МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ Н.Э. БАУМАНА
Факультет "Информатика и системы управления"
Кафедра "Системы обработки информации и управления"

Лабораторная работа № 6

"Знакомство с платформой для анализа данных Loginom"

Москва — 2019

Оглавление

1. Цель работы

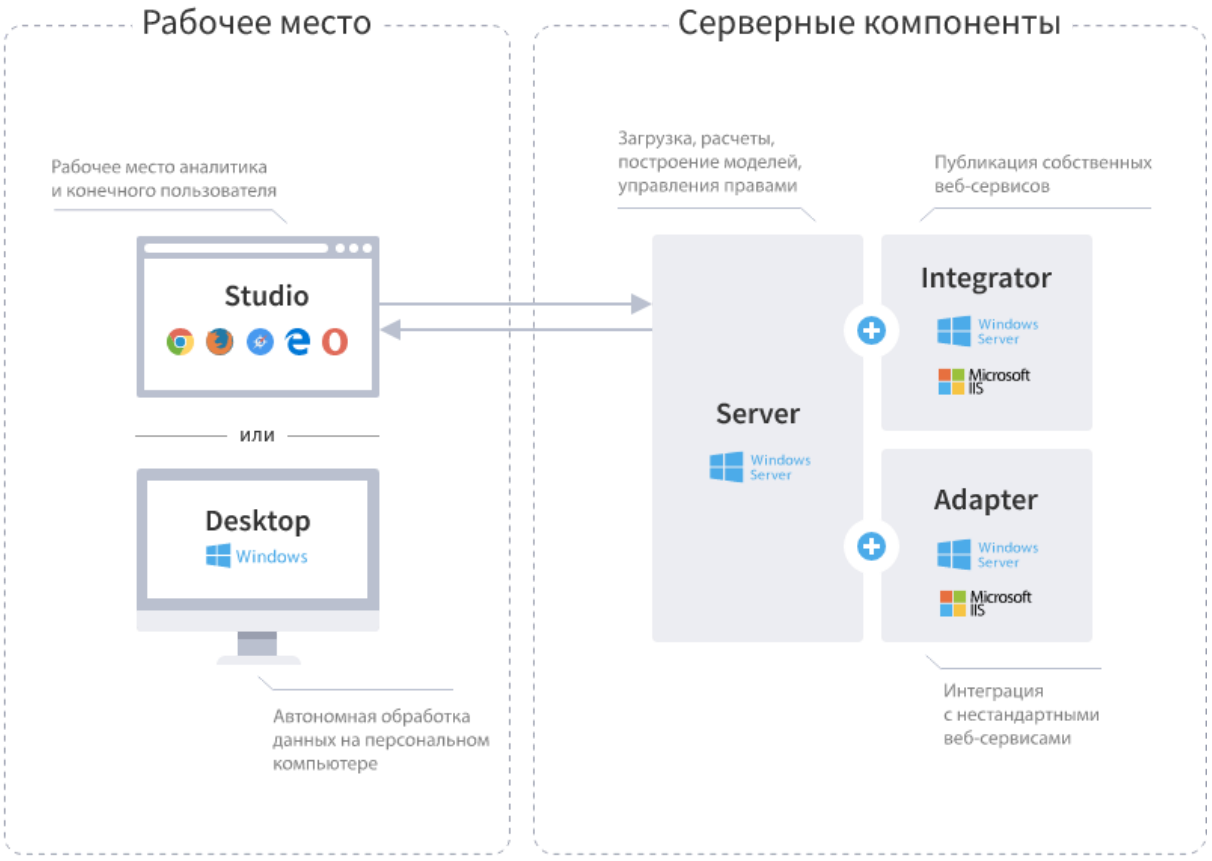
Целью работы является знакомство студентов с аналитической средой Loginom и приобретение ими навыков проведения бизнес-анализа данных и практического применения простейших алгоритмов машинного обучения.

2. Теоретическая часть

2.1. Общие сведения

Loginom — аналитическая платформа, позволяющая в единой среде выполнить все этапы бизнес-анализа от консолидации данных и построения моделей до визуализации и интеграции в бизнес-процесс.

Компоненты платформы Loginom



Для решения задач анализа Studio позволяет импортировать данные из различных источников и применять к ним необходимые алгоритмы

обработки. Результаты можно просмотреть в самой системе или экспортировать в сторонние приемники данных.

Таким образом, Studio может использоваться как для создания автономных аналитических решений, так и для разработки модулей, интегрируемых со сторонними системами.

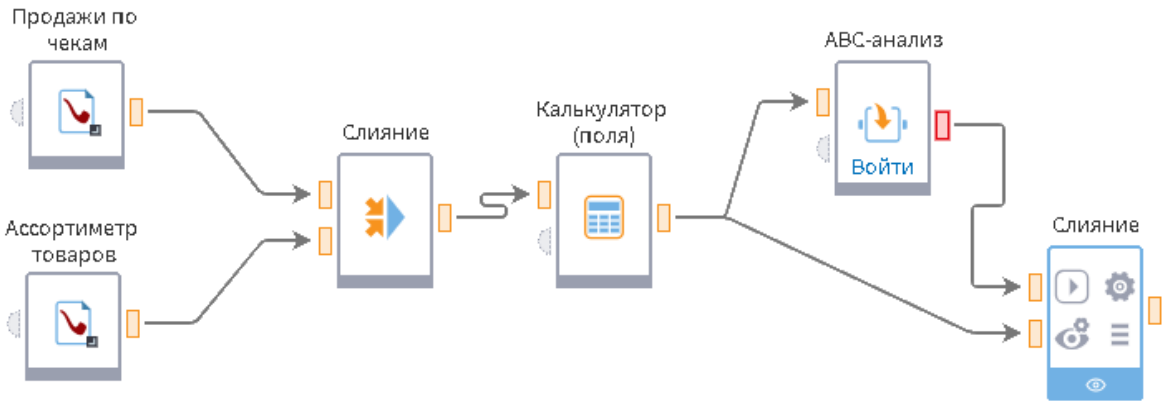
Одной из основных концепций, на которых базируется Studio, является Сценарий.

Сценарий — последовательность действий, которые необходимо провести для анализа данных. Он представляет собой комбинацию узлов обработки данных, настраиваемую пользователем для решения конкретной задачи.

Узел сценария выполняет отдельную операцию над данными. Перечень возможных операций представлен палитрой готовых *компонентов*.

Последовательность обработки задается соединением выхода предыдущего узла сценария с входом последующего. Входом и выходом обработчика являются *входные* и *выходные порты*.

Пример сценария



Узлы сценария создаются из компонентов 2-х типов:

- **Стандартные компоненты** — предоставляются в рамках платформы;

- **Производные компоненты** — создаются и настраиваются пользователем.

Производный компонент можно создать из комбинации узлов сценария, реализующих произвольную логику обработки.

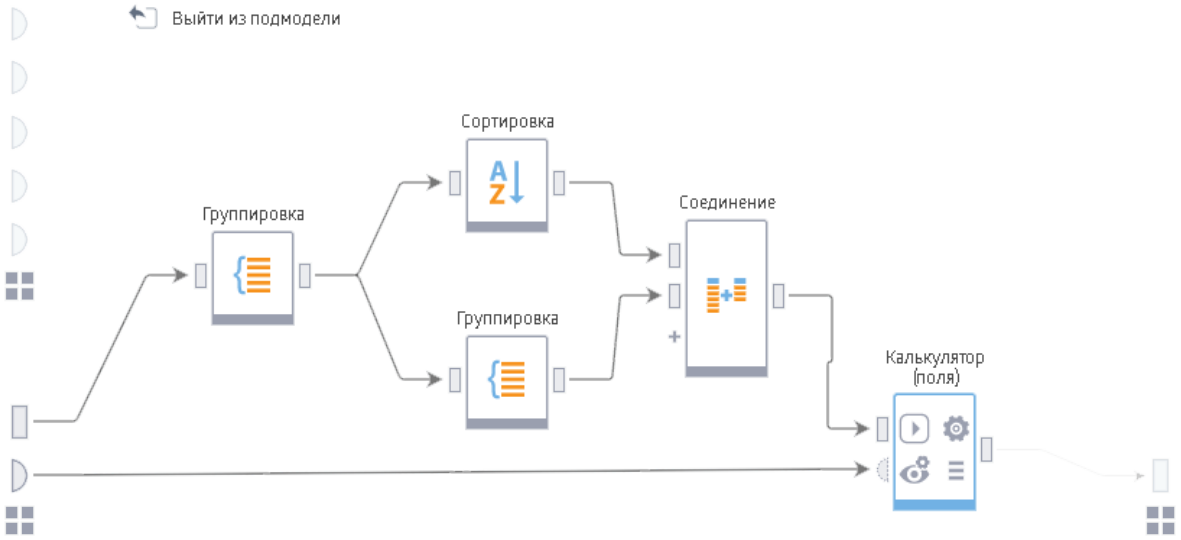
Таким образом, набор средств для реализации различной логики обработки данных не ограничивается стандартными компонентами платформы и может быть расширен самим пользователем.

Чаще всего для создания производного компонента используется Подмодель.

Подмодель является специальным узлом, способным включать в себя другие узлы сценария. Реализованная в подмодели логика может быть произвольной, при этом разработчик сценария может рассматривать её как «чёрный ящик». Подмодель принимает информацию через входные порты, производит обработку и выдает результат на выходные порты. Входные и выходные порты задаются пользователем.

На рисунке «Пример сценария» узел “ABC-анализ” является производным компонентом — подмоделью.

Узлы подмодели



2.2. Назначение и структура пакета

Все действия с проектом в Studio осуществляются в рамках Пакета, который является минимальной единицей поставки и представляет собой контейнер для компонентов, сценариев, подключений и т.д.

Пакеты сохраняются по-отдельности в виде файлов с расширением .lgp, и включают в себя Ссылки и Модули.

Ссылки применяются для подключения других пакетов с целью использования созданных в них производных компонентов и подключений в текущем проекте. Соответствующие объекты доступны только в том случае, когда они опубликованы для общего доступа.

Каждый пакет содержит хотя бы один модуль. Модуль включает в себя:

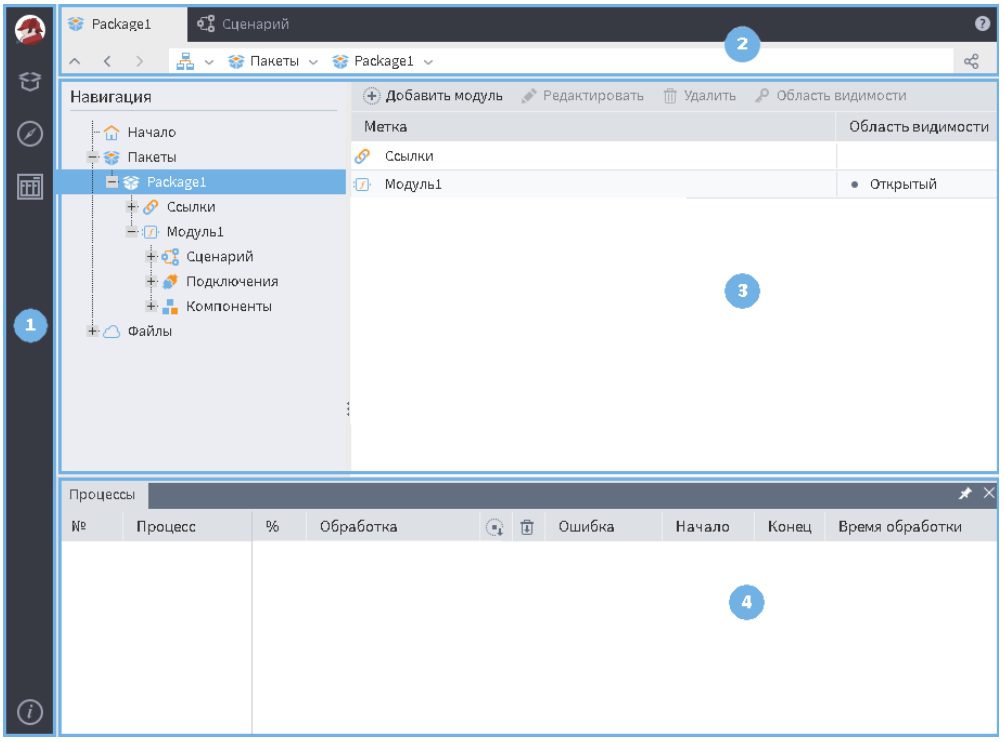
- Сценарий — содержит последовательность узлов обработки данных;
- Подключения — в них представлен список внешних источников и приемников данных, к которым можно подключиться;
- Компоненты — включают в себя доступные для работы подмодели, как созданные в рамках текущего пакета, так и заимствованные из других пакетов через ссылки.

