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

УДК 004.9

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ АЛГОРИТМОВ ДЛЯ БРОНИРОВАНИЯ НОМЕРОВ В ГОСТИНИЦАХ

Иванов Никита Владимирович,
Студент кафедры ИУК5 «Системы обработки информации»
Московский государственный технический университет имени Н.Э. Баумана
destira@mail.ru

Косова Ксения Алексеевна,
Студентка кафедры ИУК5 «Системы обработки информации»
Московский государственный технический университет имени Н.Э. Баумана
ksenya.kosova.04@mail.ru

Вершинин Евгений Владимирович,
К.ф.-м.н., доцент, заведующий кафедрой ИУК5 «Системы обработки информации»
Московский государственный технический университет имени Н.Э. Баумана
vershinin@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 статье описано исследование, целью которого являлась разработка методики взаимодействия с базой данных (открытым датасетом) по запросам арендаторов номеров в гостиницах с целью прогнозирования вероятности бронирования. Программа реализована в виде интеллектуального алгоритма на языке Python с использованием библиотек Pandas, Statsmodels, Scikit-learn, XGBoost, LightGBM, CatBoost. Проведён корреляционный анализ признаков и построена модель логистической регрессии. Реализованы модели градиентного бустинга и несколько других, способные предсказывать вероятность бронирования для каждого отеля. Произведено сравнение метрик качества предсказаний для каждой модели и выбран алгоритм, дающий наибольшую точность прогнозирования вероятности бронирования номеров.

Ключевые слова: персонализированное бронирование, предиктивная аналитика, машинное обучение, Python, Pandas, Statsmodels, Scikit-learn, XGBoost, LightGBM, learning-to-rank.

APPLYING INTELLIGENT ALGORITHMS TO HOTEL ROOM RESERVATIONS

Ivanov Nikita Vladimirovich,
Student of the Department of IUK5 «Information Processing Systems»
Bauman Moscow State Technical University
destira@mail.ru

Kosova Ksenia Alekseevna,
Student of the Department of IUK5 «Information Processing Systems»
Bauman Moscow State Technical University
ksenya.kosova.04@mail.ru

Vershinin Evgeny Vladimirovich,
Ph.D., Associate Professor, Head of the Department of IUK5 «Information Processing Systems»
Bauman Moscow State Technical University
vershinin@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 article describes a study aimed at developing a methodology for interacting with a database (open dataset) at the request of hotel room tenants in order to predict the probability of booking. The program is implemented as an intelligent algorithm in Python using the Pandas, Statsmodels, Scikit-learn, XGBoost, LightGBM, CatBoost libraries. A correlation analysis of the features was performed and a logistic regression model was constructed. Gradient boosting models and several others have been implemented that can predict the probability of booking for each hotel. Prediction quality metrics were compared for each model and an algorithm was selected that gives the highest accuracy in predicting the probability of booking rooms.

Keywords: personalized booking, predictive analytics, machine learning, Python, Pandas, Statsmodels, Scikit-learn, XGBoost, LightGBM, learning-to-rank.

Введение

В современных системах онлайн-бронирования гостиниц традиционно применяются статистические методы и базовые алгоритмы машинного обучения. Классические подходы часто ограничены в точности прогнозирования пользовательских предпочтений [1]. Существующие решения обычно фокусируются либо на задаче классификации, либо на ранжировании, но не объединяют эти подходы комплексно [2].

Актуальность исследования обусловлена ростом конкуренции на рынке онлайн-бронирований и необходимостью повышения точности персонализированных рекомендаций. Современные платформы требуют интеллектуальных систем, способных не только предсказывать вероятность бронирования, но и оптимально ранжировать предложения для каждого пользователя.

Цель исследования — разработка комплексной методики прогнозирования бронирований гостиниц, сочетающей задачи классификации и ранжирования с использованием современных алгоритмов машинного обучения, а также сравнение точности предсказаний вероятности бронирования с использованием разных интеллектуальных моделей.

Материалы и методы исследования

В работе использованы программные технологии Python с библиотеками Pandas для обработки данных [3], Scikit-learn для базовых алгоритмов, XGBoost, LightGBM и CatBoost для градиентного бустинга [4].

В основу исследования положен датасет Expedia Personalized Sort [5], содержащий информацию о поисковых запросах пользователей и их взаимодействии с результатами поиска.

Вначале была осуществлена многоэтапная предварительная обработка исходных данных, включающая удаление аномальных значений, фильтрацию выбросов и создание производных признаков для улучшения предсказательной способности моделей.

