# Аналитический отчет

**Цель исследования**: с помощью данных о покупках клиентов и их социально-демографических признаках проанализировать эффективность уже проведённых ранее маркетинговых кампаний и выявить факторы, способные повысить продажи.

В ходе проведения исследования были проведены:

- предобработка имеющихся данных магазина спортивных товаров;
- использование бинарной классификации, для заполнения потерянных данных;
- проведения A/B-тестирования для определения эффективности маркетинговой кампании;
- разбиение аудитории на кластеры, для дальнейшей эффективной персональной работы с каждым;
- построение модели склонности к покупке.

## Исследовательский анализ данных

Данные представлены в базе данных, содержащей следующие таблицы:

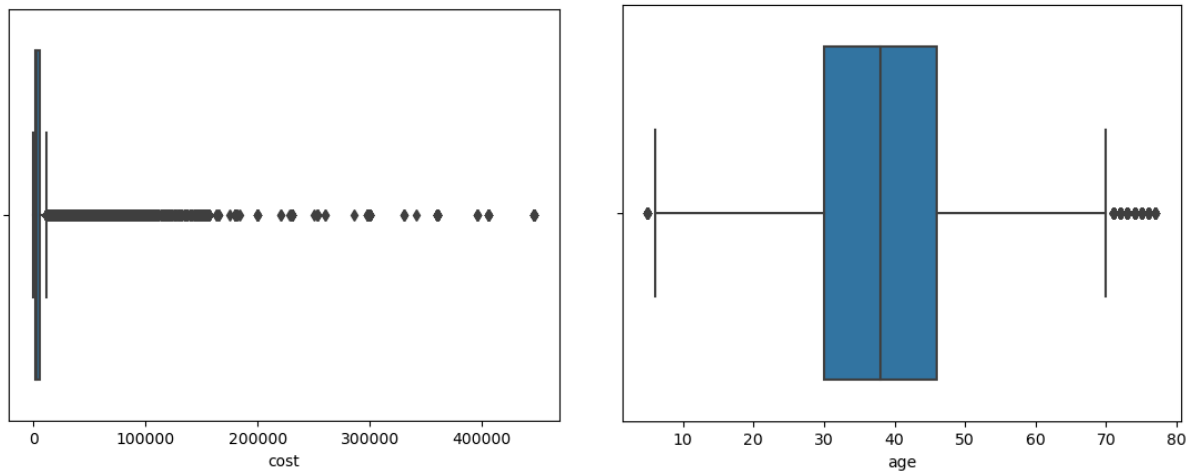
- **personal_data** — ID клиентов, их пол, возраст, образование, страна и город проживания;
- **personal_data_coeffs** — данные с персональными коэффициентами клиентов, которые рассчитываются по некоторой закрытой схеме;
- **purchases** — данные о покупках: ID покупателя, название товара, цвет, стоимость, гендерная принадлежность потенциальных покупателей товара, наличие скидки (поле base_sale. Значение 1 соответствует наличию скидки на момент покупки) и дата покупки.

Также отдельный файл **personal_data.csv.gz**, в котором некоторые данные утеряны и требуют восстановления.

*Подготовка и очистка данных*.

- пропуски в столбце product_sex заменены на «2», подразумевая товар unisex;
- пропуски в столбце colour были заменены на «нет данных». Там, где в цветах присутствовал символ «/», было заменено на «в нескольких цветах». Также все варианты цветов с приставками (темно-, светло-, ярко-, бледно- и другие), различные оттенки были скорректированы и приведены к более стандартной палитре, где в итоге оказалось 12 позиций;
- был создан столбец name_product, где отображается первое слово из столбца product, в котором очень много информации и с которым сложно работать в дальнейшем;
- добавлен столбец brand, в котором отображается бренд товара. Редко встречающиеся бренды были объединены в категорию «другое»;
- добавлен столбец category_product, в котором товары отсортированы по определенным категориям (одежда, обувь, спортинвентарь, аксессуары/уход/карты, отдых/туризм, спортивное питание);
- в столбце education, там, где возраст клиентов был меньшее 22 лет, значение «высшее» было заменено на «среднее»;

- в столбце cost, есть значительные выбросы (Иллюстрация 1). Но после оценки рынка, было принято решение оставить эти значения, так как стоимость оправдана;
- также были отсеяны клиенты, возраст которых меньше 14 лет и клиенты старше 70 лет (Иллюстрация 2).




