ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

магистерская диссертация

Сервис прогнозирования отмены в системе интернет-бронирования отелей

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

На протяжении десятилетий гостиничная индустрия была ключевой отраслью во многих развитых странах. Ожидается, что с непрерывным ростом международного туризма гостиничная индустрия будет расти устойчивыми темпами. Несмотря на эту перспективу, в отрасли давно существуют проблемы. Одной из наиболее важных проблем является отмена бронирования отелей. Данная проблема может быть серьезной для владельцев этих отелей, особенно если это происходит регулярно. Когда клиенты отменяют бронирование в последний момент, отель может потерять значительную прибыль и иметь проблемы с управлением загрузкой номеров и точным прогнозированием своих будущих доходов. Владельцы отелей также могут столкнуться с проблемами, связанными с управлением своей репутации. Если клиенты недовольны отменой бронирования или политикой возврата денег, они могут оставить отрицательный отзыв в Интернете, что может повредить репутации отеля и привести к убыткам. Таким образом, частые отмены могут быстро стать дорогостоящими и негативно повлиять на работу отеля. И в целом, владельцы отелей должны принимать меры для снижения рисков и управления своими бронированиями более эффективно.

В свете этого были проведены исследования, чтобы предсказать, произойдет ли отмена бронирования или нет. Мы стремимся расширить такие исследования, используя методы машинного обучения, чтобы классифицировать, будет ли отменено бронирование или нет. Чтобы решить эту проблему бинарной классификации, мы используем логистическую регрессию, дерево решений, случайный лес и классификатор голосования.

Модели не только предсказывают, будет ли отменено бронирование или нет, они также предоставят дополнительную информацию о том, какие независимые переменные играют важную роль в определении того, будет ли отменено бронирование. Такая информация будет полезна менеджерам и владельцам отелей. Это позволит более точно прогнозировать доходы и эффективно применять методы управления доходами. Кроме того,

эта информация может иметь ключевое значение для предотвращения будущих отмен, поскольку менеджеры будут лучше понимать переменные, которые играют роль в отмене бронирования.

Текущие практики показывают, что отели склонны использовать строгие правила отмены бронирования и стратегии чрезмерного бронирования, что, как правило, негативно сказывается на доходах и репутации отеля. Наша модель системы, основанная на машинном обучении, может помочь гостиницам преодолеть эту проблему. В последнее время машинное обучение приобрело большое значение в мире автоматизации. Машинное обучение не только сокращает объем ручной работы, но и генерирует соответствующие результаты. Вместо жесткого кодирования машинное обучение помогает машине самой учиться, принимать решения и выдавать точные результаты. Для обработки данных для получения лучших результатов используются сценарии реального времени. После того как модель готова, ее необходимо обучить на наборе данных. Это используется для проверки точности алгоритма и способности предсказывать результат. Получив желаемые результаты, легко предпринять необходимые шаги, чтобы избежать дальнейших отмен и предложить клиентам лучший опыт.

Глава 1. Постановка цели и задач

Понятно, что после отмены заказа практически ничего неможно делать. Это создает дискомфорт для многих компаний в мире электронной коммерции в целом и для систем интернет-бронирования отелей в частности. Это вызывает желание принять меры предосторожности. Таким образом, прогнозирование заказов, которые могут быть отменены, и потенциальные попытки предотвратить эти отмены создадут прибавочную стоимость для компаний.

Отмены также ограничивают производство точных прогнозов, что является важным инструментом с точки зрения эффективности управления доходами. Чтобы обойти проблемы, вызванные отменой заказа, мы пытаемся помочь бизнесу, построив сервис, который с высокой точностью прогнозирует, будет ли отменен заказ. Результаты позволяют бизнесу точно прогнозировать чистый спрос и строить более точные прогнозы, улучшать политику отмены, определять лучшие тактики и, таким образом, использовать более настойчивые стратегии ценообразования и распределения запасов.