2.3. Рабочее пространство программы

Всё рабочее пространство данного инструмента можно разделить на четыре основных области:

1. Слева расположено главное меню с кнопками: Меню, Пакеты, Навигация, Файлы, Процессы.
2. Верхняя часть отображает вкладки открытых пакетов, содержит адресную строку и элементы для навигации по пакетам и их составляющим.
3. Справа от главного меню располагается рабочий стол. Он включает левую панель, где отображаются рабочие компоненты и структура решения (пакеты и их составные части), а также непосредственно область построения сценария и визуализации данных.

4. В нижней части окна расположена панель Процессы. По умолчанию она скрыта, но ее можно закрепить.



или :

1. Главное меню — панель с кнопками для манипуляции с различными настройками;
2. Адресная строка — строка, содержащая путь к открытому объекту;
3. Рабочая панель — панель компонентов, панель инструментов и область построения Сценариев;
4. Панель “Процессы” — панель, содержащая подробную информацию о происходящих/происходивших процессах текущей сессии

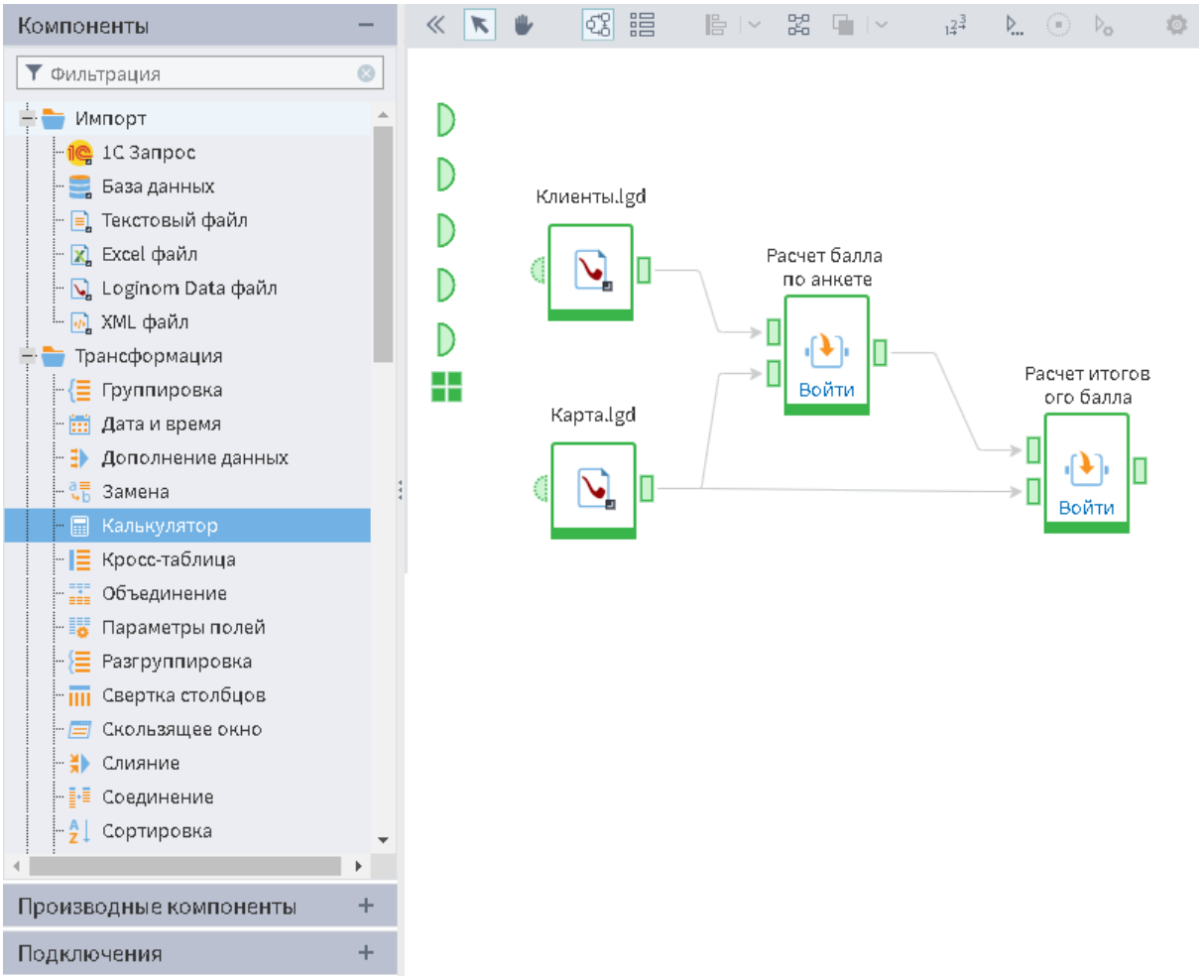
Для того чтобы использовать в сценарии какой-либо компонент, его необходимо перенести мышью из панели компонентов в область построения сценария.

В дальнейшем подмодель, выполняющая заданную пользователем функцию, может быть опубликована как производный компонент и

наравне со стандартными компонентами многократно использоваться в других сценариях.

2.3.1. Рабочее пространство

Рабочее пространство состоит из панели компонентов, области построения Сценария и панели инструментов.



Слева находится панель компонентов, состоящая из следующих категорий:

- Компоненты — стандартная библиотека базовых компонентов Loginom;
- Производные компоненты — создаваемые пользователями компоненты на основе базовых;
- Подключения — источники данных.

По центру расположена область построения Сценария — полотно, содержащее узлы Сценария и связи между ними.

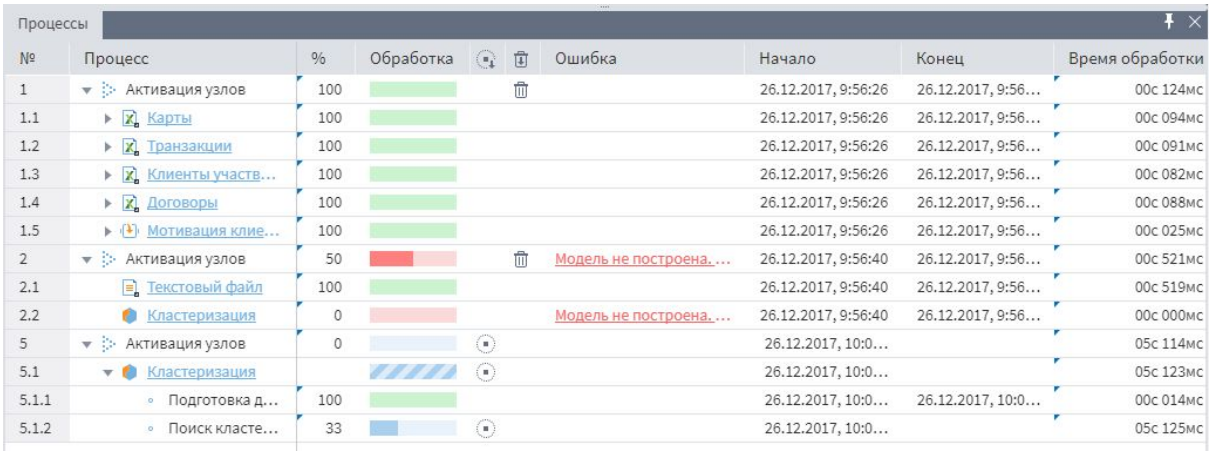
Сверху расположена панель инструментов, содержащая следующие операции для манипуляции с областью построения и ее составляющими:

- Показать/Скрыть панель компонентов — позволяет открыть или закрыть панель компонентов;
- Режим выбора объекта — режим, использующийся для построения Сценария с помощью стандартных манипуляций;
- Режим навигации по сценарию — режим, использующийся для навигации по области построения Сценария с помощью мышки;
- Показать в виде сценария — отображает Сценарий в стандартном виде (в виде направленного графа);
- Показать в виде таблицы — компактное отображание Сценария в виде таблицы, содержащей используемые элементы;
- Вертикальное выравнивание — позволяет выравнить вертикально узлы Сценария на области построения. имеются следующими виды вертикального выравнивания:
    - По левому краю;
    - По середине;
    - По правому краю;
    - По верхнему краю;
    - По центру;
    - По нижнему краю.
- Автоматическое упорядочивание узлов — автоматическое расположение узлов на области Сценария в соответствии с их последовательностью обработки данных;
- Переместить выделенные узлы — выставляет выделенные узлы и их подписи на:
    - Передний план;
    - Задний план.
- Настроить порядок выполнения — позволяет задать собственный порядок выполнения узлов;
- Выполнить все — выполнить все узлы Сценария;

- Активировать/Деактивировать узел — активировать/деактивировать узел;
- Переобучить узел — переобучает выделенный узел;
- Настроить узел — заходит в настройки выделенного узла;
- Настроить режим активации узла — настройка режима активации выделенного узла;
- Клонировать узел — клонирование выделенного узла;
- Развернуть/Свернуть подмодель — позволяет свернуть выделенные узлы в Подмодель или развернуть выделенную Подмодель на составные узлы;
- Удалить выбранное — удаляет выделенные узлы/связи Сценария;
- Создать производный компонент — создает Производный компонент на основе выделенного узла;
- Показать родительские узлы для производных — при наличии производных узлов показывает родительские узлы;
- Показать исходные узлы для Узлов-ссылок — при наличии Узлов-ссылок показывает узлы, на основе которых они создавались;
- Показать карту сценария — для навигации открывается уменьшенная копия области построения Сценария с возможностью масштабирования.

2.3.2. Панель “Процессы”

Панель «Процессы» предназначена для получения дополнительной информации о процессах обработки данных узлами Сценариев в рамках текущей сессии Loginom. В панели иерархически фиксируются процессы и выполняющиеся в них узлы/подузлы, им присваиваются порядковые номера. Каждый новый процесс начинается со строки "Активация узлов", это обусловлено тем, что при старте работы Сценария параллельно могут выполняться сразу несколько узлов.



