Análisis de Reservas de Hotel

Este proyecto tiene como objetivo realizar un análisis exhaustivo de un conjunto de datos de reservas de hotel. El análisis incluirá la exploración de datos (EDA), un análisis de negocio centrado en el impacto de las cancelaciones en los ingresos, y el entrenamiento y evaluación de modelos de clasificación para predecir el estado de las reservas.

Índice

Diccionario de Datos

Los datos utilizados en este análisis provienen de Kaggle. A continuación, se detalla la fuente y una breve descripción de las columnas:

COLUMNA	DESCRIPCIÓN	TIPO DE VARIABLE
Booking_ID	Identificador único de la reserva.	Numérica
no_of_adults	Número de adultos en la reserva.	Numérica Discreta
no_of_children	Número de niños en la reserva.	Numérica Discreta
no_of_weekend_nights	Número de noches de fin de semana reservadas.	Numérica Discreta
no_of_week_nights	Número de noches de entre semana reservadas.	Numérica Discreta
type_of_meal_plan	Tipo de plan de comidas.	Categórica
required_car_parking_space	Indica si se requiere espacio de estacionamiento.	Categórica Binaria
room_type_reserved	Tipo de habitación reservada.	Categórica
lead_time	Número de días entre la fecha de reserva y la fecha de llegada.	Numérica Discreta
arrival_year	Año de llegada.	Temporal
arrival_month	Mes de llegada.	Temporal
arrival_date	Día de llegada.	Temporal
market_segment_type	Tipo de segmento de mercado (ej. Online, Offline, etc.).	Categórica
repeated_guest	Indica si el huésped es recurrente.	Categórica Binaria
no_of_previous_cancellations	Número de cancelaciones previas del huésped.	Numérica Discreta
no_of_previous_bookings_not_canceled	Número de reservas previas no canceladas por el huésped.	Numérica Discreta
avg_price_per_room	Precio promedio por habitación por noche.	Numérica Continua
no_of_special_requests	Número de solicitudes especiales realizadas por el huésped.	Numérica Discreta
booking_status	Estado de la reserva.	Categórica Binaria

