АНАЛИТИЧЕСКИЙ ОТЧЕТ ИТОГОВОГО ПРОЕКТА

Был проведен целостный многоэтапный анализ данных крупного магазина спортивных товаров, которые отражают покупки клиентов за 2 месяца и их социально-демографических признаки.

Перед нами стояли следующие **задачи**:

1. Очистить и подготовить данные к дальнейшей работе (отфильтровать данные, заполнить пропуски, обработать информацию о купленных товарах и их цветах, восстановить утерянные данные с помощью модели классификации.
2. Проанализировать эффективность маркетинговых кампаний, провести A/B-тестирование первой рекламной кампании и рассчитать основные метрики, сформулировать и обосновать бизнес-рекомендации.
3. Построить модель кластеризации на основе собранных данных, чтобы выявить какие товары предпочитают различные кластеры клиентов и насколько на покупку влияет наличие скидки.
4. Построить модель склонности клиента к покупке, основанную на данных о профилях клиентов, данных товаров и данных о прошлых маркетинговых кампаниях.

1. Предобработка данных

Входные данные - данные о покупках клиентов за два месяца. Для исключений утечки, вся информация была зашифрована. Данные хранятся в базе данных, которая содержит 3 таблицы:

- **personal_data** – данные о покупателях: ID клиентов, их пол, возраст, образование, страна и город проживания;
- часть информации о клиентах из таблицы **personal_data** была утеряна. Поэтому, помимо базы данных, был предоставлен файл с утерянными данными;
- **personal_data_coeffs** – данные с персональными коэффициентами клиентов, которые рассчитываются по некоторой закрытой схеме, нам потребуется коэффициент personal_coef;
- **purchases** – данные о покупках: ID покупателя, название товара, цвет, стоимость, гендерная принадлежность потенциальных покупателей товара, наличие скидки и дата покупки.

В датасетах **‘personal_data’** и **‘lost_personal_data’** были отфильтрованы данные, касающиеся только тех людей, которые относятся к стране с кодовым цифровым значением 32.

Для восстановления данных в ‘lost_personal_data’ в столбце ‘gender’ была построена модель ‘случайного леса’ на полных данных из 'personal_data' и после этого применена к утерянным данным. Корректность классификации оценивалась с помощью метрики «F-мера» и составила 0.53.

После восстановления датасеты с персональными данными были объединены.

В датасете ‘**purchase’** представлены данные о покупках товаров. В некоторых столбцах таблицы содержатся пропуски, информация о названии товаров неоднородна, а в данных о цветах попадаются наборы цветов, записанные через косую черту (/).

Данные о купленных товарах были приведены к общему виду, а данные о производителе выделены в отдельный признак ‘manufacturer’.

В столбце ‘colour’ пропущенные значения были удалены, т.к. нет возможности их восстановить, и их доля не является существенной. А также были приведены названия цветов к общему виду.

В колонке ‘product_sex’ отражается гендерная принадлежность потенциальных покупателей товара. Строки, которые не заполнены по признаку ‘product_sex’ были заполнены значениями ‘gender’ из таблицы ‘personal_data_full’, а также по полу купленных товаров. Оставшиеся незаполненные строки были удалены.

После всех проведенных операций над данными они были объединены в общий датасет, отражающий произведенные покупки и социально-демографическую информацию о каждом клиенте.

2. Анализ проведения маркетинговых кампаний

Первая маркетинговая кампания проводилась в период с 5-го по 16-й день и включала в себя предоставление персональной скидки 5 000 клиентов через email-рассылку.

ID участвовавших в ней пользователей содержались в отдельном файле ids_first_company_positive.txt. Для проведения A/B-тестирования, помимо людей, которым предлагалась персональная скидка, были отобраны люди со схожими социально-демографическими признаками и покупками, которым скидку не предложили. ID этих клиентов лежали в аналогичном файле ids_first_company_negative.txt.

Для проведения A/B-тестирования были извлечены ID контрольной и тестовой групп из предоставленных файлов.