Структура панели следующая:

- № — номер по порядку.
- Процесс — иерархически указаны процессы (наименования) и их составные части. При нажатии на название процесса в Сценарии будет найден и выделен узел, который отвечает за его выполнение.
- **%** — процент выполнения процесса.
- Обработка:
  - — идет процесс обработки, и время его окончания невозможно рассчитать заранее;
  - — прогресс выполнения текущего процесса по мере обработки данных;
  - — успешное выполнение (обработка завершена);
  - — информация о результатах выполнения процесса еще не получена;
  - — при выполнении процесса произошла ошибка (обработка не завершена);
  - — выполнение процесса не начиналось, произошла ошибка.

Ошибка — указывается текст ошибки в случае ее возникновения. При нажатии на данный текст на экран выведется полная формулировка ошибки.

- Начало — дата и время начала процесса.
- Конец — дата и время окончания процесса.

- Время обработки — разница между началом и концом процесса обработки.

При выборе процесса на информационной панели можно вызвать контекстное меню, содержащее следующие действия:

- **Отменить** — останавливает выполнение выбранного процесса;
- **Отменить выполнение всех процессов...** — останавливает выполнение всех процессов;
- **Удалить из списка** — удаляет процесс из списка информационной панели;
- **Удалить все заверенные процессы...** — удаляет все завершенные процессы из списка информационной панели;
- **Показать узел** — выделяет выбранный узел на области построения Сценариев;
- **Подробнее** — выводит на экран сообщение об ошибке передаваемое узлом;
- **Отображать завершенные процессы** — изменяет отображение процессов на панели:

По умолчанию информационная панель скрыта. Ее можно открыть, нажав на кнопку в левом нижнем углу.

2.4. Проектирование

Проект — комплекс сценариев, файлов, источников данных и прочих элементов, предназначенных для решения отдельной аналитической задачи.

Проект может объединять в себе несколько пакетов благодаря тому, что каждый пакет имеет возможность предоставлять свои объекты другим пакетам через механизм ссылок.

В основе построения проекта лежит методология *структурного проектирования* — представление алгоритма в виде иерархической структуры блоков.

Каждый блок на своем уровне иерархии может быть представлен в виде «черного ящика», выполняющего независимую подзадачу. Механизм решения подзадачи внутри «черного ящика» можно изменить, но в целом проект при этом останется работоспособным и будет выполнять поставленные задачи.

Спроектированный таким образом проект имеет четкую, легко читаемую структуру. Все это позволяет создавать и сопровождать сложные проекты, а также делегировать решение выделенных подзадач.

Особенностью подобного подхода является проектирование «сверху вниз» — от общей постановки задачи к отдельным подзадачам. На первом этапе проектирования описывают решение поставленной задачи, выделяя независимые подзадачи. На следующем аналогично описывают подзадачи, формулируя при этом элементы следующего уровня.

Таким образом, на каждом шаге происходит уточнение функций проекта. Процесс продолжают, пока не доходят до подзадач, алгоритмы, решения которых очевидны.

2.5. Декомпозиция

Структура Проекта может быть представлена в иерархическом виде:

- Проект может состоять из связанных между собой Пакетов — это возможно благодаря тому, что *каждый пакет может предоставлять свои объекты другим пакетам* через механизм ссылок.
- Пакет включает в себя Модули — декомпозиция пакета на уровне модулей.
- **Модуль** — сам по себе не содержит узлов обработки данных, но предоставляет отдельное пространство для Сценариев и Подключений к различным источникам данных.
- **Сценарий** — содержит последовательность узлов обработки данных. Сценарий может:

    - Включать в себя подпрограммы — Подмодель.
    - Получать данные от узлов из других сценариев и пакетов через механизм Узел-ссылка.
    - Использовать настройки и обученные модели узлов из других сценариев и пакетов через механизм Выполнение узла.
    - Использовать готовые алгоритмы обработки данных, созданные в других сценариях и пакетах через механизм Производные компоненты.
- **Подмодель** — включает в себя другие узлы, предоставляя, таким образом, отдельное пространство для реализации произвольного алгоритма обработки данных. Подмодель в сценарии представлена в качестве узла, имеющего заданные пользователем входные и выходные порты. Может содержать в себе иерархию вложенных подмоделей. На базе подмодели может быть создан Производный компонент.

2.6. Построение скоринговых карт

Скоринг уже давно хорошо зарекомендовал себя в банках и микрофинансовых организациях, однако, этот подход применяется и в других отраслях: например, в маркетинге, где посредством скоринга можно разделять покупателей на группы, согласно их ценности для магазинов, или в медицине, для предсказания вероятности тех или иных заболеваний.

Скоринговый подход стал стандартом де-факто в финансовой отрасли, так как выдача кредитов без должной оценки платежеспособности приводит к катастрофическому росту рисков и другим негативным последствиям.

Loginom на данный момент включает в себя все возможные этапы построения скоринговых карт:

Подготовка скоринговой выборки:

- загрузка счетов;
- загрузка характеристик;
- загрузка просрочек (опционально);
- загрузка флагов хороший/плохой (опционально);
- мониторинг корректности загрузки данных;
- формирование начальных выборок.

Анализ жизненных циклов:

- аудит данных по просрочкам;
- винтажный анализ;
- построение матрицы миграции;
- определение статуса счета;
- балансировка классов;
- загрузка статусов в витрину данных.

Двумерный анализ:

- квантование непрерывных данных;
- квантование дискретных данных;
- отбор значимых факторов;
- загрузка преобразованных атрибутов в витрину данных.

Моделирование:

- построение модели;
- оценка качества модели;
- преобразование коэффициентов регрессии в скоринговые баллы;
- загрузка карты в витрину данных.

2.7. Нормализация непрерывных данных

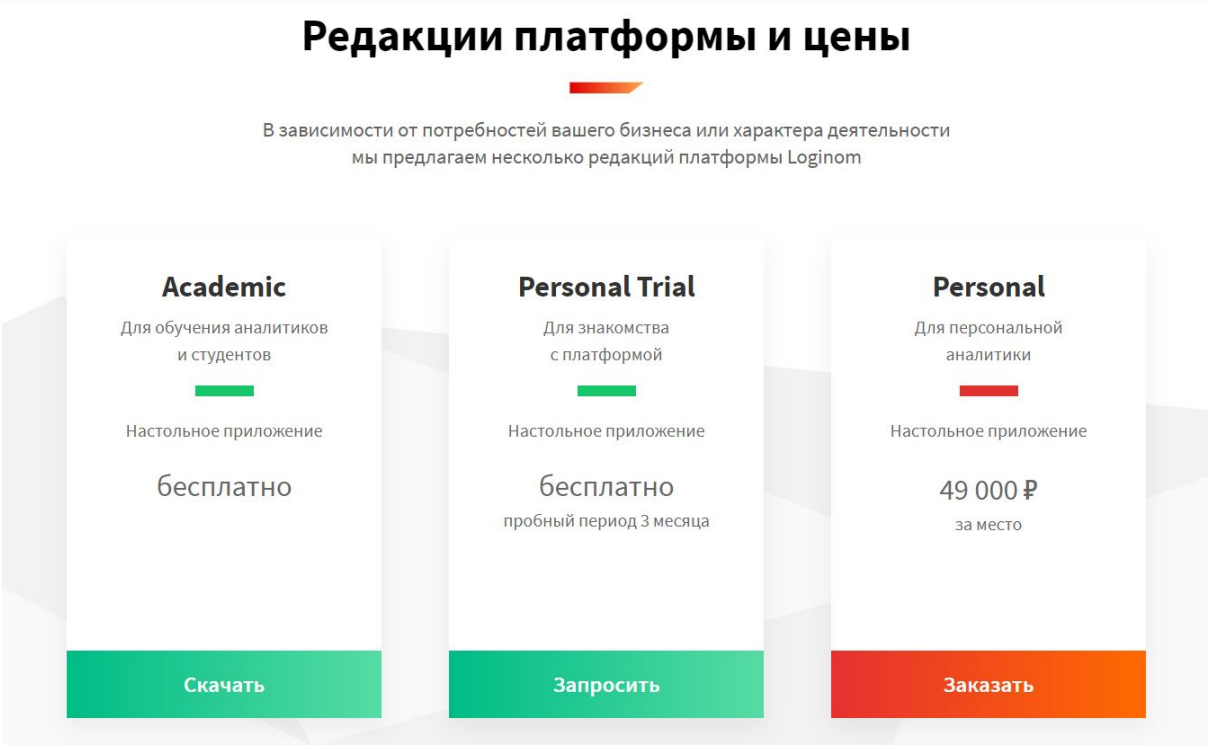
Варианты нормализации данных в Loginom:

- **Нет** — отсутствие нормализации. В таком случае данные поступают в основной алгоритм без предварительной обработки.
- **Масштабирование [min;max]** — приведение данных линейным преобразованием к заданному пользователем диапазону [min;max]:
  - **Минимум** — минимальное значение;
  - **Максимум** — максимальное значение.
- **Масштабирование [-1;1]** — приведение данных линейным преобразованием к диапазону [-1;1].
- **Масштабирование [0;1]** — приведение данных линейным преобразованием к диапазону [0;1].
- **Абсолютное масштабирование** — каждое значение делится на максимальное абсолютное значение.
- **Стандартизация** — из каждого значения вычитается среднее значение и делится на стандартное отклонение.
- **Отношение** — каждое значение делится на статистический показатель либо на заданное пользователем значение:
  - Делитель:

- Статистический показатель;
- Заданное значение.

3. Подготовка к выполнению работы

1. Установите программное обеспечение Loginom, перейдя по ссылке LOGINOM и выбрав для установки необходимую редакцию платформы (для данной работы подойдет Loginom Academic)



2. Загрузите и расположите в удобной директории используемые в работе источники данных, приложенные в архиве с настоящими методическими указаниями.

4. Ход выполнения работы

4.1. Прогнозирование стоимости недвижимости методом линейной регрессии

4.1.1. Импорт исходной таблицы данных

Создайте новый пакет Loginom в удобной директории. На появившуюся область построения сценария из списка компонентов в левой части рабочего пространства добавьте узел "Текстовый файл".

В настройках узла (нажатие левой кнопкой мыши по узлу — и затем по шестеренке) импортируйте файл housing.csv с заголовком в первой строке и нажмите "Далее".

В настройках формата импорта выберите точку в качестве десятичного разделителя, остальные настройки оставьте по умолчанию.

В параметрах импорта установите разделителем столбцов запятую. Установите вещественный **тип** данных для полей *longitude, latitude, median_income, median_house_value,* строковый тип — для поля *ocean_proximity,* а для остальных полей — целый тип. Для поля *ocean_proximity* установите дискретный **вид** данных, для остальных полей — непрерывный.

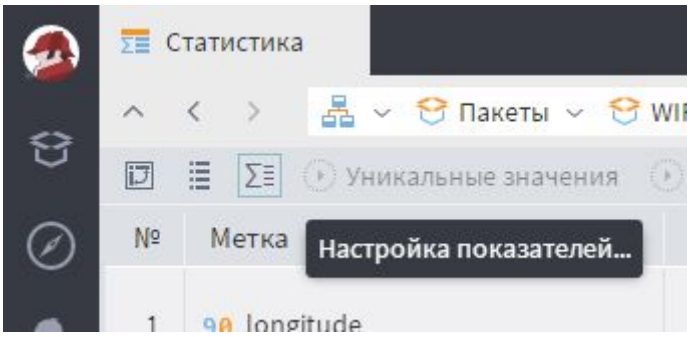
В настройках соответствия между столбцами в пункте связи убедитесь, что каждому входному столбцу сопоставлен подходящий выходной столбец.

Метку узла определите удобным образом (например, "Исходные данные о недвижимости").

