**Аналитический отчёт**

**Цель исследования**

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

**Источники данных**

Для работы были предоставлены следующие данные:

* shop\_database.db — база данных, содержащая три 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 (вам потребуется коэффициент personal\_coef);
* 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.
* personal\_data.csv.gz — утерянные данные о клиентах из таблицы personal\_data

Данные для проведения A/B-тестирования:

* ids\_first\_company\_positive.txt — пользователи, на которых проводилась первая маркетинговая кампания в период с 5-го по 16-й день. Эта кампания включала в себя предоставление персональной скидки 5 000 клиентов через email-рассылку.
* ids\_first\_company\_negative.txt— в этом файле были отобраны люди со схожими социально-демографическими признаками и покупками, которым скидку не предложили.

**1. Предобработка данных**

На начальном этапе были проведены комплексные действия по очистке и подготовке данных для дальнейшего анализа:

* **Очистка и проверка данных:** удалены дубликаты записей, проверены типы данных для всех столбцов.
* **Обработка пропусков:** числовые признаки заполнялись медианами по группе, категориальные — модой.

Результат — чистый и структурированный датафрейм, готовый к применению алгоритмов машинного обучения и статистических тестов.

**2. Бинарная классификация**

Целью данного этапа стало восстановление отсутствующих значений признака gender (пол клиента) на основе имеющихся данных. Для этой задачи использовался алгоритм Random Forest Classifier, показавший хорошую устойчивость к шуму и способность работать с большим числом признаков. Исходные данные были предварительно очищены: удалены строки с пропущенным значением gender для обучения, из признаков исключены идентификатор клиента (id) и сам целевой столбец (gender), а категориальные переменные были преобразованы в числовые с помощью One-Hot Encoding (pd.get\_dummies).

Пропущенные значения и бесконечности были заменены на медианные значения соответствующих столбцов, а признаки lbt\_coef и personal\_coef, потенциально вводящие искажения, были исключены.

Датасет был разделён на обучающую и тестовую выборки в пропорции 80/20. Модель обучалась на 200 деревьях с фиксированным random\_state=42 для воспроизводимости. После обучения была получена подробная метрика качества классификации (classification\_report), позволяющая оценить точность, полноту и F1-меру для каждого класса. По итогу вероятность составляла 66-67%.

Также был выполнен контроль пересечения ID между train и test, проверка на дубликаты, а затем проведено предсказание для всего набора данных, включая изначально неполные записи. Предсказанные значения были сохранены в новый столбец predicted\_gender, после чего недостающие значения в исходном gender были заполнены полученными прогнозами, формируя итоговый признак gender\_filled. Данный шаг позволил повысить полноту данных и обеспечить корректность дальнейшего анализа, в том числе для кластеризации и построения модели склонности клиента к покупке.

**3. A/B-тестирование**

В рамках третьего этапа работы было проведено A/B-тестирование для оценки влияния маркетинговой кампании на ключевые бизнес-метрики. Для этого в проект были загружены два списка пользователей — участников тестовой и контрольной групп. Данные идентификаторы клиентов были предварительно подготовлены в отдельных текстовых файлах:

* ids\_first\_company\_positive.txt — пользователи тестовой группы (получившие воздействие кампании)
* ids\_first\_company\_negative.txt — пользователи контрольной группы (не участвовавшие в кампании)

Файлы были считаны в датафреймы, после чего всем пользователям в основном наборе данных был присвоен признак group со значением "test" или "control". Далее был сформирован отдельный датафрейм df\_ab, содержащий только эти две группы.

**Метрики для анализа**:

В рамках теста были рассчитаны три ключевые метрики:

1. Конверсия (CR) — доля пользователей, совершивших хотя бы одну покупку.  
   Расчёт показал, что конверсия составила 100% в обеих группах. Это означает, что все пользователи, как в тестовой, так и в контрольной группе, совершили хотя бы одну покупку. Соответственно, кампания не смогла повлиять на этот показатель, так как он изначально был максимальным.
2. Средний чек (cost\_mean) — среднее значение переменной cost\_mean (средняя стоимость одной покупки). В тестовой группе средний чек оказался немного ниже, чем в контрольной(control: 6011.553524, test: 5446.618035). Это может указывать на то, что пользователи тестовой группы либо покупали более дешёвые товары, либо совершали больше мелких покупок.
3. Средний доход на пользователя (ARPU) — среднее значение cost\_sum (общая сумма покупок на одного клиента). Показатель оказался практически одинаковым в обеих группах(control: 62498.179431, test: 62892.592742). Это говорит о том, что, несмотря на небольшое снижение среднего чека в тестовой группе, общая сумма дохода от пользователей существенно не изменилась.

Была также проведена проверка гипотезы по ключевой метрике: среднему доходу на пользователя (ARPU). Для ARPU применялся двусторонний t-тест Стьюдента с поправкой на неравенство дисперсий между группами. Результаты показали, что статистическая значимость различий отсутствует (p-value > 0.05), что свидетельствует о том, что рассылка не оказала влияния на доход пользователей.