Была выдвинута гипотеза о том, что выручка среди тех, кому была предоставлена персональная скидка, равна выручке среди тех, кому эту скидку не предоставили.

После проведения статистического t-теста было обнаружено, что p-value очень мало, следовательно стоит принять альтернативную гипотезу, что выручка среди клиентов, которым предоставили скидку через email-рассылку выше, чем среди тех, кому ее не предоставили. На рисунке 2.1 изображено изменение выручки в обеих группах во время проведения маркетинговой кампании.

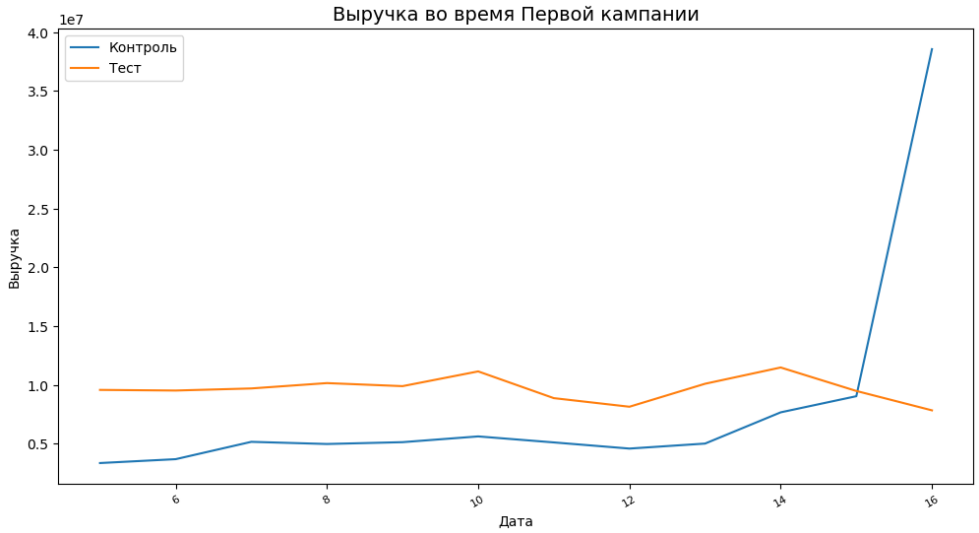

Рисунок 2.1 - Изменение выручки в обеих группах во время проведения маркетинговой кампании

Далее были рассчитаны основные метрики:

Конверсия в покупку в контрольной группе оказалась выше, чем в тестовой на 0,15 %.

Сравнение конверсий отражено на графике 2.2. Похожие значения конверсии могут быть по тому, что в ходе дальнейшего анализа было выявлено, что наличие скидки не влияет на свершение покупки.

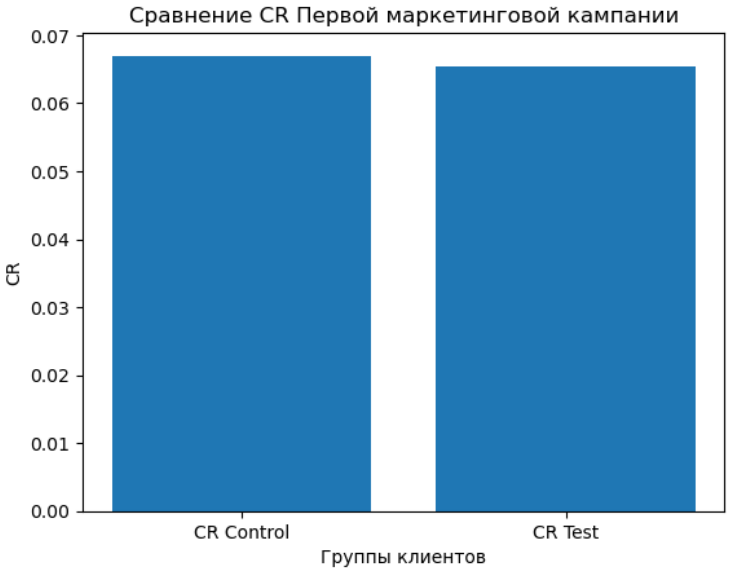

Рисунок 2.2 - Сравнение конверсий в покупку

Средний чек, среди тех, кому персональную скидку не предоставили, оказался выше на 336,96 ден. ед. Сравнение средних чеков размещено на графике 2.3.