Сохраните и запустите узел, нажав на кнопку "Выполнить".

4.1.2. Визуализация исходных данных

В настройках визуализаторов созданного узла (глаз с шестеренкой) добавьте статистику. В статистике в настройках показателей поставьте галочку напротив настройки "Значения", не снимая остальных галочек.

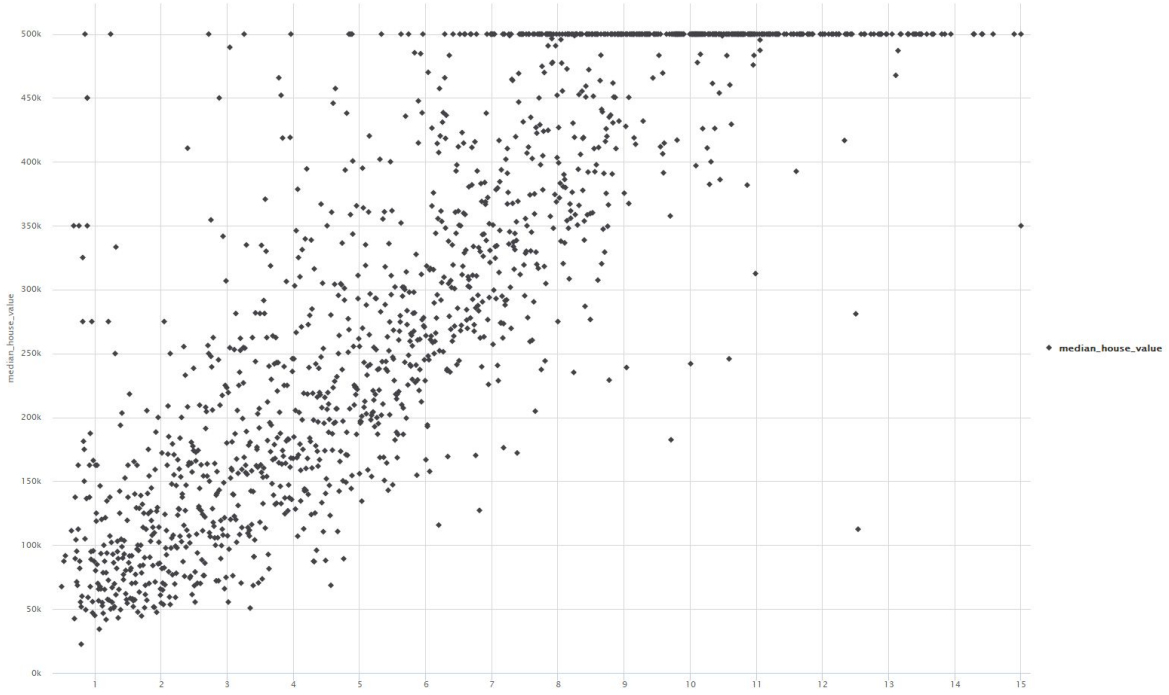


В столбце значений обнаруживается, что в поле *total_bedrooms* не хватает данных. Следовательно, вскоре Вам придётся это исправить.

В качестве второго визуализатора добавьте диаграмму. Перетащите поле *latitude* на диаграмму и создайте из него серию типа "Разброс" с полем цвета *median_house_value*. Поле *longitude* используйте как поле абсцисс.

Таким образом создана диаграмма географического распределения медианных стоимостей недвижимостей — за широту и долготу отвечают абсциссы и ординаты точек данных, а за стоимость — их цвет.

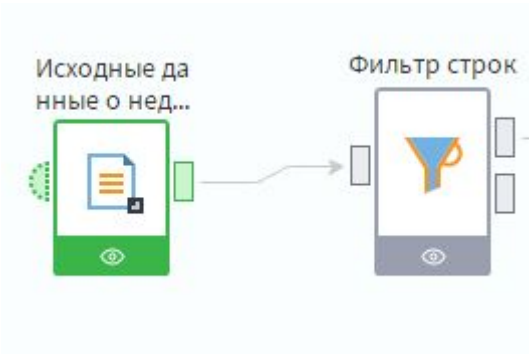
Далее создайте ещё одну диаграмму для этого узла. Перетащите на вторую диаграмму поле *median_house_value* и сделайте из него разброс-серию. В качестве абсцисс задайте поле *median_income*.



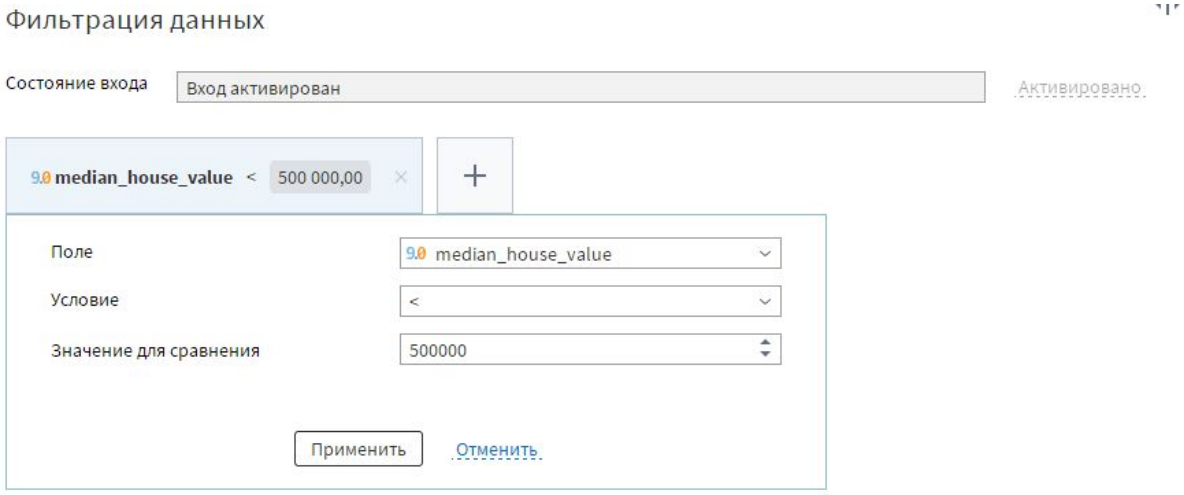
На новой диаграмме наблюдается неприятная закономерность набора данных — стоимость недвижимости была искусственно ограничена 500 тысячами долларов. Эти данные, выпадающие из общего ряда, могут усложнить обучение модели прогнозирования, поэтому от них следует избавиться.

4.1.3. Фильтр строк

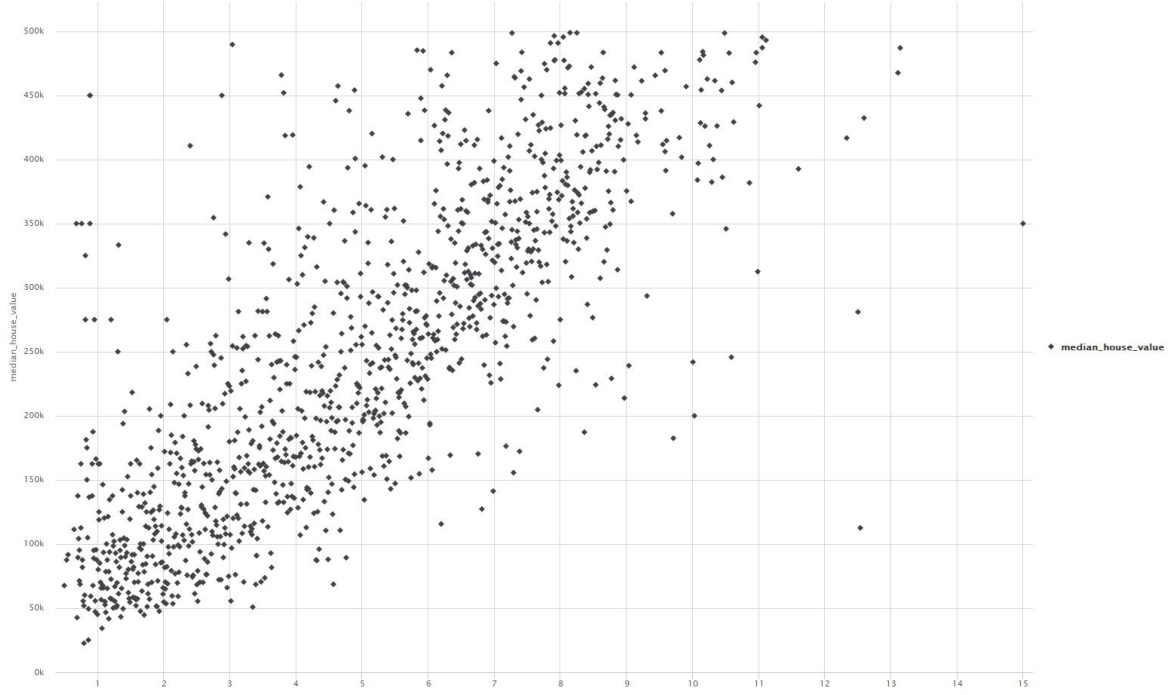
В области проектирования сценария добавьте узел "Фильтр строк". Соедините выходной порт узла с исходными данными с входным портом нового узла.



В настройках узла активируйте вход и добавьте условие "*median_house_value < 500000*".



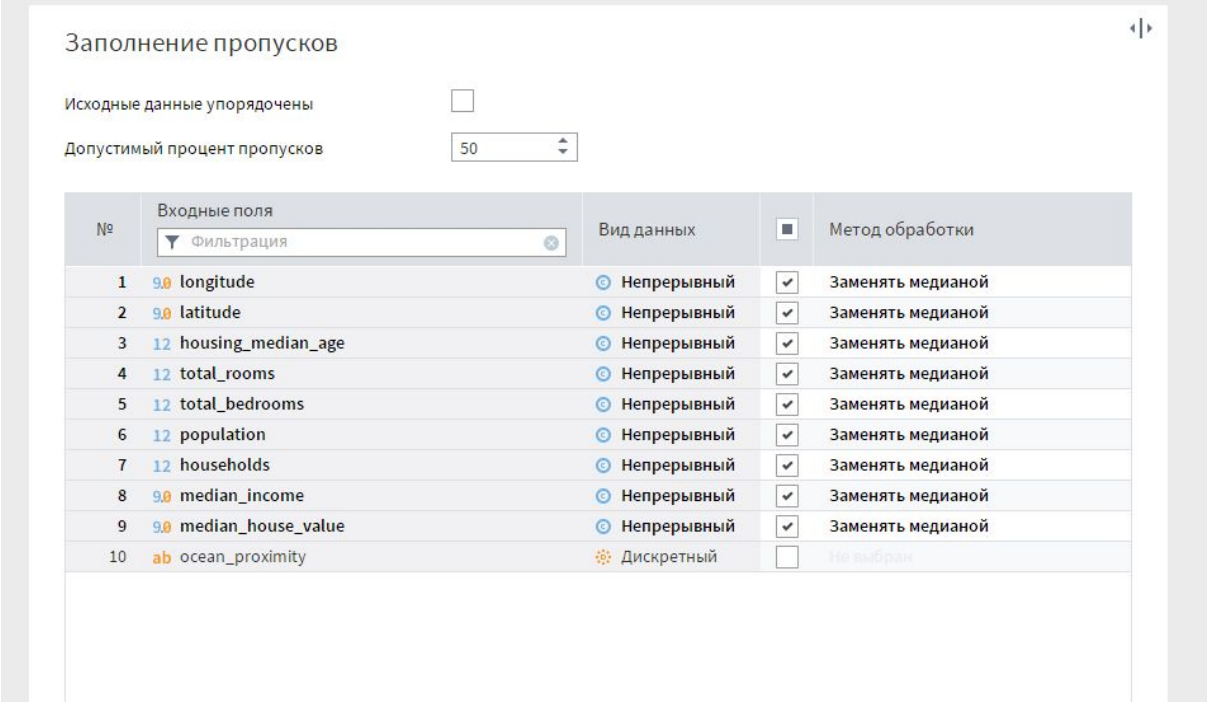
Сохраните и выполните узел. Добавьте к нему визуализатор, аналогичный второй диаграмме из узла с исходными данными. В итоге этих преобразований горизонтальная полоса, присутствовавшая вверху изначальной диаграммы, исчезнет.



