

**Отчёт о технических элементах**

**оценки результатов AB-тестов проведённых рекламных кампаний и построения модели склонности клиентов к покупке**

**в федеральной сети спортивных товаров**

**СОДЕРЖАНИЕ**

1. Извлечение данных из предоставленных источников и их предобработка.

2. Восстановление значений признака gender машинным обучением.

3. Оценка проведённых рекламных кампаний.

4. Кластеризация данных о покупках.

5. Построение модели машинного обучения для определения склонности клиентов к покупке.

Аргентов Сергей

[https://argentov.pro](https://argentov.pro/)

2023 год.

**1. ИЗВЛЕЧЕНИЕ ДАННЫХ ИХ ПРЕДОБРАБОТКА**

Загрузка и предобработка данных представлена в Разделе 1 исследования [файл «**Разделы 1 и 2. Загрузка с предобработкой данных и заполнение признака gender.ipynb**» — далее ноутбук].

Для проведения работы по формированию хранилища данных (витрины данных) подразделениями компании представлены предварительно не обработанные данные в 3 таблицах базы данных **shop\_database.db**. Так как требований по производительности и администрированию хранилища данных не задано, и работа с данным в ходе исследования производится средствами python3 – для формирования хранилища данных выбрана встроенная в python3 система управления базой данных sqlite3. Для хранения данных, полученных и преобразованных в ходе исследования создана база данных **purchases.db**.

Таблица 1. Характеристики таблиц данных, представленных подразделениями компании.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Наименование таблицы** | **Размер**  **(строк, столбцов) — объём** | **Дубликаты** | **Пропуски** |
| 1 | **purchases** | (786 260, 7) — 42.0 MB | 0 | 434 236 |
| 2 | **personal\_data\_coeffs** | (104 989, 5) — 4.0 MB | 0 | 0 |
| 3 | **personal\_data** | (89 241, 6) — 4.1 MB | 0 | 0 |

Пропущенные значения в таблице **purchases** содержаться в признаках **product\_sex** (15%) и **colour** (40%). Это может быть следствием технической/организационой ошибок при заполнении данных или может свидетельствовать о следующем:

- для признака **product\_sex** в случае унисекс-товаров данные о гендерной принадлежности потенциальных покупателей товара не проставлялась;

- для признака **colour** часть информации могла попасть в наименование товара, часть информации может быть получена из информации по одноимённым товарам, а в случае когда товар не имеет цветовой характеристики (пищевые продукты, лекарственные средства и т.п. или товары с не зафиксированным признаком цвета по техническим или организационным причинам) цвет товара не проставлялся.

Проверка наличия ссылки на пол в наименовании товара показала, что в отдельных товарах имеются ссылки на пол потенциального пользователя товара (мужской или женский), при этом все такие позиции соответственно уже были размечены в признаке **product\_sex**. Таким образом, все пропущенные значения данного признака, были размечены как дополнительное значение данного признака — 'унисекс' (присвоено значение 2.0) [см. п.1.2.1 ноутбука 1].

Восстановление пропущенных значений признака **colour** проводилось в четыре этапа

[см. п.1.2.2 ноутбука 1]:

- проверка наличия признака в наименовании товара (восстановлено 0.16%);

- проверка наличия признака в одноимённых товарах (восстановлено 14.62%);

- визуальная проверка наличия признака в названии товара (восстановлено 0.42%);

- визуальная проверка однородности значений признака (восстановлено 0.01%).

При заполнении признака для одноимённых товаров наиболее частым значением использовалось одновременное формирование двух объектов **DataFrame.Pandas**, что привело к необходимости использования многопроцессорного режима расчётов с целью сокращения времени расчётов. В результате время выполнения сложного сравнения **DataFrame.Pandas** на 9 процессорах составило всего 1.1 минуты. Этот способ позволил устранить 95% пропусков. Не восстановленным значениям признака **colour** было присвоено значение «**неопределённый**». Таким образом в данных исключены все пропущенные значения.

**2. ВОССТАНОВЛЕНИЕ ЗНАЧЕНИЙ ПРИЗНАКА gender МОДЕЛЯМИ МАШИННОГО ОБУЧЕНИЯ**

Расчёты и разработка модели машинного обучения для восстановления отсутствующих значений признака **gender** представлены в Разделе 2 исследования [ноутбук 1].

Дополнительно к персональным данным клиентов (таблица **personal\_data** базы данных **shop\_database.db**) подразделения компании предоставили таблицу с восстановленными данными. Предоставленная информация увеличила персональные данные на +17,6% (+15 748 записей) по 5 признакам из 6.

Для 15 тыс. записей признака **gender**, оставшихся невосстановленными, разработана модель машинного обучения. Для обучения все таблицы базы данных объединены в датасет, размер которого составил 786 260 записей. Незаполненные значения признака **gender** составили 15% от объединённого датасета. Перед обучением модели данные проверены на аномалии:

1. Использованы три метода выявления аномалий, дополнившие друг друга результатами.
2. При проверке данных на аномалии не рассматривались значения признаков **id**, **lbt\_coef**, **ac\_coef**, **sm\_coef** – так как значения этих признаков уникальны почти для каждого покупателя, а значит не принесут пользы при факторной оценке данных и обучении модели. Не учитывался признак **gender** – так как он является целевой переменной для обучения модели.
3. При использовании методов попризнакового выявления аномалий признаки **product\_sex**, **base\_sale**, **education** не учитывались – так как они имеют понятные, уникальные значения без аномалий (выявлено визуально).

