

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ  
Федеральное государственное автономное образовательное учреждение  
высшего образования  
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
*ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК*

**ОТЧЁТ**  
по практике

**ТЕХНОЛОГИЧЕСКАЯ (ПРОЕКТНО-ТЕХНОЛОГИЧЕСКАЯ) ПРАКТИКА**  
Магистратура

**РАЗРАБОТКА ВЕБ-СЕРВИСА ПРОГНОЗИРОВАНИЯ  
ОТМЕНЫ БРОНИРОВАНИЯ ОТЕЛЕЙ**

09.04.03 Прикладная информатика «информационные системы анализа данных»

Выполнил  
обучающийся 2 курса,  
М.М. Гола

---

(подпись)

---

(ФИО)

Руководитель  
к.т.н., доцент  
М.С. Цыганова

---

(подпись)

---

(ФИО)

Тюмень, 2023

# ОГЛАВЛЕНИЕ

## ВВЕДЕНИЕ

### Глава 1. Постановка цели и задач

### Глава 2. Подготовка входного набора

#### 2.1 Архитектура интеграции сервиса с системой интернет-бронирования отелей

#### 2.2 Извлечение данных из первичного источника

#### 2.3 Анализ исходных данных

##### 2.3.1 Работа с отсутствующими значениями и преобразование типа данных

##### 2.3.2 Feature Selection и Feature Engineering

### Глава 3. Формирование входных признаков модели прогнозирования

#### 3.1 Разведочный анализ данных (EDA)

##### 3.1.1 Характеристики бронирования

##### 3.1.2 Характеристики клиентов

#### 3.2 Реализация процедур очистки данных

#### 3.2 Формирование входного набора данных

#### 3.3 Исследование значимости признаков

### Глава 4. Разработка модуля прогнозирования отмены бронирования

#### 4.1 Методы избавления от дисбаланса классов

#### 4.2 Выбор модели прогнозирования

#### 4.3 Показатели качества прогноза

#### 4.4 Построение модели прогнозирования

#### 4.5 Интеграция модуля прогнозирования система интернет-бронирования отелей

## ЗАКЛЮЧЕНИЕ

## СПИСОК ЛИТЕРАТУРЫ

## Приложение

## 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 ухудшением репутации отеля. Если клиенты недовольны отменой бронирования или политикой возврата денег, они могут оставить отрицательный отзыв в Интернете, что может повредить имидж 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 нельзя сделать. 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 лучше определять чистый спрос и строить более точные прогнозы на распределение ресурсов компании, улучшать политику отмены, выявлять 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:**

Конечная цель исследования состоит в повышении эффективности управления бронированием в отеле. Для этого будет разработан сервис, который предсказывает вероятность отмены бронирования, что позволит оптимизировать политику отмены, улучшить тактики управления, и использовать более эффективные стратегии ценообразования и управления запасами.

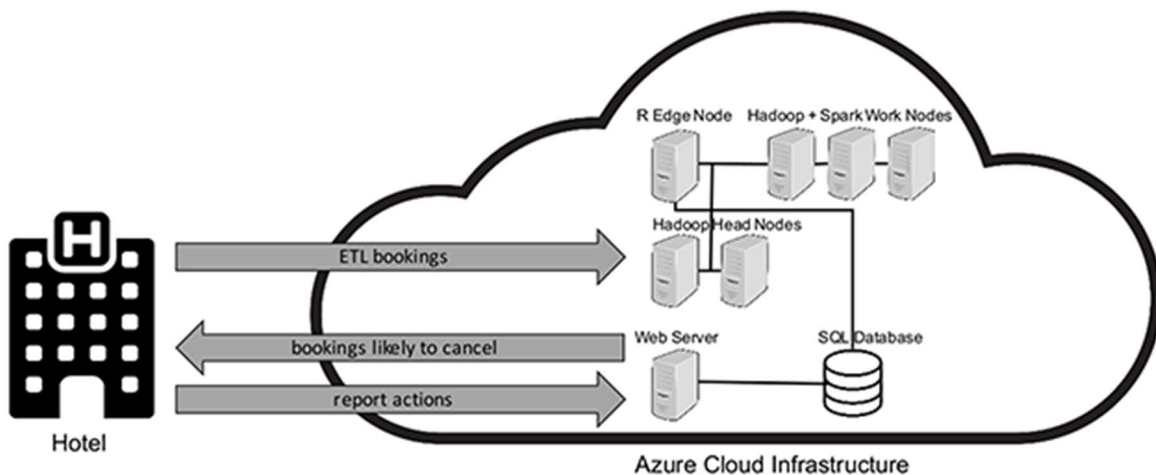
Для достижения поставленной цели необходимо решение следующих **задач**:

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

## Глава 2. Подготовка входного набора

### 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 источника (Hotel Booking Demand Datasets):

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 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.
3. 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.

