Аналитический отчет
Об эффективности проведённых маркетинговых кампаний

Этап 1. Предобработка данных

1. Сбор данных и обогащение новыми признаками

В соответствии с требованиями вышестоящего руководства требуется создать витрину данных, которая в обобщённой форме будет отражать данные о покупках клиентов и их социально-демографических признаках, позволяющих проанализировать эффективность уже проведённых ранее маркетинговых кампаний и выявить факторы, способные повысить продажи.

В рамках оставленной задачи были выполнены следующие действия:

а. Создана единая витрина данных из разрозненных источников информации – несколько файлов с информацией о клиентах и их взаимодействии объединены в единый набор данных.

б. Добавлен признак *`personal_coef`*, содержащий персональную информацию о клиенте.

в. Устранены пропуски в потерянных данных о поле клиента, путем обучения модели GradientBoostingClassifier на имеющихся полных данных.

г. Проведена очистка данных о видах и группах товаров, приобретаемых клиентами.

Сформирована единая витрина данных, позволяющая провести анализ эффективности проведенной маркетинговой компании.

2. Изучение данных и устранение имеющихся недостатков

Изучение данных в полученном фрейме проводилось по трем направлениям: оценка правильности, полноты и валидности данных, а также в разрезе типов данных (категориальные и числовые признаки). Получены следующие результаты по направлениям исследования:

А. Сведения о клиентах.

1. Правильность данных. На основании информации о типах данных, содержащихся в датасете можно сделать вывод об их соответствии описываемым показателям и правильность данных можно оценить как **высокую.**

2. **Полнота данных**. В признаке *`gender`* содержаться пропуски. Их количество достаточно велико (15%), поэтому не представляется возможным просто удалить нулевые строки. Необходимо продумать стратегию заполнения пропусков. Предположительно можно использовать модель машинного обучения (бинарной классификации) для предсказания пола.

3. Валидность данных. При изучении признака *‘age’* были выявлены незначительные аномалии (рисунок 1).

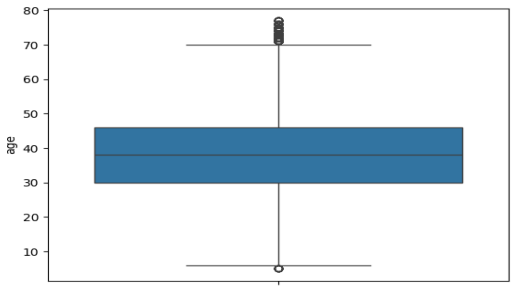

Рисунок 1 – Анализ признака *‘age’*

Из представленной диаграммы видно, что возраст клиентов находится в пределах от 30 лет до 45 лет. Так же есть данные, которые находятся выше верхней границы распределения возрасов клиентов. Однако, эти данные можно отнести скорее к аномалиям, чем к выбросам, так как они находятся в промежутке от 70 лет до 80 лет, что вполне может быть (пожилые клиенты, ведущие активный образ жизни).

Для изучения персональных коэффициентов клиентов так же была построена диаграмма распределения (рисунок 2).

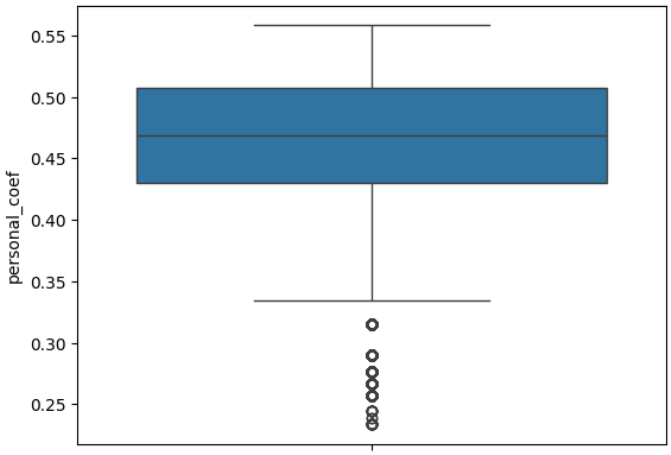

Рисунок 2 – Анализ признака ‘*personal_coef*

Персональные коэффициенты клиентов находятся в промежутке значений от 0,43 до 0,52. При этом наблюдаются значения, находящиеся ниже минмальной границы распределения. Эти значения так же можно отнести к аномалиям - данные значения коэффициентов могут быть присвоены клиентам, совершающим покупки на минимальные суммы и не слишком часто.

