Лабораторная работа

# Ознакомление с инструментарием Orange Data Mining для анализа данных

Выполнила: Короткова Инга Сергеевна

2020 год

**Цель**:

Получить навыки работы с инструментарием Orange Data Mining для задач анализа данных.

**Задачи**:

− Установить Orange Data Mining.

− Загрузить в рабочую зону предоставленный набор данных.

− Назначить целевую переменную.

− Применить различные методы визуализации данных.

− Разделить выборку на обучающую и тестовую.

− Построить дерево принятия решений и оценить его эффективность с помощью различных метрик.

− Визуализировать полученное дерево.

− Подготовить отчет по результатам работы, включающий титульный лист, задание, описание используемых данных, иллюстрации построенных схем блоков и результаты работы блоков для каждого пункта алгоритма выполнения, заключение по работе и выводы.

Цель анализа данных этого набора состоит в предсказание, будет ли клиент подписывать срочный депозит, на основе профиля клиента, который содержит такие атрибуты, как возраст, тип работы, военное положение, образование, информация о предыдущих кредитах и другие

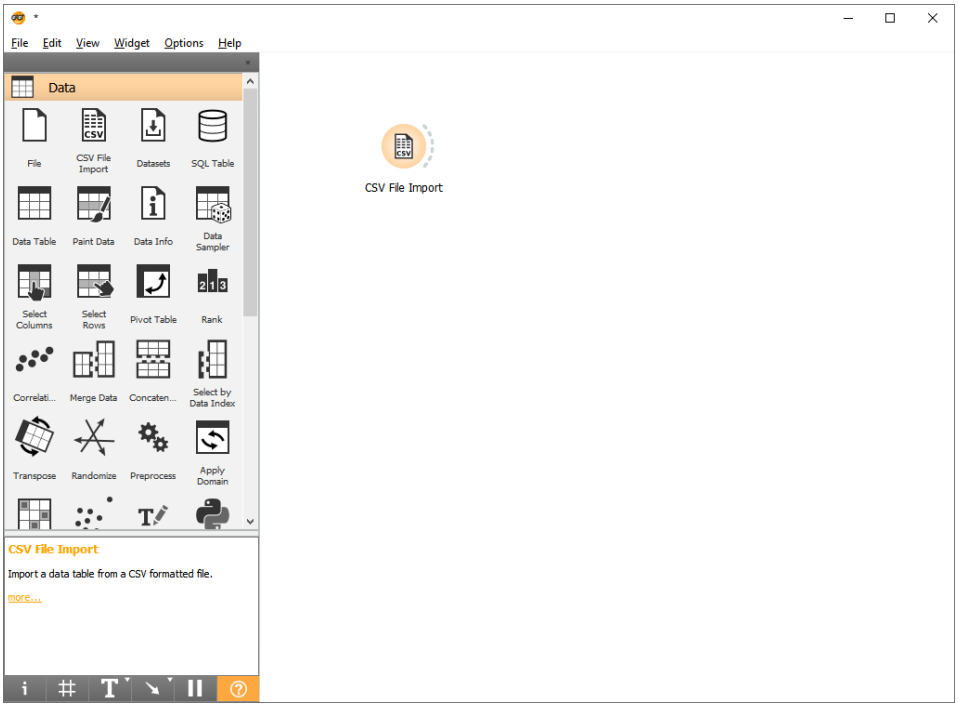
Установим пакет для Data Mining - Orange:

pip install orange3

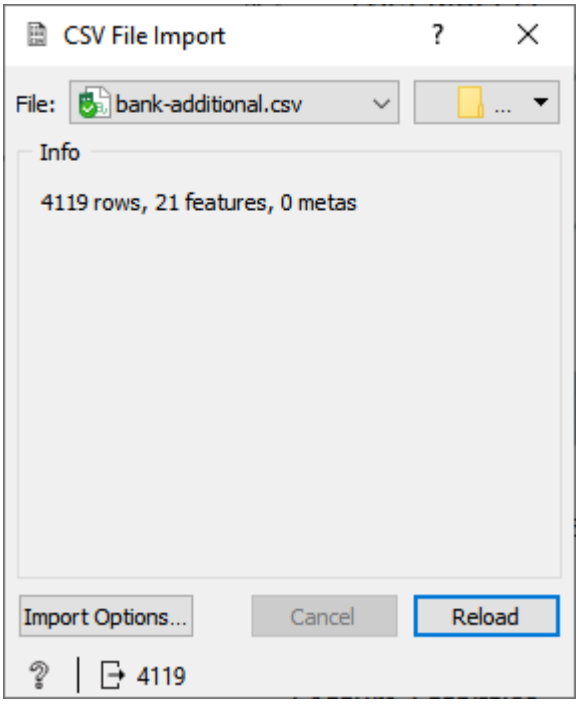
Запустим Orange в интерпретаторе:

python -m Orange.canvas

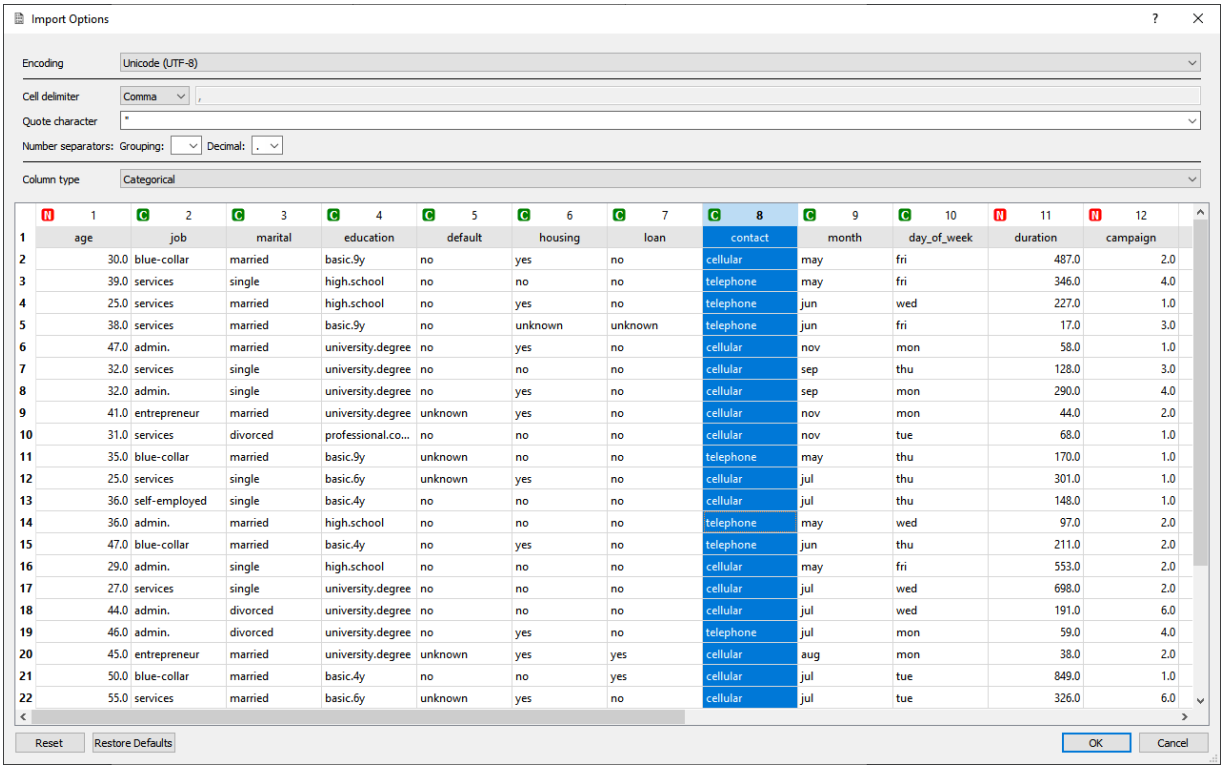
Добавим элемент для импорта файла с данными.



Загрузим csv с данными.



