МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК
Кафедра программной и системной инженерии

РЕКОМЕНДОВАНО К ЗАЩИТЕ В ГЭК
Заведующий кафедрой
д.т.н., профессор
_____ А.Г. Ивашко
_____ 2023 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
магистерская диссертация

РАЗРАБОТКА WEB СЕРВИСА "ПРОГНОЗИРОВАНИЯ ЗАГРУЖЕННОСТИ ОТЕЛЕЙ"

09.04.03 Прикладная информатика
Магистерская программа «Информационные системы анализа данных»

Выполнил работу
студент 2 курса
очной формы обучения — *(Подпись)* — Гола Момен М А

Научный руководитель
к.т.н, доцент — *(Подпись)* — Цыганова Мария Сергеевна

Рецензент
к.ф.-м.н, доцент — *(Подпись)* — Семихин Дмитрий Витальевич

г. Тюмень
2023

ОГЛАВЛЕНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

PMS – Property Management System (Система управления недвижимостью)

HOS – Hotel Operating System (Система управления отелем)

PNR – Personal Name Record (Данные физического лица)

EDA – Exploratory Data Analysis (Разведочный анализ данных)

GDS – Global Distribution System (Глобальная система распределения)

ROC – Receiver Operating Characteristic (Рабочая характеристика приёмника)

AUC – Area Under Curve (Площадь под кривой)

SVM – Support Vector Machine (Метод опорных векторов)

RF – Random Forest (Случайный лес)

VC – Voting Classifier (Классификатор голосования)

LG – Logistic Regression (Логистическая регрессия)

LR – Linear Regression (Линейная регрессия)

DT – Decision Tree (Древо решений)

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.

Для успешной подготовки и защиты выпускной квалификационной работы обучающимся использовались средства и методы физической культуры и спорта с целью поддержания должного уровня физической подготовленности, обеспечивающую высокую умственную и физической работоспособность. В режим

рабочего дня включались различные формы организации занятий физической культурой (физкультпаузы, физкультминутки, занятия избранным видом спорта) с целью профилактики утомления, появления хронических заболеваний и нормализации деятельности различных систем организма.

В рамках подготовки к защите выпускной квалификационной работы автором созданы и поддерживались безопасные условия жизнедеятельности, учитывающие возможность возникновении чрезвычайных ситуаций.

Глава 1. Постановка цели и задач

1.1 Описание предметной области

Управление доходами определяется как «применение информационных систем и стратегий ценообразования для выделения нужной мощности нужному клиенту по правильной цене в нужное время». [1] Первоначально разработанное в 1966 году в авиационной отрасли, управление доходами постепенно было перенято другими отраслями услуг, такими как отели, аренда автомобилей, поля для гольфа и казино. [2]

В индустрии гостеприимства (разделение номеров) определение управления доходами было адаптировано к «предоставлению нужного номера нужному гостю по правильной цене в нужное время через правильный канал распределения» [3]. Поскольку отели имеют фиксированный инвентарь и продают скоропортящийся «продукт», чтобы предоставить нужный номер нужному гостю в нужное время, отели принимают бронирование заранее.

Бронирование представляет собой договор между клиентом и отелем [4]. Этот договор дает клиенту право использовать услугу в будущем по установленной цене, обычно с возможностью расторгнуть договор (отменить бронирование) до предоставления услуги.

Хотя заблаговременное бронирование считается основным предиктором прогнозируемой эффективности отеля [5], этот вариант отмены услуги подвергает отель риску, поскольку отель должен гарантировать номера клиентам, которые соблюдают свои бронирования, но в то же время, должен нести альтернативные издержки свободных мощностей, когда клиент отменяет бронирование или не появляется. Несмотря на то, что между неявкой и отменой есть некоторые различия, для целей настоящего исследования и то, и другое будет рассматриваться как

отмена. Отмена происходит, когда клиент расторгает договор до своего прибытия, а неявка происходит, когда клиент не информирует отель и не регистрируется или не заселится.

Конечно, некоторые из этих отмен бронирования происходят по понятным причинам: перенос деловых встреч, перенос отпусков, болезни, плохие погодные условия и другие факторы. Но, как определили Чен и Се [6] и Чен, Шварц и Варгас [7], в настоящее время большая часть этих отмен происходит из-за клиентов, стремящихся к выгодным сделкам, которые настроены на поиск лучших предложений. Иногда эти клиенты продолжают искать более выгодные предложения того же продукта/услуги после размещения бронирования.

В некоторых случаях клиенты даже делают несколько бронирований, чтобы сохранить свои варианты, а затем отменяют все, кроме одного. Как поясняют Таллури и Ван Райзин [4], «клиенты также ценят возможность отмены бронирования. Действительно, бронирование с возможностью отмены дает клиентам лучшее из обоих миров — преимущество заблаговременной фиксации доступности и возможность отказа в случае изменения их планов или предпочтений».

В качестве способа управления рисками, связанными с отменой бронирования, отели применяют сочетание политик избыточного бронирования и отмены. Тем не менее, как избыточное бронирование, так и политика отмены бронирования могут нанести ущерб отелю. Избыточное бронирование, когда клиенту не разрешается регистрироваться в отеле, который он забронировал ранее, вынуждает отель отказывать клиенту в предоставлении услуг, что может быть ужасным опытом для клиента.

Этот опыт может оказать негативное влияние как на имидж отеля, так и на немедленный доход [8], не говоря уже о потенциальной потере будущих доходов от

недовольных клиентов, которые больше не будут бронировать проживание в отеле [3]. Правила отмены бронирования, особенно правила с невозвратной оплатой, могут не только сократить количество бронирований, но и уменьшить доход из-за значительных скидок на цену [5].

Чтобы преодолеть негативное влияние, вызванное избыточным бронированием, и введением жестких правил отмены бронирования, которые могут составлять до 20% от общего числа бронирований, полученных отелями [9], или до 60% в аэропортах/придорожных отелях [10], авторы предлагают использовать технологическую основу, основанную на модели прогнозирования отмены бронирования, разработанную в рамках науки о данных.

Эта модель, предсказывающая вероятность отмены каждого бронирования, может помочь сделать более точные прогнозы и уменьшить неопределенность в управленческих решениях. Это очень важно в контексте управления доходами, для распределения запасов и принятия решений о ценообразовании, но также важно в других контекстах управления, таких как кадровое обеспечение, закупки расходных материалов или решения о прибыльности/движении денежных средств [11]. В то же время, разработав модель прогнозирования классификации, т. е. модель, которая классифицирует вероятность отмены каждого бронирования, позволить гостиницам действовать в отношении этих конкретных бронирований, чтобы попытаться избежать их отмены или, в некоторых случаях, заставить ее.

Разработка модели прогнозирования отмены бронирования соответствует тому, что было признано Чианом и другими [2], согласно которому управление доходами должно использовать математические и прогнозные модели, чтобы лучше использовать имеющиеся данные и технологии. Это также подтверждается исследованием, проведенным Каймсом [12] на пятистах специалистах по управлению доходами. Это исследование показывает, что управление доходами будет все больше ориентироваться на стратегию и технологии и что специалисты

по управлению доходами должны обладать лучшими аналитическими и коммуникативными навыками. Исследование показало, что все специалисты по управлению доходами должны обладать аналитическими и коммуникативными навыками, которые лежат в основе науки о данных: прикладная математика, операционные исследования, машинное обучение, статистика, базы данных, интеллектуальный анализ данных, визуализация данных и отличные беглость общения/презентации, дополненная глубоким пониманием предметной области [13, 14, 15]. Как относительно новая дисциплина, наука о данных использует огромные объемы данных, которыми мы располагаем, и доступность более качественных и дешевых вычислительных мощностей. Эти факторы сделали возможным улучшение существующих алгоритмов прогнозирования и способствовали разработке новых и лучших алгоритмов, особенно в области машинного обучения.

Используя данные системы управления недвижимостью (PMS), это исследование направлено на то, чтобы продемонстрировать, как наука о данных может применяться в контексте управления загруженностей отелей и доходами для прогнозирования отмен бронирований. Более того, чтобы показать, что отмена бронирования не обязательно означает неопределенность в прогнозировании занятости номеров и прогнозировании доходов. Это достигается за счет:

1. Определение того, какие признаки в базах данных PMS отеля способствуют прогнозированию вероятности отмены бронирования.
2. Построение модели для классификации бронирований с высокой вероятностью отмены и использование этой информации для прогнозирования отмены.
3. Понимание того, подходит ли одна модель прогнозирования для всех отелей или для каждого отеля необходимо создавать особую модель.

Мехротра и Раттли [3] признали, что «хорошее прогнозирование спроса является ключевым аспектом управления доходами» (стр. 8). Таллури и Ван Райзин

[4] также признали важность прогнозирования в управлении доходами, заявив, что системы управления доходами требуют прогнозирования количества и что «его эффективность в решающей степени зависит от качества этих прогнозов» (стр. 407). Эти и другие авторы, такие как Иванов и Жечев [16] или Моралес и Ван [9], определили прогнозирование спроса как один из аспектов, в которых прогнозирование важно. За этой необходимостью прогнозирования спроса стоят отмены бронирований, потому что в индустрии гостеприимства, как и в других сферах услуг, которые работают с предварительными бронированиями, они не отражают истинный спрос на их услуги, поскольку часто имеет место значительное количество отмен [9, 10]. Чистый спрос — спрос за вычетом отмен необходимо точно прогнозировать, чтобы можно было принимать соответствующие решения по управлению спросом.

Отмена бронирования уже имеет хорошо известную совокупность знаний в области управления доходами, применяемых в сфере услуг, и в частности в индустрии гостеприимства. Тем не менее, в последние годы, с ростом влияния Интернета на то, как клиенты ищут и покупают туристические услуги [8, 17], количество исследований по этой теме увеличилось, особенно исследований по темам, связанным с контролем, используемым для смягчения последствий отмены при распределении доходов и запасов, правила отмены и избыточное бронирование [4, 11, 18]. Тем не менее, существует мало литературы по теме прогнозирования отмены бронирования для индустрии гостеприимства.

Помимо работ Хуанг, Чанг, Хо и других [19], которые используют данные о ресторанах, Юн, Ли и Сонг [20], которые используют смоделированные данные об отелях, и Лю [10], которые используют реальные данные об отелях, все в других работах используются данные физического лица (PNR) — стандарт, разработанный Международной ассоциацией воздушного транспорта (Международная организация гражданской авиации, 2010). Использование данных PNR не является

странной практикой, поскольку исследования прогнозов отмен в основном доступны для авиационной отрасли или не являются специфичными для отрасли, но используют данные авиакомпаний.

Это преобладание работ, касающихся авиационной отрасли, можно объяснить не только более длительным применением управления доходами, но и высоким уровнем отмен и неявок при бронировании авиабилетов, которые составляют от 30% [21] до 50% [4] всех бронирований. Хотя авиаперевозки и гостиничный бизнес являются сферами услуг и имеют много общего, есть аспекты, которые их отличают, например факторы, побуждающие клиентов выбирать поставщиков услуг. В авиационной отрасли ключевыми факторами являются цена, качество обслуживания (в полете), имидж авиакомпании (особенно с точки зрения безопасности), программы лояльности и доступность к транспортным узлам в конечном пункте назначения [22, 23], в то время как в гостиничном бизнесе значение этих факторов меняется, и в игру вступают другие факторы, такие как социальная репутация, местоположение, чистота.

В области науки о данных, особенно в области машинного обучения, задачи прогнозного моделирования с учителем обычно делятся на два типа задач [24]: регрессия, когда измерение результата является количественным (например, прогнозирование процента отмены бронирований от общего числа бронирований), или в качестве классификации, когда результатом является класс/категория (например, прогнозирование того, будет ли конкретное бронирование «отменено» или «не будет отменено»).

В то время как некоторые из ранее опубликованных работ по прогнозированию отмены бронирования рассматривают эту проблему как проблему классификации, в большинстве работ она рассматривается как проблема регрессии. Тем не менее, даже некоторые из первых, такие как Моралес и Ван [9], сосредоточились на прогнозе глобальной частоты отмен, а не на вероятности

отмены каждого бронирования. На самом деле, Моралес и Ван заявили, что «трудно представить, что можно с высокой точностью предсказать, будет ли бронирование отменено или нет, просто взглянув на информацию PNR» (стр. 556). Однако, как показано в следующих разделах, возможна классификация того, будет ли бронирование отменено, особенно если подходящие данные PMS используются в сочетании с существующими алгоритмами прогнозирования машинного обучения.

Еще одна причина рассматривать прогноз отмены бронирований как проблему прогнозирования классификации заключается в том, что из результата прогнозирования класса/категории также можно получить количественный результат. Например, количество бронирований, которые, по прогнозам, будут отменены в определенный день, можно вычесть из спроса и получить чистый спрос или рассчитать коэффициент отмены бронирований путем деления общего количества бронирований, которые, по прогнозам, могут быть отменены, на общее количество бронирование на сутки.

Моралес и Ван [9] также заявили, что «в контексте управления доходами классификация или даже вероятность отмены отдельного бронирования не имеет значения» (стр. 556), что не соответствует тому, что обычно утверждается в управлении доходами. теория. В управлении доходами регистрация отмен, по крайней мере, по сегментам рынка или типам бронирований, является важным инструментом для выявления закономерностей и создания более точных прогнозов [3, 18] и более эффективных политик избыточного бронирования и отмены. Что касается избыточного бронирования, как описано Таллури и Ван Райзин [4], причина этого «с исторической точки зрения, избыточное бронирование является старейшей — и, с финансовой точки зрения, одной из самых успешных — практик управления доходами» (стр. 129). В прошлом некоторые авторы считали жесткую политику аннулирования эффективным инструментом против аннулирования [25] (политики, которые требовали полной оплаты или какой-либо гарантии в момент

бронирования или, по крайней мере, налагали какие-то финансовые санкции в случае отмены) . В настоящее время правила отмены, предусматривающие такие виды штрафов или строгие условия отмены, считаются препятствием для продаж и могут негативно сказаться на доходах.

Используя науку о данных для разработки модели прогнозирования отмены бронирования как все более важной проблемы в контексте управления доходами и как части структуры системы доходов, это исследование демонстрирует важность сочетания науки и технологий в процессе принятия решений. Таллури и Ван Райзин [4] признали, что «наука и технологии позволяют управлять спросом в таких масштабах и сложности, которые были бы немыслимы с помощью ручных средств» (стр. 5).

1.2 Описание 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 и описание входного набора данных

2.1 Архитектура интеграции сервиса

Что касается архитектуры интеграции сервиса с системой интернет-бронирования отелей, и чтобы соответствовать требованиям и спецификациям и сделать систему технически надежной и способной к адекватной производительности, система будет построена на взаимодействии PMS-отелей и веб-сервиса, который будет предоставлять прогнозы на основе модели машинного обучения, которую предстоит построить. Воспользовавшись несколькими компонентами и технологиями с открытым исходным кодом, а также платной подпиской на инструмент Obviously AI (SaaS). Архитектуру высокого уровня можно увидеть на рисунке 1 ниже. Отели должны будут предоставлять свои данные о бронировании в соответствии с той же структурой ввода данных модели машинного обучения через POST-запрос вебхука, а взамен смогут получить вероятность отмены бронирования.

