

Классификация сегментации клиентов

Автор: Серегин Константин Александрович

Руководитель: Заграновская Анна Васильевна

Исходные данные

	Пол	Семейное положение	Возраст	Высшее образование	Профессия	Опыт работы	Уровень расходов	Размер семьи	Анонимная категория	Целевая категория
0	М	not married	22	нет	Медик	1.0	Высокий	4	Cat_4	D
1	Ж	married	38	есть	Инженер	NaN	Средний	3	Cat_4	A
2	Ж	married	67	есть	Инженер	1.0	Низкий	1	Cat_6	B
3	М	not married	67	есть	Юрист	0.0	Высокий	2	Cat_6	B
4	Ж	married	40	есть	Актер	NaN	Высокий	6	Cat_6	A

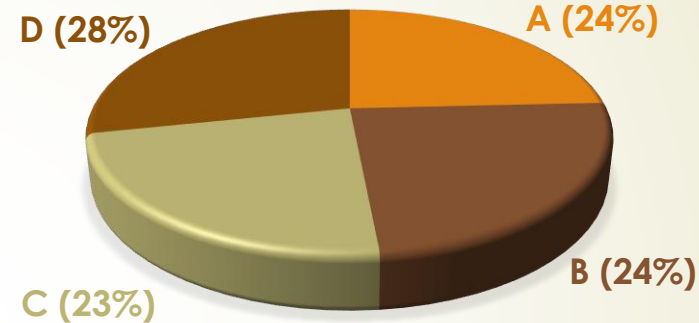
Постановка цели

1. Оценить имеющуюся разбивку клиентов на группы
2. При необходимости скорректировать сегментацию рынка
3. Обучить модель для классификации новых клиентов

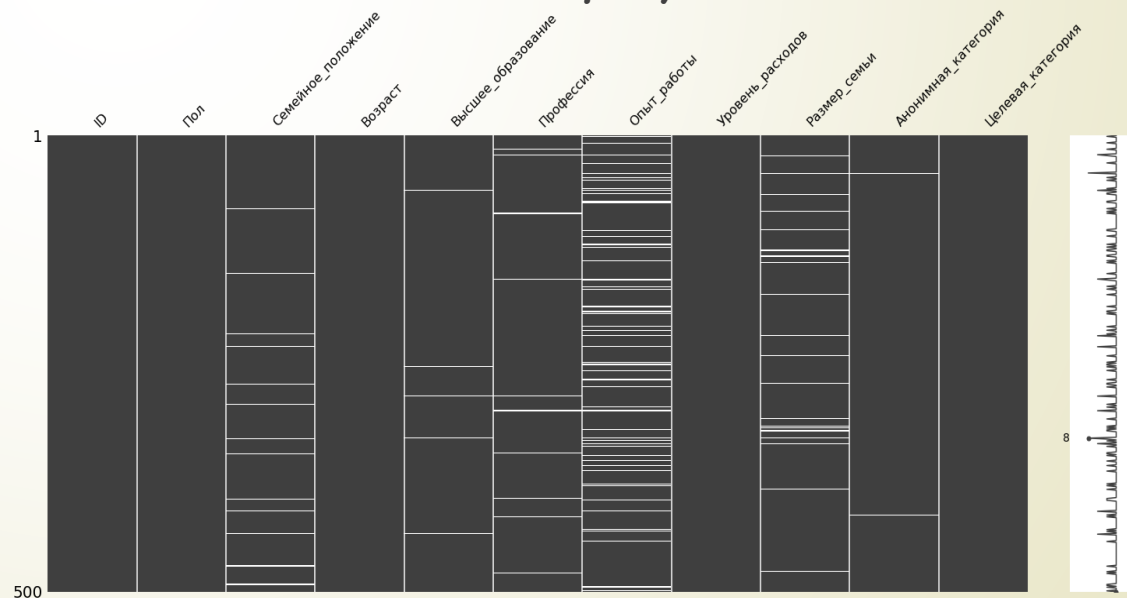
Первичный анализ данных

1. Исходные данные состоят из 8000 строк и 10-ти столбцов, из которых 9 с факторными признакам и 1 результативный (целевой) столбец.
2. Выбросы отсутствуют.
3. Четыре целевых категории распределены равномерно.
4. Факторные признаки представляют собой в основном категориальные данные.
5. Данные имеют пропущенные значения.

