Министерство науки и высшего образования РФ   
Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский технологический университет «МИСиС»

Институт ИТКН

Кафедра АСУ

**ОТЧЕТ ПО УЧЕБНОЙ**

**ПРАКТИКЕ**

**Выполнил:** Коншин Кирилл Павлович  
**ст. гр.** БИВТ-21-10

Подпись:

**Проверил:**

Москва 2023

Оглавление

[Введение 3](#_Toc140821459)

[2. Постановка Задачи 5](#_Toc140821460)

[2.1. Платформа 5](#_Toc140821461)

[2.2. Задача 7](#_Toc140821462)

[2.3. Описание данных 7](#_Toc140821463)

[3. Предобработка и анализ данных 9](#_Toc140821464)

[3.1. Предобработка 9](#_Toc140821465)

[3.2. Анализ 11](#_Toc140821466)

[4. Создание и тестирование модели 28](#_Toc140821467)

[5. Заключение 31](#_Toc140821468)

[6. Список использованных источников 33](#_Toc140821469)

# Введение

Машинное обучение является одной из наиболее захватывающих и перспективных областей в современных компьютерных науках. Эта область исследует разработку и применение алгоритмов, которые позволяют компьютерам извлекать полезные знания и делать предсказания на основе данных, не требуя явного программирования. Машинное обучение играет ключевую роль в различных областях, включая компьютерное зрение, естественный язык, автономные технологии, финансы, медицину, рекомендательные системы и многое другое.

Целью данной практической работы является изучение и практическое применение основных концепций и методов машинного обучения. В ходе работы мы будем изучать различные типы моделей машинного обучения, такие как линейные модели, деревья решений, а также методы обработки данных, разделения выборки на тренировочную и тестовую, оценки качества моделей и техники предварительной обработки данных.

Также мы разработаем и реализуем практические примеры применения машинного обучения на реальных наборах данных. Будем рассматривать задачи классификации, регрессионного анализа, чтобы полноценно понять потенциал и ограничения машинного обучения в решении разнообразных задач.

Практическая работа включает реализацию алгоритмов и моделей с использованием популярных библиотек и фреймворков машинного обучения, таких как SciPy, scikit-learn и PyTorch. Мы также обсудим важные аспекты, связанные с выбором алгоритмов, настройкой гиперпараметров и оптимизацией производительности моделей.

В результате данной практической работы я планирую получить практические навыки работы с машинным обучением, которые смогу применять их в реальных проектах и исследованиях. Это позволит мне лучше понимать потенциал машинного обучения и способствовать его дальнейшему развитию и применению в различных областях человеческой деятельности.

Кроме того, данная практическая работа также ставит перед собой цель развить навыки анализа данных, предобработки и визуализации, которые являются важными этапами в подготовке данных для применения алгоритмов машинного обучения, научить проводить исследовательский анализ данных, обнаруживать аномалии, работать с пропущенными значениями, масштабировать признаки и приводить данные к нужному формату.

Важным аспектом практической работы является также работа с реальным набором данных. Это позволит столкнуться с реальными проблемами и осознать важность правильного выбора алгоритмов и подходов в решении конкретных задач

Кроме того, в рамках практической работы мне предоставляется возможность реализовать свои собственные идеи и экспериментировать с различными подходами к решению задач машинного обучения. Это способствует развитию творческого мышления и позволяет проявить свои инновационные подходы к решению сложных задач.

Итак, данная практическая работа предоставляет уникальную возможность познакомиться с основами и практическими аспектами машинного обучения, получить опыт работы с популярными библиотеками и фреймворками, а также развить навыки анализа данных и принятия решений на основе результатов моделей машинного обучения. Все это является важным вкладом в формирование будущего специалиста, готового применять машинное обучение для решения реальных задач и способных продвигать науку и технологии вперед.

# Постановка Задачи

## Платформа

В данной лабораторной работе передо мной стояла задача познакомиться с машинным обучением, а также принять участие в соревновании на платформе, предоставляющей доступ к какким-либо данным.

Потому как машинным обучением я уже занимался, то я заранее сделал свой выбор в пользу DataCamp -– одну из ведущих платформ по машинному обучению. Считаю нужным выделить плюсы данной платформы:

1. Интерактивные курсы: DataCamp предлагает интерактивные курсы, которые обеспечивают активное участие студента. Пользователи работают с реальными наборами данных, выполняют кодирование и решают задачи, что позволяет им немедленно применить изученные концепции на практике.
2. Широкий выбор тем: Платформа DataCamp предлагает курсы по различным аспектам анализа данных и машинного обучения, включая языки программирования Python и R, а также такие технологии, как SQL, Tableau, TensorFlow, PyTorch и многое другое.
3. Профессиональное обучение: DataCamp помогает профессионалам развивать свои навыки и продвигаться в карьере. Платформа предлагает специализированные курсы и практические задания, которые помогают участникам освоить сложные темы и стать более востребованными на рынке труда.
4. Соревнования и проекты: DataCamp предлагает участникам возможность применить свои знания в реальных проектах и соревнованиях. Это дает студентам возможность применить изученные концепции на практике и показать свои навыки на публичной платформе.

