charges|Стд. откл.	charges|T-статистика	charges|Значимость
1	<Константа>	<null>	<null>	-13 236,84	1 215,01	-10,89	2,84e-26
2	age	age	<null>	259,49	13,23	19,62	2,42e-73
3	sex	sex	female	253,80	372,52	0,68	0,50
4	sex	sex	male	0,00			
5	bmi	bmi	<null>	339,19	32,05	10,58	5,88e-25
6	children	children	<null>	439,17	155,59	2,82	0,00
7	smoker	smoker	no	0,00			
8	smoker	smoker	yes	23 653,43	465,26	50,84	1,21e-286
9	region	region	northeast	1 119,29	531,48	2,11	0,04
10	region	region	northwest	647,62	538,06	1,20	0,23
11	region	region	southeast	0,00			
12	region	region	southwest	-126,64	525,16	-0,24	0,81

Рисунок 4.10 - Коэффициенты регрессионной модели

Если в качестве уровня значимости выбрать 0,05 (это вероятность ошибочно признать параметр при переменной значимым, когда на самом деле соответствующий параметру входной показатель не влияет на выход в достаточной мере, чтобы принимать его во внимание), то и sex, и большинство районов (т.е. замещающих показатель region бинарных переменных) мало влияют на медицинские расходы. Показателями sex и region можно было бы пренебречь, исключив их из входных столбцов (рис. 4.5). Можно добавить в сценарий еще один узел Линейная регрессия и настроить для него входные столбцы с учетом проведенного анализа. Можно перенастроить узел Линейная регрессия, который уже есть в сценарии. Но если настройки узла из категории Data Mining изменяются, то в обязательном порядке необходимо повторно переобучить узел.

Сводка по модели (рис. 4.11) позволяет провести анализ качества построенной модели. Например, коэффициент детерминации (пятая строка в сводке) показывает, какую часть изменений выходного показателя можно объяснить вариативностью входных показателей.

В построенной модели этот показатель равен 0,75, т.е. 75 % изменений в медицинских расходах обусловлено влиянием входных показателей.

Выход регрессии | Коэффициенты регрессионной модели | Сводка | Линейная регрессия

№	Имя	Метка	Значение
1	12 TotalSamples	Всего примеров	1 338
2	12 TotalSelectedSamples	Всего отобранных примеров	1 338
3	12 TrainSamples	Примеров в обучающем множестве	1 070
4	9.0 LogLikelihood	Логарифм функции правдоподобия	-10 832,25
5	9.0 R2	Коэффициент детерминации	0,75
6	9.0 AdjustedR2	Скорректированный коэффициент детерминации	0,75
7	9.0 StdDev	Стандартное отклонение	6 056,48
8	12 DFE	Число степеней свободы ошибки	1 061
9	12 ModelDF	Число степеней свободы модели	8
10	9.0 FStatistic	F-статистика	394,91
11	9.0 AIC	Информационный критерий Акаике	20,26
12	9.0 AICc	Информационный критерий Акаике скорректированный	20,26
13	9.0 BIC	Информационный критерий Байеса	20,31
14	9.0 HQC	Информационный критерий Ханнана-Куинна	20,28

Рисунок 4.11 - Сводка

Проведем настройку узла Нейросеть (регрессия) (шестеренка внутри узла). Мастер настройки для этого узла во многом похож на мастер настройки для узла Линейная регрессия.

Все показатели, кроме charges (медицинских расходов), являются входными (рис. 4.12). При редактировании значений столбца Назначение для дискретных переменных возможен выбор только варианта входной (по умолчанию у всех переменных значение не задан, т.е. не используется в регрессионной модели), для непрерывных переменных оба варианта (и входной, и выходной) доступны для выбора. Это связано с тем, что выходом в модели Нейросеть (регрессия) может быть только количественный непрерывный показатель.

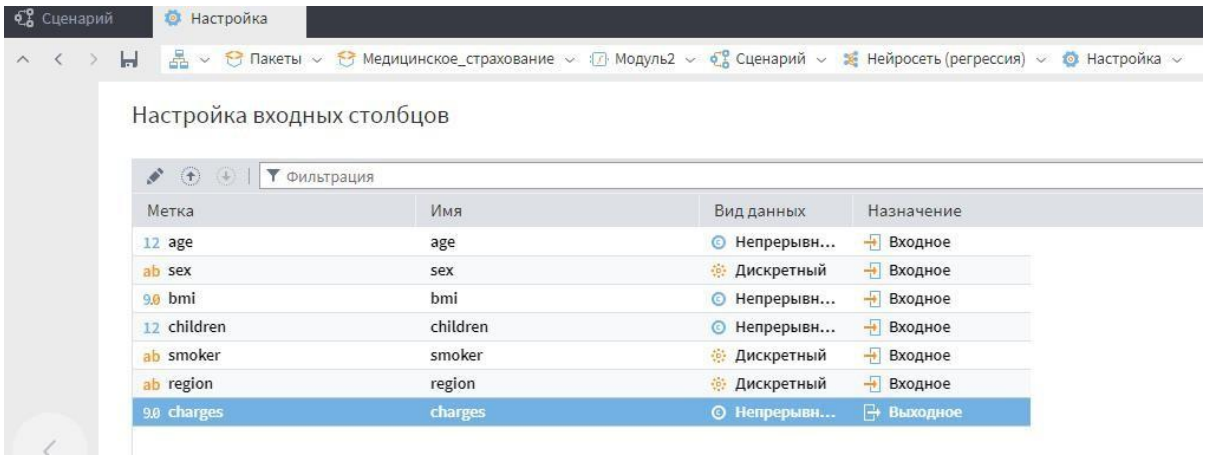


Рисунок 4.12 - Настройка входных столбцов

Настройки нормализации пропустим. В разбиении на множества разделим набор данных случайным образом в отношении 8/2 (рис. 4.13):

– 80 % наблюдений будут использоваться для построения модели;

– 20 % – для тестирования ее качества, т.е. определения того, насколько сильно отклоняются для наблюдений тестового множества прогноз выходного параметра и его значение в наборе данных для соответствующего наблюдения.

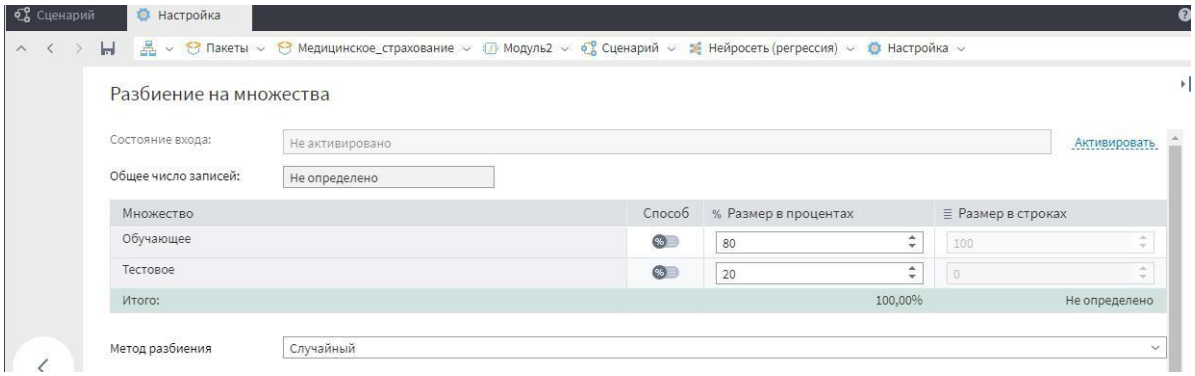


Рисунок 4.13 - Разбиение на множества

В настройках параметров нейросети изменим количество нейронов в первом скрытом слое на 3 (рис. 4.14). Чем больше скрытых слоев и нейронов в этих слоях – тем продолжительнее обучение модели и тем точнее прогноз. Но для нейросети характерна проблема переобучения: с ростом точности прогноза выходного параметра на обучающем множестве данных теряется ее гибкость и адаптируемость к новым данным, т.е. на новых данных переобученная нейросеть прогноз выполняет хуже, с большей погрешностью. Поэтому необходимо соблюдать баланс и не увеличивать сложность структуры модели для небольшого объема данных.

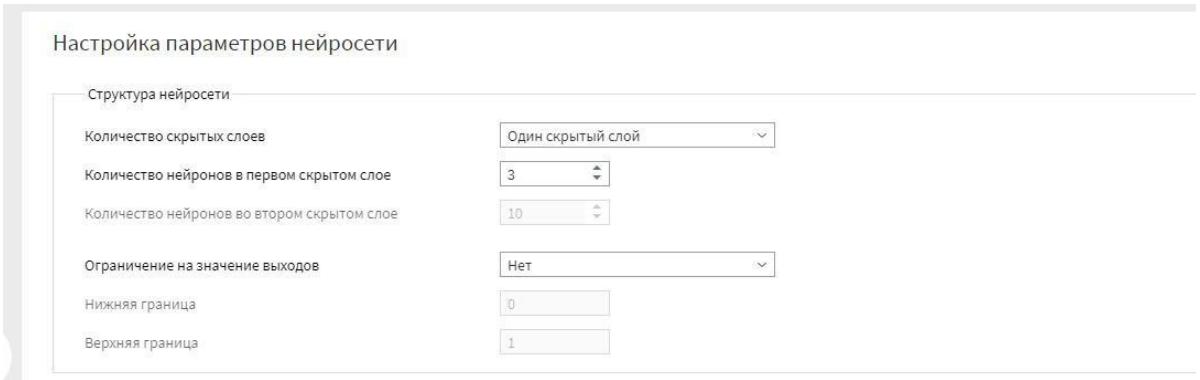
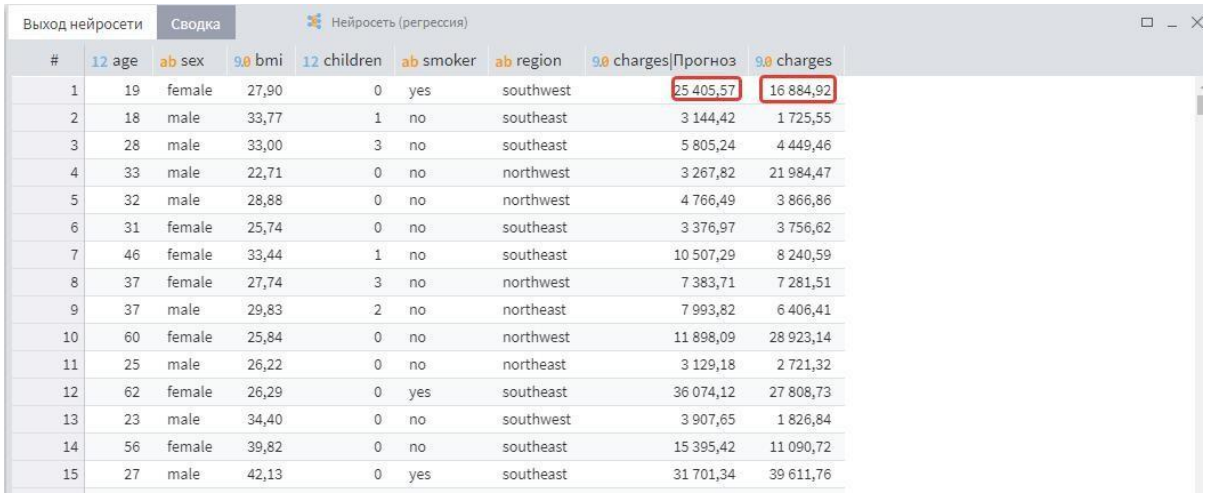


Рисунок 4.14 - Настройка параметров нейросети

Настройки автоматического подбора параметров Нейросети пропустим. Описание дополнительных параметров настройки узла Нейросеть (регрессия) можно найти в соответствующем источнике [8].