**Таблица 1. Результаты исключения аномальных записей.**

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Наименование метода,**  **которым выявлялись аномалии** | **Количество**  **аномалий** | **%** |
| 1 | **Quantile** | 8 928 | 1,33 |
| 2 | **LocalOutlierFactor** | 35 630 | 5,32 |
| 3 | **IsolationForest** | 61 030 | 9,11 |
|  | ИТОГО | 82 150 | 12,26 |

Проверка сбалансированности датасета показала соотношение **55% к 45%** позитивного класса к негативному. Такая пропорция была учтена гиперпараметром **class\_weight**, который позволят обучать модели с разбалансированностью обучающей выборки до **57% к 43%**.

Оценены три модели машинного обучения на достижение наиболее точного результата моделирования: Логистическая регрессия, Случайный лес, Нейронная сеть «Перцептрон».

**Таблица 2. Результаты тренировочного обучения моделей.**

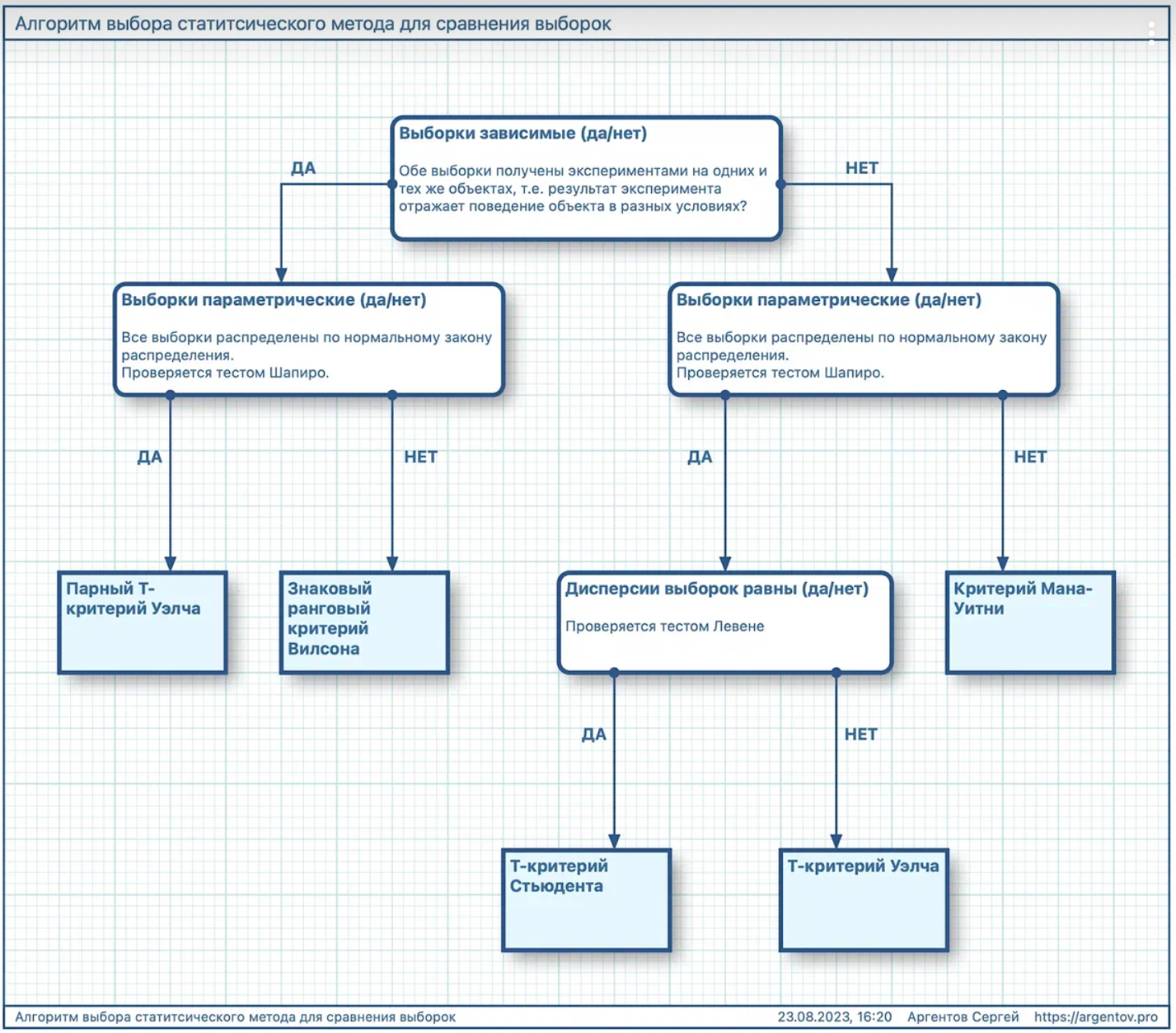
|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Наименование**  **модели** | **Точность по метрике accuracy\_score, %** | **Время обучения,**  **секунды** |
| 1 | **LogisticRegression()** | 99,1 | 0,5 |
| 2 | **RandomForestClassifier()** | 100,0 | 12 |
| 3 | **MLPClassifier()** | 100,0 | 18 |

Метрики трёх алгоритмов машинного обучения показали высокое качество предсказаний, что может свидетельствовать о наличии признаков, тесно связанных с целевой переменной. При решении нашей задачи — это положительный фактор. Для восстановления пропущенных данных в признаке **gender** выбран алгоритм «**Случайный лес**» так как данный алгоритм показал максимальное значение метрики (100%) при минимальном времени обучения (12 сек.). При проверке модели на независимых данных метрикой **f1\_score** получена точность модели 99,98%, что является достаточно высоким показателем, который может быть следствием полного восстановления пропусков в данных в ходе предобработки информации, полученной от подразделений компании [см. раздел 1 настоящего отчёта]. Датасет с восстановленными данными записан в базу данных **pursheses.db** в таблицу **full\_table**.