Важность этой работы заключается в том, что на управление и обработку каждого бронирования компания тратит ресурсы. При этом система интернет-бронирования отелей не получает прибыль, если клиенты отменяют бронирование. Поэтому наша задача — построить поведенческую модель покупателей для прогнозирования вероятности отмены. Такая модель позволит компании оптимизировать бизнес-процесс и минимизировать убытки в зависимости от стадии обработки бронирования.

Наша работа заключается в том, чтобы решить задачу прогнозирования вероятности отказа клиента от бронирования. В результате у нас будет модель решающая поставленную задачу.

Цель работы:

1. Уменьшение количества отмен бронирования отелей и повышение удовлетворенности клиентов отелей и увеличение числа повторных бронирований.

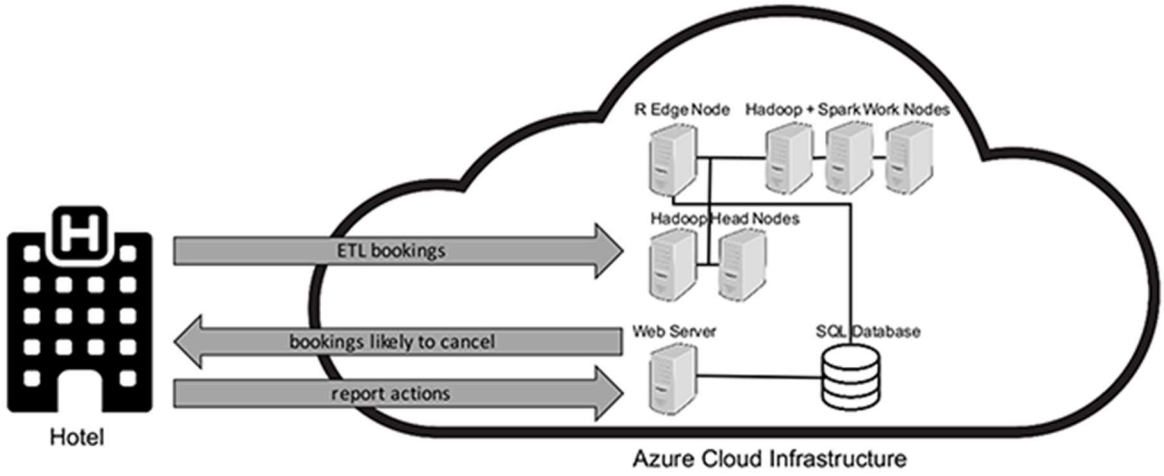
Для достижения поставленной цели необходимо решение следующих **задач**:

1. Анализ причин отмены бронирования и выявление основных факторов, влияющих на это.
2. Сбор и анализ данных о клиентах, которые отменяют бронирование, чтобы определить, какие параметры клиентов связаны с более вероятной отменой бронирования.
3. Разработка модели машинного обучения для прогнозирования вероятности отмены бронирования на основе имеющихся данных.
4. Создание системы уведомлений, которая будет оповещать отели о возможных отменах бронирования клиентами.
5. Изучение опыта других отелей и анализ лучших практик по снижению отмен бронирования.
6. Тестирование и оптимизация модели с целью улучшения ее точности и эффективности.
7. Внедрение модели в систему бронирования отеля.
8. Оценка эффективности модели и ее влияния на снижение отмен бронирования, повышение удовлетворенности клиентов и увеличение числа повторных бронирований.

Глава 2. Предварительная обработка данных

2.1 Архитектура интеграции сервиса с системой интернет-бронирования отелей

Чтобы соответствовать требованиям и спецификациям и сделать систему технически надежной и способной к адекватной производительности, система будет построена на базе облачной платформы Microsoft Azure, используя преимущества нескольких компонентов и технологий с открытым исходным кодом, доступных в качестве сервисов на этой платформе (Рисунок 1): один кластер HDInsight на базе Linux, Hadoop и Spark с Python Server. Этот компонент обеспечивает обработку больших данных на базе Hadoop/Spark, позволяет использовать Python в контексте Spark и использует преимущества производительности XGBoost [1], используя возможности кластера для распределения обработки между различными машинами; одна база данных SQL для обработки и хранения журналов всех операций. В этом компоненте также будут храниться все результаты прогнозирования.