Сохраним настройки. Вызовем контекстное меню для узла Нейросеть (регрессия) и переобучим узел. У узла Нейросеть (регрессия) два выхода: Выход нейросети и Сводка. Двойным кликом мышки на любом из них выведем в рабочую область сценария окно с этими вкладками (рис. 4.15).



#	12 age	ab sex	9.0 bmi	12 children	ab smoker	ab region	9.0 charges|Прогноз	9.0 charges
1	19	female	27,90	0	yes	southwest	25 405,57	16 884,92
2	18	male	33,77	1	no	southeast	3 144,42	1 725,55
3	28	male	33,00	3	no	southeast	5 805,24	4 449,46
4	33	male	22,71	0	no	northwest	3 267,82	21 984,47
5	32	male	28,88	0	no	northwest	4 766,49	3 866,86
6	31	female	25,74	0	no	southeast	3 376,97	3 756,62
7	46	female	33,44	1	no	southeast	10 507,29	8 240,59
8	37	female	27,74	3	no	northwest	7 383,71	7 281,51
9	37	male	29,83	2	no	northeast	7 993,82	6 406,41
10	60	female	25,84	0	no	northwest	11 898,09	28 923,14
11	25	male	26,22	0	no	northeast	3 129,18	2 721,32
12	62	female	26,29	0	yes	southeast	36 074,12	27 808,73
13	23	male	34,40	0	no	southwest	3 907,65	1 826,84
14	56	female	39,82	0	no	southeast	15 395,42	11 090,72
15	27	male	42,13	0	yes	southeast	31 701,34	39 611,76

Рисунок 4.15 - Выход нейросети

Как видим, столбец прогнозных значений выходного параметра в этой таблице называется charges|Прогноз, в отличие от выхода регрессии, где столбец предсказанных значений по регрессии назывался charges|Регрессия (см. рис. 4.9).

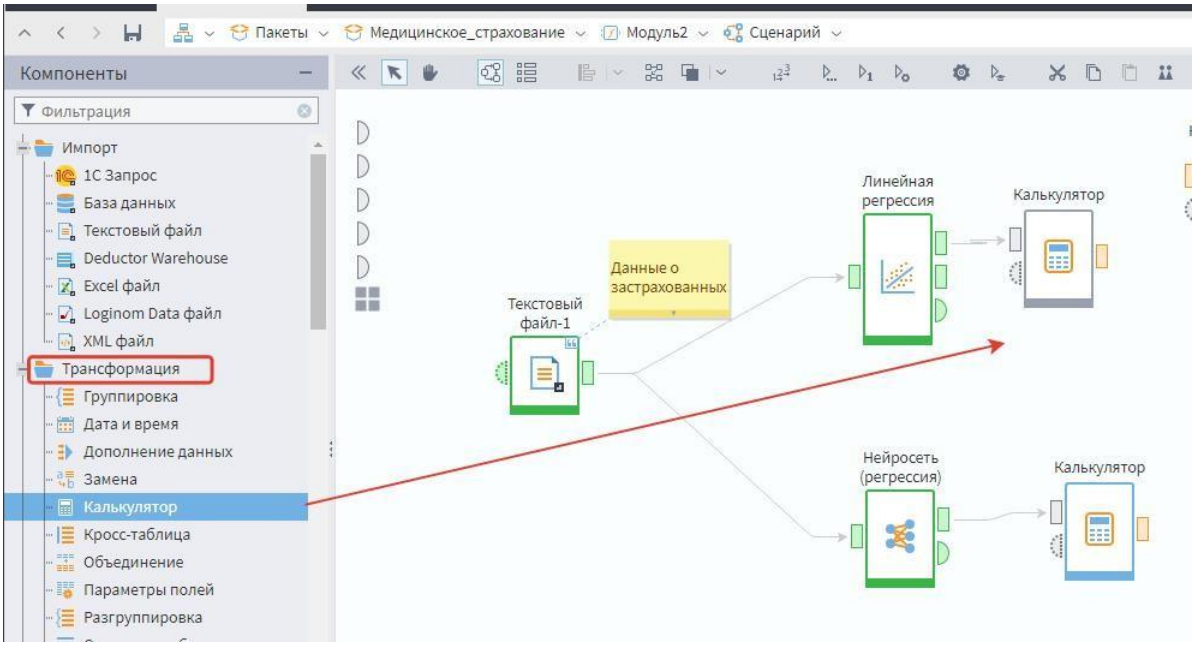


Рисунок 4.16 - Добавление узлов Калькулятор в Сценарий

Сравним точность предсказания по модели. Для этого рассчитаем ошибку аппроксимации для каждой из моделей: насколько выход модели отличается от реальных значений этого показателя.

Для этого добавим в сценарий два узла Калькулятор, соединим Выход регрессии с Входным источником данных для первого узла Калькулятор и Выход нейросети с Входным источником данных для второго узла Калькулятор.

Проведем настройку узлов Калькулятор. Для этого из значения столбца charges вычтем значение столбца прогноза по модели (рис. 4.17, 4.18). Выражение charges – charges_regression (рис. 4.17). Выражение charges – charges_predicted (рис. 4.18).

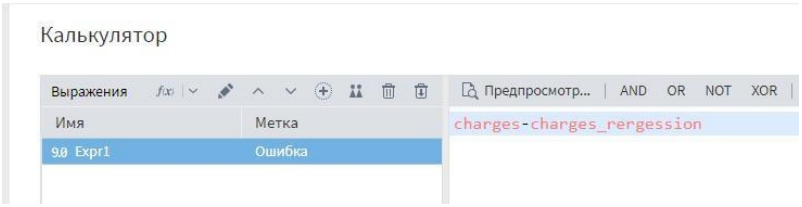


Рисунок 4.17 - Настройка узла Калькулятор, принимающего на вход выход регрессии

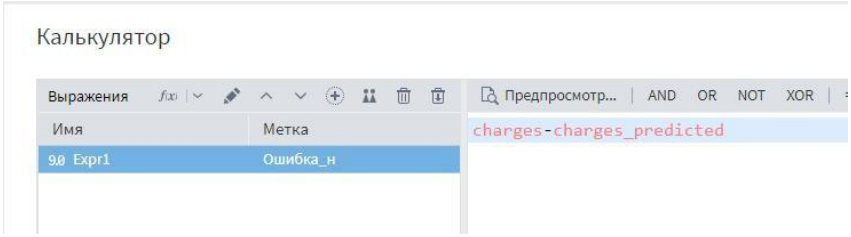


Рисунок 4.18 - Настройка узла Калькулятор, принимающего на вход выход нейросети

Добавим в сценарий узел Слияние, чтобы сравнить качество прогноза по двум моделям (рис. 4.19).

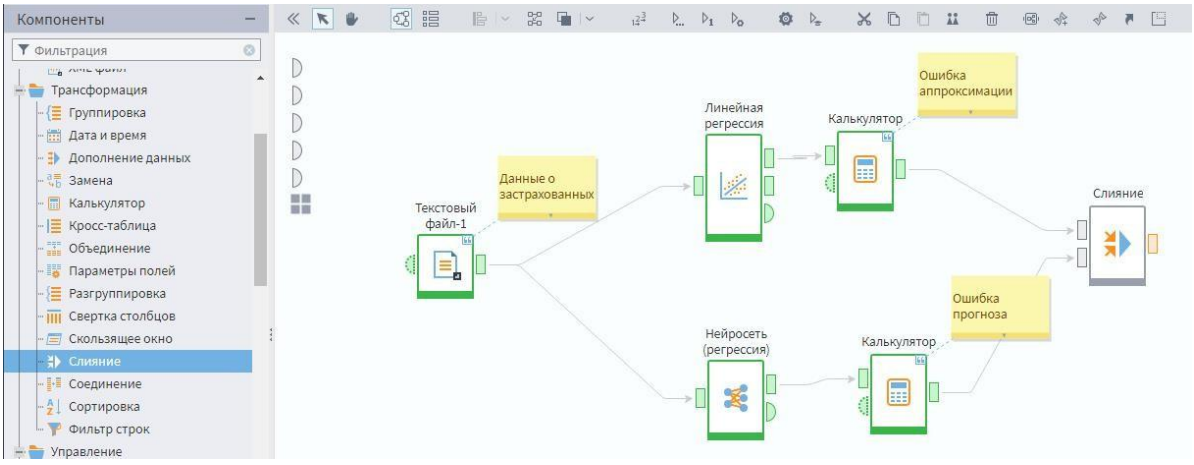


Рисунок 4.19 - Добавление узла Слияние

Настроим узел Слияние (рис. 4.20). Необходимо выбрать столбцы, по которым будет происходить объединение данных. Ключевых столбцов в этих таблицах нет. Два столбца обладают потенциально уникальными значениями – это bmi и charges, в остальных столбцах

значения повторяются и на роль потенциально ключевых они не подходят. Соединим в таблицах столбцы charges (кликнем мышкой по имени столбца в первой таблице и протащим указатель мышки на столбец с тем же именем во второй).