**3. ОЦЕНКА ЭФФЕКТИВНОСТИ РЕКЛАМНЫХ КАМПАНИЙ**

Расчёты и анализ результатов AB-тестов представлены в Разделе 3 исследования [файл «**Раздел 3. Оценка эффективности рекламных кампаний.ipynb**» — далее ноутбук 2].

**3.1. Разработка вспомогательного инструментария «класс сравнения выборок на наличие статистически значимой разницы»**

Для оценки статистической значимости различия исследуемых распределений в тестовой и контрольных группах клиентов — был разработан класс **SampleComparisonTest**, который выбирает соответствующий статистический тест для сравнения двух заданных пользователем выборок. Класс учитывает являются ли заданные выборки «зависимыми», «параметрическими» и «равными по дисперсии». Данный класс размещён в пакете «**package.statistics**» в модуле «**select\_run\_test**». Полная русифицированная документация по классу может быть получена методом «\_\_doc\_\_». Для удобства разработчиков в классе **SampleComparisonTest** реализован методом **show\_algorithm\_scheme()**, с помощью которого можно в любой момент посмотреть визуализацию следующего алгоритма выбора статистического теста:

**3.2. Первая рекламная кампания**

Эффективность первой рекламной кампании проверялась по следующим метрикам:

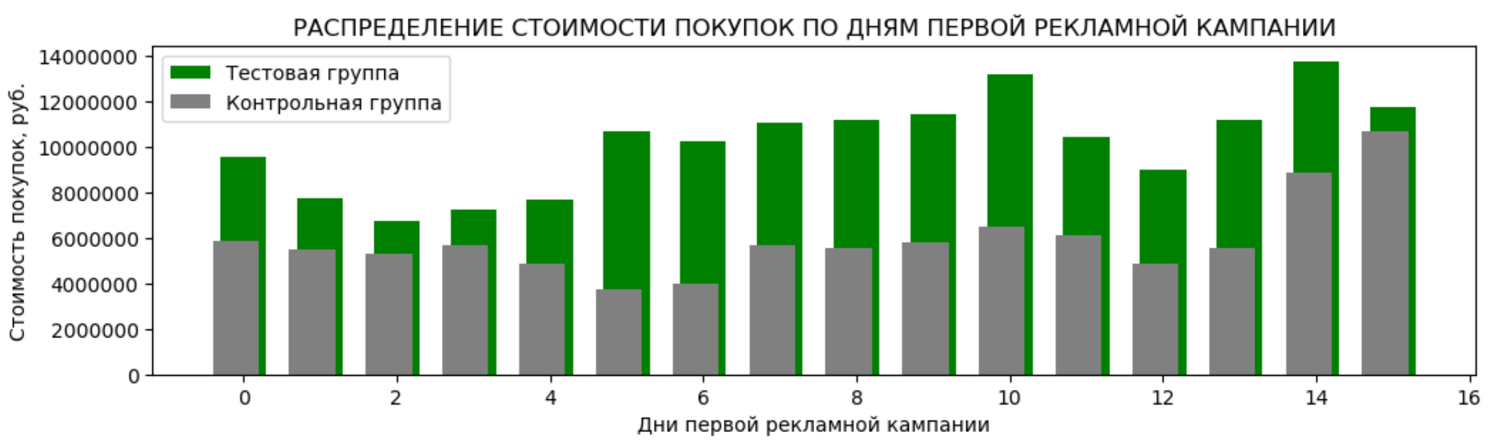
- Стоимость покупок покупателями.

- Частота покупок покупателями.

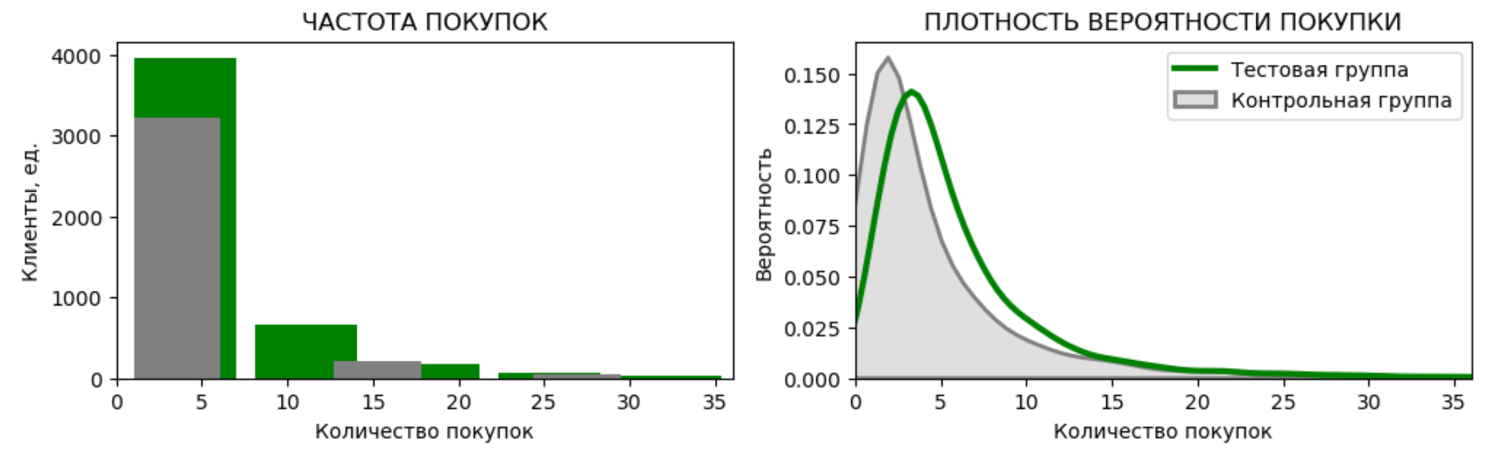
- Стоимость и частота покупок наиболее состоятельной группой покупателей (возрастная категория от 31 до 61 года). Сведения о состоятельности разных возрастных групп взяты из данных Сбербанка, представленных в статье: <https://blog.domclick.ru/ipoteka/post/kazhduyu-desyatuyu-ipoteku-v-sbere-berut-molodye-lyudi-analitika-domklik>