Рисунок 2.3 - Сравнение средних чеков

На основании результатов анализа A/B-тестирования, можно сделать следующие выводы и дать бизнес-рекомендации:

Выручка: Полученное значение t-теста указывает на то, что выручка среди клиентов, которым предоставили скидку через email-рассылку, выше, чем среди тех, кому скидку не предоставили. Это говорит о том, что персональные скидки через email-рассылку могут быть эффективным инструментом для увеличения выручки.

Конверсия в покупку: Изменение конверсии в покупку после проведения теста на уровне -0.15 п.п. говорит о том, что маркетинговая кампания не привела к значительному увеличению числа покупателей.

Размер среднего чека: Изменение размера среднего чека после проведения теста на уровне -336.96 указывает на то, что средний чек уменьшился. Возможно, это связано с тем, что клиенты стали покупать более дешевые товары или в меньшем количестве из-за предоставленных скидок.

Бизнес-рекомендации:

- Продолжать использовать персональные скидки через email-рассылку для увеличения выручки.
- Провести дополнительные исследования для повышения конверсии в покупку.
- Оптимизировать товарный ассортимент и ценовую политику для увеличения размера среднего чека.

Вторая маркетинговая кампания проводилась на жителях города 1 134 и представляла собой баннерную рекламу на билбордах: скидка всем каждое 15-е число месяца.

После ее проведения были рассчитаны основные метрики отдельно для 15 и 45 дня покупок.

По сравнению с 15 днем покупок на 45 день произошло снижение выручки на 14 946 492 ден. ед. При этом размер среднего чека вырос на 147,68 ден. ед. с учетом того, что количество покупателей снизилось практически в 2 раза.

Исходя из результатов проведенной кампании баннерной рекламы на билбордах с предложением скидки каждое 15-е число месяца для жителей города 1 134, можно сделать следующие выводы и дать бизнес-рекомендации:

Выручка: Уменьшение размера выручки на 14 946 492 единиц говорит о том, что кампания не привела к увеличению доходов компании. Возможно, предложение скидки каждое 15-е число месяца не оказалось достаточно привлекательным для целевой аудитории. Сравнение выручек в разные дни скидок изображено на графике 2.4.

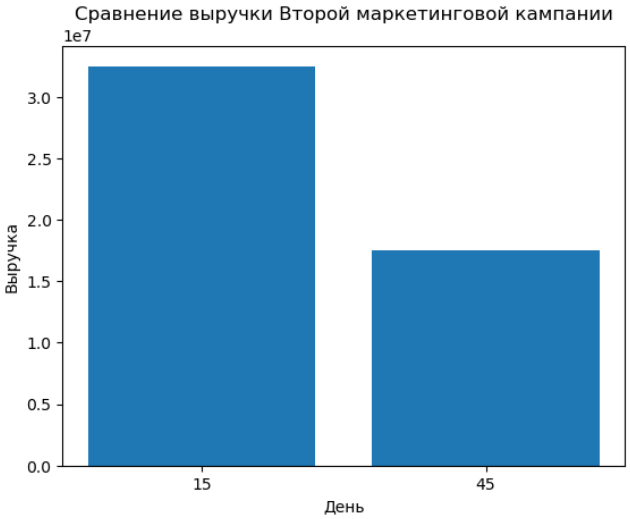

Рисунок 2.4 - Сравнение выручки

Размер среднего чека: Увеличение размера среднего чека на 147.68 указывает на то, что клиенты стали делать более крупные покупки в рамках кампании. Это может быть связано с привлекательностью предложения скидки и стимулированием клиентов делать дополнительные покупки. Размеры средних чеков отражены на графике 2.5.