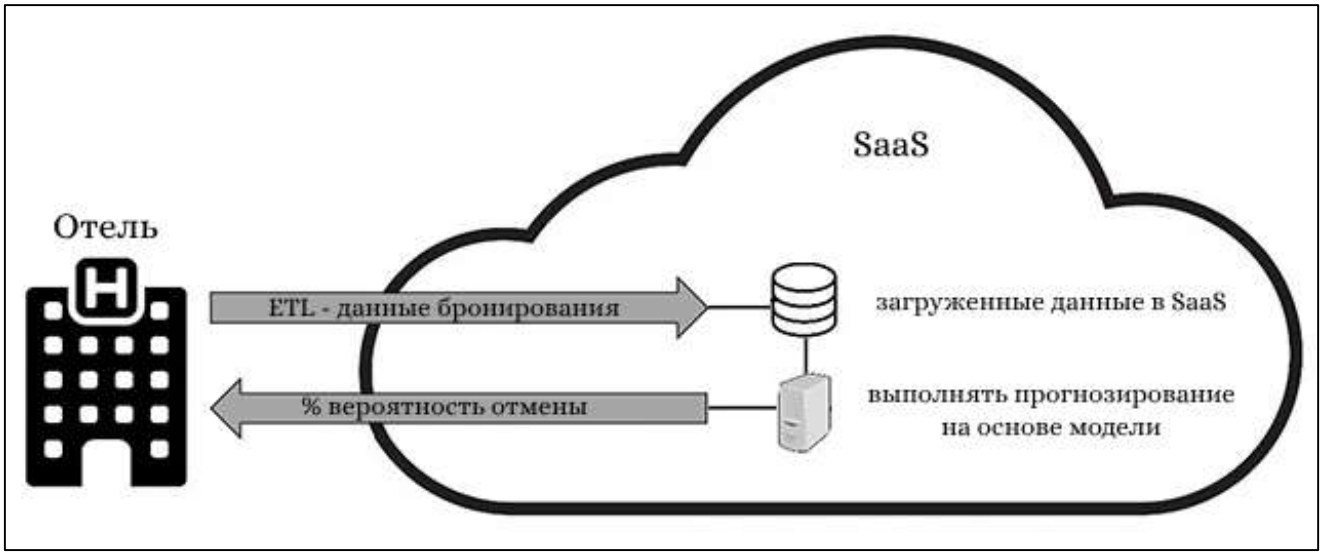


Рисунок 1. Архитектура интеграции сервиса с системой бронирования отелей.

Для построения модели, которая решает проблему отмены бронирования отелей с помощью машинного обучения, необходимо получить следующие данные из источника (Hotel Booking Demand Datasets) [29]:

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 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, [27]);
- 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 [28]. Подробное описание извлеченных переменных, их происхождения и инженерных процедур, использованных при их создании, полностью описано в данной статье [29].

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.

Таблица 1. Структура исходного набора данных.

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: • Undefined/SC - без пакета питания; • BB - Завтрак в постель (Bed & Breakfast); • HB - Полупансион (Half board) (завтрак и еще один прием пищи - обычно ужин); • FB - Полный пансион (Full board) (завтрак, обед и ужин)
country	character (categorial)	Страна происхождения. Категории представлены в формате ISO 3155-3:2013 [35]
market_segment	character (categorial)	Обозначение сегмента рынка. • Прямой (Direct) • Корпоративный (Corporate) • Онлайн турагенты (Online TA) • Офлайн турагенты/туроператоры (Offline TA/TO) • Дополнительно (Complementary) • Группы (Groups) • Не определено (Undefined) • Авиация (Aviation)
distribution_channel	character (categorial)	Канал распространения бронирования. • Прямой (Direct) • Корпоративный (Corporate) • Турагенты/туроператоры (TA/TO) • Не определено (Undefined) • Глобальная система распределения (GDS)
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, представлен вместо обозначения в целях анонимности, и включает эти категории: A,B,C,D,E,F,G,H,L,P
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, и включает следующие категории: A,B,C,D,E,F,G,H,I,K,L,P
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 (Contract) - если с бронированием связано выделение или другой тип контракта; • Группа (Group) - если бронирование связано с группой;

		• Скоротечный (Transient) - если бронирование не является частью группы или контракта и не связано с другим скоротечным бронированием; • Скоротечный-партнер (Transient-party) - если бронирование является скоротечным, но связано как минимум с другим скоротечным 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

2.4 Разведочный анализ данных (Exploratory Data Analysis, EDA)

Разведочный анализ данных (Exploratory Data Analysis, EDA) — подход, используемый здесь для анализа набора данных этого проекта и обобщения его основных характеристик, это делается с использованием статистических графиков

и других методов визуализации данных. В качестве инструментов для анализа данных использовались пакеты Numpy и Pandas, а для визуализации и исследования данных использовались пакеты Matplotlib и Seaborn. EDA в основном используется для того, чтобы увидеть, что данные могут сказать нам помимо формального моделирования, и тем самым противопоставляется традиционной проверке гипотез. Прежде чем мы зададим вопросы об отелях, было бы полезно понять демографические данные гостей, то есть страну, дату прибытия, месяц, тип клиента и т. д. Важно изучить эти переменные, чтобы лучше понять данные. Обследование такого масштаба обычно имеет тенденцию к некоторому смещению отбора.

В контексте исследовательского анализа данных, касающихся темы исследования, может возникнуть ряд вопросов, в том числе:

- Откуда приходят гости?
- Сколько гости платят за номер в сутки?
- Как меняется цена в сутки в течение года?
- Какой самый загруженный месяц?
- Процент бронирования за каждый год
- Как долго люди остаются в отелях?
- Бронирования по сегментам рынка
- Сколько бронирований было отменено?
- В каком месяце больше всего отмен?
- Эффект повторного гостя при отмене.
- Количество ночей, проведенных в отелях.
- Тип отеля с более длительным проживанием.
- Самый бронируемый тип отеля.
- Влияние депозита на отмены по сегментам.

- Связь lead time[1] с отменой.
- Ежемесячные клиенты и отмены.
- Соотношение бронирований между разными типами отелей.

В контексте этого раздела будут даны ответы на большинство этих вопросов и запросов, чтобы лучше понять данные этого исследовательского проекта.

Первый вывод, как видно на рисунке 2, который не вызывает удивления, заключается в том, что постоянные гости не отменяют свои бронирования. Конечно, есть некоторые исключения. Также большинство клиентов не являются повторными гостями.

Рисунок 2. Количество отмен по сравнению с повторными гостями.

Затем проводится анализ того, в каком отеле люди любят останавливаться и проводить больше времени. В связи с этим, это будет проверяться по будням и выходным отдельно, т.к. здесь может наблюдаться перекос распределения. Поэтому

[1] lead time: количество дней, прошедших между датой ввода бронирования в PMS и датой прибытия

сначала строится блочная диаграмма для рыночного сегмента, и она будет проверена по количеству ночей в неделю. Сегмент рынка покажет, какой это тип пребывания. Затем то же самое будет сделано для пребывания в выходные дни.

Видно на рисунке 3, что большинство групп сегментов рынка имеют распределение, близкое к нормальному, а некоторые из них имеют высокую асимметрию. Глядя на распределение, можно сделать вывод, что большинство людей не предпочитают оставаться в отеле более 1 недели. Но кажется обычным пребывание в курортных отелях до 12-13 дней. Хотя это зависит от сегментов, пребывание дольше 15 дней, безусловно, создает выбросы для каждого сегмента. Если бы признак общего времени был создан путем суммирования ночей выходного и буднего дня, это было бы яснее, но его можно ясно увидеть, если посмотреть на две визуализации вместе.

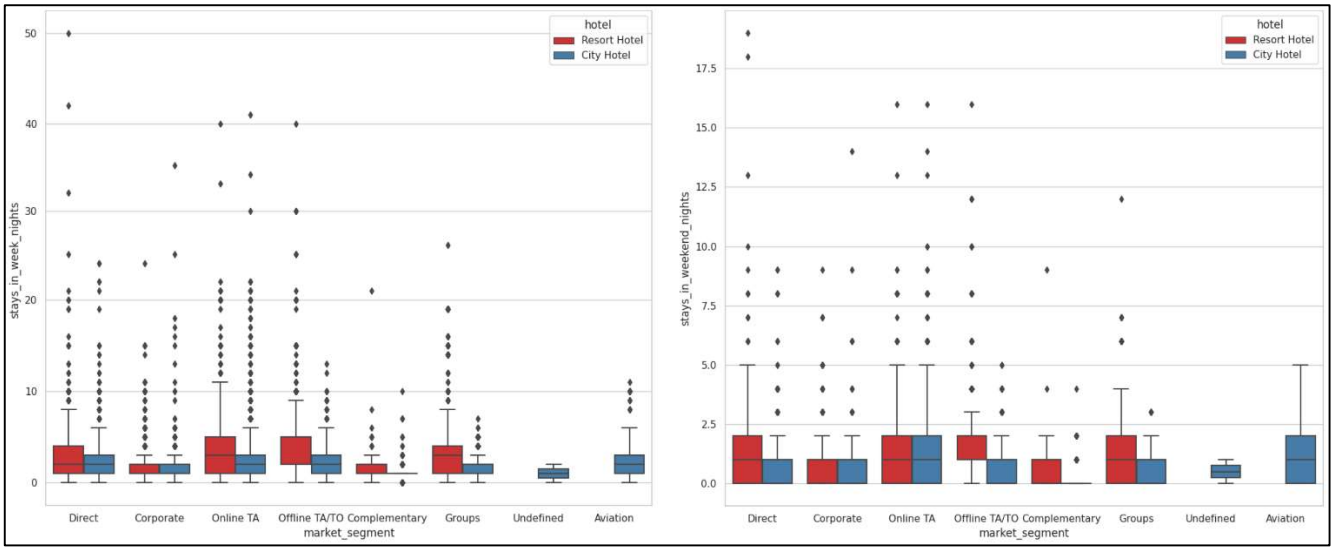


Рисунок 3. Продолжительность пребывания в отелях по сегментам рынка.

Как оказалось, клиенты из авиационного сегмента, похоже, не останавливаются в курортных отелях и имеют относительно меньшее среднее количество дней. Кроме того, средние значения выходных и будних дней примерно

равны. Клиенты в авиационном сегменте, скорее всего, прибудут в ближайшее время в связи с работой. Также, вероятно, большинство аэропортов находятся немного вдали от моря и, скорее всего, ближе к городским отелям. Очевидно, что когда люди едут в курортные отели, они предпочитают оставаться там дольше.

Еще одним важным аспектом является анализ влияния депозита на отмены по сегментам. При рассмотрении офлайновых турагентов и туроператоров (Offline TA/TO) и групп (Groups) единственными обстоятельствами, при которых была получена оплата, были те, при которых прибыли группы. Внесение депозита за значительное количество гостей, которые будут занимать значительную часть вместимости отеля, вполне логично.

Первоначально предполагалось, что сегменты рынка, где применяется депозит, будут иметь более низкий уровень отмены, чем сегменты, где депозит не применяется. Однако оказывается, что это не так, когда мы рассматриваем сокращения по отрезкам в другом представлении.

Как видно на рисунке 4:

- Уровень отмены для групп превышает 50%.
- Уровень отмены для офлайн-TA/TO и онлайн-TA превышает 33%.
- Прямые сегменты имеют более низкий уровень отмены.

Интересно, что, несмотря на внесение депозита, процент отмен в этих краях значительный. Этот сценарий может быть в некоторой степени объяснен тем фактом, что отмена производится совместно, как и бронирование. Показатели отмены онлайн-бронирования типичны для динамичной среды со значительным тиражом.

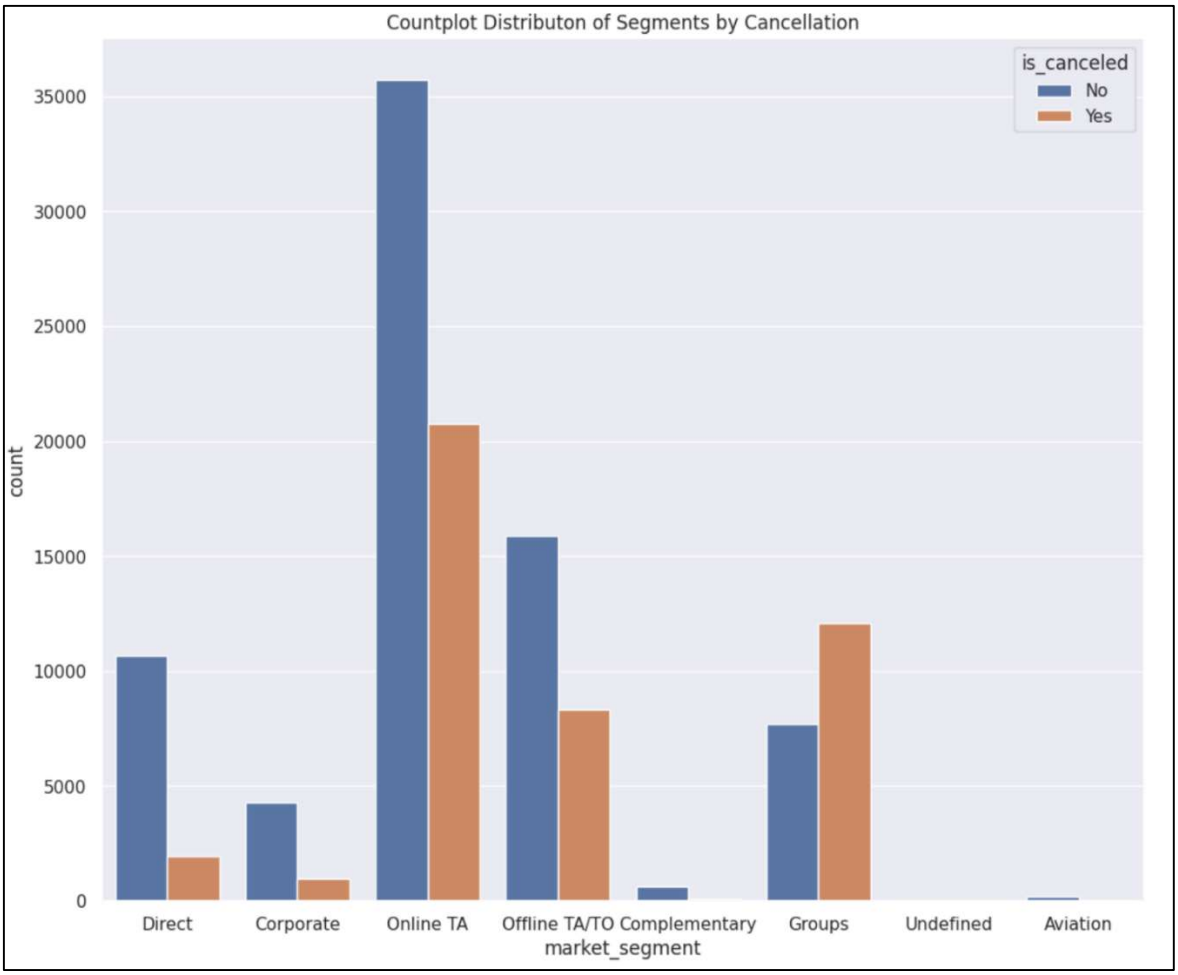


Рисунок 4. Распределение сегментов рынка по отмене бронирования.