DataCamp активно используется как начинающими аналитиками данных и специалистами, так и опытными профессионалами, которые хотят развивать свои навыки и быть в курсе последних тенденций в анализе данных и машинном обучении. Платформа имеет обширное сообщество пользователей, которые делятся своими знаниями и опытом, что делает DataCamp важным ресурсом для тех, кто стремится стать успешным в анализе данных и машинном обучении.

Все свои манипуляции с данными я проводил в среде для Data-Science разработчиков DataSpell от российской компании JetBrains. Считаю данную среду наиболее удобной, благодаря удобному интерфейсу, большому количеству шорткатов и сниппетов, знакомых мне по опыту java-разработки.

## Задача

**Задача -- Прогнозирование отмены бронирования отелей**

Информация  
Вы поддерживаете отель в рамках проекта, направленного на увеличение доходов от бронирования номеров. Они верят, что могут использовать науку о данных, чтобы помочь им сократить количество отмен.  
  
Они попросили вас использовать любую подходящую методологию и провести анализ данных, чтобы определить, что влияет на то, будет ли бронирование выполнено или отменено. Они намерены использовать результаты вашей работы, чтобы уменьшить вероятность того, что кто-то отменит их бронирование.

В данном случае перед нами стоит задача проанализировать данные, а также построить модель, которая наиболее точно предсказывает классификацию новых наблюдений(будет отменено/успешно).

## Описание данных

Данные содержат 36275 записей о бронировании номеров в отелях с 2017 по 2018 года. В данном наборе присутствует 19 предикторов(столбцов):

* Booking\_ID: уникальный идентификатор каждого бронирования
* no\_of\_adults: Количество взрослых
* no\_of\_children: Количество детей
* no\_of\_weekend\_nights: Количество ночей выходного дня (суббота или воскресенье), на которые гость останавливался или забронировал проживание в отеле
* no\_of\_week\_nights: Количество ночей недели (с понедельника по пятницу), на которые гость останавливался или забронировал проживание в отеле
* type\_of\_meal\_plan: Тип плана питания, забронированного клиентом:
* required\_car\_parking\_space: Требуется ли клиенту место для парковки автомобиля? (0 - Нет, 1- Да)
* room\_type\_reserved: Тип номера, забронированного клиентом. Значения зашифрованы (закодированы) гостиницами INN.
* lead\_time: Количество дней между датой бронирования и датой  
  прибытия
* arrival\_year: Год даты прибытия
* arrival\_month: Месяц даты прибытия
* arrival\_date: Дата месяца
* market\_segment\_type: Обозначение сегмента рынка.
* repeated\_guest: Является ли клиент повторным гостем? (0 - Нет, 1- Да)
* no\_of\_previous\_cancellations: количество предыдущих бронирований, которые были отменены клиентом до текущего бронирования
* no\_of\_previous\_bookings\_not\_canceled: Количество предыдущих бронирований, не отмененных клиентом до текущего бронирования
* avg\_price\_per\_room: Средняя цена за сутки бронирования; цены на номера являются динамическими. (в евро)  
  no\_of\_special\_requests: Общее количество специальных запросов, сделанных клиентом (например, высокий этаж, вид из номера и т.д.)
* booking\_status: Флаг, указывающий, было ли бронирование отменено или нет.

# Предобработка и анализ данных

## Предобработка

На начальном этапе перед анализом данных нужно заняться их предобработкой. Для начала с помощью графиков посмотрел, как вообще данные «выглядят». Получил следующее:



Рисунок 1

Данные перед предобработкой

На этапе предобработки я выделил следующие задачи для преобразования данных к приемлемому виду:

1) Удалить все null-значения, а также данные, содержащие пусты ячейки  
2) удалить все значения не имеющие смысла (не выбран тип питания, в резервировании нет ни одного взрослого)  
3)преобразовать нужные данные, и далее удалить ненужные ячейки (создать новый признак "общее время пребывания" и если он равен нулю – удалить значение)  
4)создать новую колонку "полное время бронирования" для удобного рассмотрения данных  
5)создать новую колонку "полное количество гостей"  
6)создать колонку "общее количество бронирований", составленную из отмененных и подтвержденных бронирований

Также стоит удалить данные, где цена за номер указана нулевой

После выполнения преобразований я получил на 10 тысяч меньше наблюдений в выборке.

## Анализ

На этапе анализа данных я решил посмотреть, как выглядят данные относительно классифицирующей переменной(отменено/нет). Ниже представлен рисунок с категориальными переменными после обработки и построенными для визуализирования их относительно классифицирующей переменной.