Затем была построена матрица корреляций (рис. 1) признаков [6] и для дальнейшей обработки выбраны наиболее коррелирующие (price_usd (отображаемая стоимость отеля для данного поискового запроса) и prop_starrating (звездность отеля по рейтингу, от 1 до 5); srch_length_of_stay (количество ночей проживания) и srch_booking_window (количество дней между датой поиска и датой заезда).

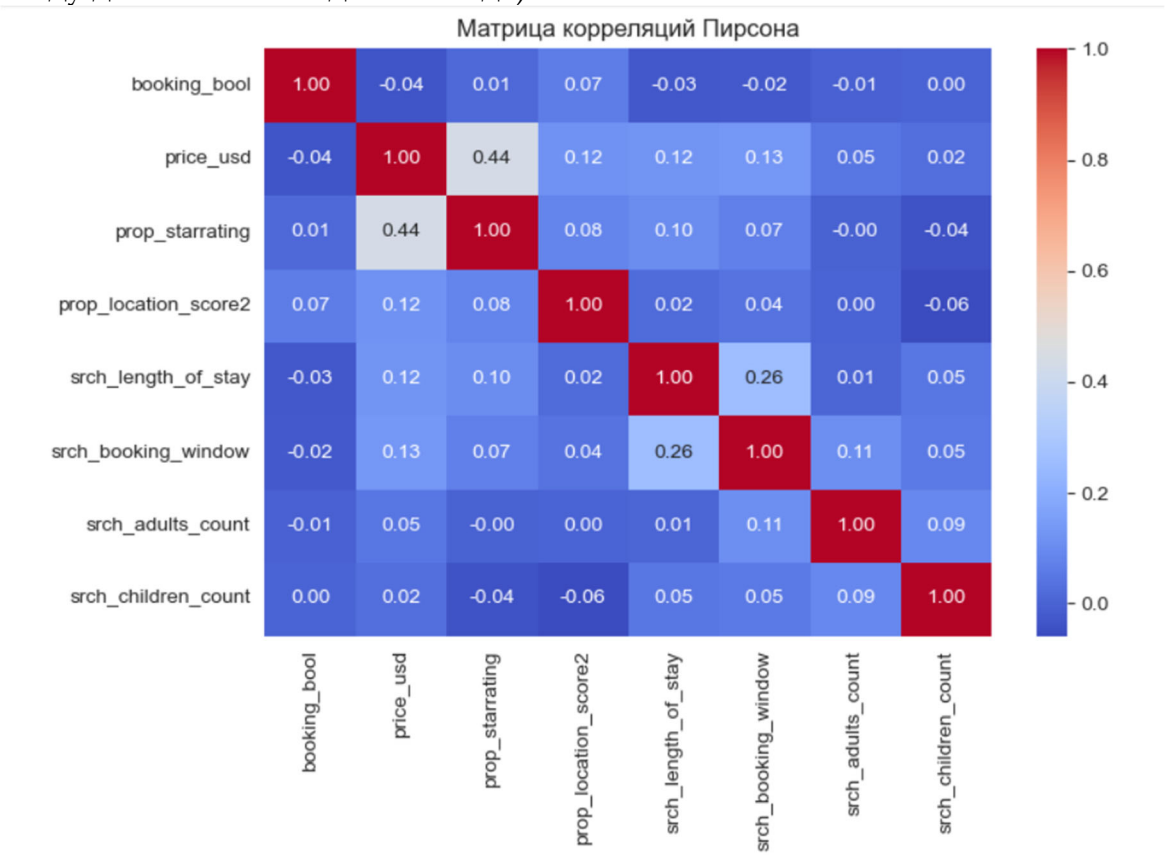

Рисунок 1. Матрица корреляций

Разработана программная система на Python, реализующая шесть различных алгоритмов машинного обучения. Каждый метод был выбран исходя из его теоретических преимуществ для решения задач классификации и ранжирования. Рассмотрим вкратце эти методы.

Логистическая регрессия - статистический метод классификации, основанный на логит-функции. Модель оценивает вероятность принадлежности к классу через линейную

комбинацию признаков с последующим применением сигмоидной функции. Основное преимущество - высокая интерпретируемость коэффициентов, что позволяет анализировать влияние каждого признака на целевую переменную. В работе использована L2-регуляризация для предотвращения переобучения.

Случайный лес - ансамблевый алгоритм, строящий множество решающих деревьев на различных подвыборках данных с использованием бутстрэп-агрегирования. Каждое дерево обучается на случайном подмножестве признаков, что повышает разнообразие ансамбля. Итоговое предсказание формируется путем голосования деревьев. Метод устойчив к переобучению и шуму в данных, эффективно обрабатывает нелинейные зависимости.

