Постановка задач

Задачи

- Написать функцию, которая будет считать retention игроков (по дням от даты регистрации игрока).
- Определить какой набор предложений можно считать лучшим по имеющимся результатам А/В теста.
- Предложите метрики для оценки результатов последнего прошедшего тематического события в игре.

Предоставленные данные:

- problem1-reg_data.csv данные о времени регистрации
- problem1-auth_data.csv данные о времени захода пользователей в игру
- problem2.csv результаты A/B теста



Выводы исследования

- Стоит подумать об акциях и/или новых механиках игры, позволяющих увеличить удержание пользователей после седьмого дня.
 - (Максимальное удержание пользователей происходит на 4-7 дни использования продукта, после чего показатель retention начинает падать)
- Изменения из тестовой группы рекомендуется запустить в прод. (По результатам A/B теста статистически значимо удалось увеличить ARPU и ARPPU)
- Список метрик для оценки результатов прошедшего тематического события в игре:
 - ASD (Average Session Duration),
 - MAU/DAU (Daily Active Users / Monthly Active Users),
 - Stickiness rate,
 - · Users Online,
 - ATV (Average Transaction Value).



Поиск аномалий данных

- Были обнаружены и убраны из датасета регистрации пользователей за период 1998-2013 годы (1.76% от общего числа регистраций) не релевантны исследованию.
- Порядка 7% данных имели аномальную частоту заходов в игру меньше 5 минут между сеансами (возможно боты или случайные заходы пользователей в игру, без самого процесса игры были удалены из датасета).
- При анализе данных для А/В-тестирования были обнаружены 123 аномально высоких значения выручки с пользователя в контрольной группе эти значения были удалены из данных, т.к. с наибольшей вероятностью они имеют ошибочную природу происхождения (технический баг, проблемы со сплитованием и т. д.).



Результаты А/В-тестов*

- Конверсия в покупку и равенство распределений значений выручки с пользователя в тестовой и контрольной группах не имеют статистически значимых различий.
- ARPU, ARPPU и распределения значений выручки с платящих пользователей в тестовой и контрольной группах статистически значимо различаются.



^{*} Для проверки гипотез использовались proportions_ztest (statsmodels.stats), непараметрический статистический U-критерий Манна-Уитни и bootstrap средних значений.