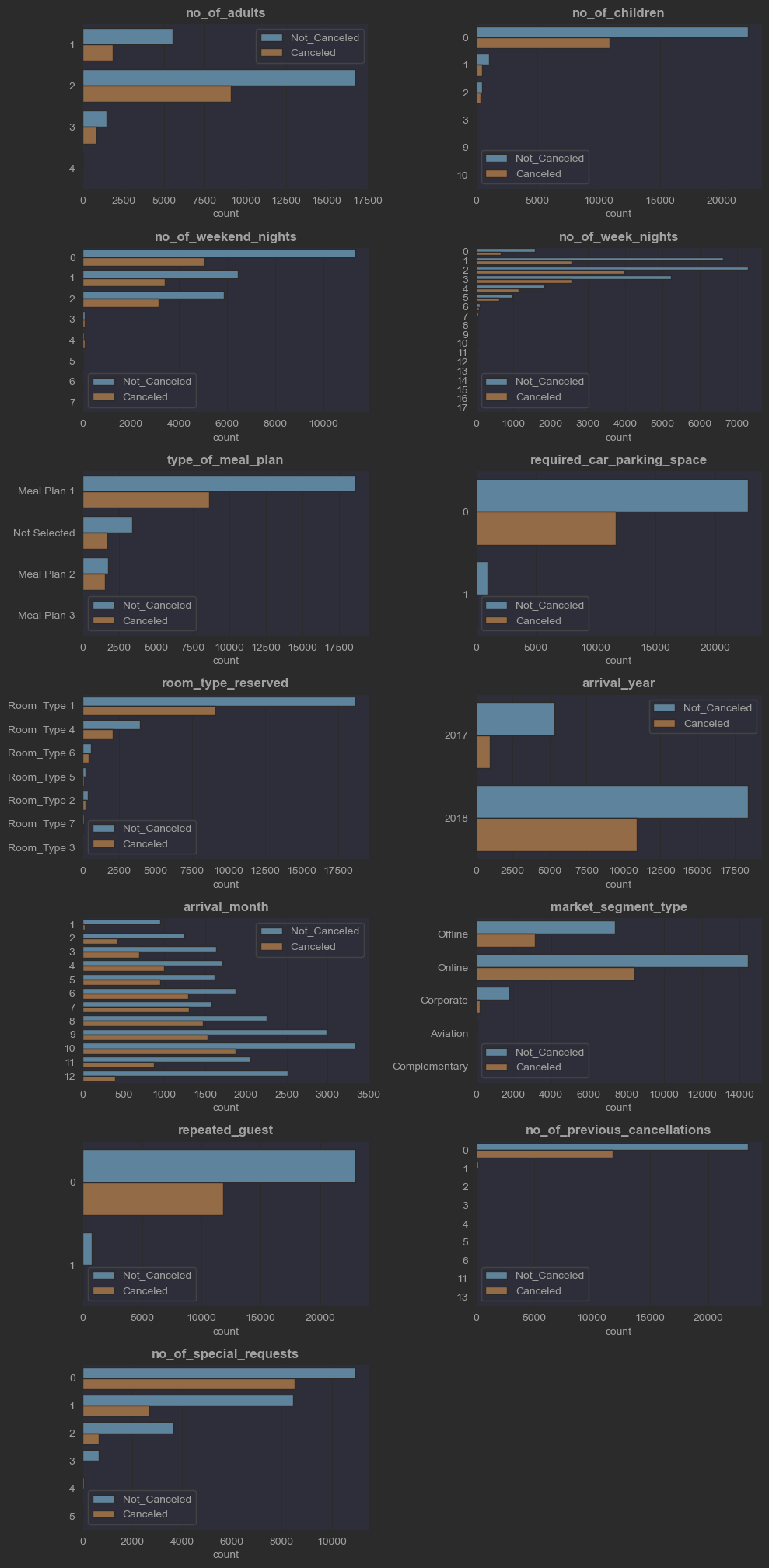


Рисунок 2

Данные после предобработки

После этого было сделано несколько выводов из графиков:

1. Если гость уже не в первый раз бронирует номер в отеле, то он не отменит свою бронь  
   2) Чем больше гость указывает предпочтений перед бронью, тем меньше шанс, что он отменит  
   3) Больше всего шанс отмены именно в онлайн - сегменте

Следом я построил гистограмму показывающую данные по дате бронирования относительно классифицирующей переменной статус бронирования и ядерной оценкой плотности для каждого из вариантов классификации.

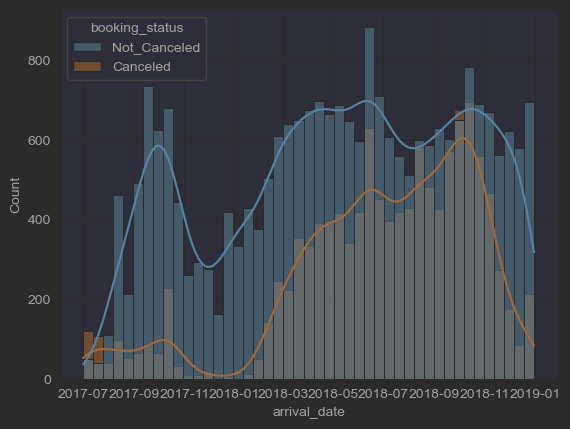


Рисунок 3

Дата бронирования и статус бронирования в датафрейме

По данному графику видно, что количество резервирований остается примерно на одном уровне, в то время как количество отмен растет.

