
# Отчёт об оценке результатов АВ-теста новой системы ценообразования страховых полисов и технических особенностях формирования витрины данных и проведения статистических тестов

**СОДЕРЖАНИЕ**



Аргентов Сергей
https://argentov.pro

2023 год.

## 1. ИЗВЛЕЧЕНИЕ ДАННЫХ ИЗ ИМЕЮЩИХСЯ ИСТОЧНИКОВ.

Для проведения работы по формированию хранилища данных (витрины данных) подразделениями компании представлены предварительно не обработанные данные в четырёх таблицах Excel, а также указан сайт Центрального банка РФ как источник данных о текущем курсе рубль/доллар.

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 создана база данных insurance.

Таблица 1. Характеристики таблиц данных, представленных подразделениями компании.

| № | Наименование таблицы | Размер (строк, столбцов) | Дубликаты | Пропуски |
|---|---|---|---|---|
| 1 | experiment_group | (14842, 3) | 0 | 2 493 |
| 2 | case_losses | (45, 4) | 0 | 0 |
| 3 | case_contracts | (3711, 10) | 0 | 0 |
| 4 | case_clients | (3711, 6) | 0 | 276 |

Данные о курсе доллара, получены через API Центрального банка РФ http://www.cbr.ru/scripts/XML_daily.asp и занесены в хранилище данных в таблицу currency_rate. Особенности работы данного API описаны по адресу: https://www.cbr-xml-daily.ru .
Для обеспечения соизмеримости данных о ценах и стоимости страховых полисов данные о соответствующих стоимостях преобразовывались в долларовый и рублёвый эквивалент с учётом актуального курса рубль/доллар на дату проведения расчётов по исследованию.

**Отчётность о предварительных результатах исследования:**

В ходе исследования таблицы базы данных группировались и обогащались информацией, после чего результирующие отчёты выводились в файл Excel (для предварительной сверки с бизнес-заказчиком), а также загружались в хранилище данных для обеспечения централизованного доступа при последующих исследованиях.

Для оптимизации оперативной памяти, используемой датафреймами после записи данных в хранилище применялась процедура очистки оперативной памяти от датафреймов.

ВНИМАНИЕ! Данная функция при первом запуске выдаёт ошибку «применение процедуры на изменяющемся локальном пространстве переменных». При повторном запуске функция отрабатывает корректно.

```
dfs = []
local_variable_scope =locals().keys()
for elem in local_variable_scope:
    if 'df' in elem:
        dfs.append(elem)
# Удаляем не используемые датафреймы
for elem in dfs:
    del locals()[elem]
```

Извлечение данных из источников и их запись в хранилище представлено в п.2 ноутбука «Исследование_страховых_полисов.jpynb» (далее - ноутбук), а также в соответствующих подразделах резюмирующих отдельные пункты исследования.

ВИМАНИЕ! Отдельные библиотеки, методами которых выполнялась кластеризация данных, конфликтовали с базовым набором библиотек JupyterNotebook. Окружение, в котором корректно работает код ноутбука исследования приложено к данному отчёту в файле requirements.txt

## 2. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ.

Обработка данных произведена в двух направлениях: исключение записей с пропусками и маркирование записей с аномалиями. Первоначальные данные по отдельным признакам содержали от 0,5% до 98,6% пропусков.

```
RangeIndex: 3315 entries, 0 to 3314
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   contract_id       3315 non-null   int64
 1   contract_num      3315 non-null   object
 2   product_name      3315 non-null   object
 3   last_name         3315 non-null   object
 4   first_name        3298 non-null   object
 5   middle_name       3089 non-null   object
 6   age               3315 non-null   int64
 7   sex               3315 non-null   object
 8   client_id         3315 non-null   int64
 9   duration          3315 non-null   int64
 10  country           3315 non-null   object
 11  price             3315 non-null   int64
 12  insurance_amount  3315 non-null   int64
 13  loss_name         45 non-null     object
 14  loss_payout_amt   45 non-null     float64
 15  currency_name     3315 non-null   object
```

При этом наибольшее количество пропусков в данных содержались в признаках об «отчестве клиента» и «выплатах при страховом случае».

