ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Экономика и анализ данных»

Проект:
Построение и анализ экономической модели

Выполнили студенты:

группы #БЭАД221, 3 курса	Смешкова Екатерина
группы #БЭАД221, 3 курса	Цисарук Мария
группы #БЭАД221, 3 курса	Миннеахметова Рената

Преподаватель:

Станкевич Иван Павлович

Москва 2024

Содержание

Введение

Цель исследования заключается в анализе набора данных о бронировании авиабилетов, полученного с веб-сайта “Ease My Trip”, и проведении различных статистических тестов для получения значимой информации. Для обучения набора данных и прогнозирования непрерывной целевой переменной будет использован статистический алгоритм "Линейная регрессия"(Linear Regression). "Ease My Trip"— это интернет-платформа для бронирования авиабилетов, которой пользуются потенциальные пассажиры для покупки билетов.

Это исследование является актуальным, поскольку с увеличением доступности и востребованности авиаперелетов пассажирам важно понимать факторы ценообразования, чтобы принимать более обоснованные решения по покупке билетов.

Мы взяли данные с сайта Kaggle https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data. Эти данные были получены с помощью инструмент для веб-скрейпинга Octoparse. Данные были собраны в двух частях: для билетов эконом-класса и для билетов бизнес-класса. Всего с сайта было извлечено 300261 уникальных вариантов бронирования авиабилетов. Сбор данных происходил в течение 50 дней, с 11 февраля по 31 марта 2022 года.

Экономическая модель

Описание имеющихся в данных переменных:

1. Авиакомпания (Airline): В столбце "airline"хранится название авиакомпании. Это категориальная переменная, содержащая 6 различных авиакомпаний. Хранится строка.

2. Рейс (Flight): Столбец "Flight"содержит информацию о коде рейса самолета. Это категориальная переменная. Хранится строка.

3. Город вылета (Source City): Город, из которого отправляется рейс. Это категориальная переменная, содержащая 6 уникальных городов. Хранится строка.

4. Время вылета (Departure Time): Это производная категориальная переменная, созданная путем группировки временных интервалов в категории. Она хранит информацию о времени вылета и имеет 6 уникальных временных меток - раннее утро, утро, полдень, день, вечер, поздний вечер, полночь. Хранится строка.

5. Пересадки (Stops): Категориальная переменная с 3 различными значениями, которая хранит информацию о количестве пересадок между городом вылета и городом назначения. Хранится строка.

6. Время прибытия (Arrival Time): Это производная категориальная переменная, созданная путем группировки временных интервалов в категории. Она содержит 6 различных временных меток и информацию о времени прибытия аналогично времени вылета.

7. Город назначения (Destination City): Город, в который прилетает самолет. Это категориальная переменная, содержащая 6 уникальных городов. Хранится строка.

8. Класс (Class): Категориальная переменная, содержащая информацию о классе места; имеет два различных значения: Бизнес и Эконом. Хранится строка.

9. Продолжительность (Duration): Непрерывная переменная, отображающая общее время, необходимое для путешествия между городами, в часах. Хранится число с плавающей точкой.

10. Оставшиеся дни (Days Left): Это производная характеристика, рассчитываемая путем вычитания даты бронирования из даты поездки. Хранится целое число

11. Цена (Price): Целевая переменная, которая хранит информацию о стоимости билета. Хранится число с плавающей точкой.

Выбор объясняющих переменных

Теперь выберем объясняющие переменные. Каждая из переменных представляет собой важный фактор, который влияет на стоимость авиабилетов. Рассмотрим каждую переменную подробнее:

- **Авиакомпания (Airline)**: Авиакомпании могут применять разные ценовые стратегии в зависимости от их бренда, репутации, качества обслуживания и уровня предоставляемых услуг. Известные и крупные авиакомпании могут позволить себе устанавливать более высокие цены, что увеличивает цену билета. Это также подтверждается экономической теорией дифференциации продукта, когда компании устанавливают цены в зависимости от их рыночной позиции и качества предоставляемых услуг.

- **Время вылета (Departure Time)**: Исследования показывают, что время вылета влияет на спрос на билеты. Например, рейсы утром или вечером могут быть более востребованы, так как они удобны для бизнес-пассажиров или туристов. Это также подтверждается концепцией эластичности спроса по времени: в пиковые часы, когда спрос высок, авиакомпании могут устанавливать более высокие цены [**Borenstein1994**].

