# РАЗДЕЛ I. ЭКОНОМИКО-МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

doi 10.17072/1994-9960-2021-4-327-345
УДК 338.488.2:004, ББК 65.43+32.8
JEL Code C6



## *ПРОГНОЗИРОВАНИЕ ОТМЕНЫ БРОНИРОВАНИЯ ОТЕЛЕЙ: СРАВНИТЕЛЬНАЯ ХАРАКТЕРИСТИКА СПЕЦИФИКАЦИЙ МОДЕЛЕЙ*

**Елена Ивановна Русакова** [a]
ORCID ID: 0000-0001-7229-9097, e-mail: elena.rusakova.2000@mail.ru
**Марина Владимировна Радионова** [b]
ORCID ID: 0000-0002-8339-3326, Researcher ID: L-9851-2015, e-mail: m.radionova812@gmail.com
[a] Национальный исследовательский университет «Высшая школа экономики», Пермский филиал
(Россия, 614070, г. Пермь, ул. Студенческая, 38)
[b] Пермский государственный национальный исследовательский университет
(Россия, 614990, г. Пермь, ул. Букирева, 15)

Неотъемлемой частью любой поездки является бронирование номера в отеле. В связи с этим за последние годы существенно возросла популярность и востребованность туристических онлайн-агентств, позволяющих клиентам сократить время и издержки прямой коммуникации с отелем, а также без штрафов и комиссий отменить бронирование. Рост количества отмен бронирований, наблюдаемый в последние несколько лет, негативно сказывается на финансовом положении и репутации отелей, которые в целях сокращения данных рисков вынуждены применять жесткую политику бронирования и стратегии овербукинга. Особую актуальность данная проблема имеет сегодня в связи с существенным сокращением туристического потока вследствие пандемии коронавируса. Решению проблемы будет способствовать разработка моделей прогнозирования отмены бронирования отелей с высокими показателями достоверности и точности прогноза. Обзор существующих решений показал, что наилучшие результаты прогнозирования обеспечивают следующие методы машинного обучения: случайный лес (*Random Forest*), нейронные сети, *CatBoost* и *XGBoost*. В связи с вышесказанным целью исследования является построение различных моделей прогнозирования отмены бронирования отелей на основе методов машинного обучения и их сравнительный анализ для обоснования выбора наилучшей модели при помощи метрик *Accuracy, Precision, Recall, F*-меры и площади под *ROC*-кривой. Информационную базу исследования составил набор данных *"Hotel Booking Demand Dataset"*, подготовленный *N. Antonio, A. de Almeida* и *L. Nunes* и опубликованный на портале *ScienceDirect*. В ходе исследования определено, что модель случайного леса (*Random Forest*) наилучшим образом предсказывает отмену бронирования отелей. В частности, на тестовой выборке данная модель показала процент правильных ответов среди всех прогнозов – 84,5 %; процент бронирований, названных классификатором отмененными и при этом действительно являющихся отмененными, – 87,3 %. В перспективе целесообразно совершенствование модели случайного леса и других моделей машинного обучения посредством включения дополнительных, ранее не учтенных гиперпараметров.

*Ключевые слова: бронирование отеля, методы прогнозирования отмены бронирования, методы машинного обучения, случайный лес, нейронная сеть, CatBoost классификация, XGBoost классификация, прогнозирование.*

**Для цитирования:**
*Русакова Е.И., Радионова М.В.* Прогнозирование отмены бронирования отелей: сравнительная характеристика спецификаций моделей // Вестник Пермского университета. Сер. «Экономика». 2021. Том 16. № 4. С. 327–345. doi: 10.17072/1994-9960-2021-4-327-345

# *PREDICTING HOTEL BOOKING CANCELLATION: A COMPARATIVE ANALYSIS OF MODELS*

**Elena I. Rusakova** [a]
ORCID ID: 0000-0001-7229-9097, e-mail: elena.rusakova.2000@mail.ru
**Marina V. Radionova** [b]
ORCID ID: 0000-0002-8339-3326, Researcher ID: L-9851-2015, e-mail: m.radionova812@gmail.com

[a] National Research University "Higher School of Economics", Perm Branch
(38, Studencheskaya st., Perm, 614070, Russia)
[b] Perm State University (15, Bukireva st., Perm, 614990, Russia)

Booking a hotel room is an integral part of any trip. Therefore, recent years are characterized by an increasing popularity of and demand for online travel agencies which save clients' time and efforts applied to the communication with the hotels, as well as cancel a booking with no fines and charges. Hotel booking cancellations are on the rise in recent several years, which has its adverse effect on the financial status and reputations of the hotels. They have to follow a strict booking policy and overbooking strategy to reduce the risks. This problem is particularly burning today due to a significant decrease in tourist flows induced by the coronavirus pandemic. This issue can be solved by developing the predictive models of hotel booking cancellation with a high confidence index and a high prediction accuracy rate. An overview of the existing solutions shows that the following machine learning methods give the best predictive results: Random Forest, neuron networks, CatBoost, and XGBoost. Thus, the purpose of the research is to develop different machine learning based predictive models for hotel booking cancellation and to compare them in order to justify the choice of the best model with such metrics as Accuracy, Precision, Recall, F-measures, and the area under the ROC curve. The information database for the research was Hotel Booking Demand Dataset prepared by N. Antonio, A. de Almeida and L. Nunes and published on ScienceDirect platform. The research found out that a Random Forest Model gives the best prediction for hotel booking cancellation. For example, this model shows the percentage of the correct answers from a text set, 84.5% is among all predictions; 87.3% is the percentage of the bookings which are actually cancelled and referred to as cancelled by a classifier. Further research is seen to be focused on improving the Random Forest Model and other models of machine learning with additional unaccounted hyperparameters.

*Keywords: hotel booking, predictive methods for booking cancellation, machine learning methods, random forest, neuron networks, CatBoost classification, XGBoost classification, prediction.*

**For citation:**
Rusakova E.I., Radionova M.V. Predicting hotel booking cancellation: A comparative analysis of models. *Perm University Herald. Economy*, 2021, vol. 16, no. 4, pp. 327–345. doi: 10.17072/1994-9960-2021-4-327-345

## ВВЕДЕНИЕ

В современных условиях бронирование отелей преимущественно осуществляется через третьих лиц: *Booking.com, AirBnb* и т. д. В связи с этим в практике гостиничного бизнеса произошли изменения, касающиеся правил отмены бронирований на сайтах туристических онлайн-агентств, предусматривающих отмену бронирования без штрафов и комиссий. Клиенты со временем привыкли к политике бесплатной отмены. Согласно результатам исследования *D-Edge Hospitality Solutions*,[1] это привело к росту доли отмененных бронирований отелей с 6 % в 2014 г. до 40 % в 2018 г. Рост числа отмен бронирований затрудняет процесс прогнозирования для отелей, что приводит к неоптимальной загрузке отелей и, следовательно, потере доходов.

[1] *D-Edge* Hospitality Solutions: How online hotel distribution is changing in Europe. URL: https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/ (дата обращения: 30.06.2021).

Несмотря на то что предварительное бронирование считается основным показателем прогнозируемой эффективности отеля [1], возможность отмены услуги создает риск, поскольку отель должен гарантировать номера всем клиентам и, соответственно, учитывать альтернативную стоимость свободных номеров в случае отмены бронирования или незаселения [2]. Как утверждают *C.-C. Chen, Z. Schwartz, P. Vargas* [3], в настоящее время большая часть отмен бронирований происходит из-за того, что клиенты продолжают искать более выгодные предложения от отелей даже после совершения бронирования. Такие клиенты делают несколько бронирований, а затем отменяют все, кроме одного, наиболее предпочтительного для них. Соответственно, клиенты ценят возможность бесплатной отмены бронирования, предоставляющей право отказа от услуг в случае изменения их планов или предпочтений. Однако возможность отмены бронирования оказывает существенное влияние на решения по управлению спросом в индустрии гостеприимства. Отмены бронирования номеров ограничивают построение точных прогнозов, что является важным инструментом управления доходами отелей. Чтобы нивелировать данные риски, отели применяют жесткую политику отмены бронирования и стратегии овербукинга (когда отель позволяет клиентам бронировать больше номеров, чем на самом деле есть в отеле), что также может негативно сказаться на доходах отеля и его репутации.