Тот факт, что уровень отмен в прямом сегменте настолько низок, является еще одной проблемой, которая привлекла внимание. В том случае, когда люди разговаривают один на один, считается, что в это время установились взаимно доверительные отношения. Это может не относиться к онлайн-транзакциям, однако на этом нельзя слишком задерживаться, так как в игре может быть психологический элемент.

Двигаясь вперед, и чтобы посмотреть на соотношение между количеством дней, прошедших между датой бронирования и датой прибытия (время выполнения)

с отменой, была построена кривая плотности времени выполнения по отмене бронирования. (Рисунок 5)

Из кривой плотности видно, что когда время выполнения заказа превышает примерно 60, гости часто отменяют свои бронирования (на кривой видно, что после этой точки процент отмен увеличивается).

Кроме того, 50% бронирований приходится на 100-дневный время выполнения. Это означает, что люди планируют и рассчитывают график отпуска или работы/командировки до 100 дней вперед.

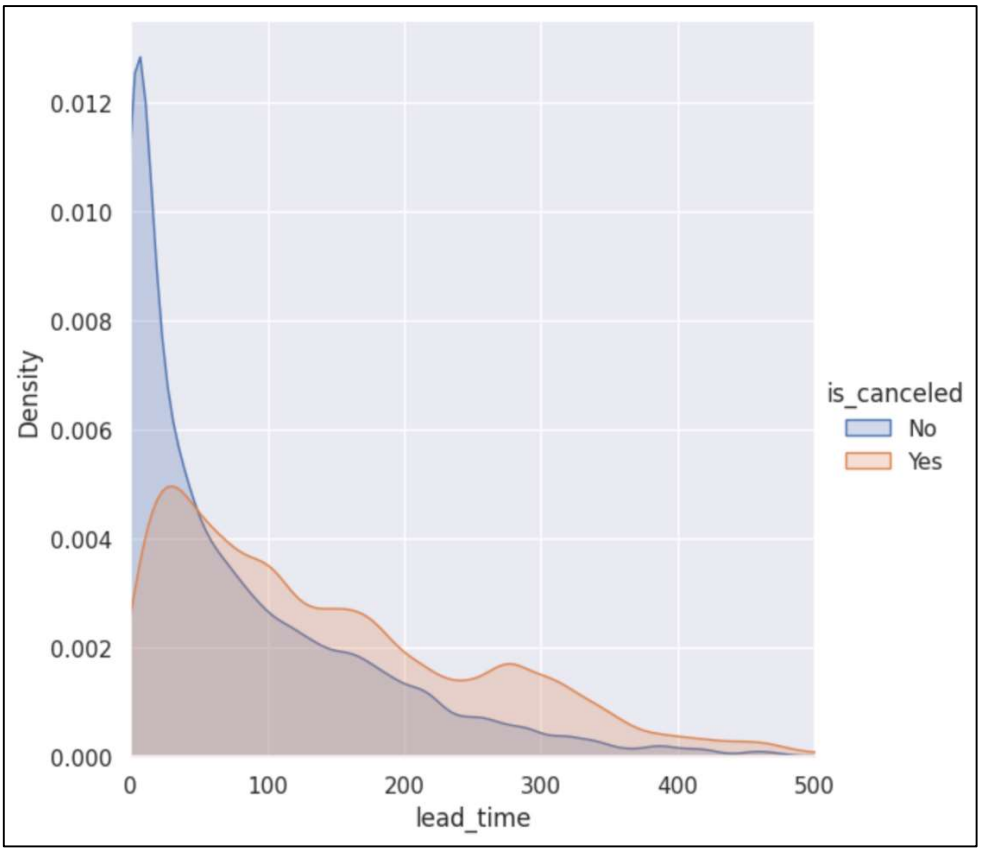


Рисунок 5. Кривая плотности времени выполнения по отмене бронирования.

Другим важным аспектом для анализа является ежемесячное количество гостей, которых принимает каждый отель. Чтобы проанализировать это, была

построена диаграмма распределения, как показано на рисунке 6. Очевидно, городские отели принимают больше гостей в течение года. Курортные отели летом кажутся немного ближе к городским отелям при пропорциональном сравнении.

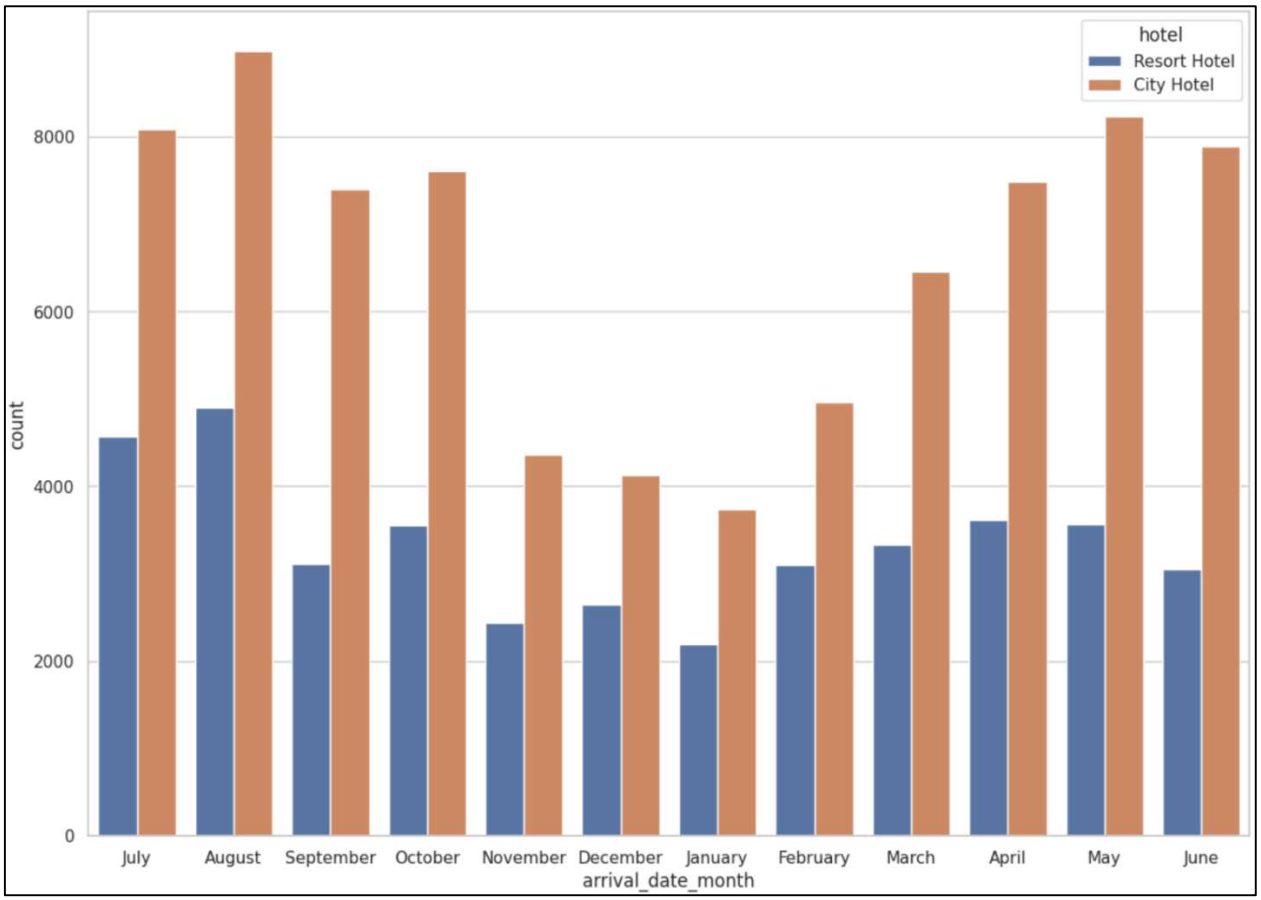


Рисунок 6. Распределение ежемесячных гостей, принятых отелями.

Теперь пришло время взглянуть на количество стран, из которых делается общее количество бронирований, и нанесем на график десять стран с наибольшим количеством бронирований. Это также будет включать бронирования, которые позже были отменены. Анализ показал, что всего в этих отелях осуществляется бронирование из 177 стран. Страны с наибольшим количеством бронирований можно определить с помощью метода value_counts. Здесь только первые 10 стран, из которых поступило наибольшее количество бронирований. (Рисунок 7)

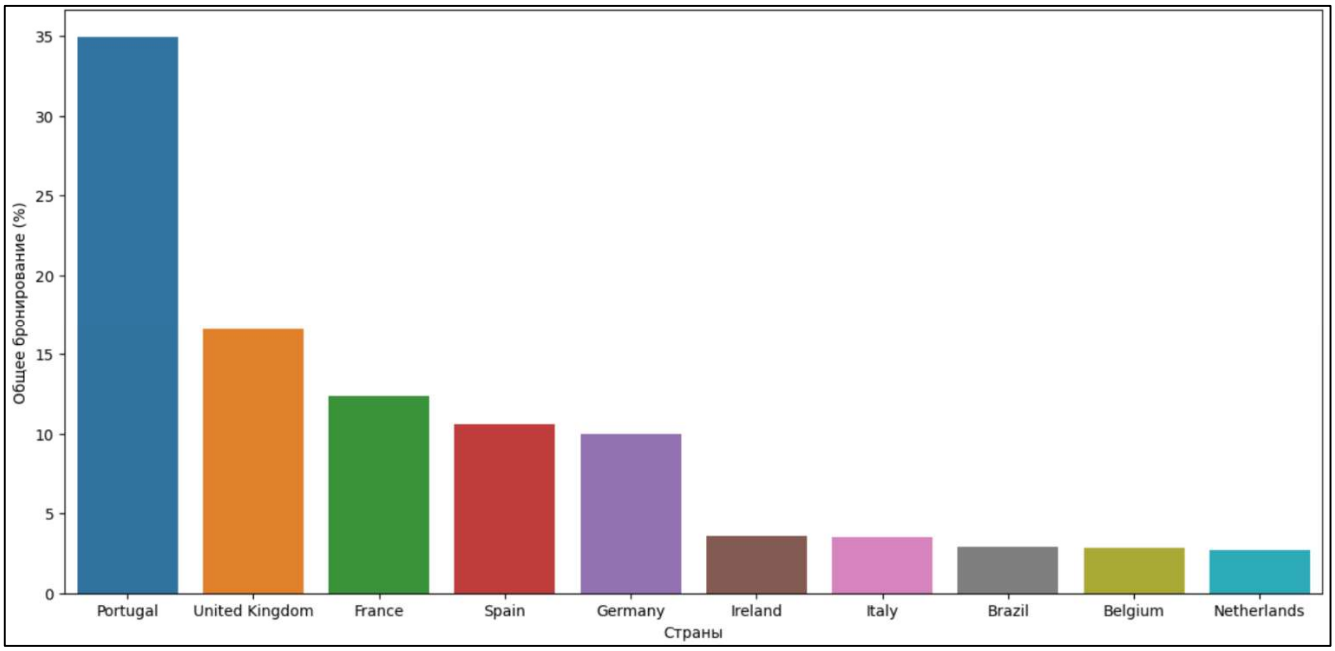


Рисунок 7. Топ стран, из которых поступило больше всего бронирований.

(включая бронирования, которые были отменены)

Здесь показаны первые 10 стран, откуда совершается наибольшее количество бронирований:

- PRT — Португалия
- GBR — Великобритания
- FRA — Франция
- ESP — Испания
- DEU — Германия
- ITA — Италия
- IRL — Ирландия
- BEL — Бельгия
- BRA — Бразилия
- NLD — Нидерланды

Теперь, если снова просмотреть данные о гостях по странам без учета отмененных бронирований, результат немного изменится. Бразилии на этом графике нет, а США заняли 10-е место по количеству гостей, фактически не отменяющих бронирование. (Рисунок 8)

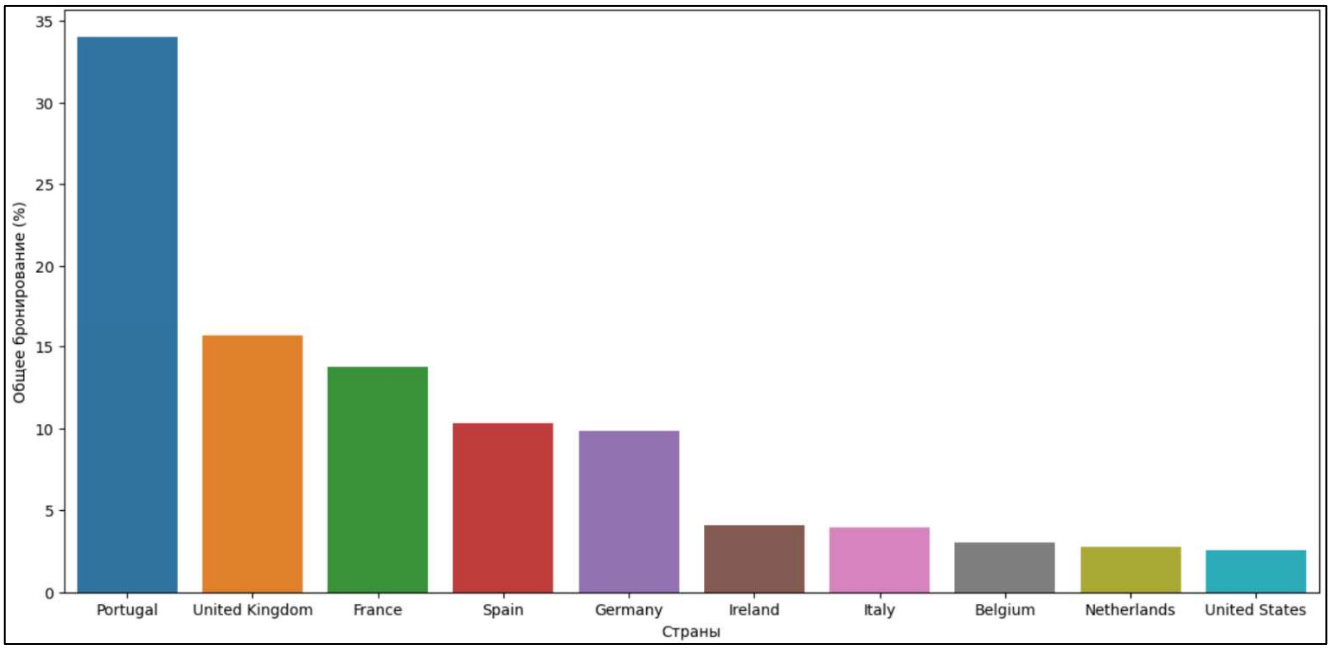


Рисунок 8. Топ стран, из которых поступило больше всего бронирований.

