Домашняя работа №9 (Итоговая)

Выполнила Наталья Кейзер, студентка потока от 29.03.2023

Практическое задание:

Проанализируйте результаты эксперимента и напишите свои рекомендации менеджеру.

Mobile Games AB Testing with Cookie Cats

Решение:

1. Анализ данных.

Сначала проанализируем данные. В том числе поищем ошибки и несоответствия в данных.

In [1]:
```
import pandas as pd
import numpy as np
import scipy.stats as stats
```

In [2]:
```
df = pd.read_csv('A_B_cookie_cats.csv')
df.head()
```

Out[2]:

	userid	version	sum_gamerounds	retention_1	retention_7
0	116	gate_30	3	0	0
1	337	gate_30	38	1	0
2	377	gate_40	165	1	0
3	483	gate_40	1	0	0
4	488	gate_40	179	1	1

1.1. Посмотрим **количество пользователей** в каждой из версий:

In [3]:
```
pd.DataFrame((df.version.value_counts().rename("Quantity"), df.version.value_counts
    .format(precision=1, thousands=" ")
```

Out[3]:

	gate_40	gate_30
Quantity	45 489.0	44 700.0
Percentage	50.4	49.6

Разделение примерно поровну.

1.2. Проверим, есть ли пользователи, которые **попали сразу в обе выборки**:

In [4]:
```
check1 = pd.pivot_table(df, index='userid', columns='version', values='sum_gamerou
```

```
check1.head()
```

Out[4]:

version	gate_30	gate_40	All
userid			
116	1.0		1
337	1.0		1
377		1.0	1
483		1.0	1
488		1.0	1

In [5]:
```
check1[check1.All>1]
```

Out[5]:

version	gate_30	gate_40	All
userid			
All	44700	45489	90189

Анализ выдает только общее значение, это означает, что каждый пользователь попал **только в одну группу** - версия 30 или версия 40.

1.3. Проверим, есть ли **задвоения в id пользователей**:

In [6]:
```
check2 = pd.DataFrame(df.userid.value_counts())
check2.head()
```

Out[6]:

	userid
116	1
6632278	1
6658202	1
6658194	1
6658134	1

In [7]:
```
check2[check2.userid>1]
```

Out[7]:

userid

Задвоений пользователей в базе не найдено.

Можно приступать к анализу полученных результатов.

2. Анализ результатов

Мы будем анализировать серии retention_1 и retention_7 (удержание пользователей через 1 и 7 дней, соответственно).

2.1 Исследуем наличие корреляции между этими показателями:

In [8]:
```
corr_auto = np.corrcoef(df.retention_1, df.retention_7)
print("Коэффициент корреляции между удержанием на 1-й и 7-й дни равен: {0:.2%}".fo
```

```
Коэффициент корреляции между удержанием на 1-й и 7-й дни равен: 32.74%
```

Используем Шкалу Чеддока для интерпретации результатов:

Показатель корреляции	**0,1-0,3**	**0,3-0,5**	**0,5-0,7**	**0,7-0,9**	**0,9-0,99**
Сила связи	слабая	умеренная	заметная	высокая	весьма высокая

Коэффициент корреляции свидетельствует о наличии умеренной зависимости между этими двумя показателями. Это означает, что нужно исследовать оба показателя, и только после этого можно интерпретировать результаты.

2.2 Проанализируем **конверсию в удержание в 1й день**.

Количество пользователей в группе gate_30:

In [9]:
```
n30 = df.userid[df.version=='gate_30'].nunique()
n30
```

Out[9]:
```
44700
```

Количество пользователей в группе gate_30 с retention_1 = 1:

In [10]:
```
r1_30 = sum(df.retention_1[df.version=='gate_30'])
r1_30
```

Out[10]:
```
20034
```

Количество пользователей в группе gate_40:

In [11]:
```
n40 = df.userid[df.version=='gate_40'].nunique()
n40
```

Out[11]:
```
45489
```

Количество пользователей в группе gate_40 с retention_1 = 1:

In [12]:
```
r1_40 = sum(df.retention_1[df.version=='gate_40'])
r1_40
```

Out[12]:
```
20119
```

In [13]:
```
print('Результаты конверсии в удержание пользователей в 1-й день:')
print('Конверсия в группе gate_30: {0:.2%}.\nКонверсия в группе gate_40: {1:.2%}.'
```

```
Результаты конверсии в удержание пользователей в 1-й день:
Конверсия в группе gate_30: 44.82%.
Конверсия в группе gate_40: 44.23%.
```

Проведем z-test:

H_0: Статистические различия между выборками отсутствуют (выборки примерно одинаковы).

H_1: Выборки показали разные результаты, разница статистически значимая.

In [14]:
```
from statsmodels.stats import proportion
```

In [15]:
```
z_score, z_pval = proportion.proportions_ztest(np.array([r1_30, r1_40]), np.array(
print('Result of z-test: z-score = {0:.2}, p-value = {1:.2}'.format(z_score, z_pval
```

```
Result of z-test: z-score = 1.8, p-value = 0.074
```

Так как $p_{value} > \alpha$ ($\alpha = 5\%$), то мы не можем отвергнуть H_0 с уровнем достоверности 95%. Таким образом, статистически значимых различий между выборками нет.

In [16]:
```
chisq, p_val, table = proportion.proportions_chisquare(np.array([r1_30, r1_40]), np
print('Result of chi-test: chi-score = {0:.2}, p-value = {1:.2}'.format(chisq, p_val
```

```
Result of chi-test: chi-score = 3.2, p-value = 0.074
```

Результаты χ-теста подтверждают полученные ранее результаты.

2.3 Проанализируем **конверсию в удержание в 7й день**.

Количество пользователей в группе gate_30 с retention_7 = 1:

In [17]:
```
r7_30 = sum(df.retention_7[df.version=='gate_30'])
r7_30
```

Out[17]:
```
8502
```

Количество пользователей в группе gate_40 с retention_7 = 1:

In [18]:
```
r7_40 = sum(df.retention_7[df.version=='gate_40'])
r7_40
```

Out[18]:
```
8279
```

In [19]:
```
print('Результаты конверсии в удержание пользователей в 7-й день:')
print('Конверсия в группе gate_30: {0:.2%}.\nКонверсия в группе gate_40: {1:.2%}.'
```

```
Результаты конверсии в удержание пользователей в 7-й день:
Конверсия в группе gate_30: 19.02%.
Конверсия в группе gate_40: 18.20%.
```

Проведем z-test:

H_0: Статистические различия между выборками отсутствуют (выборки примерно одинаковы).
H_1: Выборки показали разные результаты, разница статистически значимая.

In [20]:
```
z_score, z_pval = proportion.proportions_ztest(np.array([r7_30, r7_40]), np.array(
print('Result of z-test: z-score = {0:.2}, p-value = {1:.2}'.format(z_score, z_pval
```

```
Result of z-test: z-score = 3.2, p-value = 0.0016
```

Так как $p_{value} < \alpha$ ($\alpha = 5\%$), то мы отвергаем H_0 с уровнем достоверности 95%. Таким образом, действительно разница между конверсией на 7-й день становится заметна и статистически значима.

Важно отметить, что конверсия уменьшается. То есть, изменения в версии 40 относительно версии 30 привели к снижению возвращений пользователей на 7-й день.

Если целью проекта было увеличение конверсии, то эффект получен противоположный.

Рекомендуется доработка проекта, и дальнейшее продолжение тестирования.

На основании тестирования данную версию 40, как она есть, внедрять не рекомендуется.