Развитие рынка интернет-бронирования отелей актуализировало интерес к исследованиям, связанным с разработкой методов и инструментов прогнозирования отмены бронирований. Помимо работ *H.-C. Huang, A.Y. Chang, C.-C. Ho* [4], которые использовали данные о ресторанах при отелях, и *M.G. Yoon, H.Y. Lee, Y.S. Song* [5], использующих смоделированные данные об отменах бронирований, в других исследованиях для прогнозирования отмены бронирования применялся стандарт *Personal Name Record data* (далее – *PNR*), разработанный Международной ассоциацией воздушного транспорта. *PNR* не позволял установить причины отмены бронирования, так как в нем преимущественно были собраны факторы, которые важны для авиакомпаний.

В исследовании *N. Antonio, A. Almeida, L. Nunes* [6] построена модель определения бронирования отелей с высокой вероятностью отмены и предложен инструментарий прогнозирования отмен бронирований. Поскольку целевая переменная принимала только двоичные значения (0 – нет; 1 – да), авторами применялись следующие алгоритмы классификации: *Boosted Decision Tree; Random Forest; Decision Jungle; Locally Deep Support Vector Machine и Neural Network*, лучшим из которых оказался алгоритм *Random Forest*.

В 2019 г. на платформе *Towards Data Science* было опубликовано исследование *E. Zeytinci "Predicting Hotel Reservation Cancellations with Machine Learning"* [7], в котором обосновывалась возможность прогнозирования отмены бронирования отелей на основе методов машинного обучения. В работе *E. Zeytinci* подчеркнута значимость предварительной обработки данных, трансформации категориальных признаков при построении модели, оптимизации модели и настройки ее гиперпараметров и построена *XGBoost* модель.

В исследовании *M. Wingen* [8] показано применение алгоритма случайного леса, деревьев решений, логистической регрессии и *XGBoost* для прогнозирования отмены бронирований. Стоит отметить, что, в отличие от других исследователей, *M. Wingen* не использовал при обучении данные о количестве изменений (предыдущих отмен) в бронированиях, поскольку эта информация может изменяться с течением времени. В качестве наиболее значимых были определены следующие переменные: тип депозита, суточная стоимость проживания и время от момента бронирования до прибытия в отель. Результаты исследования свидетельствуют о том, что наилучшим образом предсказать результат на тестовой выборке удалось с помощью алгоритма случайного леса.

В работе [9] применялись следующие три алгоритма для прогнозирования отмены бронирования: *Adaptive Boosting; Gradient Boosting; Random Forest.* Наилучший результат на тестовой выборке показала

*Random Forest* модель. Было выявлено, что количество дней, прошедших между датой ввода бронирования и датой прибытия, средняя суточная стоимость проживания и тип депозита оказывают наиболее сильное влияние на отмену бронирования отелей.

В книге *"XAI stories"* [10] были собраны результаты студенческих проектов по курсу машинного обучения университетов *University of Warsaw* и *Warsaw University of Technology*. Третья глава книги *"Story Hotel Booking Cancellations: eXplainable predictions for booking cancellation"* посвящена построению моделей прогнозирования отмены бронирования отелей, наибольшую достоверность из которых продемонстрировали модели *LightGBM, Naive Bayes* и логистическая регрессия. Поясним также, что в связи с тем, что около 40 % бронирований совершалось жителями Португалии со средним процентом отмененных бронирований 38 %, в ходе исследования было построено по две модели каждого типа: для прогнозирования отмен бронирований жителями Португалии и жителями других стран. После разделения набора данных и обучения двух различных моделей каждого типа (*LightGBM, Naive Bayes* и логистической регрессии) была значительно повышена точность прогнозирования для каждой модели. Самую высокую точность показала *LightGBM* модель.

В работе *M. Banza* 2020 г. [11] был применен метод *Power Predict Score (PPS)*, позволяющий оценить силу зависимости не только между числовыми, но и между категориальными переменными. Наилучшей в данном исследовании оказалась модель, в основу которой лег алгоритм *CatBoost*[1], представляющий собой градиентный бустинг на деревьях решений на тестовой выборке.

В статье *J. Kelman* [12] описан исследовательский анализ данных, кластеризация данных по клиентам, которые совершали бронирование, и построена модель прогнозирования отмены брони отеля на основе нейронной сети, точность которой равна 97 %. Кластеризация клиентов позволила получить дополнительные сведения о клиентах и причинах отмены бронирований. Самой подходящей для создания кластеров оказалась модель *K-prototypes*, учитывающая числовые и категориальные переменные. Стоит отметить, что в этом исследовании впервые было уделено внимание дате отмены бронирования: в среднем клиенты отменяют бронь за 3 дня до предполагаемой даты заезда. У сотрудников отеля практически не остается времени, чтобы найти нового гостя или скорректировать свою работу. Это является еще одним свидетельством необходимости построения модели прогнозирования отмены бронирования с высокой степенью достоверности.

В источнике[2] для прогнозирования отмены бронирования отелей применялись модели *Decision Tree* и *Random Forest*, точность которых оказалась примерно одинаковой (*Accuracy* = 78 %). Результаты исследования были следующим образом прокомментированы в источнике: «Небольшое различие можно игнорировать, потому что оно могло быть результатом случайного подбора параметра *Random Forest* модели. Таким образом, технически обе модели могут быть использованы». В исследовании был сделан вывод, что тип депозита, количество поправок в бронировании, общее количество специальных запросов и средняя суточная стоимость проживания оказывают наибольшее влияние на прогнозы модели.

Таким образом, изучение литературы по вопросу отмены бронирования показало, что модели прогнозирования с алгоритмами машинного обучения могут позволить менеджерам отелей минимизировать потери доходов от отмены бронирований и снизить риски, связанные с овербукингом, а также применять менее жесткие правила отмены бронирования.

Обзор существующих решений по моделированию прогнозирования отмены бронирований за последние годы показал, что использование алгоритмов машинного обучения позволяет обеспечить высокую досто-

[1] *CatBoost* is a high-performance open source library for gradient boosting on decision trees. URL: https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus (дата обращения: 15.05.2021).

[2] *Hotel* Bookings Cancellation. 2020. URL: https://rpubs.com/rogate16/hotel-bookings (дата обращения: 30.06.2021).

верность результатов. При этом наивысший показатель *Accuracy* демонстрируют случайный лес (*Random Forest*), нейронные сети, *CatBoost и XGBoost*. Однако в приведенных исследованиях указанные методы машинного обучения не сравнивались между собой.

В связи с вышесказанным целью настоящего исследования является построение различных моделей прогнозирования отмены бронирования отелей на основе методов машинного обучения и их сравнительный анализ для обоснования выбора наилучшей модели.

## МЕТОДЫ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В настоящем исследовании была использована база данных *"Hotel Booking Demand Dataset"* [13], подготовленная *N. Antonio, A. de Almeida* и *L. Nunes* и опубликованная в феврале 2019 г. на портале *ScienceDirect*. Этот набор данных содержит реальные данные о бронировании двух отелей: курорта в регионе Алгарве и городского отеля в Лиссабоне, Португалия. База данных состоит из 32 переменных и 119 390 наблюдений. Каждое наблюдение представляет собой бронирование отеля в период с 1 июля 2015 г. по 31 августа 2017 г., включая как отмененные, так и неотмененные бронирования. Поскольку в наборе данных содержались реальные данные отеля, все данные, относящиеся к идентификации отеля или клиента, были удалены. Зависимая (целевая) переменная *in_canceled* принимает только два значения: 1 – если бронь была отменена, 0 – если нет.

В табл. 1 приведены факторные переменные, используемые в исследовании.

Таблица 1. **Обозначение переменных**

Table 1. **Variable notation**

| Переменная | Тип данных | Описание |
|---|---|---|
| *is_canceled* | Категориальный | Переменная, которая показывает, было отменено (1) бронирование или нет (0) |
| *hotel* | Категориальный | Тип отеля (H1 = «Курортный отель» или H2 = «Городской отель») |
| *lead_time* | Числовой | Количество дней, прошедших между датой ввода бронирования в *PMS* и датой прибытия |
| *adr* | Числовой | Средняя дневная ставка, определяемая делением суммы всех транзакций по размещению на общее количество ночей проживания |
| *adults* | Числовой | Количество взрослых, на которое бронируется номер |
| *children* | Числовой | Количество детей (в бронируемом номере) |
| *babies* | Числовой | Количество младенцев / маленьких детей (в бронируемом номере) |
| *agent* | Категориальный | *ID* туристического агентства, через которое было оформлено бронирование |
| *arrival_date_day_of_month* | Числовой | День месяца даты прибытия |
| *arrival_date_month* | Категориальный | Месяц прибытия – 12 уникальных значений (Категории: Январь, Февраль, Март и т. д.) |
| *arrival_date_week_number* | Числовой | Номер недели даты прибытия |
| *arrival_date_year* | Числовой | Год даты прибытия |
| *assigned_room_type* | Категориальный | Код для типа номера, назначенного для бронирования |
| *booking_changes* | Числовой | Количество изменений / дополнений, внесенных в бронирование до момента заселения или отмены |
| *company* | Категориальный | *ID* компании / юридического лица, совершившего бронирование или ответственного за его оплату |
| *country* | Категориальный | Страна проживания клиента (категории в формате *ISO* 3155–3:2013) |