- **Класс (Class)**: Класс обслуживания является одним из ключевых факторов, который напрямую влияет на цену билета. .

- **Время прибытия (Arrival Time)**: Время прибытия также может влиять на цену билета, так как рейсы, прибывающие в утренние или вечерние часы, могут быть более удобными для пассажиров, что увеличивает спрос на такие рейсы. Это связано с тем, что пассажиры склонны предпочитать рейсы, которые позволяют им максимально эффективно использовать время по прибытию, что может повлиять на цену билета [**Borenstein1992**].

- **Продолжительность (Duration)**: Длительность рейса влияет на его стоимость из-за операционных затрат авиакомпаний, таких как топливо, обслуживание и персонал. Более долгие рейсы требуют больше ресурсов и, соответственно, могут стоить дороже. Это согласуется с теорией издержек, где продолжительность рейса прямо пропорциональна затратам на его обслуживание [**Morrell2008**].

- **Оставшиеся дни (Days Left)**: Цена билета часто изменяется в зависимости от того, сколько времени осталось до вылета. В теории предложения и спроса это объясняется тем, что ближе к дате вылета, спрос на билеты может возрасти, особенно если количество доступных мест уменьшается. Это также связано с тем, что авиакомпании, как правило, повышают цены на билеты по мере приближения даты рейса, что подтверждается многочисленными исследованиями в области динамического ценообразования [**Zohar2002**].

- **Топ-3 города (Top_3_City)**(создадим сами): Переменная Top_3_City — это дамми-переменная, принимающая значения 0 или 1, где 1 означает, что город входит в топ-3, а 0 — что не входит. Эта переменная указывает, входит ли город вылета или назначения в топ-3 самых популярных городов по количеству рейсов. Популярные города, как правило, имеют высокий спрос на авиаперевозки, что может приводить к более высоким ценам на билеты. Это связано с тем, что более популярные маршруты обычно предлагают больше рейсов и могут быть связаны с большей конкуренцией среди авиакомпаний, что позволяет устанавливать более высокие цены. Таким образом, включение переменной `Top_3_City` помогает учитывать влияние популярности города на стоимость билетов.

- **Количество остановок (Has_Stops)** (сделаем сами): Переменная `Has_Stops` — это дамми-переменная, принимающая значения 0 или 1, где 1 означает, что рейс имеет хотя бы одну остановку, а 0 — что рейс без остановок. Эта переменная указывает, имеет ли рейс пересадки, что часто влияет на цену билета. Рейсы с пересадками могут стоить дороже по сравнению с прямыми рейсами, поскольку они обычно связаны с более длительным временем в пути, дополнительными сборами и потенциальными неудобствами для пассажиров. Таким образом, включение переменной `Has_Stops` помогает учитывать влияние наличия остановок на стоимость билетов, выделяя рейсы с пересадками и без них, что важно для анализа ценообразования и предпочтений пассажиров.

Гипотезы:

1. Гипотеза об отношении времени бронирования к цене билета. Чем больше дней остается до даты вылета (переменная "Days Left"), тем ниже будет цена билета (переменная "Price").

2. Рейсы с пересадками будут иметь более высокую цену по сравнению с прямыми рейсами. Мы предполагаем, что рейсы с одной или более остановками будут стоить дороже из-за дополнительных затрат на обслуживание, времени в пути и неудобств для пассажиров.

3. Гипотеза о влиянии популярных направлений на стоимость билетов. Рейсы между крупными городами или популярными туристическими направлениями (переменные "Source City"и "Destination City") стоят дороже.

Анализ данных

Целевая переменная: Цена (Price)

- **Средняя цена билета**: 20,889.66.
- **Минимальная цена**: 1,105 — это могут быть выбросы или билеты со скидками/акциями.
- **Максимальная цена**: 123,071 — цены для премиальных рейсов.
- **Разброс (стандартное отклонение)**: 22,697.77 — это свидетельствует о значительном варьировании цен в наборе данных.