(исключая бронирования, которые были отменены)

Теперь, чтобы выяснить, сколько гости платят за номер в сутки. Была сделана приблизительная оценка, так как в отелях разные типы номеров и разное питание. Сезонные факторы также имеют значение. Так что цены сильно разнятся. Кроме того, поскольку информация о валюте не предоставляется, но первые 10 стран, из которых прибыли гости, являются европейскими странами, будет сделано предположение, что все цены указаны в евро. Кроме того, только взрослые и дети будут считаться платными гостями, а не младенцы. Для каждой строки в наборе данных применялись эти формулы расчёта средней цены в сутки на человека:

$$\text{Среднесуточный тариф на человека в курортном отеле} = \frac{\text{среднесуточный тариф для курортного отеля}}{(\text{кол. взрослых гостей} + \text{количество детей})}$$

$$\text{Среднесуточный тариф на человека в городском отеле} = \frac{\text{среднесуточный тариф для городского отеля}}{(\text{кол. взрослых гостей} + \text{количество детей})}$$

Дальше, среднее значение $\bar{x}$ для всех строк данных было рассчитано для каждого из этих двух значений,чтобы получить среднюю цену в сутки на человека для каждого типа отеля.

После получения средней цены в сутки на человека для всех неотмененных бронирований, по всем типам номеров и различному питанию, средняя цена в сутки на человека как в курортном, так и в городском отеле:

- курортный отель — 47,49 € в сутки на человека
- городской отель — 59,23 € в сутки на человека

Теперь можно анализировать изменение цены в день в течение года. Для простоты указана средняя цена за сутки с человека, независимо от типа номера и питания.

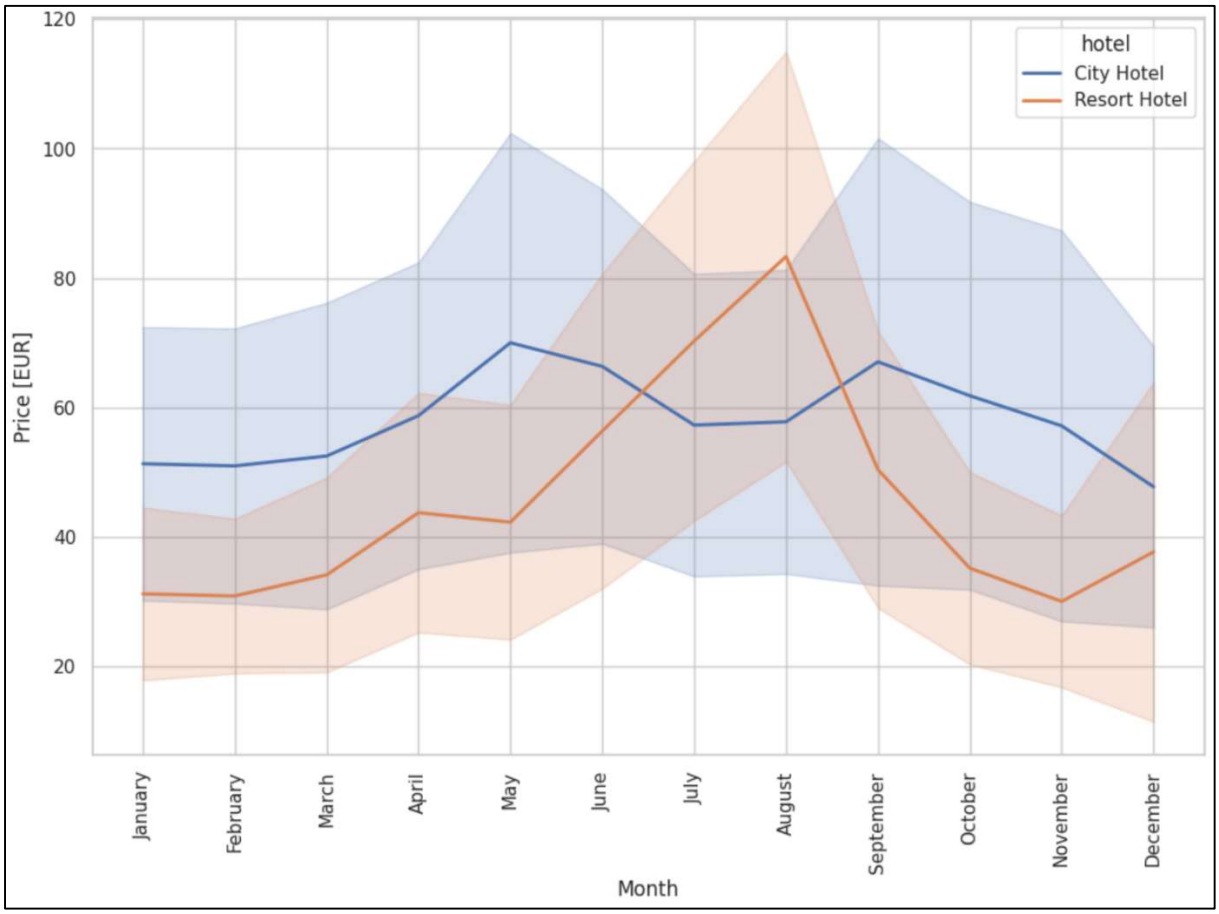


Рисунок 9. Стоимость номера в сутки на человека в течение года.

Анализ ясно показывает (Рисунок 9), что цены в курортном отеле намного выше в августе. В это время летний сезон. Цена городского отеля варьируется меньше и является самой дорогой в мае и сентябре, когда есть весенний и осенний сезоны соответственно. Кроме того, если посмотреть, какие месяцы являются самыми загруженными (Рисунок 10), городские отели отмечают увеличение посетителей весной и осенью, когда цены также самые высокие. Меньше людей приезжает в июле и августе, когда цены еще ниже. С июня по сентябрь, когда цены самые высокие, в курортных отелях останавливается меньше гостей. Зимний сезон привлекает меньше всего посетителей в оба типа отелей.

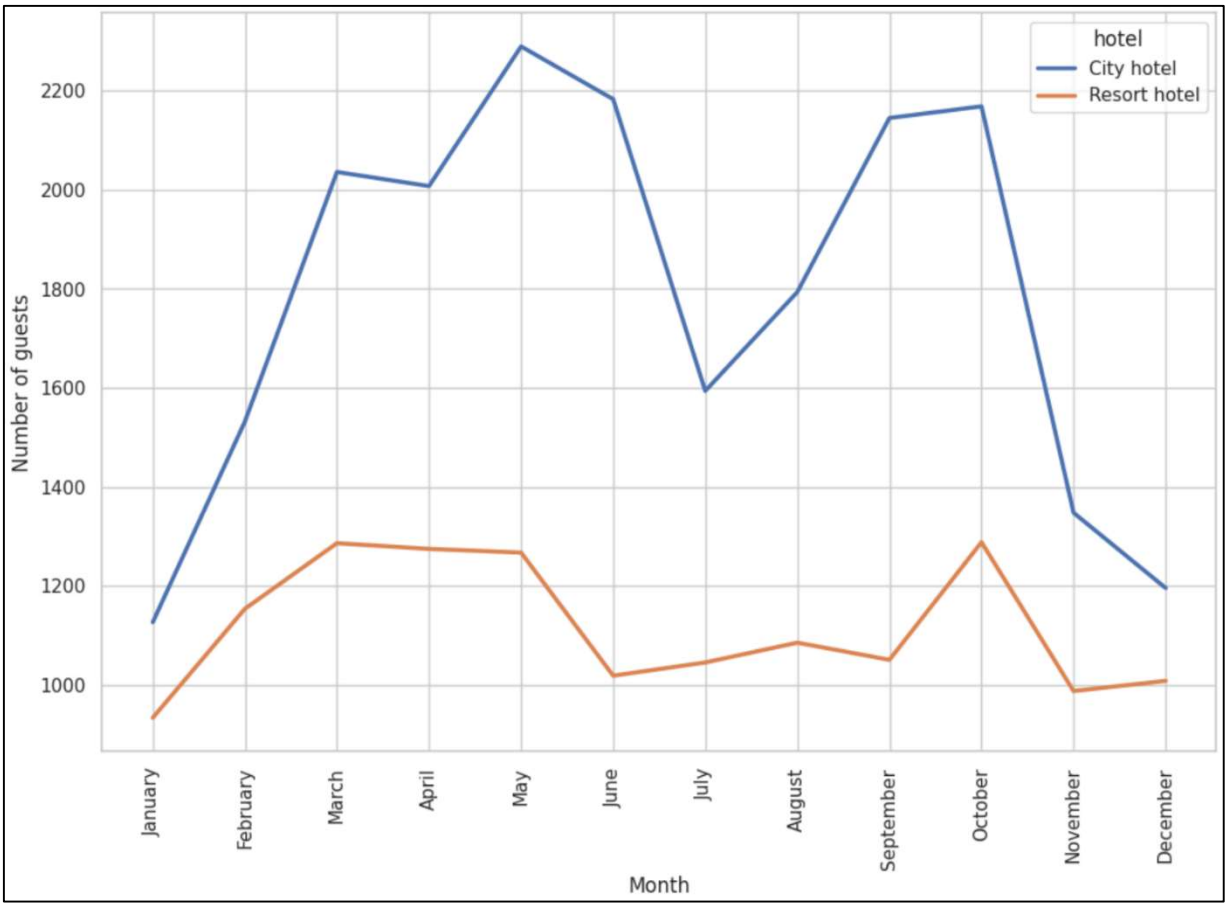


Рисунок 10. Распределение загруженности отелей в течение года.

Выводы разведочного анализа данных

- Повторные гости с меньшей вероятностью отменят или даже не отменят свои бронирования. Конечно, есть некоторые исключения. Кроме того, большинство клиентов не являются повторными гостями.
- Когда время выполнения превышает 60, гости часто отменяют свои заказы (после этой точки времени выполнения 60 дней, процент отмен увеличивается). Это означает, что среднее время выполнения значительно выше для отмененных бронирований.

- 50% бронирований приходится на 100-дневный время выполнения. Это означает, что люди планируют и рассчитывают график отпуска или работы/командировки до 100 дней вперед.
- Из всех неотмененных бронирований, по всем типам номеров и питанию средние цены составляют: в курортном отеле: 47,56 € в сутки на человека, а в городском отеле: 59,29 € в сутки на человека.
- Большинство людей не предпочитают оставаться в отеле более 1 недели. Но кажется нормальным пребывание в курортных отелях до 12-13 дней. Хотя это меняется в зависимости от сегмента рынка, пребывание дольше 15 дней, безусловно, создает выбросы для каждого сегмента.
- Клиенты из авиационного сегмента, по-видимому, не останавливаются в курортных отелях и имеют относительно меньшую среднюю продолжительность пребывания.
- Клиенты в авиационном сегменте, скорее всего, прибудут в ближайшее время из-за работы, другими словами, у них более короткое время выполнения. Кроме того, вероятно, большинство аэропортов находятся немного в стороне от моря и, скорее всего, ближе к городским отелям.
- Городские отели принимают больше гостей в течение года. Курортные отели только летом кажутся немного ближе к городским отелям при пропорциональном сравнении.
- Цены в курортном отеле намного выше в августе, который является высоким сезоном в это время. Цена городского отеля варьируется в меньшей степени и является самой дорогой в мае и сентябре, когда наступает весенний и осенний сезоны соответственно.
- Весной и осенью в городских отелях наблюдается увеличение посетителей, когда цены также самые высокие. Меньше людей приезжает в июле и августе, когда цены еще ниже.

- С июня по сентябрь, когда цены самые высокие, в курортных отелях останавливается меньше гостей. Зимний сезон привлекает меньше всего посетителей в оба типа отелей.
- Уровень отмены для групп превышает 50%. Тем временем, уровень отмены для офлайн-ТА/ТО (турагенты/туроператоры) и онлайн-ТА (турагенты) превышает 33%. Напротив, прямой сегмент имеет более низкий уровень отмены.
- Средние значения продолжительности пребывания как в выходные, так и в будние дни примерно равны.
- Клиенты, которые относятся к типу «скоротечных клиентов (Transient)», с большей вероятностью отменят бронирование. Однако чем больше количество специальных запросов, тем меньше вероятность отмены бронирования.
- Бронирования, поступающие от онлайн-турагентов, с большей вероятностью будут отменены, чем бронирования из любого другого сегмента рынка.

Глава 3. Формирование входных признаков модели прогнозирования

3.1 Реализация процедур очистки данных

Работа с отсутствующими значениями и преобразование типа данных:

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.

```
# Удалим строки, в которых нет взрослых, детей и младенцев
df = df.drop(df[(df.adults+df.babies+df.children)==0].index)


# Если идентификатор агента или компании не равен нулю, просто заменим его на 0
df[['agent','company']] = df[['agent','company']].fillna(0.0)


# Для отсутствующих значений в столбце "country" заменим его "OTHER"
df['country'].fillna("OTHER", inplace=True)


# Для отсутствующего значения детей, заменим его округленным средним значением
df['children'].fillna(round(data.children.mean()), inplace=True)
```

Для работы с отсутствующими значениями были приняты следующие процедуры:

- Удалим строки, в которых нет взрослых, детей и младенцев;
- Если идентификатор агента/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.

```
# Преобразуем тип данных этих столбцов из float в integer
df[['children', 'company', 'agent']] = df[['children', 'company', 'agent']].astype('int64')
```

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

- Преобразуем тип данных столбцов (children, company, agent) из float в integer.

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.

От признака previous_cancellations, который указывает количество предыдущих бронирований, которые были отменены клиентом до текущего бронирования, и также от признака previous_bookings_not_canceled, который указывает количество предыдущих бронирований, не отмененных клиентом до текущего бронирования, и для дальнейшего анализа необходимо выяснить, является ли существенной информация о предыдущих отменах бронирования.

1. Определим, какова доля отмен бронирования для клиентов, которые уже отменяли бронирования ранее.
2. Сравним ее с таким же показателем, но для клиентов, которые не отменяли бронирование но хотя бы раз бронировали отель ранее.

```
df_canceled = df[df['previous_cancellations'] > 0]
df_canceled.groupby('is_canceled').size() / len(df_canceled)
```

```
is_canceled
0    0.07489
1    0.92511
dtype: float64
```

```
df_not_canceled_before = df[(df['previous_cancellations'] == 0) & (df['previous_bookings_not_canceled'] > 0)]
df_not_canceled_before.groupby('is_canceled').size() / len(df_not_canceled_before)
```

```
is_canceled
0    0.98
1    0.02
dtype: float64
```

Видим, что более чем в 90% случаев, клиенты, которые отменяли бронирование ранее, отменяют его снова. В то время как для тех клиентов, которые ни разу не отменяли бронирование эта вероятность чуть более 2%. Очевидно, что это важный показатель, который нельзя сбрасывать о счетов.

Поэтому, на основе количества предыдущих бронирований, которые были отменены клиентом до текущего бронирования, и количества предыдущих бронирований, не отмененных клиентом до текущего бронирования, были созданы два новых признака:

- 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;
- новый признак cancellation_ratio на основании характеристик previous_cancellations и previous_bookings_not_canceled который показывает процент отмен.