4.1.4. Заполнение пропусков

Создайте узел "Заполнение пропусков" и соедините с его входным портом выходной порт "Соответствуют условию" фильтрующего узла.

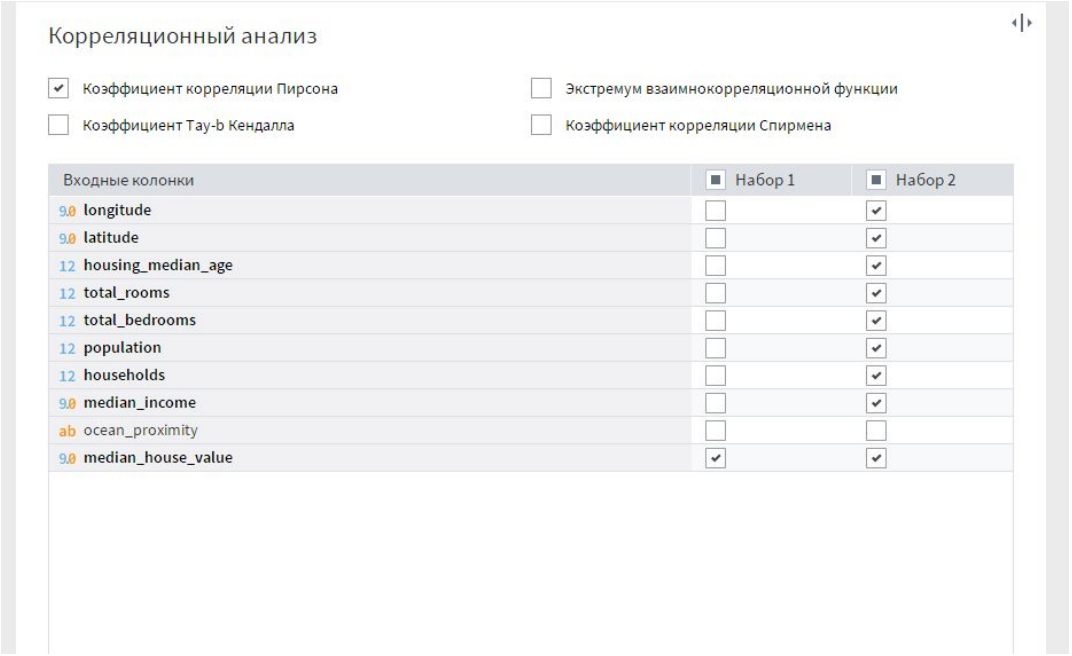
Для всех полей с непрерывным видом данных установите метод обработки "Заменять медианой".



Сохраните и выполните узел. Визуализируйте статистику и убедитесь, что в поле *total_bedrooms* пустые значения были замещены.

4.1.5. Вычисление корреляций

Создайте узел "Корреляционный анализ", свяжите его с выходом узла заполнения пропусков и настройте его следующим образом.

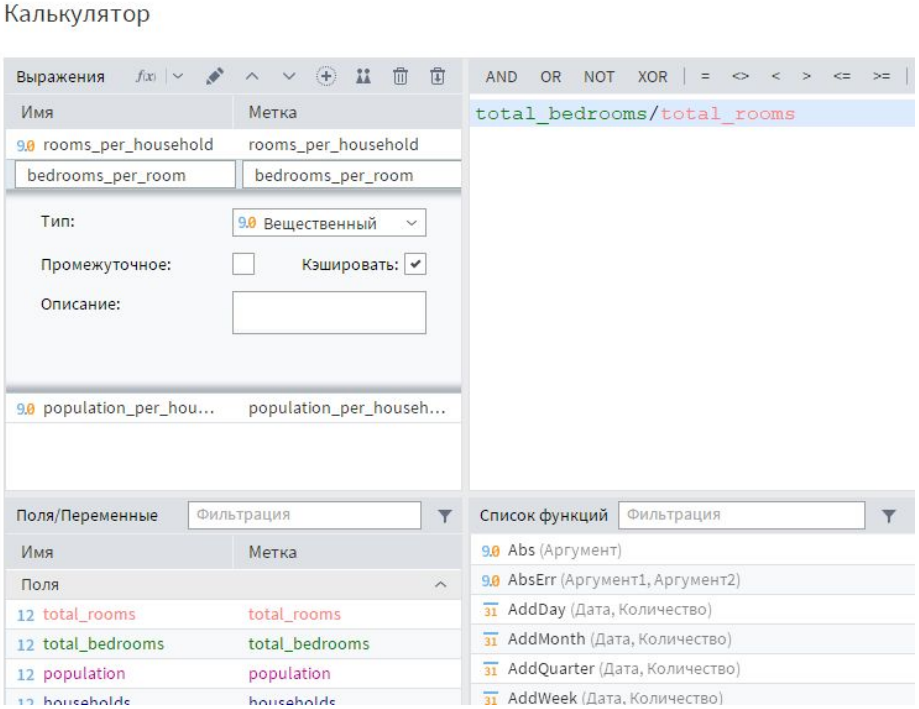


Создайте узел "Сортировка". Свяжите его вход с выходом узла корреляционного анализа. Настройте узел на сортировку столбца "Пирсона" по убыванию. Добавьте табличный визуализатор в узел. Определите, какие признаки оказывают наибольшее влияние на стоимость недвижимости.

4.1.6. Введение новых признаков

Создайте узел "Калькулятор" и свяжите его вход с выходом узла заполнения пропусков. В окне выражений в настройках узла создайте следующие выражения:

Имя / Метка	*Выражение*
rooms_per_household	`total_rooms/households`
bedrooms_per_room	`total_bedrooms/total_rooms`
population_per_household	`population/households`

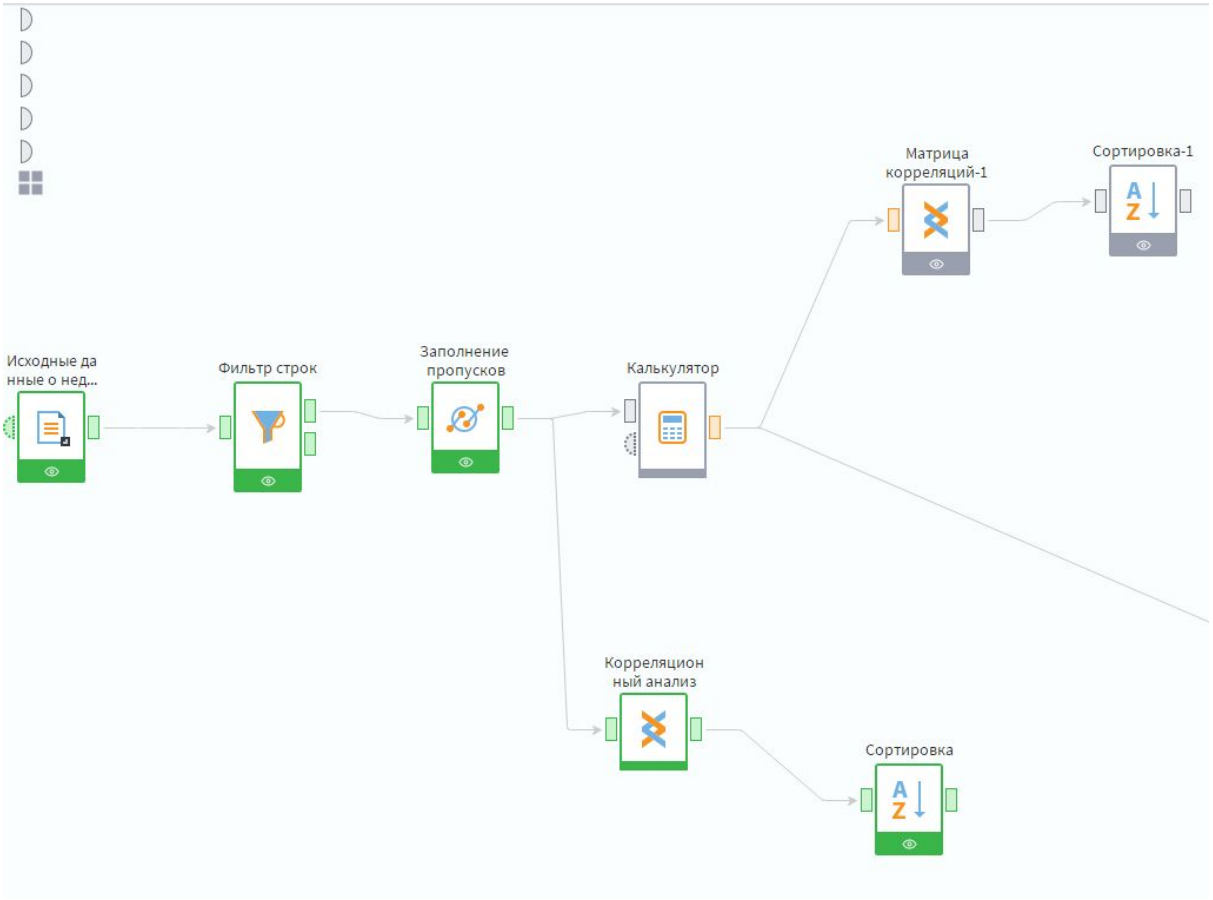


Имена полей в редактор выражений можно быстрее вставлять двойным левым кликом по полю в списке полей/переменных. Для ускорения дальнейших вычислений рекомендуется включать кэширование новых выражений.

Сохраните и выполните узел.

Аналогично пункту 4.1.5., создайте вторую отсортированную таблицу корреляций, включающую все имеющиеся непрерывные признаки. Определите, какой из новых признаков оказывает наибольшее влияние на стоимость недвижимости.

К данному моменту времени созданный сценарий должен выглядеть примерно следующим образом.