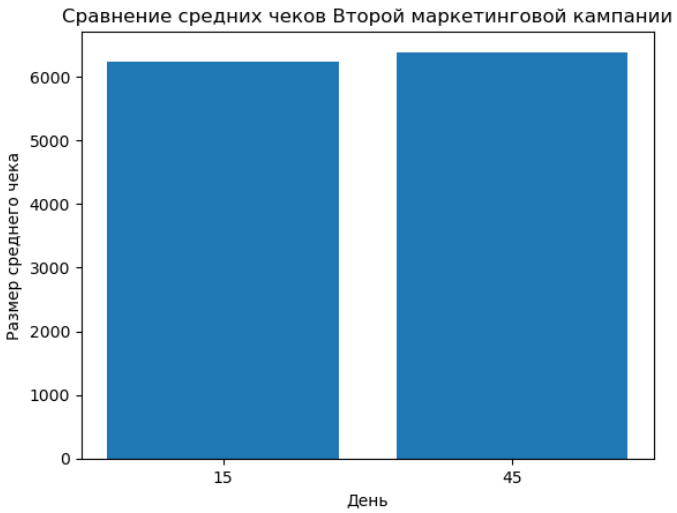

Рисунок 2.5 - Сравнение средних чеков

Количество покупателей: Снижение количества покупателей на 1467 человек может быть связано с неэффективностью баннерной рекламы на билбордах в привлечении новых клиентов. На рисунке 2.6 изображено сравнение количества чеков.

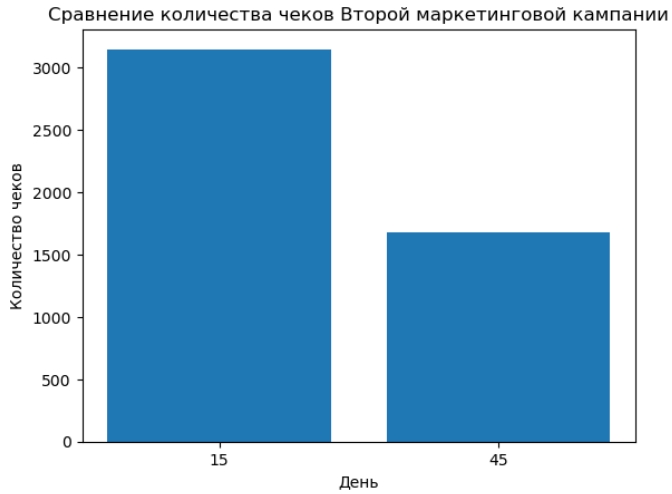

Рисунок 2.6 - Сравнение количества покупателей

Бизнес-рекомендации:

- Провести анализ предпочтений и потребностей жителей города для разработки более целевых маркетинговых стратегий.
- Разнообразить маркетинговые каналы и использовать комбинацию онлайн и офлайн инструментов для привлечения новых клиентов.
- Продолжать стимулировать клиентов на совершение дополнительных покупок через предложения скидок и акций.

3. Кластеризация

Необходимо было выполнить кластеризацию аудитории и выявить какие товары предпочитают различные кластеры клиентов и насколько на покупку влияет наличие скидки.

Кластеризацию начали с подготовки данных – сначала закодировали категориальные признаки с помощью LabelEncoder, затем стандартизировали весь датасет.

С помощью метода снижения размерности t-SNE визуализировали группы данных.

Определили с помощью ‘метода локтя’ оптимальное количество кластеров и применили вероятностный тип кластеризации для разделения данных на 5 кластеров (рисунок 3.1).