Анализируя тепловую карту (Рисунок 11), 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.

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 11), чтобы увидеть корреляцию с колонками.

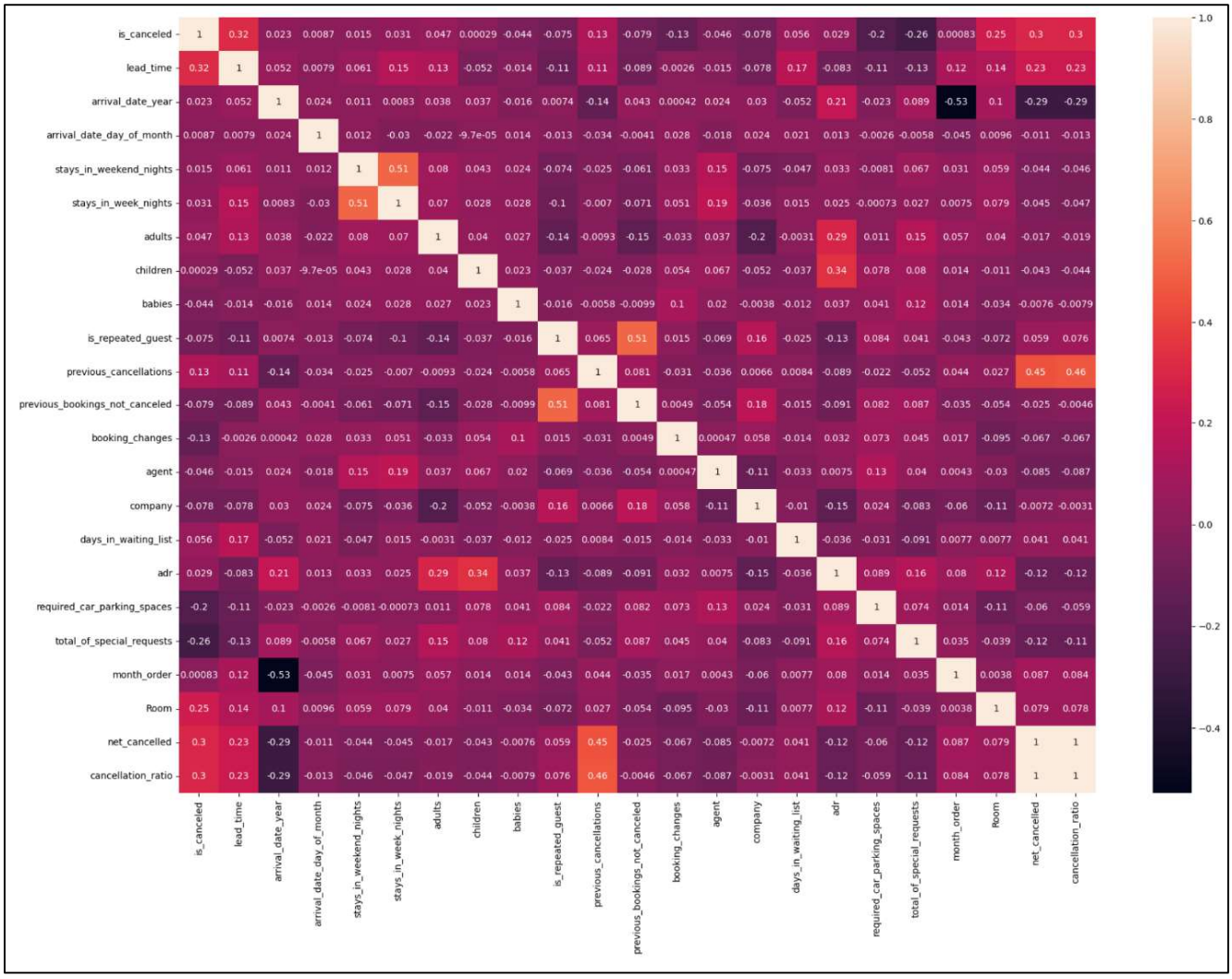


Рисунок 11. Тепловая карта, показывающая корреляцию между признаками.

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

3.3 Формирование входного набора данных

Первичными входными признаками для модели прогнозирования будут являться – признаки из исходного источника данных в дополнение к новым созданным признакам:

- 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:
    - Undefined/SC – без пакета питания;
    - BB – Завтрак в постель (Bed & Breakfast);
    - HB – Полупансион (Half board) (завтрак и еще один прием пищи /обычно ужин);
    - FB – Полный пансион (Full board) (завтрак, обед и ужин)
- market_segment: Обозначение сегмента рынка. В категориях термин:
    - Прямой (Direct)
    - Корпоративный (Corporate)
    - Онлайн турагенты (Online TA)
    - Офлайн турагенты/туроператоры (Offline TA/TO)
    - Дополнительно (Complementary)
    - Группы (Groups)
    - Не определено (Undefined)
    - Авиация (Aviation)
- distribution_channel: Канал распространения бронирования:
    - Прямой (Direct)
    - Корпоративный (Corporate)
    - Турагенты/туроператоры (TA/TO)
    - Не определено (Undefined)
    - Глобальная система распределения (GDS)

- 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 (Contract) – если с бронированием связано выделение или другой тип контракта;
    - Группа (Group) – если бронирование связано с группой;
    - Скоротечный (Transient) – если бронирование не является частью группы или контракта и не связано с другим скоротечным бронированием;
    - Скоротечный-партнер (Transient-party) – если бронирование является скоротечным, но связано как минимум с другим скоротечным 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 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 12.

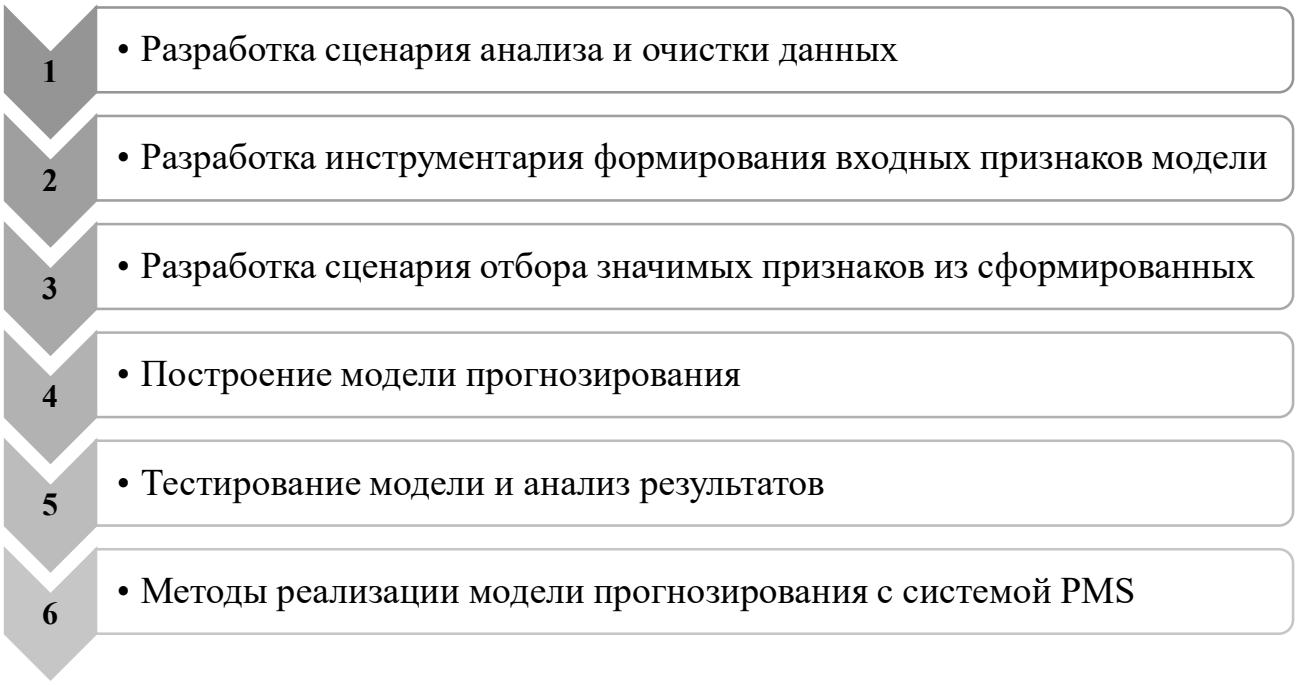


Рисунок 12. Блок-схема по основным этапам разработки модуля прогнозирования.

Глава 4. Разработка модуля прогнозирования отмены бронирования

4.1 Выбор модели прогнозирования и показатели качества прогноза

Задача прогнозирования – предсказание будущих событий. Применительно к текущим сформулированным целям прогнозирование заключается в прогнозировании вероятности отмены бронирования, другими словами, прогнозировании риска потери сделки с потенциальным гостем отеля.

Исходя из вышеописанных целей, задачу прогнозирования можно свести к бинарной классификации со следующими описаниями значений целевой переменной:

- 1 – существует значительная вероятность отмены бронирования;
- 0 – вероятность отмены бронирования незначительна или отсутствует.

Говоря о классификации, в статистике классификация — это процесс определения того, к какому классу принадлежат данные входные данные, другими словами, прогнозирование качественной переменной результата. Качественная переменная результата может быть бинарной или мультиклассовой [36].

Типичная проблема бинарной классификации, например, заключается в определении того, сможет ли физическое лицо погасить свой кредит, то есть дефолт или нет. Примером проблемы мультиклассовой классификации является предсказание исхода футбольного матча, где результатом может быть победа, поражение или ничья. В этом исследовании мы работаем с проблемой бинарной классификации, потому что бронирование отеля либо отменено, либо не отменено.

Показатели качества прогноза используются для определения качества классификации, например, если модель предсказывает правильный класс наблюдения, предсказанный класс будет равен истинному классу. Если

предсказанный класс не соответствует истинному классу, была сделана ложная классификация. Это можно выразить следующим образом:

$$I(y_i = \hat{y}_i) \text{ или } I(y_i \neq \hat{y}_i)$$

где y_i — правильный класс для i-го наблюдения, а $\hat{y}_i$ — прогнозируемый класс для i-го наблюдения. Более того, $I(\text{A})$ — индикаторная функция, которая будет равна единице, если произойдет событие A, и нулю в противном случае. Доля ошибочных классификаций называется коэффициентом ошибок и рассчитывается по формуле:

$$\frac{1}{N}\sum_{n=1}^{N} I(y_i \neq \hat{y}_i)$$

где N представляет размер выборки. Модель прогнозирования, которая пытается решить проблему бинарной классификации, может сделать два типа ошибочных классификаций: ложноположительные (False Positive, FP) и ложноотрицательные (False Negative, FN) [37]. Следующий пример будет использоваться для объяснения значения ложноположительных и ложноотрицательных результатов. Предположим, есть модель, которая пытается предсказать, будет ли у человека сердечное заболевание. Если модель предсказывает, что человек будет страдать сердечным заболеванием, но на самом деле этого не произойдет, модель предсказывает ложное срабатывание. Напротив, если модель предсказывала, что человек здоров, то есть не страдает сердечным заболеванием, но на самом деле будет страдать сердечным заболеванием, модель возвращала ложноотрицательный результат. Обе ситуации являются примерами неправильной классификации, но подчеркивают разницу между ошибками. Ложноположительные и ложноотрицательные результаты почти всегда встречаются в модели прогнозирования, но их важность будет различаться в зависимости от контекста. В приведенном выше примере с сердцем более серьезно

классифицировать больного человека как здорового, чем здорового человека как больного, т. е. сделать ложноотрицательный результат хуже. Ложноположительный результат можно также назвать ошибкой 1-го типа, а ложноотрицательный — ошибкой 2-го типа.

Оценка того, насколько хорошо модель работает, анализируя показатель точности классификации, ложноположительные и ложноотрицательные результаты, интересна только в том случае, если модель тестируется на данных, которых она раньше не видела, однако в этом проекте это невозможно из-за отсутствия дополнительных данных, которые можно было бы использоваться для проведения этого типа тестирования.

Однако коэффициент ложных срабатываний (FPR) рассчитывается путем деления количества ложных срабатываний (FP) на количество ложных срабатываний (FP) плюс количество истинных отрицательных (TN) результатов. Ссылаясь на приведенный выше пример, это означало бы деление лиц, ошибочно классифицированных как больные, на количество здоровых людей. Коэффициент ложноотрицательных (FNR) результатов рассчитывается путем деления количества ложноотрицательных (FN) результатов на количество ложноотрицательных (FN) результатов плюс количество истинно положительных (TP) результатов.

Ссылаясь на приведенный выше пример, это означало бы разделение людей, которые ошибочно классифицируются как здоровые, на всех людей, страдающих сердечными заболеваниями.

Пусть FP — количество наблюдений, ошибочно классифицированных как положительные, FN — количество наблюдений, ложно классифицированных как отрицательные, TN — количество наблюдений, правильно классифицированных как отрицательные, а TP — количество наблюдений, правильно классифицированных как положительные. Затем частота ложноположительных

результатов и частота ложноотрицательных результатов определяются следующим образом:

$$FPR = \frac{FP}{(FP + TN)} \qquad FNR = \frac{FN}{(FN + TP)}$$

Чтобы вычислить общую точность (accuracy) модели бинарной классификации, нужно просто разделить правильное количество классификаций на все классификации, то есть:

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

Распространенный способ суммировать производительность модели, пытающейся решить проблему бинарной классификации, — использовать матрицу ошибок, показанную на рисунке 13. На этом рисунке по диагонали от левого верхнего угла к правому нижнему отображаются правильные значения, что означает, что прогностические значения соответствуют истинным значениям. Диагональ из нижнего левого угла в верхний правый показывает ложноположительные и отрицательные результаты [37].

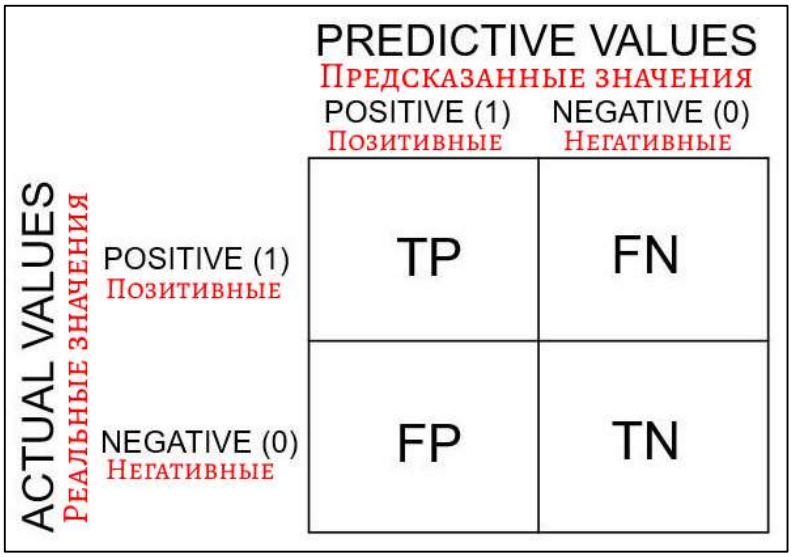


Рисунок 13. Матрица ошибок.