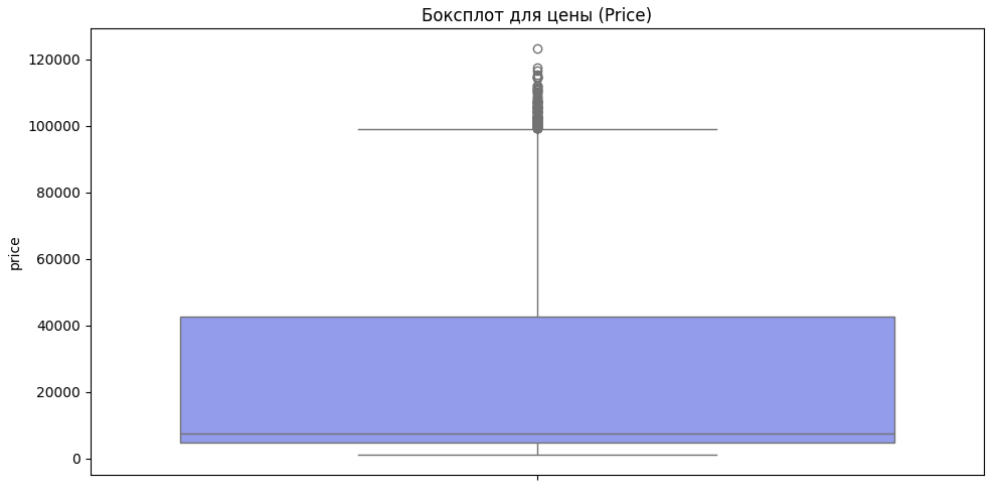

Рис. 0.1: Боксплот для цены (Price)

На графике присутствуют выбросы, которые сильно выше верхней границы "усов"(приблизительно 100,000). Эти выбросы могут быть связаны с премиальными рейсами или необычными ситуациями (например, сильно завышенные цены для отдельных рейсов или акции) (Рис. 0.1).

Логарифмирование цены имеет экономическое обоснование, поскольку оно помогает моделировать нелинейные зависимости и снижает влияние экстремальных значений, таких как очень дорогие билеты. Это также делает данные более нормальными, что улучшает результаты анализа и повышает точность моделей. Кроме того, логарифмирование преобразует гистограмму в более симметричное распределение, что делает визуализацию данных более понятной и интерпретируемой.

Итоговая гистограмма логарифма цены Рис. 0.2:

Численные показатели

Теперь перейдем к численным объясняющим переменным.

Объясняющий фактор: Длительность (Duration)

- **Средняя длительность полета**: 12.22 часа.
- **Минимальная длительность**: 0.83 часа — это могут быть короткие рейсы или ошибки в данных.
- **Максимальная длительность**: 49.83 часа — дальние международные рейсы с пересадками.

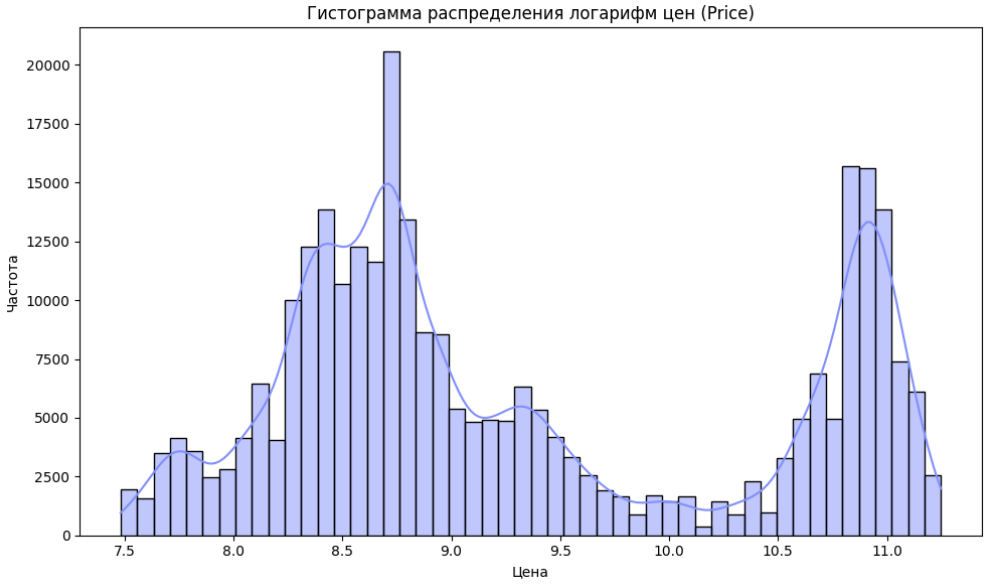

Рис. 0.2: Гистограмма распределения логарифм цен (Price)

- **Разброс (стандартное отклонение)**: 7.19 часа — это указывает на значительное разнообразие в длительности рейсов.