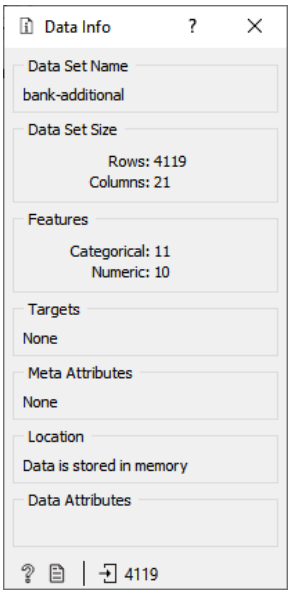
Определим для каждой колонки подходящий тип данных:




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | age | job | marital | education | default | housing | loan | contact | month | day_of_week | duration | campaign |
| 2 | 30.0 | blue-collar | married | basic.9y | no | yes | no | cellular | may | fri | 487.0 | 2.0 |
| 3 | 39.0 | services | single | high.school | no | no | no | telephone | may | fri | 346.0 | 4.0 |
| 4 | 25.0 | services | married | high.school | no | yes | no | telephone | jun | wed | 227.0 | 1.0 |
| 5 | 38.0 | services | married | basic.9y | no | unknown | unknown | telephone | jun | fri | 17.0 | 3.0 |
| 6 | 47.0 | admin. | married | university.degree | no | yes | no | cellular | nov | mon | 58.0 | 1.0 |
| 7 | 32.0 | services | single | university.degree | no | no | no | cellular | sep | thu | 128.0 | 3.0 |
| 8 | 32.0 | admin. | single | university.degree | no | yes | no | cellular | sep | mon | 290.0 | 4.0 |
| 9 | 41.0 | entrepreneur | married | university.degree | unknown | yes | no | cellular | nov | mon | 44.0 | 2.0 |
| 10 | 31.0 | services | divorced | professional.co... | no | no | no | cellular | nov | tue | 68.0 | 1.0 |
| 11 | 35.0 | blue-collar | married | basic.9y | unknown | no | no | telephone | may | thu | 170.0 | 1.0 |
| 12 | 25.0 | services | single | basic.6y | unknown | yes | no | cellular | jul | thu | 301.0 | 1.0 |
| 13 | 36.0 | self-employed | single | basic.4y | no | no | no | cellular | jul | thu | 148.0 | 1.0 |
| 14 | 36.0 | admin. | married | high.school | no | no | no | telephone | may | wed | 97.0 | 2.0 |
| 15 | 47.0 | blue-collar | married | basic.4y | no | yes | no | telephone | jun | thu | 211.0 | 2.0 |
| 16 | 29.0 | admin. | single | high.school | no | no | no | cellular | may | fri | 553.0 | 2.0 |
| 17 | 27.0 | services | single | university.degree | no | no | no | cellular | jul | wed | 698.0 | 2.0 |
| 18 | 44.0 | admin. | divorced | university.degree | no | no | no | cellular | jul | wed | 191.0 | 6.0 |
| 19 | 46.0 | admin. | divorced | university.degree | no | yes | no | telephone | jul | mon | 59.0 | 4.0 |
| 20 | 45.0 | entrepreneur | married | university.degree | unknown | yes | yes | cellular | aug | mon | 38.0 | 2.0 |
| 21 | 50.0 | blue-collar | married | basic.4y | no | no | yes | cellular | jul | tue | 849.0 | 1.0 |
| 22 | 55.0 | services | married | basic.6y | unknown | yes | no | cellular | jul | tue | 326.0 | 6.0 |

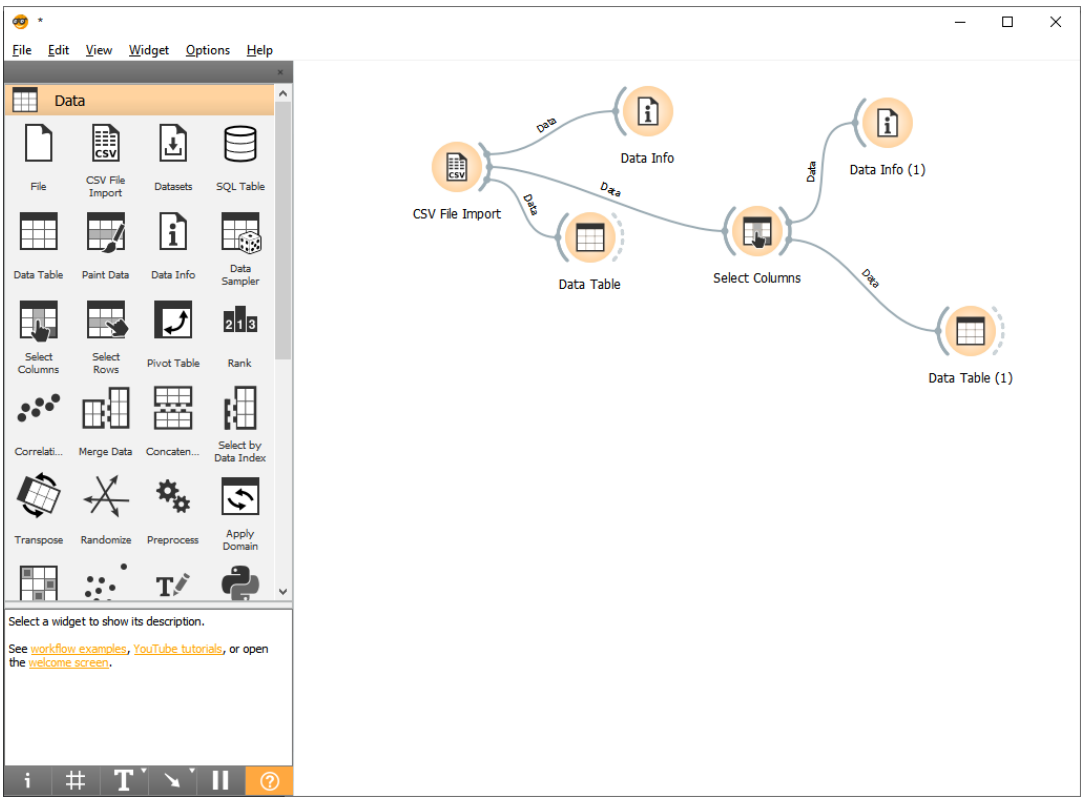
Данные представляют собой различные категориальный и численные значения.

Посмотрим общее описание датасета при помощи блока Data Info.

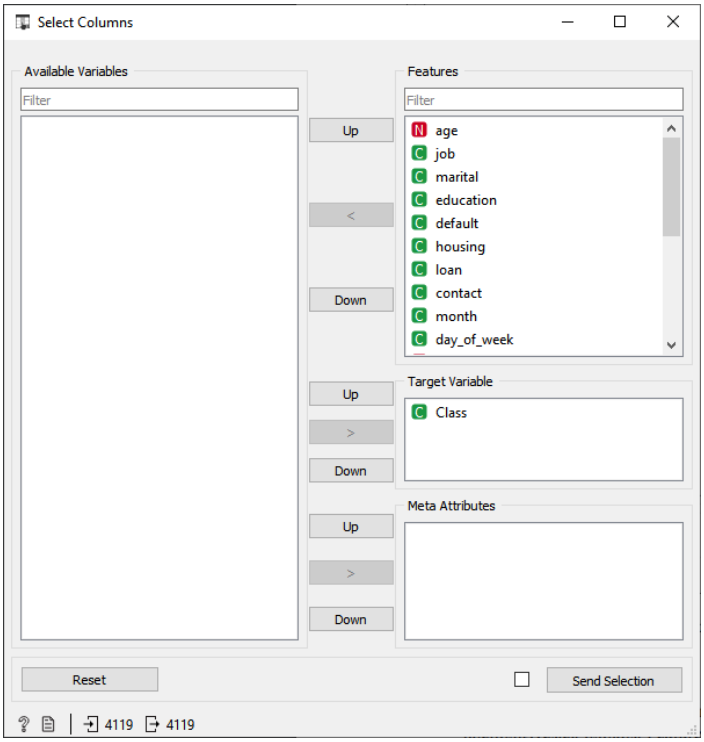



Всего 21 колонка из 4119 наблюдений.

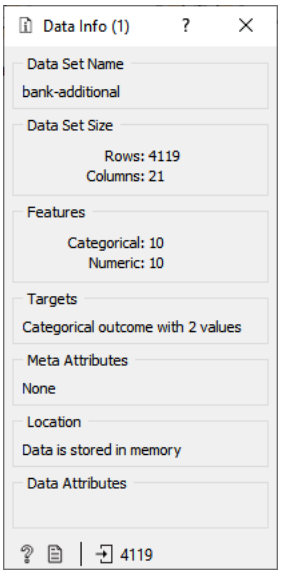
Добавим блок Select Columns, для того, чтобы отметить целевую переменную.



Отметим класс в качестве целевой переменной.