РАСПРЕДЕЛЕНИЕ
РЕЗУЛЬТАТИВНОГО
ПРИЗНАКОВ

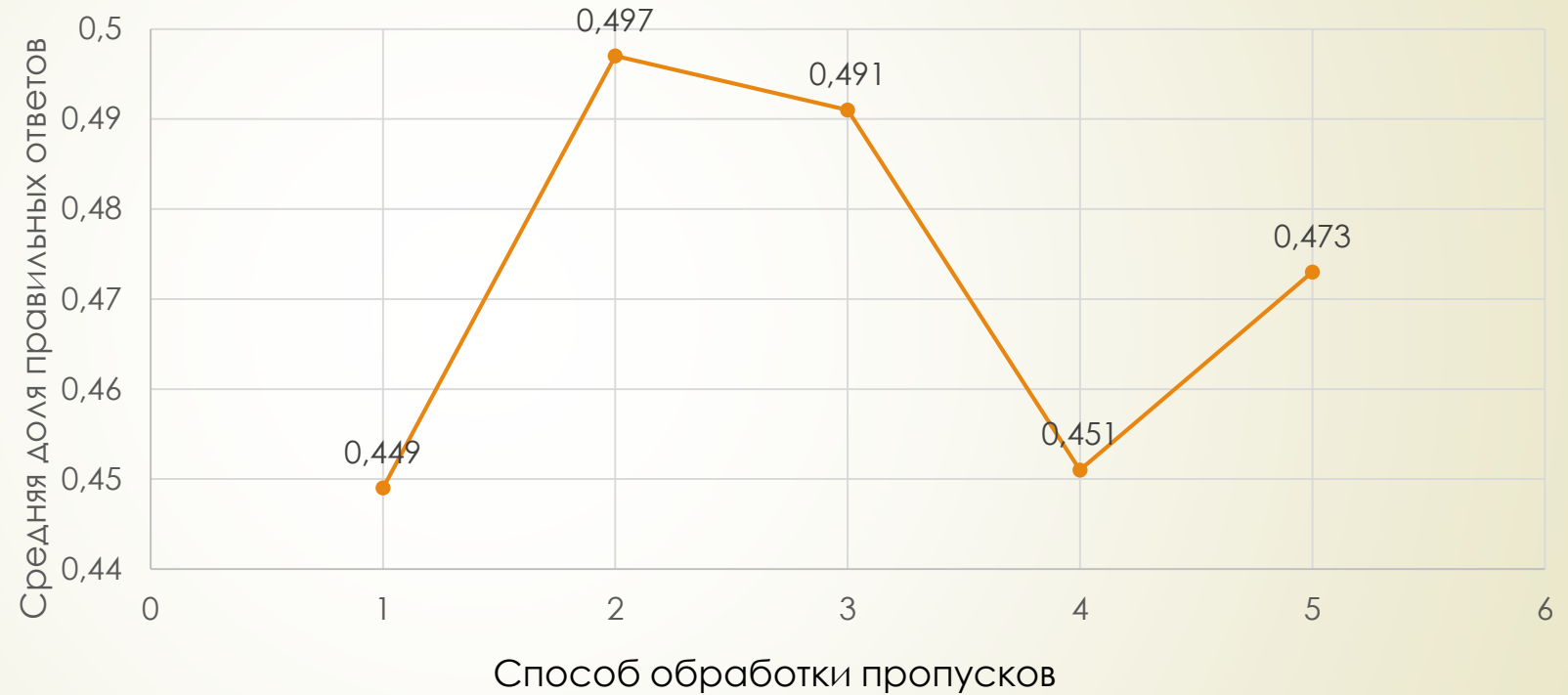


➤ Наличие пропусков



Выбор способа удаления пропусков с помощью модели случайного леса

1. Данные без удаления пропусков.
2. Пропуски заполняются соседними значениями.
3. Все строки с пропущенными данными удаляются.
4. Все строки с пропусками удаляются и удаляются дубликаты.
5. Все строки с пропусками удаляются, удаляются дубликаты и удаляются противоречивые данные.