Если также представить выборку относительно месяцев бронирования, можно сделать вывод Исходя из графика видно, что наибольшее число отмен брони происходят в период с августа по октябрь, что также связано с увеличением количества броней в общем случае. В общем случае повышенный шанс отмены происходит в конце весны и всё лето, что также связано с повышением спроса на номера в отелях.

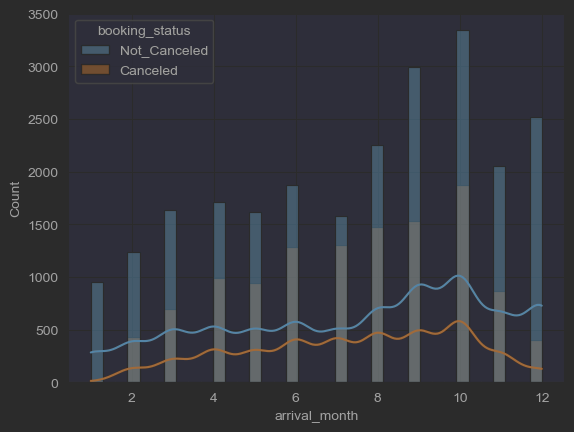


Рисунок 4

Данные по месяцам

Далее я решил посмотреть данные относительно переменных "общее количество прошлых бронирований", "количество отмененных бронирований" и "количество прошлых бронирований "

Percentage Canceled for 1 person/people: 24.87%

Percentage Canceled for 2 person/people: 34.89%

Percentage Canceled for 3 person/people: 36.44%

Percentage Canceled for 4 person/people: 43.98%

Percentage Canceled for 5 person/people: 38.46%

Percentage Canceled for 10 person/people: 0.00%

Percentage Canceled for 11 person/people: 100.00%

Percentage Canceled for 12 person/people: 0.00%

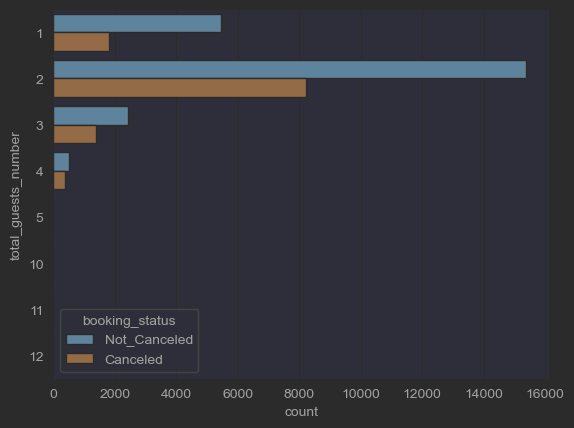


Рисунок 5

Количество гостей и статус бронирования

Как мы видим, самый низкий процент отмен у меньшего количества гостей (данные, где больше 4 гостей брать в такой расчет не стоит за маленьким объемом выборки)

Рассмотрев данные относительно прошлых бронирований сделал вывод, что если человек уже бронировал номер, то в новом бронировании более чем с 90% вероятностью в новом наблюдении бронирование отменено не будет.

Далее показан просмотр данных по общему времени, на которое забронирован номер. Результаты дают понять, что средний процент отмены находится на уровне 30-40%, потому как ликвидными для анализа являются данные, где количество дней пребывания меньше 10(размер выборки позволяет делать вывод в данном случае позволяет делать какие-либо выводы).