Окончание табл. 1

| Переменная | Тип данных | Описание |
|---|---|---|
| *customer_type* | Категориальный | Тип бронирования, предполагающий одну из четырех категорий:<br>*Contract*;<br>*Group* – бронирование группой клиентов;<br>*Transient* – бронирование не является групповым или совершенным в рамках контракта и не связано с другим временным бронированием;<br>*Transient-party* – бронирование является временным, но связано как минимум с другим временным бронированием |
| *days_in_waiting_list* | Числовой | Количество дней, в течение которых бронирование находилось в листе ожидания, прежде чем оно было подтверждено клиенту |
| *deposit_type* | Категориальный | Тип залога:<br>*No Deposit* – депозит не производился;<br>*Non Refund* – внесен залог в размере полной стоимости проживания;<br>*Refundable* – внесен залог в размере, меньшем общей стоимости проживания |
| *distribution_channel* | Категориальный | Канал «распространения» бронирования: «ТА» – «Туристические агенты» или «ТО» – «Туроператоры» |
| *is_repeated_guest* | Категориальный | Значение, указывающее, были ли уже бронирования от этого клиента (1) или нет (0) |
| *market_segment* | Категориальный | Сегмент рынка: «ТА» – «Туристические агенты» или «ТО» – «Туроператоры» |
| *meal* | Категориальный | Тип забронированного питания:<br>*Undefined/SC* – нет определенного питания;<br>*BB* – только завтрак;<br>*HB* – полупансион (завтрак и еще один прием пищи, обычно ужин);<br>*FB* – полный пансион (завтрак, обед и ужин) |
| *previous_bookings_not_canceled* | Числовой | Количество предыдущих бронирований, которые не были отменены клиентом до текущего бронирования |
| *previous_cancellations* | Числовой | Количество предыдущих бронирований, которые были отменены клиентом до текущего бронирования |
| *required_car_parking_spaces* | Числовой | Количество парковочных мест, которые требуются для клиента |
| *reservation_status* | Категориальный | Последний статус бронирования, допускающий одну из трех категорий:<br>*Canceled* – бронирование было отменено заказчиком;<br>*Check-Out* – клиент зарегистрировался, но уже уехал;<br>*No-Show* – клиент не прошел регистрацию и не проинформировал отель о причине |
| *reservation_status_date* | Дата | Дата, когда был установлен последний статус. Эта переменная может использоваться вместе с *reservation_status*, чтобы установить, когда было отменено бронирование или когда клиент уехал из отеля |
| *reserved_room_type* | Категориальный | Код типа забронированного номера |
| *stays_in_weekend_nights* | Числовой | Количество ночей в выходные (суббота или воскресенье), которые клиент проживал в отеле или забронировал номер для проживания |
| *stays_in_week_nights* | Числовой | Количество ночей в неделю (с понедельника по пятницу), в которые клиент останавливался в отеле или забронировал номер для проживания |
| *total_of_special_requests* | Числовой | Количество особых запросов, сделанных клиентом (например, две односпальные кровати или высокий этаж) |

На рис. 1 показано, что исходный набор данных являлся не сбалансированным по целевой переменной.

После заполнения пропусков, удаления выбросов и ошибочных значений в выборке осталось 117 244 наблюдения.

В табл. 2 приведены описательные статистики количественных переменных.

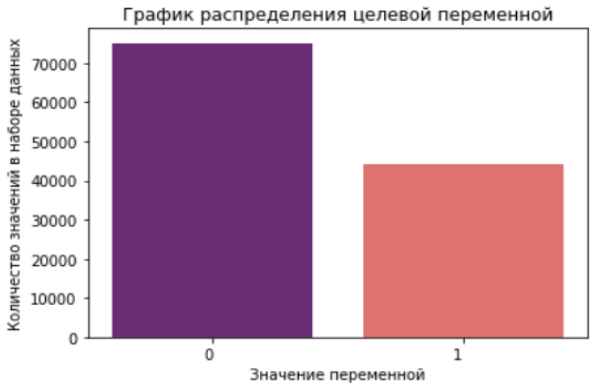


Рис. 1. **Распределение целевой переменной**

Fig. 1. **Distribution of the target variable**

Таблица 2. **Описательные статистики количественных переменных**

Table 2. **Summary statistics of the quantitative variables**

| Переменная | Среднее | Среднеквадратичное отклонение | Мода | Минимум | Медиана | Максимум |
|---|---|---|---|---|---|---|
| *lead_time* | 104,5 | 105 | 0 | 0 | 70 | 594 |
| *arrival_date_week_number* | 27,1 | 13,6 | 33 | 1 | 27 | 53 |
| *arrival_date_day_of_month* | 15,8 | 8,7 | 17 | 1 | 16 | 31 |
| *stays_in_weekend_nights* | 0,94 | 1 | 0 | 0 | 1 | 19 |
| *stays_in_week_nights* | 2,5 | 1,9 | 2 | 0 | 2 | 50 |
| *adults* | 1,86 | 0,48 | 2 | 0 | 2 | 4 |
| *children* | 0,1 | 0,4 | 0 | 0 | 0 | 3 |
| *babies* | 0,008 | 0,1 | 0 | 0 | 0 | 2 |
| *previous_cancellations* | 0,087 | 0,85 | 0 | 0 | 0 | 26 |
| *previous_bookings_not_canceled* | 0,125 | 1,45 | 0 | 0 | 0 | 72 |
| *booking_changes* | 0,22 | 0,64 | 0 | 0 | 0 | 18 |
| *days_in_waiting_list* | 2,34 | 17,7 | 0 | 0 | 0 | 391 |
| *adr* | 103,55 | 46,7 | 62 | 0,3 | 95 | 510 |
| *required_car_parking_spaces* | 0,06 | 0,25 | 0 | 0 | 0 | 8 |
| *total_of_special_requests* | 0,57 | 0,79 | 0 | 0 | 0 | 5 |

Набор данных содержит переменные разного масштаба, что следует из данных описательной статистики. Количественные признаки были стандартизованы для дальнейшего использования и обучения моделей с помощью трансформера *StandardScaler*[1].

Для кодирования категориальных признаков был применен трансформер из *sklearn* – *OneHotEncoder*[2]. Для кодируемого категориального признака создается $N$ новых признаков, где $N$ – количество категорий. Каждый $i$-й новый признак – бинарный характеристический признак $i$-й категории.

Для всех количественных переменных была построена корреляционная матрица в виде *heatmap*-графика[3] (рис. 2).

[1] *StandardScaler*. URL: sklearn.preprocessing.StandardScaler – scikit-learn 0.24.1 documentation (scikit-learn.org) (дата обращения: 15.05.2021).

[2] *OneHotEncoder*. URL: sklearn.preprocessing.OneHotEncoder – scikit-learn 0.24.1 documentation (scikit-learn.org) (дата обращения: 10.06.2021).

[3] *Wetschoreck F.* RIP correlation. Introducing the predictive power score. 2020. URL: https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598 (дата обращения: 20.05.2021).