XGBoost (Extreme Gradient Boosting) - продвинутая реализация градиентного бустинга, характеризующаяся добавлением регуляризации в функцию потерь. Алгоритм последовательно строит деревья, где каждое последующее дерево минимизирует ошибки предыдущих. Особенности включают встроенную обработку пропущенных значений, поддержку параллельных вычислений и автоматическую кросс-валидацию. Оптимизирован для работы с большими объемами данных.

LightGBM - оптимизированная версия градиентного бустинга, использующая гистограммный метод для ускорения обучения и уменьшения потребления памяти. Алгоритм применяет Exclusive Feature Bundling для комбинации разреженных признаков и Gradient-based One-Side Sampling для выборки объектов с большими градиентами. Особенно эффективен при работе с категориальными признаками и большими наборами данных.

CatBoost - алгоритм градиентного бустинга, специализирующийся на работе с категориальными признаками. Использует Ordered Boosting для борьбы со смещением предсказаний и автоматически обрабатывает категориальные переменные без предварительного кодирования. Отличается устойчивостью к переобучению и высокой точностью на табличных данных с смешанными типами признаков.

Для решения только задачи классификации использован метод LightGBM Ranker — это расширение градиентного бустинга LightGBM, специально предназначенное для задач ранжирования. Метод ориентирован на обучение моделей, способных эффективно упорядочивать элементы согласно заданным критериям.

Для обучения моделей использовалось стратифицированное разделение данных на обучающую (80%) и тестовую (20%) выборки. Гиперпараметры моделей настраивались с помощью перекрестной проверки, для бустинговых алгоритмов применялось ранняя остановка для предотвращения переобучения. Все эксперименты проводились на одинаковых данных для обеспечения корректности сравнения.

Результаты и их обсуждение

Экспериментальное исследование проводилось на выборке 780 483 записей после предварительной обработки данных. Сравнительный анализ шести алгоритмов машинного обучения выявил существенные различия в их эффективности.

Новизна разработки заключается в комбинированном подходе, объединяющем задачу бинарной классификации (бронирование/не бронирование) и задачу ранжирования отелей внутри поисковой сессии. Предложена оригинальная методика оценки, использующая две группы метрик: AUC-ROC для оценки качества классификации и NDCG@5 для оценки качества ранжирования, что соответствует подходу использования данных и алгоритмов для прогнозирования будущих событий [7].

Метрика AUC-ROC (Area Under the Curve Receiver Operating Characteristic - «Площадь под кривой ошибок операционного показателя»). Это мера качества бинарной классификации модели машинного обучения. Она показывает способность модели

различать положительные и отрицательные классы независимо от порога принятия решения. Чем ближе значение AUC-ROC к единице, тем лучше модель классифицирует объекты.

Метрика NDCG@k (Normalized Discounted Cumulative Gain at k –«Нормализованный накопленный выигрыш»), используется для оценки ранжирования списков рекомендаций или поисковых результатов. Она учитывает порядок выдачи объектов и важность каждого элемента в списке. Значение k обозначает глубину списка, до которой оценивается качество (например, @5 означает первые пять позиций). Метрика рассчитывает общую ценность предложенных элементов с учетом их места в выдаче и нормализует её относительно наилучшего возможного результата.

Значения полученных метрик AUC и NDCG@5 для всех рассмотренных методов предиктивной аналитики показаны на рис. 2.