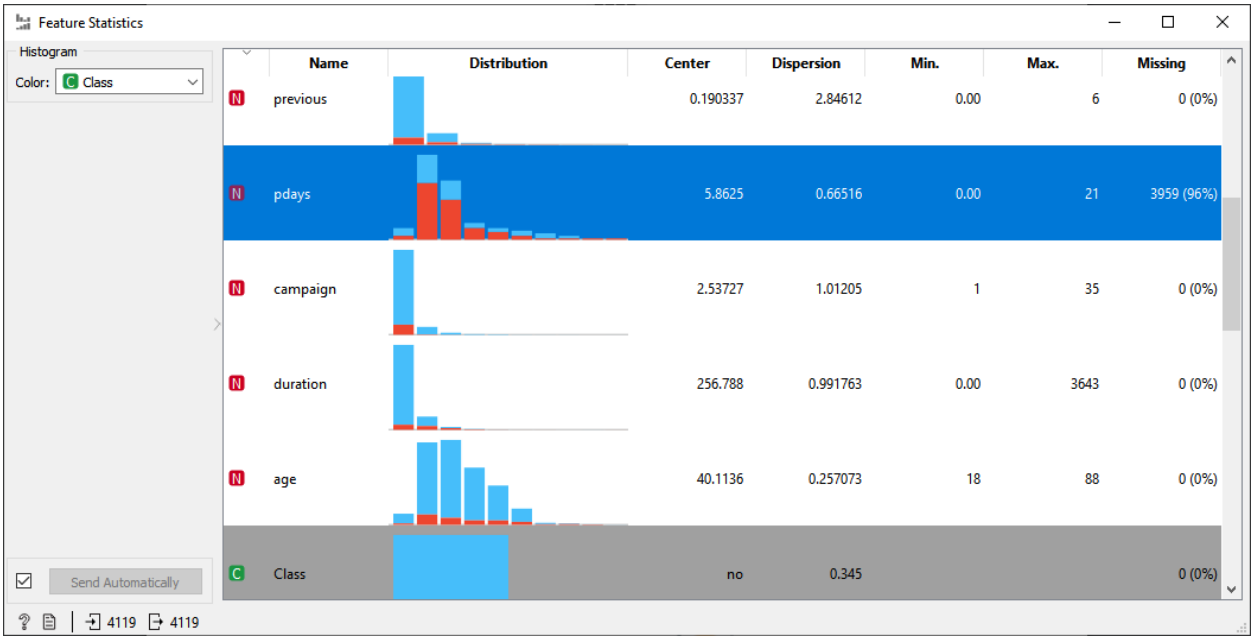



Теперь целевая переменная отображается корректно.

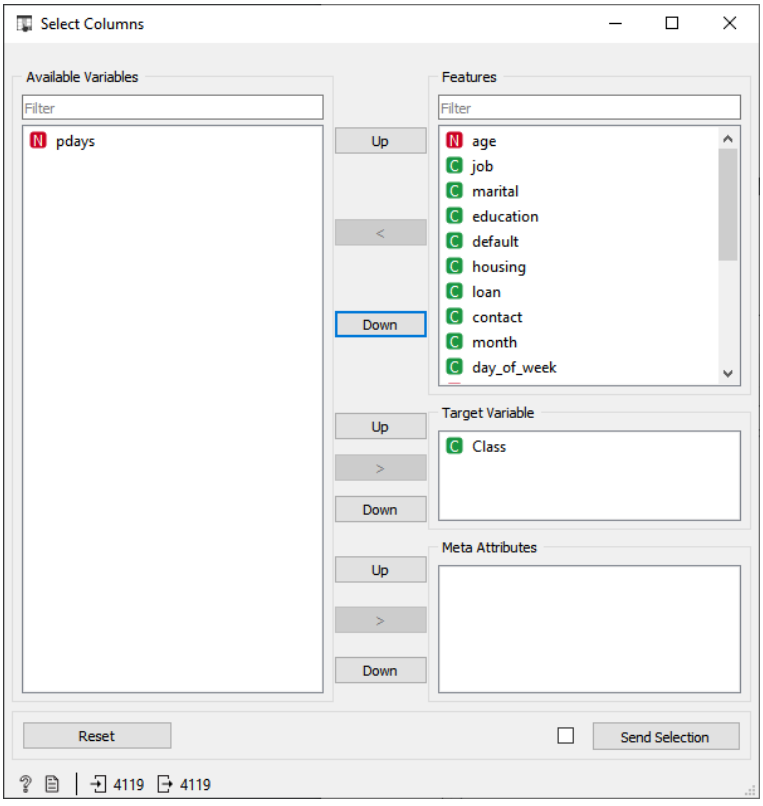


Воспользуемся блоком Feature Statistics для того, чтобы обзорно посмотреть на данные.

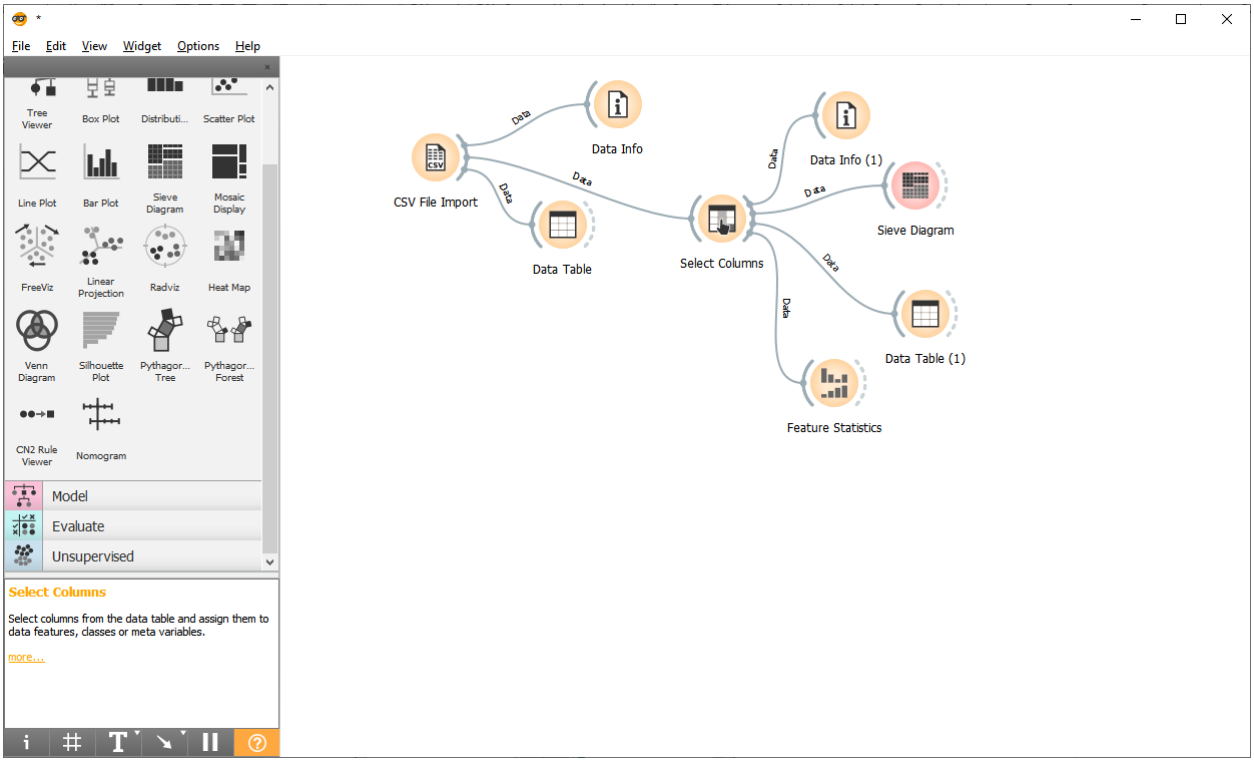
Колонка pdays содержит 96% пропусков.




Удалим его из датасета.



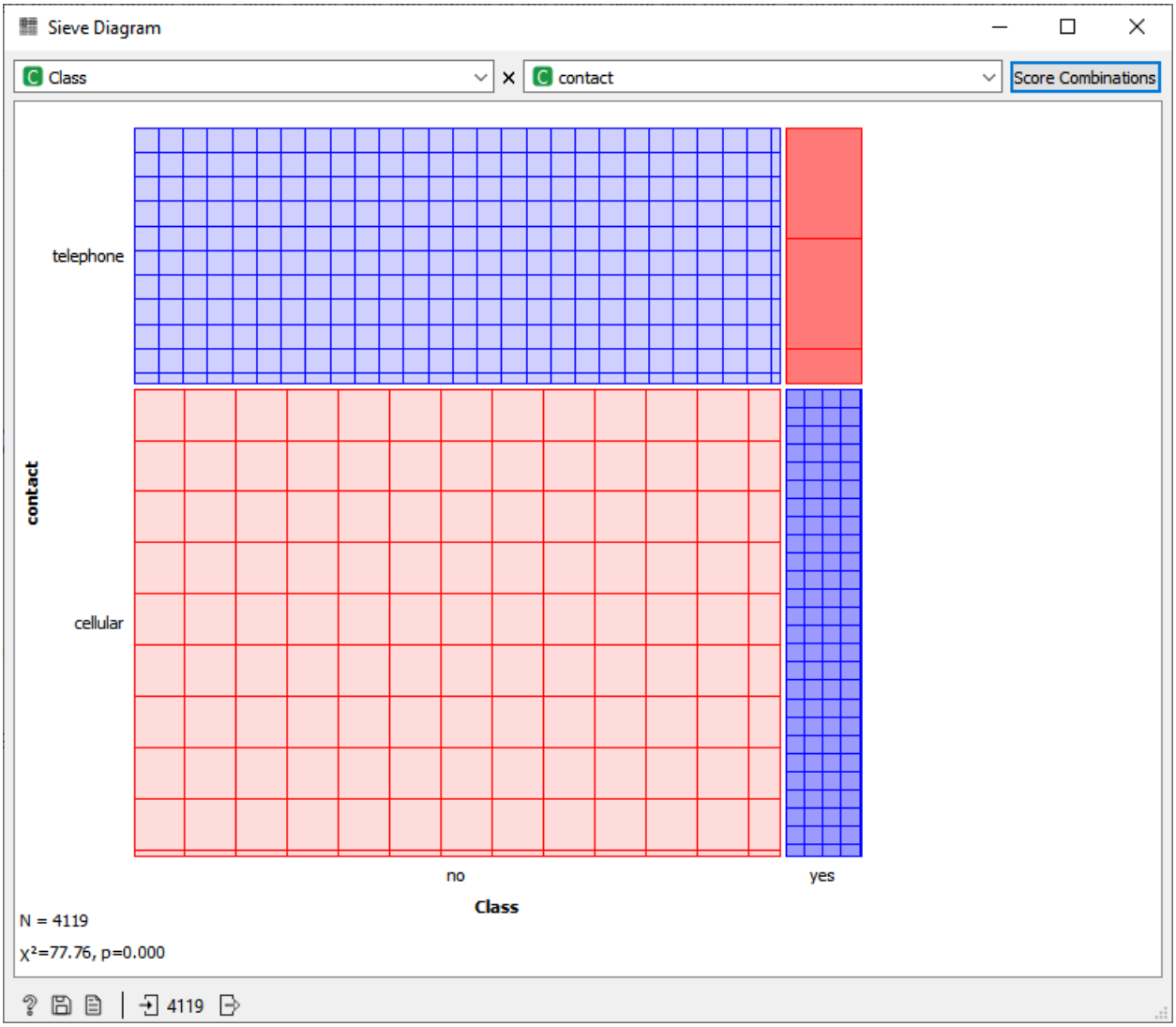
Воспользуемся Sieve Diagram для визуального визуализации данных.