Иллюстрации 1, 2. Визуализации (boxplot) распределения cost и age.

**Бинарная классификация**

Для восстановления данных с полом клиента (gender) использовалась бинарная классификация. В моделях использовались следующие признаки:

- категориальные – education, закодированы с помощью OneHotEncoder;
- числовые - age, city, country, personal_coef, стандартизированы с помощью StandardScaler.

Значения метрик для моделей прогнозирования представлены в Таблице 1.

Таблица 1.

| Модель | F1-score | Precision | Recall | Accuracy |
|---|---|---|---|---|
| Логистическая регрессия | 0,98 | 0,96 | 1,0 | 0,98 |
| Дерево решений | 1,0 | 1,0 | 1,0 | 1,0 |
| Случайный лес | 0,99 | 0,99 | 0,99 | 0,99 |

В результате, в качестве заполнения пропусков, использовалась модель дерева решений. Файл с заполненными данными сохранен в «Предсказание пола покупателей.csv».

**A/B – тестирование**

Кампания проводилась в период с 5-го по 16-й день, ID участвовавших в ней пользователей содержатся в файле ids_first_company_positive.txt. Эта кампания включала в себя предоставление персональной скидки 5 000 клиентов через email-рассылку. Помимо людей, которым предлагалась персональная скидка, были отобраны люди со схожими социально-демографическими признаками и покупками, которым скидку не предложили (файл ids_first_company_negative.txt).

В тестовой и контрольной группах предварительно убираем дубликаты и ID, которые попали и в ту, и в другую группу.

Используемые метрики, для оценки тестирования:

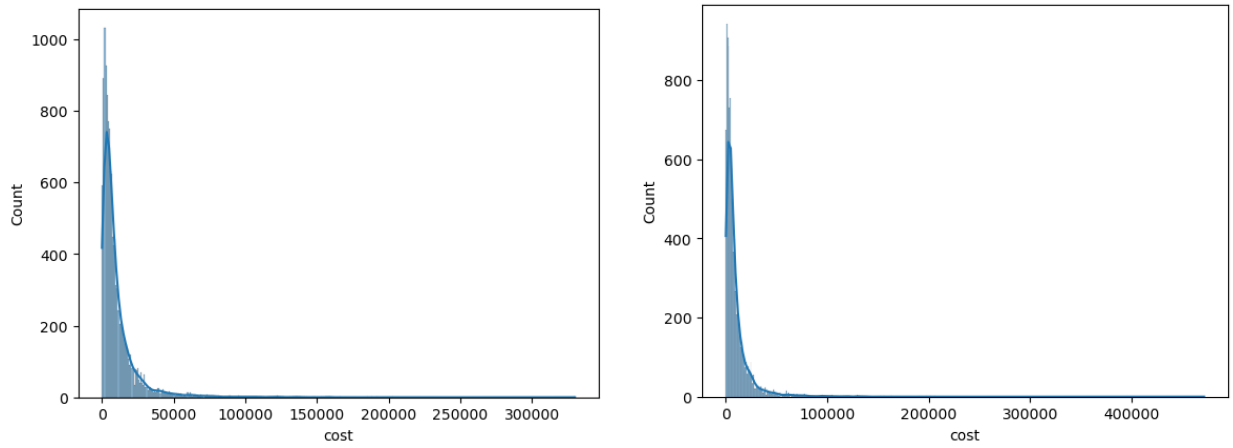
- средний чек (так как отсутствует идентификатор одного заказа можно предположить, что один день = одному чеку для конкретного клиента);
- выручка;
- число покупок на клиента;
- количество людей с повторными покупками.

Результаты тестирования представлены в Таблице 2.