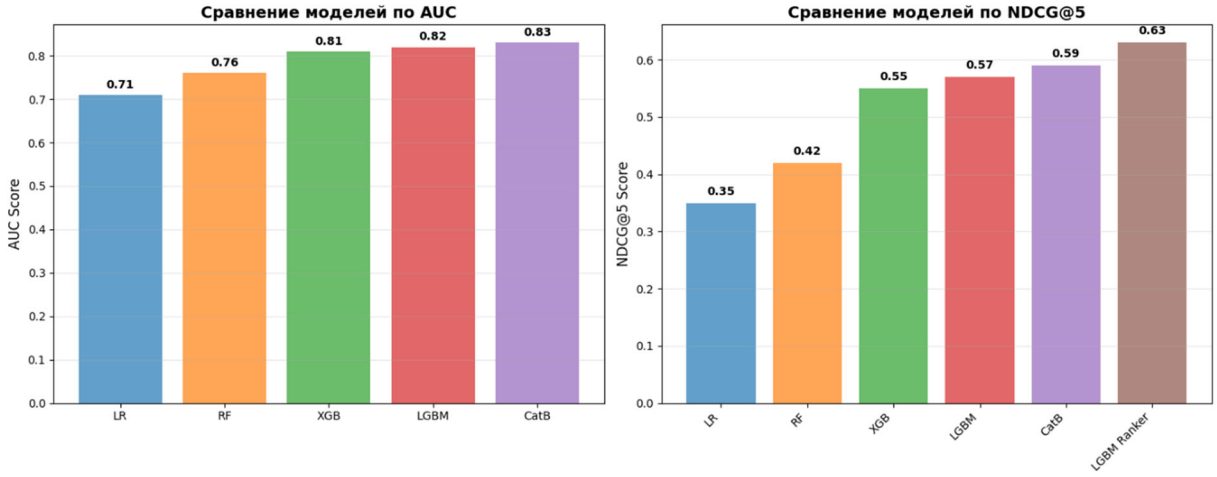

Рисунок 2. Зависимость значений полученных метрик от типа модели

Логистическая регрессия показала AUC = 0,71 и NDCG@5 = 0,35. Эти результаты соответствуют ожиданиям для линейной модели, которая хотя и обеспечивает хорошую интерпретируемость, но ограничена в учете сложных нелинейных зависимостей в данных.

Алгоритм случайного леса демонстрирует устойчивость к шумам и способность улавливать нелинейные зависимости. В нашем исследовании это подтвердилось значениями AUC = 0,76 и NDCG@5 = 0,42, что на 7% и 20% соответственно превышает показатели логистической регрессии.

Наибольший интерес представляют результаты алгоритмов градиентного бустинга. XGBoost достиг AUC = 0,81 и NDCG@5 = 0,55. Более современная реализация LightGBM показала незначительное улучшение - AUC = 0,82 и NDCG@5 = 0,57. Алгоритм CatBoost продемонстрировал наилучшие результаты среди классификаторов - AUC = 0,83 и NDCG@5 = 0,59.

Специализированный алгоритм LightGBM Ranker показал исключительную эффективность в метрике NDCG@5 = 0,63, что на 23% превышает результат CatBoost и на 80% - логистической регрессии. Это объясняется тем, что алгоритм изначально оптимизирован для задач ранжирования и учитывает групповую структуру данных поисковых сессий.

Далее для подтверждения результатов были построены ROC-кривые (Receiver Operating Characteristic) - графики, используемые для визуализации производительности бинарной классификации модели. Эти графики отображают зависимость доли правильно распознанных положительных примеров (чувствительность, True Positive Rate, TPR) от доли неправильно распознанных отрицательных примеров (ложноположительная ошибка, False Positive Rate, FPR).

На рис. 3 изображены графики оценки качества бинарных классификаторов (ROC-кривые) [8], полученные по результатам оценки алгоритмов градиентного бустинга (XGBoost, LightGBM, CatBoost).

