

Когортный анализ поведения пользователей на основе синтетических данных с использованием Python

Андриенко Иван Сергеевич

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

В данной работе представлен когортный анализ пользовательской активности, выполненный на основе синтетически сгенерированных данных. Исследование охватывает пользователей, зарегистрированных в 2023 году, с отслеживанием их активности по месяцам. С помощью языка программирования Python реализован расчет показателя удержания (retention rate), а также проведён анализ средней активности пользователей. Построены тепловая карта удержания и графики, демонстрирующие динамику посещений и покупок. Работа подтверждает пригодность синтетических данных для тестирования аналитических подходов в маркетинге и пользовательской аналитике.

Ключевые слова: когортный анализ, удержание, retention rate, синтетические данные, Python, визуализация, пользовательская активность

Cohort analysis of user behavior based on synthetic data using Python

Andrienko Ivan Sergeevich

Sholom-Aleichem Priamursky State University

Student

Abstract

This paper presents a cohort analysis of user activity based on synthetically generated data. The study covers users registered in 2023, tracking their activity by month. Using the Python programming language, the retention rate is calculated, and the average user activity is analyzed. A heat map of retention and graphs showing the dynamics of visits and purchases have been built. The work confirms the suitability of synthetic data for testing analytical approaches in marketing and user analytics.

Keywords: cohort analysis, retention, retention rate, synthetic data, Python, visualization, user activity

1 Введение

1.1 Актуальность

В условиях высокой конкуренции и растущих затрат на привлечение новых пользователей важно понимать динамику поведения уже существующих клиентов. Когортный анализ позволяет выявлять

закономерности в активности пользователей в зависимости от времени их регистрации. Использование синтетических данных для такого анализа исключает риски утечки персональной информации и даёт возможность безопасного тестирования гипотез.

1.2 Обзор исследований

Когортный анализ широко применяется в современных исследованиях как мощный инструмент для изучения поведения групп и долгосрочных тенденций. В работе Е.И. Паршиной когортный анализ рассматривается в качестве ключевого метода маркетинговой аналитики, позволяющего оценивать клиентскую активность и оптимизировать бизнес-процессы. Автор особо подчёркивает его значимость для анализа retention rate и прогнозирования LTV в условиях ограниченных маркетинговых бюджетов [1]. Практическим аспектам применения когортного анализа посвящено исследование А.А. Черник, где метод представлен как основа юнит-экономики. На конкретных примерах демонстрируется, как когортный анализ помогает оценивать рентабельность различных сегментов клиентов и принимать обоснованные управленческие решения. Особое внимание уделяется взаимосвязи между когортным анализом и показателями CAC и LTV [2]. В контексте некоммерческих исследований Н.А. Кондакова детально рассматривает применение когортного анализа для оценки динамики здоровья детей. Работа демонстрирует универсальность метода, его способность выявлять устойчивые тенденции на протяжённых временных периодах, что особенно ценно для социально-значимых исследований [3]. Теоретические основы и методологические аспекты когортного анализа глубоко проанализированы в работе О.В. Терещенко. Автор систематизирует подходы к формированию когорт и интерпретации результатов, подчёркивая особую ценность метода для longitudinal studies в социальных науках. Приводятся конкретные примеры применения метода для изучения социальной мобильности и изменений ценностных ориентаций [4].

1.3 Цель исследования

Цель исследования – построение когортного анализа с использованием синтетических данных, моделирующих поведение пользователей в течение года после регистрации, и визуализация ключевых метрик, характеризующих уровень удержания и активности.

2 Результаты и обсуждения

На первом этапе были сгенерированы синтетические данные, имитирующие активность 1000 пользователей, зарегистрированных в течение 2023 года. Для каждого пользователя указывалась дата регистрации (RegistrationDate), а также ежемесячная активность в течение 12 месяцев с момента регистрации или до конца 2023 года. Таким образом, пользователь, зарегистрированный в январе, имел до 12 записей активности, тогда как зарегистрированный в декабре – только одну.

Сгенерированные данные включают следующие поля:

- UserID – уникальный идентификатор пользователя;
- RegistrationDate – дата регистрации (формат ГГГГ-ММ-ДД);
- ActivityMonth – месяц активности;
- Visits – количество посещений;
- Purchases – количество покупок.

Визуальная проверка таблицы показала корректное распределение дат регистрации по месяцам и наличие активности в соответствующих периодах.

	A	B	C	D	E	F	G	
1	UserID,RegistrationDate,ActivityMonth,Visits,Purchases							
2	1,2023-06-04,2023-06,5,0							
3	1,2023-06-04,2023-07,2,0							
4	1,2023-06-04,2023-08,4,0							
5	1,2023-06-04,2023-09,2,0							
6	1,2023-06-04,2023-10,4,0							
7	1,2023-06-04,2023-11,3,0							
8	1,2023-06-04,2023-12,5,0							
9	2,2023-02-09,2023-02,6,1							
10	2,2023-02-09,2023-03,4,0							
11	2,2023-02-09,2023-04,2,0							
12	2,2023-02-09,2023-05,5,1							
13	2,2023-02-09,2023-06,4,0							
14	2,2023-02-09,2023-07,4,0							
15	2,2023-02-09,2023-08,4,0							
16	2,2023-02-09,2023-09,2,0							
17	2,2023-02-09,2023-10,2,0							
18	2,2023-02-09,2023-11,0,0							
19	2,2023-02-09,2023-12,4,0							
20	3,2023-09-24,2023-09,7,1							
21	3,2023-09-24,2023-10,1,0							

Рис. 1. Фрагмент таблицы

В исследовании для проведения когортного анализа был написан Python-скрипт, который обрабатывает синтетически сгенерированные данные и рассчитывает ключевые метрики пользовательской активности. Работа начинается с загрузки набора данных из CSV-файла, содержащего информацию о пользователях, датах их регистрации, активности по месяцам, количестве посещений и совершённых покупок. На этапе предобработки данные о датах регистрации и месяцах активности преобразуются в формат `datetime`, а затем – в формат периодов (`Period`) по месяцам (рис. 2, 3).

```

# Загрузка данных
df = pd.read_csv("cohort_data.csv", encoding="utf-8-sig")

# Преобразование дат
df["RegistrationDate"] = pd.to_datetime(df["RegistrationDate"])
df["Cohort"] = df["RegistrationDate"].dt.to_period("M")
df["ActivityMonth"] = pd.to_datetime(df["ActivityMonth"], format="%Y-%m").dt.to_period("M")

# Расчет количества месяцев с момента регистрации (cohort index)
df["CohortIndex"] = (df["ActivityMonth"].dt.year - df["Cohort"].dt.year) * 12 + (df["ActivityMonth"].dt.month - df["Cohort"].dt.month) + 1

# Вычисление retention: уникальные пользователи в каждом месяце для каждой когорты
cohort_data = df.groupby(["Cohort", "CohortIndex"])["UserID"].nunique().unstack(1)

# Размер когорты (кол-во уникальных пользователей в 1-й месяц)
cohort_sizes = cohort_data.iloc[:, 0]
retention = cohort_data.divide(cohort_sizes, axis=0).round(3) * 100 # в процентах

# Расчет средних метрик
avg_visits = df.groupby("Cohort")["Visits"].mean().round(2)
avg_purchases = df.groupby("Cohort")["Purchases"].mean().round(2)
avg_metrics = pd.concat([avg_visits, avg_purchases], axis=1)
avg_metrics.columns = ["Visits", "Purchases"]

(retention, avg_metrics)

```

Рис. 2. Фрагмент кода обработки данных

(CohortIndex	1	2	3	4	5	6	7	8	9	\
Cohort										
2023-01	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
2023-02	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
2023-03	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
2023-04	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
2023-05	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	NaN
2023-06	100.0	100.0	100.0	100.0	100.0	100.0	100.0	NaN	NaN	NaN
2023-07	100.0	100.0	100.0	100.0	100.0	100.0	NaN	NaN	NaN	NaN
2023-08	100.0	100.0	100.0	100.0	100.0	NaN	NaN	NaN	NaN	NaN
2023-09	100.0	100.0	100.0	100.0	NaN	NaN	NaN	NaN	NaN	NaN
2023-10	100.0	100.0	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-11	100.0	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-12	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

CohortIndex	10	11	12
Cohort			
2023-01	100.0	100.0	100.0
2023-02	100.0	100.0	NaN
2023-03	100.0	NaN	NaN
2023-04	NaN	NaN	NaN
2023-05	NaN	NaN	NaN
2023-06	NaN	NaN	NaN
2023-07	NaN	NaN	NaN
2023-08	NaN	NaN	NaN
2023-09	NaN	NaN	NaN
2023-10	NaN	NaN	NaN
2023-11	NaN	NaN	NaN
2023-12	NaN	NaN	NaN

	Visits	Purchases
Cohort		
2023-01	2.82	0.11
2023-02	3.10	0.13
2023-03	3.14	0.14
2023-04	3.24	0.16
2023-05	3.43	0.17
2023-06	3.47	0.18
2023-07	3.57	0.20
2023-08	3.76	0.22
2023-09	3.84	0.23
2023-10	4.19	0.34
2023-11	4.00	0.28
2023-12	4.38	0.34)

Рис. 3. Результат кода обработки данных

Следующим шагом стало построение когорт и расчёт показателя retention rate, отражающего долю пользователей, оставшихся активными в каждом последующем месяце после регистрации. Для этого каждому пользователю была присвоена когорта – месяц его регистрации, а также

рассчитан индекс месяца активности относительно даты регистрации (период пребывания).