Если вывести зависимость, подписан ли срочный депозит (да/нет) от типа связи с клиентом, то можно выявить:

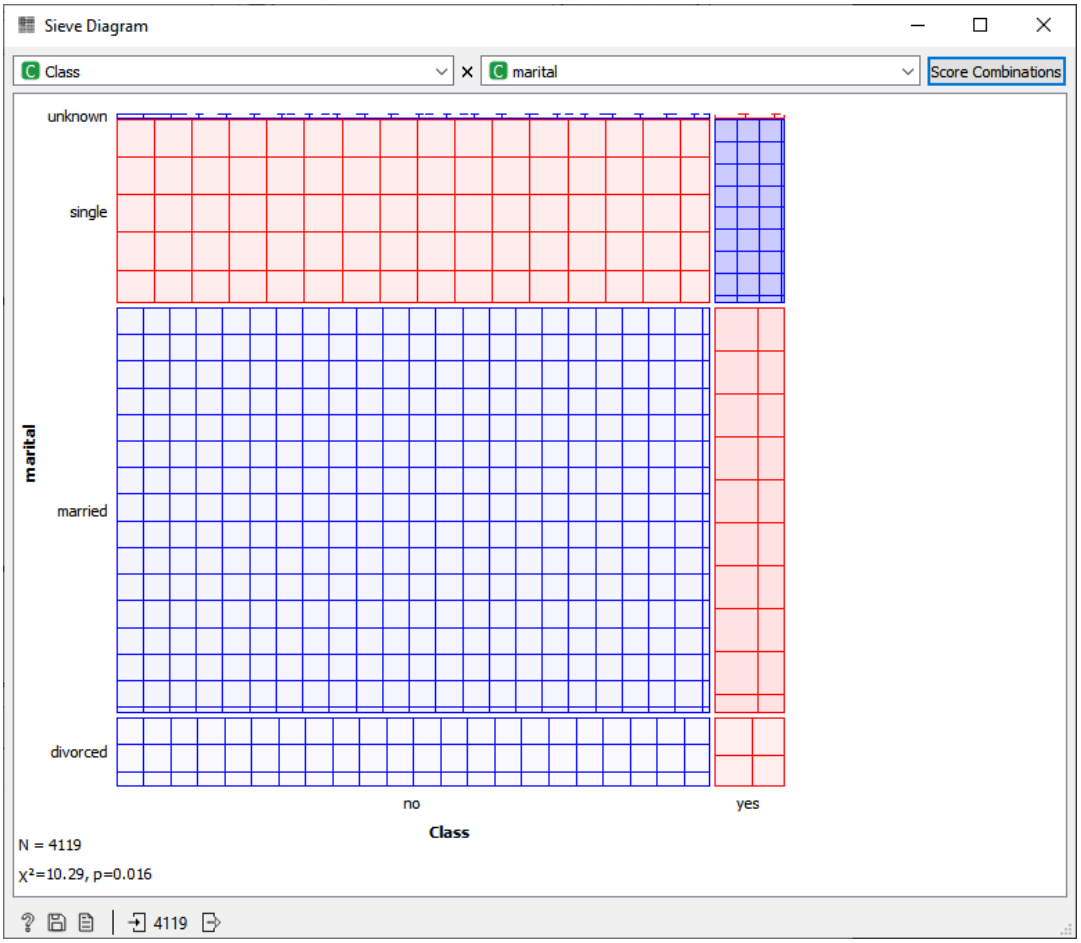
- если связь с клиентом производится по мобильному телефону, то клиент чаще согласен взять заем, нежели чем по стационарному телефону.

- чаще общение с клиентом происходит по мобильному телефону.



Если вывести зависимость, подписан ли срочный депозит (да/нет) от их семейного статуса, то видно:

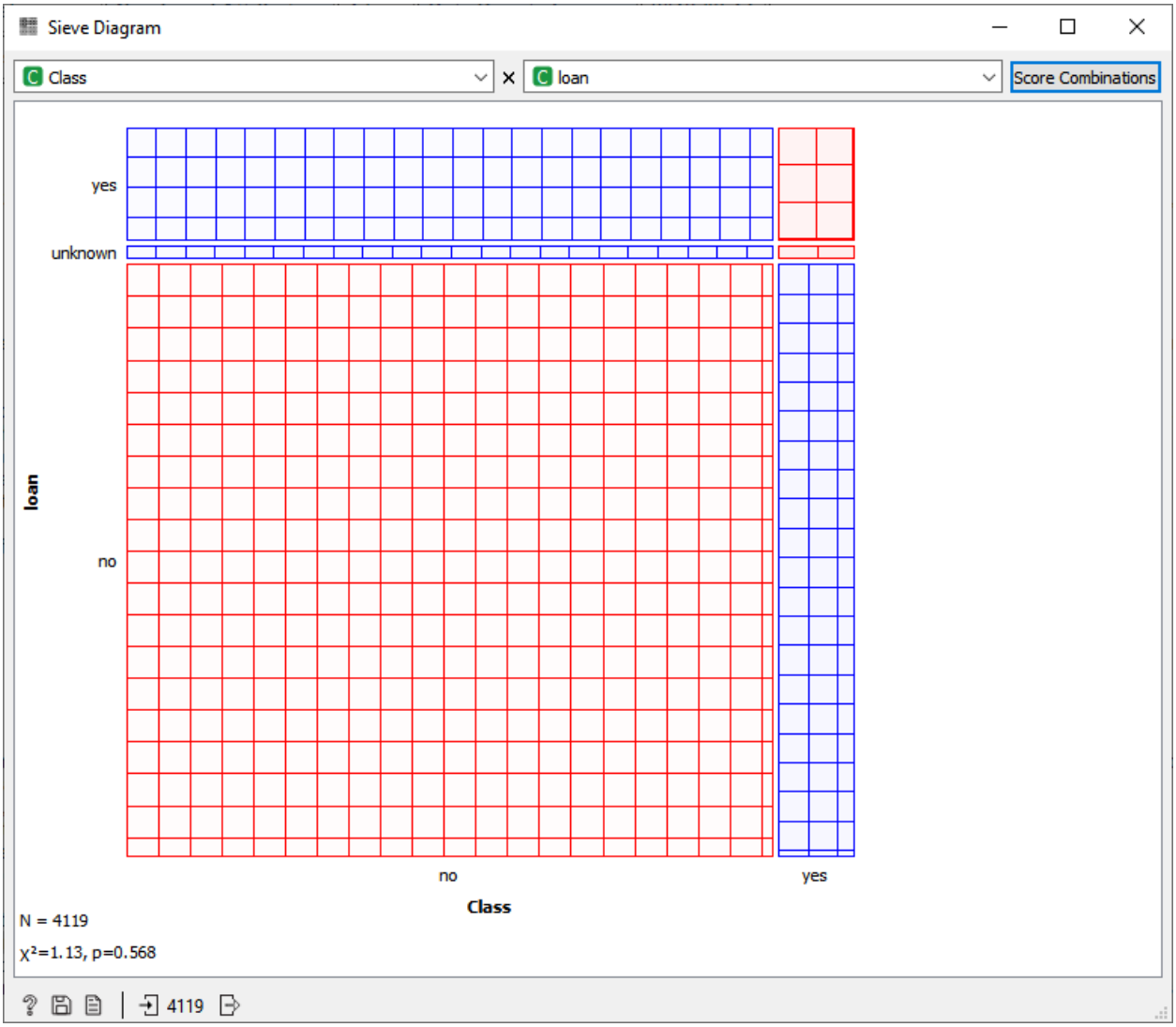
Согласно данным, клиенты в браке чаще согласны взять заем, нежели разведенные или не в браке.