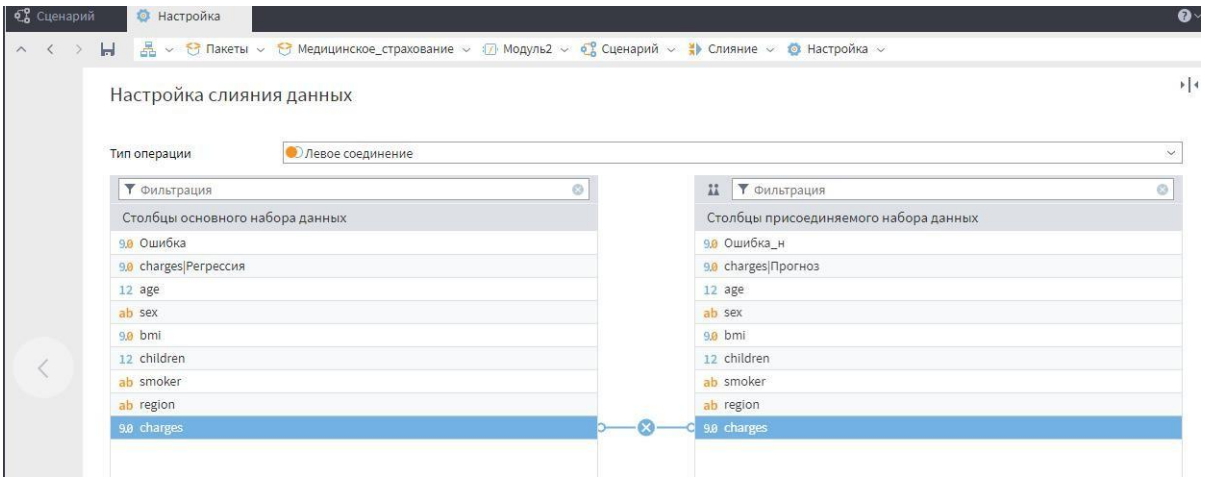
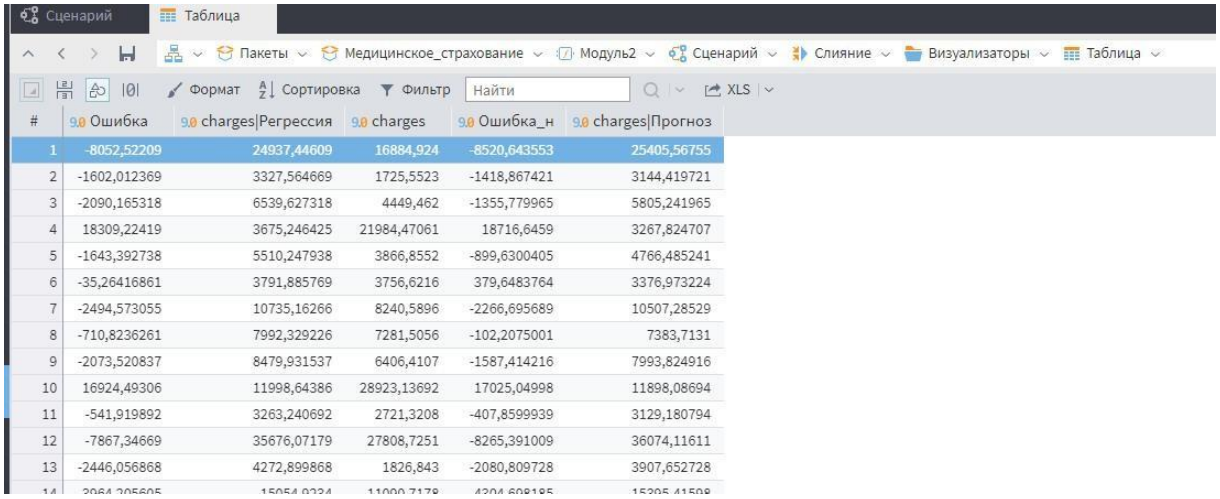


Рисунок 4.20 - Слияние данных

Настроим для узла Слияние два визуализатора: Таблицу и Диаграмму. Войдем в визуализатор Таблица и с помощью контекстного меню любого столбца (стрелка рядом с именем столбца) уберем с экрана все лишнее), оставив только ошибки, прогнозные значения по двум моделям и исходные данные показателя charges (рис. 4.21).



#	Ошибка	charges|Регрессия	charges	Ошибка_н	charges|Прогноз
1	-8052,52209	24937,44609	16884,924	-8520,643553	25405,56755
2	-1602,012369	3327,564669	1725,5523	-1418,867421	3144,419721
3	-2090,165318	6539,627318	4449,462	-1355,779965	5805,241965
4	18309,22419	3675,246425	21984,47061	18716,6459	3267,824707
5	-1643,392738	5510,247938	3866,8552	-899,6300405	4766,485241
6	-35,26416861	3791,885769	3756,6216	379,6483764	3376,973224
7	-2494,573055	10735,16266	8240,5896	-2266,695689	10507,28529
8	-710,8236261	7992,329226	7281,5056	-102,2075001	7383,7131
9	-2073,520837	8479,931537	6406,4107	-1587,414216	7993,824916
10	16924,49306	11998,64386	28923,13692	17025,04998	11898,08694
11	-541,919892	3263,240692	2721,3208	-407,8599939	3129,180794
12	-7867,34669	35676,07179	27808,7251	-8265,391009	36074,11611
13	-2446,056868	4272,899868	1826,843	-2080,809728	3907,652728

Рисунок 4.21 - Визуализатор Таблица для узла Слияние

Как видим, прогноз по двум моделям достаточно близок друг к другу. Сохраним настройки и перейдем в визуализатор Диаграмма. Переместим в область построения диаграммы два поля: Ошибка (ошибка для регрессии) и Ошибка_н (ошибка нейросети). Ощущения при просмотре числовых значений в таблице сложились верные – графики ошибок накладываются друг на друга. Можно щелчком на легенде изменять порядок отображения показателей (рис. 4.21, 4.22).

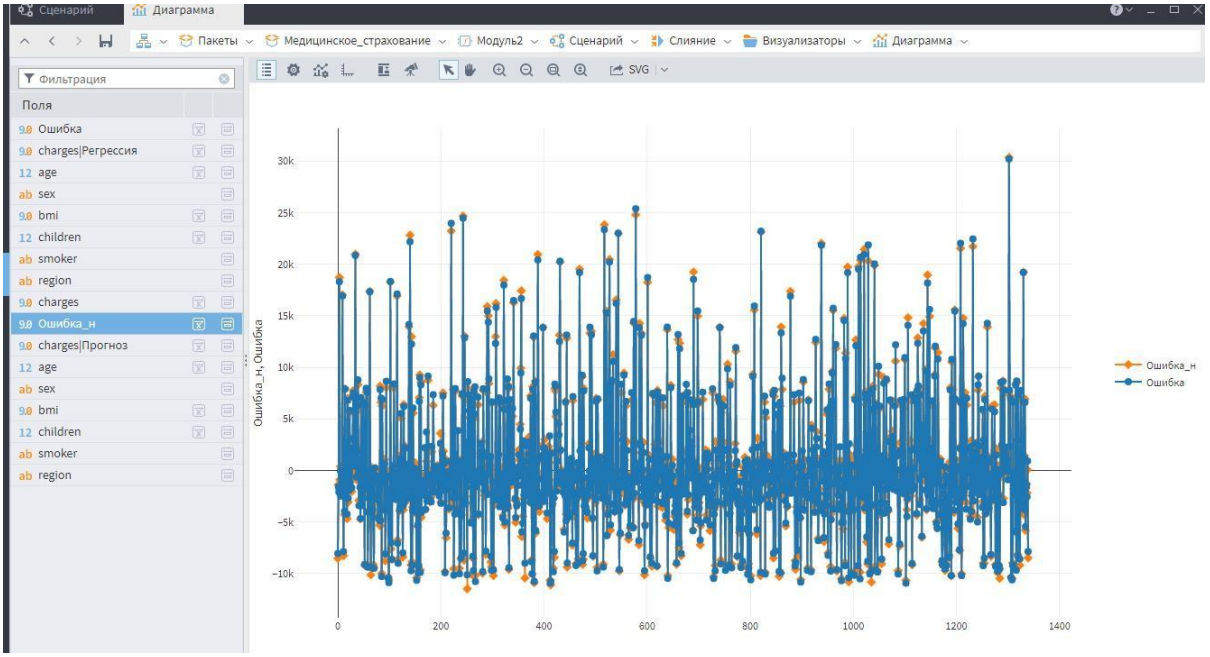


Рисунок 4.21 - Диаграмма ошибок по двум моделям

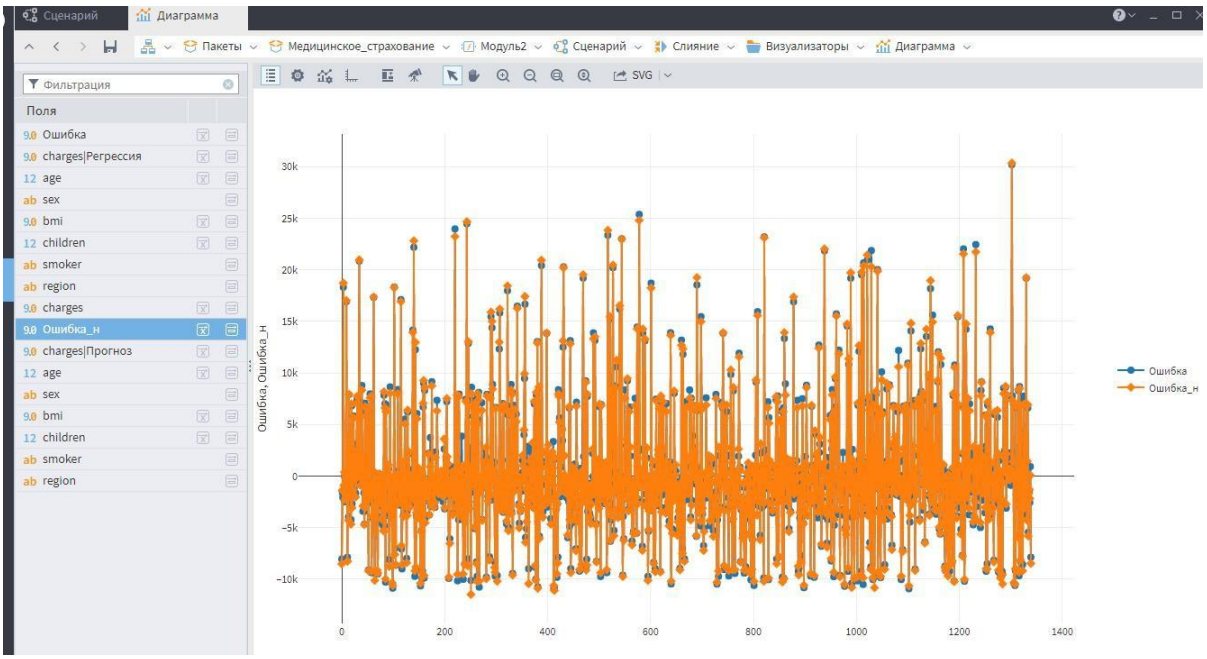


Рисунок 4.22 - Диаграмма ошибок по двум моделям

Результат в узле Линейной регрессии на тех же самых данных будет стабилен, т.е. предсказания при переобучении модели не изменятся. Нейросеть с каждым переобучением будет давать несколько иной вариант предсказаний.

Контрольные вопросы

1. Что представляет собой линейная регрессия.
2. В чем заключается обучение нейронной сети?
3. Как провести настройку узла Линейная регрессия, основные шаги.

Лабораторная работа 4. Нейросеть (классификация) и логистическая регрессия

Цель работы: ознакомиться принципами построения нейросетей (классификация) в Loginom.

Содержание работы:

Метод регрессии можно адаптировать для решения задачи классификации, когда отклик – это категориальная переменная (т.е. дискретный показатель), в простейшем случае отклик принимает два возможных значения, которые можно закодировать метками 1 и 0.

Для решения задачи классификации с помощью регрессии используется логистическая регрессия, которая является линейным классификатором. Ценность логистической регрессии в том, что она не просто предсказывает, к какому классу относится предъявленный объект, но и прогнозирует вероятность уверенности в отнесении объекта к этому классу [9].

Основная идея линейного классификатора заключается в том, что признаковое пространство может быть разделено гиперплоскостью на два полупространства, в каждом из которых прогнозируется одно из двух значений целевого класса.

Логистическая регрессия оценивает вероятность того, что событие наступит для конкретного объекта/экземпляра наблюдения (болен/здоров, ушел/остался, возврат кредита / дефолт и т.д.).

Для нейросети (классификации) основное отличие от нейросети (регрессии) заключается в том, что выходное поле может быть только дискретным.