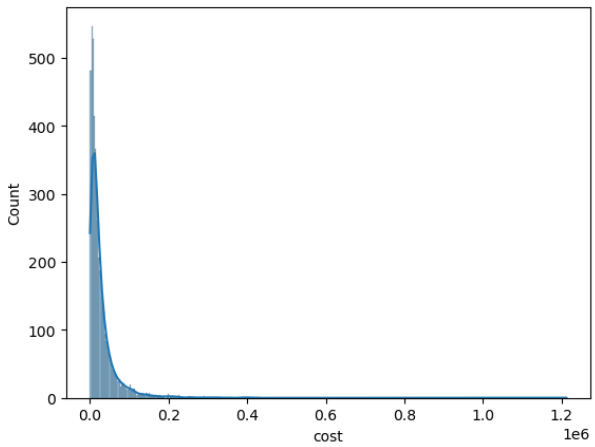
Таблица 2.

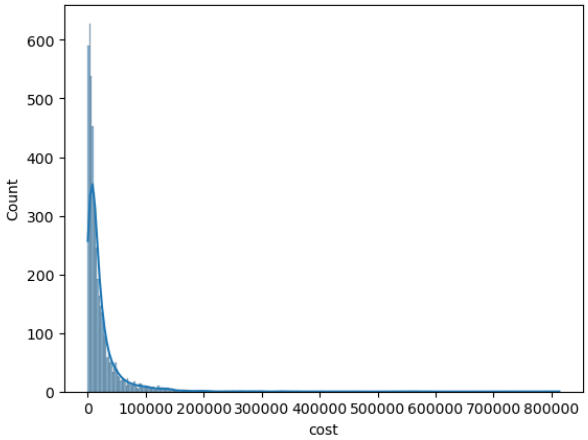
| Группа | Средний чек | Выручка | Число покупок на клиента | Количество людей с повторными покупками |
|---|---|---|---|---|
| test | 10590 | 131190662 | 5,23 | 3381 |
| control | 10071 | 109621740 | 4,00 | 2945 |

Различия в метриках оценивались с помощью статистических тестов. Так для значений среднего чека, выручки, и числа покупок распределение ненормальное (Иллюстрации 3-8), использовался тест Манна-Уитни. Так как в случае количества людей с повторными покупками мы считаем долю, то можем использовать Z-тест. Результаты представлены в Таблице 3.

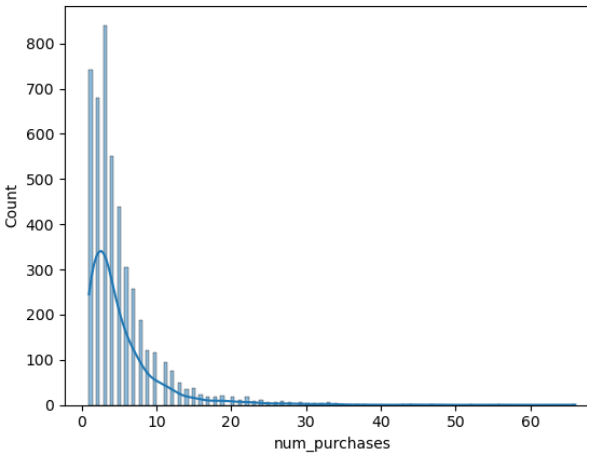


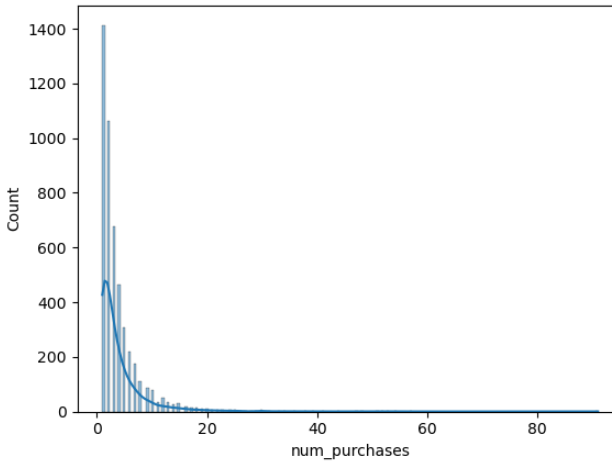

Иллюстрации 3,4. Распределения среднего чека тестовой и контрольной выборок.