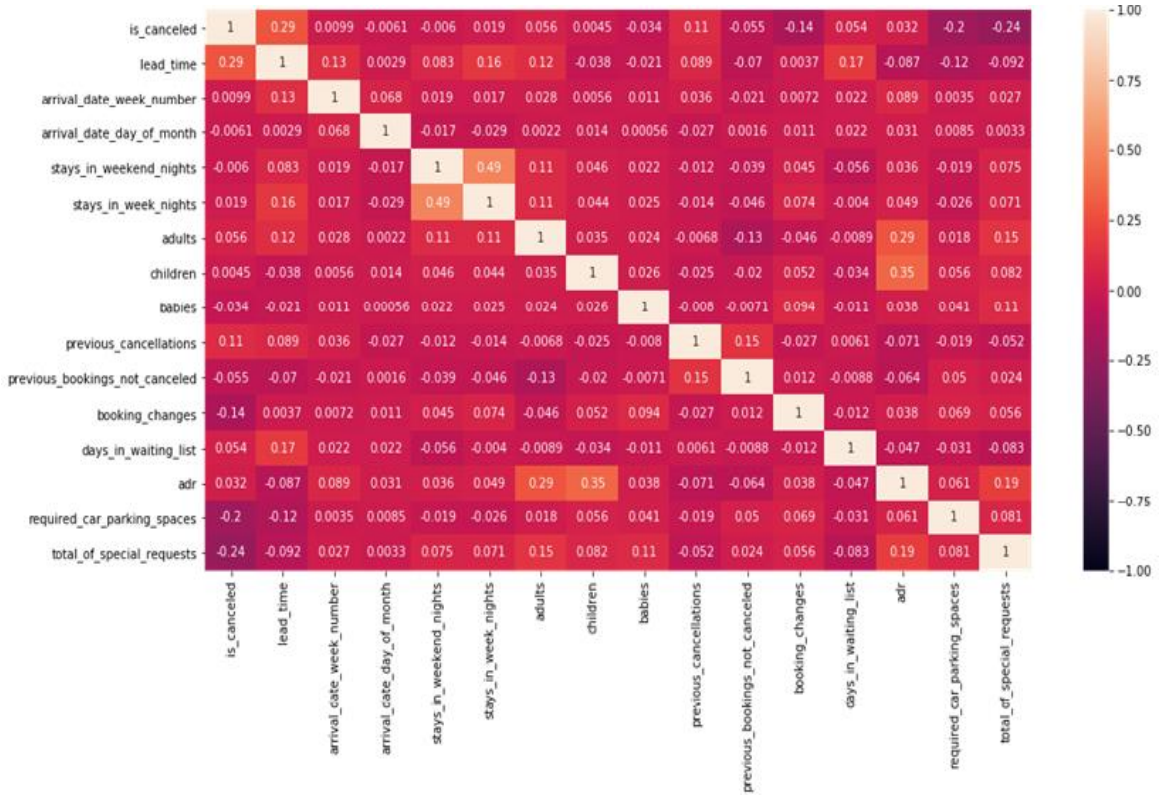


Рис. 2. **Корреляционная матрица**

Fig 2. **Correlation matrix**

Исходя из рис. 2 между целевой переменной *is_canceled* и остальными факторами есть определенные прямые и отрицательные зависимости. Наиболее сильная связь наблюдается между переменной *is_canceled* и следующими факторами:

– *lead_time*;
– *total_of_special_requests*;
– *required_car_parking_spaces*;
– *booking_changes*;
– *previous_cancellations*.

Полученное значение коэффициента корреляции 0,29 свидетельствует о наличии прямой слабой связи между отменой бронирования и временем между совершением бронирования и планируемой датой заселения в отель: когда клиент бронирует номер заранее, выше вероятность, что он отменит бронирование. Коэффициент корреляции –0,24 говорит о наличии слабой прямой обратной связи между количеством специальных пожеланий клиента к номеру и отменой бронирования: чем больше клиент оставляет специальных пожеланий к номеру, тем меньше вероятность, что гость отменит бронирование. Слабая обратная прямая связь есть между количеством требующихся клиенту парковочных мест и отменой бронирования, об этом свидетельствует коэффициент корреляции, равный – 0,2: чем больше нужно парковочных мест, тем меньше вероятность отмены бронирования (парковочные места часто указывают гости, относящиеся к категории «Группа», когда бронирование совершается сразу на несколько человек).

Коэффициент корреляции –0,14 говорит о наличии слабой обратной связи между изменениями, внесенными в бронь, и отменой бронирования: чем больше изменений и дополнений клиент внес в бронь, тем меньше вероятность, что он отменит бронирование. Полученное значение коэффициента корреляции, равное 0,11, свидетельствует о наличии прямой слабой связи между отменой бронирования и количеством бронирований, отмененных клиентом ранее: чем больше броней ранее отменял клиент, тем выше вероятность, что бронирование будет отменено.

На рис. 3 проранжированы значения коэффициентов корреляции с зависимой переменной *in_canceled.*

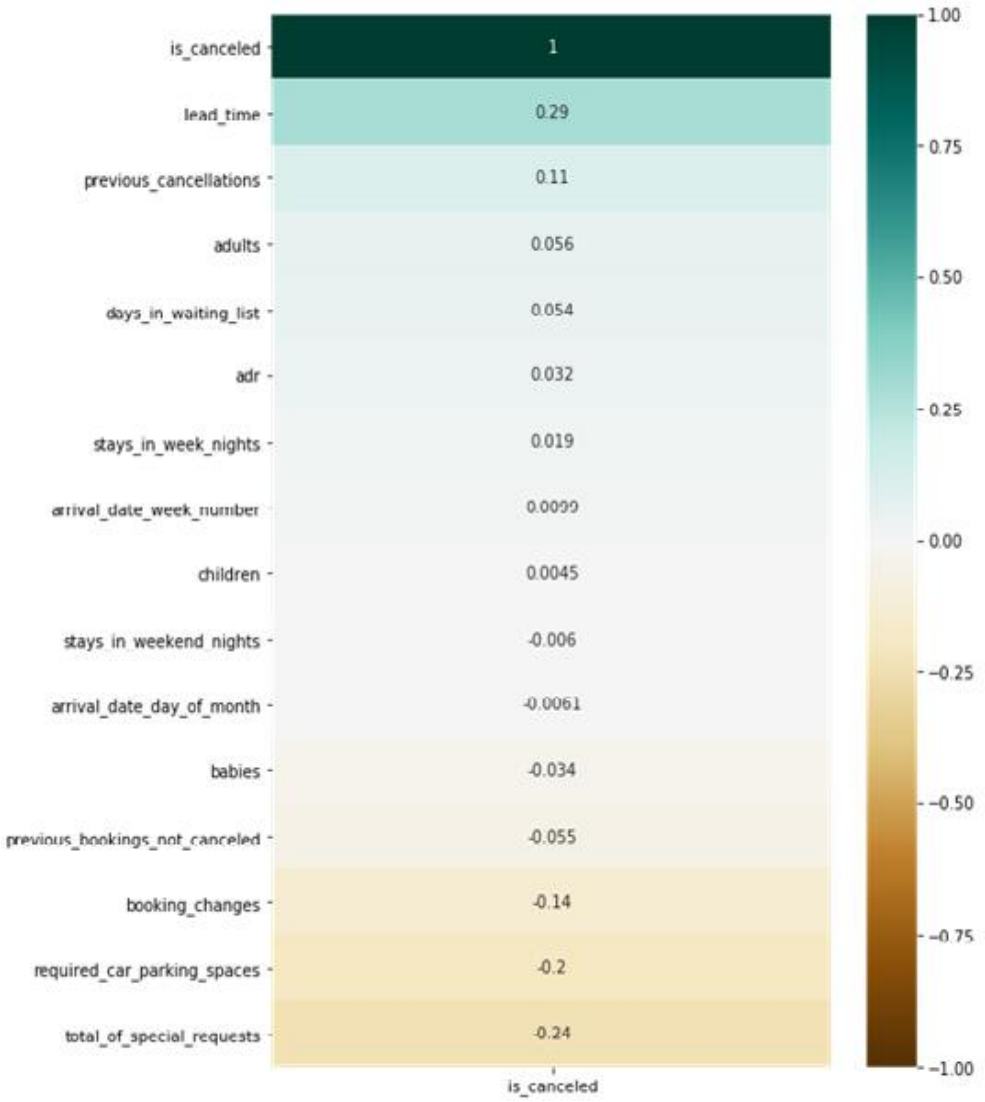


Рис. 3. **Коэффициенты корреляции параметров с целевой переменной**

Fig. 3. **Correlation coefficients of the parameters with the target variable**

В наборе данных, благодаря построению корреляционной матрицы и графиков для качественных переменных, были выявлены закономерности между описывающими переменными и целевой переменной, что позволило исключить незначимые переменные. После этого в наборе осталось 24 объясняющих переменных:

– количественные переменные: *adr, lead_time, stays_in_weekend_nights, adults, stays_in_week_nights, children, babies, previous_cancellations, booking_changes, previous_bookings_not_canceled, required_car_parking_spaces, arrival_date_week_number, days_in_waiting_list, total_of_special_requests, arrival_date_day_of_month*;

– категориальные переменные: *hotel, arrival_date_month, meal, market_segment, distribution_channel, is_repeated_guest, deposit_type, customer_type*.

Перед обучением моделей машинного обучения для прогнозирования отмены бронирования выборка была разбита на обучающую (80 % совокупности) и тестовую (20 %). Обучающая выборка была использована для обучения четырех моделей, а тестовая выборка – для оценки качества этих моделей. Для оценки качества прогнозной силы моделей использовались матрицы ошибок (табл. 3), которые показывают количество ложно и истинно предсказанных исходов [14].