## 2.2 Извлечение данных из первичного источника

Для обучения модели в этом проекте использованы данные, изначально взятые 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 сбора 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 (categorical)	Отель (H1 = курортный отель (Resort Hotel) или H2 = городской отель (City Hotel))
is_canceled	double (categorical)	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 (categorical)	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 (categorical)	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 (categorical)	Страна происхождения. Категории представлены в формате ISO 3155-3:2013
market_segment	character (categorical)	Обозначение сегмента рынка. В категориях термин "TA" означает "Турагенты (Travel Agents)", а "TO" - "Туроператоры (Tour Operators)"
distribution_channel	character (categorical)	Канал распространения бронирования. "TA" означает "Турагенты (Travel Agents)", а "TO" - "Туроператоры (Tour Operators)"
is_repeated_guest	double (categorical)	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 (categorical)	Код типа забронированного номера. Код представлен вместо обозначения в целях анонимности
assigned_room_type	character (categorical)	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 (categorical)	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 (categorical)	Идентификатор туристического агентства, осуществившего бронирование
company	character (categorical)	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 (categorical)	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 (categorical)	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
-------------------------	---------------	--

### 2.3.1 Работа с отсутствующими значениями и преобразование типа данных

Для работы с отсутствующими значениями были найдены следующие пропущенные значения:

- В столбце company было найдено 112 593 пропущенных значения;
- В столбце agent было найдено 16 340 пропущенных значений;
- В столбце country было найдено 488 пропущенных значений;
- И в столбце children было найдено только 4 пропущенных значения.

Для работы с отсутствующими значениями были приняты следующие процедуры:

- Удалим строки, в которых нет взрослых, детей и младенцев;
- Если идентификатор агента или компании не равен нулю, просто заменим на 0;
- Для отсутствующих значений в столбце country заменим его OTHER;
- Для отсутствующего значения детей, заменим его округленным средним значением.

Что касается преобразования типа данных, то была предпринята только следующая процедура.

- Преобразуем тип данных этих столбцов из float в integer.

### 2.3.2 Feature Selection и Feature Engineering

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.

- reservation\_status: 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

Удаляем колонку reservation\_status потому что она сообщает нам, было ли бронирование отменено.

- 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.

