**Основная задача:** Проанализировать клиентов в выборке, провести кластеризацию, сделать описание сегментов.

**Алгоритм решения:**

1. Загрузка данных.
2. Построение scatter\_matrix, оценка распределения признаков.
3. Масштабирование данные (MaxAbsScaler).
4. Выбор метода кластеризации.
5. Кластеризация данных с различными параметрами выбранного алгоритма.
6. Оценка полученных результатов

**Результаты:**

Признаки "возраст" и "стаж вождения" сильно коррелированы, и на проекциях с другими признаками дают сильно похожие матрицы разброса данных. Значит их можно рассматривать как один признак (например, стаж).

Проанализировав построенную матрицу разброса можно утверждать, что для нашей выборки можно с высокой степенью точности посчитать наиболее удачное количество итоговых кластеров. Однако это, скорее, частный случай, который мы имеем в следствие достаточно хорошо разделимой выборки и небольшой степени размерности признаков.

Рассмотрим более общий случай, когда мы не можем качественно оценить число итоговых кластеров. Тогда лучшим выбором в качестве метода кластеризации представляется DBSCAN, который достаточно хорошо работает, когда у нас небольшая выборка, и предположительно, небольшое количество кластеров различной формы.

Примеры разбиения выборки с различными значениями параметра eps метода DBSCAN. В качестве значений признаков взяты средние значения по кластеру

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| eps | Кластер | Возраст, лет | Стаж, лет | Убыточность, % | Уровень заработной платы, руб/год | Размер кластера |
| 0.5 | 1 | 27 | 5 | 345 | 424 954 | 242 |
| 2 | 63 | 37 | 109 | 1 610 847 | 242 |
|  | | | | | | |
| 0.3 | 1 | 27 | 5 | 345 | 424 459 | 242 |
| 2 | 70 | 46 | 152 | 242 960 | 121 |
| 3 | 57 | 30 | 66 | 2 978 736 | 121 |
|  | | | | | | |
| 0.13 | 1 | 25 | 3 | 200 | 700 447 | 121 |
| 2 | 70 | 46 | 152 | 242 960 | 121 |
| 3 | 30 | 7 | 487 | 149 461 | 121 |
| 4 | 57 | 30 | 66 | 2 978 736 | 121 |
|  | | | | | | |
| 0.10 | 1 | 25 | 3 | 203 | 700 447 | 121 |
| 2 | 70 | 46 | 152 | 242 960 | 121 |
| 3 | 30 | 7 | 487 | 149 461 | 121 |
| 4 | 54 | 30 | 63 | 2 773 575 | 69 |
| 5 | 54 | 31 | 68 | 2 946 672 | 23 |
| 6 | 64 | 30 | 68 | 3 686 482 | 25 |
| 7 | 68 | 22 | 80 | 2 278 719 | 4 |

Разбиение на большее число кластеров не выглядит целесообразным (такой вывод можно сделать если посмотреть на кластеры №4, 5 в таблице для eps = 0.10).

Данный вывод также подтверждает применение алгоритмов аггломеративной кластеризации и k-means для числа кластеров больше 4-х.

Полученные кластеры можно интерпретировать следующим образом:

|  |  |
| --- | --- |
| 1 кластер | * низкий возраст * низкий стаж, * убыточность ниже средней, * невысокая З/П |
| 2 кластер | * очень высокий возраст * очень высокий стаж, * убыточность ниже средней, * низкая З/П |
| 3 кластер | * невысокий возраст * невысокий стаж, * высокая убыточность, * низкая средняя З/П |
| 4 кластер | * высокий возраст * высокий стаж, * низкая убыточность, * очень высокая З/П |

Примечание: низкий < невысокий < высокий < очень высокий: