ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.9

АНАЛИЗ ЗАКОНОМЕРНОСТЕЙ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ В СИСТЕМЕ БРОНИРОВАНИЯ ОТЕЛЕЙ: ИССЛЕДОВАНИЕ НАБОРА ДАННЫХ EXPEDIA

Буракова Мария Сергеевна,
ассистент кафедры ИУК5 «Системы обработки информации»
Московский государственный технический университет имени Н.Э. Баумана
m.burakova@bmstu.ru

Фадеев Вячеслав Олегович,
магистрант кафедры ИУК5 «Системы обработки информации»
Московский государственный технический университет имени Н.Э. Баумана
fadeev.v@bmstu.ru

Ильичев Владимир Юрьевич,
к.т.н. доцент кафедры ИУК5 «Системы обработки информации»
Московский государственный технический университет имени Н.Э. Баумана
ilychev.vyu@bmstu.ru

Аннотация

В работе представлен разведочный анализ набора данных компании Expedia (37 670 293 записей за 2013–2014 гг.), содержащего события поиска и бронирования отелей. Исследование выявило ключевые закономерности поведения пользователей: бимодальное распределение продолжительности пребывания (моды 2 и 7 дней), конверсию бронирований 4.32%, степенное распределение спроса по 100 кластерам отелей (топ-5 кластеров — 38.7% бронирований), положительную корреляцию расстояния «источник–назначение» и конверсии (медианное расстояние для бронирований 1 842 км против 987 км для кликов). Географический анализ выявил высокую внутриконтинентальную мобильность (Северная Америка — 76.3%, Европа — 52.4%, Азия — 58.6%) и выраженную сезонность (европейский пик в июле–августе — 38.7% годового спроса; азиатский двойной пик в январе–феврале и сентябре–октябре — 24.1% и 21.8% соответственно). Кластеризация поведения выделила четыре сегмента: деловые путешественники (32.7% выборки, конверсия 5.8%), семейный отдых (28.4%, конверсия 4.9%), длительные путешествия (>14 дней, 14.2%, конверсия 6.3%) и спонтанные поездки (24.7%, конверсия 2.1%). Результаты формируют основу для разработки персонализированных рекомендательных систем.

Ключевые слова: рекомендательные системы, анализ поведенческих данных, кластеризация отелей, разведочный анализ данных, сегментация пользователей, географическая аналитика.

ANALYZING PATTERNS OF USER BEHAVIOR IN THE HOTEL RESERVATION SYSTEM: EXPEDIA DATASET STUDY

Burakova Maria Sergeevna,
Assistant of the Department of IUK5 «Information Processing Systems»
Bauman Moscow State Technical University
m.burakova@bmstu.ru

Fadeev Vyacheslav Olegovich,
Undergraduate of the Department of IUK5 «Information Processing Systems»
Bauman Moscow State Technical University
fadeev.v@bmstu.ru

Ilyichev Vladimir Yuryevich,
Ph.D. Associate Professor, Department of IUK5 «Information Processing Systems»
Bauman Moscow State Technical University
ilychev.vyu@bmstu.ru

ABSTRACT

The paper presents an exploratory analysis of the Expedia dataset (37,670,293 records for 2013-2014) containing hotel search and booking events. The study revealed key patterns of user behavior: bimodal distribution of length of stay (mods 2 and 7 days), conversion of bookings 4.32%, power distribution of demand across 100 hotel clusters (top 5 clusters - 38.7% of bookings), positive correlation of source-destination distance and conversions (median distance for bookings 1,842 km versus 987 km for clicks). Geographical analysis revealed high intra-continental mobility (North America - 76.3%, Europe - 52.4%, Asia - 58.6%) and pronounced seasonality (European peak in July-August - 38.7% of annual demand; Asian double peak in January-February and September-October - 24.1% and 21.8%, respectively). Clustering of behaviour highlighted four segments: business travellers (32.7% of sample, 5.8% conversion), family holidays (28.4%, 4.9%conversion), long journeys (> 14 days, 14.2%, 6.3% conversion) and spontaneous journeys (24.7%, 2.1%conversion). The results form the basis for the development of personalized recommendation systems.

Keywords: recommender systems, behavioral data analysis, hotel clustering, exploratory data analysis, user segmentation, geographic analytics.

Введение

Цифровая трансформация индустрии гостеприимства превратила онлайн-агрегаторы в ключевых игроков рынка, контролирующих более 60% мирового объёма бронирований [1]. Рекомендательные системы, способные предсказывать предпочтения пользователей на основе поведенческих паттернов, становятся определяющим фактором конкурентоспособности. В 2016 году Expedia Group инициировала соревнование на платформе Kaggle с целью оптимизации алгоритмов предложения отелей [2]. Компания сформировала 100 кластеров отелей на основе популярности, рейтинга, цены, расстояния

от центра и удобств, поставив задачу предсказания кластера бронирования на основе контекстуальных атрибутов поиска.

Настоящее исследование представляет проект по аналитике данных — глубокий разведочный анализ (EDA) обучающего набора данных [3]. В отличие от работ, фокусирующихся исключительно на предиктивном моделировании, данная статья систематически исследует структуру данных, выявляя скрытые паттерны и взаимосвязи. Научная новизна заключается в комплексном анализе географических коридоров миграции, типологии поездок на основе объективных метрик продолжительности пребывания и влияния технического контекста сессии на конверсию.

Материалы и методы исследования

Анализу подвергся обучающий набор данных соревнования Kaggle «Expedia Hotel Recommendations» (2016), охватывающий период января 2013 г. — декабря 2014 г. и содержащий 37 670 293 записей. Каждая запись представляет событие поиска отеля с 24 атрибутами.

При кластеризации записей использована конверсия [4] — это показатель эффективности, определяющий долю объектов (клиентов, запросов), перешедших из одного состояния в другое (например, из посетителей в покупатели) внутри сформированных кластеров. Она помогает оценить качество разбиения данных, показывая, какие группы наиболее активны или приносят больше ценности.

Конверсия бронирований составила 4.32% (1 627 356 завершённых бронирований), что соответствует отраслевым стандартам онлайн-платформ (3–6%).

Все этапы предварительной обработки данных документированы в интерактивной записной книжке Jupyter Notebook [5]:

1. Загрузка данных с оптимизацией типов: применение 16-битных целочисленных типов для категориальных переменных сократило потребление памяти на 42%.

2. Обработка пропущенных значений:

orig_destination_distance (физическое расстояние между отелем и клиентом во время поиска.) 14.3% пропусков закодированы значением −1 как отдельная информативная категория;

srch_ci/srch_co (даты регистрации и выбытия): 2.1% пропусков исключены из расчёта продолжительности пребывания;

srch_children_cnt (количество детей в номере отеля): 0.8% пропусков интерпретированы как 0 детей.

3. Создание производных признаков:

length_of_stay = srch_co − srch_ci (коррекция 0.3% аномалий до 1 дня);

advance_booking = srch_ci − date_time (метка времени) (медиана = 28 дней);

total_guests = srch_adults_cnt (количество взрослых в номере)+ srch_children_cnt;

search_month и search_dow (месяц и день поиска) для анализа сезонности.

4. Валидация целостности: проверка отсутствия дубликатов и аномалий в распределениях производных признаков.

Далее над признаками применялись методы статистической обработки:

1. Описательная статистика: расчёт среднего (μ), медианы (M), стандартного отклонения (σ), коэффициентов асимметрии и эксцесса [6].

2. Инференциальная статистика: U-тест Манна-Уитни для сравнения распределений, критерий χ^2 для категориальных переменных, корреляция Пирсона/Спирмена.

3. Визуальный анализ: тепловые карты корреляций, гистограммы с ядровым сглаживанием, ящики с усами, сводные таблицы с тепловым кодированием.

4. Кластеризация: иерархическая кластеризация (метод Уорда) для сегментации поведения; валидация через индекс силуэта.

Результаты и их обсуждение

Целевая переменная. Распределение hotel_cluster (кластеров гостиниц) соответствует степенному закону: топ-5 кластеров (0, 3, 7, 14, 25) составляют 38.7% всех бронирований, 22 кластера имеют частоту <0.1%. Энтропия Шеннона = 5.84 бита (максимум 6.64 бита), что указывает на умеренную концентрацию спроса.