Архитектура интеграции сервиса с системой интернет-бронирования отелей. Рисунок 1.

Для построения модели, которая решает проблему отмены бронирования отелей с помощью машинного обучения, необходимо получить следующие данные из различных источников:

1. Данные о бронировании отелей: Для анализа проблемы отмены бронирования отелей необходимо иметь данные о бронировании отелей, включая даты бронирования, количество гостей, тип номера, длительность пребывания и стоимость бронирования.

2. Данные о клиентах: Для определения причин отмены бронирования необходимо иметь данные о клиентах, которые отменяют бронирование, включая возраст, пол, место жительства, цель поездки, источник информации об отеле и т.д.

3. Данные об отменах и типе бронирования: Для построения модели машинного обучения необходимы данные об отменах и типе бронирования, включая даты отмены, причины отмены, количество отмен и т.д. Например, если гость забронировал номер с возможностью бесплатной отмены бронирования до определенной даты, то вероятность отмены может быть выше, чем если бы гость забронировал номер с неотменяемой бронью. Поэтому данные о типе бронирования, включая условия бронирования, такие как сроки отмены, размер штрафа за отмену и т.д., могут помочь в определении причин отмены и выработке стратегии уменьшения количества отмен бронирования.

4. Данные о компании, осуществившей бронирование: Эти данные также могут быть важны для построения модели, которая решает проблему отмены бронирования отелей. Например, гости, которые бронируют через определенную компанию посредника, могут иметь отличающиеся предпочтения и потребности, чем гости, бронирующие напрямую на сайте отеля. Также, компании посредники могут предоставлять свои услуги с дополнительными условиями, которые могут повлиять на вероятность отмены бронирования.

5. Данные об услугах и ценах конкурентов: Данные об услугах и ценах конкурентов могут помочь определить, почему гости выбирают один отель вместо другого, и что можно сделать для увеличения конкурентоспособности отеля.

2.2 Извлечение данных из первичного источника

Данные, использованные в этом проекте, были изначально взяты из статьи "Наборы данных для бронирования отелей" (Hotel Booking Demand Datasets), написанной Нуно Антонио, Аной Алмейдой и Луисом Нунесом для журнала “Кратко о данных” (Data in Brief), выпуск 22, февраль 2019 года. Данные были загружены и обработаны Томасом Моком и Антуаном Бишатом в рамках проекта #TidyTuesday, который проходил в течение недели с 11 февраля 2020 года, в последующем они были опубликованы на Github и Kaggle.

Эта статья описывает два набора данных, содержащих данные о бронировании отелей. Один из отелей (H1) – это курортный отель, а другой (H2) – городской отель. Оба набора данных имеют одинаковую структуру, с 31 переменной, описывающей 40 060 наблюдений H1 и 79 330 наблюдений H2. Каждое наблюдение представляет собой бронирование отеля. Оба набора данных содержат информацию о бронированиях за три года и охватывают бронирования, предполагаемые на период с 1 июля 2015 года по 31 августа 2017 года, включая как выполненные бронирования, так и отмененные. Однако городской отель (41,90%) показал более высокий уровень отмены бронирования, чем курортный отель (27,69%). Так как это настоящие данные от отелей, все элементы данных, касающиеся идентификации отеля или клиента, были удалены. В связи с нехваткой настоящих данных о бизнесе для научных и образовательных целей, эти наборы данных не только полезны для этого проекта, но также могут иметь важное значение для исследований и обучения в области управления доходами, машинного обучения или сборе данных, а также в других областях.