Объясняющий фактор: Количество дней до вылета (Days Left)

- **Среднее количество дней до вылета**: 26 дней.
- **Минимальное значение**: 1 день — для срочных бронирований.
- **Максимальное значение**: 49 дней — возможно, существует ограничение для дальних бронирований.
- **Медианное значение**: Большая часть значений (50% пассажиров) бронирует билеты за 15–38 дней до вылета.

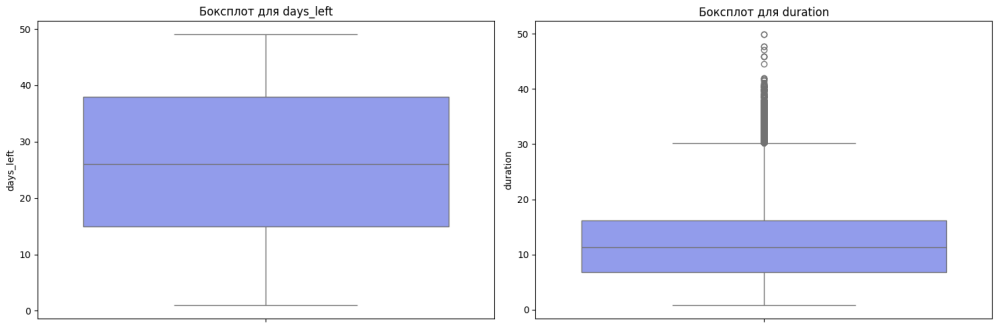

Рис. 0.3: Боксплоты для daysleft и duration

Фактор "количество дней до вылета"не имеет значительных выбросов, что означает, что данные в основном сбалансированы и соответствуют обычной практике бронирования. Фактор "длительность"имеет выбросы, которые, скорее всего, отражают экстремальные случаи с длительными рейсами, что важно учитывать при дальнейшей обработке данных или анализе. Выбросы можно оставить.

В ходе анализа данных были рассчитаны корреляции между различными переменными, что позволило сделать следующие выводы Рис. 0.4(однако пока не можем точно оценить влияние этих факторов, так как в моделях, учитывающих другие переменные, эффект может быть незначительным):

Продолжительность рейса (duration) и цена (price)

Корреляция между продолжительностью рейса и ценой составляет 0.26, что указывает на умеренную положительную связь. Это означает, что с увеличением длительности рейса цена, как правило, возрастает. Однако данная связь не является слишком сильной, что предполагает влияние других факторов, таких как класс обслуживания, авиакомпания и маршрут.

Количество дней до вылета (daysleft) и цена (price)

Корреляция между количеством дней до вылета и ценой составляет -0.19, что указывает на слабую отрицательную связь. Это говорит о том, что существует небольшая тенденция к снижению цены при увеличении числа дней до вылета. Этот результат подтверждает политику авиакомпаний предлагать более низкие цены при раннем бронировании.