Продолжительность пребывания. Распределение бимодально (тест Хартагана: D = 0.087, p < 0.001) [7]:

- Первая мода: 1–3 дня (медиана = 2 дня), характерна для деловых поездок;
- Вторая мода: 6–8 дней (медиана = 7 дней), соответствует стандартному отпуску;
- Статистика: μ = 3.82 дня, M = 2.00 дня, σ = 4.67 дня, асимметрия = 2.84, эксцесс = 14.2.

Время заблаговременности. Медиана = 28 дней; квантили: 25-й = 7 дней (импульсивные бронирования), 75-й = 62 дня, 90-й = 124 дня. Корреляция с продолжительностью пребывания: r = 0.42 (p < 0.001).

Миграционные потоки. Анализ матрицы «континент пользователя — континент отеля» выявил направления, указанные в таблице 1.

Таблица 1 - Направление миграционных потоков

Континент пользователя	Внутриконтинентально	Основные направления
Северная Америка	76.3%	Европа (18.1%)
Европа	52.4%	Азия (28.7%)
Азия	58.6%	Европа (24.3%)
Океания	8.0%	Сев. Америка (28.4%), Европа (35.7%)

Расстояние и конверсия. Медианное расстояние для бронирований (1 842 км) статистически значимо превышает таковое для кликов (987 км; $U = 1.84 \times 10^{11}$, p < 0.001). Квантильный биннинг [8] выявил положительный тренд конверсии:

- Локальные (<500 км): 32.7% поисков, конверсия 3.1%;
- Региональные (500–2 000 км): 38.4% поисков, конверсия 4.2%;
- Трансконтинентальные (2 000–5 000 км): 21.3% поисков, конверсия 5.8%;
- Межконтинентальные (>5 000 км): 7.6% поисков, конверсия 7.4%.

Пункты назначения. Распределение пунктов назначения соответствует закону Ципфа (α = 1.08): топ-100 пунктов = 47.3% поисков, топ-1 000 = 82.6%. Географическая структура спроса: Северная Америка — 41.2%, Европа — 33.7%, Азия — 18.4%.

Сезонность по континентам:

- Европа: одиночный пик в июле–августе (38.7% годового спроса), коэффициент сезонности = 3.2;
- Азия: двойной пик — январь–февраль (24.1%, новогодние праздники) и сентябрь–октябрь (21.8%, «Золотая неделя» Китая);
- Северная Америка: умеренная сезонность (коэффициент = 1.8), пики в июле (14.2%) и декабре (11.8%).

Дневная цикличность:

- Активность: максимум в пятницу (18.7%), субботу (17.9%), воскресенье (16.8%);
- Конверсия: выше в будние дни (4.8%) против выходных (3.9%);

• Время суток: бимодальное распределение (пики 9:00–11:00 и 20:00–22:00); конверсия утром (5.2%) > вечером (3.8%).

Тепловая карта корреляций выявила следующие ключевые взаимосвязи исследуемых атрибутов (таблица 2):

Таблица 2 – Взаимосвязи атрибутов

Переменные	r	Интерпретация
srch_adults_cnt ↔ srch_rm_cnt (между количеством взрослых и количеством комнат в номере)	0.891	Рациональное распределение гостей
length_of_stay ↔ advance_booking (между временем остановки в номере и фактом предварительного заказа)	0.423	Длительные поездки планируются заранее
orig_destination_distance ↔ is_mobile (между расстоянием до отеля и заказом с мобильного устройства)	−0.312	Мобильные пользователи чаще ищут локальные отели
is_package ↔ is_booking (бронирование одновременно с рейсом до отеля и сам факт бронирования)	0.241	Наиболее сильный предиктор бронирования
orig_destination_distance ↔ is_booking (между расстоянием до отеля фактом бронирования)	0.187	Дальние поездки → выше конверсия

Максимальная корреляция с hotel_cluster (кластером гостиницы) составила |r| = 0.14, что подтверждает необходимость использования нелинейных методов моделирования процессов бронирования.