В целом, данные, содержащиеся в датасете `*df_clients*`, являются валидными. Дополнительных действий не требуется.

Б. Сведения о покупках

1. Правильность данных. На основании информации о типах данных, содержащихся в датаете можно сделат вывод об их соответсвии описываемым показателям и правильность данных можно оценить как **высокую.**

2. Полнота данных. В признаке `*colour*` содержаться пропуски. Их количество достаточно велико (15%), поэтому не представляется возможным просто удалить нулевые строки.

Также пропуски содержаться в признаке `*product_sex*` - пропущено около 40% значений. Так как отсутствующие данные достаточно разнородны по своей природе, то требуется их детальное изучения для определения стратегии заполнения пропусков.

3. Валидность данных. Изучение данных по стоимости, базовой скидке и длительности между покупками клиентов на предмет выбросов и аномальности данных показало следующее.

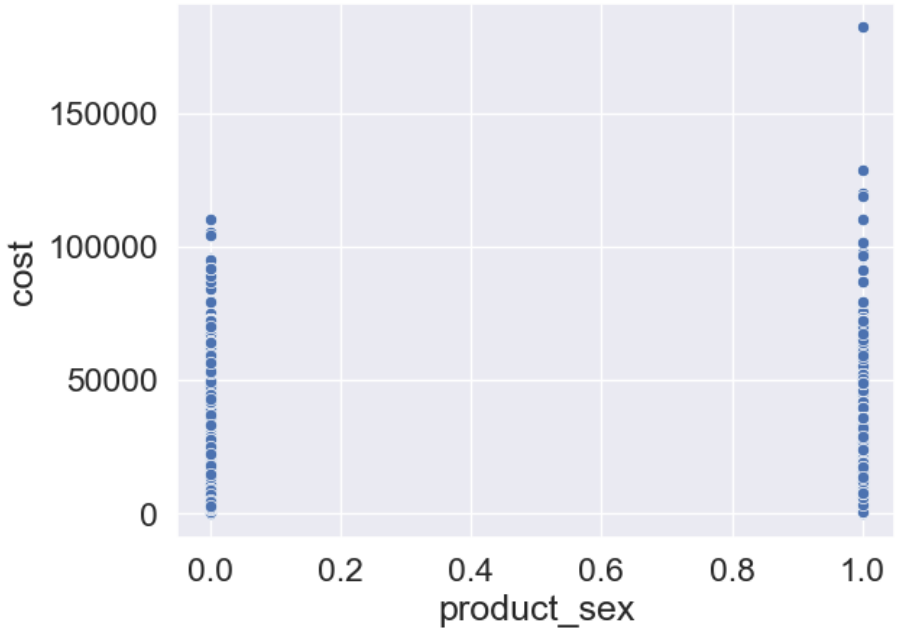

Рисунок 3 – Изучение распределения стоимости покупок в разрезе гендерного признака

Полученные данные показывают, что выбросов и аномалий в стоимостях товаров нет. На рисунке 4 представлено распределение скидок клиентов

Рисунок 4 – Распределение скидок

Диаграмма показывает, что все скидки клиентов находятся в пределах от 0% до 100%, что в целом соответствует реальности.

Диаграмма 4 демонстрирует распределение длительности (в днях) между покупками клиентов (рисунок 5).

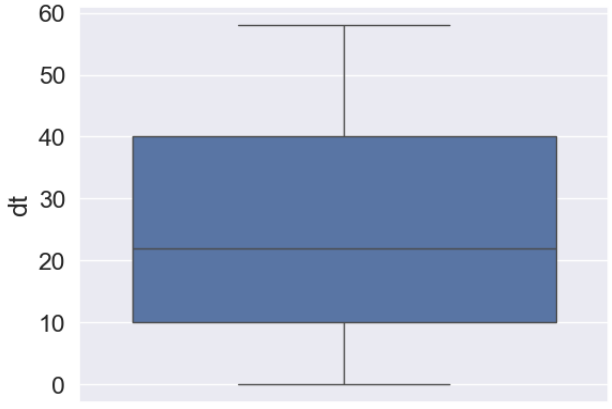

Рисунок 5 – Диапазон длительности между покупками клиентов

Из диаграммы видно, что длительность между покупками находится в пределах от 10 дней до 40 дней. Выбросов и аномалий в данном признаке не наблюдается.