Иллюстрации 5, 6. Распределения выручки тестовой и контрольной выборок.






Иллюстрации 7, 8. Распределения числа покупок тестовой и контрольной выборок.

Следует дополнительно изучить 16 день акции, так как количество покупок только в этот день в контрольной группе гораздо выше (Иллюстрация 9).

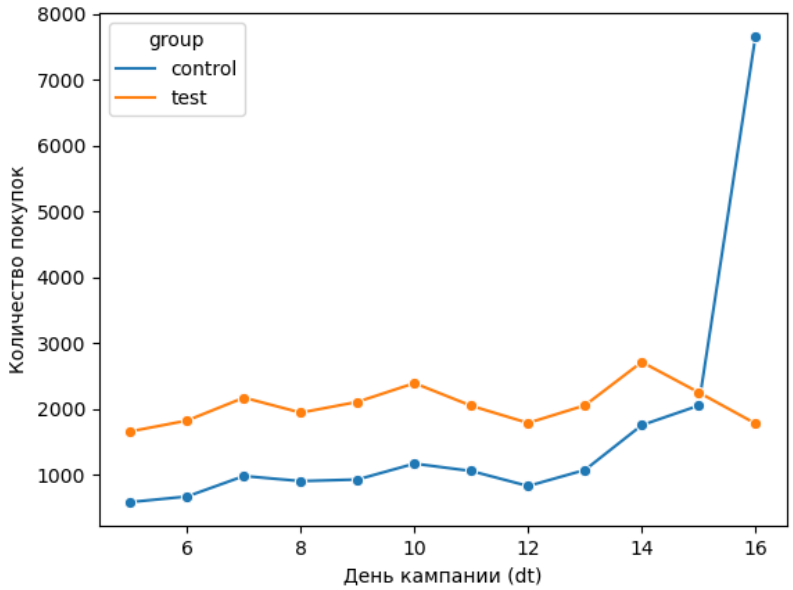



Иллюстрация 9. Динамика покупок по группам.

Таблица 3.

| Метрика | Тест | p-value |
|---|---|---|
| Средний чек | Манна-Уитни | 2.34e-10 |
| Выручка | Манна-Уитни | 4.89e-39 |
| Число покупок на клиента | Манна-Уитни | 4.93e-86 |
| Количество людей с повторными покупками | Z-тест | 3.06e-32 |

Различия в метриках в тестовой и контрольной группах статистически значимы.

Бизнес-рекомендация:

Не смотря на наличие скидки, средний чек вырос, это означает, что люди покупают больше товаров либо переходят в более дорогой сегмент. Число покупок и количество повторных покупок тоже больше, следовательно, акция привлекательно влияет на клиентов, провоцируя их покупать больше и возвращаться. Общая выручка выросла почти на 20% это прямое подтверждение того, что кампания эффективна. Перед тем как внедрить кампанию на всю аудиторию нужно:

- протестить данную кампанию на более широкую аудиторию (или другой сегмент), увеличить период и проверить стабилен ли положительный эффект;
- если средний чек вырос, то возможно стоит скорректировать размер скидки, сделать его меньше, тем самым снизить затраты, но сохранить заинтересованность клиентов в покупке;
- выделить категории покупателей, которые откликнулись на данную кампанию и запускать ее более персонализировано;
- обратить внимание на клиентов, которые давно не совершали покупки и на тех, кто покупает часто, но с невысоким чеком и попробовать реализовать акцию на них;
- перед масштабированием посчитать ROI (на данный момент это невозможно, так как не знаем затраты на рекламу).

Общая рекомендация – продолжать и масштабировать кампанию с дальнейшей сегментацией и оптимизацией скидки.

**Кластеризация**

Так как данные смешанные, были принято решение использовать для кластеризации метод K-Prototypes. Числовые признаки стандартизированы с помощью StandardScaler, категориальные признаки для t-SNE закодированы с помощью LabelEncoder.

Для кластеризации использовались следующие признаки: category_product, cost, base_sale, gender, age, personal_coef.