Таблица 3. **Матрица ошибок**

Table 3. **Error matrix**

| | | *Actual class* | |
|---|---|---|---|
| | | *Positive* (0) | *Negative* (1) |
| *Predicted class* | *Positive* (0) | *True positives (TP)* | *False positives (FP)* |
| | *Negative* (1) | *False negatives (FN)* | *True negatives (TN)* |

Как известно, наиболее распространенным критерием качества модели является ее *точность* (доля верных предсказаний – *Accuracy, ACC*). Помимо точности модели, также проверяется ее чувствительность и специфичность. Под *чувствительностью* (*True Positives Rate, TPR*) понимается доля истинно положительных классификаций, под *специфичностью* (*True Negatives Rate, TNR*) – доля отрицательных значений. Данные показатели рассчитываются по формулам:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$TNR = \frac{TN}{TN + FP}, \quad (2)$$

$$TPR = \frac{TP}{TP + FN}. \quad (3)$$

Показатель чувствительности отражает долю истинно положительных результатов классификации. Показатель специфичности отражает точность работы алгоритма классификации, т. е. определяет долю истинно отрицательных значений, определенных методом классификации. Классификация, обладающая высокой специфичностью, обеспечивает большую вероятность правильного распознавания неотмененных бронирований.

Поскольку выборка не сбалансирована, то в таких условиях чаще всего применяют показатели *точность* и *полнота*, которые не зависят от соотношения классов, в отличие от доли верных ответов. При этом существует риск возникновения противоречия, для устранения которого применяется усредненная метрика, так называемая *F*-мера – среднее гармоническое показателей *точность* и *полнота*. С помощью *F*-меры определяют важность конкретной метрики по формуле

$$F_\beta = (1 + \beta^2) \cdot \frac{TNR \cdot TPR}{(\beta^2 \cdot TNR) + TPR}. \quad (4)$$

Параметр $\beta \in [0, \infty)$ определяет вес точности в метрике. Так, при $\beta = 0$ получаем точность модели, при $\beta = 1$ – непараметрическую *F*-меру, при $\beta = \infty$ – полноту модели. Наилучшей признается та классификация, при которой *F*-мера принимает наибольшее значение. Также модели сравниваются по площади *AUC* под *ROC*-кривой.

Далее представим результаты построения моделей случайного леса *(Random forest), XGBoost, CatBoost*, нейронной сети и их сравнения в целях определения наилучшей модели прогнозирования отмены бронирования отелей. Для подбора параметров моделей машинного обучения использовался инструмент, который находит наилучшие параметры путем перебора: создает модель для каждой возможной комбинации заданных пользователем параметров.

***Построение модели случайного леса***

Метод случайных лесов (*Random Forest*) основан на бэггинге над решающими деревьями. В первую очередь рассмотрим решающие деревья (*Decision Trees*) – семейство моделей, которые позволяют восстанавливать нелинейные зависимости произвольной сложности. Алгоритм бинарных решающих деревьев начинается в корневой вершине и вычисляет в ней значение функции. Если значение функции равно нулю, то алгоритм переходит в левую вершину дерева, если не равно нулю, то в правую вершину, далее вычисляет значение предиката в текущей вершине и делает переход или влево, или вправо. Процесс продолжается, пока не будет достигнута листовая вершина. Алгоритм возвращает тот класс, который приписан листовой вершине[1]. У решающих деревьев есть существенный недостаток: поскольку дерево может быть глубоким, оно пытается уловить самые сложные зависимости, что приводит к переобучению модели. Такое дерево не сможет показать хорошие результаты на новых данных. Деревья решений чувствительны к шумам во входных данных: небольшие изменения обу-

[1] *Соколов Е.А.* Решающие деревья. URL: https://github.com/esokolov/ml-course-hse/blob/master/2020-fall/lecture-notes/lecture07-trees.pdf (дата обращения: 15.06.2021).

чающей выборки могут привести к глобальным корректировкам модели, что, определенно, скажется на интерпретируемости модели[1].

Сейчас решающие деревья редко используются как отдельные методы классификации. Однако композиции из решающих деревьев – *Random Forest* – более устойчивы к изменениям в данных и могут показывать очень хорошие результаты.

В ходе исследования на первом этапе были построены модели случайного леса *Random Forest* с различными входными параметрами. Анализ показал, что наилучшей моделью является модель с максимальной глубиной дерева, равной 18, количеством деревьев в ансамбле 100 и количеством параметров, которые следует учитывать при поиске наилучшего разделения, заданным как корень из количества параметров, на которых обучается модель. При данных значениях параметров модели значения метрик классификации, полученные на обучающем наборе данных, были максимально приближены к значениям метрик классификации, полученным на тестовом наборе данных. Время обучения данной модели составило 12,6 с.

Матрица ошибок классификации, полученная на тестовой выборке, представлена в табл. 4.

Таблица 4. **Матрица ошибок модели *Random Forest***

Table 4. **Random Forest error matrix**

| | | *Actual class* | |
|---|---|---|---|
| | | *Positive* (0) | *Negative* (1) |
| *Predicted class* | *Positive* (0) | 13 804 | 873 |
| | *Negative* (1) | 2 755 | 6 017 |

На основании матрицы ошибок вычислены значения для метрик классификации (табл. 5) и построена *ROC*-кривая (рис. 4). В табл. 5 добавлены значения метрик классификации, которые были получены на обучающей выборке.

Анализ табл. 5 показывает, что значения метрик на обучающей выборке немного превышают значения на тестовой выборке, поэтому можно говорить о том, что модель могла немного переобучиться.

Таблица 5. **Метрики классификации алгоритма *Random Forest***

Table 5. **Classification metrics of the Random Forest algorithm**

| **Показатели** | **На обучающей выборке** | **На тестовой выборке** |
|---|---|---|
| *Accuracy* | 0,867 | 0,845 |
| *Precision* | 0,895 | 0,873 |
| *Recall* | 0,731 | 0,686 |
| *F*-мера | 0,805 | 0,768 |
| *AUC-ROC* | 0,96 | 0,92 |

Модель случайного леса (*Random forest*) на тестовой выборке показала высокий процент правильных ответов среди всех прогнозов – 84,5 %. Процент бронирований, названных классификатором отмененными и при этом действительно являющихся отмененными, – 87,3 %. Модель действительно «покрыла» все отмененные бронирования на 68,6 %. Высокое значение *AUC-ROC* говорит о том, что модель хорошо ранжировала объекты.

Наибольшие веса в модели имеют следующие категориальные переменные: *meal_HB* (0,162789), *meal_SC* (0,134908), *market_segment_Offline_TA/TO* (0,104166), *customer_type_Group* (0,080403), *deposit_type_Refundable* (0,051064).

Согласно модели вероятность отмены бронирования снижается, если:

– клиент планирует оплачивать полупансион (*HB*) или тип питания строго не определен (*SC*);

– клиент совершал бронирование офлайн через туристическое агентство или туроператора;

– бронирование совершено группой клиентов (тип клиента – *Group*);

– была внесена хотя бы неполная предоплата (тип предоплаты – *Refundable*).

Среди количественных переменных наибольший вес имеет *lead_time* (0,007906): чем больше время с момента бронирования до заселения в отель, тем выше вероятность отмены бронирования.

[1] *Соколов Е.А.* Бэггинг, случайные леса и разложение ошибки на смещение и разброс. URL: https://github.com/esokolov/ml-course-hse/blob/master/2020-fall/lecture-notes/lecture08-ensembles.pdf (дата обращения: 30.06.2021).

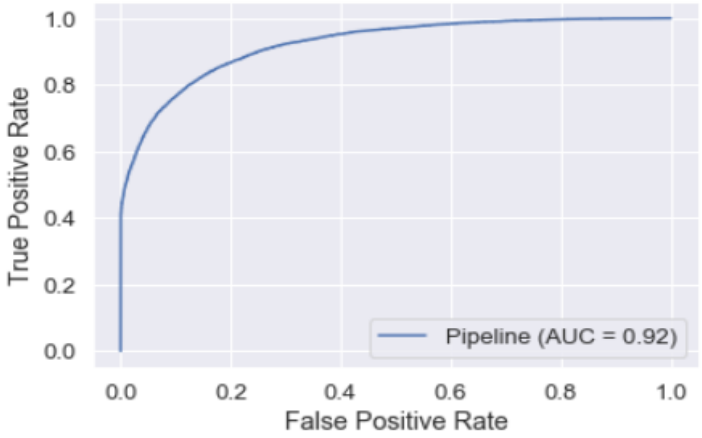


Рис. 4. ***ROC*-кривая для алгоритма *Random Forest***

Fig. 4. **ROC curve for the Random Forest algorithm**

***Построение XGBoost модели***

*XGBoost* (*Extreme Gradient Boosting*) – это алгоритм машинного обучения, основанный на композиции деревьев решений с использованием градиентного бустинга[1]. Идея градиентного бустинга заключается в том, чтобы каждая следующая модель исправляла ошибки предыдущей модели. В отличие от случайного леса, который создает дерево решений для каждой выборки, в градиентном бустинге деревья создаются последовательно, предыдущие деревья в модели не изменяются. В градиентном бустинге уменьшается смещение базовых моделей и на каждом шаге вычисляются производные функции потерь по прогнозу модели.