Отбор числа признаков



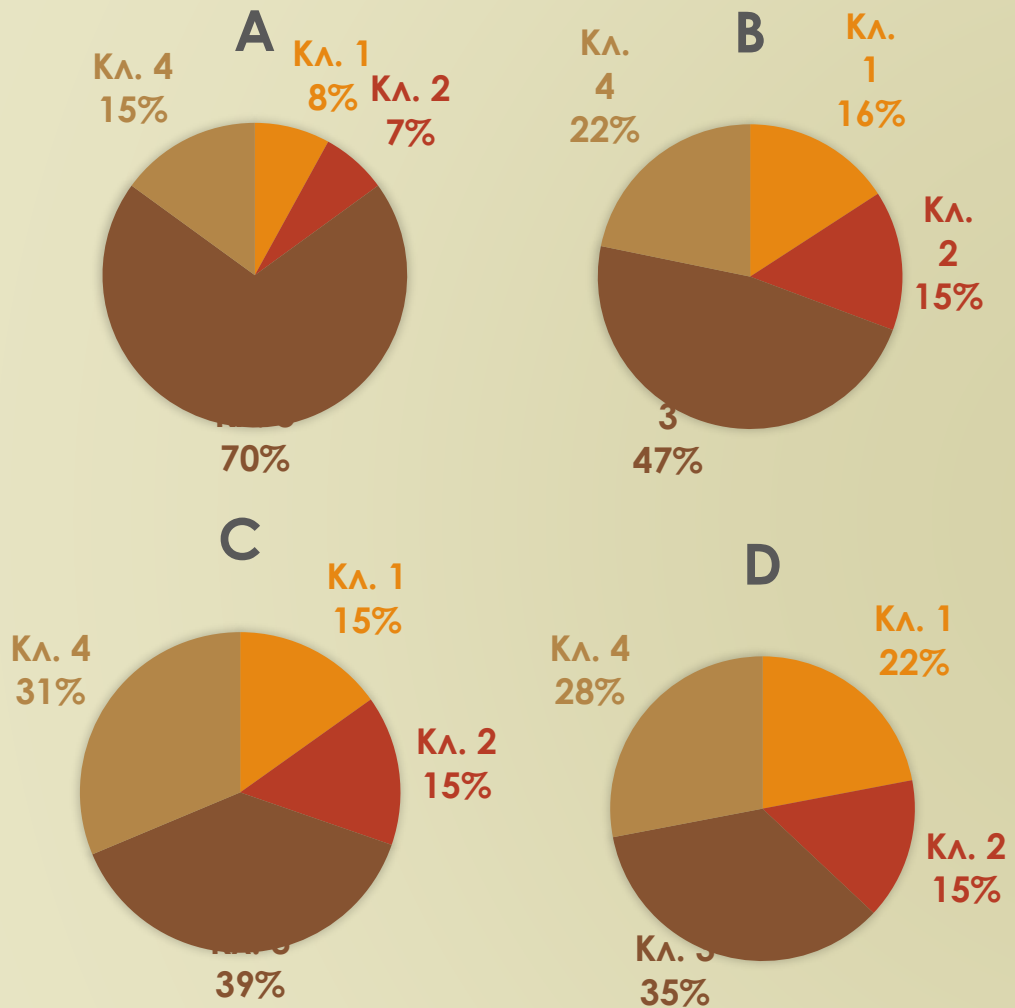
Важные признаки: «Возраст», «Семейное положение», «Образование»

Несущественные признаки: «Пол», «Опыт работы», «Категория товара»