Ценность этого набора данных:

- Описательная аналитика может быть использована для дальнейшего понимания закономерностей, тенденций и аномалий в данных;

- Можно использовать для проведения исследований по различным проблемам, таким как: прогнозирование отмены бронирования, сегментация клиентов, удовлетворенность клиентов, сезонность и многим другим;
- Исследователи могут использовать эти наборы данных для сравнения моделей прогнозирования отмены бронирования с известными результатами (например, [2]);
- Исследователи в области машинного обучения могут использовать наборы данных для оценки производительности различных алгоритмов для решения одного и того же типа проблем (классификация, сегментация или т.д.);
- Преподаватели могут использовать наборы данных для решения задач классификации или сегментации в машинном обучении;
- Преподаватели могут использовать наборы данных для обучения статистике или интеллектуальному сбору данных.

Данные были получены непосредственно с серверов баз данных “PMS отелей (Property Management System)” путем выполнения запроса на языке TSQL (Транзакционно-структурированный язык запросов) в SQL Server Studio Manager - интегрированной среде для управления базами данных Microsoft SQL[3]. Подробное описание извлеченных переменных, их происхождения и инженерных процедур, использованных при их создании, полностью описано в данной статье [4].

2.3 Анализ исходных данных

Структура полученного набора данных показана в Таблице 1. Набор данных содержит случаи бронирования за три года и охватывает бронирования, которые должны поступить в период с июля 2015 года по август 2017 года. Первичный набор данных содержал 119 390 записей.

Структура исходного набора данных. Таблица 1.

Переменные	Класс (тип данных)	Описание поля
hotel	character (categorial)	Отель (H1 = курортный отель (Resort Hotel) или H2 = городской отель (City Hotel)
is_canceled	double (categorial)	Значение, указывающее, было ли бронирование отменено (1) или нет (0)
lead_time	double (integer)	Количество дней, прошедших между датой ввода бронирования в PMS и датой прибытия
arrival_date_year	double (integer)	Год даты прибытия
arrival_date_month	character (categorial)	Месяц даты прибытия
arrival_date_week_number	double (integer)	Номер недели года для даты прибытия
arrival_date_day_of_month	double (integer)	День даты прибытия
stays_in_weekend_nights	double (integer)	Количество ночей в выходные дни (суббота или воскресенье), когда гость останавливался или бронировал проживание в отеле
stays_in_week_nights	double (integer)	Количество ночей в неделю (с понедельника по пятницу), когда гость останавливался или бронировал проживание в отеле
adults	double (integer)	Количество взрослых
children	double (integer)	Количество детей
babies	double (integer)	Количество младенцев
meal	character (categorial)	Тип заказанного питания. Категории представлены в виде стандартных пакетов питания в отелях: 1. Undefined/SC - без пакета питания; 2. BB - Завтрак в постель (Bed & Breakfast); 3. HB - Полупансион (Half board) (завтрак и еще один прием пищи - обычно ужин); 4. FB - Полный пансион (Full board) (завтрак, обед и ужин)
country	character (categorial)	Страна происхождения. Категории представлены в формате ISO 3155-3:2013

market_segment	character (categorial)	Обозначение сегмента рынка. В категориях термин "TA" означает "Турагенты (Travel Agents)", а "TO" - "Туроператоры (Tour Operators)"
distribution_channel	character (categorial)	Канал распространения бронирования. "TA" означает "Турагенты (Travel Agents)", а "TO" - "Туроператоры (Tour Operators)"
is_repeated_guest	double (categorial)	Значение, указывающее, было ли название бронирования от повторного гостя (1) или нет (0)
previous_cancellations	double (integer)	Количество предыдущих бронирований, которые были отменены клиентом до текущего бронирования
previous_bookings_not_canceled	double (integer)	Количество предыдущих бронирований, не отмененных клиентом до текущего бронирования
reserved_room_type	character (categorial)	Код типа забронированного номера. Код представлен вместо обозначения в целях анонимности
assigned_room_type	character (categorial)	Код типа номера, назначенного для бронирования. Иногда назначенный тип номера отличается от забронированного по причинам работы отеля (например, избыточное бронирование) или по запросу клиента. Код представлен вместо обозначения в целях анонимности
booking_changes	double (integer)	Количество изменений/поправок, внесенных в бронирование с момента ввода бронирования в PMS до момента заселения или аннуляции
deposit_type	character (categorial)	Индикация о том, внес ли клиент депозит для гарантии бронирования. Эта переменная может принимать три категории: No Deposit - депозит не вносился; Non Refund - депозит был внесен в размере общей стоимости проживания; Refundable - депозит был внесен в размере меньше общей стоимости проживания
agent	character (categorial)	Идентификатор туристического агентства, осуществившего бронирование
company	character (categorial)	Идентификатор компании/организации, осуществившей бронирование или ответственной за оплату бронирования. Идентификатор указывается вместо обозначения в целях анонимности
days_in_waiting_list	double (integer)	Количество дней, в течение которых бронирование находилось в списке ожидания, прежде чем оно было подтверждено клиенту
customer_type	character (categorial)	Тип бронирования, предполагающий одну из четырех категорий: Контракт - если с бронированием связано выделение или другой тип контракта; Группа - если