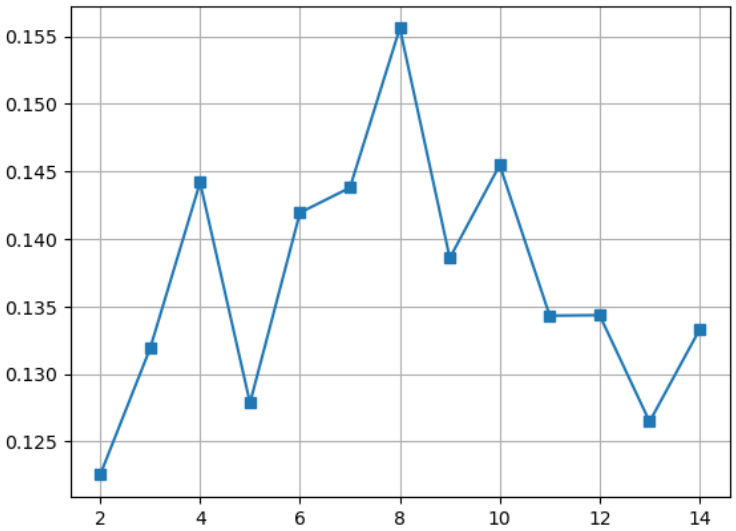

Рисунок 3.1 – Метод локтя

Разделение данных с помощью метода KPrototypes проиллюстрировано на рисунке 3.2.

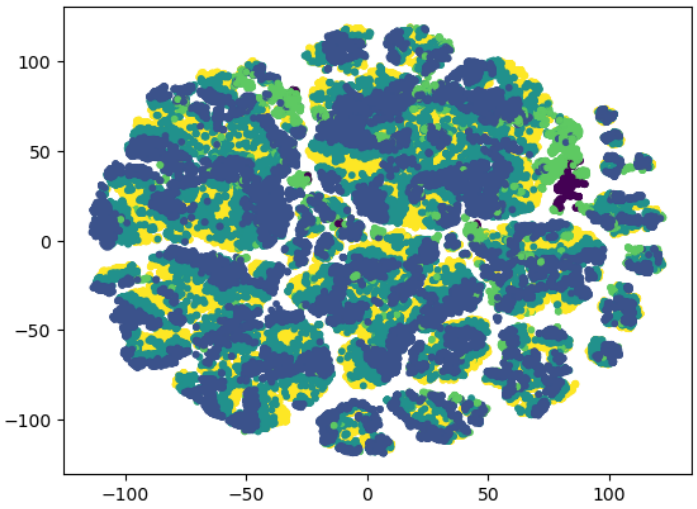

Рисунок 3.2 – Кластеризация данных методом KPrototypes

Разделение данных с помощью метода KMeans проиллюстрировано на рисунке 3.3.

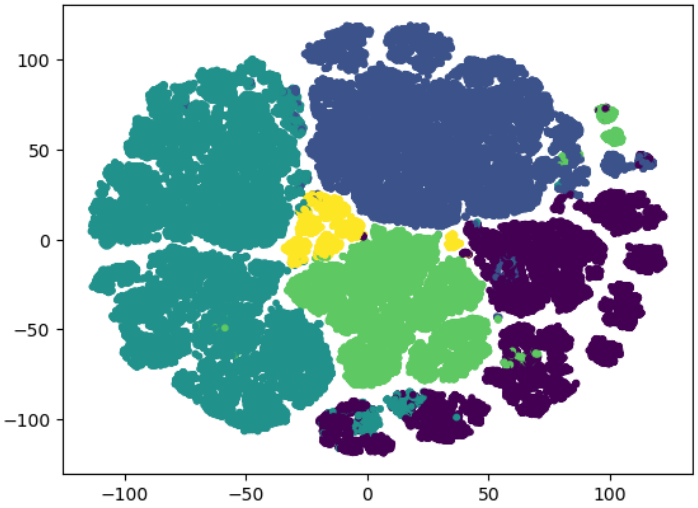

Рисунок 3.3 – Кластеризация данных методом KMeans

Для дальнейшей работы с данными будем использовать кластеризацию метода KMeans, т.к. разделение на кластеры имеет более четкие границы. Оптимальное количество кластеров определили выше с помощью метода локтя.

Создадим график среднего возраста в разных кластерах. И по графику видно, что, возраст во всех кластерах разный, и уже на основании этого признака можно предположить о разбиении на кластеры на основании и этого признака.