*XGBoost* – это конкретная реализация градиентного бустинга, характеризующаяся следующими особенностями:

1. Базовый алгоритм, стремясь минимизировать ошибки, приближает направление, рассчитанное с учетом вторых производных функции потерь.

2. Функционал регуляризируется, чтобы избежать переобучения модели посредством включения штрафов за количество листьев (чем больше листьев, тем сложнее разделяющая поверхность дерева) и за норму коэффициентов (чем сильнее коэффициенты отличаются от 0, тем сильнее базовый алгоритм будет влиять на итоговый прогноз композиции деревьев).

3. При построении дерева используется критерий информативности, зависящий от оптимального вектора сдвига. Критерий остановки при обучении дерева также зависит от оптимального сдвига.

4. Алгоритм *XGBoost* в процессе обучения может заполнять пропущенные значения в зависимости от значения потерь и использует свой собственный метод кросс-валидации на каждой итерации [15].

В рамках построения *XGBoost* модели прогнозирования отмены бронирования отелей в исследовании использовались следующие параметры: максимальная глубина дерева равна 7, количество деревьев в ансамбле равно 70.

При данных значениях параметров значения метрик классификации, полученные на обучающем наборе данных, были максимально приближены к значениям метрик классификации, полученным на тестовом наборе данных. Время обучения данной модели составило 6 с.

Матрица ошибок, полученная на тестовой выборке, представлена в табл. 6.

Таблица 6. **Матрица ошибок алгоритма *XGBoost***

Table 6. **Error matrix of the XGBoost algorithm**

| | | *Actual class* | |
|---|---|---|---|
| | | *Positive* (0) | *Negative* (1) |
| *Predicted class* | *Positive* (0) | 13 609 | 1 068 |
| | *Negative* (1) | 2 584 | 6 188 |

[1] *Соколов Е.А.* Градиентный бустинг. URL: https://github.com/esokolov/ml-course-hse/blob/master/2020-fall/lecture-notes/lecture09-ensembles.pdf (дата обращения: 15.05.2021).

На основании матрицы ошибок вычислены значения для метрик классификации и построена *ROC*-кривая (рис. 5). В табл. 7 также добавлены значения метрик классификации, которые были получены на обучающей выборке.

Таблица 7. **Метрики классификации алгоритма *XGBoost***

Table 7. **Classification metrics for the XGBoost algorithm**

| Показатели | На обучающей выборке | На тестовой выборке |
|---|---|---|
| *Accuracy* | 0,857 | 0,844 |
| *Precision* | 0,8699 | 0,853 |
| *Recall* | 0,729 | 0,705 |
| *F*-мера | 0,793 | 0,772 |
| *AUC-ROC* | 0,93 | 0,93 |

Согласно табл. 7, судя по тому, что значения метрик на обучающей выборке близки к значениям на тестовой выборке, можно говорить о том, что модель не переобучилась.

Алгоритм *XGBoost* на тестовой выборке показал процент правильных ответов среди всех прогнозов – 84,4 %. Процент бронирований, названных классификатором отмененными и при этом действительно являющихся отмененными, составил 85,3 %. Модель «покрыла» все действительно отмененные бронирования на 70,5 %. Высокое значение *AUC-ROC* говорит о том, что модель хорошо ранжировала объекты.

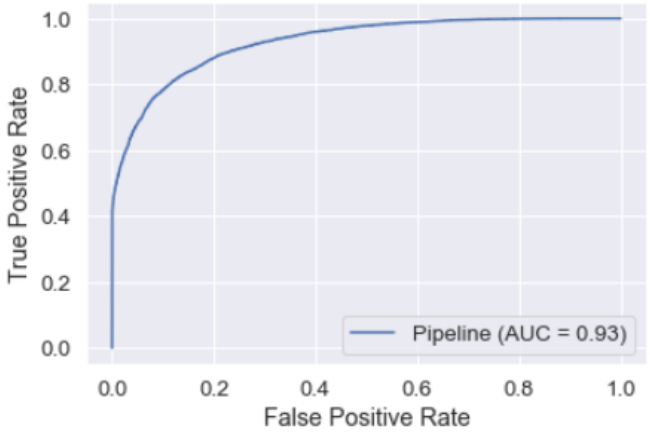


Рис. 5. ***ROC*-кривая для алгоритма *XGBoost***

Fig. 5. **ROC curve for the XGBoost algorithm**

Наибольшие веса в модели имеют следующие категориальные переменные: *meal_SC, customer_type_Contract, is_repeated_guest_0, arrival_date_month_March и customer_type_Group.* Среди количественных переменных наибольший вес имеет *lead_time.*

Согласно модели вероятность отмены бронирования снижается, если:

– тип питания строго не определен (*SC*);

– бронирование совершено группой клиентов или сопровождается заключением контракта (типы клиента – *Group, Contract*);

– бронирование планируется на март.

Вероятность отмены бронирования увеличивается:

– если клиент не заселялся в отель ранее;

– если отмечен большой промежуток времени с момента бронирования до заселения в отель.

***Построение CatBoost модели***

*CatBoost* – это библиотека градиентного бустинга над деревьями решений, созданная инженерами компании Яндекс. Она использует небрежные или «симметричные» (*oblivious*) деревья решений, чтобы построить сбалансированное дерево. Отличитель-

ной чертой небрежных деревьев является то, что одни и те же функции используются для расщепления (создания левых и правых веток) во всех промежуточных узлах в пределах одного уровня дерева. Небрежное дерево глубины $k$ имеет ровно $2^k$ листьев, а индекс листа можно вычислить простыми битовыми операциями [16].

Данный метод машинного обучения обладает рядом важных преимуществ. В частности, очень часто возникает необходимость обучаться на наборах данных с качественными переменными, которые, в отличие от количественных, не всегда можно сравнивать между собой. Это создает определенные трудности при работе с решающими деревьями, поскольку в классическом случае при «расщеплении» номер категории переводится в числовой вид и теряет изначальный смысл. В такой ситуации библиотека *CatBoost* позволяет получить отличные результаты на тестовых наборах данных с параметрами по умолчанию, что сокращает время, необходимое для настройки гиперпараметров.

Кроме того, *CatBoost* умеет «по умолчанию» обрабатывать пропущенные значения. Способ обработки пропущенных значений зависит от типа данных и выбранного пакета.

*CatBoost* также обеспечивает высокую точность прогноза за счет уменьшения переобучения: каждую итерацию *CatBoost* проверяет количество итераций с момента начала обучения, сравнивая его с оптимальным значением функции потерь. Модель считается переобученной, если количество итераций превышает значение, указанное в параметрах.

В настоящем исследовании при построении *CatBoost* модели была получена следующая матрица ошибок, полученная на тестовой выборке (табл. 8).

Таблица 8. **Матрица ошибок алгоритма *CatBoost***

Table 8. **Error matrix for the CatBoost algorithm**

| | | *Actual class* | |
|---|---|---|---|
| | | *Positive* (0) | *Negative* (1) |
| *Predicted class* | *Positive* (0) | 13 601 | 1 070 |
| | *Negative* (1) | 2 592 | 6 186 |

На основании матрицы ошибок вычислены значения для метрик классификации (табл. 9) и построена *ROC*-кривая (рис. 6). В табл. 9 добавлены значения метрик классификации на обучающей выборке.

Таблица 9. **Метрики классификации для алгоритма *CatBoost***

Table 9. **Classification metrics for the CatBoost algorithm**

| **Показатели** | **На обучающей выборке** | **На тестовой выборке** |
|---|---|---|
| *Accuracy* | 0,854 | 0,843 |
| *Precision* | 0,868 | 0,855 |
| *Recall* | 0,720 | 0,698 |
| *F*-мера | 0,788 | 0,768 |
| *AUC-ROC* | 0,93 | 0,91 |

Значения метрик на обучающей выборке близки к значениям на тестовой выборке, следовательно, можно утверждать, что модель не переобучилась.