Изучение признака `*colour*` показало, что количество значений в признаке `colour` достаточно велико - 1693 шт. Вывлены следующие проблемы:

1. Присутствуют одинаковые по смыслу значения, но с разным написанием. Например, `чёрный` - `черный`. Возможное решение - заменить символ `ё` на `е`.

2. Присутствуют значения признаков с несколькими цветами. Например, `черный/красный/лаймовый`. Возможное решение - можно предположить, что первый цвет в описании является доминирующим, поэтому оставляем его.

3. Присутствуют нулевые значения признака. Возможное решение - так как их количество достаточно велико, заменим их на значение `другой`.

Изучение признака `product_sex` показало, что нет возможности достоверно определить категорию товара (мужской или женский) из его описания. Возможное решение - ввести новое значение признака - унисекс и закодировать его значением 2.

Изучение признака `product` показало, что число значений в признаке достаточно велико - 23145 шт. Можно предположить, что такое количество обусловлено наличием детального описания каждого товара (указание бренда, материала, типа, размера, модели...). Возможное решение - понизить количество товаров, объеденив их в товарные группы (оставить в названии товара только русский текст и привести все к единому регистру).

В остальном, данные в `df_purchases` являются валидными. Дополнительых действий не требуется.

Для устранения выявленных недостатков с данными были проведены следующие манипуляции:

а. Для заполнения пропусков в признаке `*gender*` было обучено несколько моделей бинарной классификации. В результате для моделирования отсутствующих значений была выбрана модель классификации GradientBoostingClassifier. Метрики качества выбранной модели следующие:

- F-мера 0,7279
- лучшие гипперпараметры модели: 'subsample': 0.7,
  'n_estimators': 500,
  'min_samples_split': 10,
  'min_samples_leaf': 9,
  'max_features': 'sqrt',
  'max_depth': 14,
  'loss': 'log_loss',
  'learning_rate': 0.1,
  'criterion': 'squared_error'

На рисунке 6 представлена матрица ошибок обученной модели.

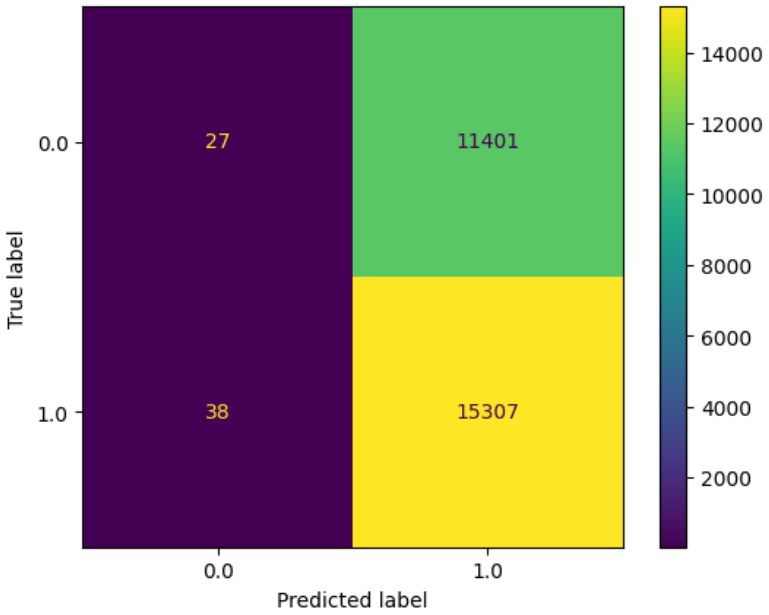

Рисунок 6 – Матрица ошибок

Основные метрики модели следующие:

Точность (precision) – 0,9975

Полнота (recall) – 0,5713

F-мера – 0,7279

б. Для устранения незаполненных данных в признаке `*colour*` были проведены следующие корректирующие действия:

- произведена замена символов `ё` на `е`;

- значение признака, состоящее из нескольких цветов заменено на значение цвета, стоящего первым в описании признака;

- отсутствующие значения в признаке заменены на значение `другой`.

В результате проведенных манипуляций количество значений признака `colour` сократилось в 5 раз – с 1693 шт до 315 шт.

в. Для заполнения отсутствующих значений в признаке `product_sex` введено новое значение признака - `унисекс` и закодировано значением 2.

г. Для снижения размерности признака `product` применено объедение товаров в товарные группы, путем оставления в названии товара только русского текста и приведения наименований к единому регистру.

В результате удалось снизить количество значений признака с 23145 шт. до 3610 шт. - более чем в 6 раз.