Поскольку *отчество* клиента не влияет на результаты исследования и также, как и фамилия и имя, должно быть зашифровано (для соблюдения требований по безопасности персональных данных) – пропуски в данных об *отчестве* заполнялись спецсимволом «-». Замена пропусков на соответствующие значения представлена в п.2.3 ноутбука.

Пропуски данных в признаке «выплаты при страховом случае» с большой вероятностью означают отсутствие таких выплат. Соответственно в данном поле пропущенные данные заменялись на значение «0» (ноль). Замена пропусков на соответствующие значения представлена в п.3.1 ноутбука.

После заполнения пропусков данных получился датасет с количеством уникальных значений в отдельных признаках от 2 до 3315 шт:

```
contract_id           3315
contract_num            75
product_name             2
client_name           1227
age                     52
sex                      2
client_id             3315
duration                23
country                 18
price                   44
insurance_amount         5
loss_name                2
loss_payout_amt         20
currency_name            2
insurance_amount_$       5
```

Для определения записей с аномалиями принят следующий алгоритм:

1. Из всех колонок с признаками, выбраны только те, которые не являются идентификаторами [contract_id, client_id] и не являются производными от других признаков [insurance_amount_$]. Это позволило оптимизировать время выполнения процедур без потери качества расчётов.
2. Все признаки закодированы с помощь метода *LabelEncoder()*.
3. Выявлены аномалии методом квантилей (отсечены данные более 3-сигма).
4. Выявлены аномалии методом Локалфактора *LocalOutlierFactor(n_neighbors=5)* и Изолеса *IsolationForest(random_state=42)*.

   Справочно: поскольку метод *IsolationForest()* не индепотентен – для обеспечения воспроизводимости результата его работы целесообразно использовать его с параметром *random_state*.
5. Сформирован словарь с индексами записей, которые попали в любой из перечней аномальных записей, выявленных вышеперечисленными методами.

Таким образом, различные методы выявления аномалий фактически дополнили друг друга. При этом необходимо учитывать что в отношении аномальных записей, выявленных методами Локалфактора и Изолеса, невозможно интерпретировать по каким именно признаками конкретная запись выделена как аномалия. Однако, учитывая что в нашем исследовании итоговой задачей является оценка проведённого АВ-теста – знание о признаках, которые относят запись к «аномальной» является необязательным.

Таблица 2. Результаты выявления аномалий различными методами.

| № | Наименование метода | Выявленные аномальные записи | |
|---|---|---|---|
| | | количество | % |
| 1 | Quantile (все записи, попадающие в диапазон 3-сигма) | 19 | 0,57 |
| 2 | LocalOutlierFactor | 394 | 11,89 |
| 3 | IsolationForest | 542 | 16,35 |
| | **ВСЕГО** | **841** | **25,37** |

По результатам исследования в данные добавлен признак anomaly. Соответствующие преобразования выполнены в п.3 ноутбука.

## 3. КЛАСТЕРИЗАЦИЯ ДАННЫХ О СТРАХОВЫХ ПОЛИСАХ.

Для выделения кластеров в имеющихся данных апробированы два варианта стратегии. Первый вариант – «Кластеризация только по числовым признакам» (см. п.4.2 ноутбука). Второй вариант – «Кластеризация по числовым и категориальным признакам» (см. п.4.3 ноутбука). При реализации второй стратегии данные предварительно кодировались методом OrdinalEncoder(). В каждой стратегии данные стандартизировались методом StandardScaler(). Далее n-мерное пространство данных преобразовывалось в двухмерное, к которому последовательно применялись методы кластеризации: иерархический метод (только для первой стратегии), метод k-средних. После проведения расчётов данные визуализировались покластерно. На визуальных изображениях кластеров определялся наиболее выраженный результат кластеризации. Данные о советующем кластере присваивались каждой записи датасета. Лучшие результаты кластеризации двух вариантов стратегий получились следующие:

### 3.1. Кластеризация только по числовым признакам

Рис.3.1.1. Двухмерное пространство.