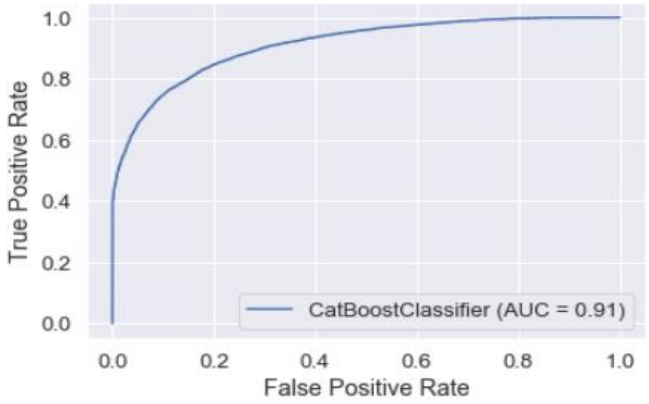


Рис. 6. ***ROC*-кривая для алгоритма *CatBoost***

Fig. 6. **ROC curve for the CatBoost algorithm**

Модель случайного леса на тестовой выборке показала процент правильных ответов среди всех прогнозов – 84,3 %. Процент бронирований, названных классификатором отмененными и при этом действительно являющихся отмененными, – 85,5 %. Модель «покрыла» все действительно отмененные бронирования на 69,8 %. Высокое значение *AUC-ROC* говорит о том, что модель хорошо ранжировала объекты.

Переменные, имеющие наибольший вес в модели *CatBoost*, приведены на рис. 7.

| | Feature Id | Importances |
|---|---|---|
| 0 | deposit_type | 32.730347 |
| 1 | required_car_parking_spaces | 13.977738 |
| 2 | previous_cancellations | 10.231279 |
| 3 | lead_time | 7.907919 |
| 4 | adr | 5.580529 |
| 5 | market_segment | 4.845450 |

Рис. 7. **Наиболее значимые переменные при алгоритме *CatBoost***

Fig. 7. **The most significant variables in the CatBoost algorithm**

***Построение нейронной сети***

Искусственные нейронные сети – это распределенные и параллельные системы из нейронов и связей между ними, способные к адаптивному обучению путем реакции на положительные и отрицательные воздействия [17; 18].

В основе построения сети лежит преобразователь, называемый искусственным нейроном или просто нейроном по аналогии с его биологическим прототипом, который представляет собой определенную функцию (линейную классификацию или регрессию) с неопределенным количеством входных данных и одним «выходом». Нейрон состоит из входов, весов и сумматора. Веса – показатели, позволяющие отмечать степень важности признаков. Числа, поступающие на входы, умножаются на соответствующие им веса (вектор параметров модели – *w*), после чего произведения суммируются в сумматоре. Для получения итогового результата путем преобразования результата, полученного в сумматоре, применяется функция активации. При этом использовать стоит ту функцию, с которой процесс обучения и сходимость будут происходить быстрее. Подбор гиперпараметров крайне важен и будет влиять на сходимость нейронной сети, для которой они выставляются. Сходимость нейронной сети говорит о том, правильно ли были подобраны гиперпараметры и архитектура самой сети относительно поставленной задачи[1].