Количество кластеров определялось по методу локтя, методу силуэта (Иллюстрации 10, 11).

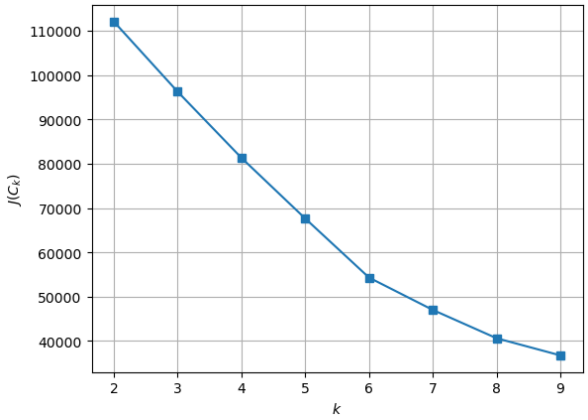



Иллюстрация 10. Метод локта

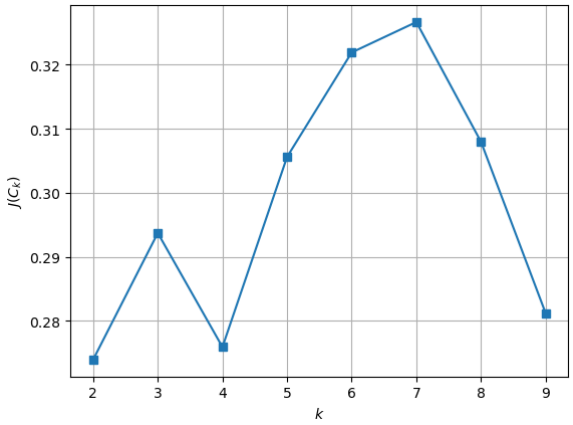



Иллюстрация 11. Метод силуэта

Согласно графикам количество кластеров равно 7.

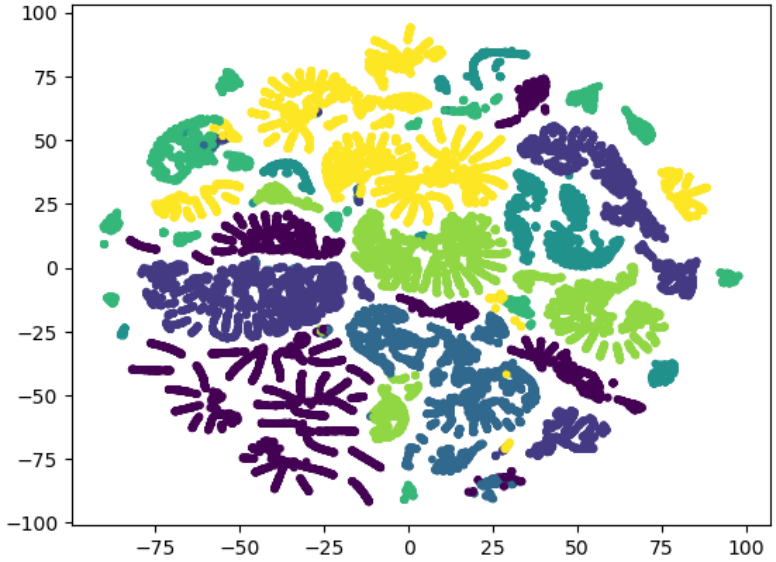
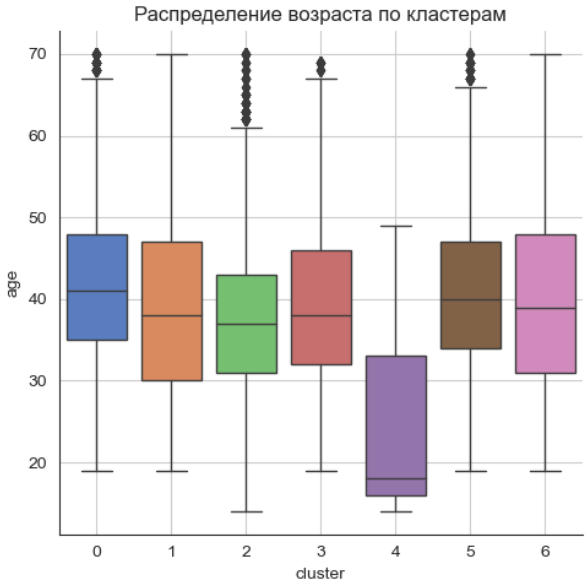
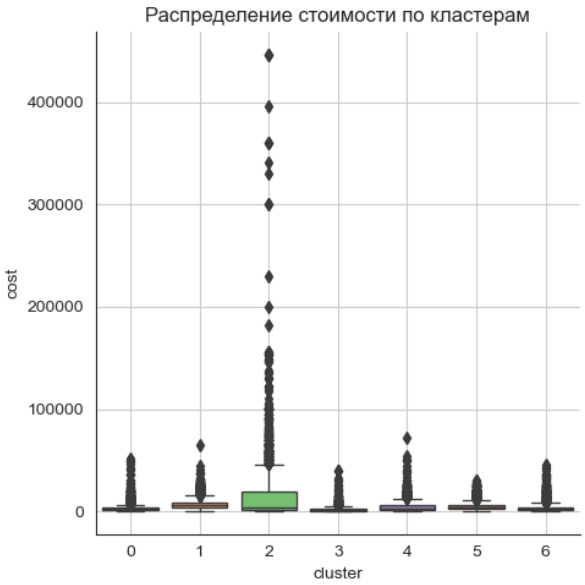