Этап 2. Проведение A/B-тестирования для оценки эффективности маркетинговой кампании

1. Метрики и гипотезы для оценки эффективности проведенной маркетинговой компании.

Для проведения A/B-тестирования в качестве метрик для оценки эффективности маркетинговой компании предлагается испольовать следующие метрики:

1. Средняя выручка на одного покупателя

Выдвигаемые гипотезы:

Нулевая гипотеза: После проведения маркетинговой компании среняя выручка на одного покуптеля в тестовой группе изменилась незначительно (нет статистической значмости).

Альтернативная гипотеза: После проведения маркетинговой компании среняя выручка на одного покуптеля в тестовой группе значительно изменилась (есть статистическая значмость).

2. Средний чек покупателя

Выдвигаемые гипотезы:

Нулевая гипотеза: После проведения маркетинговой компании сумма среднего чека на одного покуптеля в тестовой группе изменилась незначительно (нет статистической значмости).

Альтернативная гипотеза: После проведения маркетинговой компании сумма среднего чека на одного покуптеля в тестовой группе значительно изменилась (есть статистическая значмость).

3. Срелнее количество покупок

Выдвигаемые гипотезы:

Нулевая гипотеза: После проведения маркетинговой компании сренее количество покупок одного покуптеля в тестовой группе изменилась незначительно (нет статистической значмости).

Альтернативная гипотеза: После проведения маркетинговой компании сренее количество покупок одного покуптеля в тестовой группе значительно изменилась (есть статистическая значмость).

2. Проверка выдвинутых гипотез

При проведении проверки выборок на нормальность распределения было установлено, что распределение в выборках отличается от нормального, а выборки являются независимыми (разделение произошло случайным образом). Для проверки выдвинутых гипотез использовался критерий `Манна-Уитни`.

1. При проверке первой гипотезы было установлено, что наблюдается статистически значимая разница в средней выручке на покупателя в тестовой и контрольной группе (рисунок 7).

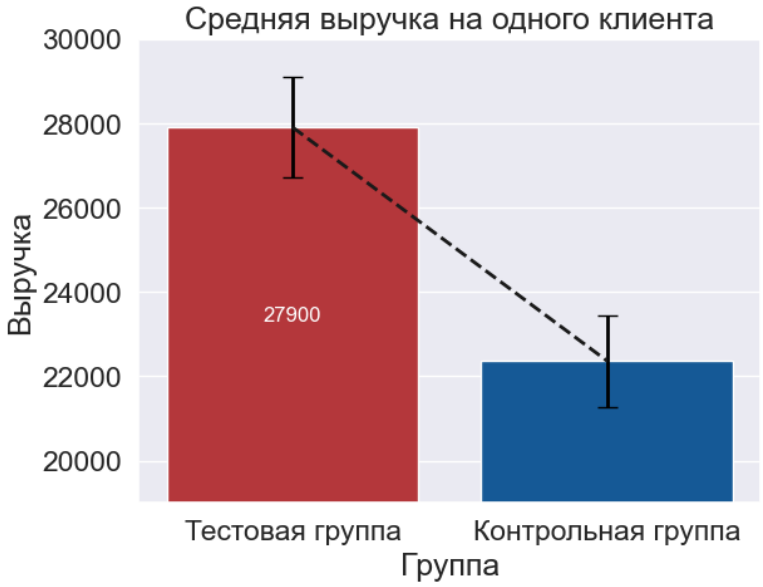

Рисунок 7 – Различия в средней выручке в тесовой и контрольной группе

2. При проверке второй гипотезы было установлено, что наблюдается статистически значимая разница в среднем чеке покупателя в тестовой и контрольной группе (рисунок 8).