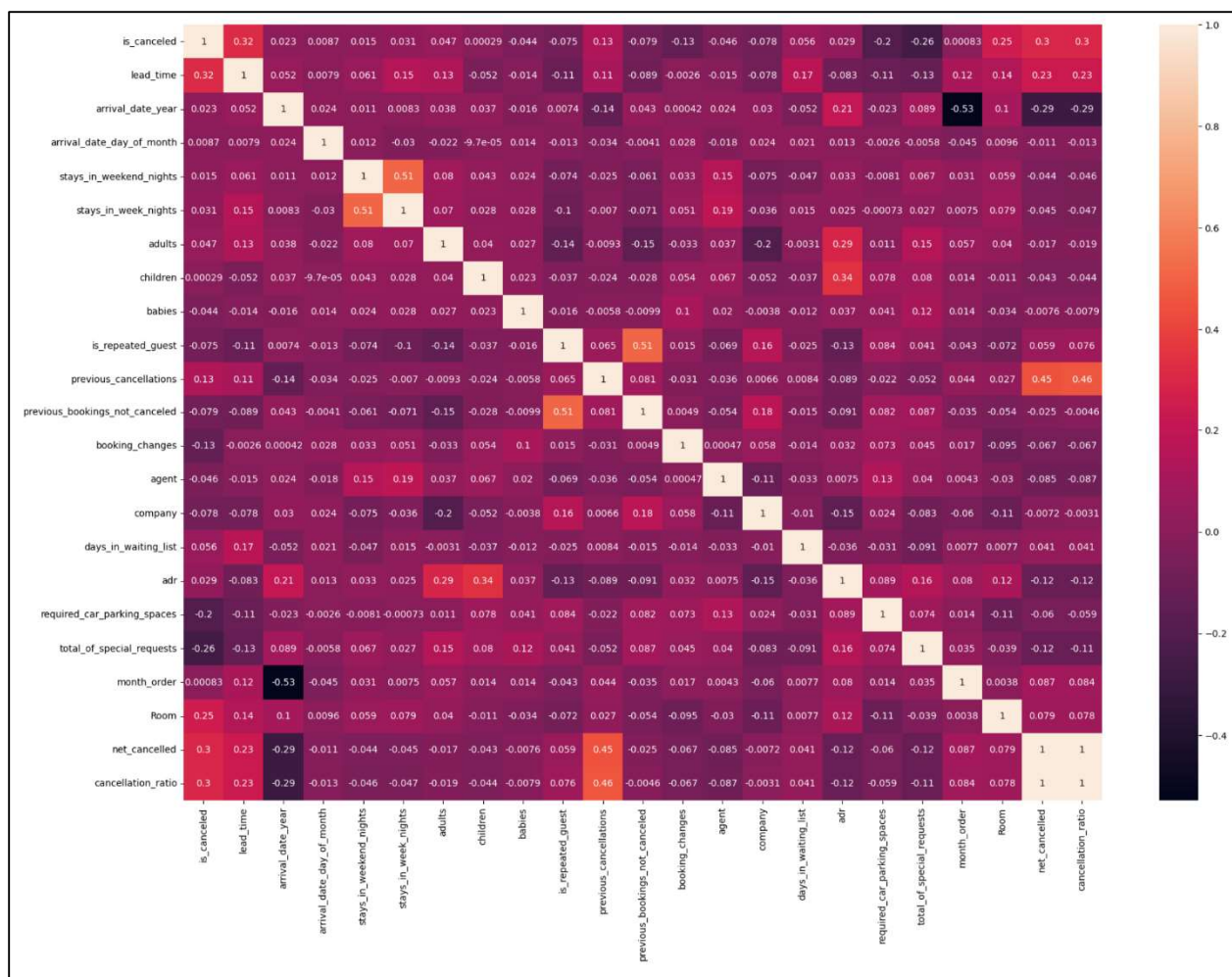
Более того, в ходе анализа выяснилось, что в признаке `country` много малочисленных категорий. Это не особо хорошо для будущей модели (по аналогии с количественными признаками имеем "тяжелый хвост"). Имеет смысл задуматься об объединении малочисленных категорий. После анализа было видно что наибольший вклад вносят первые 5 стран. Было решено объединить остальные страны в одну категорию `OTHER`. Это позволит в будущем получить модель, более устойчивую к изменению данных. Значения первых 5 стран и всех остальных стран, объединенных в одну категорию `OTHER`, теперь присваиваются новому признаку `countries`, после чего столбец `country` может быть удален.

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

- Как описано выше, создаем новый признак `countries`, где он включает в себя все первые 5 стран из столбца `country`, а также все остальные страны, объединенные в одну категорию `OTHER`;
- Создаем новый признак `Room`, который содержит 1, если гость получил тот же номер, который был зарезервирован, иначе 0;
- Создаем новый признак `net_cancelled`, который будет содержать 1, если гость отменил больше бронирований в прошлом чем количество бронирований, которые он не отменял, в противном случае 0;
- И наконец, создаем новый признак `cancellation_ratio` на основании характеристик `previous_cancellations` и `previous_bookings_not_canceled` который показывает процент отмен.

Для дальнейшего анализа была создана тепловая карта (Рисунок 2), чтобы увидеть корреляцию с колонками.





Тепловая карта, показывающая корреляцию между столбцами признаков. Рисунок 2.

Анализируя тепловую карту видим, что новая характеристика `cancellation_ratio` гораздо лучше коррелирует с целевой меткой `is_canceled` чем `previous_bookings_not_canceled` и `previous_cancellations`. В то же время видим, что корреляция между `cancellation_ratio` и `net_cancelled` составляет 1. Таким образом мы можем удалить все три показателя `cancellation_ratio`, `previous_cancellations` и `previous_bookings_not_canceled`, оставив только `net_cancelled`.