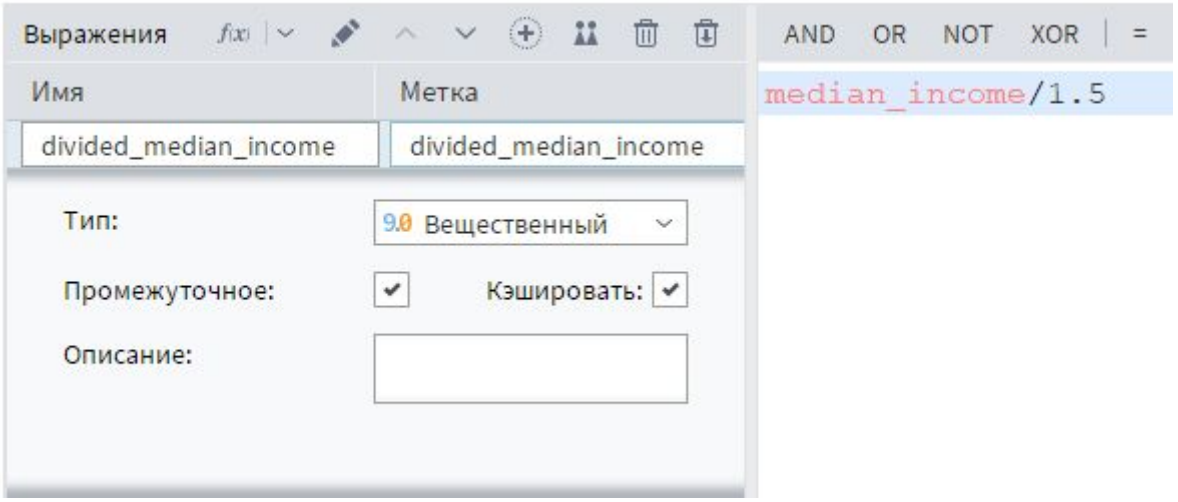


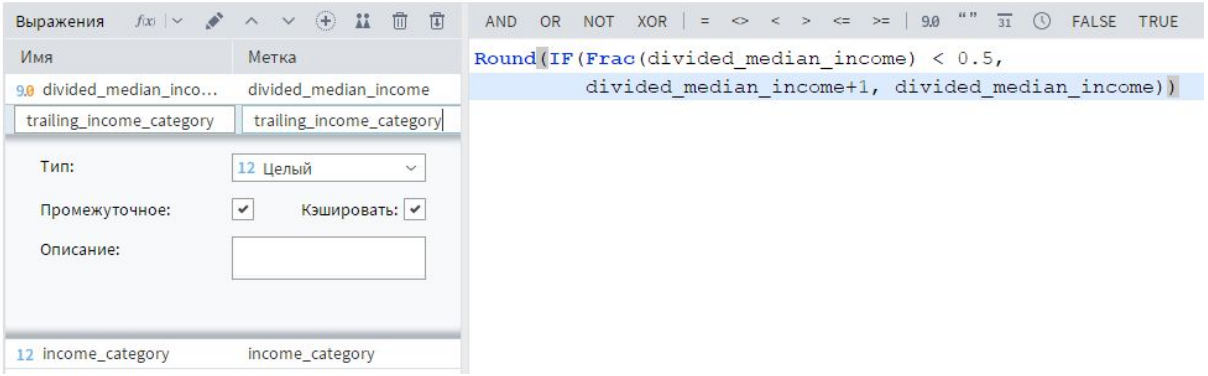
4.1.7. Стратификация данных

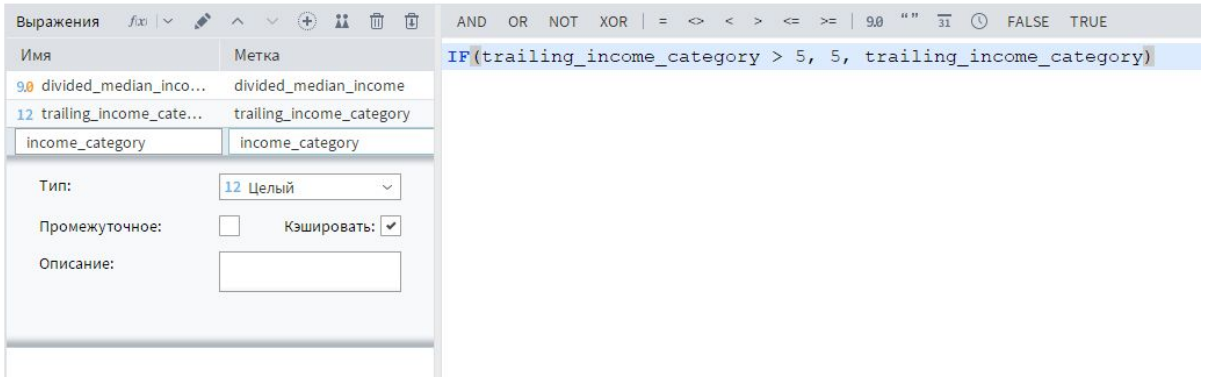
Прежде чем разбивать данные на обучающий и тестовый наборы, следует их стратифицировать с целью обеспечения наиболее

репрезентативной выборки обоих наборов. Для этого введите новый признак — *income_category* следующим образом:

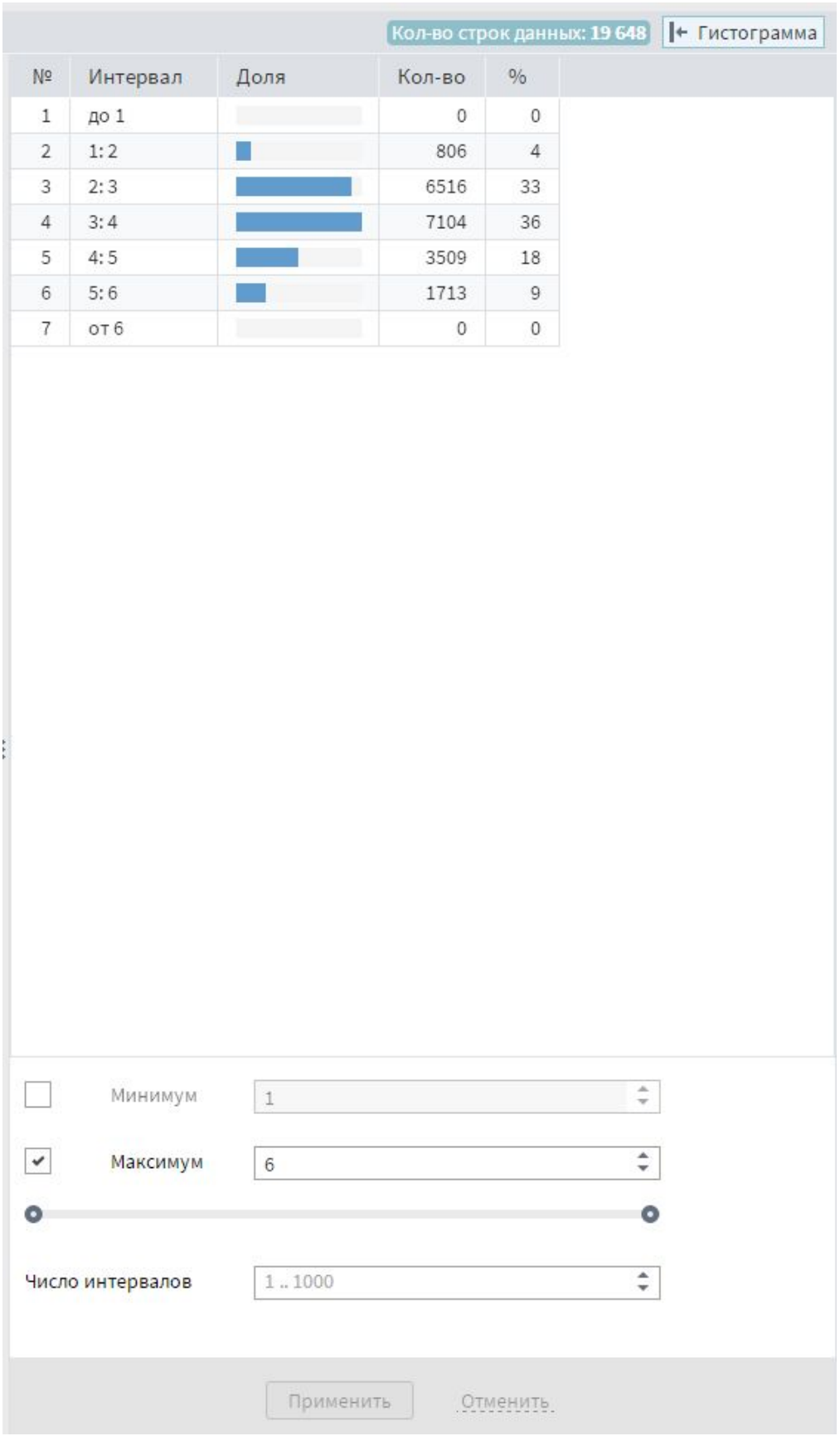
Создайте узел "Калькулятор" и свяжите его с предыдущим таким же узлом. В новом узле создайте следующие выражения:





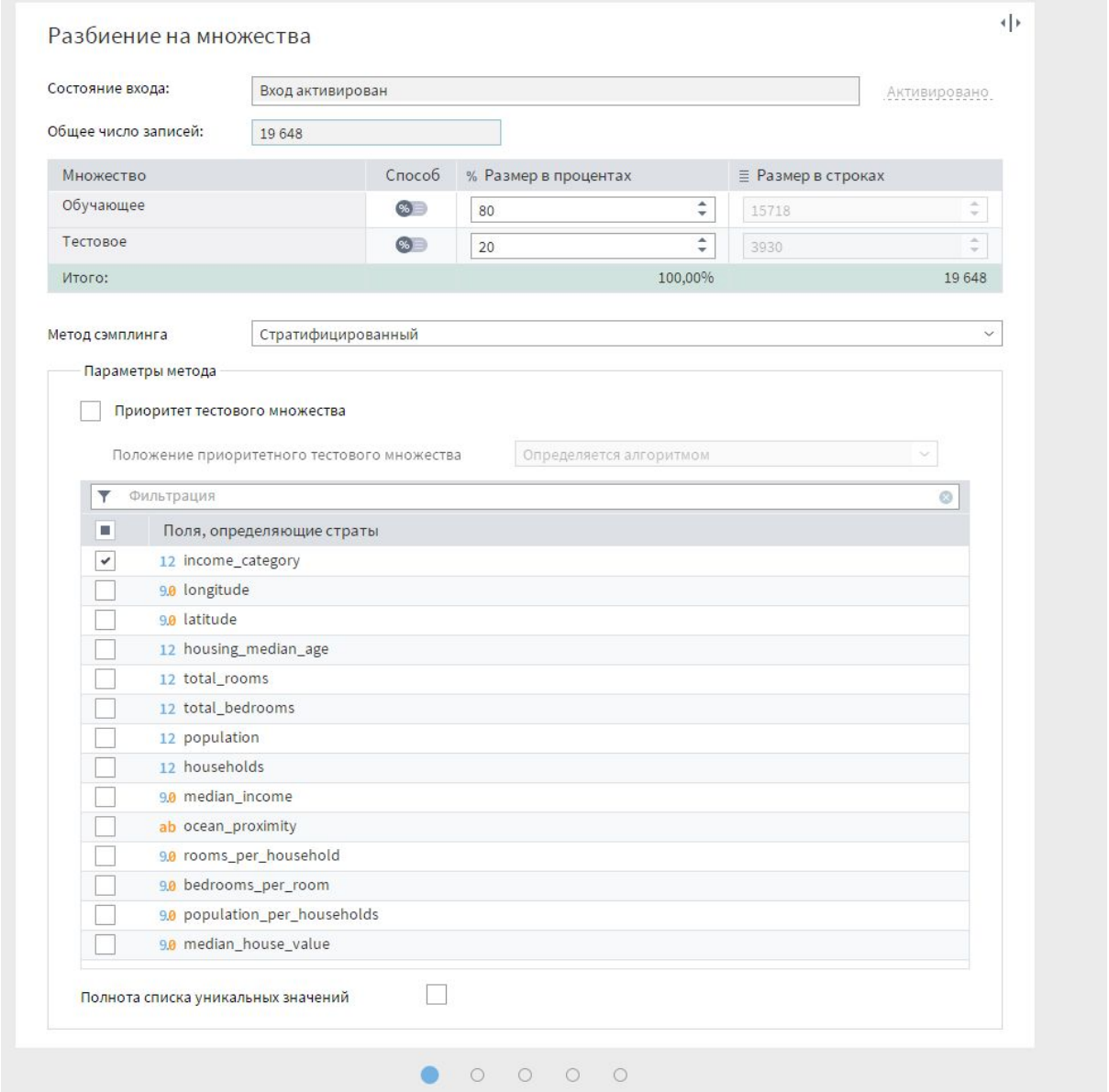


Сохраните и выполните узел. Добавьте в него визуализатор статистики. В статистике узла для поля *income_category* настройте гистограмму (в верхнем правом углу), установив настройку "максимум" в значение 6.