Рабочая характеристика приемника, также известная как ROC-кривая, является еще одним показателем качества прогноза и широко используемым инструментом для визуализации возможного компромисса между ошибками типа 1 и типа 2, то есть ложноположительными (FP) и ложноотрицательными (FN) для модели классификации.

Эта метрика показывает изменение типов ошибок по отношению к порогу T классификатора [36]. Модель, используемая для задачи бинарной классификации, будет выводить оценку или вероятность того, что данное наблюдение принадлежит к классу 1, который мы называем P, и наблюдение классифицируется как 1 в случае $P > T$, в противном случае 0.

Таким образом, пороговое значение T является минимальным значением P, которое необходимо отнести к классу 1. Таким образом, наиболее интуитивно понятный порог для моделей, которые выводят вероятность, например логистической регрессии, составляет 0,5, так что каждое наблюдение относится к наиболее вероятному классу.

Однако, например, снижение порога до 0,4 приведет к тому, что больше наблюдений будет отнесено к классу 1, что затем увеличит общее количество правильных прогнозов для класса 1, но, в свою очередь, уменьшит количество правильных прогнозов для класса 0. Изменение поэтому порог классификатора соответствует изменению критериев, когда модель относит наблюдение к любому классу.

Пример ROC-кривой показан на рисунке 14, взятом из Википедии и созданном (cmglee, MartinThoma) в 2018 году, где синие, зеленые и оранжевые линии представляют собой гипотетические ROC-кривые.

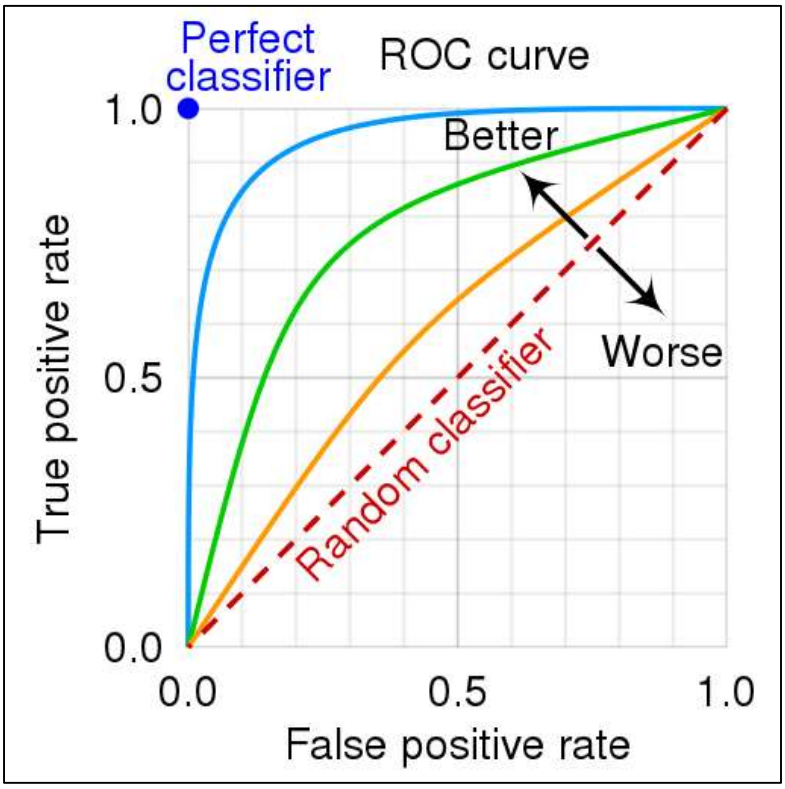


Рисунок 14. ROC-кривая.

Вдоль кривых на рисунке 14 находятся все возможные значения порогов между 0 и 1, и поскольку эти возможные значения непрерывны, то и кривая непрерывна. Порог уменьшается по мере того, как кривая идет от нижнего левого угла к верхнему правому углу. В левом нижнем углу у нас есть порог 1, что означает, что мы присваиваем всем наблюдениям класс 0. Следовательно, у нас есть 0% истинных положительных результатов и 0% ложных положительных результатов, поскольку никакие наблюдения не прогнозируются как «положительные».

Для противоположного варианта в правом верхнем углу порог равен 0, что означает, что все наблюдения классифицируются как 1 или «положительные». Это приводит к тому, что у нас будет 100 % истинных положительных результатов, но также у нас будет 100 % ложноположительных результатов. Таким образом, перемещение между этими двумя крайностями будет способом балансирования

между увеличением истинно положительного уровня и снижением уровня ложноположительного результата. Этот компромисс визуализируется с помощью ROC-кривой [36].

Для данного уровня порога ROC-кривая иллюстрирует соответствующую скорость классификации для каждого класса. При прогнозировании цель состоит в том, чтобы найти оптимальный порог, который приводит к желаемому соотношению между ошибками типа 1 и типа 2. Однако, поскольку этот оптимальный порог может различаться в зависимости от модели и данных, общим показателем для сравнения моделей является площадь под кривой или AUC [36]. Поскольку оптимальная ROC-кривая касается оси у и верхней линии на рисунке 14, показанной синей точкой, поэтому AUC, равная 1, является оптимальной. Для сравнения, мы ожидаем, что классификатор, не обладающий предсказательной силой, например случайные назначения, будет иметь AUC 0,5, и на рисунке 14 это соответствует красной пунктирной диагональной линии.

Возвращаясь теперь к бинарной классификации, она используется для решения вопроса о принадлежности некоторого объекта к одному из двух классов. Как показывает практика, множество задач классификации в Data Mining может быть сведено к бинарным. При их использовании удается упростить модель и снять некоторые ограничения, связанные с большим числом возможных состояний выходного значения [30].

Наиболее часто в качестве используемых на практике моделей бинарной классификации или прогнозирования можно встретить следующие:

- Логистическая регрессия (Logistic Regression, LG)
- Случайный лес (Random Forest, RF)
- Метод опорных векторов (Support Vector Machine, SVM)
- Классификатор голосования (Voting Classifier, VC)

- Линейная регрессия (Linear Regression, LR)
- Древо решений (Decision Tree, DT)

После более чем много итераций экспериментов с признаками и разными типами моделей и их параметрами было решено использовать в итогах следующие:

- Случайный лес (Random Forest, RF)
- Логистическая регрессия (Logistic Regression, LG)
- Метод опорных векторов (Support Vector Machine, SVM)
- Классификатор голосования (Voting Classifier, VC)

Экспериментируя с моделями, изначально использовалась линейная регрессия, но вместо нее было решено использовать логистическую регрессию. Проблема в том, что нужно прогнозировать вероятность, т.е. число в интервале (0,1). Логистическая регрессия это обеспечивает, а вот линейная может прогнозировать любое число. Более того, дерево решений также было заменено случайным лесом. Одиночные деревья обычно не используют, т.к. они склонны к переобучению. Случайный лес заведомо будет лучше, чем одиночное дерево.

4.1.1 Случайный лес (Random Forest, RF)

Метод случайного леса (Random Forest, RF) заключающийся в использовании ансамбля решающих деревьев. Алгоритм применяется для задач классификации, регрессии и кластеризации. Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим. Случайный лес представляет из себя ансамбль независимых деревьев решений. Каждое дерево обучается независимо от других на случайной выборке данных. Это помогает сделать модель более надежной, чем одна модель

дерева решений, и меньшая вероятность получить переобучение [31]. Обычно, в RF есть два параметра – количество деревьев-оценщиков и количество элементов в случайной листе. В целом, это простой в использовании алгоритм машинного обучения, который в большинстве случаев дает отличный результат даже без настройки гиперпараметров.

Плюсы:

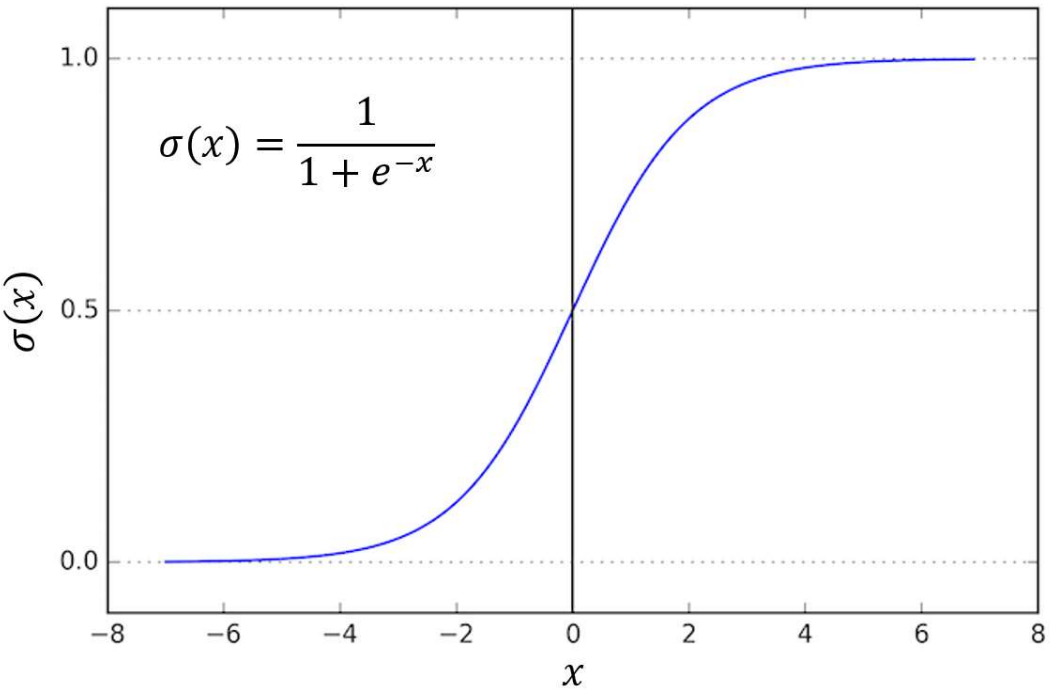
- Случайный лес можно использовать как для задач классификации, так и для регрессии;
- Случайный лес хорошо работает как с категориальными, так и с числовыми данными. Обычно не требуется масштабирование или преобразование переменных;
- Способность эффективно обрабатывать данные с большим числом признаков и классов;
- Случайный лес в значительной степени не подвержен влиянию выбросов. Он делает это, объединяя переменные;
- Случайный лес очень гибки и обладают очень высокой точностью.

Минусы:

- Случайный лес может потребовать больших вычислительных ресурсов для больших наборов данных;
- Случайный лес нелегко интерпретировать. Он обеспечивает важность признаков, но не обеспечивает полной видимости коэффициентов в виде линейной регрессии;
- Случайный лес подобен алгоритму черного ящика, что означает, что он позволяет очень мало контролировать то, что делает модель.

4.1.2 Логистическая регрессия (Logistic Regression, LG)

Классификация является одной из важнейших областей машинного обучения, а логистическая регрессия (Logistic Regression, LG) — одним из его основных методов. Это относится к технике машинного обучения с учителем. Он используется для прогнозирования категориальной зависимой переменной с использованием заданного набора независимых переменных. Логистическая регрессия предсказывает выход категориальной зависимой переменной. Она модель линейной классификации, которая используется для моделирования двоичной целевой переменной. Она используется для прогнозирования вероятности того, что событие произойдет. Для определения класса на основания полученного значения задается некий порог P так, что если $p \geq P$, то на выходе будет 1, а иначе 0. Используемая регрессией сигмоидальная функция ограничена двумя горизонтальными асимптотами, из-за чего значение вероятности находится в промежутке [0 , 1] [33]. Часто под сигмоидой понимают логистическую функцию:



Плюсы:

- Логистическую регрессию легче реализовать, интерпретировать и очень эффективно обучать;
- Она не делает никаких предположений о распределении классов в пространстве признаков;
- Её можно легко распространить на несколько классов (полиномиальная регрессия) и естественное вероятностное представление прогнозов классов;
- Она не только обеспечивает меру того, насколько уместным является предиктор (размер коэффициента), но также и направление его связи (положительное или отрицательное);
- Она очень быстро классифицирует неизвестные записи и выдает хорошо откалиброванные вероятности вместе с результатами классификации;
- Хорошая точность для многих простых наборов данных, и она хорошо работает, когда набор данных линейно разделим, а также на низкоразмерных данных;
- Она может интерпретировать коэффициенты модели как индикаторы важности признаков;
- Логистическая регрессия менее склонна к переобучению, но может переобучать наборы данных большой размерности. Можно рассмотреть методы регуляризации (L1 и L2), чтобы избежать переподгонки в этих сценариях.

Минусы:

- Нелинейные задачи не могут быть решены с помощью логистической регрессии, поскольку она имеет линейную поверхность решения;
- Она предполагает линейность между зависимыми и независимыми переменными;

- С помощью логистической регрессии сложно получить сложные взаимосвязи, а также требуется средняя мультиколлинеарность или отсутствие мультиколлинеарности между независимыми переменными;
- Она переобучается на многомерных данных;
- Следует использовать только важные и релевантные признаки, иначе прогностическая ценность модели ухудшится.

4.1.3 Метод опорных векторов (Support Vector Machine, SVM)

Метод опорных векторов (Support Vector Machine, SVM) — один из наиболее популярных методов обучения, который применяется для решения задач классификации и регрессии. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора [34]. SVM связан с алгоритмами обучения, которые анализируют данные для классификации и регрессионного анализа. Это набор контролируемых методов обучения, используемых для классификации, регрессии и обнаружения выбросов. Преимущества метода опорных векторов: эффективен в пространствах большой размерности. По-прежнему эффективен в случаях, когда количество измерений больше, чем количество выборок.

Плюсы:

- SVM работает относительно хорошо, когда между классами существует четкая граница разделения;
- SVM более эффективен в многомерных пространствах;

- SVM эффективен в случаях, когда количество измерений больше, чем количество выборок;
- SVM относительно эффективен с точки зрения вычислительной памяти.

Минусы:

- Алгоритм SVM не подходит для больших наборов данных;
- SVM не очень хорошо работает, когда в наборе данных больше шума, т. е. целевые классы перекрываются;
- В тех случаях, когда количество признаков для каждой точки данных превышает количество выборок обучающих данных, SVM будет работать хуже;
- Поскольку классификатор опорных векторов работает, помещая точки данных выше и ниже классифицирующей гиперплоскости, нет вероятностного объяснения классификации.

4.1.4 Классификатор голосования (Voting Classifier, VC)

Классификатор голосования (Voting Classifier, VC) — это модель машинного обучения, которая обучается на ансамбле многочисленных моделей и прогнозирует результат (класс) на основе их наивысшей вероятности выбора класса в качестве результата. Если использовать VC, то рекомендуется выбирать между логистической регрессией, случайным лесом и добавить еще один алгоритм — метод опорных векторов (SVM). Классификатор голосования часто используется кагглерами (Kagglers), чтобы повысить производительность своей модели и подняться по лестнице рангов [32]. Классификатор голосования также можно использовать для реальных наборов данных для повышения производительности, но он имеет некоторые ограничения. По сути, это просто принимает наиболее

распространенный вывод за вывод окончательной оценки, которая может работать лучше, если доступно несколько «кажущихся хорошими» моделей, но иногда они дают неверные прогнозы, поэтому использование классификатора голосования помогает откалибровать результат.