По результатам A/B-тестирования можно заключить, что проведённая кампания не оказала значимого влияния на конверсию и средний доход на пользователя. Единственное заметное изменение — небольшое снижение среднего чека в тестовой группе, которое, однако, не привело к падению общей выручки. С точки зрения бизнеса, тест не показал убедительных оснований для масштабирования данной кампании в текущем виде.

**4. Кластеризация**

Для сегментации клиентской базы была проведена кластеризация методом K-Means. На исходных данных предварительно была выполнена подготовка: числовые признаки, включая возраст, заполненные данные по полу (gender\_filled), коэффициенты лояльности и активности, метрики покупок (cost\_mean, cost\_sum, purchases\_count, sale\_ratio), были стандартизированы с помощью StandardScaler, а категориальные признаки (уровень образования и город) закодированы с применением OneHotEncoder. Для подбора оптимального количества кластеров использовались два критерия — метод локтя (анализ инерции) и коэффициент силуэта, что позволило выявить подходящее число кластеров равным трём.

После обработки данных и создания датафрейма для этого этапа модель была обучена на выборке из 100 000 клиентов, и каждому клиенту был присвоен номер кластера. Анализ средних значений признаков и структуры кластеров показал, что:

* **Кластер 0** включает наименьшее количество клиентов (855 человек), которые демонстрируют высокую покупательскую активность, самый высокий средний чек и низкую долю покупок со скидкой (23%). Эти клиенты относятся к категории «премиум» — они склонны к приобретению дорогих товаров (палатки, велосипеды, роликовые коньки) и менее чувствительны к скидкам.
* **Кластер 1** насчитывает около 43 тысяч клиентов. Средний чек здесь является самым низким (~4 тыс.), зато доля покупок со скидкой достигает 42%, что указывает на высокую чувствительность к акциям. Основные покупки связаны с повседневной и женской одеждой (футболки, брюки, легинсы, купальники).
* **Кластер 2** является крупнейшим (около 56 тысяч клиентов) и занимает промежуточную позицию по среднему чеку между кластерами 0 и 1. Доля покупок со скидкой составляет 31%, а ассортимент преимущественно включает мужскую одежду (шорты, футболки, толстовки).

Влияние скидок на покупательское поведение чётко различается между сегментами: премиальные клиенты (кластер 0) совершают покупки преимущественно без скидок, что делает маркетинговые акции для них менее актуальными, тогда как клиенты кластера 1 проявляют наибольшую ценовую чувствительность и являются ключевой целевой группой для стимулирования сбыта через скидочные предложения. Кластер 2 демонстрирует умеренную реакцию на скидки, и для него целесообразно тестировать сбалансированные маркетинговые подходы.

**5. Модель склонности клиента к покупке**

На данном этапе была построена модель для оценки склонности клиентов к совершению покупки на основе исторических данных о покупках и характеристиках клиентов. Для начала был выбран датафрейм df\_sample, после чего была произведена очистка данных: все строки с пропусками были удалены, а индексы пересчитаны. Это позволило исключить влияние неполных данных на обучение модели и повысить её стабильность.

Для дальнейшего анализа был выделен целевой сегмент клиентов, относящийся к конкретной стране и городу (country = 32, city = 1188). Целевая переменная purchase была сформирована бинарным способом: если значение base\_sale было больше нуля, то клиент считался покупателем. Этот подход позволяет формально разделить выборку на покупателей и непокупателей для задачи классификации.

Далее данные были разделены на числовые и категориальные признаки. К числовым отнесены характеристики, такие как возраст, количество покупок, средняя стоимость и сумма покупок, коэффициенты маркетинговых взаимодействий (lbt\_coef, ac\_coef, sm\_coef) и некоторые персональные коэффициенты. К категориальным признакам отнесены образование, продукт и цвет. Для корректной обработки категориальных признаков был выполнен их перевод в строковый формат.

Модель строилась с использованием пайплайна Pipeline, который включал два этапа: предварительная обработка данных и классификация. На этапе предварительной обработки числовые признаки стандартизировались с помощью StandardScaler, а категориальные признаки кодировались через OneHotEncoder с игнорированием неизвестных значений. Для классификации использовался Random Forest Classifier с 200 деревьями, что обеспечивает устойчивость модели и способность учитывать сложные зависимости между признаками.

Выборка была разделена на обучающую и тестовую в пропорции 80/20, после чего модель была обучена на тренировочных данных. После обучения была проведена проверка качества модели на тестовой выборке. Классификационный отчёт показал высокие показатели точности и полноты для обеих классов, а значение ROC-AUC составило 0.825. Это говорит о том, что модель хорошо различает покупателей и непокупателей, и её прогноз вероятности покупки можно использовать независимо от выбранного порога классификации.

В завершение прогнозируемые вероятности покупки были добавлены к исходному датафрейму df\_target, что позволяет оценивать склонность каждого клиента к покупке и использовать эти данные для дальнейшей персонализации маркетинговых кампаний или анализа клиентской базы.

В рамках проекта реализован полный аналитический цикл — от очистки и трансформации данных до построения интерпретируемых и высокоточных моделей. Результаты могут быть интегрированы в маркетинговую стратегию компании для повышения рентабельности и качества клиентской работы.