Из обобщённого датасета выделены тестовая и контрольная группы, данные о которых представлены подразделениями кампании. Проверка сбалансированности групп показала что в каждой из них участвовало по 5 тыс. клиентов. Таким образом AB-тестирование проведено корректно с точки зрения количества участников. Сравнение стоимостей покупок клиентами тестовой и контрольной групп по дням проведения рекламной кампании показало следующее.

Статистически значимо изменилась стоимость покупок [см. пункт 3.2.1. ноутбука 2]:



Статистически значимо изменилась частота покупок [см. пункт 3.2.2. ноутбука 2]:



**Однако, ни стоимость, ни частота покупок наиболее состоятельных групп относительно менее состоятельных - не произошло [см. пункт 3.2.3. ноутбука 2].**

Таким образом:

1) Рекламная кампания мотивировала большее количество клиентов на повторные покупки и значительно увеличила выручку. Это с одно стороны позволит снизить издержки и на инфраструктуру обслуживания клиентов, так как она будет загружена более равномерно, а с другой стороны увеличит доходы за счёт повторного возврата клиентов, ожидающих дополнительных акций. Для поддержания эффекта целесообразно повторять такие рекламные кампании с сезонной периодичностью.

2) Однако, в отношении наиболее состоятельных групп клиентов — в будущих кампаниях рекомендуется акцентировать внимание на конкретных возрастных группах, таким образом, чтобы наиболее сильно мотивировать к повторным покупкам группы клиентов со средним и высоким доходом (от 31 до 61 года), которые смогут в будущем совершать более дорогие и более частые покупки.