		бронирование связано с группой; Транзитный - если бронирование не является частью группы или контракта и не связано с другим переходным бронированием; Транзитный-партнер - если бронирование является транзитным, но связано как минимум с другим транзитным бронированием
adr	double (numeric)	Среднесуточный тариф (Average Daily Rate), определяемый путем деления суммы всех операций по размещению на общее количество ночей проживания
required_car_parking_spaces	double (integer)	Количество парковочных мест, необходимых клиенту
total_of_special_requests	double (integer)	Количество специальных запросов, сделанных клиентом (например, двухместная кровать или высокий этаж)
reservation_status	character (categorial)	Последний статус бронирования, предполагающий одну из трех категорий: Canceled - бронирование было отменено клиентом; Check-Out - клиент зарегистрировался в отеле, но уже уехал; No-Show - клиент не зарегистрировался в отеле и не проинформировал отель о причине
reservation_status_date	double (date)	Дата, на которую был установлен последний статус. Эта переменная может быть использована вместе со статусом бронирования (reservation_status), чтобы понять, когда было отменено бронирование или когда клиент выехал из отеля

Из данных, приведенных в Таблице 1 не все признаки будут являться входными, поскольку для прогнозирования не важны следующие значения:

- arrival_date_year: Год даты прибытия
- arrival_date_month: Месяц даты прибытия
- arrival_date_week_number: Номер недели года для даты прибытия
- arrival_date_day_of_month: День даты прибытия

Удаляем эти четыре признака, потому что они представляют дату прибытия с различными уровнями детализации.

- reserved_room_type: Код типа забронированного номера. Код представлен вместо обозначения в целях анонимности
- assigned_room_type: Код типа номера, назначенного для бронирования. Иногда назначенный тип номера отличается от забронированного по причинам работы отеля (например, избыточное бронирование) или по запросу клиента. Код представлен вместо обозначения в целях анонимности

Эти два признака представляют тип назначенного и зарезервированного номера. Они часто коррелируют, и удаление одного из них поможет уменьшить мультиколлинеарность в наших данных

- reservation_status_date: Дата, на которую был установлен последний статус. Эта переменная может быть использована вместе со статусом бронирования (ReservationStatus), чтобы понять, когда было отменено бронирование или когда клиент выехал из отеля

Удаляем этот признак, так как он описывает дату обновления статуса бронирования и имеет слабую связь с отменой бронирования.

- previous_cancellations: Количество предыдущих бронирований, которые были отменены клиентом до текущего бронирования
- previous_bookings_not_canceled: Количество предыдущих бронирований, не отмененных клиентом до текущего бронирования

Удаляем эти две переменные, так как их влияние на текущее бронирование может быть незначительным. Они могут быть полезными при анализе поведения клиентов, но не являются критическими факторами при прогнозировании отмены текущего бронирования.

Удаление этих менее важных характеристик позволит нам сосредоточиться на наиболее важных факторах, которые влияют на отмену бронирования, улучшить производительность модели и упростить интерпретацию результатов.