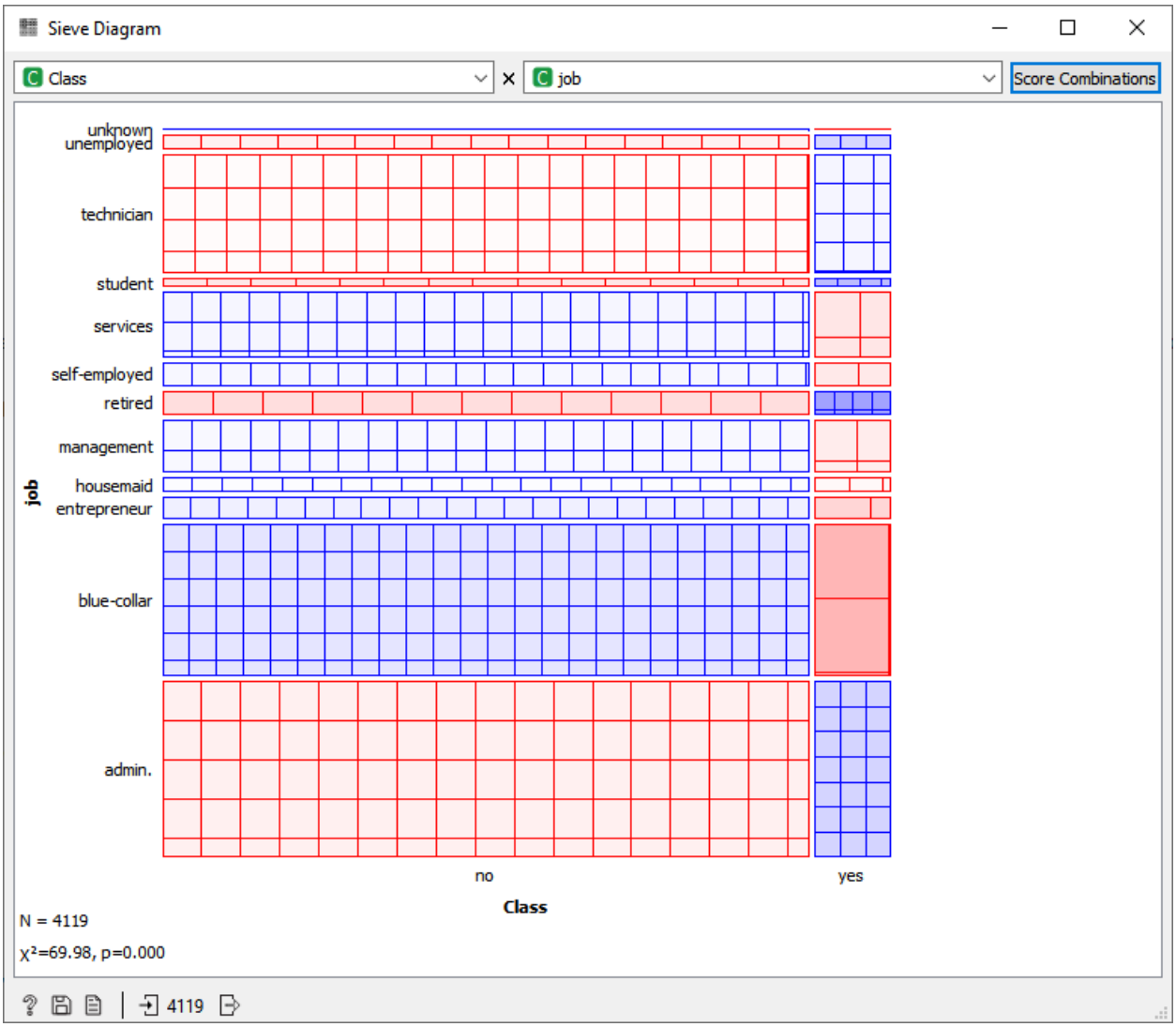


Среди общего количества опрошенных клиентов:

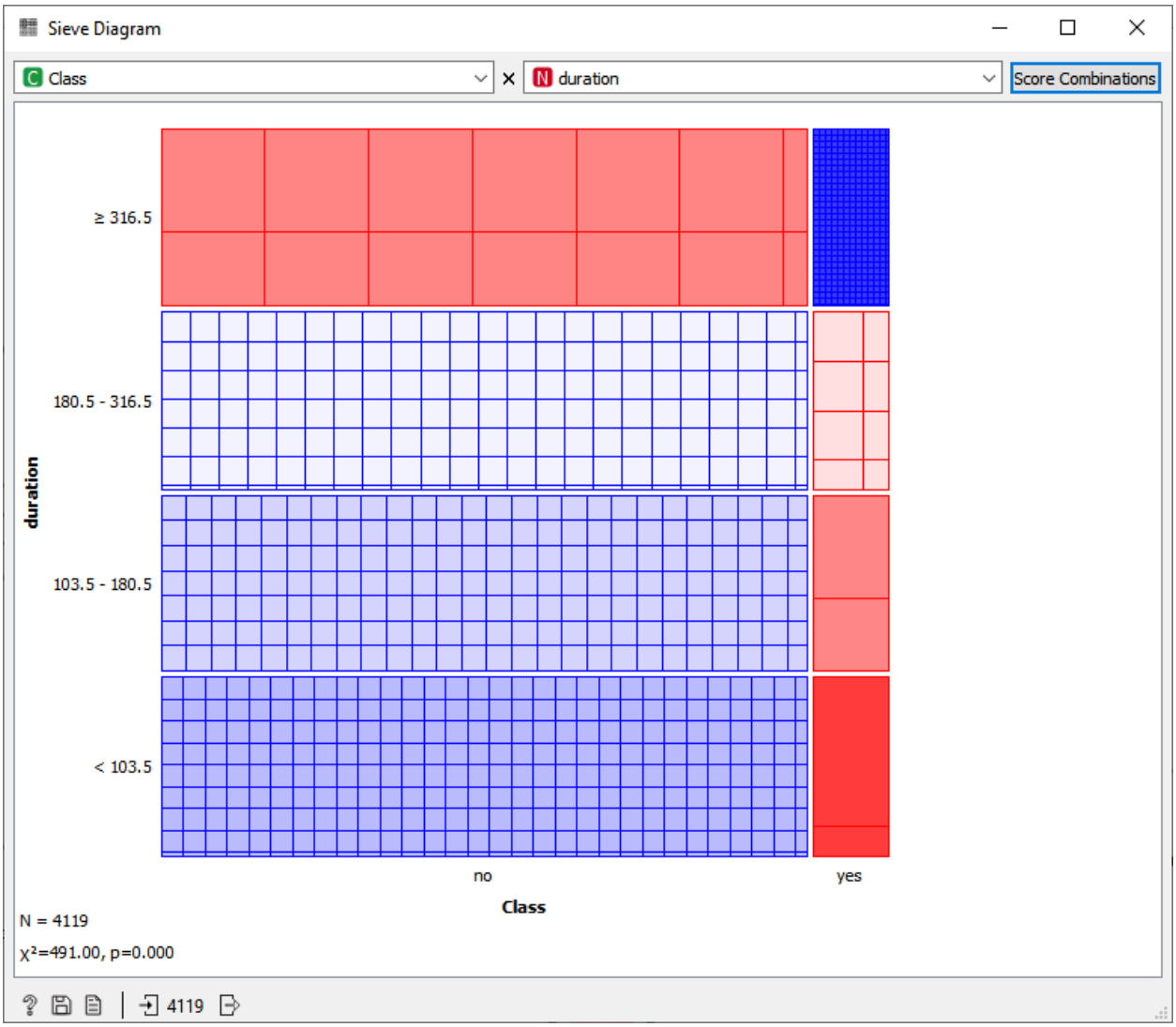
- больше всего их было в мае-августе, и на эти месяцы приходится основное количество тех, кто согласны взять заем.

Те, кто не имеет займа на текущий момент, более склонны взять его.