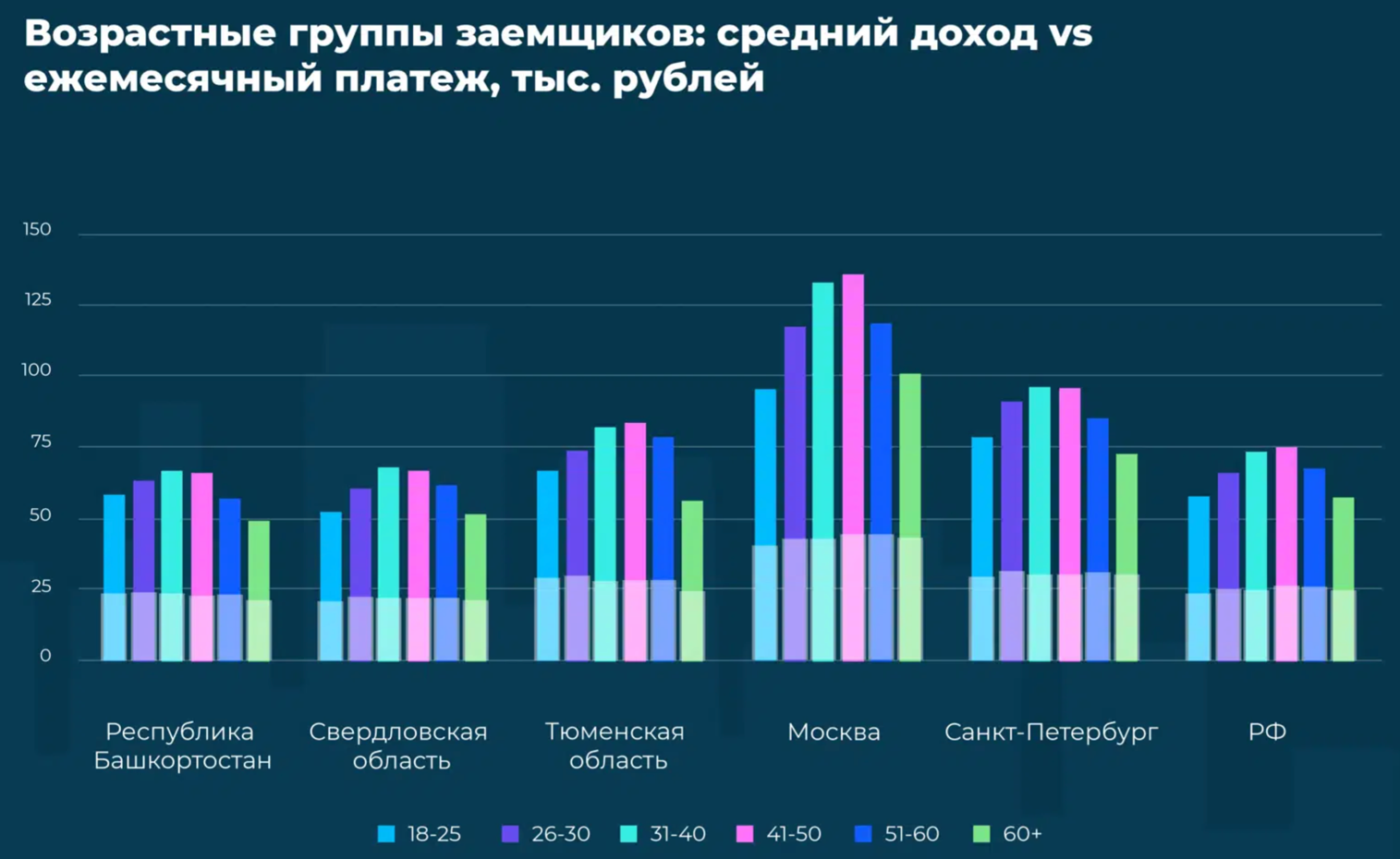
**3.3. Вторая рекламная кампания**

Поскольку вторая рекламная кампания проводилась на отдельной территории (город №1134) её эффективность проверялась по изменившейся плотности вероятности покупок покупателями в дни акции в данном городе в сравнении с другими городами. Количество таких покупателей в городе 1134 составило 796 человек, а в других городах 1 973 человека. При этом балансировка размеров тестовой и контрольной выборок не целесообразны так как оценка данной рекламной кампании проводилась по относительному показателю (плотность вероятности покупок). Такая оценка является единственно возможной так как вторая рекламна кампания проводилась через канал в конкретном городе, количество пользователей которого неконтролируемо (банер), а в качестве контрольной группы для такой кампании можно выбрать только покупателей других территорий.

В результате данной рекламной кампании плотность вероятности разовой покупки покупателем снизилась что является положительным фактором, в случае роста вероятности нескольких покупок, однако этого не произошло:



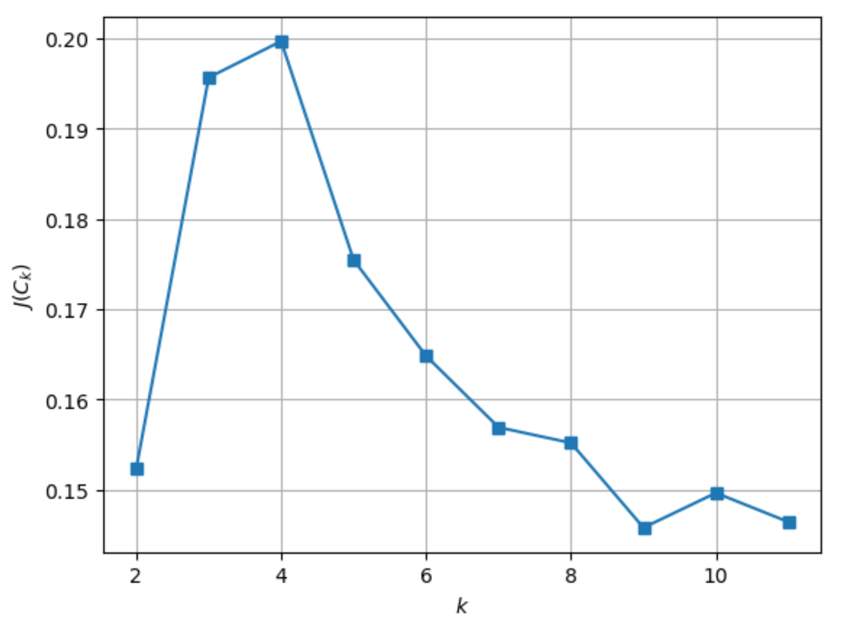
Таким образом кампанию можно признать неудачной. В будущем её необходимо переработать, например, сориентировавшись на группы с большей доходностью, тыс.руб.:



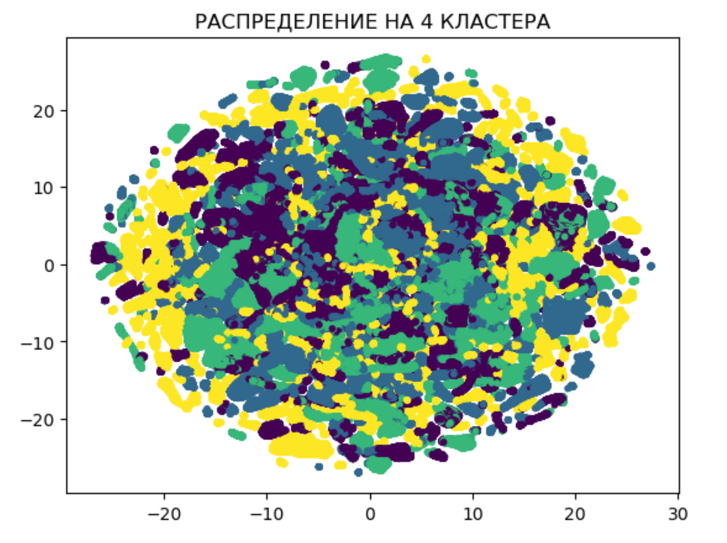
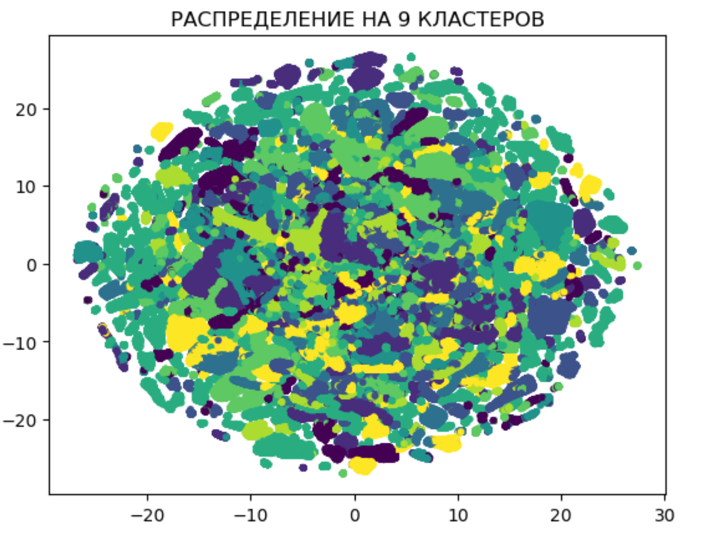
**4. КЛАСТЕРИЗАЦИЯ ДАННЫХ О ПОКУПКАХ**