Поэтому, первичными входными признаками для модели прогнозирования будут являться:

- …
- …
- …
- …
- …

Полученных признаков недостаточно для дальнейшего прогнозирования, ввиду чего необходимо сформировать другие входные признаки на основании текущих.

- …

В рамках рассматриваемой задачи подготовка данных включает в себя:

- анализ данных на наличие пропусков, ошибок, дубликатов, противоречий и т. п.;
- реализацию процедур обработки пропусков;
- обработку ошибочных записей;
- удаление дубликатов;
- поиск и удаление противоречий.

Таким образом, основные этапы реализации модуля прогнозирования представлены на Рисунке 2.

2.4 Реализация процедур очистки данных

Глава 3. Формирование входных признаков прогнозирования

3.1 Разведочный анализ данных (EDA)

A thorough awareness of client wants, and preferences is necessary for the competitive and continually expanding hospitality sector to deliver top-notch customer service. The amount of data hotels create has dramatically expanded with the introduction of Internet reservation systems. Here we will perform an exploratory data analysis (EDA) of our dataset used in this research. The data consists of information on hotel reservations, client demographics, and data on cancellations. This analysis aims to examine the features of hotel reservations and clients to offer more insights into our dataset.

We performed EDA using Python, together with the Pandas and Matplotlib packages. A Jupyter notebook was used to import and perform initial data cleaning. The distributions and connections between the variables in the dataset were then examined using univariate and bivariate analyses.

The dataset's information is arranged into 32 variables, which may be further broken down into the following four groups:

- hotel information,
- booking information,
- guest information,
- and transaction information.

The hotel data includes the hotel's type, address, and ID. The reservation information includes the reservation date, the arrival date, the lead time (i.e., the number of days between the reservation date and the arrival date), the total number of guests, including children and infants, the kind of accommodations, and the type of meals. The information on the guest also contains the

type of visitor, such as transient (i.e., a one-time client), contract (i.e., the reservation is necessary in accordance with the terms of the contract), or group (i.e., the booking is part of a group booking). The transaction information consists of the full cost of the reservation, the amount of the deposit, and the type of deposit made (i.e., refundable or non-refundable). Details on the visitor's needs and the sector from which the reservation was made are also included in the collection (e.g., travel agent, corporate, direct booking, etc.). Whether or not the reservation was canceled, as well as the day it occurred, are also included in the dataset.

3.1.1 Характеристики бронирования

The variables that offer information on bookings, such as hotel type, booking status, and booking details, are covered in the section on booking characteristics. The following variables were examined:

Hotel Type

The dataset has two different types of hotels: resort hotels and city hotels. There are 40,060 reservations for the Resort Hotel and 79,330 for the City Hotel. Due to their location in cities and ease of access for business travelers, city hotels are more well-liked than resort hotels.

Bookings by Hotel Type. Figure 3.

From the data on hotel type, and as shown in the figure below, the cancellations made by customers in city hotels is higher in comparison with the cancellations made in the resort. According to the dataset, the city hotel had more reservations than the resort hotel, with a total of 79,330 reservations as opposed to 40,060 reservations for the resort hotel. Yet, compared to the resort hotel's cancellation rate of 27.7%, the city hotel's cancellation rate was marginally higher at 37.04%. This shows that even while the city hotel had a greater number of reservations overall, it also experienced more cancellations than the resort hotel.

Cancellations by hotel type. Figure 4.