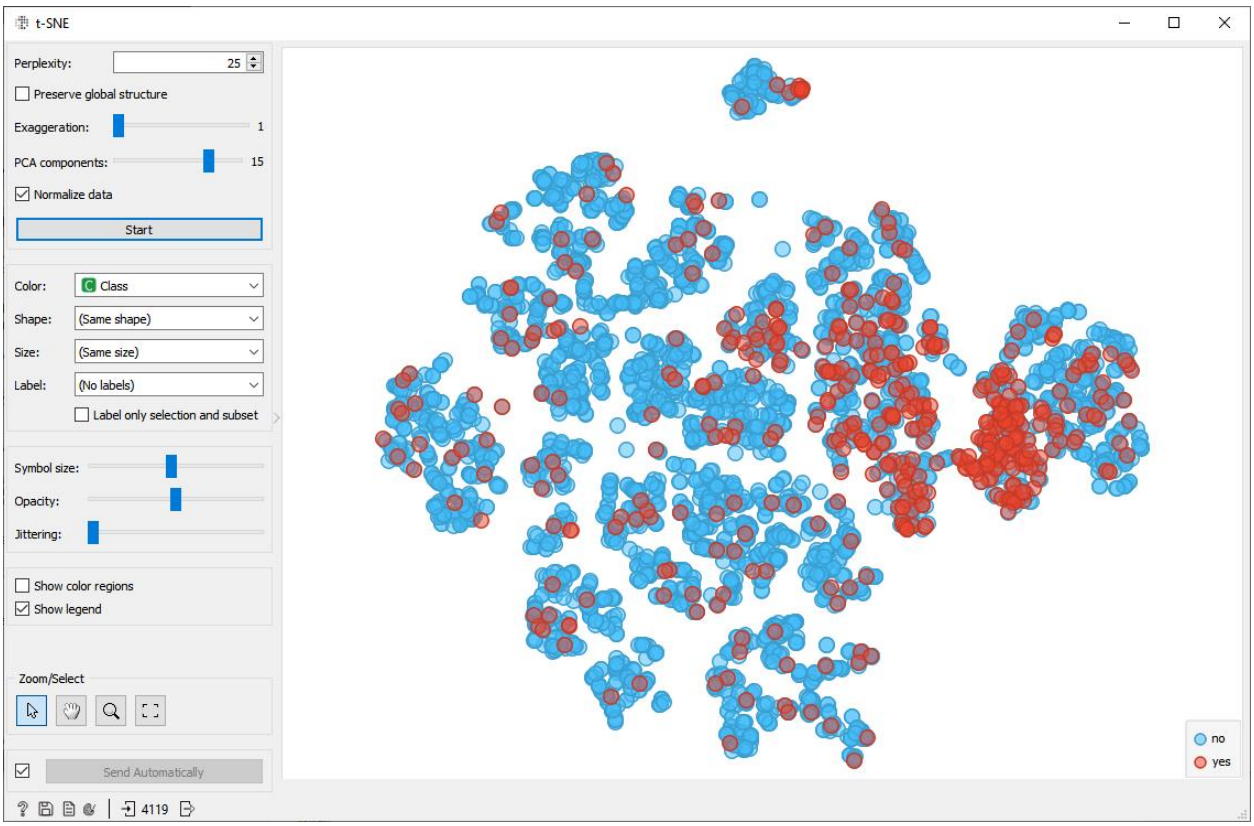

В зависимости от трудоустройства, чаще всего согласный взять заем административные работники или люди, занятые ручным трудом.




Как видно из графика, чем дольше продлился телефонный звонок, тем с больше шанс того, что клиент согласится на заем.

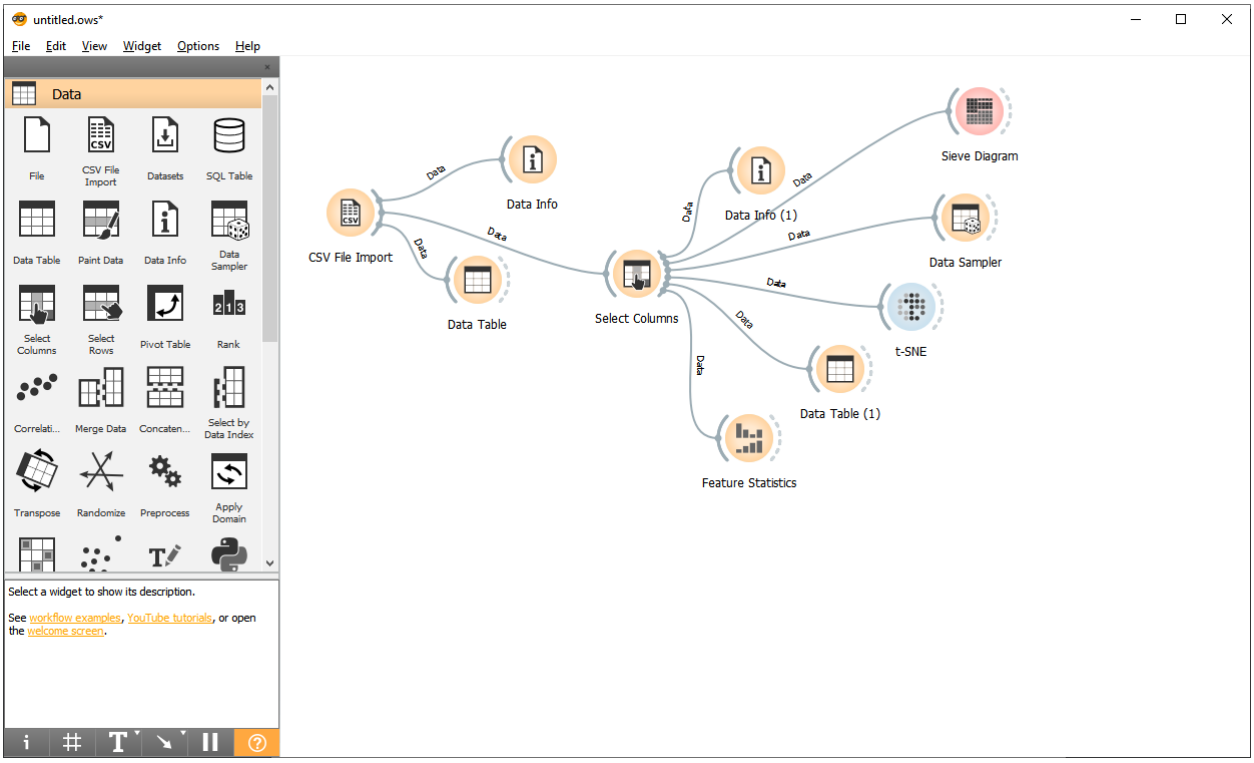
Попробуем также визуализировать данные.

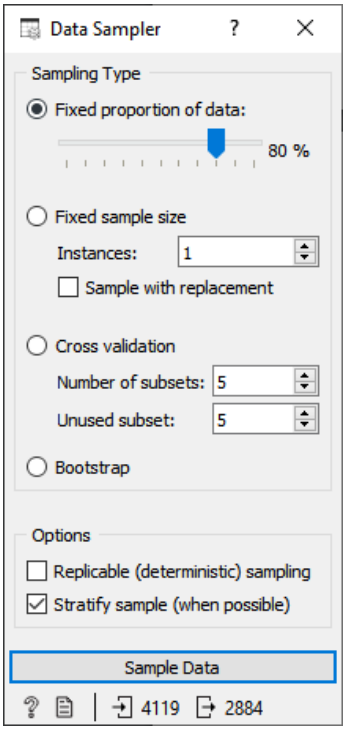
Инструмент t-SNE хорошо подходит для визуализации данных, за счёт того, что производит сокращения размерности данных до двух переменных, что приводит к тому, что можно оценить, как близки/далеки данные.




Классы между собой довольно сильно перемешаны, но и есть зона концентрации клиентов, которые согласились подписать заем.

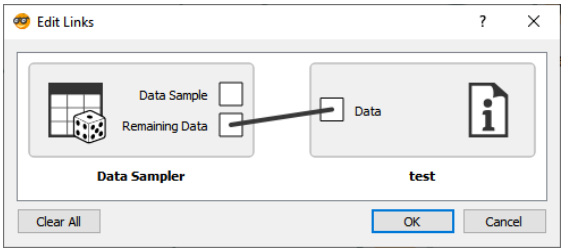
Добавим Data Sampler, чтобы разделить данные со стратификацией (сохранить требуемую пропорцию данных, чтобы в трейн-тест попали требуемые доли объектов классов).