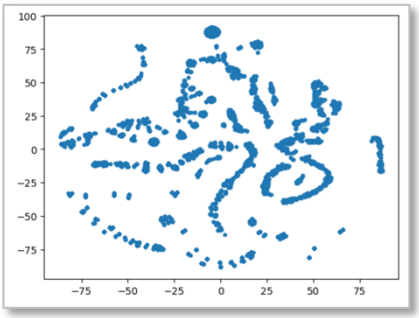



Рис.3.1.2. Иерархический метод.

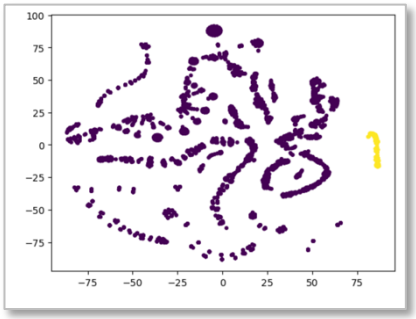
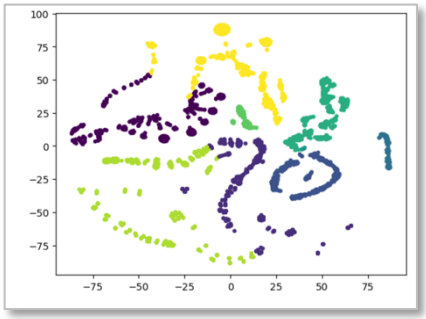



Рис.3.1.3. Методом k-средних.




### 3.2. Кластеризация по числовым и категориальным признакам

Рис.3.2.1. Двухмерное пространство.

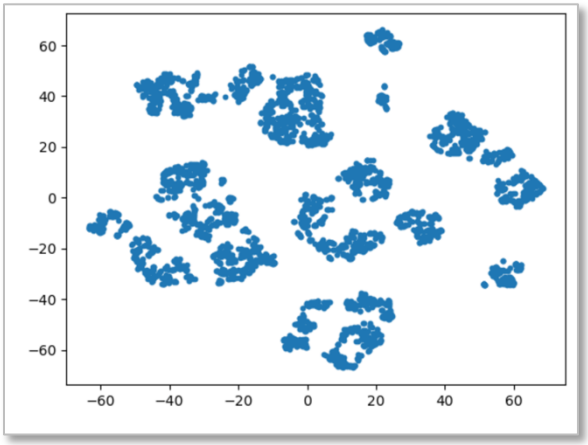
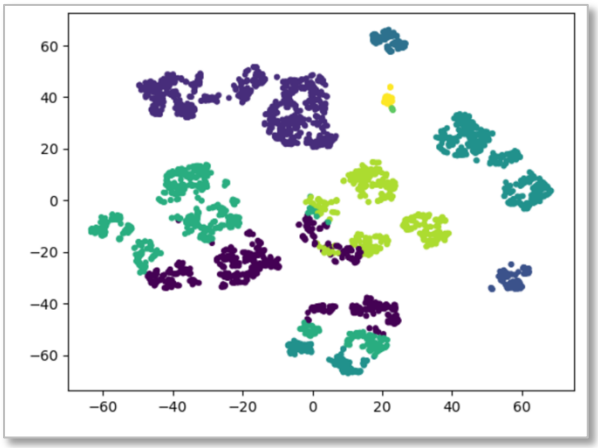



Рис.3.2.2. Методом k-средних.




Как видно из графического представления результатов кластеризации разными стратегиями и разными методами – наиболее ярко-выраженные кластеры получаются при проведении кластеризации по второй стратегии методом k-средних. Наибольшее число кластеров, получившееся при оптимизации – 9 кластеров.