Для визуального анализа была построена тепловая карта retention rate (рис 4). Каждая строка на карте представляет когорту, а каждый столбец – номер месяца с момента регистрации. Значения в ячейках отражают процент пользователей, остававшихся активными в соответствующий месяц. В рамках данного синтетического набора данные по всем когортам демонстрируют стабильный 100% уровень удержания, что является следствием заданного алгоритма генерации, предусматривающего непрерывную активность до конца 2023 года.

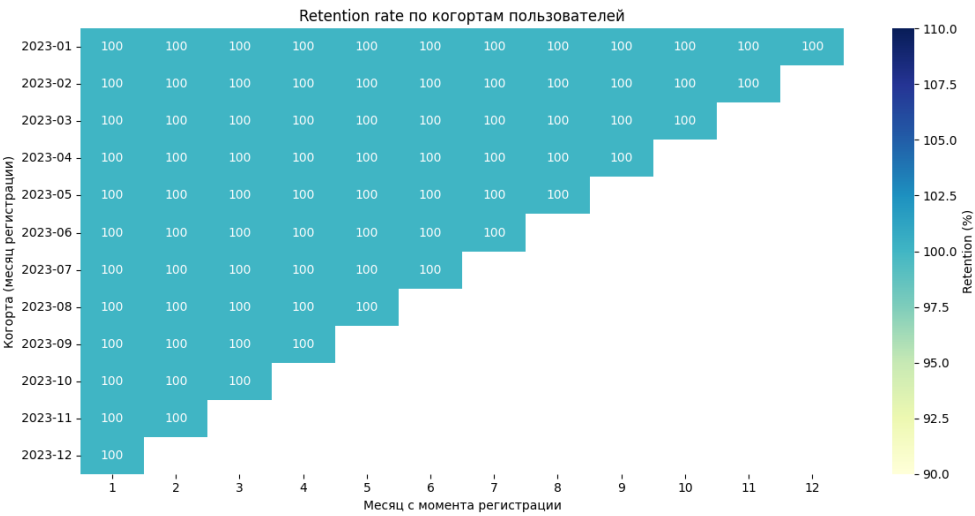


Рис. 4 Тепловая карта retention rate

Также был рассчитан средний уровень активности в разрезе когорт – как по посещениям, так и по покупкам. Полученные результаты визуализированы в виде гистограммы (рис. 5). Как видно, в более поздних когортах (например, ноябрь и декабрь 2023 года) характерен более высокий средний уровень активности, что может объясняться меньшим количеством месяцев наблюдения и отсутствием выраженного эффекта «усталости пользователя».



Рис. 5. Гистограмма активности

Таким образом, когортный анализ на основе синтетических данных позволяет выявить условные поведенческие паттерны, а визуальные инструменты – эффективно интерпретировать результаты.

3 Выводы

В рамках исследования был проведён когортный анализ синтетических данных с использованием Python. Сгенерированный набор охватывал пользователей 2023 года и отражал их месячную активность. С помощью разработанного кода были рассчитаны показатели удержания, а также средние значения посещений и покупок по когортам. Результаты визуализированы в виде тепловой карты и графика, что позволило наглядно оценить поведенческие различия между группами. Несмотря на искусственный характер данных, анализ подтвердил эффективность когортного подхода для выявления динамики активности пользователей и формирования маркетинговых стратегий.

Биографический список

1. Паршина Е.И. Когортный анализ как одна из основ маркетинговой аналитики бизнеса // Электронная наука. 2023. Т. 4. № 1.
2. Черник А.А. Использование когортного анализа как основного приема юнит-экономики в оценке эффективности деятельности организации // Шаг в будущее: искусственный интеллект и цифровая экономика. Техноэкономика: трансформация платформ: Сборник научных статей Международного научного форума. В 3-х томах. Москва, 2023. С. 253–259.
3. Кондакова Н.А. Оценка здоровья детей Вологодской области за 1995–2019 гг.: метод когортного анализа // Социальные аспекты здоровья населения. 2019. Т. 65. № 6. С. 5.
4. Терещенко О.В. Метод когортного анализа в социальных исследованиях // Социология: методология, методы, математическое моделирование. 2009. № 29. С. 172–185.