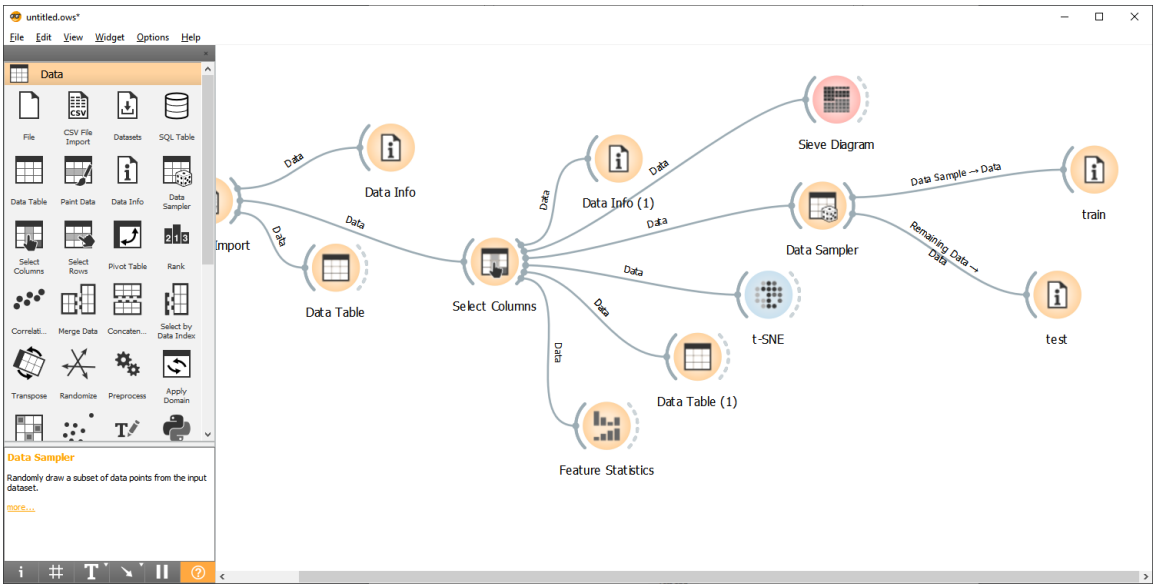


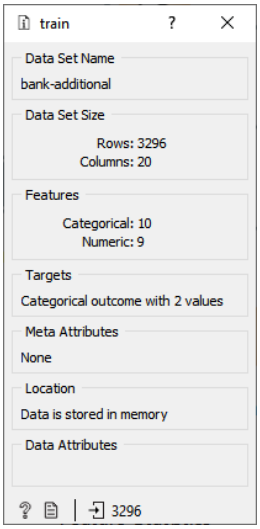



Перед этим важно удостовериться, что настройки связи выставлены корректно.

Пример для тестовой выборки.



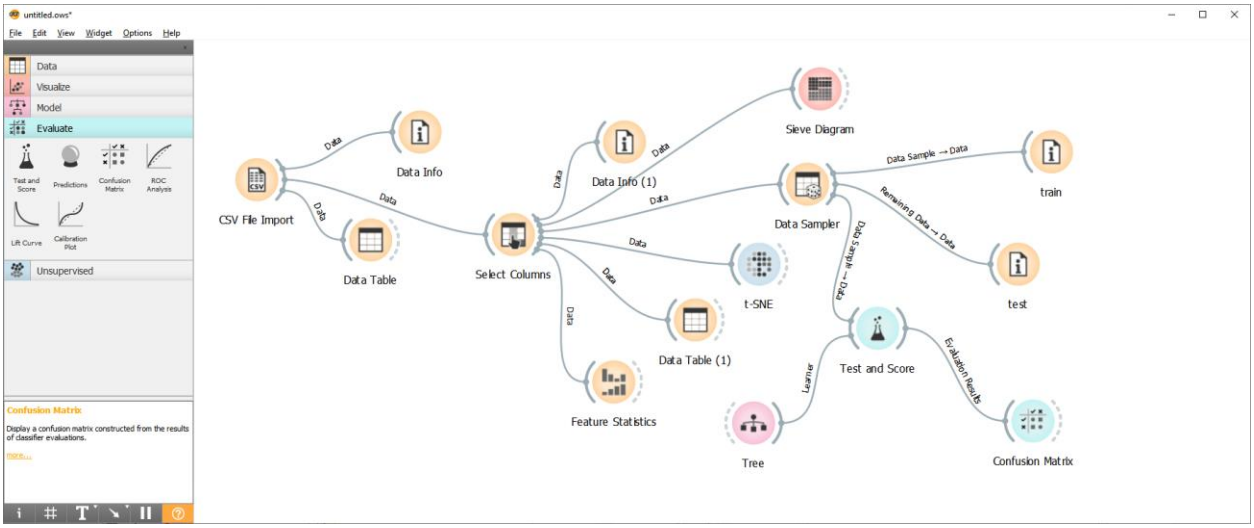



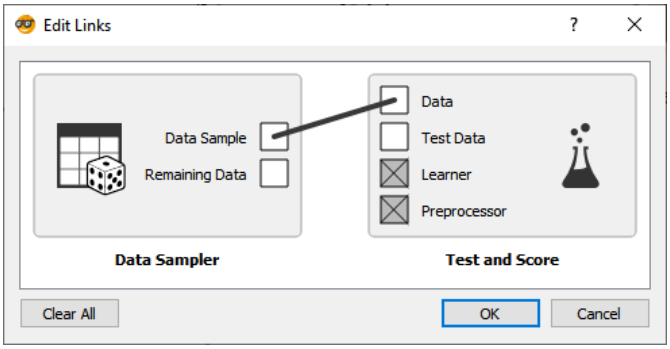

С помощью data info, можно удостовериться, что в трейн попало 80% данных (3296 записей).

Для оценки эффективности классификации добавим из раздела Evaluate блок Test and Score.

Также добавим Tree (решающее дерево), в качестве алгоритма классификации.

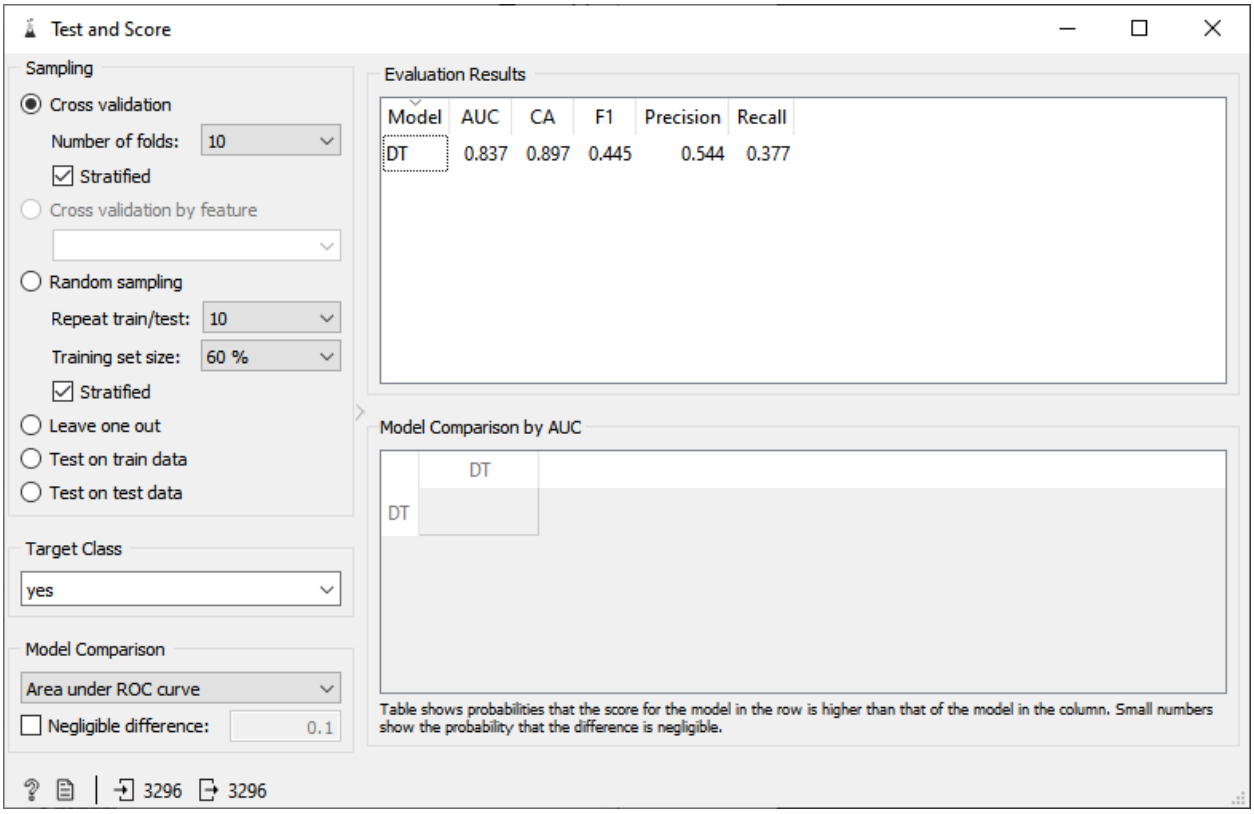



Убедимся, что параметры связи выставлены так, чтобы в блок Test and Score попадал тренировочный набор данных.