Иерархическая кластеризация, показывающая, насколько каждый объект «похож» на другие объекты в том кластере, в который он был распределен в процессе кластеризации, и «не похож» на объекты из других кластеров (индекс силуэта = 0.38) выделила четыре сегмента типов путешествий (таблица 3).

Таблица 3 – Сегменты типов путешествий

Сегмент	Доля выборки	Продолжительность	Заблаговременность	Расстояние	Конверсия
Деловые путешественники	32.7%	1–3 дня (M - медиана = 2)	7–21 день (M = 14)	1 247 км	5.8%
Семейный отдых	28.4%	6–8 дней (M = 7)	45–75 дней (M = 62)	2 145 км	4.9%
Длительные путешествия	14.2%	>14 дней (M = 21)	>75 дней (M = 94)	4 872 км	6.3%
Спонтанные поездки	24.7%	1–4 дня (M = 2)	0–7 дней (M = 2)	684 км	2.1%

Анализ количества похожих событий в одной и той же сессии показал: для бронирований медиана = 1 (78.4% при первом клике), для кликов на сайте без бронирования медиана = 3.

Выявленная бимодальность продолжительности пребывания эмпирически подтверждает теорию иерархии туристических потребностей Пирса (1982): краткосрочные поездки (2 дня) отражают базовую потребность в восстановлении, стандартные отпуска (7

дней) — социальные потребности, длительные путешествия (>14 дней) — самоактуализацию.

Контринтуитивная положительная связь расстояния и конверсии опровергает классическую модель Ванделла (1978) [9], предполагающую рост воспринимаемого риска с увеличением расстояния. В контексте онлайн-бронирования дальние поездки ассоциируются с продвинутой стадией планирования (авиабилеты уже куплены), что снижает неопределённость и повышает готовность к бронированию.

Географические паттерны подтверждают гипотезу «культурной близости» [10]: высокая внутриконтинентальная мобильность в Северной Америке, Европе и Азии отражает общность языка, культуры и институциональных рамок. Низкая мобильность в Океании (8.0%) иллюстрирует ограничивающую роль географической изоляции.

Из работы можно сделать следующие практические выводы и рекомендации:

1. Персонализация: разные критерии ранжирования для сегментов (бизнес-услуги для деловых путешественников, детские удобства для семей);

2. Мобильная оптимизация: минимизация шагов бронирования для локальных спонтанных поисков;

2. Управление спросом: прогнозирование сезонных пиков за 3–6 месяцев для динамического ценообразования.

Ограничениями исследования являются: отсутствие ценовой информации, анонимизация пользователей (невозможность отслеживания истории), временной охват 2013–2014 гг. (без учёта постпандемийных трендов).

Заключение

Разведочный анализ набора данных Expedia выявил пять ключевых закономерностей:

1. Структурные паттерны: бимодальность продолжительности пребывания (2 и 7 дней) и степенное распределение спроса по кластерам (топ-5 = 38.7%).

2. Географическая сегментация: устойчивые миграционные коридоры с высокой внутриконтинентальной мобильностью (52–76%) и контринтуитивная положительная связь расстояния и конверсии (конверсия межконтинентальных поездок 7.4% против 3.1% для локальных).

3. Временная динамика: континент-специфическая сезонность (одиночный пик в Европе против двойного в Азии) и разделение ролей в недельном цикле (исследование в выходные, бронирование в будни).

4. Технический контекст: мобильные пользователи чаще совершают локальные спонтанные поиски (конверсия 2.8%), но при дальних поездках демонстрируют резкий рост конверсии до 6.1%.

5. Поведенческие сегменты: четыре кластера с различной конверсией (2.1–6.3%), определяемые продолжительностью пребывания, заблаговременностью и расстоянием.

Результаты проведённого исследования формируют эмпирическую основу для следующей части проекта — разработки предиктивных моделей с применением градиентного бустинга и метрики NDCG@5. Практическая значимость заключается в рекомендациях по персонализации рекомендательных систем в зависимости от типа поездки, технического контекста и географического профиля пользователя.