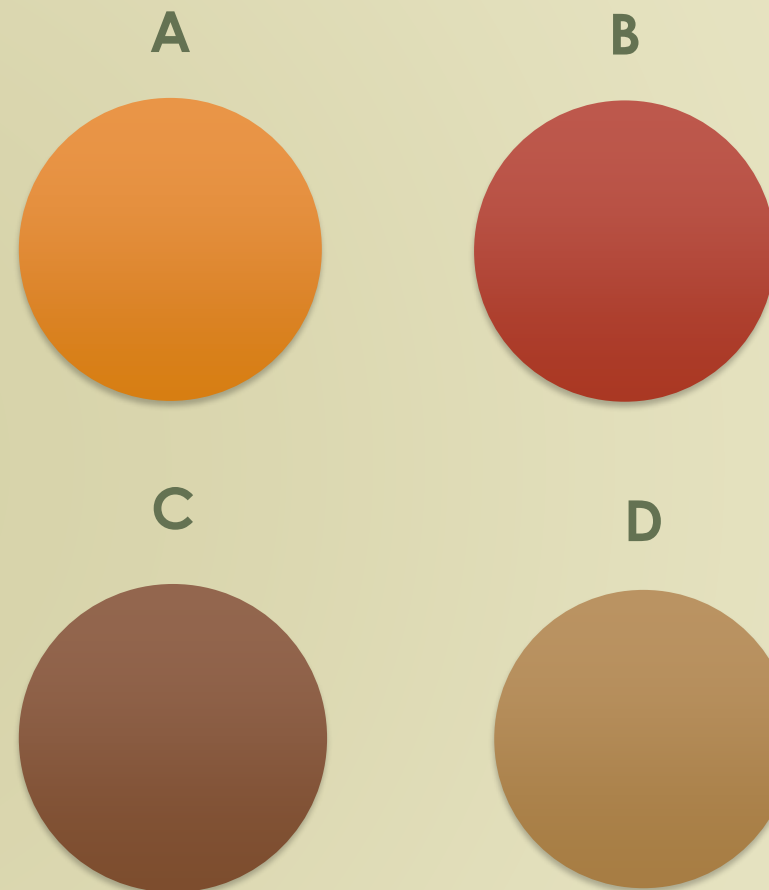


Кластерный анализ

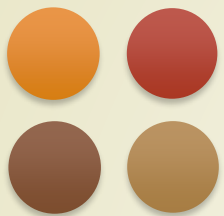
Кластеризация на исходных данных



Ожидаемая кластеризация

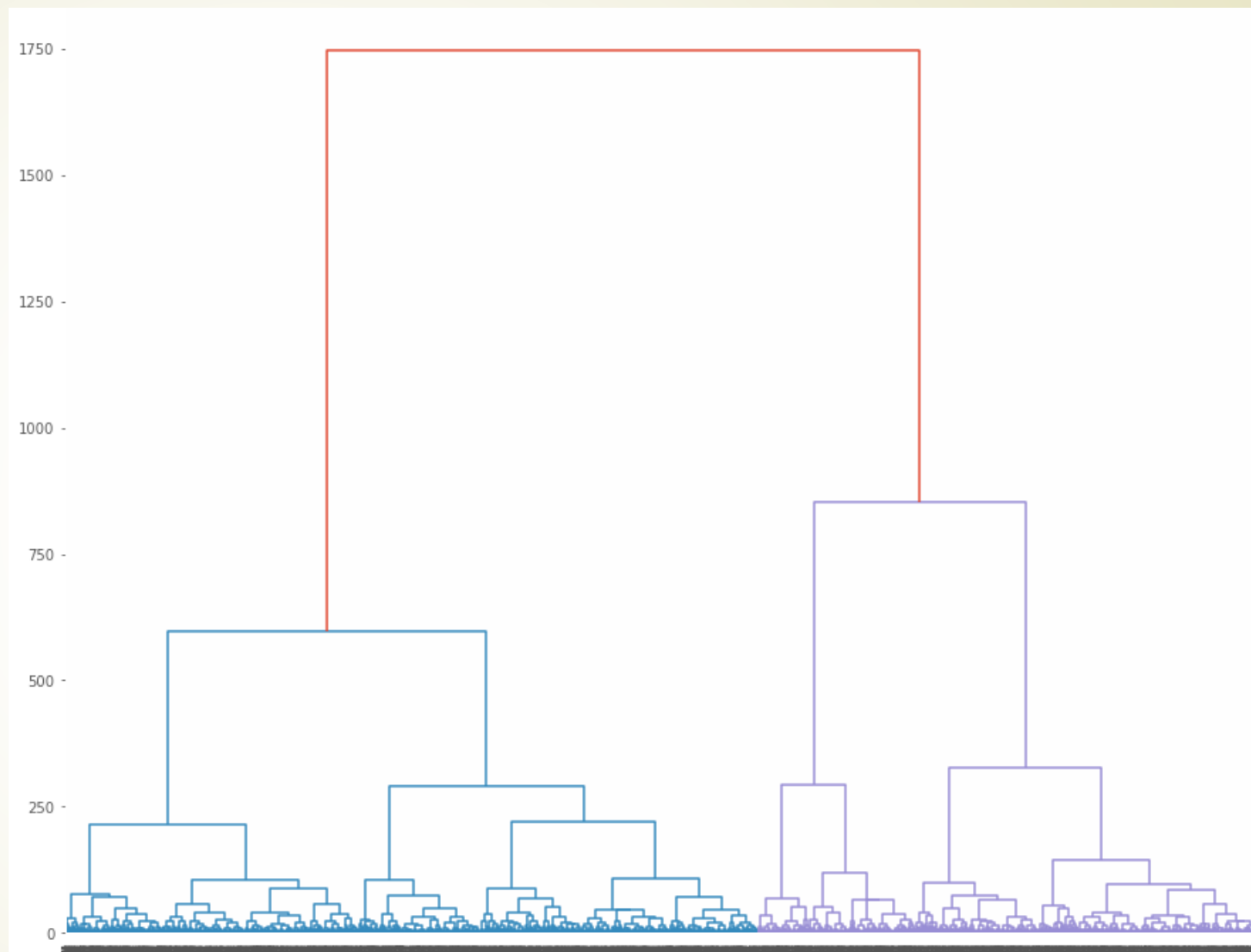



Кластерный анализ на основе k лучших признаков,
где $k = 1, \dots, 8$



Иерархическая кластеризация

- ▶ Дендрограмма показывает, что данные уместней всего разделить 2, 3 или 4 группы.





Классификация клиентов по
исходной разбивке на классы
(обучение моделей)

Для классификации были использованы:

Линейные модели:

- Модель логистической регрессии (`LogisticRegression()`)
- Метод опорных векторов (`SVC()`)