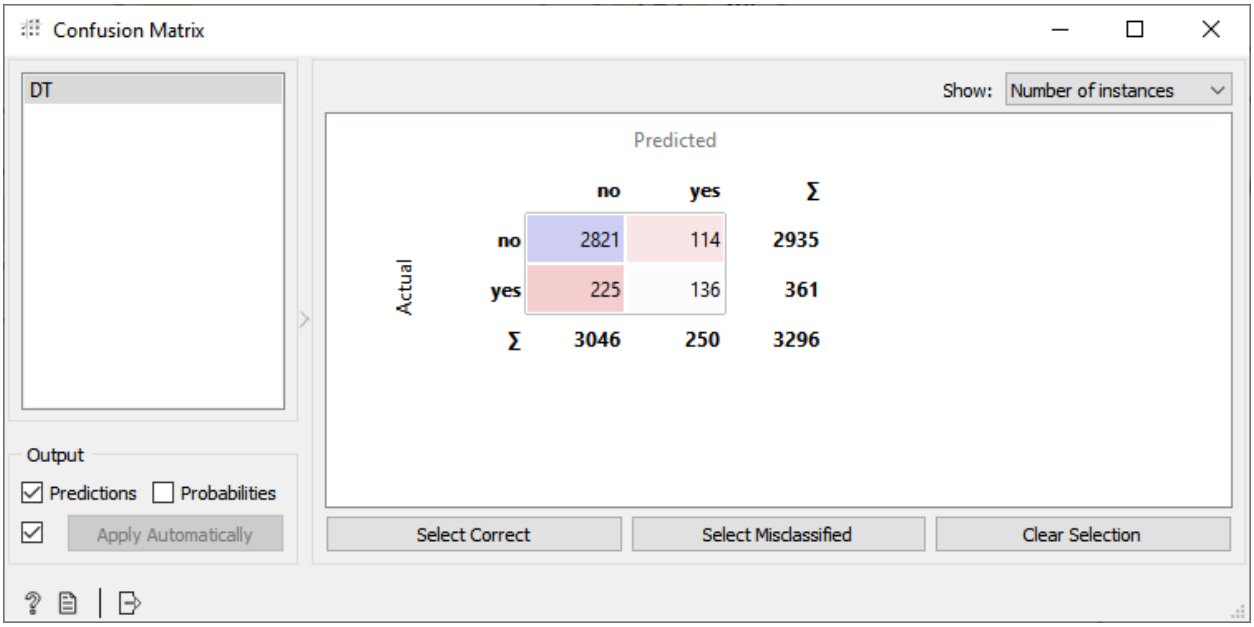


В настройках блока Test and Score необходимо выберем Cross validation, выставим Target Class в значение yes.

Кросс-валидация будет происходит на 10 фолдах со стратификацией.

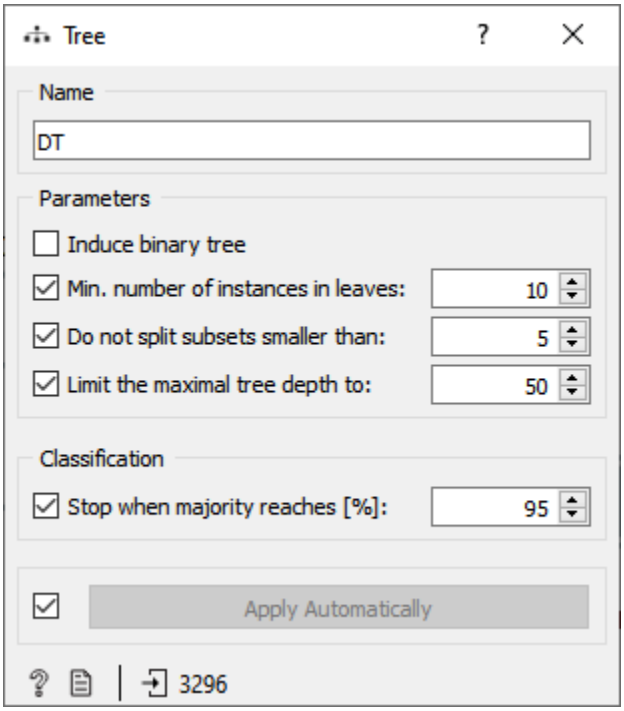



Добавим из Evaluate блок Confusion Matrix:

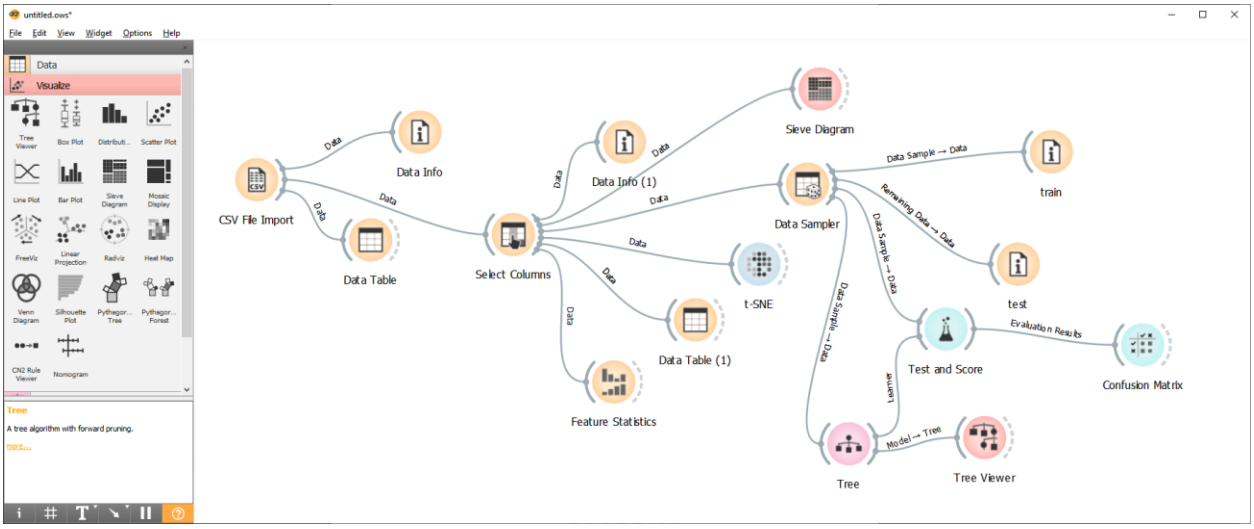


Данный блок позволяет посмотреть матрицу ошибок классификации.

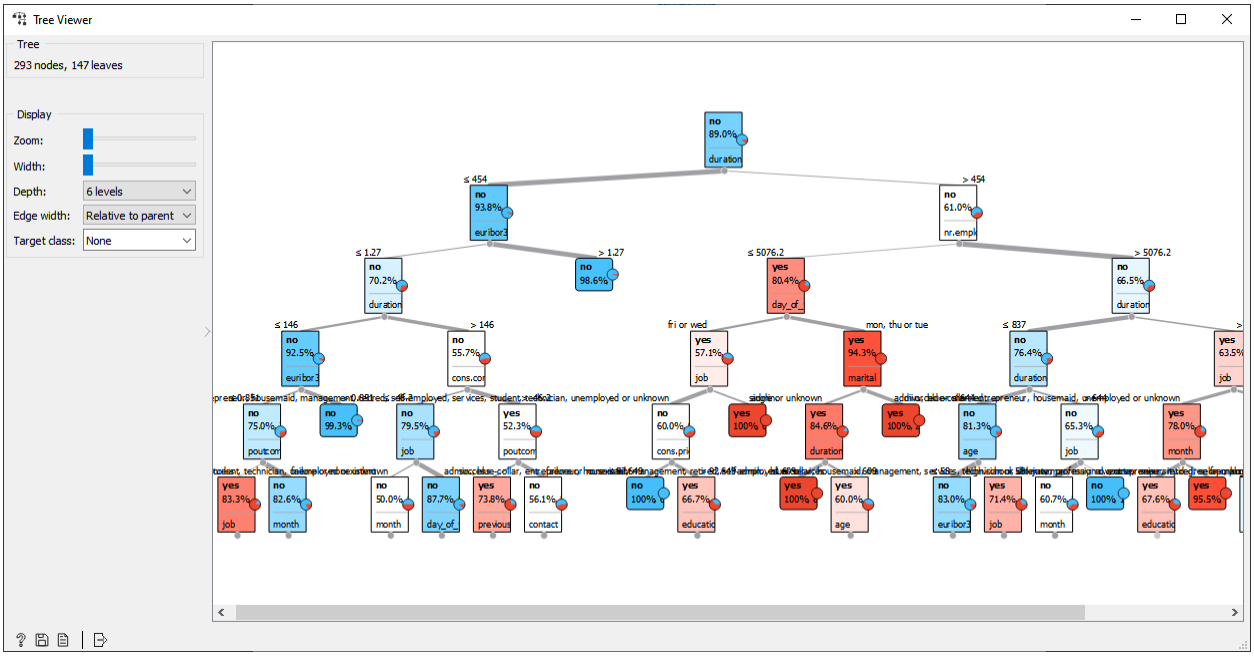
При следующих настройках алгоритма удалось добиться наилучших результатов.



Используем также блок Tree Viewer из раздела Visualize.




В данном блоке можно изучить построенное дерево принятия решений.




Как видно из дерева, первое разбиение идет по длительности звонка, это достаточно хорошо видно по графику Sieve Diagram, затем по количеству сотрудников.

Заключение.

В данной работе мы познакомились с таким инструментом как Orange, который служит для анализа данных.

В ходе работы были произведены следующие действия:

1. Установка Orange Data Mining.
2. Загрузка набора данных bank-additional.csv.
3. Произведен интеллектуальный анализ данных.
4. Построено и обучено дерево принятия решений.
5. Оценили качество построенной модели на тестовой выборке.
6. Визуализировали дерево принятия решений и изучили его структуру.