Reservation Status

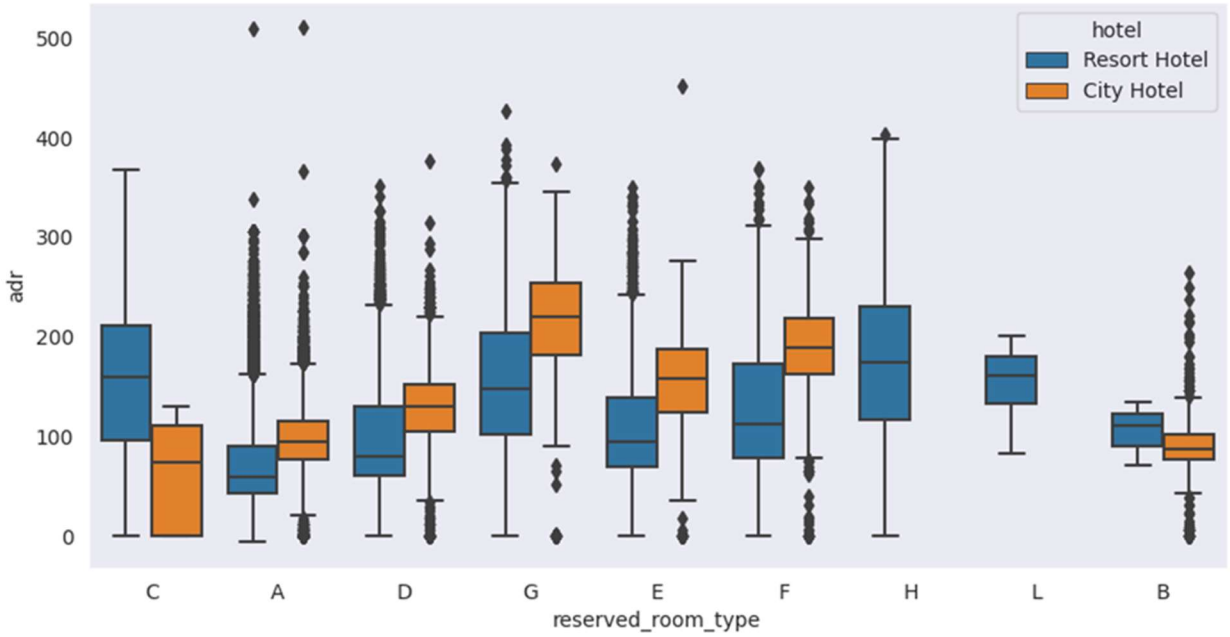
Canceled, Check-Out, and No-Show are the three possible booking statuses in the dataset. 63% of reservations were either Check-Out or No-Shows, while 37% were Canceled. This suggests that cancellations are a serious hotel problem and that management and reduction techniques are required.

Booking Details

The stay's start and end dates, the type of accommodation reserved, and the total number of guests are all factors in the booking details. With a standard deviation of 106 days, the lead time for reservations is typically 104 days. With a standard variation of 2.0 nights, the average length of stay is 3.3 nights. Type A reservations are the most frequent, followed by type D reservations. The dataset also details the number of adults, kids, and babies in each reservation. Bookings typically include 1.9 adults, 0.1 kids, and 0.01 newborns.

Room Type

The dataset includes details about the reservation's room type. The breakdown of reservations by room type is shown in the figure below. The bulk of reservations are for resort hotel basic rooms. This is seen in Figure too.



Bookings by Room Type. Figure 5.

3.1.2 Характеристики клиентов

3.2 Формирование входного набора данных

3.3 Исследование значимости признаков

СПИСОК ЛИТЕРАТУРЫ

1. Chen, T and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp. 785–794. Available at http://dl.acm.org/citation.cfm?id=2939785 [Last accessed 22 January 2017]. DOI: https://doi.org/10.1145/2939672.2939785
2. N. Antonio, A. Almeida, L. Nunes, Predicting hotel bookings cancellation with a machine learning classification model, in: Proceedings of the 16th IEEE International Conference Machine Learning Application, IEEE, Cancun, Mexico, pp. 1049–1054. Available at https://ieeexplore.ieee.org/document/8260781 DOI: https://doi.org/10.1109/ICMLA.2017.00-11
3. Microsoft, SQL Server Management Studio (SSMS), (2017). 〈https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms〉 (accessed 24 March 2018).
4. Antonio N, de Almeida A, Nunes L. Hotel booking demand datasets. Data Brief. 2018 Nov 29;22:41-49. DOI: https://doi.org/10.1016/j.dib.2018.11.126. PMID: 30581903; PMCID: PMC6297060.