Набор данных, который будем использовать для построения модели логистической регрессии, – это данные по абонентам сотового оператора, которые можно скачать из репозитория https://github.com/Smeilz/ML-Class/tree/master/R/Trees. Имя файла Churn_binned.csv.

Описание переменных набора:

– longdist: длительность междугородних звонков в минутах (порядковый предиктор);

– local: длительность местных звонков в минутах (порядковый предиктор);

– int_disc: скидка за междугородние звонки (номинальный предиктор);

– billtype: тип местных звонков (номинальный предиктор);

– pay: способ оплаты (номинальный предиктор);

– gender: пол (номинальный предиктор);
– marital: семейное положение (номинальный предиктор);
– incomecat: категория дохода (порядковый предиктор);
– agecat: возрастная категория (порядковый предиктор);
– churn: наличие оттока (номинальный предиктор).

Скачаем этот набор на локальный диск. Создадим в Loginom новый пакет Оператор_сотовой_связи. Разместим в сценарии первый узел Текстовый файл и проведем его настройку (рис. 5.1).

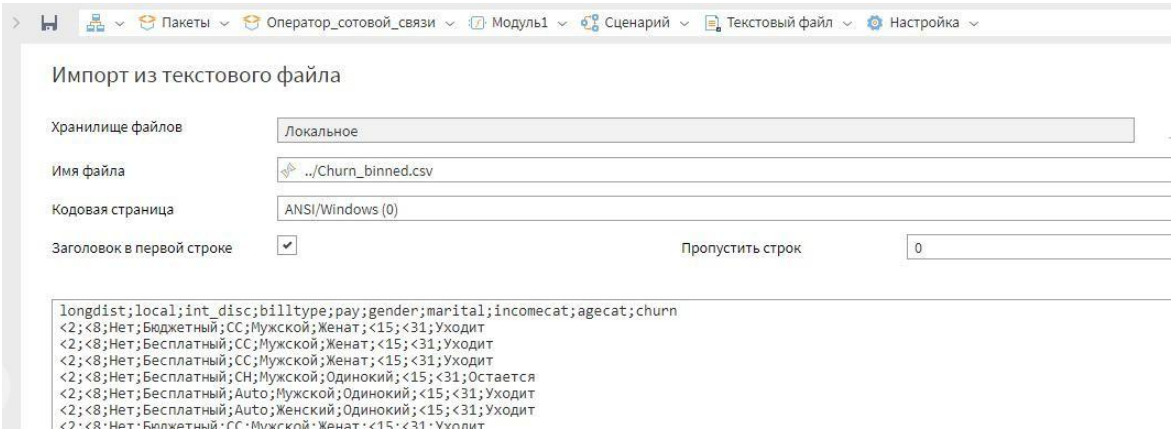


Рисунок 5.1 - Импорт из текстового файла

При настройке форматов импорта для узла Текстовый файл разделитель столбцов не соответствует используемому в импортируемом файле, из-за чего строка для наблюдения не разбилась на значения отдельных полей, а целиком была отнесена к одному полю (рис. 5.2). При выборе верного разделителя разнесение отдельных значений в строке по полям будет корректным (рис. 5.3).

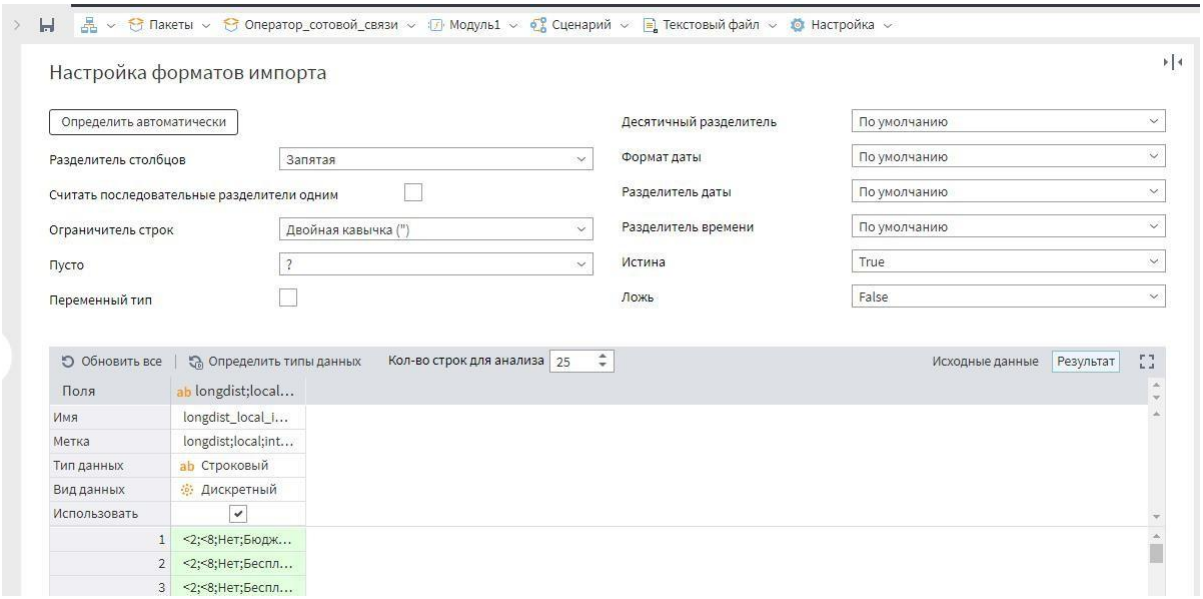


Рисунок 5.2 - Настройка форматов импорта

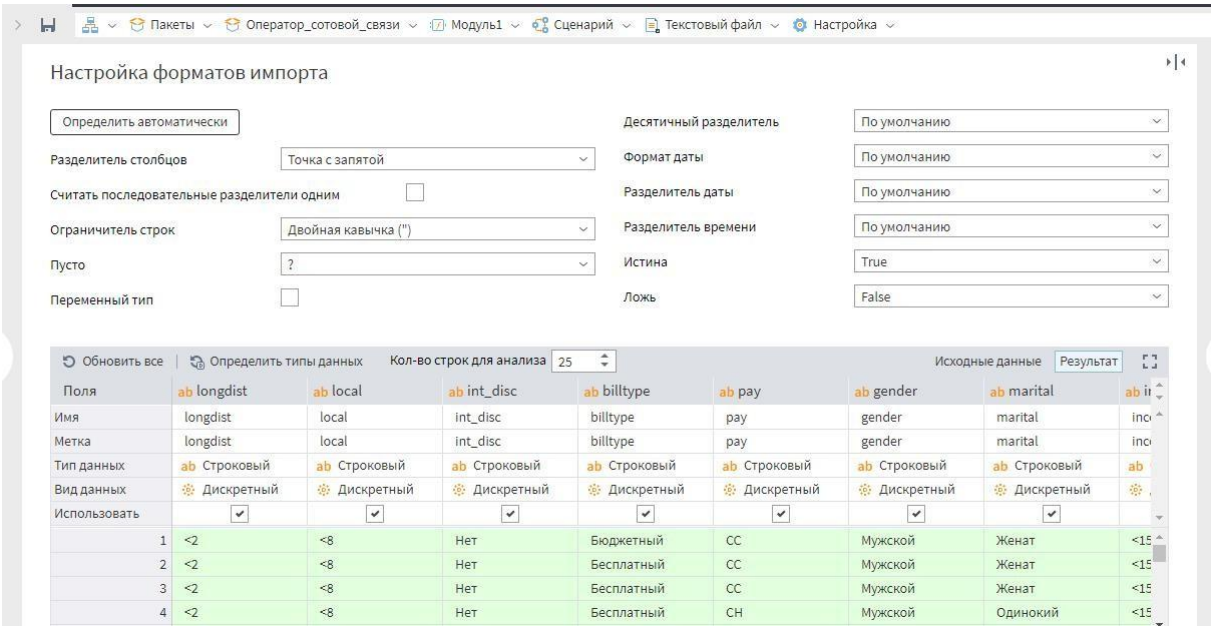


Рисунок 5.3 - Настройка форматов импорта

Все поля представлены категориальными дискретными показателями, даже возраст. Любой непрерывный показатель с помощью процедуры квантования может быть представлен в дискретном формате, для этого интервал его значений нужно разбить на части и каждому подинтервалу присвоить некоторую метку.

Разместим в Сценарии два узла: Логистическая регрессия и Нейросеть (классификация) – и соединим выходной порт узла Текстовый файл с входным портом каждого из добавленных узлов (рис. 5.4).

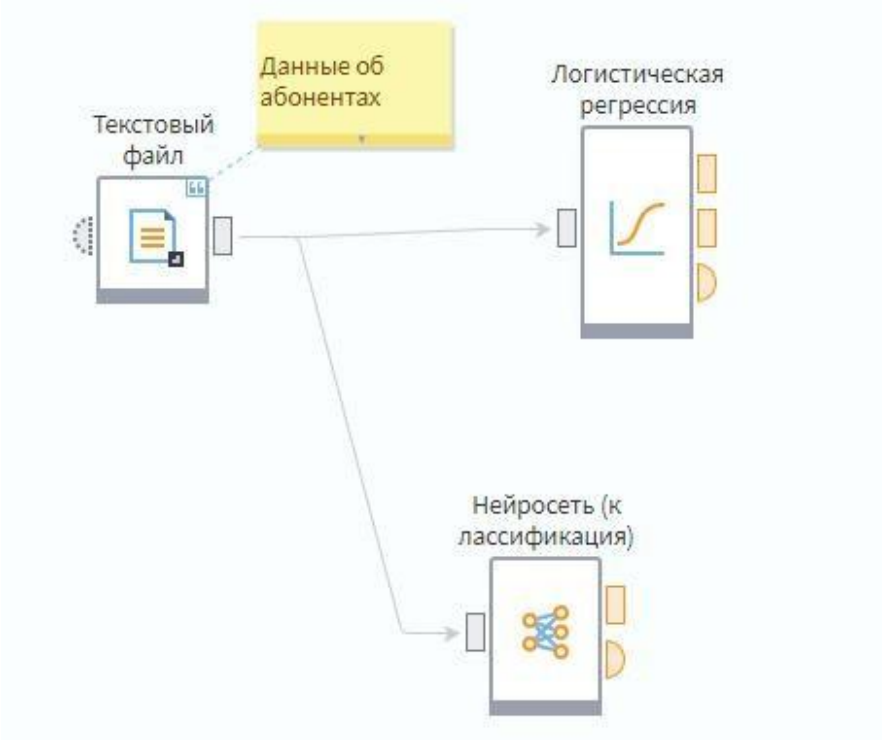


Рисунок 5.4 - Добавление узлов в сценарий

Проведем настройку узла Логистическая регрессия (шестеренка внутри узла). Все показатели, кроме churn (оттока абонентов), являются входными (рис. 5.5). При редактировании значений столбца Назначение для дискретных переменных возможен выбор и варианта входной, и варианта выходной, так как в логистической регрессии выходом может быть только категориальный дискретный показатель, а все переменные набора именно такие. В качестве входа в логистической регрессии могут быть как дискретные, так и непрерывные переменные. Просто набор данных, используемый в этом пакете, непрерывных количественных показателей не содержит.