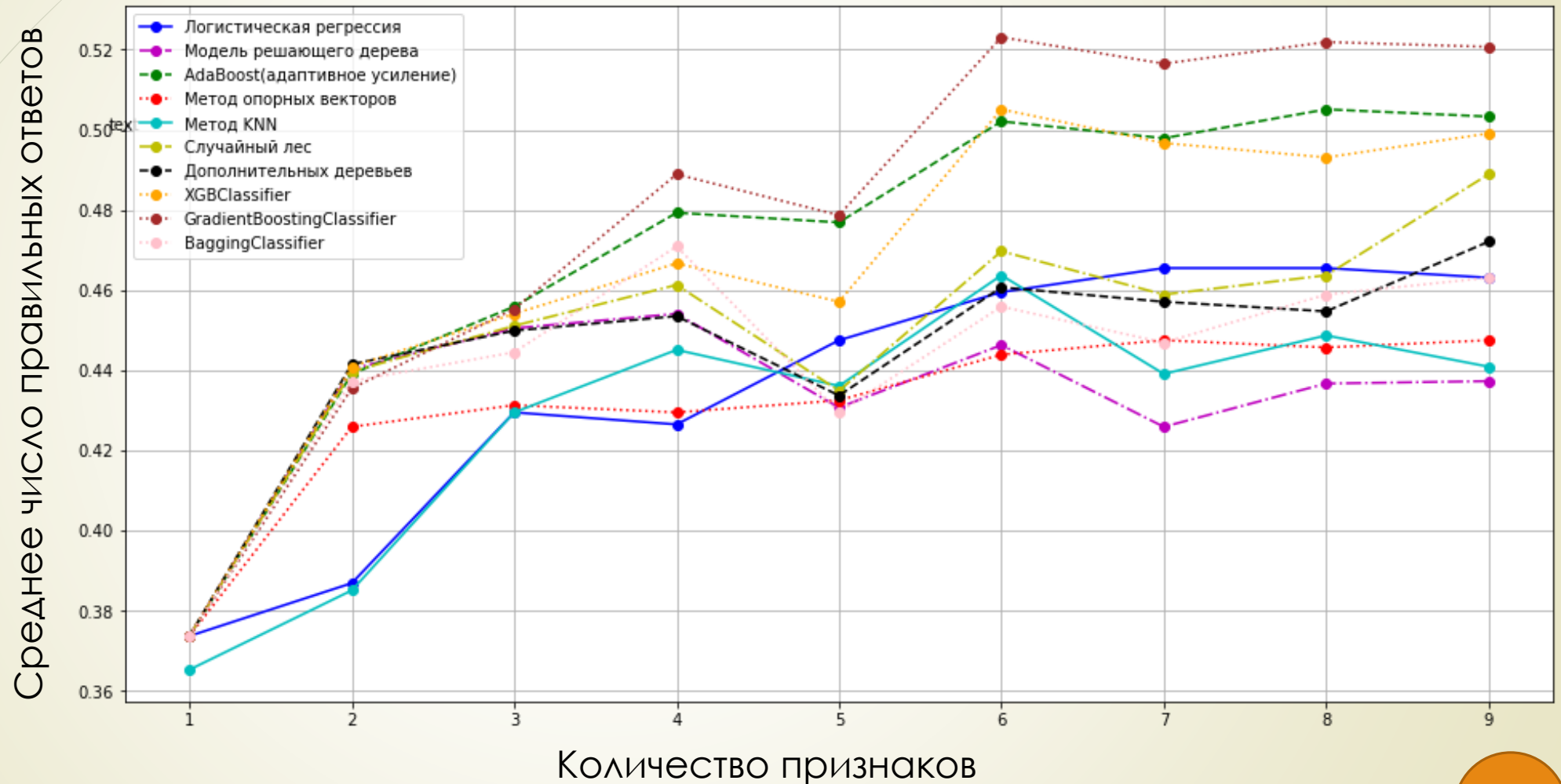
Нелинейные модели:

- Метод kNN (`KNeighborsClassifier()`)
- Модель решающего дерева (`DecisionTreeClassifier()`)