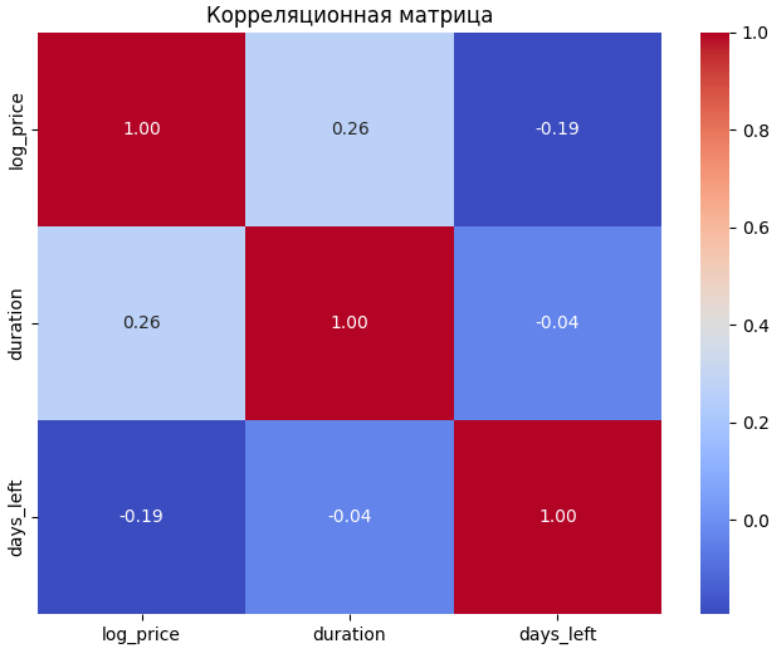

Рис. 0.4: Корреляционная матрица

Категориальные показатели

Теперь перейдем к категориальным показателям. При анализе категориальных переменных, таких как 'airline', 'departure_time', 'class', и 'arrival_time', не обнаружено супер редких категорий, которые бы встречались исключительно редко в данных. Хотя присутствуют некоторые категории, которые составляют менее 1% от общего числа наблюдений, их количество не настолько велико, чтобы существенно влиять на результаты анализа. В целом, данные категориальные переменные

имеют сбалансированное распределение, и не требуется исключать или объединять редкие категории для дальнейшего анализа.

Анализ топ-3 популярных городов

В ходе анализа данных была добавлена новая дамми-переменная `Top_3_City`, которая указывает, входит ли город вылета или назначения в топ-3 самых популярных городов по количеству рейсов.

1. Определение топ-3 городов: Для каждого города подсчитано общее количество рейсов, как для города вылета, так и для города назначения. На основе этих данных были определены три города с наибольшим количеством рейсов, которые можно считать наиболее популярными.

2. Добавление дамми-переменной: - Для каждой строки данных (рейса) была создана переменная `Top_3_City`, которая принимает значение 1, если хотя бы один из городов (выходной или назначенный) входит в топ-3. - Если город присутствует в обоих столбцах (и вылет, и назначение), переменная также будет равна 1. - В противном случае, если город не входит в топ-3, переменная получает значение 0.

Таким образом, добавленная дамми-переменная позволяет быстро идентифицировать, связаны ли рейсы с наиболее популярными направлениями.

Анализ количества отстановок

Теперь проделаем схожую процедуру с количеством отстановок. В анализе данных о рейсах переменная "Stops"указывает на количество пересадок, которые делает рейс. Эта переменная может принимать одно из трех значений: 'zero' (без пересадок), 'one' (одна пересадка) и 'two_or_more' (две или более пересадки). Однако для упрощения анализа и создания бинарной классификации было решено создать дамми-переменную Has_Stops, которая будет указывать, имеется ли пересадка в рейсе или нет.

Новая переменная Has_Stops принимает значение 1, если рейс включает хотя бы одну пересадку (то есть если значение в столбце "Stops"равно 'one' или 'two_or_more'), и 0, если рейс не имеет пересадок (значение 'zero'). Это упрощает дальнейший анализ, позволяя выделить рейсы с пересадками и без них и исследовать их влияние на цену билета и другие характеристики рейса.

Оценка модели

Основные метрики модели

Коэффициент детерминации (R-squared) составляет 0.91, что указывает на высокую объяснительную способность модели. Также, скорректированный R-squared равен 0.91. Лучше смотреть на скорректированный R-squared. Мы используем логарифмическую линейную модель (log-linear model).

Оценка значимости переменных

- **Нулевая гипотеза** (H0): Переменная не имеет статистически значимого влияния на зависимую переменную.
- **Альтернативная гипотеза** (H1): Переменная имеет статистически значимое влияние на зависимую переменную.

Для оценки значимости переменных мы использовали p-значения. Если p-value меньше 0.05, переменная считается статистически значимой.

$$y_i = \beta_0 + \beta_1 \cdot \text{duration}_i + \beta_2 \cdot \text{days\_left}_i + \cdots + \beta_k \cdot X_i + \epsilon_i$$

Где: - y_i — логарифм цены билета, - β_0 — константа, - $\beta_1, \ldots, \beta_k$ — коэффициенты переменных, - X_i — значения независимых переменных, - ϵ_i — ошибка.

Для проверки значимости коэффициентов используется t-статистика, которая вычисляется как:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

где:

- $\hat{\beta}_j$ — оценка коэффициента,
- $SE(\hat{\beta}_j)$ — стандартная ошибка коэффициента.