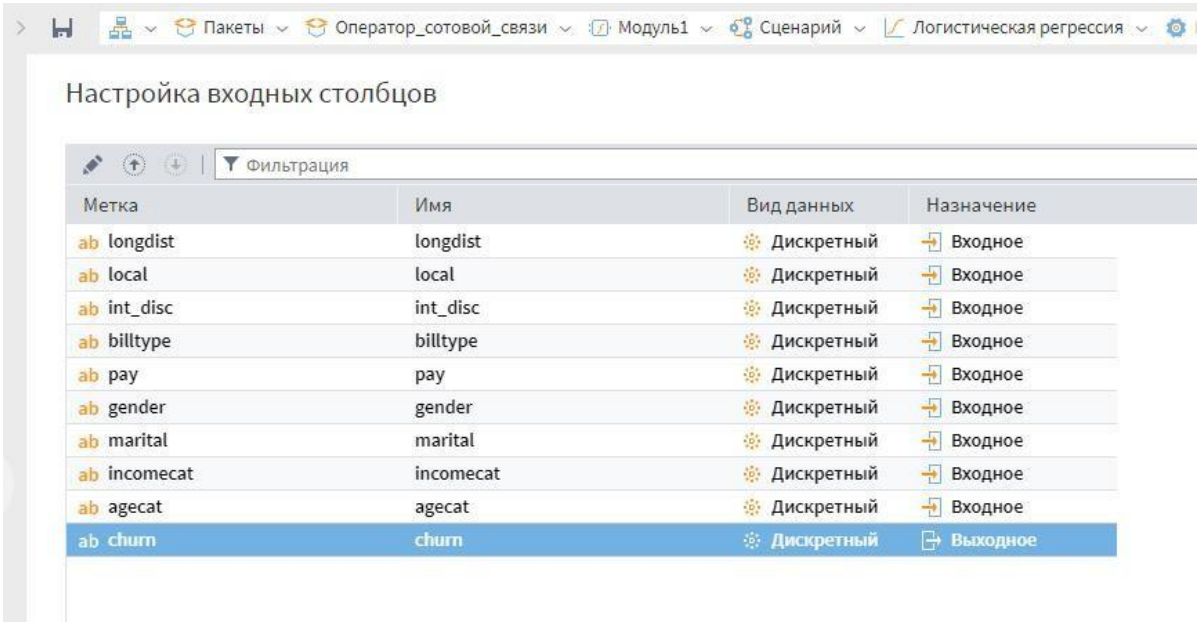


Рисунок 5.5 - Настройка входных столбцов

Настройки нормализации пропустим. В разбиении на множества разделим набор данных случайным образом в отношении 8/2 (рис. 5.6):

– 80 % наблюдений будут использоваться для построения модели, т.е. для оценки параметров логистической регрессии;

– 20 % – для тестирования ее качества, т.е. определения того, насколько сильно отклоняются для наблюдений тестового множества прогноз выходного параметра и его значение в наборе данных для соответствующего наблюдения.

Настройки логистической регрессии оставим такие, какие есть в варианте по умолчанию (рис. 5.7).

Описание дополнительных параметров настройки узла Логистическая регрессия можно найти в соответствующем источнике [11].

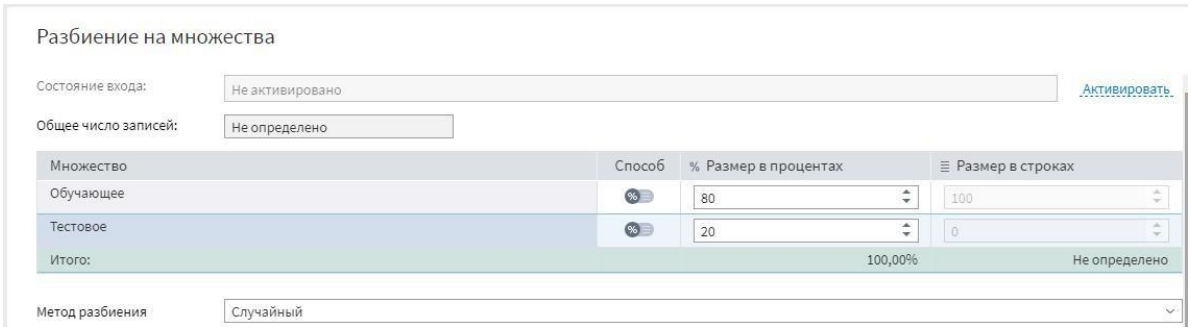


Рисунок 5.6 - Разбиение на множества

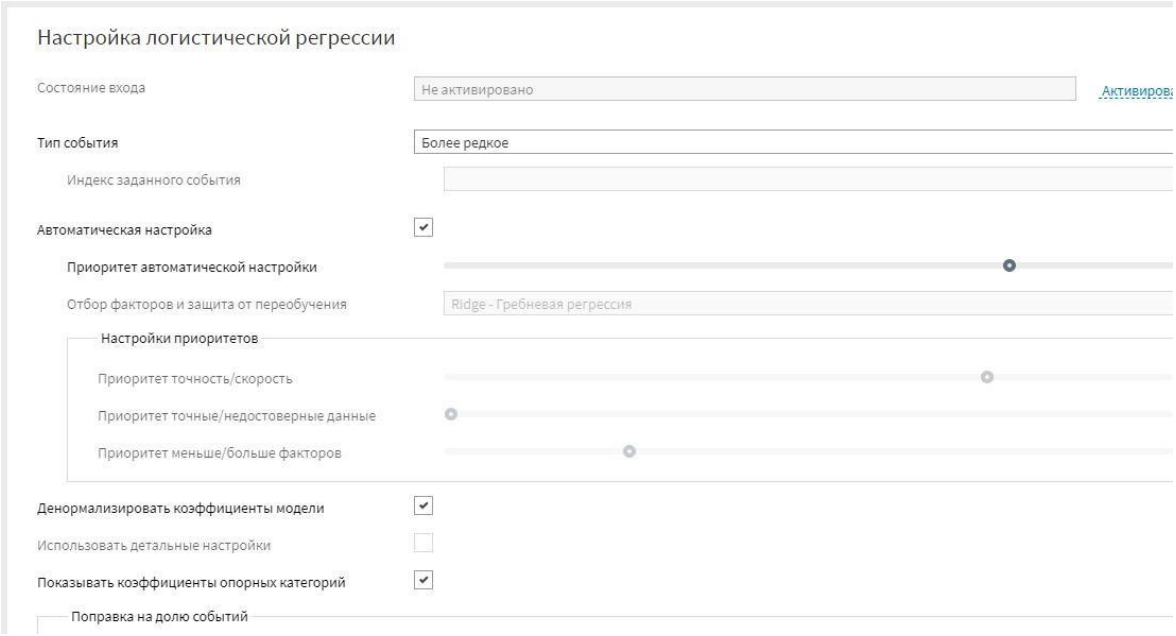
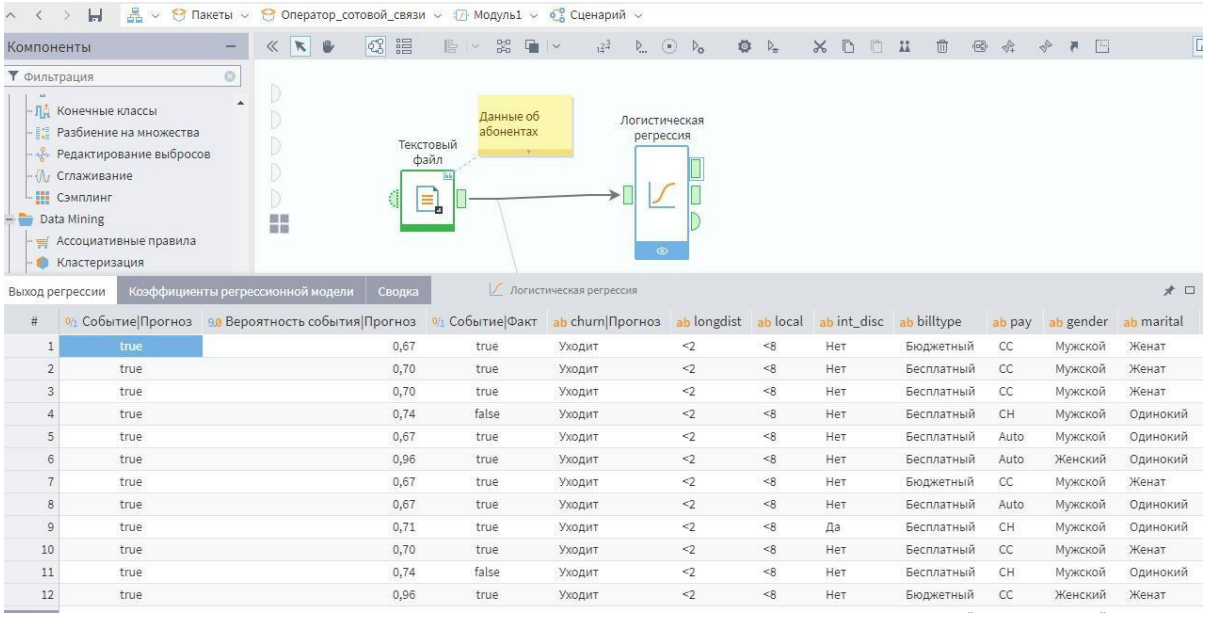


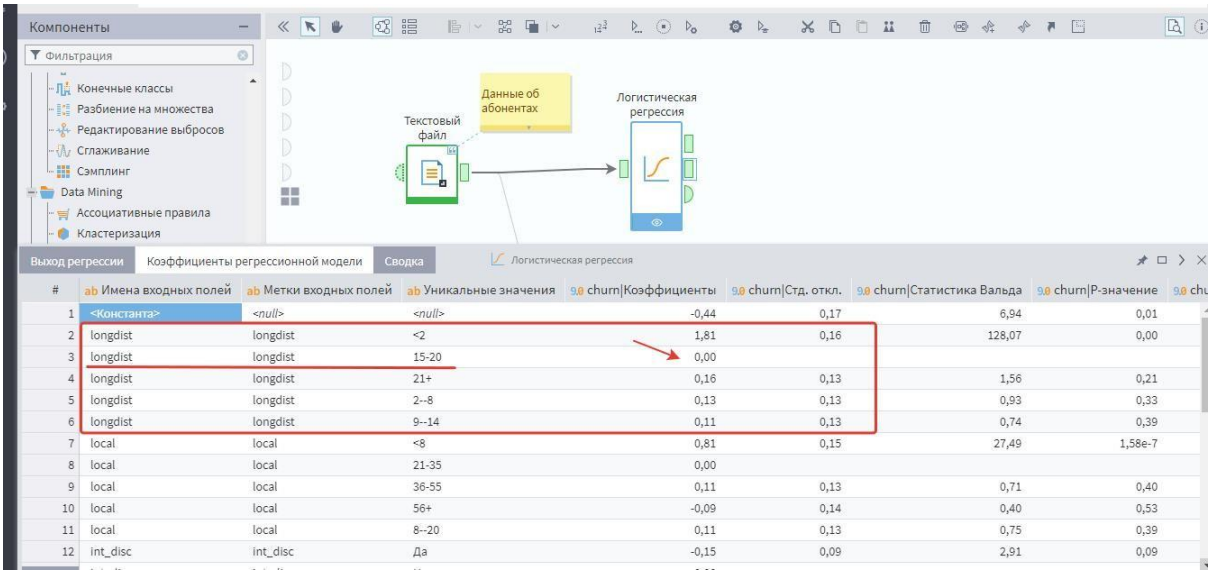
Рисунок 5.7 - Настройки логистической регрессии

Сохраним настройки. Вызовем контекстное меню для узла Логистическая регрессия и переобучим узел. Кликнув мышкой на любом из выходных портов Логистической регрессии, вызовем окно с вкладками этих выходов в рабочую область Сценария. Как и у узла Линейной регрессии их три: Выход регрессии (рис. 5.8), Коэффициенты регрессионной модели (рис. 5.9) и Сводка (рис. 5.10). Как и при построении линейной регрессии в логистической категориальные переменные заменяются dummy-переменными, принимающими значения 0/1, количество которых на единицу меньше уровней категориальной переменной. Так, у переменной longdist пять уникальных значений (пять уровней) и ее замещают четыре dummy-переменные, в каждой из которых только один из уровней отличен от нуля (рис. 5.9).