4.2 Построение модели прогнозирования и оценка качества прогноза

Для реализации модели прогнозирования использовались следующие средства разработки:

- язык программирования Python 3.9;
- среда разработки Google Colab;
- библиотека scikit-learn.

Полученный набор данных был разделен на «набор признаков» и «целевая переменная». Значения «целевой переменной», указывающее, было ли бронирование отменено – 1, или нет – 0.

Для задачи предсказания отмены бронирования номеров в отеле обычно используется метод train_test_split для разделения данных на обучающую и тестовую выборки, поскольку он предоставляет следующие преимущества:

- Быстрота и простота: Использование train_test_split является простым и быстрым методом разделения данных, который обеспечивает хороший компромисс между эффективностью и точностью модели. Это позволяет сосредоточиться на самом процессе обучения и подборе оптимальных гиперпараметров, а также упрощает поддержку и внедрение модели в продуктивную среду.

- Предотвращение переобучения: train_test_split перемешивает данные перед разделением, что предотвращает возможное переобучение модели. Это гарантирует, что обучающая выборка содержит достаточно разнообразие, чтобы модель могла уловить основные закономерности, а тестовая выборка позволяет оценить обобщающую способность модели на новых данных.

Таким образом, использование train_test_split для разделения данных на обучающую и тестовую выборки обеспечивает эффективный и надежный способ для обучения и оценки модели предсказания отмены бронирования номеров в отеле. Этот метод обеспечивает хороший баланс между простотой, скоростью и точностью.

Однако метод train_test_split является независимой временной структурой, т. е. использование train_test_split обеспечивает случайное разделение данных, что позволяет получить обучающую и тестовую выборки, представляющие разнообразные примеры безотносительно времени. В результате с помощью этого метода модели будут обучаться на более поздних данных для предсказания более ранних. Это приводит к обманчивым результатам, которые показывают, что модели предсказывают с высокой точностью, но это может быть не так, когда есть новые данные. В тех случаях, когда мы имеем дело с данными, упорядоченными по времени, этого допускать нельзя.

Потому что, заметно, что на практике будут известны только данные за прошедшие периоды и прогнозы мы должны строить опираясь на них. В этом случае в качестве обучающей выборки будем брать только данные из первой части набора данных, а в качестве тестовой – вторую часть, чтобы моделировать эту ситуацию.

Поэтому, при разделении данных на обучающую и тестовую выборку нужно сделать так, чтобы в обучающей выборке были более ранние, а в тестовой - более

поздние данные. Например, упорядочив набор по времени и проведя разделение не используя train_test_split. И это именно то, что было сделано здесь, чтобы избежать этой проблемы — сортировка данных по времени перед разбиением. Ниже приведен фрагмент кода, показывающий, как были реализованы разделение и кодирование данных. Кодирование, в котором использовалось одномоментное кодирование (OneHotEncoder) категориальных переменных.

```
def data_split(df, label, split_on=0):
    '''
    Разделим входной массив данных на обучающий и тестовый наборы и выполнить
    одномоментное кодирование для категориальных переменных.

    Args:
        df (pandas.DataFrame): Входной фрейм данных.
        label (str): Имя столбца целевой метки.

    Returns:
        Кортеж из четырех элементов, содержащий обучающие и тестирующие признаки и метки.
    '''
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import OneHotEncoder

    X = df.drop(label, axis=1)
    Y = df[label]

    # Разделение на обучающее и тестирующее множества
    if split_on == 0: # Используем train_test_split с перемешиванием
      x_train, x_test, y_train, y_test = train_test_split(X, Y, random_state=0)
    else: # Делим не перемешивая, а потом перемешиваем обучающую и тестовую выборки отдельно
      x_train = df_subset.iloc[:split_on].sample(frac=1)
      x_test = df_subset.iloc[split_on:].sample(frac=1)
      y_train = x_train.pop(label)
      y_test = x_test.pop(label)

    # Одномоментное кодирование категориальных переменных
    categorical_cols = X.select_dtypes(include=['object']).columns
    if len(categorical_cols) > 0:
        encoder = OneHotEncoder(handle_unknown='ignore')
        encoder.fit(x_train[categorical_cols])
        x_train_encoded = encoder.transform(x_train[categorical_cols]).toarray()
        x_test_encoded = encoder.transform(x_test[categorical_cols]).toarray()
        x_train = x_train.drop(categorical_cols, axis=1)
        x_test = x_test.drop(categorical_cols, axis=1)
        x_train = np.hstack((x_train, x_train_encoded))
        x_test = np.hstack((x_test, x_test_encoded))

    return x_train, x_test, y_train, y_test
```

Что касается выбора наиболее подходящих гиперпараметров моделей, с помощью объекта GridSearch библиотеки scikit-learn и перебором были подобраны такие гиперпараметры модели, с которым оценка качества достигает максимума.

```
from sklearn.model_selection import GridSearchCV

def train(x_train, y_train, model_type='random_forest', random_state=0, estimators=[]):
    """
    Обучаем модели классификации на входных признаках и метках с помощью заданного алгоритма.

    Args:
        x_train (pandas.DataFrame): Фрейм данных, содержащий входные признаки.
        y_train (pandas.Series): Серия, содержащая целевые метки.
        x_test (pandas.DataFrame): Фрейм данных, содержащий входные признаки для тестирования.
        y_test (pandas.Series): Серия, содержащая целевые метки для тестирования.
        model_type (str): Тип используемой модели. Может быть 'random_forest',
        'voting_classifier', 'logistic_regression' или 'SVM'. По умолчанию - 'random_forest'.
        random_state (int): Случайное зерно, используемое для воспроизводимости. По умолчанию 0.

    Returns:
        Обученный объект модели scikit-learn или CatBoost.
    """
    if model_type == 'random_forest':
        model = RandomForestClassifier(random_state=random_state)
    elif model_type == 'voting_classifier':
        model = VotingClassifier(estimators=estimators, voting='soft')
    elif model_type == 'logistic_regression':
        model = LogisticRegression(solver='newton-cholesky', max_iter=500)
    elif model_type == 'SVM':
        model = SVC(kernel='rbf', C=15000, gamma='scale', probability=True)
    else:
        raise ValueError(f"Неверный тип модели: {model_type}")

    model.fit(x_train, y_train)

    return model
```

Для оценки качества модели использовалась ROC_{AUC} классификации. Результат представлен на рисунке 15. Можно предположить, что модель справляется очень хорошо с бинарной классификацией и способна прогнозировать с высокой вероятностью того, что будет ли бронирование отменено.

Наилучший результат показал классификатор голосования, как видно на рисунке 15, удалось достигнуть $ROC_{AUC} \approx 0{,}865$ что показывает, что Voting Classifier действительно играет свою роль, так как он явно повысило

производительность модели. По экспертной шкале построенная модель является отличным классификатором, так как:

- от 0,6 до 0,7 — считается приемлемым;
- от 0,7 до 0,8 — считается хорошим;
- от 0,8 до 0,9 — считается отличным;
- а более 0,9 — выдающимся.

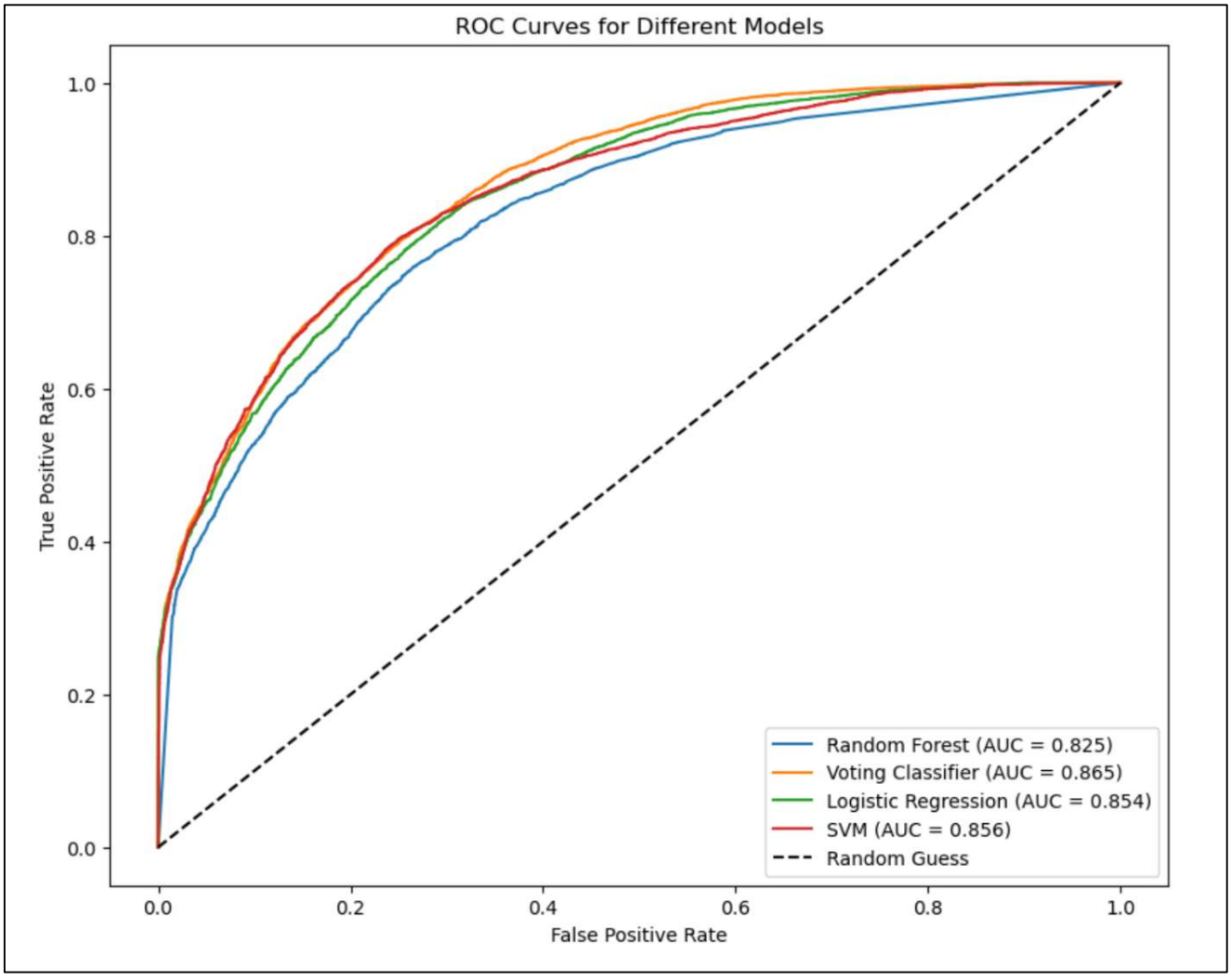


Рисунок 15. ROC-кривые для построенных моделей.

4.3 Методы реализации модели прогнозирования в системе PMS

В этом исследовании было применено несколько методов машинного обучения, в которых данные о бронировании отелей обрабатывались для получения хороших результатов прогнозирования. На рисунке 16 показано, как модели использовались в этом наборе данных, и исследовано влияние этих моделей. Предлагаемый подход описывается более подробно следующим образом:

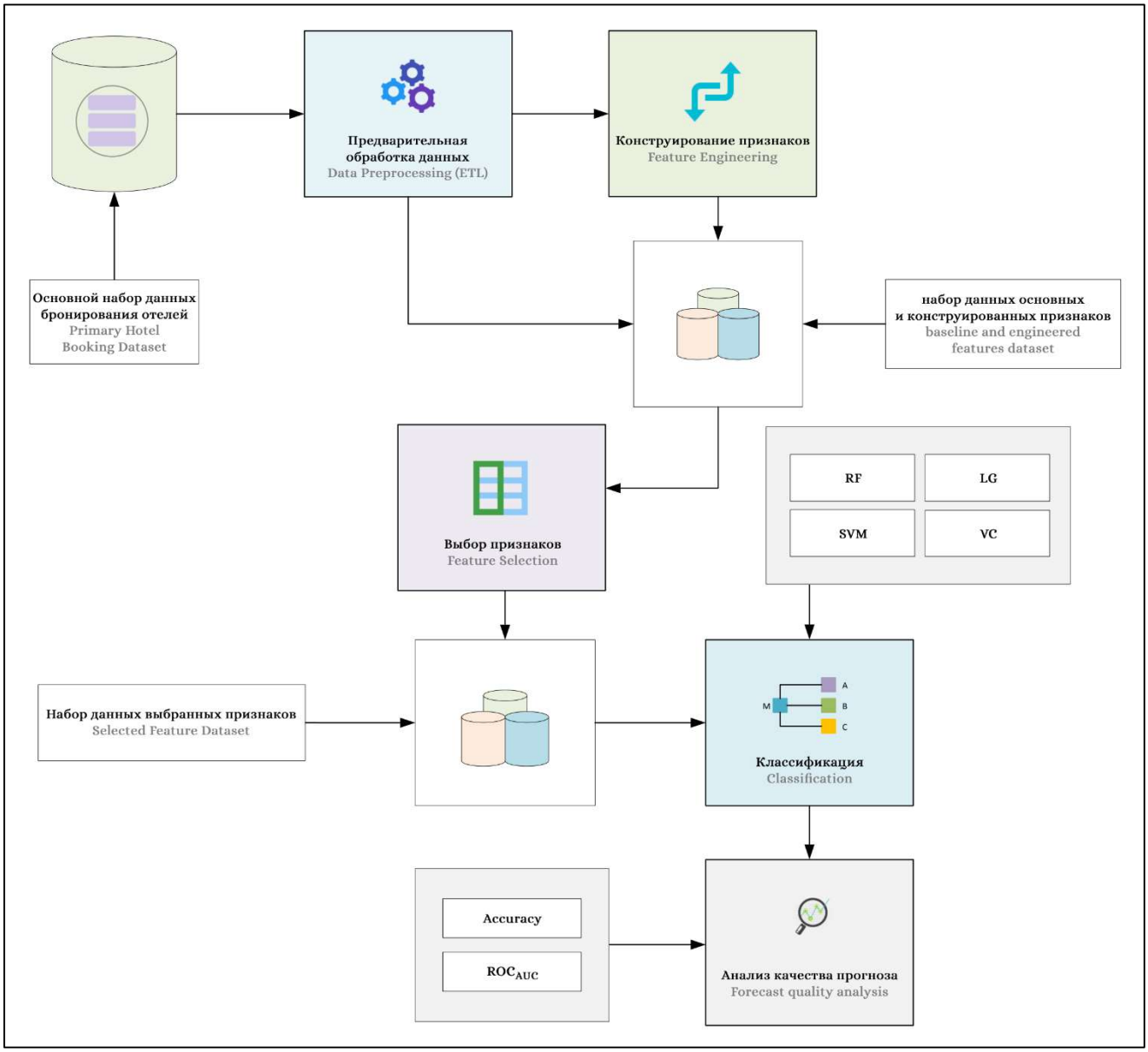
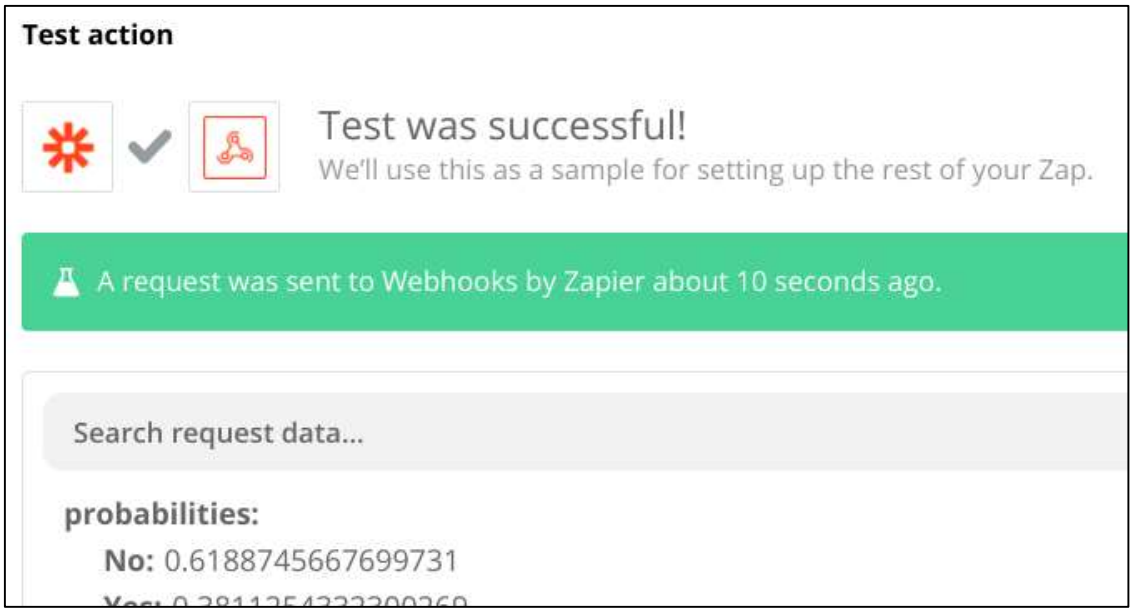