Кластеризация данных о покупках представлена в Разделе 4 исследования [файл «**Раздел 4. Кластеризация транзакций покупок.ipynb**» — далее ноутбук 3].

Для ускорения расчёта кластеризации данные обобщённого датасета стандартизированы и нормализованы. Получено двухмерное пространство признаков. Рассчитана графическая кривая «колен» оптимального количества кластеров, полученных методом k-средних:



Для анализа принято минимальное оптимальное количество кластеров (4 кластера), так как при увеличении количества кластеров их графически наблюдаемое разделение не улучшается:



**Таблица 3. Выявленная корреляция кластеров с признаками датасета.**

|  |  |  |  |
| --- | --- | --- | --- |
| № | Признак | Значение корреляции | Описание |
| 1 | **education** | 0.739 | Сильная корреляция |
| 2 | **personal\_coef** | 0.443 | Средняя корреляция |
| 3 | **gender** | 0.137 | Очень слабая корреляция |
| 4 | age | 0.136 | Очень слабая корреляция |
| 5 | product\_sex | 0.129 | Очень слабая корреляция |

Таким образом, разбиение на 4 кластера выявило признаки, по группе которых можно распределять клиентов чтобы они проявляли одинаковые поведенческие патерны. Это значит, что для повышения эффективности рекламных кампаний целесообразно делать их направленными на покупателей, с разными признаками: **education**, **personal\_coef**, **gender**, **age**, **product\_sex** – так как отличие именно этих признаков покупателей имеет корреляцию с разделением клиентов на кластеры.

Для практического использования в рекламных кампаниях разделения покупателей на кластеры — разработана модель классификации клиентов, которая позволит отнести каждого клиента к одному из кластеров. В том числе:

- уточнены признаки, по которым покупатель отнесён к кластеру (с важностью признака >1%);

- определён перечень топ-10 товаров, приобретаемых покупателями каждого кластера.

**Таблица 4. Характеристики четырёх выделенных кластеров покупателей.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№** | **Характеристика** | **Кластер 1** | **Кластер 2** | **Кластер 3** | **Кластер 4** |
| 1 | **Наиболее важные признаки** | product: 0.393403  cost: 0.226707  age: 0.157966  city: 0.136804  education: 0.042333  product\_sex: 0.019962  personal\_coef: 0.0188 | product: 0.605358  cost: 0.125969  age: 0.096099  education: 0.069049  city: 0.067916  personal\_coef: 0.0191  product\_sex: 0.0129 | product: 0.606876  cost: 0.112318  age: 0.10063  education: 0.082844  city: 0.067435  personal\_coef: 0.014711  product\_sex: 0.013682 | education: 0.530191  age: 0.308106  personal\_coef: 0.14485  gender: 0.013632 |
| 2 | **Топ-10 приобретаемых товаров** |  |  |  |  |

**Выводы:**

1. При определении каждого кластера клиентов важными являются признаки **age**, **education**, **personal\_coef**. Это означает, что при разработке рекламных кампаний вместе с персональным коэффициентом целесообразно, в первую очередь, акцентировать внимание на конкретные возрастные группы. А учитывая рекомендации из раздела 3 настоящего отчёта — дополнительно стоит учитывать статистику доходов возрастных групп:

- Так клиентам от 31 до 61 года целесообразно предлагать наиболее дорогие и бОльшие по комплектации товары.