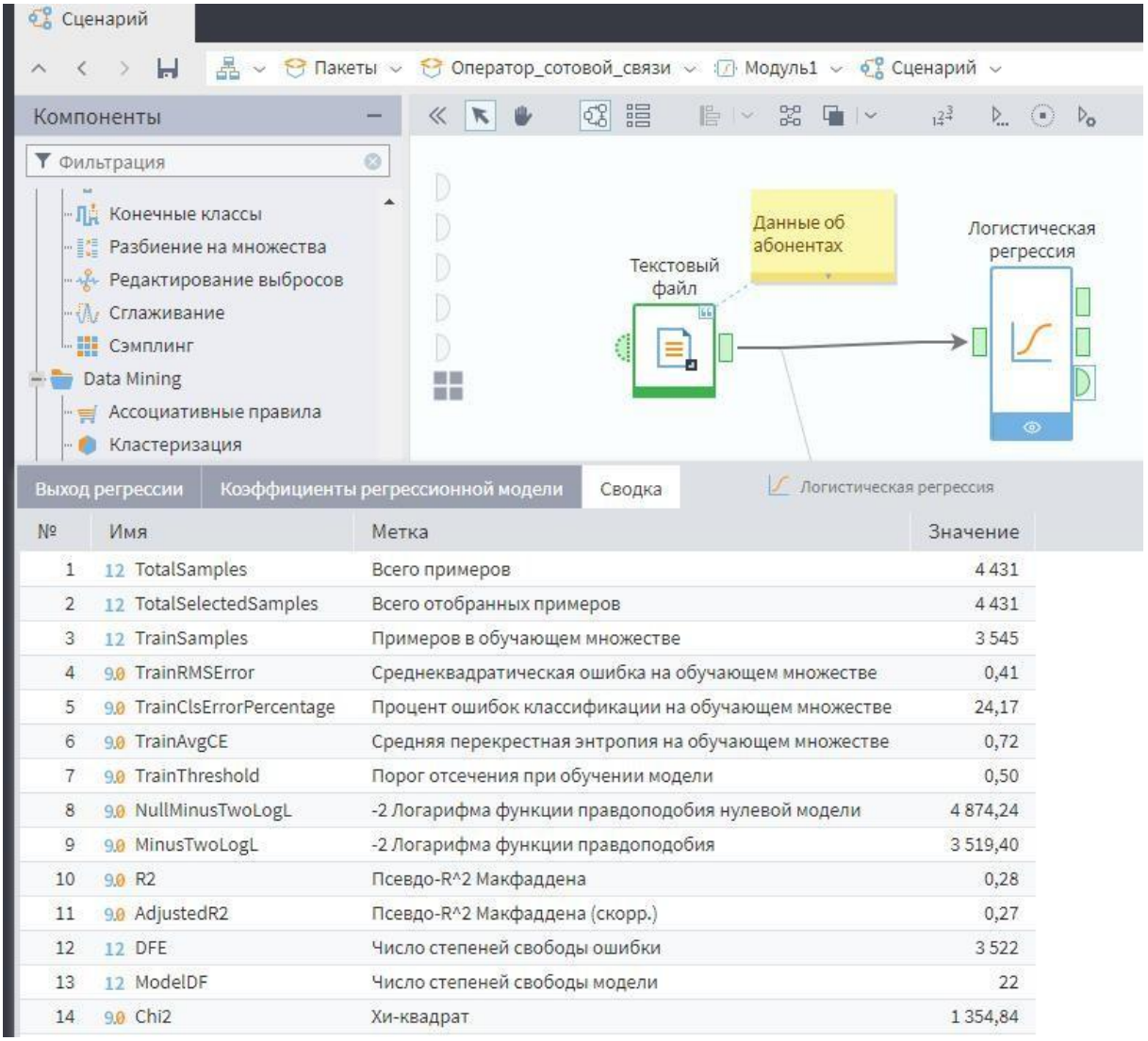
#	Событие|Прогноз	Вероятность события|Прогноз	Событие|Факт	churn|Прогноз	longdist	local	int_disc	billtype	pay	gender	marital
1	true	0,67	true	Уходит	<2	<8	Нет	Бюджетный	CC	Мужской	Женат
2	true	0,70	true	Уходит	<2	<8	Нет	Бесплатный	CC	Мужской	Женат
3	true	0,70	true	Уходит	<2	<8	Нет	Бесплатный	CC	Мужской	Женат
4	true	0,74	false	Уходит	<2	<8	Нет	Бесплатный	CH	Мужской	Одинокий
5	true	0,67	true	Уходит	<2	<8	Нет	Бесплатный	Auto	Мужской	Одинокий
6	true	0,96	true	Уходит	<2	<8	Нет	Бесплатный	Auto	Женский	Одинокий
7	true	0,67	true	Уходит	<2	<8	Нет	Бюджетный	CC	Мужской	Женат
8	true	0,67	true	Уходит	<2	<8	Нет	Бесплатный	Auto	Мужской	Одинокий
9	true	0,71	true	Уходит	<2	<8	Да	Бесплатный	CH	Мужской	Одинокий
10	true	0,70	true	Уходит	<2	<8	Нет	Бесплатный	CC	Мужской	Женат
11	true	0,74	false	Уходит	<2	<8	Нет	Бесплатный	CH	Мужской	Одинокий
12	true	0,96	true	Уходит	<2	<8	Нет	Бюджетный	CC	Женский	Женат

Рисунок 5.8 - Выход регрессии



#	Имена входных полей	Метки входных полей	Уникальные значения	churn|Коэффициенты	churn|Стд. откл.	churn|Статистика Вальда	churn|P-значение
1	<Константа>	<null>	<null>	-0,44	0,17	6,94	0,01
2	longdist	longdist	<2	1,81	0,16	128,07	0,00
3	longdist	longdist	15-20	0,00			
4	longdist	longdist	21+	0,16	0,13	1,56	0,21
5	longdist	longdist	2--8	0,13	0,13	0,93	0,33
6	longdist	longdist	9--14	0,11	0,13	0,74	0,39
7	local	local	<8	0,81	0,15	27,49	1,58e-7
8	local	local	21-35	0,00			
9	local	local	36-55	0,11	0,13	0,71	0,40
10	local	local	56+	-0,09	0,14	0,40	0,53
11	local	local	8--20	0,11	0,13	0,75	0,39
12	int_disc	int_disc	Да	-0,15	0,09	2,91	0,09

Рисунок 5.9 - Коэффициенты регрессионной модели



№	Имя	Метка	Значение
1	TotalSamples	Всего примеров	4 431
2	TotalSelectedSamples	Всего отобранных примеров	4 431
3	TrainSamples	Примеров в обучающем множестве	3 545
4	TrainRMSError	Среднеквадратическая ошибка на обучающем множестве	0,41
5	TrainClsErrorPercentage	Процент ошибок классификации на обучающем множестве	24,17
6	TrainAvgCE	Средняя перекрестная энтропия на обучающем множестве	0,72
7	TrainThreshold	Порог отсечения при обучении модели	0,50
8	NullMinusTwoLogL	-2 Логарифма функции правдоподобия нулевой модели	4 874,24
9	MinusTwoLogL	-2 Логарифма функции правдоподобия	3 519,40
10	R2	Псевдо-R^2 Макфаддена	0,28
11	AdjustedR2	Псевдо-R^2 Макфаддена (скорр.)	0,27
12	DFE	Число степеней свободы ошибки	3 522
13	ModelDF	Число степеней свободы модели	22
14	Chi2	Хи-квадрат	1 354,84

Рисунок 5.10 - Сводка

Но для узла Логистическая регрессия есть особый визуализатор, который позволяет получить все параметры качества построенной модели на одном экране (рис. 5.11).

Войдем в этот визуализатор (рис. 5.12). Правая верхняя таблица называется Оценки классификации (размер визуализатора больше экрана).

Можем проанализировать, сколько ошибок допускает классификатор, какие это ошибки: неверно определен ушедший абонент или оставшийся. Можем оценить качество прогноза на тестовом множестве и на обучающем.

Приступим к настройке узла Нейросеть (классификация). Все показатели, кроме churn (оттока абонентов), являются входными (рис. 5.13). При редактировании значений столбца Назначение для дискретных переменных возможен выбор и варианта входной и варианта выходной, так как в модели Нейросеть (классификация) выходом может быть

только категориальный дискретный показатель, а все переменные набора именно такие. В качестве входа в узле Нейросеть (классификация) могут быть как дискретные, так и непрерывные переменные.