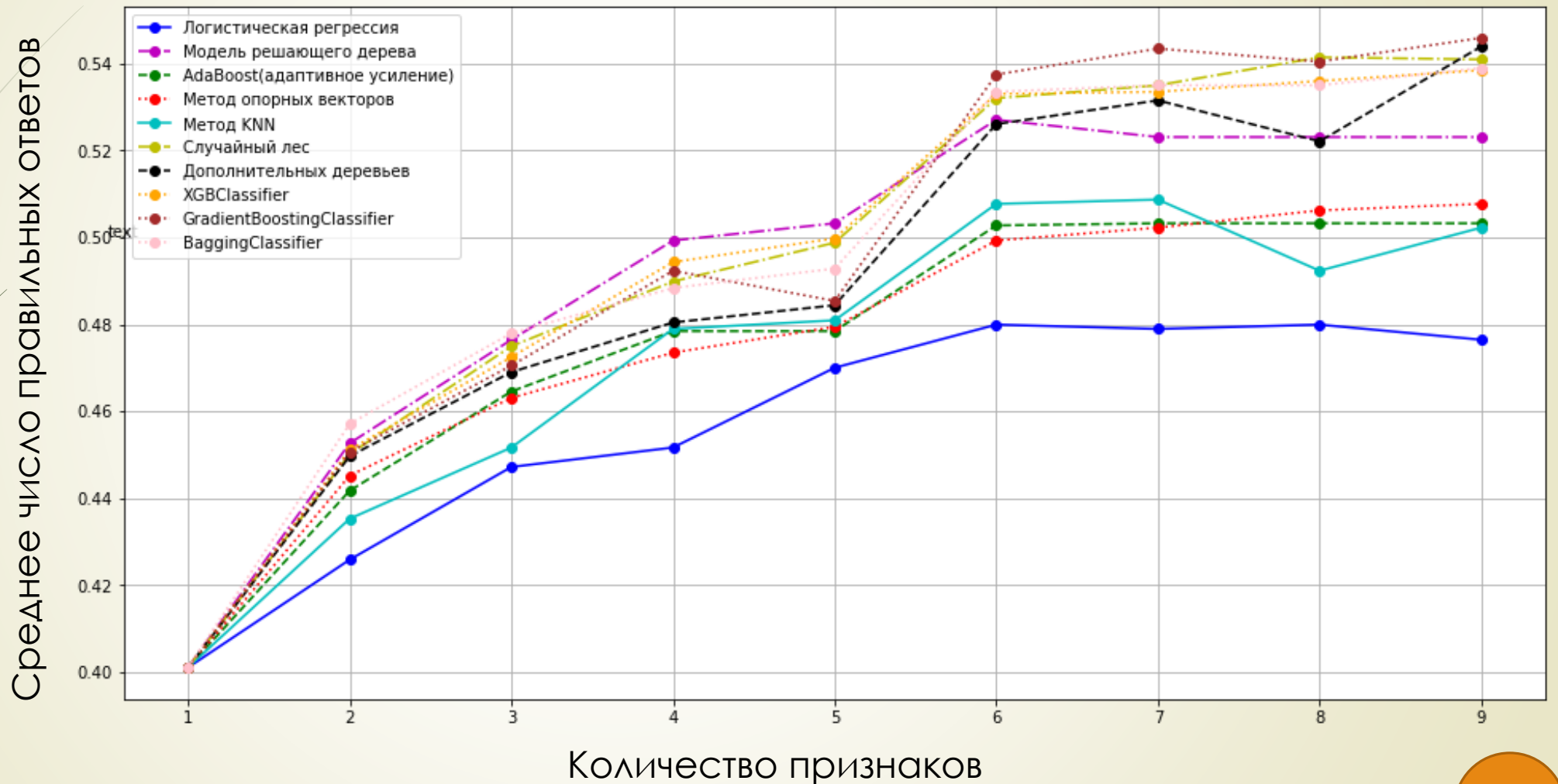
Ансамблевые методы:


- Метод пакетирования `BaggingClassifier()`
- Метод случайного леса `RandomForestClassifier()`
- Метод дополнительных деревьев (`ExtraTreesClassifier()`)
- Метод `AdaBoostClassifier()`
- Метод `GradientBoostingClassifier()`
- Метод `XGBClassifier()`