Пример работы функции расчёта Retention

```
Retention Heatmap
def retention(data, reg pos=2, auth pos=0, uid pos=1, min date=None, max date=None):
   # делаем выборку нужных колонок
                                                                                                                      Day 0
                                                                                                                                   Day 1
                                                                                                                                               Day 2
                                                                                                                                                            Day 3
                                                                                                                                                                         Day 4
                                                                                                                                                                                      Day 5
                                                                                                                                                                                                  Day 6
                                                                                                                                                                                                               Day 7
                                                                                                                                                                                                                            Day 8
                                                                                                                                                                                                                                        Day 9
   df = data.iloc[:, [uid pos, reg pos, auth pos]].copy(deep=True)
                                                                                                                      100.00%
                                                                                                                                   6.61%
                                                                                                                                               13.43%
                                                                                                                                                            13.84%
                                                                                                                                                                         15.08%
                                                                                                                                                                                     21.07%
                                                                                                                                                                                                  18.39%
                                                                                                                                                                                                              21.07%
                                                                                                                                                                                                                           13.43%
                                                                                                                                                                                                                                        17.77%
                                                                                                       2020-09-01
   # проверка наличия и определение параметра min date
                                                                                                                      100.00%
                                                                                                                                   8.00%
                                                                                                                                               12.40%
                                                                                                                                                            14.80%
                                                                                                                                                                        18.00%
                                                                                                                                                                                     19.80%
                                                                                                                                                                                                  22.60%
                                                                                                                                                                                                              20.00%
                                                                                                                                                                                                                           15.40%
                                                                                                                                                                                                                                        13.00%
                                                                                                       2020-09-02
   if min_date is None:
                                                                                                                                                                                                  24.68%
                                                                                                                      100.00%
                                                                                                                                   6.72%
                                                                                                                                               15.61%
                                                                                                                                                            13.61%
                                                                                                                                                                        18.87%
                                                                                                                                                                                     20.15%
                                                                                                                                                                                                              18.33%
                                                                                                                                                                                                                          15.79%
                                                                                                                                                                                                                                        18.15%
                                                                                                       2020-09-03
       min date = df.reg_ts.min()
                                                                                                                                   6.31%
                                                                                                                                               11 44%
                                                                                                                                                            11 44%
                                                                                                                                                                        16 17%
                                                                                                                                                                                     20.12%
                                                                                                                                                                                                  22.88%
                                                                                                                                                                                                              15 58%
                                                                                                                                                                                                                          16 77%
                                                                                                                                                                                                                                        13.61%
                                                                                                       2020-09-04
                                                                                                                      100.00%
   # проверка наличия и определение параметра max date
                                                                                                       2020-09-05
                                                                                                                      100 00%
                                                                                                                                   8 58%
                                                                                                                                               12.28%
                                                                                                                                                            14 42%
                                                                                                                                                                        16 96%
                                                                                                                                                                                     18.32%
                                                                                                                                                                                                  23.20%
                                                                                                                                                                                                              15 20%
                                                                                                                                                                                                                          15.79%
                                                                                                                                                                                                                                        16 96%
   if max date is None:
       max_date = df.reg_ts.max()
                                                                                                       2020-09-06
                                                                                                                      100.00%
                                                                                                                                   9.69%
                                                                                                                                               13.37%
                                                                                                                                                            16.86%
                                                                                                                                                                        17.44%
                                                                                                                                                                                     18.22%
                                                                                                                                                                                                  21.12%
                                                                                                                                                                                                              18.80%
                                                                                                                                                                                                                           18.02%
                                                                                                                                                                                                                                        17.05%
                                                                                                                                   9.11%
                                                                                                                                                                                     19.77%
                                                                                                                                                                                                  22.67%
                                                                                                                                                                                                              16.28%
                                                                                                       2020-09-07
                                                                                                                      100.00%
                                                                                                                                               13.18%
                                                                                                                                                            16.67%
                                                                                                                                                                        15.70%
                                                                                                                                                                                                                           12.79%
                                                                                                                                                                                                                                        17.44%
   # выносим даты регистрации и входа в приложение в отдельные колонки
   df['date_reg'] = df.reg_ts.dt.date
                                                                                                       2020-09-08
                                                                                                                      100.00%
                                                                                                                                   5.39%
                                                                                                                                               12.14%
                                                                                                                                                            12.14%
                                                                                                                                                                        15 03%
                                                                                                                                                                                     23.51%
                                                                                                                                                                                                  23.31%
                                                                                                                                                                                                              20 42%
                                                                                                                                                                                                                          15.99%
                                                                                                                                                                                                                                        15.99%
   df['date_auth'] = df.auth_ts.dt.date
                                                                                                       2020-09-09
                                                                                                                      100 00%
                                                                                                                                   5 69%
                                                                                                                                               10 57%
                                                                                                                                                            15 24%
                                                                                                                                                                        18 70%
                                                                                                                                                                                     21 75%
                                                                                                                                                                                                  18 70%
                                                                                                                                                                                                              19 31%
                                                                                                                                                                                                                           14 23%
                                                                                                                                                                                                                                        16 46%
   # делаем выборку по диапазону дат
                                                                                                       2020-09-10