Соответствующий p-value для каждой переменной можно вычислить через распределение t-статистики. Если p-value < 0.05, то переменная считается статистически значимой на уровне 5%.

Результаты оценки переменных

- **Has_Stops (Наличие остановок)**: Коэффициент $\hat{\beta}_{\text{Has\_Stops}} = 0.460$ (стандартная ошибка: 0.003). Это означает, что наличие хотя бы одной пересадки увеличивает цену билета на 46%. Переменная имеет значимость, так как p-value < 0.05.
- **Top_3_City (Город в топ-3)**: Коэффициент $\hat{\beta}_{\text{Top\_3\_City}} = -0.022$ (стандартная ошибка: 0.002). Это указывает на то, что города, входящие в топ-3 по количеству рейсов, имеют несколько более низкие цены на билеты. Переменная статистически значима с p-value < 0.05.
- **airline_Air_India (Авиакомпания Air India)**: Коэффициент $\hat{\beta}_{\text{airline\_Air\_India}} = 0.474$ (стандартная ошибка: 0.003). Рейсы авиакомпании Air India имеют более высокие цены, чем базовая авиакомпания, на 47%. Переменная имеет высокую значимость (p-value < 0.05).

	Dependent variable: log_price
	(1)
Has_Stops	0.460***
	(0.003)
Top_3_City	-0.022***
	(0.002)
airline_Air_India	0.474***
	(0.003)
airline_GO_FIRST	0.379***
	(0.004)
airline_Indigo	0.285***
	(0.004)
airline_SpiceJet	0.449***
	(0.005)
airline_Vistara	0.600***
	(0.003)
arrival_time_Early_Morning	-0.079***
	(0.004)
arrival_time_Evening	0.048***
	(0.002)
arrival_time_Late_Night	0.013***
	(0.004)
arrival_time_Morning	-0.009***
	(0.003)
arrival_time_Night	0.030***
	(0.002)
class_Economy	-2.015***
	(0.002)
const	10.158***
	(0.005)
days_left	-0.014***
	(0.000)
departure_time_Early_Morning	0.016***
	(0.002)
departure_time_Evening	-0.009***
	(0.002)
departure_time_Late_Night	0.059***
	(0.011)
departure_time_Morning	0.037***
	(0.002)
departure_time_Night	-0.021***
	(0.003)
duration	0.005***
	(0.000)
Observations	240122
R^2	0.908
Adjusted R^2	0.908
Residual Std. Error	0.338 (df=240101)
F Statistic	118231.208*** (df=20; 240101)
Note:	*p<0.1; **p<0.05; ***p<0.01

- **airline_GO_FIRST (Авиакомпания GO FIRST)**: Коэффициент $\hat{\beta}_{\text{airline\_GO\_FIRST}} = 0.379$ (стандартная ошибка: 0.004). Рейсы авиакомпании GO FIRST имеют более высокие цены на 37.9%. Переменная статистически значима с p-value < 0.05.

- **airline_Indigo (Авиакомпания Indigo)**: Коэффициент $\hat{\beta}_{\text{airline\_Indigo}} = 0.285$ (стандартная ошибка: 0.004). Рейсы авиакомпании Indigo стоят на 28.5% больше по сравнению с базовой авиакомпанией. Статистическая значимость подтверждается p-value < 0.05.

- **airline_SpiceJet (Авиакомпания SpiceJet)**: Коэффициент $\hat{\beta}_{\text{airline\_SpiceJet}} = 0.449$ (стандартная ошибка: 0.005). Переменная указывает на то, что рейсы авиакомпании SpiceJet стоят на 44.9% дороже. Это значимо с p-value < 0.05.

- **airline_Vistara (Авиакомпания Vistara)**: Коэффициент $\hat{\beta}_{\text{airline\_Vistara}} = 0.600$ (стандартная ошибка: 0.003). Рейсы Vistara стоят на 60% дороже, чем базовая авиакомпания. Значимость переменной подтверждается p-value < 0.05.

- **arrival_time_Early_Morning (Время прибытия — раннее утро)**: Коэффициент $\hat{\beta}_{\text{arrival\_time\_Ear}}$ -0.079 (стандартная ошибка: 0.004). Это указывает на снижение цены билета на 7.9% для рейсов, прибывающих ранним утром. Переменная статистически значима.