Обучение моделей с базовыми гиперпараметрами



Обучение моделей с гиперпараметрами, отобранными методом GridSearchCV



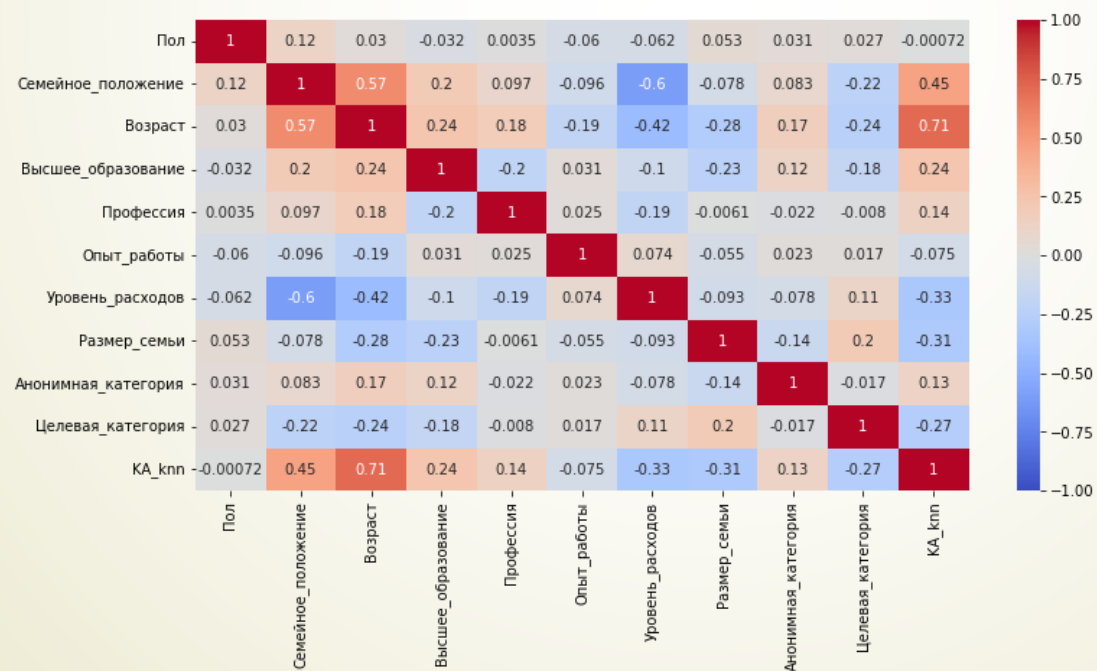


Классификация клиентов по
разбивке на классы в соответствии с
результатами кластерного анализа

Таблица средних значений

cluster	Пол	Семейное_положение	Возраст	Высшее_образование	Профессия	Опыт_работы	Уровень_расходов	Размер_семьи	Анонимная_категория	Целевая_категория
1	0.519298	0.616541	39.005514	0.768421	3.046115	3.528822	2.472682	2.495238	5.095739	2.196992
2	0.534489	0.181114	26.625231	0.376614	3.320177	3.030247	2.840649	3.462929	4.897824	3.081151
3	0.581767	0.951128	74.921992	0.625000	4.957707	1.118421	1.946429	2.038534	5.599624	2.328008
4	0.571180	0.859817	52.659556	0.784502	2.962995	2.075316	2.215934	2.794515	5.252068	2.375272

Корреляционная матрица по результатам кластеризации





Наименования кластеров

- ▶ Кластер 1 – Зрелый возраст (≈ 39 лет)
- ▶ Кластер 2 – Молодежь (≈ 26 лет)
- ▶ Кластер 3 – Пожилые люди (≈ 75 лет)
- ▶ Кластер 4 – Средний возраст (≈ 52 года)

Результаты классификации на основе модели GBM для кластеризованных данных

Метод `classification_report`

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	572
1.0	1.00	1.00	1.00	255
2.0	1.00	1.00	1.00	528
3.0	1.00	0.99	1.00	662
accuracy			1.00	2017
macro avg	1.00	1.00	1.00	2017
weighted avg	1.00	1.00	1.00	2017

Метод `confusion_matrix`

```
[[571  0  0  1]
 [ 0 255  0  0]
 [ 0  0 528  0]
 [ 5  0  0 657]]
```

Средняя доля правильных ответов: **0.9981405**

ВЫВОДЫ

1. Классификация исходных данных не корректна
2. Разбивка на 4 класса возможна
3. Главный критерий – возраст
4. Модели на исходных данных дают низкую точность – 55%
5. Модели на кластеризованных данных дают 100%-ю точность