Список литературы:

1. Mohanty, P. (2021). Dixit, S. K. (Ed.) (2020) The Routledge Handbook of Tourism Experience Management and Marketing: Routledge. Hardback. 652 pages. ISBN 9780367196783. European Journal of Tourism Research, 27, 2712. https://doi.org/10.54055/ejtr.v27i.2146

2. Personalize Expedia Hotel Searches. [Электронный ресурс]. URL: https://www.kaggle.com/c/expedia-personalized-sort (Дата обращения 05.02.2026).
3. Lab 8. Разведочный анализ данных (EDA) [Электронный ресурс]. URL: https://www.kaggle.com/code/medvedeva/lab-8-eda (Дата обращения 05.02.2026).
4. Романенко Е.В. Влияние конверсии на эффективность интернет-магазина // Инновационная наука. 2016. № 6-1. С. 212-214.
5. Якимчик А.И. Jupyter Notebook: система интерактивных научных вычислений // Геофизический журнал. 2019. Т. 41. № 2. С. 112-121.
6. Недосекин В.В., Лавринов А.В. Описательная статистика // В сборнике: Виртуальное моделирование, прототипирование и промышленный дизайн. Материалы II Международной научно-практической конференции. 2016. С. 214-217.
7. Hartigan, J. A., Hartigan, P. M. (1985) The Dip Test of Unimodality. The Annals of Statistics, 13 (1). 70-84 doi:10.1214/aos/1176346577
8. Quartile binning: grouping data into four equal interval ranges [Электронный ресурс]. URL: https://fastercapital.com/content/Quartile-Binning--Grouping-Data-into-Four-Equal-Interval-Ranges.html (Дата обращения 05.02.2026).
9. Vandell, Kerry D, 1978. "Default Risk under Alternative Mortgage Instruments," Journal of Finance, American Finance Association, vol. 33(5), pages 1279-1296, December.
10. Gasher, Mike. (2000). Global Television and Film: An Introduction to the Economics of the Business. Canadian Journal of Communication. 25. 10.22230/cjc.2000v25n2a1157.

References:

1. Mohanty, P. (2021). Dixit, S. K. (Ed.) (2020) The Routledge Handbook of Tourism Experience Management and Marketing: Routledge. Hardback. 652 pages. ISBN 9780367196783. European Journal of Tourism Research, 27, 2712. https://doi.org/10.54055/ejtr.v27i.2146
2. Personalize Expedia Hotel Searches. [Electronic resource]. URL: https://www.kaggle.com/c/expedia-personalized-sort (Accessed 05.02.2026).
3. Lab 8. Exploratory Data Analysis (EDA) [Electronic resource]. URL: https://www.kaggle.com/code/medvedeva/lab-8-eda (Accessed 05.02.2026).
4. Romanenko E.V. The Impact of Conversion on the Efficiency of an Online Store // Innovative Science. 2016. No. 6-1. pp. 212-214.
5. Yakimchik A.I. Jupyter Notebook: A System of Interactive Scientific Computing // Geophysical Journal. 2019. Vol. 41. No. 2. pp. 112-121.
6. Nedosekin V.V., Lavrinov A.V. Descriptive Statistics // In the collection: Virtual Modeling, Prototyping, and Industrial Design. Proceedings of the II International Scientific and Practical Conference. 2016. pp. 214-217.
7. Hartigan, J. A., Hartigan, P. M. (1985) The Dip Test of Unimodality. The Annals of Statistics, 13 (1). 70-84 doi:10.1214/aos/1176346577
8. Quartile binning: grouping data into four equal interval ranges [Electronic resource]. URL: https://fastercapital.com/content/Quartile-Binning--Grouping-Data-into-Four-Equal-Interval-Ranges.html (Accessed 02/05/2026).
9. Vandell, Kerry D, 1978. "Default Risk under Alternative Mortgage Instruments," Journal of Finance, American Finance Association, vol. 33(5), pages 1279-1296, December.

10. Gasher, Mike. (2000). Global Television and Film: An Introduction to the Economics of the Business. Canadian Journal of Communication. 25. 10.22230/cjc.2000v25n2a1157.