Из тепловой карты также можно сделать вывод, что, чтобы избежать переобучения модели удалим так же наименее значимые характеристики корреляция которых с целевой меткой is\_canceled меньше 0,1

- stays\_in\_weekend\_nights: Количество ночей в выходные дни (суббота или воскресенье), когда гость останавливался или бронировал проживание в отеле
- stays\_in\_week\_nights: Количество ночей в неделю (с понедельника по пятницу), когда гость останавливался или бронировал проживание в отеле
- adults: Количество взрослых
- children: Количество детей
- babies: Количество младенцев
- is\_repeated\_guest: Значение, указывающее, было ли название бронирования от повторного гостя (1) или нет (0)
- agent: Идентификатор туристического агентства, осуществившего бронирование
- company: Идентификатор компании/организации, осуществившей бронирование или ответственной за оплату бронирования. Идентификатор указывается вместо обозначения в целях анонимности
- days\_in\_waiting\_list: Количество дней, в течение которых бронирование находилось в списке ожидания, прежде чем оно было подтверждено клиенту
- adr: Среднесуточный тариф (Average Daily Rate), определяемый путем деления суммы всех операций по размещению на общее количество ночей проживания

Поэтому, первичными входными признаками для модели прогнозирования будут являться – признаки из исходного источника данных плюс все новые признаки, описанные выше, которые были созданы:

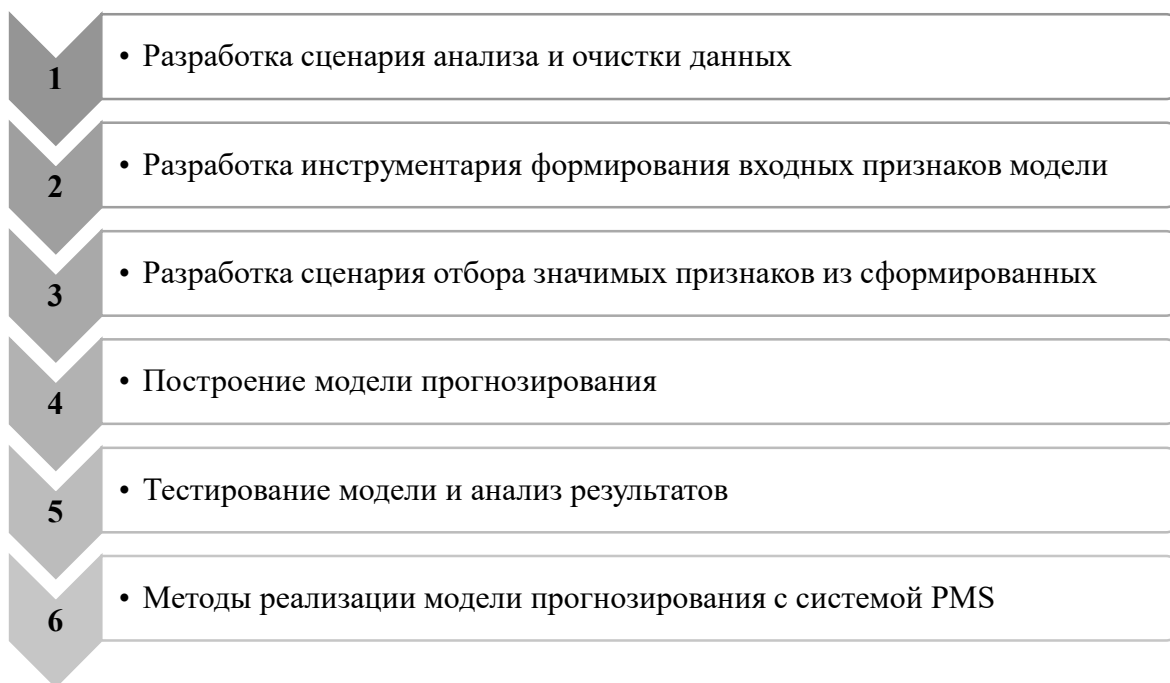
- hotel: Отель (H1 = курортный отель (Resort Hotel) или H2 = городской отель (City Hotel))
- lead\_time: Количество дней, прошедших между датой ввода бронирования в PMS и датой прибытия

- meal: 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)
- market\_segment: Обозначение сегмента рынка. В категориях термин "ТА" означает "Турагенты (Travel Agents)", а "ТО" – "Туроператоры (Tour Operators)"
- distribution\_channel: Канал распространения бронирования. "ТА" означает "Турагенты (Travel Agents)", а "ТО" – "Туроператоры (Tour Operators)"
- booking\_changes: Количество изменений/поправок, внесенных в бронирование с момента ввода бронирования в PMS до момента заселения или аннуляции
- deposit\_type: 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
- customer\_type: 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
- required\_car\_parking\_spaces: Количество парковочных мест, необходимых клиенту
- total\_of\_special\_requests: Количество специальных запросов, сделанных клиентом (например, двухместная кровать или высокий этаж)

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 3.



Блок-схема по основным этапам разработки модуля прогнозирования. Рисунок 3.

## 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.