Для выявления признаков, от которых наиболее сильно зависит отнесение каждой конкретной записи датасета к конкретному кластеру проведён корреляционный анализ методом factorize(x):

Рис.3.3. Матрица корреляций признаков и принадлежности записи к кластеру.

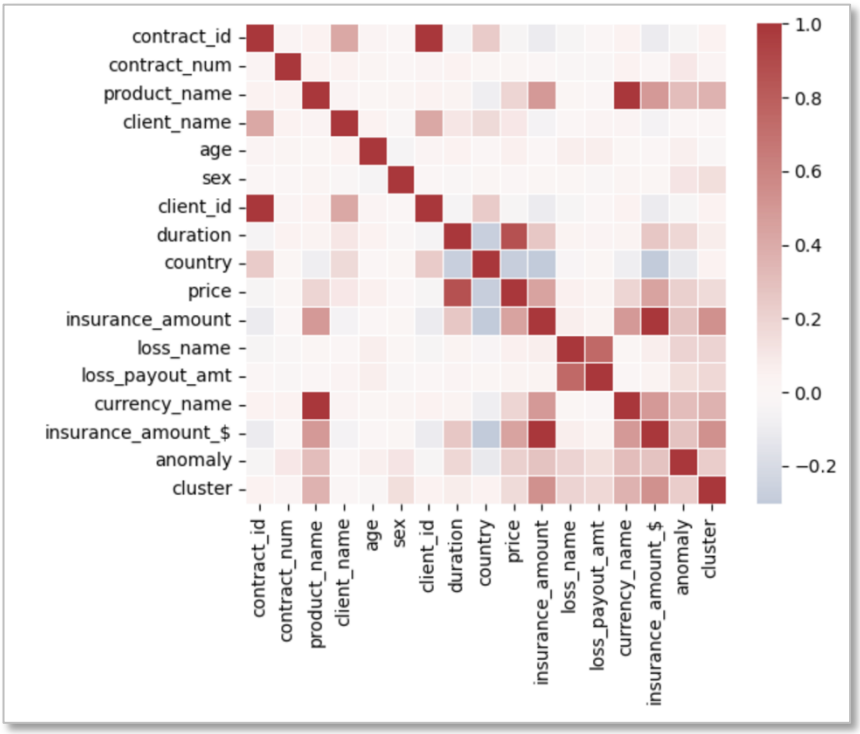



Рис. 3.4. Степень корреляции признаков и принадлежности к кластеру.