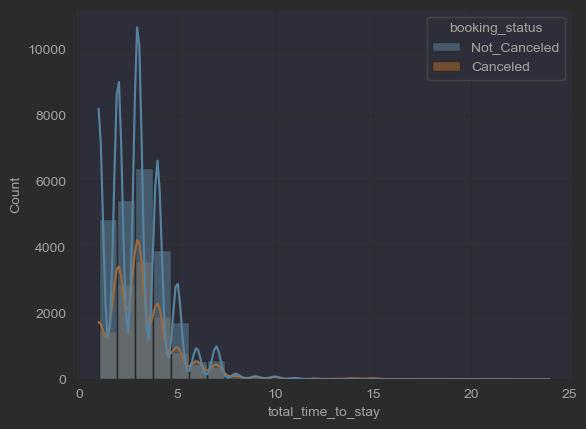


Рисунок 6

Общее время брони номера

Данные относительно даты пребывания

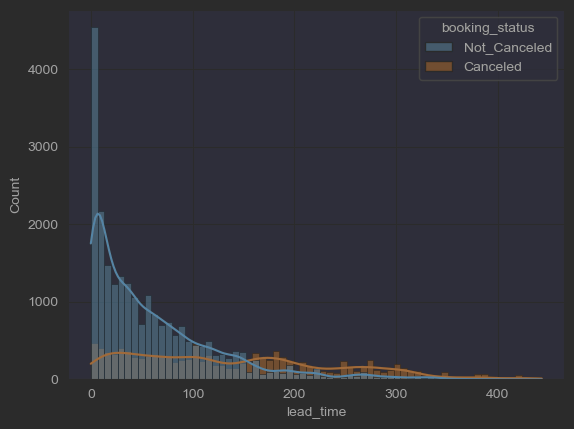


Рисунок 7

Дата пребывания

Как мы видим, данные имеют log-нормальное распределение. Видно, что чем меньше времени остается до даты бронирования, тем меньше процент отмены, особенно в начальных сегментах(что может быть связано с резкой необходимостью в жилье на отдыхе). Также стоит отметить, что приемлемый уровень отмен бронирования сохраняет свою тенденцию примерно до промежутка в 100 дней, далее наблюдается резкое возрастание уровня отмен брони(что может быть связано с тем, что в большем промежутке между временем бронирования и заселением может произойти большее количество форс-мажорных ситуаций, которые могут побудить людей отменить бронирование).

Следующий блок я полностью посвятил анализу данных по переменной «средняя цена за номер», так как именно эта переменная представляет наибольший интерес для обычного обывателя. Разбив весь промежуток данных по ценам на интервалы по 15 евро, получилось достичь приемлемого вида графика, а также точечных оценок мат.ожидания для цены.

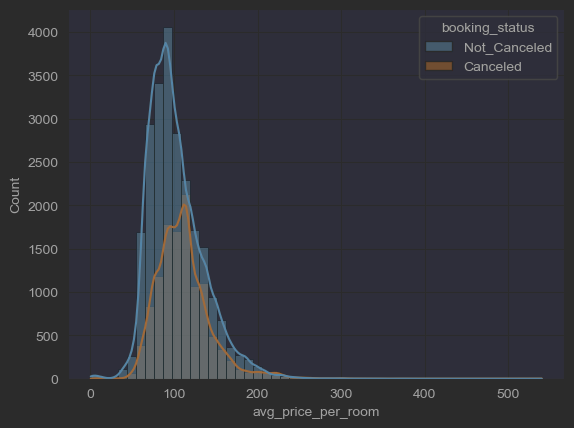


Рисунок 8

Гистограмма средней цены за номер

Промежуток цен: (0, 15], Процент отмененных бронирований: 10.00%

Промежуток цен: (15, 30], Процент отмененных бронирований: 50.00%

Промежуток цен: (30, 45], Процент отмененных бронирований: 3.27%

Промежуток цен: (45, 60], Процент отмененных бронирований: 19.02%

Промежуток цен: (60, 75], Процент отмененных бронирований: 21.89%

Промежуток цен: (75, 90], Процент отмененных бронирований: 26.93%

Промежуток цен: (90, 105], Процент отмененных бронирований: 32.50%

Промежуток цен: (105, 120], Процент отмененных бронирований: 46.89%

Промежуток цен: (120, 135], Процент отмененных бронирований: 40.91%

Промежуток цен: (135, 150], Процент отмененных бронирований: 34.44%

Промежуток цен: (150, 165], Процент отмененных бронирований: 35.16%

Промежуток цен: (165, 180], Процент отмененных бронирований: 34.00%

Промежуток цен: (180, 195], Процент отмененных бронирований: 26.08%

Промежуток цен: (195, 210], Процент отмененных бронирований: 38.36%

Промежуток цен: (210, 225], Процент отмененных бронирований: 60.59%

Промежуток цен: (225, 240], Процент отмененных бронирований: 42.86%

Промежуток цен: (240, 255], Процент отмененных бронирований: 48.48%

Промежуток цен: (255, 270], Процент отмененных бронирований: 63.64%

Промежуток цен: (270, 285], Процент отмененных бронирований: 33.33%

Промежуток цен: (285, 300], Процент отмененных бронирований: 53.85%

Промежуток цен: (300, 315], Процент отмененных бронирований: 33.33%

Промежуток цен: (315, 330], Процент отмененных бронирований: 100.00%

Промежуток цен: (330, 345], Процент отмененных бронирований: 0.00%

Промежуток цен: (345, 360], Процент отмененных бронирований: 0.00%

Промежуток цен: (360, 375], Процент отмененных бронирований: 100.00%

Промежуток цен: (375, 390], Процент отмененных бронирований: 0.00%

Промежуток цен: (525, 540], Процент отмененных бронирований: 100.00%

Средний процент отмененных бронирований по фрагментам: 39.09%

Данные по бронированию номеров относительно средней цены за номер подчинены закону нормального распределения (действительные и отмененные бронирования также), о чем свидетельствуют ядерные оценки плотности данного набора. Я разделил данные на промежутки по 15, чтобы лучше посмотреть в более доверительной форме, каков процент отмен по сегментам(также привел среднее значение для всего набора, но среднее наиболее подвержено выбросам и экстремальным значениям, потому и приведены данные по сегментам).