Иллюстрация 12. Визуализации кластеров (k = 7) методом t-SNE для кластеризации методом K-Prototypes.

Значения метрики Silhouette для данного метода = 0,44.

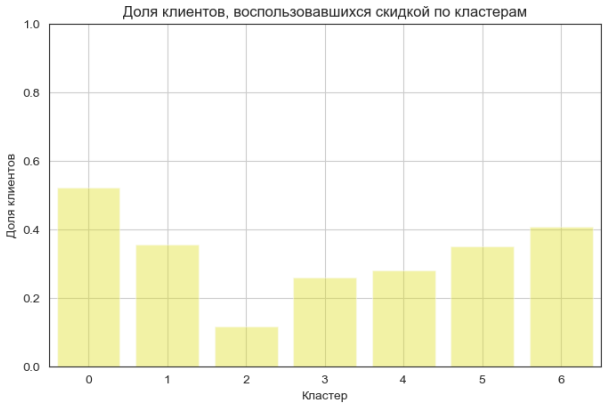
Визуализации распределения признаков по кластерам представлены на Иллюстрациях 13-17.

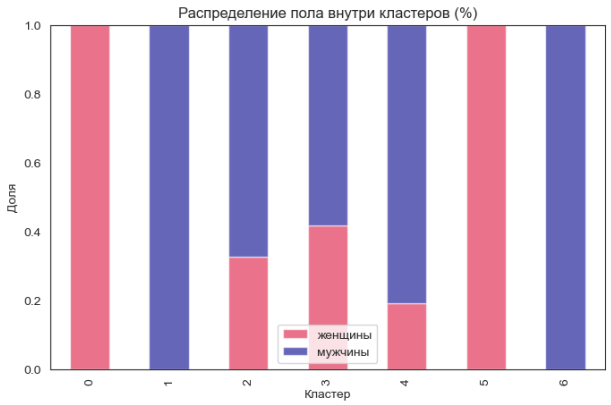



Иллюстрации 13, 14. Распределения возраста и стоимости товара по кластерам.

Стоит отметить, что самые молодые клиенты, чей возраст от 16 до 32 лет, находятся в 4 кластере. Товары с самой высокой стоимостью во 2 кластере.




Иллюстрации 15-16. Доли клиентов, воспользовавшихся скидкой и распределение пола клиентов по кластерам.