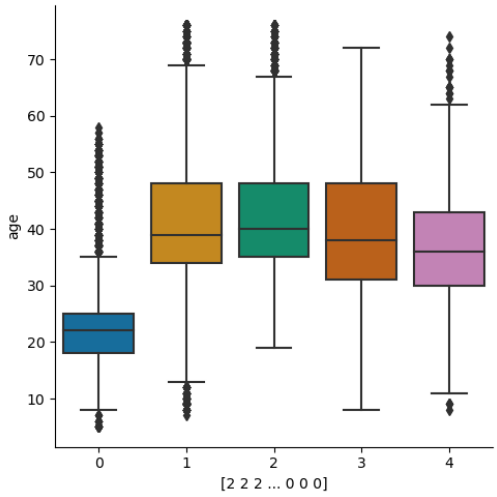

Рисунок 3.5 - Боксплоты возраста клиентов в разных кластерах

Глядя на визуальное представление профиля усредненного объекта для каждого кластера, можно сказать, что для всех кластеров первые 5 признаков почти полностью совпадают, а потом в кластерах 0, 2, 3 и 4 на разных отрезах идет разбиение. Провал на 12 признаке в кластере 3 связан с тем, что этот признак имеет большое влияние на разделение, именно в кластере 3 средний возраст составляет 15 лет, и является самым низким. Графическое представление изображено на рисунке 3.6.

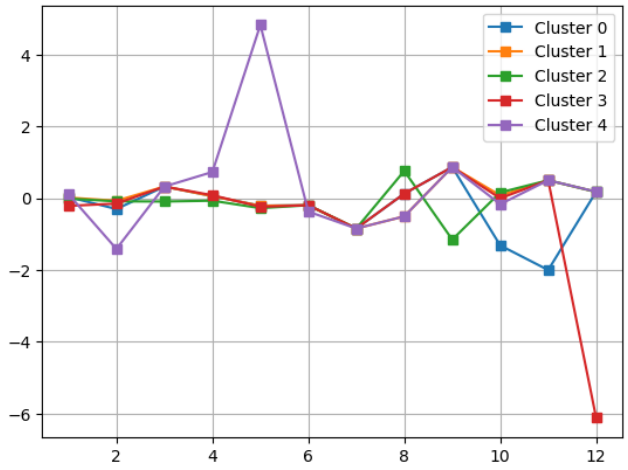

Рисунок 3.6 - Визуализация профиля усредненного объекта кластеров

После проведения кластеризации данные датасета были распределены по кластерам.

В кластер 0 попали 2460 покупателей, преимущественно мужчины только с высшим образованием и средним возрастом около 26 лет. С 2 персональными коэффициентами - 0.4304, 0,5072 и 0.4688. В кластере 0 покупатели предпочитают товары без скидки (их почти в 2 раза больше), среди часто покупаемых товаров – кроссовки, кеды, майки, футболки, куртки, ветровки.

В кластер 1 попали 7919 покупателей, преимущественно мужчины в большей части со средним образованием и средним возрастом около 41 лет. С 3 персональными коэффициентами - 0.4688 и 0.5584. В кластере 1 покупатели предпочитают товары без скидки, среди часто покупаемых товаров – кроссовки, кеды, куртки, различное спортивное оборудование.

В кластер 2 попали 413 покупателей, преимущественно мужчины (с небольшим перевесом) в большей части со средним образованием и средним возрастом около 38 лет. Со всеми персональными коэффициентами. В кластере 2 покупатели почти на одинаковом уровне приобретают товары на скидке и без нее, среди часто покупаемых товаров – товары категории обувь (кроссовки, сабо, сандалии), футболки, худи, но и присутствуют товары из категории ремкомплекты, лодки, палатки.

В кластер 3 попали 897 покупателей, преимущественно молодые люди (перевес мужчин по количеству в 2 раза) в большей части с высшим образованием и средним возрастом около 15 лет. Со 4 персональными коэффициентами. В кластере 3 покупатели чаще приобретают товары без скидки, среди часто покупаемых товаров – товары категории обувь (кроссовки, кеды, полуботинки, шлепанцы), футболки, леггинсы, пальто.