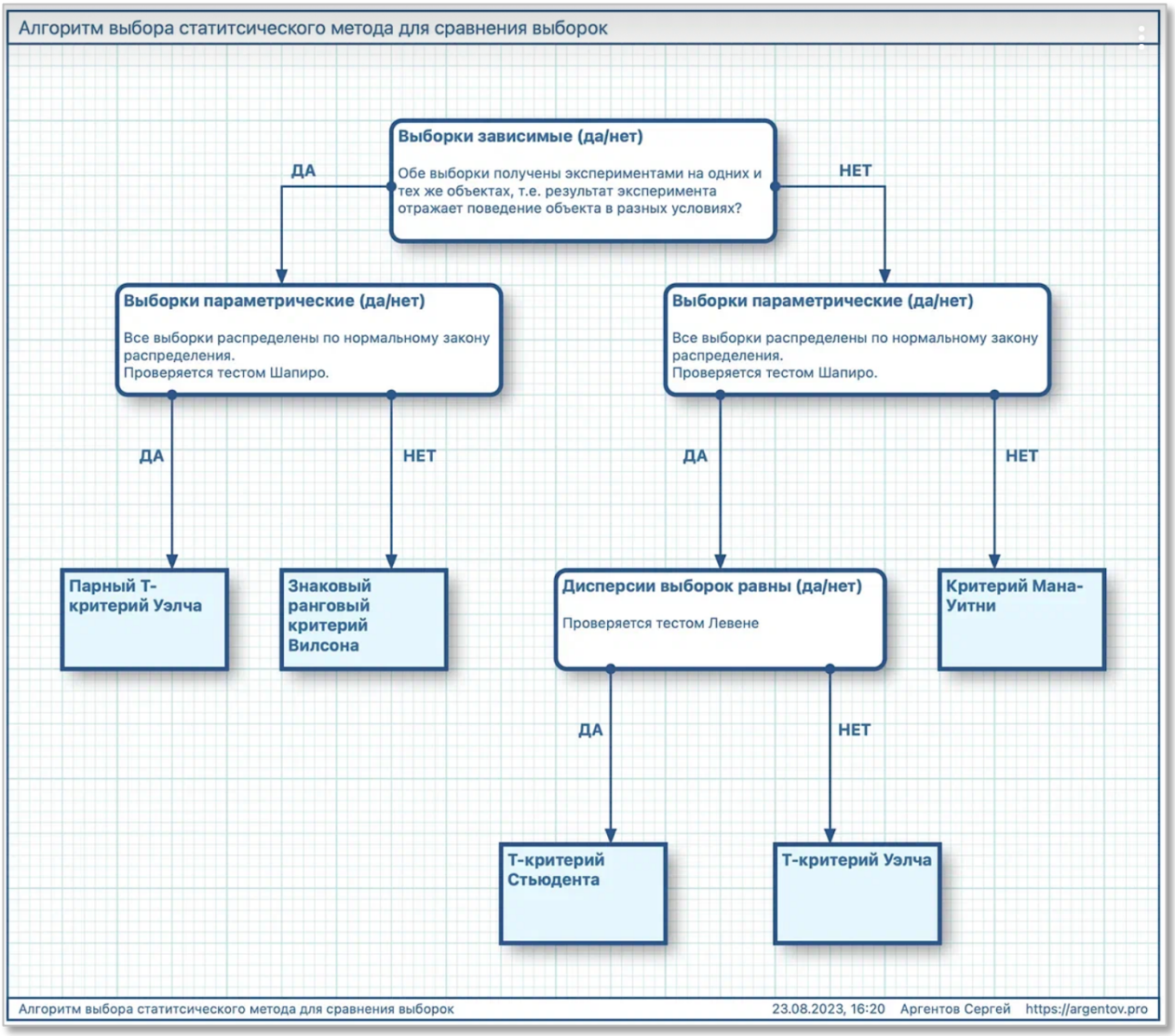
| | |
|---|---|
| insurance_amount_$ | 0.534974 |
| insurance_amount | 0.534974 |
| product_name | 0.368315 |
| currency_name | 0.368315 |
| anomaly | 0.231943 |
| loss_name | 0.202928 |
| loss_payout_amt | 0.175056 |
| price | 0.164177 |
| sex | 0.144487 |
| duration | 0.077750 |
| country | 0.041644 |
| client_id | 0.040513 |
| contract_id | 0.040513 |
| contract_num | 0.022877 |
| client_name | −0.000081 |
| age | −0.017539 |

Таким образом, кластеры выделены по средней корреляции страховой суммы полисов, пола, наименований продукта-полиса. Остальные признаки не являются «значимо выделяющими» отдельные группы приобретённых клиентами полисов.