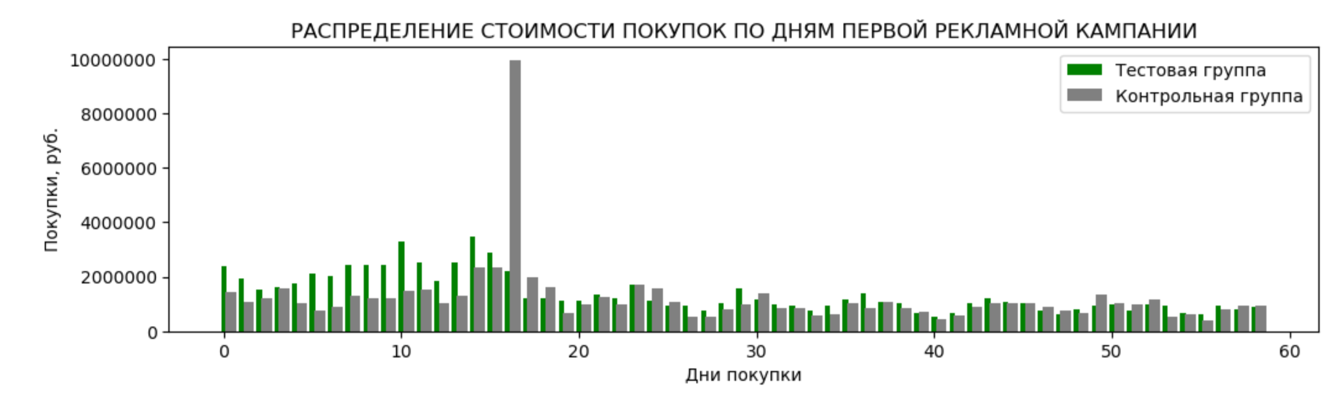
- При достаточности ресурсов на рекламные кампании целесообразно разработать и провести отдельные кампании на каждую возрастную группу, определённую признаком доходности: менее 31 года, от 31 до 61 года, старше 61 года.

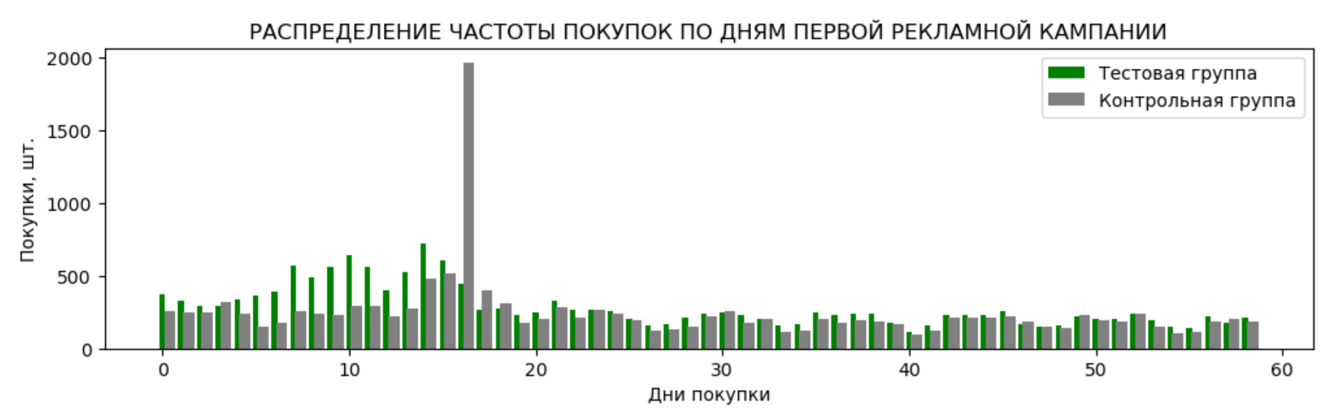
- При недостаточности ресурсов на рекламные кампании целесообразно разработать и провести кампанию для группы клиентов с наибольшим доходом - от 41 до 51 года. Для повышения имиджа компании как социально направленной можно исключит верхний диапазон акции и провести её для всех клиентов старше 41 года.

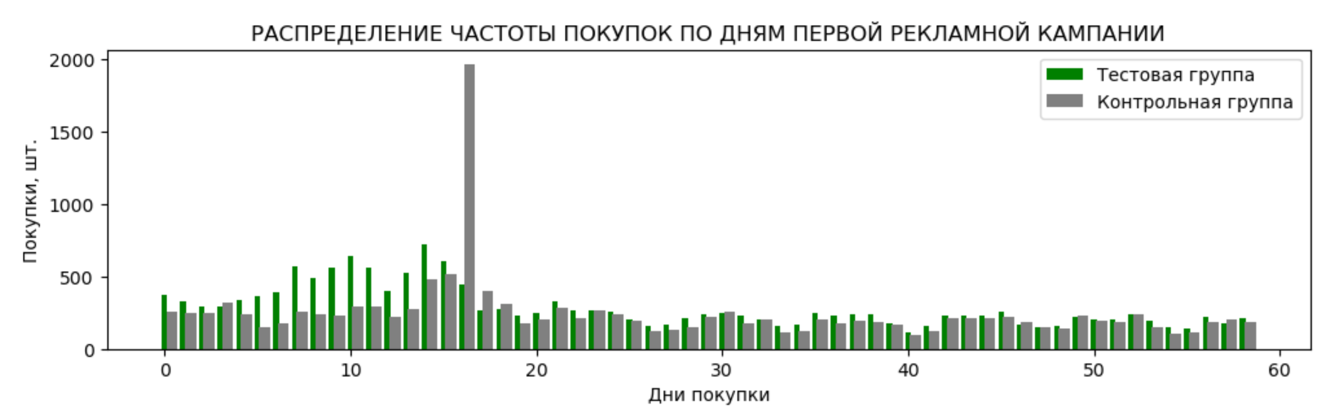
2. Для клиентов из кластера №4 одним из важных признаков является **gender**. Соответственно целесообразно разработать отдельные рекламные компании для женщин и мужчин и провести их в разные периоды времени, чтобы продавцы-консультанты справлялись с потоком покупателей и концентрировались на продажах более узкого спектра товаров, что, в свою очередь, повысит качество предпродажных консультаций и соответственно повысит количество продаж.