Кол-во строк данных: 19 648
Гистограмма
№ Интервал Доля Кол-во %
1 до 1 0 0
2 1: 2 806 4
3 2: 3 6516 33
4 3: 4 7104 36
5 4: 5 3509 18
6 5: 6 1713 9
7 от 6 0 0
Минимум
1
Максимум
6
Число интервалов
1 .. 1000
Применить
Отменить

4.1.8. Генерация обучающего и тестового наборов данных

Создайте узел "Разбиение на множества" и свяжите его со стратифицированными данными. В настройках узла активируйте вход, установите размеры обучающего и тестового наборов на 80% и 20% соответственно и выберите стратифицированный метод сэмплинга по полю *income_category*.



Сохраните и выполните узел. При возникновении ошибки, требующей обучения узла, нажмите на узел правой клавишей мыши и переобучите его.

Добавьте к узлу визуализаторы статистик обучающего и тестового выходных наборов. В них настройте гистограммы образом, аналогичным

пункту 4.1.7. и убедитесь, что выборка данных соответствует стратификации исходного набора.

4.1.9. Обучение модели линейной регрессии

Создайте узел "Линейная регрессия" и на его вход подайте **обучающий** набор данных. Признак *median_house_value* сделайте выходным, а все остальные, **за исключением *income_category***, — входными. (Признак *income_category* остаётся неиспользуемым, так как он был необходим исключительно для осуществления стратифицированной выборки. Использование его в модели прогнозирования может исказить конечный результат.)

Для всех непрерывных входных признаков установите нормализатор "Стандартизация", а для *ocean_proximity* — "Индикатор". Настройку приоритета автоматической регрессии выберите в зависимости от производительности Вашего компьютера (чем больше точность — тем вычисления более требовательны ко временным и системным ресурсам). Все дальнейшие параметры оставьте по умолчанию.

Сохраните узел и обучите его.

4.1.10. Анализ обученной модели (индивидуальное задание)

Построенная модель, очевидно, далека от идеального средства прогнозирования. Однако, Вы можете исследовать ошибки регрессии и выдвинуть гипотезы о том, что именно могло пойти не так в процессе её проектирования. Для этого Вы можете самостоятельно исследовать результаты обучения (например, рассчитать различные погрешности прогнозируемой стоимости от фактической) и визуализировать/проанализировать их удобным для вас способом.

4.2. Построение скоринговых карт. Организация практического применения Loginom

Для правильного выполнения следующих пунктов НЕОБХОДИМО ознакомиться с теоретической частью данных методических указаний (пункты 2.6. И 2.7.)

4.2.1. Кластеризация и визуализация данных.

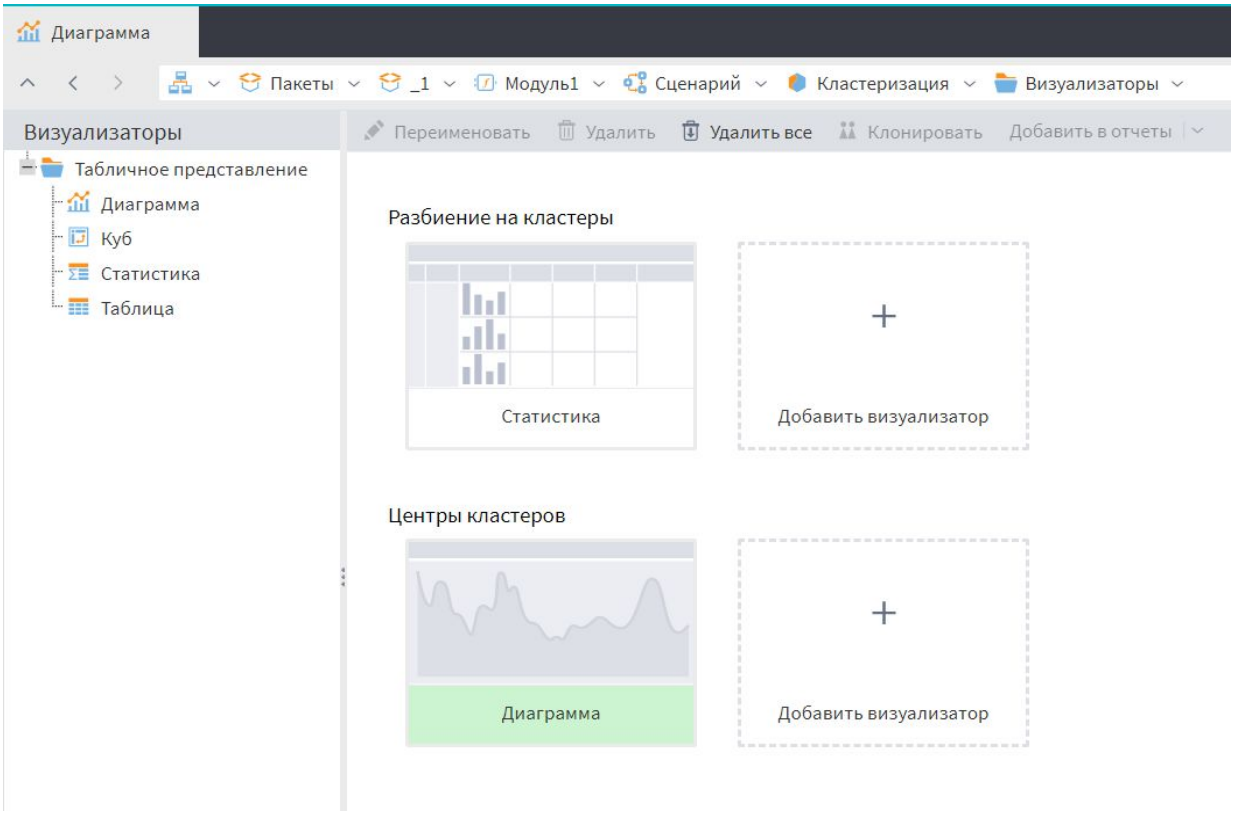
Для дальнейшей работы произведите импорт текстового файла Credit.txt, как делали это в предыдущих практических заданиях. Все пункты выполнения импорта можно оставить по умолчанию, название файла выбрать самостоятельно. После чего перейдите в панель компонентов и используйте Кластеризацию. Сконфигурируйте входной порт и перейдите в настройки параметров кластеризации.

В параметре настройки входных столбцов установите значение поля “Назначение” Используемым у следующих признаков:

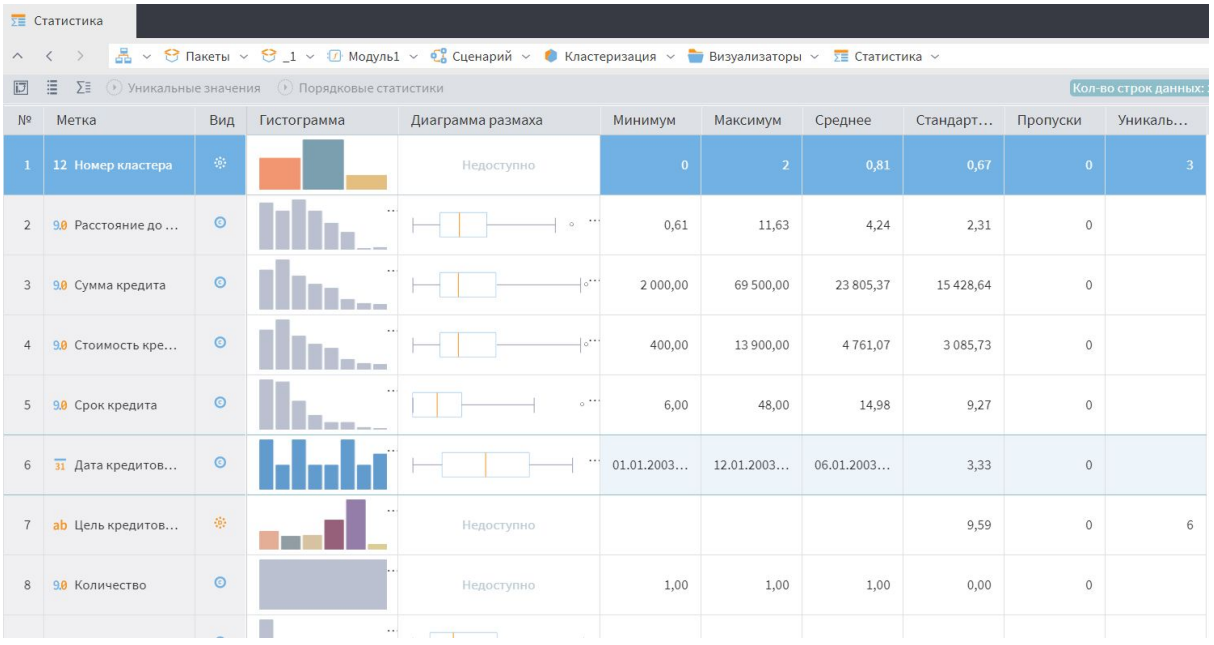
- Сумма кредита;
- Стоимость кредита;
- Срок кредита;
- Возраст;
- Среднемесячный доход;

В настройках нормализации постарайтесь самостоятельно настроить нормализации (здесь пригодится теоретический материал) таким образом, чтобы кластеры были распределены по возрасту наиболее логичным образом. Необходимо получить предположительно 3 кластера.

После успешного выполнения предыдущего задания перейдите в визуализатор кластеризации. Добавьте статистику и диаграмму как это показано ниже.