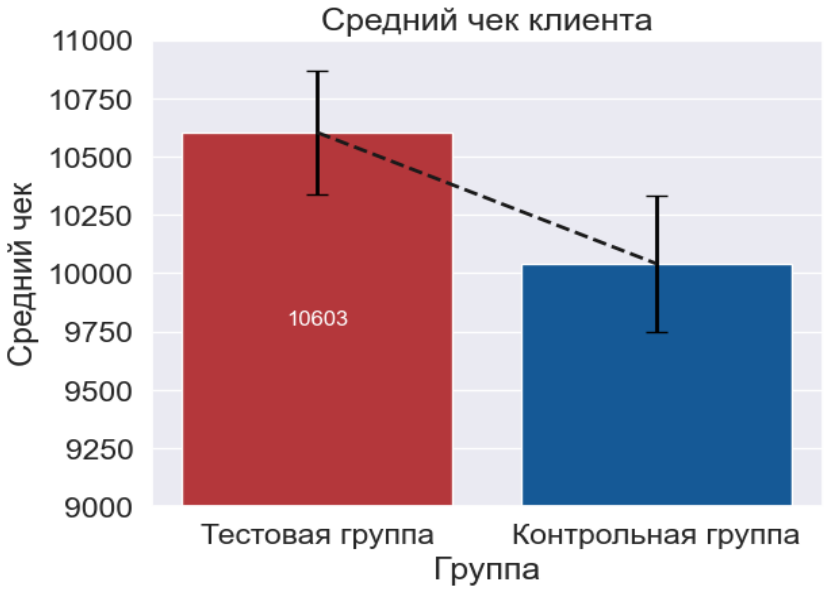

Рисунок 8 – Различия в среднем чеке в тестовой и контрольной группе

3. Проверка третьей гипотезы выявила, что наблюдается статистически значимая разница в количестве покупок в тестовой и контрольной группе (рисунок 9).

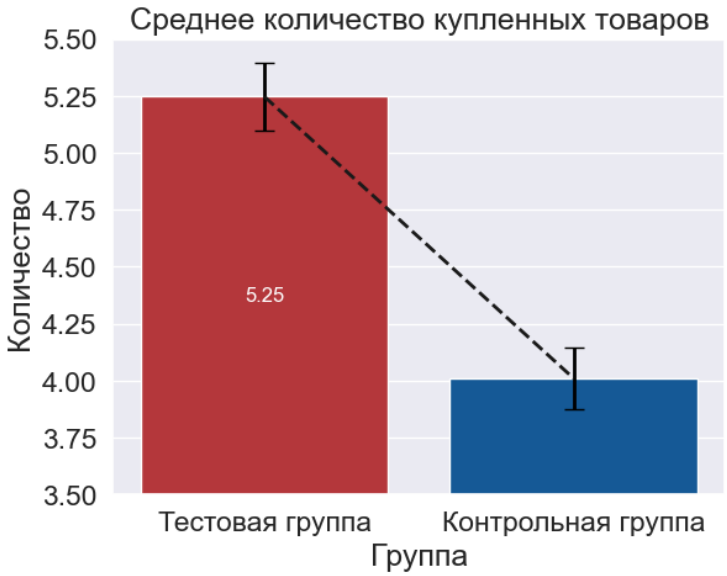

Рисунок 9 – Различия в среднем количестве товаров на покупателя в тестовой и контрольной группе

Выводы по проведенному A/B-тестированию.

По результатам проведения тестирования выявлена статистически значимая разница во всех трех метриках в расчете на одного клиента, выбранных для оценки эффективности маркетинговой кампании:

- средняя выручка в тестовой группе выше на 25%;
- средний чек выше на 6%;
- среднее количество купленных товаров выше на 25%.

Вывод - маркетинговая кампания эффективна.

Рекомендации - продолжать данную маркетинговую кампанию.

Этап 3. Кластеризация клиентов

1. Выбор модели кластеризации

Для проведения кластеризации были применены 3 вида алгоритмов: модель К-средних, модель кластеризации DBSCAN и модель KPrototypes. При этом моделирование производилось как без учета категориальных признаков, так и с их использованием. Все модели показали, что наилучшее количество кластеров 4, а наилучшую кластеризацию продемонстрировала модель Kprototypes: метрика силуэт-скор равна 0,65, что свидетельствует о достаточно плотной группировке объектов внутри кластеров (рисунок 10).