- **arrival_time_Evening (Время прибытия — вечер)**: Коэффициент $\hat{\beta}_{\text{arrival\_time\_Evening}} = 0.048$ (стандартная ошибка: 0.002). Для рейсов, прибывающих вечером, цена увеличивается на 4.8%. Переменная значима.

- **arrival_time_Late_Night (Время прибытия — поздняя ночь)**: Коэффициент $\hat{\beta}_{\text{arrival\_time\_Late\_N}}$ 0.013 (стандартная ошибка: 0.004). Небольшое увеличение цены на 1.3% для рейсов, прибывающих поздно ночью. Переменная значима.

- **arrival_time_Morning (Время прибытия — утро)**: Коэффициент $\hat{\beta}_{\text{arrival\_time\_Morning}} = -0.009$ (стандартная ошибка: 0.003). Цена для рейсов, прибывающих утром, уменьшается на 0.9%. Значимость переменной подтверждается.

- **arrival_time_Night (Время прибытия — ночь)**: Коэффициент $\hat{\beta}_{\text{arrival\_time\_Night}} = 0.030$ (стандартная ошибка: 0.002). Для рейсов, прибывающих ночью, цена увеличивается на 3%. Значимость переменной подтверждается.

- **class_Economy (Класс — эконом)**: Коэффициент $\hat{\beta}_{\text{class\_Economy}} = -2.015$ (стандартная ошибка: 0.002). Эконом-класс стоит на 201.5% дешевле по сравнению с другими классами. Переменная статистически значима.

- **const (Константа)**: Коэффициент $\hat{\beta}_{\text{const}} = 10.158$ (стандартная ошибка: 0.005). Константа является важной переменной, так как она задает базовый уровень цены билета. Статистически значима.

- **days_left (Оставшиеся дни)**: Коэффициент $\hat{\beta}_{\text{days\_left}} = -0.0145$ (стандартная ошибка: 0.000). Увеличение дней до вылета снижает цену на 1.45% для каждого дополнительного дня. Переменная значима.

- **departure_time_Early_Morning (Время вылета — раннее утро)**: Коэффициент $\hat{\beta}_{\text{departure\_time\_}}$ 0.016 (стандартная ошибка: 0.002). Цена увеличивается на 1.6% для рейсов, вылетающих рано утром. Статистическая значимость.

- **departure_time_Evening (Время вылета — вечер)**: Коэффициент $\hat{\beta}_{\text{departure\_time\_Evening}} = -0.009$ (стандартная ошибка: 0.002). Для рейсов, вылетающих вечером, цена снижается на 0.9%. Значимость.

- **departure_time_Late_Night (Время вылета — поздняя ночь)**: Коэффициент $\hat{\beta}_{\text{departure\_time\_Lat}}$ 0.059 (стандартная ошибка: 0.011). Цена увеличивается на 5.9% для рейсов, вылетающих поздней ночью. Переменная значима.

- **departure_time_Morning (Время вылета — утро)**: Коэффициент $\hat{\beta}_{\text{departure\_time\_Morning}} = 0.037$ (стандартная ошибка: 0.002). Цена увеличивается на 3.7% для рейсов, вылетающих утром. Значимость.

- **departure_time_Night (Время вылета — ночь)**: Коэффициент $\hat{\beta}_{\text{departure\_time\_Night}} = -0.021$ (стандартная ошибка: 0.003). Для рейсов, вылетающих ночью, цена снижается на 2.1%. Статистическая значимость.

- **duration (Продолжительность рейса)**: Коэффициент $\hat{\beta}_{\text{duration}} = 0.005$ (стандартная ошибка: 0.000). Увеличение продолжительности рейса на 1 час увеличивает цену на 0.5%. Значимость.

Выводы

На основе коэффициентов линейной регрессии можно выделить переменные, которые оказывают наибольшее влияние на цену билета:

- **airline_Vistara**: $\hat{\beta}_7 = 0.600$, что означает, что рейсы авиакомпании Vistara имеют на 60% более высокие цены по сравнению с базовой авиакомпанией. Это переменная с самым высоким коэффициентом, что указывает на её сильное влияние на цену.
- **Has_Stops**: $\hat{\beta}_3 = 0.460$, что говорит о том, что рейсы с пересадками имеют на 46% более высокие цены по сравнению с прямыми рейсами.
- **airline_Air_India**: $\hat{\beta}_4 = 0.474$, что также означает, что рейсы авиакомпании Air India имеют на 47.4% более высокие цены по сравнению с базовой авиакомпанией.