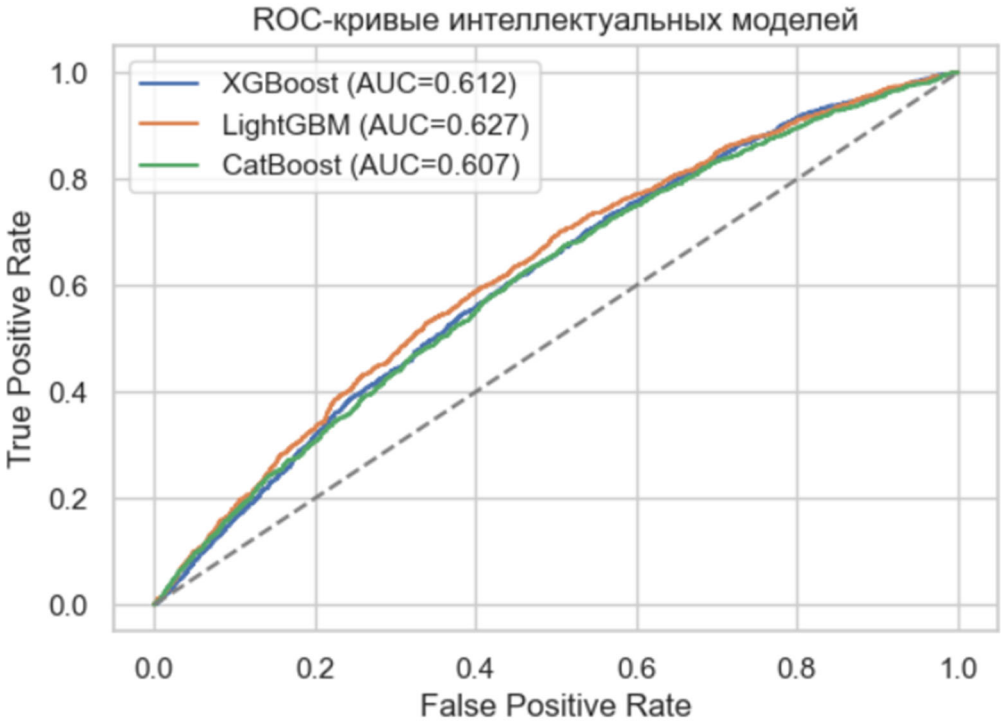

Рисунок 3. Графики оценки качества бинарных классификаторов

Эти графики показывают, что метод LightGBM даёт наилучшее качество прогнозирование вероятности бронирования номеров в случае решения только проблемы классификации (при выборе единственного номера для бронирования, когда ранжирование не имеет значения).

Заключение

Разработанное решение комбинированной оценки моделей позволяет существенно повысить точность прогнозирования бронирований гостиниц по сравнению с традиционными подходами. Комбинирование задач классификации и ранжирования обеспечивает более полное решение бизнес-задачи персонализации рекомендаций.

LightGBM Ranker рекомендуется использовать в системе по прогнозированию бронирования, где критически важно точное ранжирование предложений. Для задач, требующих вероятностной оценки бронирования каждого отдельного отеля, более подходят CatBoost и LightGBM. Логистическую регрессию целесообразно применять как базовую модель для быстрого прототипирования и анализа значимости признаков.

Список литературы:

1. Friedman J., Hastie T., Tibshirani R. The Elements of Statistical Learning. Springer, 2017.
2. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference, 2016. https://doi.org/10.48550/arXiv.1603.02754
3. Ильичев В.Ю., Драч В.Е., Пацукевич А.Н. Использование технологий глубокого обучения для формирования моделей ценообразования. // Системный администратор. 2024. № 5 (258). С. 92-96.

4. Ильичев В.Ю., Жукова Ю.М., Шамов И.В. Использование технологии градиентного бустинга для создания аппроксимационных моделей. // Заметки ученого. 2021. № 12-1. С. 62-67.
5. Personalize Expedia Hotel Searches. [Электронный ресурс]. URL: https://www.kaggle.com/c/expedia-personalized-sort (Дата обращения 22.10.2025).
6. Рысин Р.А., Макотра А.В., Непомнящий В.Д. Анализ корреляции данных и отбор признаков в естественных и машинных задачах. // В сборнике: Научное сообщество студентов XXI столетия. Технические науки. Сборник статей по материалам CXLV студенческой международной научно-практической конференции. Новосибирск, 2025. С. 53-59.
7. Liu T.-Y. Learning to Rank for Information Retrieval // Foundations and Trends in Information Retrieval, 2009. Vol. 3: No. 3, pp 225-331. http://dx.doi.org/10.1561/1500000016.
8. Белоусов Н.С. ROC кривая, как оценка качества модели бинарной классификации. // Вестник современных исследований. 2018. № 5.1 (20). С. 388-391.

References:

1. Фридман Дж., Хасти Т., Тибширани Р. Элементы статистического обучения. Спрингер, 2017.
2. Чен Т., Гестрин К. XGBoost: масштабируемая система повышения древовидности // Материалы 22-й международной конференции ACM SIGKDD, 2016. https://doi.org/10.48550/arXiv.1603.02754
3. Ильичев В.Ю., Драч В.Е., Пацукевич А.Н. Использование технологий глубокого обучения для формирования моделей ценообразования. // Системный администратор. 2024. № 5 (258). С. 92-96.
4. Ильичев В.Ю., Жукова Ю.М., Шамов И.В. Использование технологии градиентного бустинга для создания аппроксимационных моделей. // Заметки ученого. 2021. № 12-1. С. 62-67.
5. Персонализируйте поиск отелей Expedia. [Электронный ресурс]. URL: https://www.kaggle.com/c/expedia-personalized-sort (Дата обращения 22.10.2025).
6. Рысин Р.А., Макотра А.В., Непомнящий В.Д. Анализируйте данные и отбирайте критерии при изготовлении и машинных задачах. // В сборнике: Научное сообщество студентов XXI века. Технические науки. Сборник статей по материалам CXLV студенческой международной научно-практической конференции. Новосибирск, 2025. С. 53-59.
7. Лю Т.-Ю. Обучение ранжированию информационного поиска // Основы и тенденции информационного поиска, 2009. Том. 3: № 3, стр. 225–331. http://dx.doi.org/10.1561/1500000016.
8. Белоусов Н.С. РПЦ кривая, оценка качества модели бинарной классификации. // Вестник современных исследований. 2018. № 5.1 (20). С. 388-391.