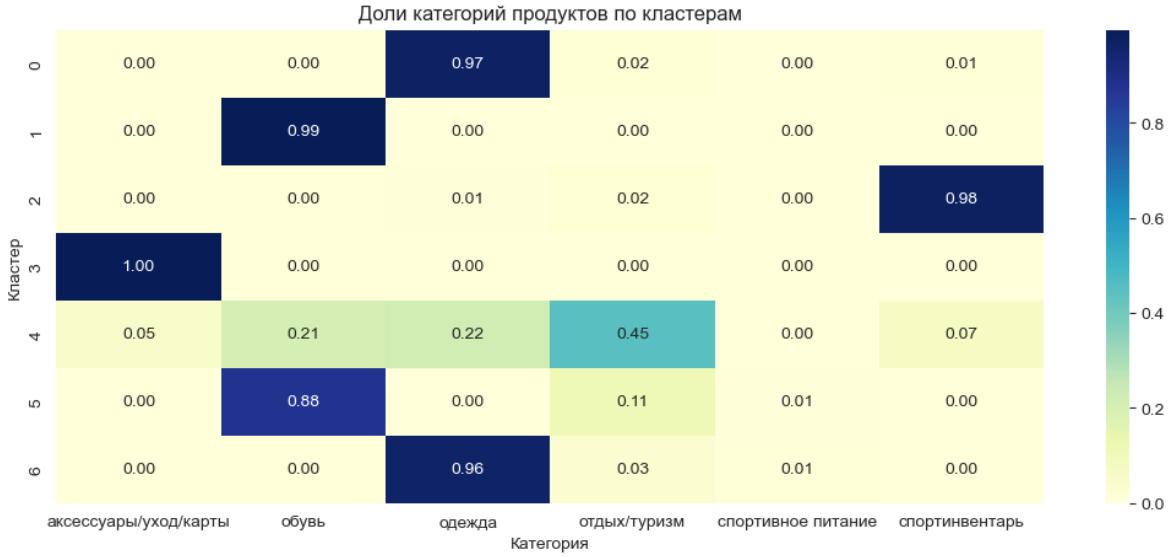



Иллюстрация 17. Категории продуктов, распределенные по кластерам.

Средние значения по кластерам приведены в таблице 4.

Таблица 4.

| Кластер | Стоимость | Наличие скидки | Пол | Возраст | Персональный коэффициент | Категория продуктов |
|---|---|---|---|---|---|---|
| 0 | 2819,98 | 0,52 | 0,00 | 42 | 0,51 | одежда |
| 1 | 7044,72 | 0,36 | 1,00 | 39 | 0,44 | обувь |
| 2 | 14704,27 | 0,12 | 0,67 | 38 | 0,46 | спортинвентарь |
| 3 | 2181,93 | 0,26 | 0,58 | 40 | 0,47 | аксессуары/уход/ карты |
| 4 | 4639,50 | 0,28 | 0,81 | 24 | 0,34 | отдых/туризм |
| 5 | 4876,66 | 0,35 | 0,00 | 41 | 0,51 | обувь |
| 6 | 3517,96 | 0,41 | 1,00 | 40 | 0,44 | одежда |

Описание кластеров:

0 - женская одежда со скидками;

1 - обувь с высоким среднем чеком для мужчин среднего возраста, реагирующих на скидки, но не сильно;

2 - дорогие спортивные товары для мужчин и женщин, не реагирующих на скидки (премиальные клиенты);

3 - товары с низким средним чеком из категории аксессуары/уход для мужчин и женщин, покупаемые без скидок;

4 - товары для активного отдыха и туризма, для молодежи, преимущественно для мужчин, не очень интересующихся скидками;

5 - обувь для женщин средней стоимости, скидки важны, но не критично;

6 - мужская одежда средней стоимости, со скидками.

*Рекомендации*:

0 - всевозможные скидочные кампании, персональные подборки из категории одежда, акции для увеличения количества товара в чеке (покупаешь 3 вещи - получаешь скидку);

1 - сезонные предложения на обувь, скидка на вторую пару обуви, персональные подборки;

2 - персональные предложения и пакеты, как для премиальных клиентов, персональная подборка товаров, не предлагать скидок так как они не нужны;

3 - предлагать покупать сопутствующий товар, предлагать дополнить корзину еще каким-то продуктом, предлагать товар по более дорогой цене или улучшенной версии, тем самым повышая средний чек;

4 - комплекты и сеты для активного отдыха, подборки товаров для путешествий, делать рекомендации на основе интересов;

5 - сезонные предложения на обувь, скидка на вторую пару, подборки «тренд сезона»;

6 - увеличивать средний чек, путем предложения акций 2+1, акции «собери себе образ», распродажи.