Рисунок 16. Предлагаемая методология.

Чтобы использовать преимущества этой модели машинного обучения для прогнозирования вероятности отмены бронирования, используется веб-сервис. Веб-сервис – программная система со стандартизированными интерфейсами (которая идентифицируется уникальным веб-адресом). Веб-сервисы могут взаимодействовать как друг с другом, так и со сторонними приложениями с помощью REST, SOAP и XML-RPC.

Чтобы продемонстрировать, как эта модель может работать в коммерческом использовании, предлагается использовать Obviously AI в качестве платного инструмента для реализации сервиса. Этот инструмент позволяет отелям извлекать данные из нескольких источников данных, включая простую загрузку CSV-файла и интеграцию с MySQL, PostgreSQL, SQL-сервером, Google Диском и т. д. После загрузки данных в инструмент загружаются конфигурации модели, после чего инструмент готов к использованию бизнес-подразделением для прогнозирования вероятности отмены бронирования. Бизнес может интегрироваться с Obviously AI с помощью вебхука (webhook), а затем Custom Request. Custom Request — это просто POST-запрос.

В результате на выходе будет вероятность отмены бронирования. С этими прогнозами теперь можно делать что угодно: отправлять его в канал Slack, отправлять уведомления по электронной почте ответственной команде, добавлять прогнозы в таблицу данных или делать что-то еще, что нужно бизнесу.



При этом инструмент Obviously AI является платным с ежемесячной подпиской, начиная с 999 долларов США в месяц. Поэтому, поскольку этот проект в настоящее время предназначен для исследовательских, а не коммерческих целей, была изучена только возможность использования инструмента Obviously AI без подписки на платный сервис.

Может быть другой вариант реализации модели машинного обучения с использованием Streamlit, который представляет собой фреймворк с открытым исходным кодом для машинного обучения и проектов по науке о данных, который помогает создавать веб-приложения из кода Python. Это может быть план Б для бизнеса, который не хочет тратить ежемесячные подписки на такие инструменты, как Obviously AI, а предпочитает инвестировать в разработку собственного индивидуального веб-приложения.

В дальнейшем, при появлении коммерческого интереса, планируется установка сервера на выделенной виртуальной машине с необходимыми требованиями и запуск веб-сервиса для интеграции модуля прогнозирования с PMS-системой отелей.

ЗАКЛЮЧЕНИЕ

Разработан модуль прогнозирования отмены бронирования в отелях, с помощью которого можно спрогнозировать загруженность отелей. Все построенные модели достигли значений AUC выше 80%, а классификатор голосования достиг 86,5%, что считается отличным. Инструментарий реализован на базе Python и включает в себя процедуры первичной загрузки и очистки данных (выявление и обработка пропусков, ошибок и противоречий), вычисление признаков прогнозирования, методы избавления от дисбаланса классов.

Разработаны инструменты позволяют менеджерам отелей:

- смягчить потерю дохода, связанную с отменой бронирования, это связано с тем, что непредсказуемые отмены ограничивают производство точных прогнозов, поэтому результаты этого исследования позволяют менеджерам отелей точно прогнозировать чистый спрос и строить более точные прогнозы;
- снизить риски, связанные с избыточным бронированием (затраты на перераспределение, денежные компенсации или компенсации за услуги и, что особенно важно сегодня, затраты на социальный имидж);
- добиться лучшей видимости приведет к увеличению продаж, поскольку менее жесткие правила отмены бронирования генерируют больше бронирований;
- принимать меры в отношении бронирований, которые определены как «потенциально подлежащие отмене», а также составлять более точные прогнозы спроса и загруженности.

При появлении коммерческого интереса, планируется установка и запуск веб-сервиса для интеграции модуля прогнозирования с PMS-системой разных отелей. По итогам выполненной работы статья готовится к публикации.

СПИСОК ЛИТЕРАТУРЫ

1. Kimes S.E., Wirtz J. Has revenue management become acceptable? Findings from an International study on the perceived fairness of rate fences. Journal of Service Research, 2003. Раздел 6(2). С. 125–135. URL: http://dx.doi.org/10.1177/1094670503257038 (дата обращения: 20.05.2023)
2. Chiang W.C., Chen J.C., Xu X. An overview of research on revenue management: current issues and future research. International Journal of Revenue Management, 2007. Раздел 1(1). С. 97–128. URL: http://dx.doi.org/10.1504/IJRM.2007.011196 (дата обращения: 20.05.2023)
3. Mehrotra R., Ruttley J. Revenue management (2nd edition). Вашингтон, округ Колумбия, США: American Hotel & Lodging Association (AHLA), 2006. URL: https://www.scribd.com/document/292066294/Revenue-Management-AHLA-2006 (дата обращения: 20.05.2023)
4. Talluri K.T., Van Ryzin G. The theory and practice of revenue management. Бостон, Массачусетс, США: Kluwer Academic Publishers, 2004. URL: https://download.e-bookshelf.de/download/0000/0046/37/L-G-0000004637-0002369102.pdf (дата обращения: 20.05.2023)
5. 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 & Tourism Marketing, 2015. Раздел 32(7). С. 886–906. URL: http://dx.doi.org/10.1080/10548408.2015.1063864 (дата обращения: 20.05.2023)
6. Chen C.C., Xie Lijia K. Differentiation of cancellation policies in the U.S. hotel industry. International Journal of Hospitality Management, 2013. Раздел 34. С. 66–72. URL: http://dx.doi.org/10.1016/j.ijhm.2013.02.007 (дата обращения: 20.05.2023)

7. 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. Раздел 30(1). С. 129–135. URL: https://doi.org/10.1016/j.ijhm.2010.03.010 (дата обращения: 20.05.2023)
8. Noone B.M., Lee C.H. Hotel overbooking: The effect of overcompensation on customers' reactions to denied service. Journal of Hospitality & Tourism Research, 2010. Раздел 35(3). С. 334–357. URL: https://doi.org/10.1177/1096348010382238 (дата обращения: 20.05.2023)
9. Morales D.R., Wang, J. Forecasting cancellation rates for services booking revenue management using data mining. European Journal of Operational Research, 2010. Раздел 202(2). С. 554–562. URL: http://dx.doi.org/10.1016/j.ejor.2009.06.006 (дата обращения: 20.05.2023)
10. Liu P.H. Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In I. Yeoman & U. McMahon-Beattie (Eds.), Revenue management and pricing: Case studies and applications (pp. 91–101). Cengage Learning EMEA, 2004. (дата обращения: 20.05.2023)
11. Hayes D.K., Miller A.A. Revenue management for the hospitality industry. Хобокен, Нью-Джерси, США: John Wiley & Sons, Inc, 2011. (дата обращения: 20.05.2023)
12. Kimes S.E. The future of hotel revenue management. Cornell Hospitality Reports, 2010. Раздел 10(14). URL: https://rb.gy/usdum (дата обращения: 20.05.2023)
13. Dhar V. Data science and prediction. Communications of the ACM, 2013. Раздел 56(12). С. 64–73. URL: https://doi.org/10.1145/2500499 (дата обращения: 20.05.2023)
14. O'Neil C., Schutt R. Doing data science. Севастополь, Калифорния, США: O'Reilly Media, 2013. URL: https://rb.gy/ks0np (дата обращения: 20.05.2023)
15. Yangyong Z., Yun X. Dataology and data science: Up to now. 2011. URL: https://rb.gy/8qcwj (дата обращения: 20.05.2023)

16. Ivanov S., Zhechev V. Hotel revenue management–A critical literature review. Turizam: Znanstveno-Strucnicasopis, 2012. Раздел 60(2). С. 175–197. URL: https://dx.doi.org/10.2139/ssrn.1977467 (дата обращения: 20.05.2023)
17. Anderson C.K. The impact of social media on lodging performance. Cornell Hospitality Report, 2012. Раздел 12(15). С. 4–11. URL: https://rb.gy/jln6h (дата обращения: 20.05.2023)
18. Ivanov, S. Hotel revenue management: From theory to practice. Варна, Болгария: Zangador, 2014. URL: https://ssrn.com/abstract=2447337 (дата обращения: 20.05.2023)
19. Huang H.C., Chang A.Y., Ho C.C. Using artificial neural networks to establish a customer-cancellation prediction model. Przeglad Elektrotechniczny, 2013. Раздел 89(1b). С. 178–180. URL: https://rb.gy/pinhs (дата обращения: 20.05.2023)
20. Yoon M.G., Lee H.Y., Song Y.S. Linear approximation approach for a stochastic seat allocation problem with cancellation & refund policy in airlines. Journal of Air Transport Management, 2012. Раздел 23. С. 41–46. URL: https://doi.org/10.1016/j.jairtraman.2012.01.013 (дата обращения: 20.05.2023)
21. Phillips R.L. Pricing and revenue optimization. Стэнфорд, Калифорния, США: Stanford University Press, 2021. URL: http://www.sup.org/books/title/?id=31628 (дата обращения: 20.05.2023)
22. Chen A.H., Peng N., Hackley C. Evaluating service marketing in airline industry and Its Influence on student passengers' purchasing behavior using Taipei–London route as an example. Journal of Travel & Tourism Marketing, 2008. Раздел 25(2). С. 149–160. URL: http://dx.doi.org/10.1080/10548400802402503 (дата обращения: 20.05.2023)
23. Park J.W., Robertson R., Wu C.L. Modelling the Impact of airline service quality and marketing variables on passengers' future behavioral intentions. Transportation Planning and Technology, 2006. Раздел 29(5). С. 359–381. URL: http://dx.doi.org/10.1080/03081060600917686 (дата обращения: 20.05.2023)

24. Hastie T., Tibshirani R., Friedman, J. The elements of statistical learning. Springer series in statistics Springer, Berlin, 2001. URL: https://hastie.su.domains/Papers/ESLII.pdf (дата обращения: 20.05.2023)
25. DeKay F., Yates B., Toh R.S. Non-performance penalties in the hotel industry. International Journal of Hospitality Management, 2004. Раздел 23(3). С. 273–286. URL: http://dx.doi.org/10.1016/j.ijhm.2003.11.003 (дата обращения: 20.05.2023)
26. Chen T., Guestrin C. Xgboost: A scalable tree boosting system. ACM, 2016. С. 785–794. URL: https://doi.org/10.1145/2939672.2939785 (дата обращения: 20.05.2023)
27. Antonio N., de Almeida A., Nunes L. Predicting hotel bookings cancellation with a machine learning classification model, 16th IEEE International Conference Machine Learning Application, IEEE, Канкун, Мексика, 2017. С. 1049–1054. URL: http://dx.doi.org/10.1109/ICMLA.2017.00-11 (дата обращения: 20.05.2023)
28. Microsoft, SQL Server Management Studio SSMS. Статья: 31.03.2023. URL: https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms (дата обращения: 20.05.2023)
29. Antonio N., de Almeida A., Nunes L. Hotel booking demand datasets. Data Brief, 2018. Раздел 22. С. 41-49. URL: https://doi.org/10.1016/j.dib.2018.11.126 (дата обращения: 20.05.2023)
30. Задачи Data Mining. Классификация и кластеризация, ИНТУИТ. URL: https://intuit.ru/studies/courses/6/6/lecture/166 (дата обращения: 20.05.2023)
31. Машинное обучение для начинающих: алгоритм случайного леса. Alex Maszański, 2021. URL: https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12 (дата обращения: 20.05.2023)
32. Use Voting Classifier to improve the performance of your ML model, Satyam Kumar, 2021. URL: https://towardsdatascience.com/use-voting-classifier-to-improve-the-performance-of-your-ml-model-805345f9de0e (дата обращения: 20.05.2023)

33. Логистическая регрессия (Logistic Regression), Loginom, 2023. URL: https://wiki.loginom.ru/articles/logistic-regression.html (дата обращения: 20.05.2023)
34. Метод опорных векторов (SVM), Интернет-вики, Студенческий командный чемпионат мира по программированию ICPC в Северной Евразии, 2022. URL: https://neerc.ifmo.ru/wiki/index.php?title=Метод_опорных_векторов_(SVM) (дата обращения: 20.05.2023)
35. ISO 3166-3:2013 Codes for the representation of names of countries and their subdivisions, 2013. URL: https://www.iso.org/standard/63547.html (дата обращения: 20.05.2023)
36. James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning, Vol. 12, Springer, 2013. URL: https://www.statlearning.com/ (дата обращения: 20.05.2023)
37. Fawcett T. An introduction to roc analysis, Pattern Recognition Letters, ROC Analysis in Pattern Recognition, 2006. Раздел 27(8). С. 861–874. URL: https://doi.org/10.1016/j.patrec.2005.10.010 (дата обращения: 20.05.2023)

ПРИЛОЖЕНИЕ

Список рисунков и таблиц