## 4. АНАЛИЗ АВ-ТЕСТА «ИЗМЕНЕНИЯ СИСТЕМЫ ЦЕНООБРАЗОВАНИЯ ПОЛИСОВ».

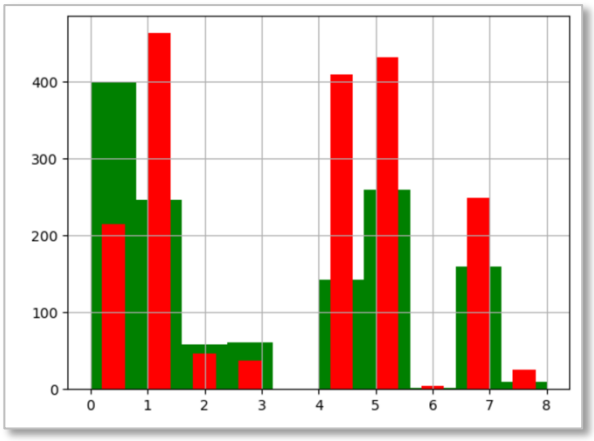
Проверка сбалансированности датасета, отражающего результаты АВ-теста, показала что датасет разбалансирован на 28,9%. Такая ситуация могла возникнуть из-за некорректного разделения полисов на тестовую и контрольную группы. Например, часть недействующих полисов могли попасть в сплитуемый датасет и таким образом не участвовать в итоговой выборке, которая формировалась только по действующм полисам.

Для оценки статистической значимости различия распределения полисов по кластерам, по цене полисов и по величине страховых выплат был написан класс SampleComparisonTest, который выбирает соответствующий статистический тест для сравнения двух выборок, в зависимости от того являются ли они «зависимыми», «параметрическими» и «равными по дисперсии». Класс размещён в пакете «asg.statistics» в модуле «select_run_test». Полная русифицированная документация по данноме 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:

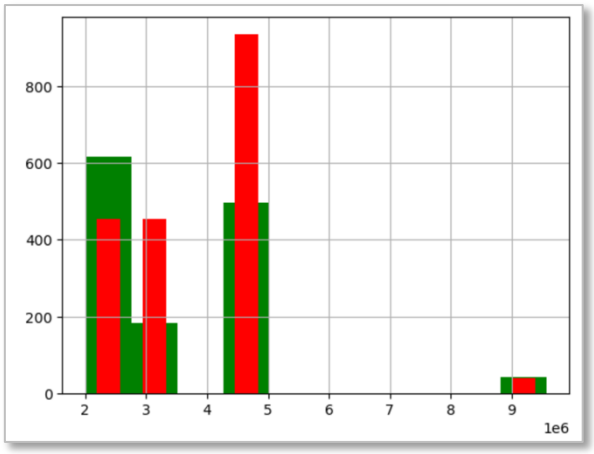


Все тестируемы выборки являются непараметрическими. При этом выборки по кластерам и по цене полисов статистически значимо различаются и свидетельствуют об ухудшении общей картины продаж полисов при изменении системы ценообразования полисов на новую систему. В частности:

а) по критерию «распределение по типам (кластерам) полисов» – полисы в тестовой выборке распределились более равномерно чем в контрольной выборке. Таким образом при переходе на новую систему ценообразования будет необходимо концентрироваться на продвижении продаж большего типа полисов. Это менее эффективно для бизнеса по затратам.




б) по критерию «распределение по ценовым группам полисов» – полисы в тестовой выборке перераспределились в сторону продаж более дешёвых полисов. Таким образом при переходе на новую систему ценообразования, возможно, снизится выручка компании и, соответственно, упадут прибыли.




Кроме того, при применении новой системы ценообразования статистически значимо не уменьшаются страховые выплаты по полисам, а соответственно сокращение расходов компании не произойдёт.

Таким образом:

1) АВ-тест целесообразно повторить, устранив причины разбалансировки датасета с результатами тестовой и контрольной групп.
2) Опираясь на полученные результаты выполненного АВ-теста можно сделать вывод об отсутствии оснований изменения действующей системы ценообразования по полисам на новую систему, так как предложенное новое изменение может привести к:
   - Увеличению кластеров клиентов, для которых необходимо будет проводить усиленные маркетинговые и рекламные кампании.
   - Снижению продаж более прибыльных полисов.

   При этом с большой вероятностью не удастся снизить убытки компании так как статистически значимого изменения по страховым выплатам не произойдёт.

Расчёт анализа проведённого АВ-теста «Изменения системы ценообразования для продажи полисов» см. в п.5 ноутбука.