**Построение модели склонности к покупке.**

В качестве таргета использовались:

1. Категория товара;
2. Склонность клиента принять участие в акции.

Целями построения моделей были, в первом случае - предсказать какой товар с наибольшей вероятностью выберет клиент, во втором - насколько клиент склонен участвовать в акциях. Это позволит сегментировать клиентов и в дальнейшем делать персональные предложения.

1) Модели предсказания категории товара, которую выберет клиент представлены в Таблице 5. В моделях использовались следующие признаки:

- категориальные – colour, education, brand, закодированы с помощью OneHotEncoder;
- числовые - cost, product_sex, base_sale, gender, age, personal_coef стандартизированы с помощью StandardScaler.

Разделение на тестовую и контрольную выборку производилось с помощью кросс-валидации. Обучение проводилось на всех данных.

Таблица 5.

| Модель | Accuracy |
|---|---|
| Логистическая регрессия | 0,60 |
| Дерево решений | 0,86 |
| Случайный лес | 0,82 |
| Многослойный перцептрон | 0,76 |

Наилучшими моделями в данном случае оказались модели дерева решений и случайный лес. Это можно объяснить тем, что такие модели хорошо работают с нелинейными зависимостями и категориальными переменными и не требуют сложного тюнинга.

Данные по продуктам были достаточно размазаны и разнообразны. Признак category_product был создан вручную на основании текстового описания товара. Нельзя точно сказать, что категории были максимально разделены правильно, но это в любом случае упростило работу модели. Также нужно учесть, что один и тот же клиент может купить совершенно разные товары, поэтому сложно полностью предсказать его поведение, поэтому данные метрики являются неплохими для данной задачи.

2) Модели предсказания склонности клиента принять участие в акции представлены в Таблице 6. В этом случае для обучения моделей использовались данные кампании, проводимые на жителях города 1 134 (представляющие собой баннерную рекламу на билбордах: скидка всем каждое 15-е число месяца (15-й и 45-й день в нашем случае)). В качестве тестовой выборки использовались все дни до 45, в качестве контрольной оставшиеся дни, начиная с 45. В моделях использовались следующие признаки:

- категориальные – category_product, brand, закодированы с помощью OneHotEncoder;

- числовые - cost, base_sale, gender, age, personal_coef стандартизированы с помощью StandardScaler.

Для устранения дисбаланса классов использовалось SMOTE (категории 0 – 66%, категории 1 – 34%)

Таблица 6.

| Модель | F1-score | Precision | Recall | ROC-AUC | Accuracy |
|---|---|---|---|---|---|
| Логистическая регрессия | 0,44 | 0,39 | 0,51 | 0.52 | 0,51 |
| Дерево решений | 0,46 | 0,45 | 0,47 | 0,57 | 0,58 |
| Случайный лес | 0,43 | 0,46 | 0,41 | 0.59 | 0,60 |
| Многослойный перцептрон | 0,46 | 0,41 | 0,52 | 0,55 | 0,54 |
| CatBoost (Categorical Boosting) | 0,53 | 0,40 | 0,82 | 0,56 | 0,46 |

Низкое качество моделей скорее всего связано с небольшим дисбалансом данных, также метрика Accuracy не совсем информативна в данном случае. Возможно, модели недостаточно данных, чтобы предсказать будет ли участвовать клиент в акции или нет. Наилучшей моделью среди представленных является CatBoost, метрика Recall у нее достаточно высокая (ловит 82% целевых клиентов), также неплохое значение F1, но низкое значение Precision (среди предсказанных «1», только 40% настоящие). Точность ниже, но охватывает нужных клиентов. Если у нашей маркетинговой кампании не стоит вопрос в стоимости привлечении каждого клиента (а это так, так как это баннерная реклама, а не персональная рассылка), то можно использовать.