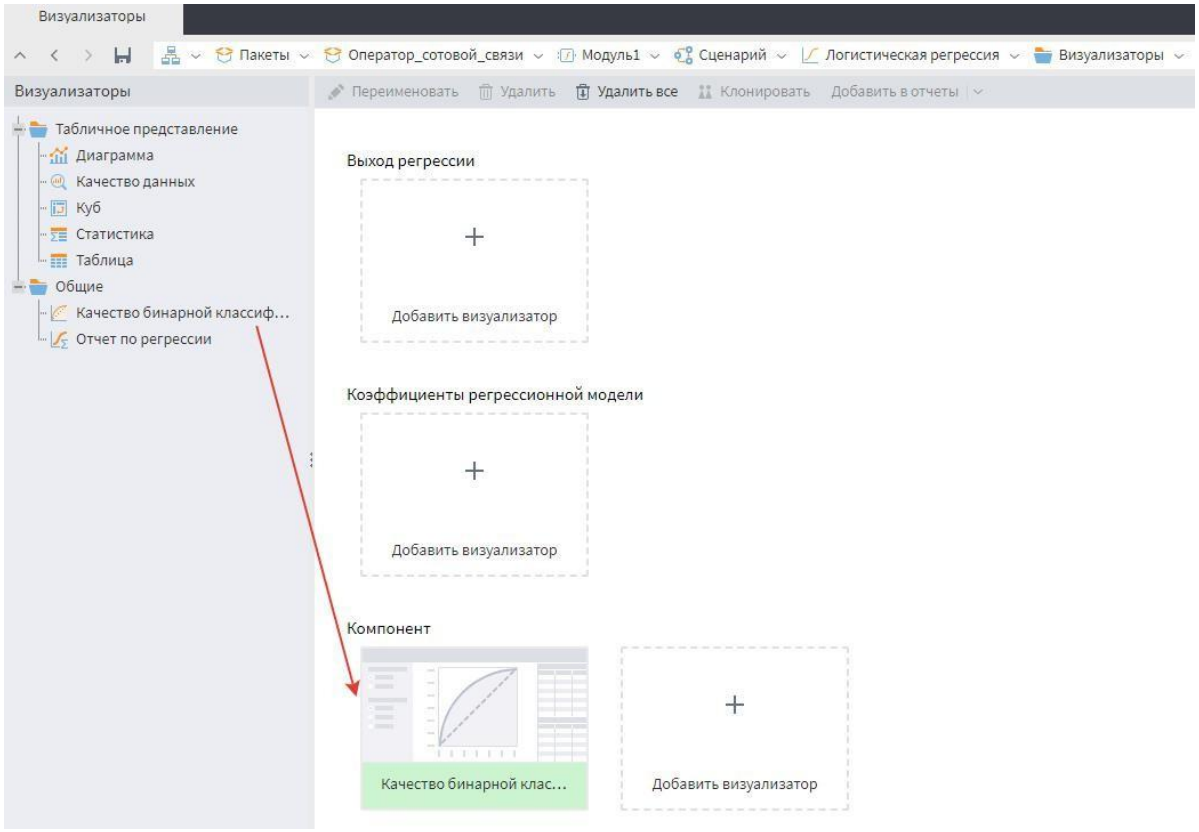


Рисунок 5.11 - Визуализатор Качество бинарной классификации

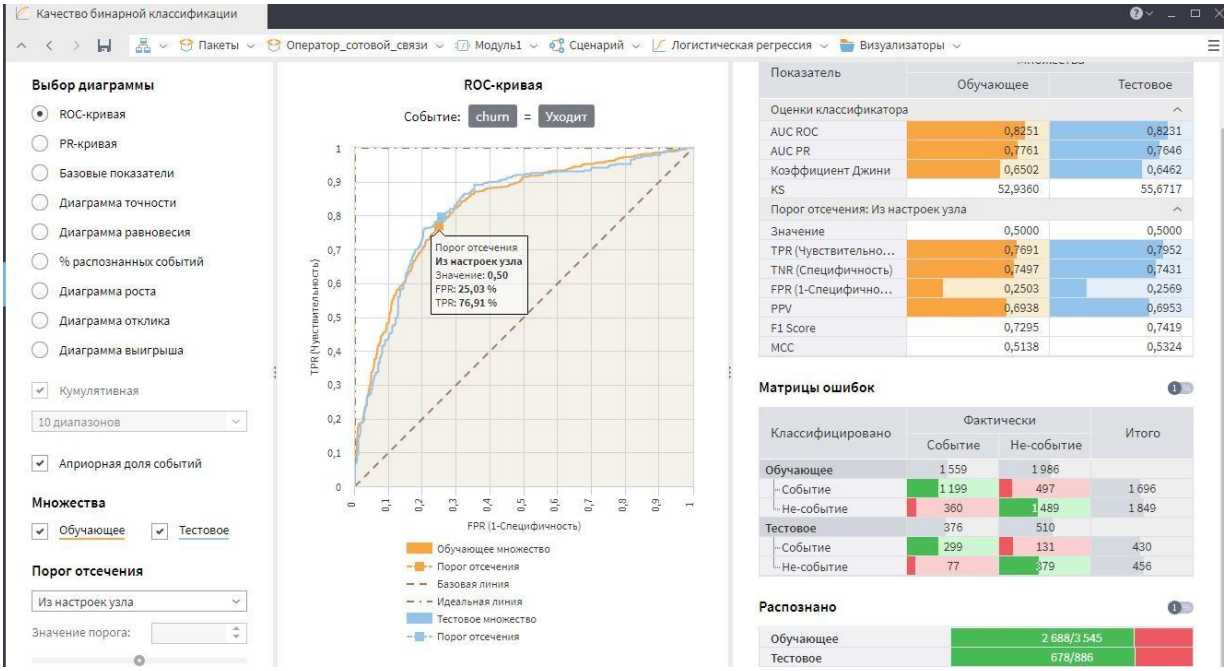


Рисунок 5.12 - Визуализатор Качество бинарной классификации

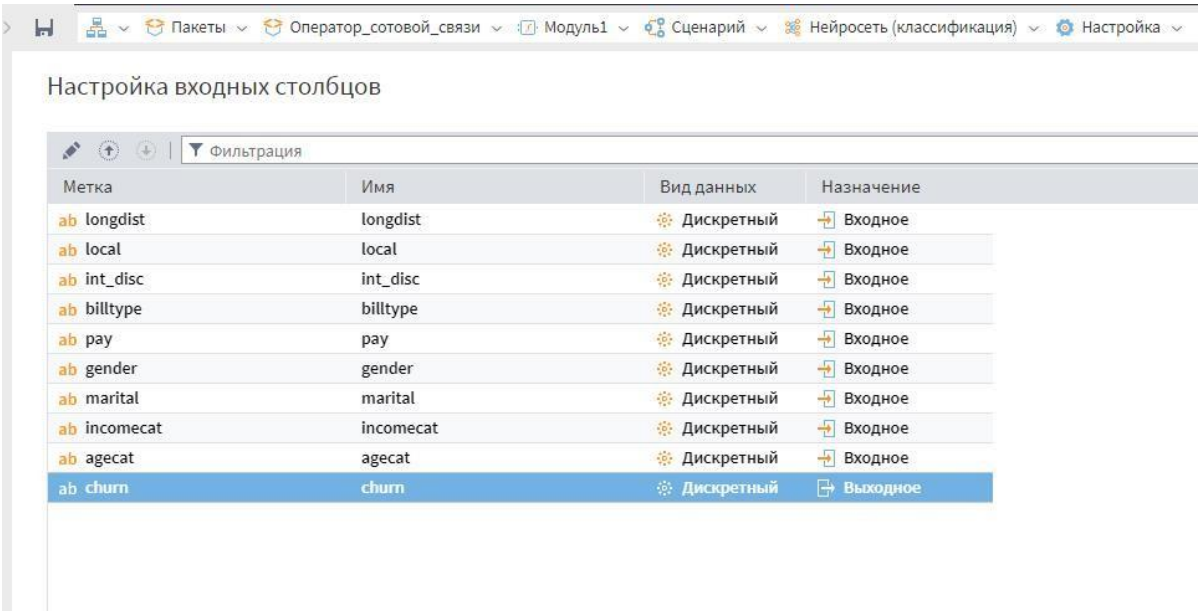


Рисунок 5.13 - Настройка входных столбцов

Настройки нормализации пропустим. В разбиении на множества разделим набор данных случайным образом в отношении 8/2:

– 80 % наблюдений будут использоваться для построения модели;

– 20 % – для тестирования ее качества, т.е. определения того, насколько сильно отклоняются для наблюдений тестового множества прогноз выходного параметра и его значение в наборе данных для соответствующего наблюдения.

В настройках параметров нейросети изменим количество нейронов в первом скрытом слое на 5 (рис. 5.14).

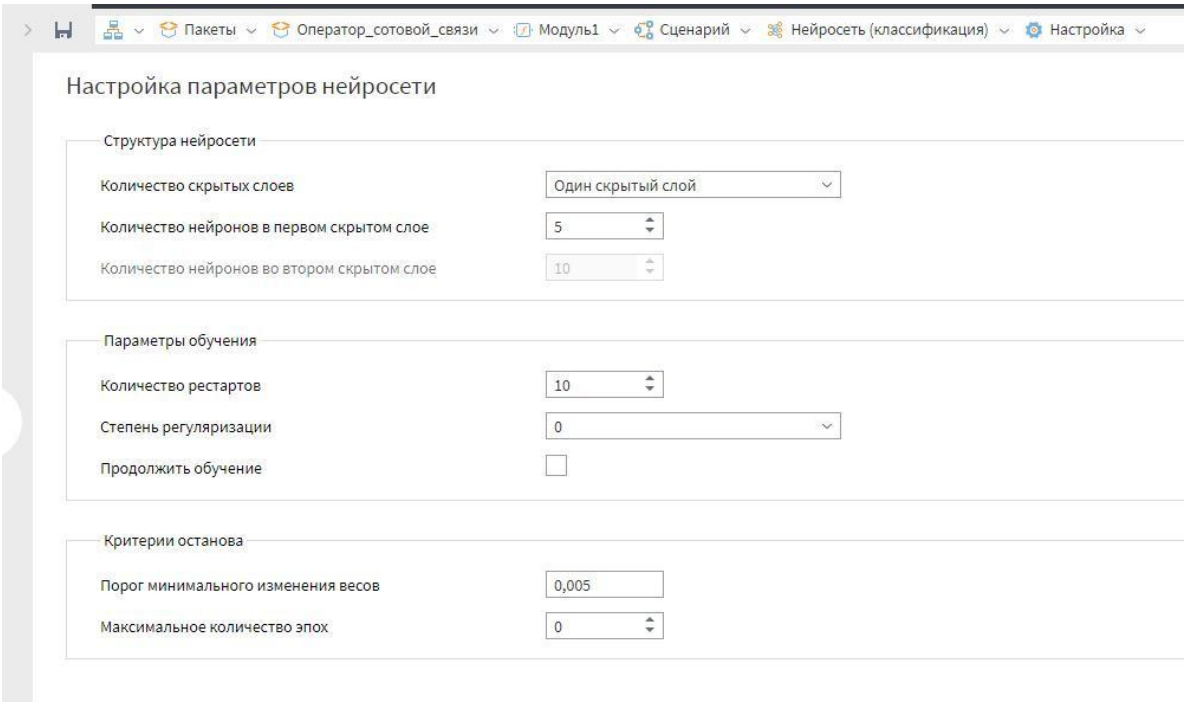


Рисунок 5.14 - Настройка параметров нейросети

Настройки автоматического подбора параметров Нейросети пропустим. Описание дополнительных параметров настройки узла Нейросеть (классификация) можно найти в соответствующем источнике [12].

Сохраним настройки. Вызовем контекстное меню для узла Нейросеть (классификация) и переобучим узел. Кликнув мышкой на любом из выходных портов Нейросеть (классификация), вызовем окно с вкладками этих выходов в рабочую область Сценария. Как и у узла Нейросеть (регрессия) их два: Выход нейросети (рис. 5.15) и Сводка.