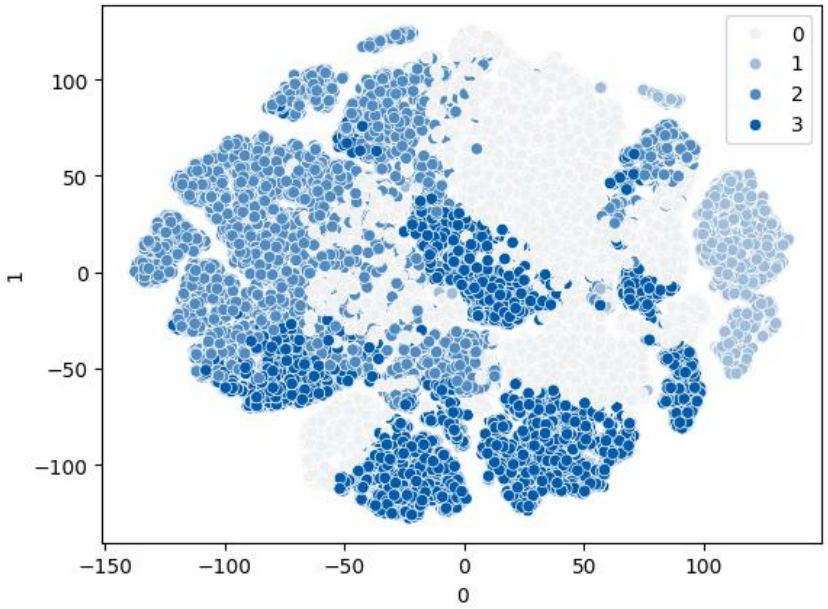

Рисунок 10 – Результат кластеризации клиентов

Значения усредненных значений признаков для каждого профиля представлены в таблице 1.

Таблица 1 – Усредненные значения признаков, характерных для каждого кластера.

	cluster 1	**cluster 2**	**cluster 3**	**cluster 4**
age	39,78	38,61	39,97	15,81
personal_coef	0,46	0,47	0,49	0,27
cost_sum	60572,66	25731,69	28715,41	37975,49
best_sale_mean	0,22	0,17	0,6	0,33
dt_max	50,3	23,23	44,19	39,46
education	0,84	0,8	0,82	0,2
gender	0,75	0,63	0,28	0,68
city	1134	1134	1134	1134
product_sex_mode	2	1	0	1
country	32	32	32	32

Результаты исследования:

В результате проведения исследования клиентской базы было выявленно 4 кластера. Профили полученных кластеров можно охарактеризовать следующим образом:

1. Первый кластер - это люди в возрасте около 40 лет, в основном мужчины (75%), имеющие среднее образованиее (16%), в 22% случаев приобретают товары со скидкой, готовы тратить в среднем 60,5 тысяч рублей и совершающие покупки один раз в полтора месяца (50 дней между поупками).

2. Второй кластер - это молодежь (средний возраст 16 лет), и имеют в основном высшее образование (80%). В основном мужчины (68%), треть покупок совершают со скидкой. Средняя сумма покупок составляет 37,9 тысяч рублей. В среднем покупки совершают раз в 40 дней.

3. Третий кластер - это люди в возрасте до 40 лет, совершающие пкупки в среднем раз в 44 дня, 18% из них имеет среднее образование, средние расходы составляют 28,7 тысячи рублей и в 40% случаев клиенты данной группы пиобретают товары по полной стоимости.

4. Четвертый кластер - в основном мужчины (63%), имеют среднее образование (20%), приобретают товары по скидкам 17%, средние расходы на пкупки составлют 25,7 тысячи рублей, возратная категория - 38+ лет.

Этап 4. Моделирование склонности клиента к покупке товара

1. Загрузка и создание датафреймов, содержащих информацию о клиента, учувствовавших в первой и второй маркетинговых компаниях

В соответствии с имеющимися данными сформированы два датафрейма по маркетинговым кампаниям:

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

Б. Вторая кампания проводилась на жителях города 1 134 и представляла собой баннерную рекламу на билбордах: скидка всем каждое 15-е число месяца (15-й и 45-й день в нашем случае).

Так как заказчика интересует конкретный регион (информация о жителях страны 32 города 1 188) была сформирована витрина данных по клиентов соответствии с заданными параметрами.

2. Обучение модели для определения склонности клиента к покупке товара

Для решения поставленной задачи была обучена модель RandomForestClassifier. Метрики качества выбранной модели следующие:

- F-мера 0,5484
- лучшие гипперпараметры модели: 'n_estimators': 300,
  'min_samples_split': 4,
  'min_samples_leaf': 6,
  'max_depth': 46,
  'class_weight': None

Результат:

С помощью полученной модели на обучающей выборке был проведен поиск склонности клиентов к покупке определенных товаров. После применения модели на тестовой выборке были получены следующие результаты по стране 32 городу 1188:

- в большей степени жители данного региона склонны к покупке кросовок;
- примерно одинаковая склонность у жителей региона к приобретению кед и бейсболок.

- не часто клиенты готовы приобретать брюки.

Именно на этих товарах необходимо сосредоточить внимание при запуске новой маркетинговой кампании в городе 1188, страны 32.