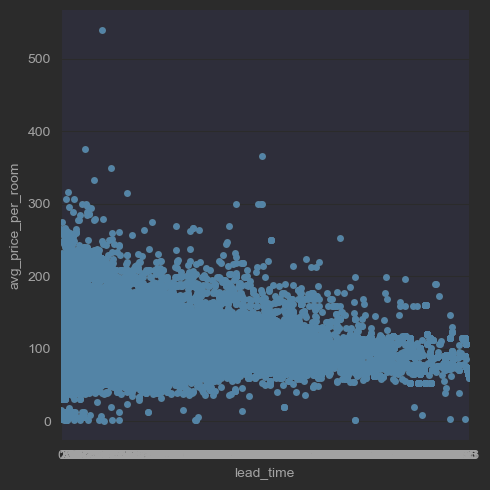


Рисунок 9

Зависимость времени между бронированием и ценой за номер

Видно, что чем раньше сделано бронирование. тем дешевле будет цена

Изображение выглядит как снимок экрана, диаграмма, 3D-моделирование, текст

Автоматически созданное описание

Рисунок 10

Сегмент рынка, где сделали бронирование, относительно цены на номер

Вывод: в среднем номера дороже приобретаются в онлайн-сегменте

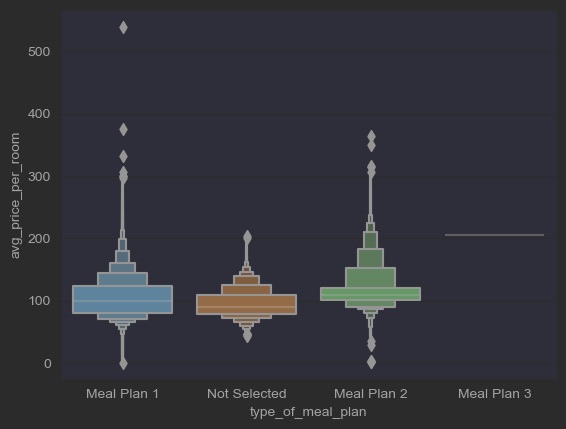


Рисунок 11

Тип питание и цена за номер

Вывод: не видно, какой-то сильной разницы в ценах по питанию, разве что второй дороже первого, а третий второго.