В нашем исследовании практические эксперименты по подбору значений параметров модели по количеству нейронов на слоях нейронной сети показали, что нейросеть будет условно оптимальной при наличии входного слоя с 64 нейронами и одним скрытым слоем с 32 нейронами. Для входных и скрытых нейронов была выбрана функция активации *relu*. На выходе нейронной сети – 1 слой с сигмоидной функцией активации для того, чтобы получить результат в диапазоне от 0 до 1. Структура нейронной сети представлена на рис. 8.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_68 (Dense)             (None, 64)                3456
_________________________________________________________________
dense_69 (Dense)             (None, 32)                2080
_________________________________________________________________
dense_70 (Dense)             (None, 1)                 33
=================================================================
Total params: 5,569
Trainable params: 5,569
Non-trainable params: 0
_________________________________________________________________
```

Рис. 8. **Структура нейронной сети**

Fig. 8. **Structure of the neural network**

В качестве функции потерь была выбрана функция бинарной кросс-энтропии. В качестве метрики оценки задана бинарная *accuracy (binary_accuracy)*. Время обучения данной модели составило 80 с.

Матрица ошибок, полученная на тестовой выборке, представлена в табл. 10.

[1] *Воронцов К.В.* Искусственные нейронные сети: градиентные методы оптимизации. URL: http://www.machinelearning.ru/wiki/images/e/e1/Voron-ML-ANN-slides.pdf (дата обращения: 15.05.2021).

Таблица 10. **Матрица ошибок для нейронной сети**

Table 10. **Error matrix for the neural network**

| | | *Actual class* | |
|---|---|---|---|
| | | *Positive* (0) | *Negative* (1) |
| *Predicted class* | *Positive* (0) | 13 625 | 1 052 |
| | *Negative (1)* | 2 898 | 5 874 |

На основании матрицы ошибок вычислены значения для метрик классификации (табл. 11).

Нейросеть на тестовой выборке показала процент правильных ответов среди всех прогнозов – 83,1 %. Процент бронирований, названных классификатором отмененными и при этом действительно являющихся отмененными, – 84,8 %. Модель «покрыла» все действительно отмененные бронирования на 67 %. Высокое значение *AUC-ROC* говорит о том, что модель хорошо ранжировала объекты [18].

Таблица 11. **Метрики классификации**

Table 11. **Classification metrics**

| Показатели | На тестовой выборке |
|---|---|
| *Accuracy* | 0,831 |
| *Precision* | 0,848 |
| *Recall* | 0,670 |
| *F*-мера | 0,748 |
| *AUC-ROC* | 0,877 |

Далее представим результаты сравнительного анализа построенных моделей.

***Сравнительный анализ моделей машинного обучения***

Построенные модели машинного обучения были оценены по результатам на тестовой выборке (которая составляет 20 % от исходной). Метрики классификации и метрики ранжирования (*AUC-ROC*) для каждой из четырех моделей машинного обучения приведены в табл. 12.

Таблица 12. **Сравнительный анализ метрик построенных моделей**

Table 12. **Comparative analysis of metrics for the developed models**

| Показатели | *Random Forest* | *XGBoost* | *CatBoost* | Нейронная сеть |
|---|---|---|---|---|
| *Accuracy* | 0,845 | 0,844 | 0,843 | 0,831 |
| *Precision* | 0,873 | 0,853 | 0,855 | 0,848 |
| *Recall* | 0,686 | 0,705 | 0,698 | 0,670 |
| *F*-мера | 0,768 | 0,772 | 0,768 | 0,748 |
| *AUC-ROC* | 0,92 | 0,93 | 0,91 | 0,877 |

Согласно данным табл. 12 все построенные модели дали приблизительно одинаковые результаты на тестовой выборке. Вместе с тем мы полагаем, что в случае прогнозирования отмены бронирования отелей наилучшей моделью должна быть модель с наибольшим *precision*. Метрика *precision* показывает долю объектов, названных классификатором положительными и при этом действительно являющихся положительными: отелю будет важно получить как можно более точное число отмененных бронирований, чтобы грамотно распределить ресурсы. Поэтому наилучшей моделью по этому критерию будет случайный лес (*Random Forest*), демонстрирующий наибольшие значения *precision* и *accuracy* и незначительно уступающий другим моделям в значениях метрик *recall*, *F*-меры и площади *AUC* под *ROC*-кривой.

## ЗАКЛЮЧЕНИЕ

Изучение литературы по проблеме отмены бронирования отелей позволило идентифицировать наиболее эффективные методы машинного обучения, позволяющие спрогнозировать вероятность отмены бронирования. Среди них особо выделяются модели случайный лес (*Random Forest*), нейронные сети, *CatBoost* и *XGBoost*.

В ходе исследования модели были протестированы на открытых данных *"Hotel Booking Demand Dataset"* портала *ScienceDirect*.

Новизна исследования состоит в построении различных моделей прогнозирования бронирования отелей на основе методов машинного обучения и обосновании выбора наилучшей из них.

В ходе сравнения полученных моделей определено, что модель случайного леса (*Random Forest*) наилучшим образом предсказывает отмену бронирования отеля. На тестовой выборке данная модель показала 84,5 % правильных ответов среди всех прогнозов, а значение метрики *precision* составило 87,3 %. Модель «покрыла» все действительно отмененные бронирования на 68,6 %.

Проведенное исследование показывает, что в современных условиях построение моделей прогнозирования отмены бронирования является актуальной задачей, поскольку на рынок онлайн-бронирования отелей оказывает влияние множество факторов, в том числе пандемия. В связи с этим в дальнейшем планируется совершенствование модели случайного леса и других построенных моделей машинного обучения посредством включения дополнительных, ранее не учтенных гиперпараметров.

**СПИСОК ЛИТЕРАТУРЫ**

1. *Smith S.J., Parsa H.G., Bujisic M., van der Rest J-P.* Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry // Journal of Travel and Tourism Marketing. 2015. № 32 (7). P. 886–906. doi: 10.1080/10548408.2015.1063864.

2. *Talluri K.T., van Ryzin G.J.* The theory and practice of revenue management. NY: Kluwer Academic Publishers. 2004. 745 p.

3. *Chen C.-C., Schwartz Z., Vargas P.* The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers // International Journal of Hospitality Management. 2011. № 30 (1). P. 129–135. doi: 10.1016/j.ijhm.2010.03.010.

4. *Huang H.-C., Chang A. Y., Ho C.-C.* Using artificial neural networks to establish a customer-cancellation prediction model // Przeglad Elektrotechniczny. 2013. № 89 (1b). P. 178–180.

5. *Yoon M.G., Lee H.Y., Song Y.S.* Linear approximation approach for a stochastic seat allocation problem with cancellation and refund policy in airlines // Journal of Air Transport Management. 2012. № 23. P. 41–46.

6. *Antonio N., Almeida A., Nunes L.* Predicting hotel booking cancellations to decrease uncertainty and increase revenue // Tourism and Management Studies. 2017. № 13 (2). P. 25–39. doi: 10.18089/tms.2017.13203.

7. *Zeytinci E.* Predicting hotel reservation cancellations with machine learning. 2019. URL: https://towardsdatascience.com/predicting-hotel-cancellations-with-machine-learning-fa669f93e794 (дата обращения: 29.02.2021).

8. *Wingen M.* EDA of bookings and ML to predict cancelations. URL: https://www.kaggle.com/marcuswingen/eda-of-bookings-and-ml-to-predict-cancelations (дата обращения: 30.06.2021).

9. *Denyse T.* Learning Pitstop: Predicting hotel booking cancellations using Classification Techniques. 2020. URL: https://medium.com/tech4she/investigating-factors-affecting-hotel-booking-cancelations-9ec9bf81b0a8 (дата обращения: 20.06.2021).

10. *Michta M., Wojciechowski K.* Story hotel booking cancellations: eXplainable predictions for booking cancellation. URL: https://pbiecek.github.io/xai_stories/story-hotel-booking-cancellations-explainable-predictions-for-booking-cancellation.html#bias-correction (дата обращения: 10.05.2021).

11. *Banza M.* Predicting hotel booking cancellations using machine learning – Step by step guide with real data and python. 2020. URL: https://www.hospitalitynet.org/opinion/4099297.html (дата обращения: 30.06.2021).

12. *Kelman J.* Predicting hotel booking cancellations using customer segmentation and neural networks. 2020. URL: https://medium.com/@julkel/predicting-hotel-booking-cancellations-using-customer-segmentation-and-neural-networks-8a31c2755f5c (дата обращения: 30.06.2021).

13. *Antonio N., Almeida A., Nunes L.* Hotel booking demand datasets // Data in Brief. 2019. Vol. 22. P. 41–49. doi: 10.1016/j.dib.2018.11.126.

14. *Breiman L.* Random forest // Machine Learning. 2001. Vol. 45. P. 5–32. doi: 10.1023/A:1010933404324.

15. *Morde V.* XGBoost algorithm: Long may she reign! 2019. URL: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d (дата обращения: 25.07.2021).

16. *Dorogush A.V., Ershov V., Gulin A.* CatBoost: Gradient boosting with categorical features support // Workshop on ML Systems at NIPS. 2017.

17. *Sharma A.V.* Understanding activation functions in neural networks. 2017. URL: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0 (дата обращения: 30.06.2021).

18. *Powers D.M.W.* Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation // Journal of Machine Learning Technologies. 2011. Vol. 2, Iss. 1. P. 37–63.

## СВЕДЕНИЯ ОБ АВТОРАХ

Елена Ивановна Русакова – студентка факультета экономики, менеджмента и бизнес-информатики, Национальный исследовательский университет «Высшая школа экономики», Пермский филиал (Россия, 614070, г. Пермь, ул. Студенческая, 38; e-mail: elena.rusakova.2000@mail.ru).

Марина Владимировна Радионова – кандидат физико-математических наук, доцент, доцент кафедры информационных систем и математических методов в экономике, Пермский государственный национальный исследовательский университет (Россия, 614990, г. Пермь, ул. Букирева, 15; e-mail: m.radionova812@gmail.com).

## REFERENCES

1. Smith S.J., Parsa H.G., Bujisic M., van der Rest J-P. Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry. *Journal of Travel and Tourism Marketing*, 2015, no. 32 (7), pp. 886–906. doi: 10.1080/10548408.2015.1063864.

2. Talluri K.T., van Ryzin G.J. *The theory and practice of revenue management*. New York, Kluwer Academic Publishers, 2004. 745 p.

3. Chen C.-C., Schwartz Z., Vargas P. The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, 2011, no. 30 (1), pp. 129–135. doi: 10.1016/j.ijhm.2010.03.010.

4. Huang H.-C., Chang A. Y., Ho C.-C. Using artificial neural networks to establish a customer-cancellation prediction model. *Przeglad Elektrotechniczny*, 2013, no. 89 (1b), pp. 178–180.

5. Yoon M.G., Lee H.Y., Song Y.S. Linear approximation approach for a stochastic seat allocation problem with cancellation and refund policy in airlines. *Journal of Air Transport Management*, 2012, no. 23, pp. 41–46.

6. Antonio N., Almeida A., Nunes L. Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism and Management Studies*, 2017, no. 13 (2), pp. 25–39. doi: 10.18089/tms.2017.13203.

7. Zeytinci E. *Predicting hotel reservation cancellations with machine learning*. Available at: https://towardsdatascience.com/predicting-hotel-cancellations-with-machine-learning-fa669f93e794 (accessed 29.02.2021).

8. Wingen M. *EDA of bookings and ML to predict cancelations*. Available at: https://www.kaggle.com/marcuswingen/eda-of-bookings-and-ml-to-predict-cancelations (accessed 30.06.2021).

9. Denyse T. *Learning Pitstop: Predicting hotel booking cancellations using Classification Techniques*. 2020. Available at: https://medium.com/tech4she/investigating-factors-affecting-hotel-booking-cancelations-9ec9bf81b0a8 (accessed 20.06.2021).

10. Michta M., Wojciechowski K. *Story hotel booking cancellations: eXplainable predictions for booking cancellation*. Available at: https://pbiecek.github.io/xai_stories/story-hotel-booking-cancellations-explainable-predictions-for-booking-cancellation.html#bias-correction (accessed 10.05.2021).

11. Banza M. *Predicting hotel booking cancellations using machine learning – Step by step guide with real data and python*. 2020. Available at: https://www.hospitalitynet.org/opinion/4099297.html (accessed 30.06.2021).

12. Kelman J. *Predicting hotel booking cancellations using customer segmentation and neural networks*. 2020. Available at: https://medium.com/@julkel/predicting-hotel-booking-cancellations-using-customer-segmentation-and-neural-networks-8a31c2755f5c (accessed 30.06.2021).

13. Antonio N., Almeida A., Nunes L. Hotel booking demand datasets. *Data in Brief*, 2019, vol. 22, pp. 41–49. doi: 10.1016/j.dib.2018.11.126.

14. Breiman L. Random forest. *Machine Learning*, 2001, vol. 45, pp. 5–32. doi: 10.1023/A:1010933404324.

15. Morde V. *XGBoost algorithm: Long may she reign!* 2019. Available at: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d (accessed 25.07.2021).

16. Dorogush A.V., Ershov V., Gulin A. CatBoost: Gradient boosting with categorical features support. *Workshop on ML Systems at NIPS*. 2017.

17. Sharma A.V. U*nderstanding activation functions in neural networks*. 2017. Available at: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0 (accessed 30.06.2021).

18. Powers D.M.W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2011, vol. 2, iss. 1, pp. 37–63.

## INFORMATION ABOUT THE AUTHORS

Elena Ivanovna Rusakova – Student of the Faculty of Economics, Management and Business Informatics, National Research University "Higher School of Economics", Perm Branch (38, Studencheskaya st., Perm, 614070, Russia; e-mail: elena.rusakova.2000@mail.ru).

Marina Vladimirovna Radionova – Candidate of Physics and Mathematics, Associate Professor, Assistant Professor at the Department of Information Systems and Mathematical Methods in Economics, Perm State University (15, Bukireva st., Perm, 614990, Russia; e-mail: m.radionova812@gmail.com).

*Статья поступила в редакцию 30.07.2021, принята к печати 15.12.2021*

*Received July 30, 2021; accepted December 15, 2021*