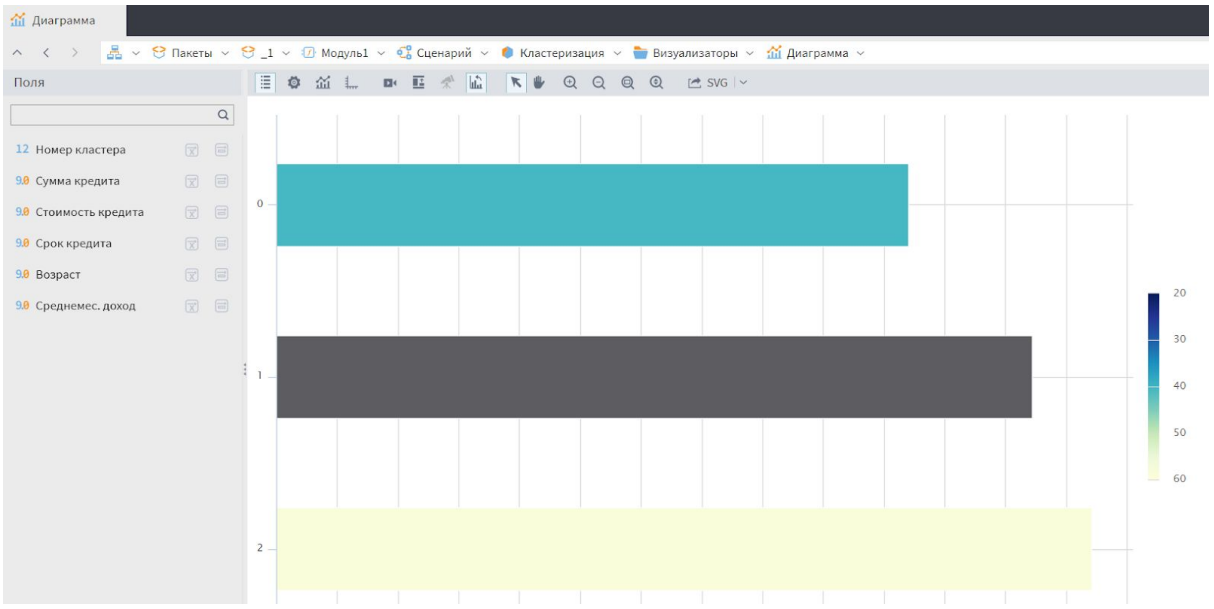


Перейдите в статистику и изучите составленные диаграммы:



№	Метка	Вид	Гистограмма	Диаграмма размаха	Минимум	Максимум	Среднее	Стандарт…	Пропуски	Уникаль…
1	12 Номер кластера			Недоступно	0	2	0,81	0,67	0	3
2	9.0 Расстояние до …				0,61	11,63	4,24	2,31	0	
3	9.0 Сумма кредита				2 000,00	69 500,00	23 805,37	15 428,64	0	
4	9.0 Стоимость кре…				400,00	13 900,00	4 761,07	3 085,73	0	
5	9.0 Срок кредита				6,00	48,00	14,98	9,27	0	
6	31 Дата кредитов…				01.01.2003…	12.01.2003…	06.01.2003…	3,33	0	
7	ab Цель кредитов…			Недоступно				9,59	0	6
8	9.0 Количество			Недоступно	1,00	1,00	1,00	0,00	0	

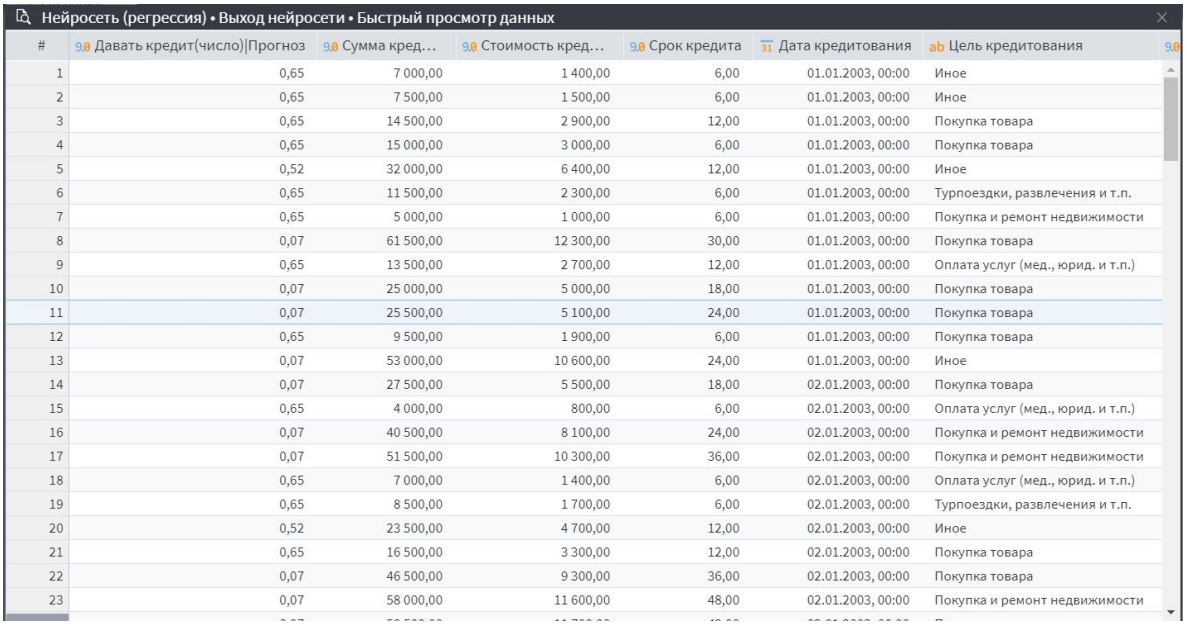
После чего откройте диаграмму центров кластеров и настройте её таким образом, чтобы можно было наглядно увидеть зависимость суммы кредита от возраста заёмщиков, это должно выглядеть приблизительно так:



4.2.2. Прогнозирование вероятности получения кредита при помощи нейросети.

К ранее используемому источнику данных примените компонент “Нейросеть (регрессия). Задайте назначение входного параметра для тех же меток, что использовались в кластеризации.

Выходным- Давать кредит (число). Нормализацией выходного параметра выберите Масштабирование [min; max]. Остальные оставьте по умолчанию. Разбиение на множества установите 95% к 5%. Количество скрытых слоёв -1; Количество нейронов в 1 скрытом слое- 5. Название каждого узла всегда остается на ваше усмотрение. Правой кнопкой мыши нажмите на узел и произведите его переобучение перед выполнением во избежание последующих ошибок. Нажмите на выход нейросети, выберите быстрый просмотр, изучите полученный прогноз.



Нейросеть (регрессия) • Выход нейросети • Быстрый просмотр данных

#	Давать кредит(число)|Прогноз	Сумма кред…	Стоимость кред…	Срок кредита	Дата кредитования	Цель кредитования
1	0,65	7 000,00	1 400,00	6,00	01.01.2003, 00:00	Иное
2	0,65	7 500,00	1 500,00	6,00	01.01.2003, 00:00	Иное
3	0,65	14 500,00	2 900,00	12,00	01.01.2003, 00:00	Покупка товара
4	0,65	15 000,00	3 000,00	6,00	01.01.2003, 00:00	Покупка товара
5	0,52	32 000,00	6 400,00	12,00	01.01.2003, 00:00	Иное
6	0,65	11 500,00	2 300,00	6,00	01.01.2003, 00:00	Турпоездки, развлечения и т.п.
7	0,65	5 000,00	1 000,00	6,00	01.01.2003, 00:00	Покупка и ремонт недвижимости
8	0,07	61 500,00	12 300,00	30,00	01.01.2003, 00:00	Покупка товара
9	0,65	13 500,00	2 700,00	12,00	01.01.2003, 00:00	Оплата услуг (мед., юрид. и т.п.)
10	0,07	25 000,00	5 000,00	18,00	01.01.2003, 00:00	Покупка товара
11	0,07	25 500,00	5 100,00	24,00	01.01.2003, 00:00	Покупка товара
12	0,65	9 500,00	1 900,00	6,00	01.01.2003, 00:00	Покупка товара
13	0,07	53 000,00	10 600,00	24,00	01.01.2003, 00:00	Иное
14	0,07	27 500,00	5 500,00	18,00	02.01.2003, 00:00	Покупка товара
15	0,65	4 000,00	800,00	6,00	02.01.2003, 00:00	Оплата услуг (мед., юрид. и т.п.)
16	0,07	40 500,00	8 100,00	24,00	02.01.2003, 00:00	Покупка и ремонт недвижимости
17	0,07	51 500,00	10 300,00	36,00	02.01.2003, 00:00	Покупка и ремонт недвижимости
18	0,65	7 000,00	1 400,00	6,00	02.01.2003, 00:00	Оплата услуг (мед., юрид. и т.п.)
19	0,65	8 500,00	1 700,00	6,00	02.01.2003, 00:00	Турпоездки, развлечения и т.п.
20	0,52	23 500,00	4 700,00	12,00	02.01.2003, 00:00	Иное
21	0,65	16 500,00	3 300,00	12,00	02.01.2003, 00:00	Покупка товара
22	0,07	46 500,00	9 300,00	36,00	02.01.2003, 00:00	Покупка товара
23	0,07	58 000,00	11 600,00	48,00	02.01.2003, 00:00	Покупка и ремонт недвижимости

Выбрав сводку можете ознакомиться с данными о тестовом множестве и полученных ошибках:

Нейросеть (регрессия) • Сводка • Быстрый просмотр данных

№	Имя	Метка	Значение
1	TotalSamples	Всего примеров	149
2	TotalSelectedSamples	Всего отобранных примеров	149
3	TrainSamples	Примеров в обучающем множестве	142
4	TrainRMSError	Среднеквадратическая ошибка на обучающем множестве	0,45
5	TrainAvgError	Средняя абсолютная ошибка на обучающем множестве	0,40
6	TrainAvgRelError	Средняя относительная ошибка на обучающем множестве	0,49
7	TestSamples	Примеров в тестовом множестве	7
8	TestRMSError	Среднеквадратическая ошибка на тестовом множестве	0,50
9	TestAvgError	Средняя абсолютная ошибка на тестовом множестве	0,44
10	TestAvgRelError	Средняя относительная ошибка на тестовом множестве	0,47
11	GTest_COL1	Сумма кредита|G-тест	224,65
12	GTestDF_COL1	Сумма кредита|Число степеней свободы G-теста	225
13	GTestPValue_COL1	Сумма кредита|P-значение G-теста	0,49
14	MutualInf_COL1	Сумма кредита|Взаимная информация	1,44
15	GTest_COL2	Стоимость кредита|G-тест	224,65
16	GTestDF_COL2	Стоимость кредита|Число степеней свободы G-теста	225
17	GTestPValue_COL2	Стоимость кредита|P-значение G-теста	0,49
18	MutualInf_COL2	Стоимость кредита|Взаимная информация	1,44
19	GTest_COL3	Срок кредита|G-тест	258,86
20	GTestDF_COL3	Срок кредита|Число степеней свободы G-теста	225
21	GTestPValue_COL3	Срок кредита|P-значение G-теста	0,06
22	MutualInf_COL3	Срок кредита|Взаимная информация	1,66
23	GTest_COL7	Возраст|G-тест	193,04

Если для понимания полученных данных этого недостаточно, перейдите к визуализатору нейросети и получите статистику, в которой

можно наглядно увидеть соотношение количества одобренных и кредитов, по которым получен отказ.

Это лишь малая часть организации скоринговой карты в Loginom, со всем объемом возможностей ознакомиться в данной лабораторной работе не позволяет ограничение времени и отсутствие допустимых источников данных.

5. Контрольные вопросы

1. Декомпозиция проекта в Loginom. Что такое пакет? Модуль? Сценарий? Узел?
2. Способы нормализации непрерывных данных.
3. Типы дискретных и непрерывных данных.
4. Какие визуализаторы определены для всех базовых компонентов Loginom?
5. С какой целью проводится очистка данных от пропущенных и критических значений?
6. С какой целью проводится стратификация исходных данных?
7. Устный вопрос по некоторому промежуточному визуализатору.
8. Устный вопрос по анализу построенной модели прогнозирования стоимости недвижимости.
9. Устный вопрос по скоринговым картам.

6. Список используемой литературы

1. Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. Пер. с англ. - СпБ.: ООО "Альфа-книга': 2018. - 688 с.: ил. - Парал. тит. англ.
   ISBN 978-5-9500296-2-2 (рус.)
2. https://help.loginom.ru/userguide/ (Дата обращения: 22 ноября 2019 г.)