Изображение выглядит как снимок экрана, диаграмма, 3D-моделирование

Автоматически созданное описание

Рисунок 12

Влияние количества гостей на цену номера

Вывод: как видим, с ростом количества гостей увеличивается и чек

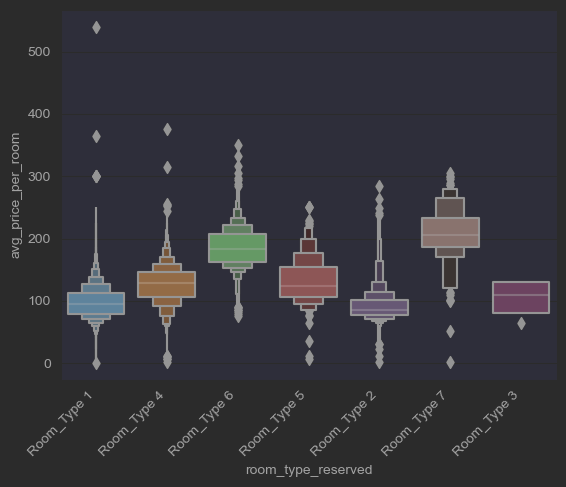


Рисунок 13

Тип комнаты и средняя цена за него

Вывод: Самые дорогие типы комнат - 6 и 7

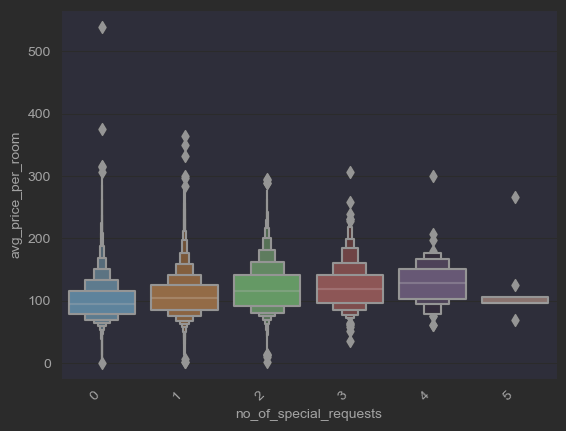


Рисунок 14

Количество специальных пожеланий и цена за номер

Количество пожеланий не влияет на цену номера

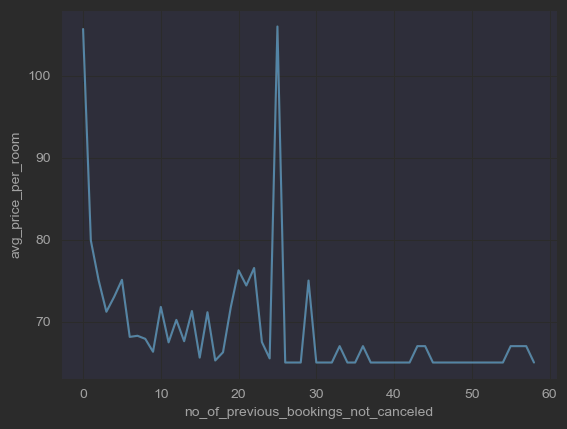


Рисунок 15

Линейная диаграмма количества предыдущих бронирований неотменённых и цена за номер

Как видно, возможно, имеется корреляция меду количеством прошлых успешных бронирований, но такая корреляция виднеется только на малых количествах прошлых бронирований.

# Создание и тестирование модели

Для начала, я решил глянуть таблицу корреляции предикторов между собой(построил тепловую диаграмму)

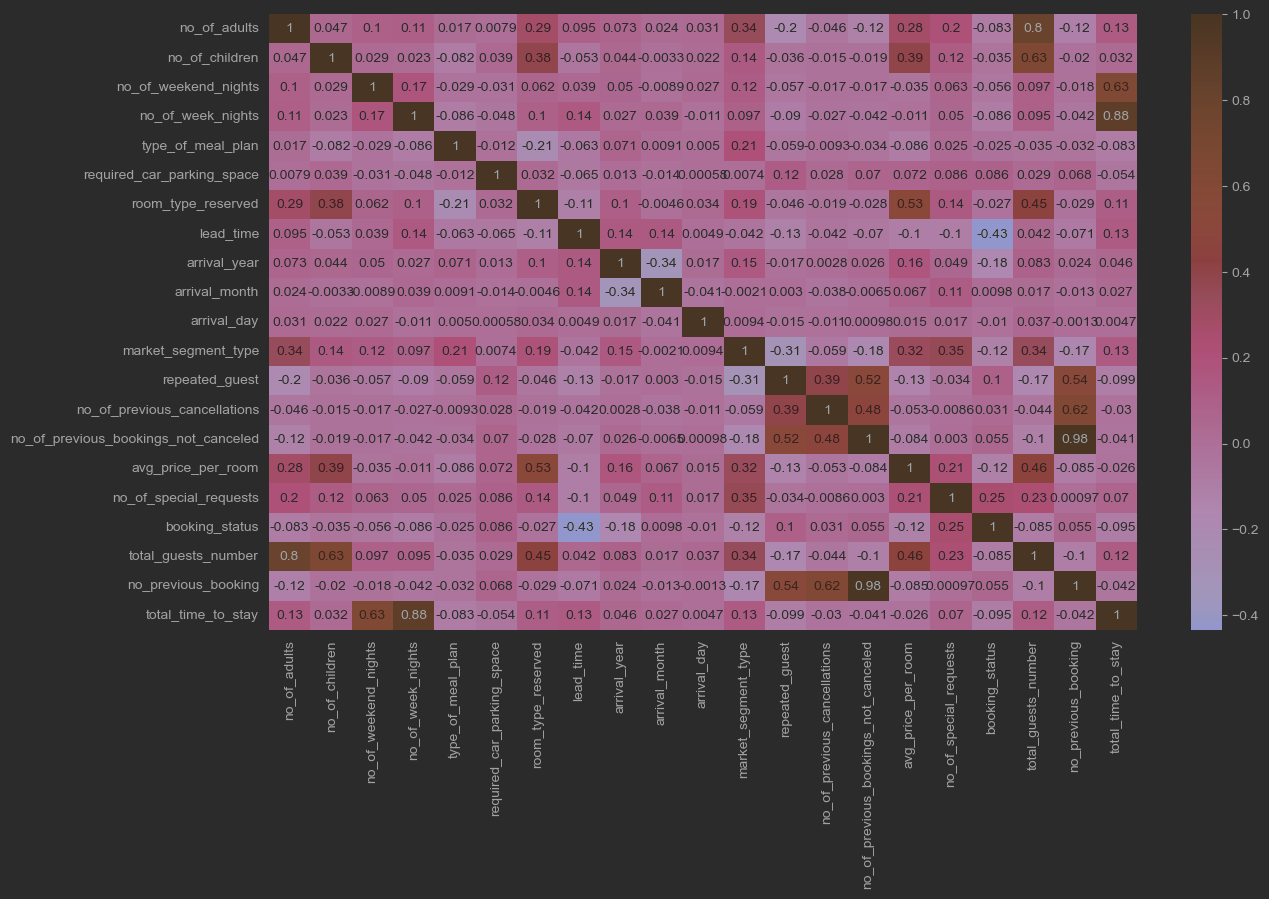


Рисунок 16

Тепловая диаграмма корреляции предикторов

Показана тепловая карта корреляций предикторов. Из интересного могу отметить наличие связи между временем между датой брони и датой заезда, средней ценой за номер и его типом, а также типом номера и количеством гостей