                                                                                                                      100 00%
                                                                                                                                   6.51%
                                                                                                                                               13 61%
                                                                                                                                                            15 19%
                                                                                                                                                                        13 41%
                                                                                                                                                                                     21 50%
                                                                                                                                                                                                  19 92%
                                                                                                                                                                                                              22 49%
                                                                                                                                                                                                                           15 38%
                                                                                                                                                                                                                                        14 40%
   df = df[(df['reg ts'] >= min date) & (df['reg ts'] <= max date)]</pre>
                                                                                                                      100.00%
                                                                                                                                   4 23%
                                                                                                                                               13.08%
                                                                                                                                                            14.69%
                                                                                                                                                                        18.31%
                                                                                                                                                                                     18.51%
                                                                                                                                                                                                  21.73%
                                                                                                                                                                                                              20.52%
                                                                                                                                                                                                                           14.29%
                                                                                                                                                                                                                                        15.09%
                                                                                                       2020-09-11
   # посчитаем количество дней между датой регистрации и входом в приложение
                                                                                                                      100.00%
                                                                                                                                   6.08%
                                                                                                                                               12.16%
                                                                                                                                                            14.90%
                                                                                                                                                                        14.71%
                                                                                                                                                                                     19.22%
                                                                                                                                                                                                  20.78%
                                                                                                                                                                                                              18.24%
                                                                                                                                                                                                                           16.67%
                                                                                                                                                                                                                                        16.27%
                                                                                                       2020-09-12
   df['days'] = df['date_auth'] - df['date_reg']
                                                                                                       2020-09-13
                                                                                                                                                            13.83%
                                                                                                                      100.00%
                                                                                                                                   6.41%
                                                                                                                                               13.23%
                                                                                                                                                                        16 63%
                                                                                                                                                                                     21.24%
                                                                                                                                                                                                  21.04%
                                                                                                                                                                                                              16.03%
                                                                                                                                                                                                                           15.23%
                                                                                                                                                                                                                                        16.03%
   # сгруппируем по дням регистрации и количеству дней использования
                                                                                                                      100.00%
                                                                                                                                   7.05%
                                                                                                                                               12.13%
                                                                                                                                                            14.29%
                                                                                                                                                                        16.83%
                                                                                                                                                                                     20.55%
                                                                                                                                                                                                  22.11%
                                                                                                                                                                                                              19.96%
                                                                                                                                                                                                                           14.09%
                                                                                                       2020-09-14
                                                                                                                                                                                                                                        9.20%
   cohorts = df.groupby(['date_reg', 'days'], as_index=False) \
                .agg({'uid': 'nunique'}) # агрегируем по уникальным пользователям
                                                                                                                                                                                                                                                              - 0.4
                                                                                                       2020-09-15
                                                                                                                      100.00%
                                                                                                                                   7.02%
                                                                                                                                               12.87%
                                                                                                                                                            14.42%
                                                                                                                                                                        19.10%
                                                                                                                                                                                     17.74%
                                                                                                                                                                                                  21.83%
                                                                                                                                                                                                              17.93%
                                                                                                                                                                                                                            7.80%
                                                                                                       2020-09-16
                                                                                                                      100.00%
                                                                                                                                   6.18%
                                                                                                                                               13.75%
                                                                                                                                                            14.94%
                                                                                                                                                                        15.14%
                                                                                                                                                                                     21.91%
                                                                                                                                                                                                  22.51%
                                                                                                                                                                                                              10.96%
   # сформируем сводную таблицу
   cohorts_pivot = cohorts.pivot(index='date_reg', columns='days', values='uid')
                                                                                                                      100.00%
                                                                                                                                   5.91%
                                                                                                                                               15.32%
                                                                                                                                                            16.19%
                                                                                                                                                                        19.47%
                                                                                                                                                                                     22.32%
                                                                                                                                                                                                  15.10%
                                                                                                       2020-09-17
   # посчитаем retention
                                                                                                       2020-09-18
                                                                                                                      100.00%
                                                                                                                                   7.60%
                                                                                                                                               18.14%
                                                                                                                                                            16.91%
                                                                                                                                                                        20.83%
                                                                                                                                                                                     14.95%
   retention = pd.DataFrame() # пустой датафрейм к которому будем прибавлять построково
                                                                                                       2020-09-19
                                                                                                                      100.00%
                                                                                                                                   9.04%
                                                                                                                                               17.47%
                                                                                                                                                            23.19%
                                                                                                                                                                        17.17%
   for i in range(cohorts_pivot.shape[0]): # итерируемся по строкам
                                                                                                                                                                                                                                                              - 0.2
      a = cohorts_pivot.iloc[i, 0] # 100% число пользователей для этой даты
                                                                                                       2020-09-20
                                                                                                                      100.00%
                                                                                                                                   13.70%
                                                                                                                                               24.32%
                                                                                                                                                            13.36%
       temp = cohorts pivot.iloc[i.:l.divide(a, axis=0).round(4) # получаем пересчет на проценты
                                                                                                       2020-09-21
                                                                                                                      100.00%
                                                                                                                                   13.42%
                                                                                                                                               21.21%
       retention = retention.append(temp) # построково добавляем посчитанные значения в итоговую таблицу
          retention
                                                                                                        2020-09-22
                                                                                                                      100.00%
                                                                                                                                   7.82%
                                                                                                                      100.00%
                                                                                                        2020-09-23
   return retention
```



Список метрик

- Использование предложенного набора метрик (ASD, MAU/DAU, Stickiness rate, Users Online, ATV) оптимально, чтобы как можно точнее следить за тем, как происходящие изменения в игровых механиках отражаются на поведении пользователей и самом продукте.
- Рекомендуется дополнительно провести исследование как перечисленные метрики коррелируют (или даже предсказывают) траты пользователей и LTV. Это будет полезно для последующей приоритизации используемых метрик.

Дополнительно

Подробности исследования, технические детали и код в Jupyter Notebook