Эти переменные оказывают наибольшее влияние на цену билетов и должны учитываться при прогнозировании стоимости рейсов.

Оценка гипотез

1. Гипотеза об отношении времени бронирования к цене билета

Гипотеза: Чем больше дней остается до даты вылета (переменная "Days Left"), тем ниже будет цена билета (переменная "Price").

Результаты модели: Переменная `days_left` имеет отрицательное влияние на логарифм цены с коэффициентом $\hat{\beta}_{\text{days\_left}} = -0.0145$ и p-value = 0.00.

Вывод: Гипотеза подтверждается. Чем больше дней остается до вылета, тем ниже цена на билет. Это логично, поскольку авиакомпании обычно предлагают скидки на билеты, которые бронируются заранее, чтобы стимулировать спрос и заполнить рейсы.

2. Гипотеза о рейсах с пересадками

Гипотеза: Рейсы с пересадками будут иметь более высокую цену по сравнению с прямыми рейсами. Мы предполагаем, что рейсы с одной или более остановками будут стоить дороже из-за дополнительных затрат на обслуживание, времени в пути и неудобств для пассажиров.

Результаты модели: Переменная `Has_Stops` имеет положительный коэффициент $\hat{\beta}_{\text{Has\_Stops}} = 0.460$ и p-value = 0.00.

Вывод: Гипотеза подтверждается. Рейсы с пересадками имеют на 46% более высокие цены по сравнению с прямыми рейсами. Это объясняется дополнительными затратами на обслуживание, возможными неудобствами для пассажиров и увеличенным временем в пути.

3. Гипотеза о влиянии популярных направлений на стоимость билетов

Гипотеза: Рейсы между крупными городами или популярными туристическими направлениями (переменные "Source City"и "Destination City") стоят дороже.

Результаты модели: Для переменной `Top_3_City` (город в топ-3) был получен коэффициент $\hat{\beta}_{\text{Top\_3\_City}} = -0.0222$ с p-value = 0.00.

Вывод: Эта гипотеза кажется частично опровергнутой. Хотя в некоторых случаях можно ожидать, что конкуренция в крупных городах и на популярных маршрутах может снизить цену, наши данные показывают, что города в топ-3 (по количеству рейсов) имеют на 2.22% более низкие

цены. Это может быть связано с большей конкуренцией между авиакомпаниями, которая приводит к снижению цен на популярных маршрутах.

Также вероятнее всего мы столкнулись с переобучением, модель слишком точно подгоняется под обучающие данные, включая случайный шум, что может привести к плохой обобщаемости на новых данных. Причина может заключаться в том, что мы использовали слишком много признаков, можно попробовать это исправить, оставив только наиболее важные признаки. Можем попробовать применить метод "корень из r"для фильтрации переменных, в итоге это не сильно повлияло на корректированный R-squared, оставив только значимые признаки, и позволило избежать переобучения.

- $R^2 = 0.826$: Модель объясняет 82.6% вариации логарифма цены, что свидетельствует о хорошей объяснительной способности модели.
- Фактор *class_Economy*: Коэффициент $\hat{\beta} = -2.1834$ показывает, что билеты в эконом-классе дешевле на 2.18 логарифмических единиц по сравнению с другими классами.
- Тесты на значимость: Все переменные с t-значениями выше порога показали статистическую значимость (p-value = 0.000).

	Dependent variable: log_price
	(1)
class_Economy	-2.183***
	(0.002)
const	10.834***
	(0.002)
Observations	300153
R^2	0.826
Adjusted R^2	0.826
Residual Std. Error	0.465 (df=300151)
F Statistic	1420906.942*** (df=1; 300151)
Note:	*p<0.1; **p<0.05; ***p<0.01

Список литературы

[1] Borenstein, S., Rose, N. L. (1994). *Competition and price dispersion in the U.S. airline industry.* Journal of Political Economy, 102(4), 653-683. Link

[2] Borenstein, S. (1992). *Airline mergers, airport dominance, and market power.* American Economic Review, 82(2), 400-404. Link

[3] Morrell, P. (2008). *The economics of the airline industry.* Ashgate Publishing. Link

[4] Zohar, D. (2002). *Airline pricing and revenue management.* Routledge. Link