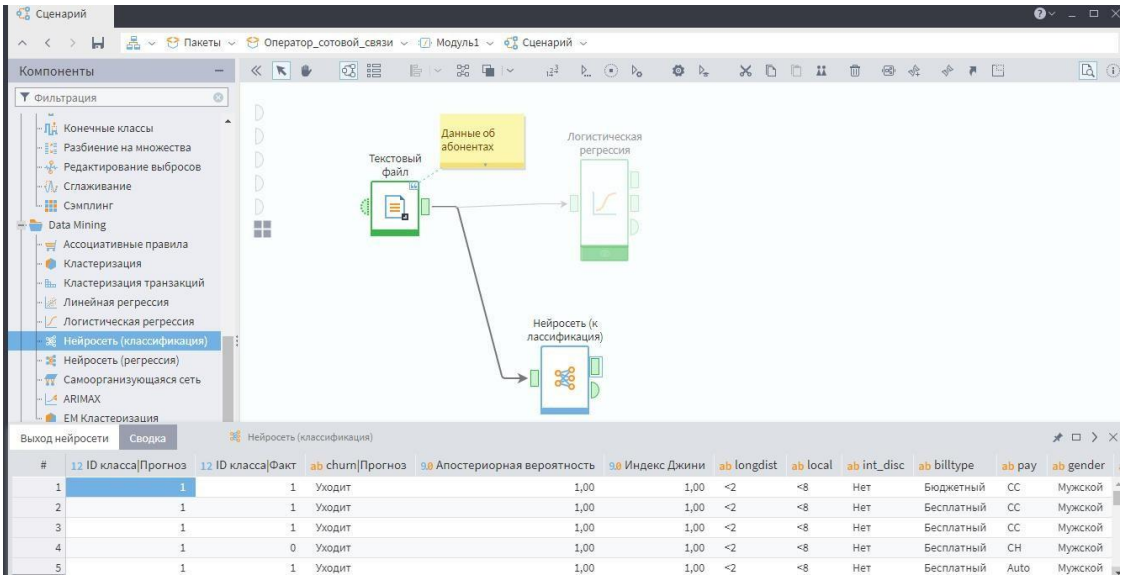


Рисунок 5.15 - Выход нейросети

Добавим в узел Нейросеть (классификация) визуализатор Куб (рис. 5.16).

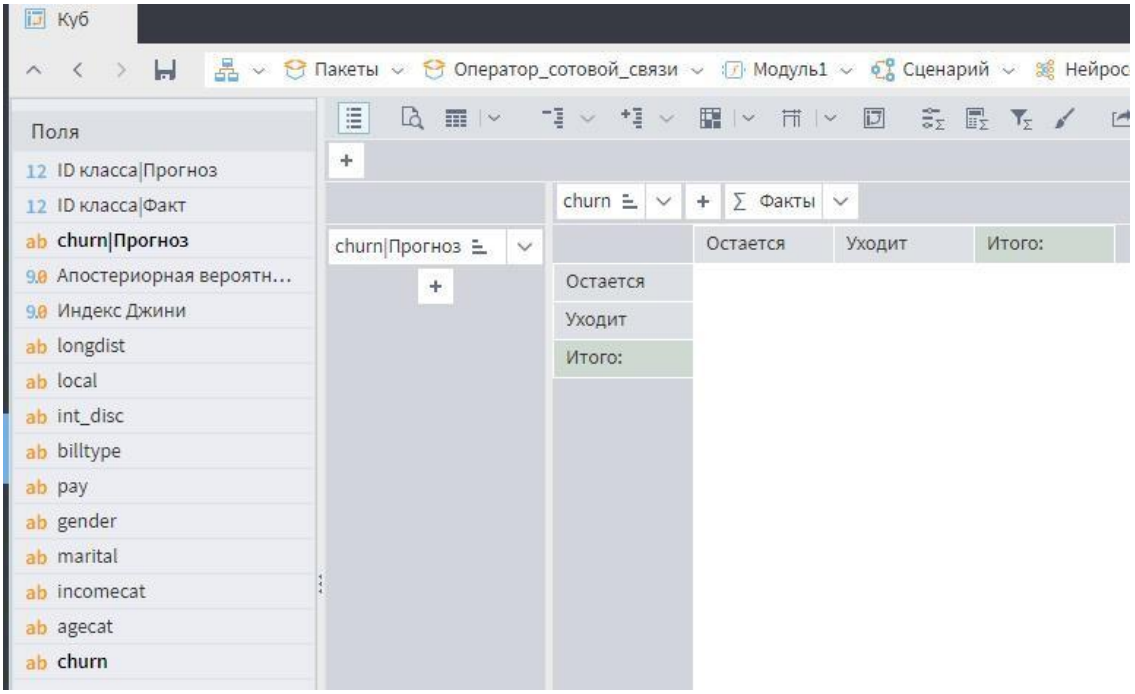


Рисунок 5.16 - Визуализатор Куб

В контекстном меню для фактов (стрелка рядом с кнопкой Факты над построенной таблицей без значений) выберем опцию Настроить факты и в соответствующем диалоговом окне поставим галочку рядом с фактом Количество (рис. 5.17).

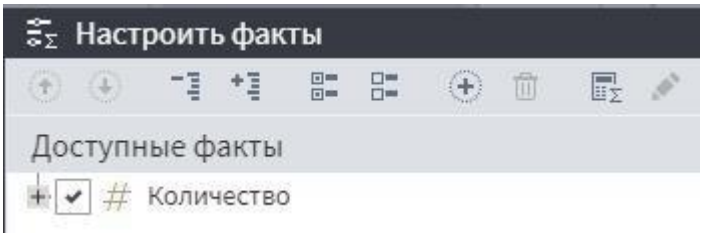


Рисунок 5.17 - Добавление факта

Получим таблицу сопряженности для выходного показателя churn и прогноза по модели Нейросеть (классификация) этого показателя (рис. 5.18). По данным этой таблицы можно рассчитать показатели качества прогноза.

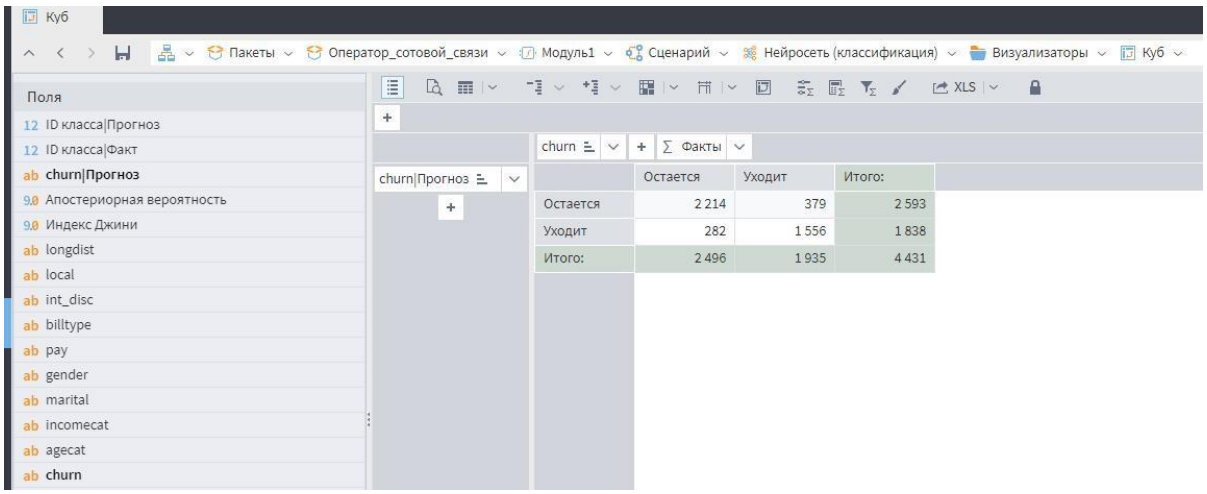


Рисунок 5.18 - Визуализатор Куб с матрицей ошибок прогноза

Контрольные вопросы

1. Что представляет собой логистическая регрессия.
2. Основная идея линейного классификатора.
3. Как провести настройку узла Нейросеть (классификация)?
4. Как провести настройку узла логистическая регрессия, основные шаги.

Лабораторная работа 5. Анализ данных с использованием Loginom

Цель работы: научиться использовать ML-платформу Loginom для решения бизнес-задач.

Формируемые знания, умения и навыки: научиться работать с данными с использованием ML-платформы Loginom, уметь создавать пакеты, создавать сценарии, настраивать узлы, визуализаторы, владеть навыками извлечения из данных полезных для бизнеса знаний.

Содержание работы:

1. С сайта https://www.kaggle.com/ (либо из найденного самостоятельно интернет-ресурса) импортировать один из наборов данных, включающий столбцы с числовыми значениями.
2. Используя возможности ML-платформы Loginom, импортировать данные, провести анализ их качества, построить визуализации (куб, статистика, диаграмма).
3. Использовать для анализа данных один из узлов категории Data Mining. Интерпретировать полученные результаты.

Контрольные вопросы

1. Как произвести импорт данных?
2. Как настраиваются визуализаторы для узла?
3. Чем отличаются друг от друга факты и измерения? Какие типы данных есть в Loginom Community? Какие виды данных определены в Loginom Community?
4. Опишите настройки узла из категории Data Mining, который использовали для анализа данных. Какие входные порты есть у узла? Какие выходные порты есть у узла? Какие визуализаторы были настроены для этого узла? Обосновать выбор визуализаторов.
5. Какую ценность для бизнеса можно извлечь из проведенного анализа данных?

Список литературы

1. Loginom. URL: https://loginom.ru/ (дата обращения: 10.11.2021).
2. Рындина С. В. Интеллектуальные информационные системы и технологии: системы Business Intelligence (Microsoft Power BI) : учеб.-метод. пособие. Пенза : Изд-во ПГУ, 2021. 64 с.
3. Руководство пользователя. URL: https://help.loginom.ru/user guide/
4. Измерение (Dimension). Loginom Wiki. URL: https://wiki.loginom.ru/articles/dimension.html?_ga=2.164394623.1710229773.1642063066-1779185901.1642063066 (дата обращения: 09.12.2021).
5. Факт (Fact). Loginom Wiki. URL: https://wiki.loginom.ru/articles/fact.html?_ga=2.239564611.1710229773.1642063066-1779185901.1642063066 (дата обращения: 09.12.2021).
6. Кластеризация. Loginom Help. URL: https://help.loginom.ru/user guide/processors/datamining/clustering.html (дата обращения: 09.12.2021).
7. Линейная регрессия. Loginom Help. URL: https://help.loginom.ru/userguide/processors/datamining/linear-regression/ (дата обращения: 09.12.2021).
8. Нейросеть (регрессия). Loginom Help. URL: https://help.loginom.ru/userguide/processors/datamining/neural-network-regression.html (дата обращения: 09.12.2021).
9. Рындина С. В. Бизнес-аналитика на основе больших данных: обучение с учителем на языках Python и R : учеб.-метод. пособие. Пенза : Изд-во ПГУ, 2020. 80 с.
10. Груздев А. В. Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес : руководство. М. : ДМК Пресс, 2018. 642 с. URL: https://e.lanbook.com/book/123700 (дата обращения: 09.12.2021).
11. Логистическая регрессия. Loginom Help. URL: https://help.loginom.ru/userguide/processors/datamining/logistic-regression/?_ga=2.38777443.1710229773.1642063066-1779185901.1642063066 (дата обращения: 09.12.2021).
12. Нейросеть (классификация). Loginom Help. URL: https://help.loginom.ru/userguide/processors/datamining/neural-network-classification.html (дата обращения: 09.12.2021).
13. Рындина С. В. Интеллектуальные информационные системы и технологии: системы Business Intelligence (ML-платформа Loginom): учеб.-метод. пособие. Пенза : Изд-во ПГУ, 2022.