В кластер 4 попали 6594 покупателя, только женщины преимущественно со средним образованием и средним возрастом около 42 лет. Со 4 персональными коэффициентами. В кластере 4 покупатели почти на одинаковом уровне приобретают товары на скидке и без нее, среди часто покупаемых товаров – товары для детей: обувь, одежда, спортивный инвентарь и сопутствующие товары.

Исходя из проведенного анализа можно сделать вывод о том, что кластеризация клиентов была проведена на основании возраста, т.к. в разных категориях он разный. В первых 4 кластерах покупатели были преимущественно мужчины, в последнем кластере покупатели – только женщины.

В кластерах 0, 1 и 3 покупатели предпочитают приобретать товары без скидки, кластерах 2 и 4 покупатели приобретают в равной степени акционные и неакционные товары.

Отдельно выделяющихся категорий товаров в кластерах нет, товары распределены примерно одинаково – лидеры продаж – обувь и одежда.

4. Построение модели склонности клиента к покупке

Для определения целевой переменной - повторность покупки, посчитаем для каждого клиента количество покупок за весь период.

Для построения модели склонности клиента к покупке было выбрано 2 метода: случайный лес и логистическая регрессия.

Точность модели «Случайный лес»:

	precision	recall	f1-score	support
0	0.61	0.04	0.07	2340
1	0.85	1.00	0.92	12733
accuracy			**0.85**	15073
macro avg	0.73	0.52	0.49	15073
weighted avg	0.81	0.85	0.78	15073

Точность модели «Логистическая регрессия»:

	precision	recall	f1-score	support
0	0.33	0.00	0.00	11724
1	0.84	1.00	0.92	63637
accuracy			**0.84**	75361
macro avg	0.58	0.50	0.46	75361
weighted avg	0.76	0.84	0.77	75361

Исходя из полученных результатов оценки качества моделей, можно сделать вывод о том, что более точной в прогнозировании повторной покупки оказалась модель случайного леса.

Выводы:

В первой части работы было ознакомление и работа над улучшением качества данных, а также подготовка данных к дальнейшему анализу:

- изучена общая информация;
- отфильтрованы, восстановлены и приведены к общему виду необходимые данные;
- сведены в общую таблицу.

Во второй части работы был осуществлен анализ проведенных маркетинговых исследований.

В первой маркетинговой кампании, проходившей с 5 по 16 день, было реализовано A/B-тестирование, тестовой группе были высланы на e-mail персональные скидки. После проведения статистического теста был сделан вывод о том, что на уровне значимости 95% выручка среди тех, кому была предоставлена персональная скидка, выше выручки тех, кому эту скидку не предоставили.

После этого рассчитаны основные метрики и сформулированы рекомендации:

- Продолжать использовать персональные скидки через email-рассылку для увеличения выручки.
- Провести дополнительные исследования для повышения конверсии в покупку.
- Оптимизировать товарный ассортимент и ценовую политику для увеличения размера среднего чека.

По результатам проведения второй маркетинговой кампании были сформулированы следующие выводы:

- Следует провести анализ предпочтений и потребностей жителей города для разработки более целевых маркетинговых стратегий.
- Разнообразить маркетинговые каналы и использовать комбинацию онлайн и офлайн инструментов для привлечения новых клиентов.
- Продолжать стимулировать клиентов на совершение дополнительных покупок через предложения скидок и акций.

В третьей части была произведена кластеризация данных двумя методами: KPrototypes и KMeans. В качестве основного был выбран метод средних. Данные были разделены на 5 кластеров, основным признаком разделения является возраст. Было выявлено, что наличие скидок на товары не влияет на их приобретение. Значимой разницы в кластерах в приобретении товаров выявлено не было.

В последней части работы было необходимо построить модель склонности клиента к покупке. В качестве целевой переменной была выбрана повторность покупки. Для моделирования склонности клиента было выбрано 2 метода: случайный лес и логистическая регрессия. После получения данных о точности этих моделей, был сделан вывод, что модель «Случайного леса» оказалась более точной в прогнозировании повторности покупки.