После еще одной обработки данных, чтобы подготовить их для построения, я остановился на списке из 6 самых популярных классификаторов:

* + - Деревья решений
    - Градиентный спуск
    - Рандомный лес
    - Дополнительные деревья решений
    - K-ближайших соседей
    - Логистистическая регрессия

Далее разбив всю выборку на обучающую(30%) и контрольную, начал тест всех классификаторов. Выведя их матрицы неточностей, а также подсчитав основные метрики для оценки моделей, я нашел лучшее решение из приведенных – рандомный лес. Его точность равна 90% что позволяет говорить о надежности данной модели.  
print(best)  
print(maxv)

RANDOM\_FOREST

0.8956533326880778

Далее для ознакомления я решил глянуть, какие же предикторы вносят наибольший вклад в предсказания.

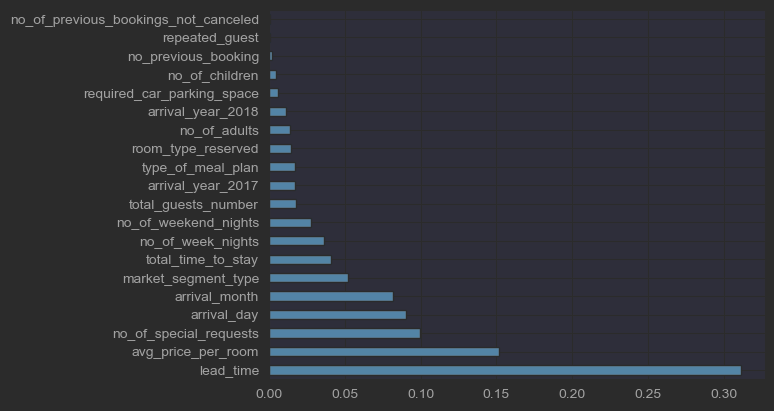


Рисунок 17

Вклад предикторов в предсказания для рандомного леса

Оказалось, что наиболее важными предикторами для правильной классификации являются время между бронью и заездом, а также средняя цена за номер.

Проведенный мною анализ гостиничного номера позволяет лучше выявлять зависимости между различными предикторами, а построенная мною модель, основанная на рандомном лесе, поможет лучше определять, откажется ли от бронирования постоялец или нет. Данные предсказания могут предоставить возможность.

# Заключение

В ходе данной работы был выполнен комплексный анализ данных по отелям с целью прогнозирования отмены бронирования. Я начал с предобработки данных, чтобы убедиться в их точности и пригодности для последующего анализа. Затем я провел исследовательский анализ данных, визуализировал результаты и выявил ключевые особенности, которые могут повлиять на решение клиента об отмене бронирования.

После этого я разделил данные на тренировочную и тестовую выборки, чтобы обучить и протестировать различные методы классификации. В ходе тестирования я использовал 6 различных методов и измерял их производительность с помощью f1-метрики, которая учитывает как точность, так и полноту предсказаний.

Исследуя результаты, я обнаружил, что наилучший метод классификации для прогнозирования отмены бронирования - это случайный лес. Этот метод показал впечатляющую f1-метрику на уровне 0.9, что говорит о его высокой точности и полноте предсказаний. Такой высокий уровень производительности говорит о том, что наша модель может быть применена для предсказания отмены бронирования с высокой степенью уверенности.

В заключение, данная работа позволила мне получить ценный опыт в области анализа данных и машинного обучения. Я освоил различные этапы работы с данными, от предобработки до выбора наилучшей модели классификации. Результаты этой работы имеют реальное применение для бизнеса и помогут отелю принимать осознанные решения для сокращения отмен бронирования и увеличения доходов.

Я благодарен за возможность участвовать в этом соревновании на платформе DataCamp, которая предоставила мне такой ценный опыт и дала возможность применить свои знания на практике. Полученные результаты и умения будут важным активом в моей карьере и помогут мне продолжать развиваться в области анализа данных и машинного обучения.

# Список использованных источников

1. Аббакумов В.Л., Анализ данных на Python в примерах и задачах.– Computer Science Center, 2018.
2. Вандер Плас Дж. Python для сложных задач: наука о данных и машинное обучение / пер. с англ. И. Пальти. — СПб.: Питер, 2019. — 572 с.
3. Вандер Плас Дж. Python для сложных задач: наука о данных и машинное обучение / пер. с англ. И. Пальти. — СПб.: Питер, 2019. — 572 с.
4. Гатман А.Дж. Becoming a Data head: how to think, speak and understand data science, statistics and machine learning — Москва, 2023