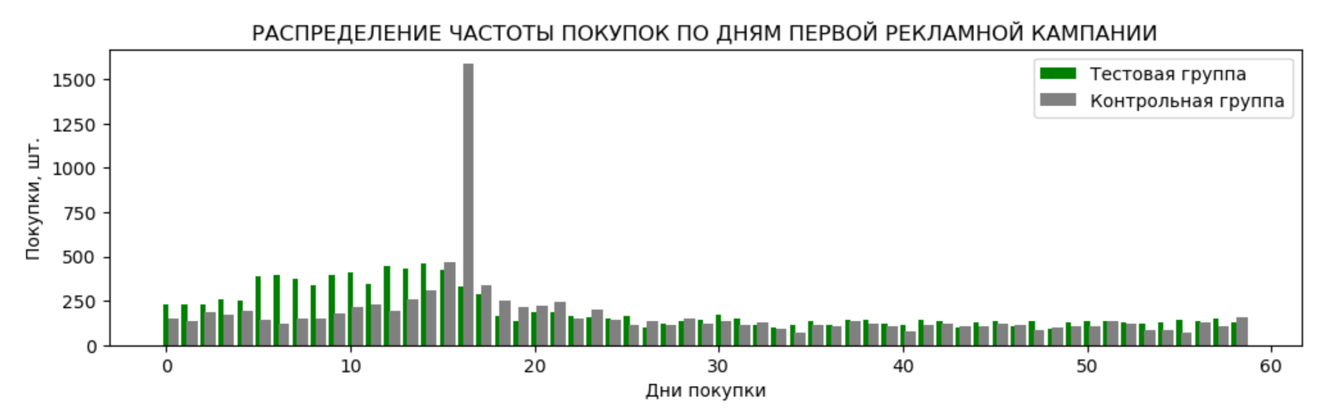
3. При контакте с клиентом целесообразно сразу определять кластер, к которому он относится и предлагать ему дополнительно приобрести перечень товаров из топ-10 товаров, наиболее часто приобретаемых покупателями соответствующего кластера (см. Топ-10 выше для каждого кластера в п.2 Таблицы 4).

**Скидка, предоставленная клиентам, статистически значимо:**

- увеличивает для 1 кластера сумму и частоту покупок:



- увеличивает для 3 кластера сумму частоту покупок:

- увеличивает для 3 кластера сумму частоту покупок:

Соответственно среди клиентов, отнесённых к кластерам 1, 2, 4 целесообразно проводить рекламные кампании подобные первой (успешной) рекламной кампании.

**5. ПОСТРОЕНИЕ МОДЕЛИ СКЛОННОСТИ КЛИЕНТОВ К ПОКУПКАМ**

Разработка модели отнесения клиентов к клиентам, склонным к покупке, представлена в Разделе 5 исследования [файл «**Раздел 5. Построение модели склонности клиентов к покупкам.ipynb**» — далее ноутбук 4].

Клиент считается склонным к покупке — если после коммуникации с ним в успешной рекламной кампании клиент купил больше одного товара. Таким образом для построения модели склонности клиента к покупке данные были размечены по принципу: клиент участвовал в первой (успешной) рекламной кампании и совершил: одну покупку (3 397) или более одной покупки (6 313 клиентов). Из полученного датасета были исключены аномальные записи тремя методами, дополняющими результаты друг друга.

**Таблица 5. Результаты исключения аномальных записей.**

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Наименование метода,**  **которым определялись аномалии** | **Количество**  **аномалий** | **%** |
| 1 | **Quantile** | 507 | 1,31 |
| 2 | **LocalOutlierFactor** | 400 | 1,04 |
| 3 | **IsolationForest** | 7 864 | 20,38 |
|  | ИТОГО | 8 207 | 21,27 |

Проверена сбалансированность датасета, которая показала что позитивный и негативный класс представлены в датасете с соотношением 66% и 34%. Такая пропорция была не достаточна чтобы учесть разбалансированность гиперпараметрами обучаемой модели машинного бучения. Таким образом из записей положительного класа были исключены записи в количестве, позволившем довести соотношение классов до допукстимого (57% к 43%).

Из трёх моделей (Логистическая регрессия, Случайный лес и Нейронная сеть «Перцептрон») наилучшее качество показала модель «Случайный лес».

**Таблица 6. Результаты тестового обучения моделей.**

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Наименование**  **модели** | **Точность по метрике accuracy\_score, %** | **Время обучения,**  **секунды** |
| 1 | **LogisticRegression** | 95,9 | 0,02 |
| 2 | **RandomForestClassifier** | 96,5 | 0,45 |
| 3 | **MLPClassifier** | 96,5 | 0,54 |

После обучения модели на всём датасете была прогнозирована склонность к покупкам клиентов, проживающих в городе 1188 страны 32. Для определения значимости склонности этих клиентов к покупкам также была вычислена склонность к покупкам клиентов, по которым проведена первая (успешная) рекламная кампания.

**Таблица 7. Результаты оценки склонности клиентов к покупкам.**

|  |  |  |  |
| --- | --- | --- | --- |
| № | Параметр оценки | Первая рекламная кампания, % | Город 1188  страны 32, % |
| 1 | **Клиенты, склонные к покупкам** | 66 | 78 |
| 2 | **Клиенты, не склонные к покупкам** | 34 | 22 |

Таким образом, в городе 1188 страны 32 целесообразно проводить рекламную кампанию подобную первой (успешной). Перед проведением кампании предпочтительна её доработка, по рекомендациям, указанным в разделах 3 и 4 настоящего отчёта.