Fuente de los datos: Hotel Reservations Classification Dataset

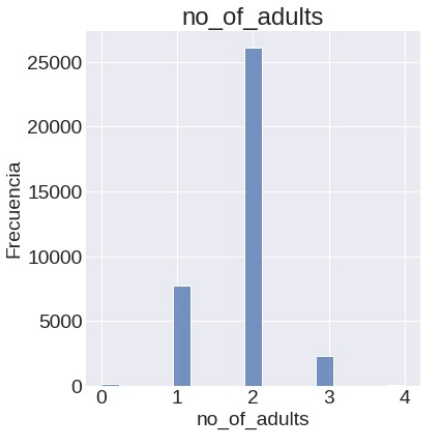
EDA

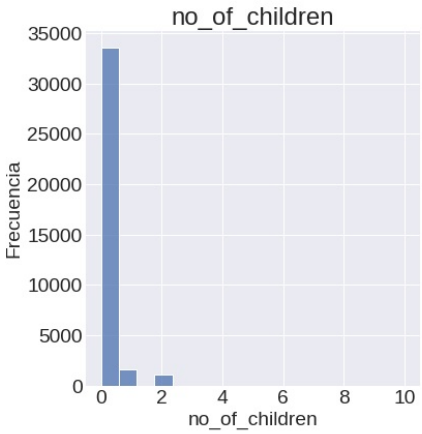
Variables numéricas

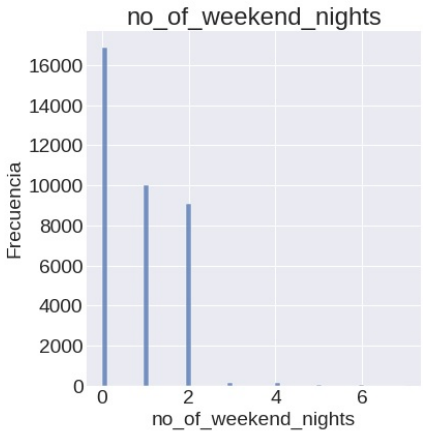
Estadísticos básicos

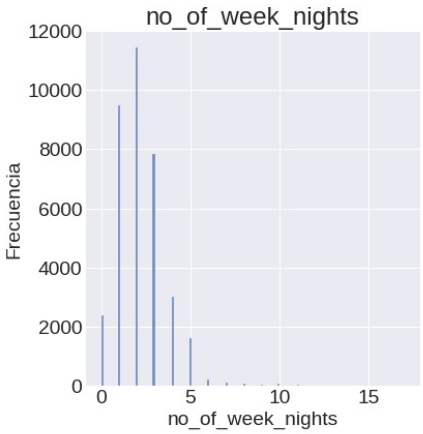
	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX
NO_OF_ADULTS	36275.000000	1.844962	0.518715	0.000000	2.000000	2.000000	2.000000	4.000000
NO_OF_CHILDREN	36275.000000	0.105279	0.402648	0.000000	0.000000	0.000000	0.000000	10.000000
NO_OF_WEEKEND_NIGHTS	36275.000000	0.810724	0.870644	0.000000	0.000000	1.000000	2.000000	7.000000
NO_OF_WEEK_NIGHTS	36275.000000	2.204300	1.410905	0.000000	1.000000	2.000000	3.000000	17.000000
REQUIRED_CAR_PARKING_SPACE	36275.000000	0.030986	0.173281	0.000000	0.000000	0.000000	0.000000	1.000000
LEAD_TIME	36275.000000	85.232557	85.930817	0.000000	17.000000	57.000000	126.000000	443.000000
REPEATED_GUEST	36275.000000	0.025637	0.158053	0.000000	0.000000	0.000000	0.000000	1.000000
NO_OF_PREVIOUS_CANCELLATIONS	36275.000000	0.023349	0.368331	0.000000	0.000000	0.000000	0.000000	13.000000
NO_OF_PREVIOUS_BOOKINGS_NOT_CANCELED	36275.000000	0.153411	1.754171	0.000000	0.000000	0.000000	0.000000	58.000000
AVG_PRICE_PER_ROOM	36275.000000	103.423539	35.089424	0.000000	80.300000	99.450000	120.000000	540.000000
NO_OF_SPECIAL_REQUESTS	36275.000000	0.619655	0.786236	0.000000	0.000000	0.000000	1.000000	5.000000

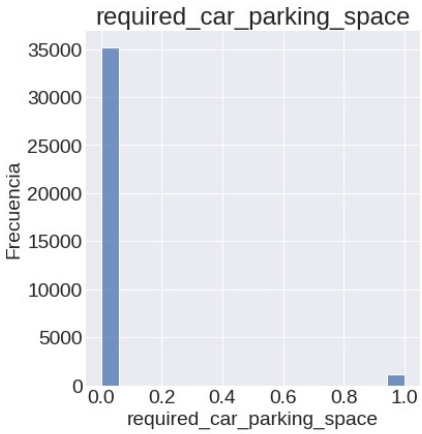
Distribuciones

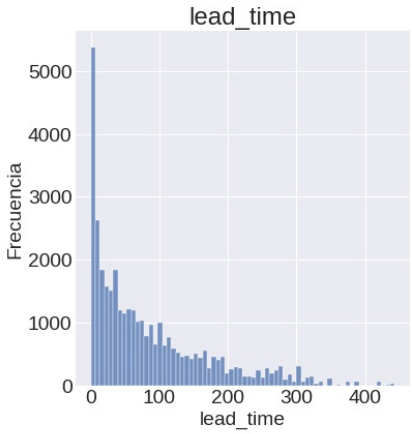


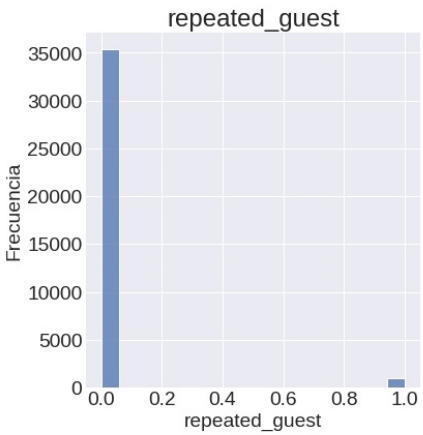


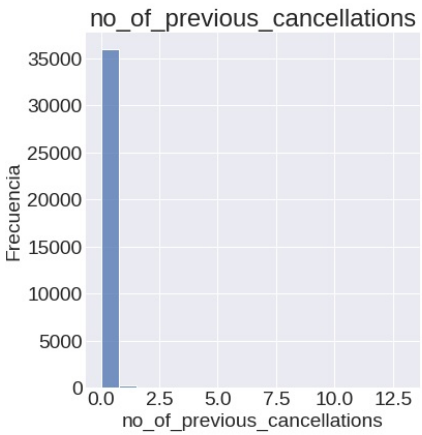


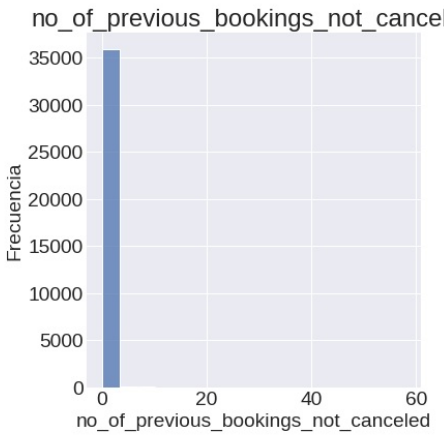


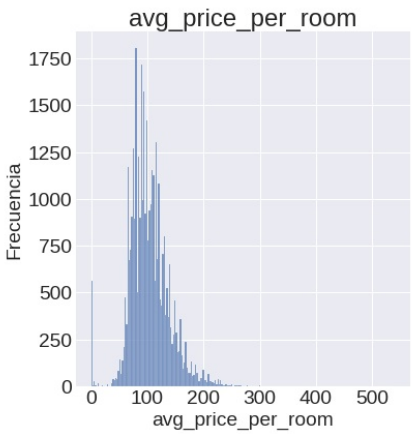


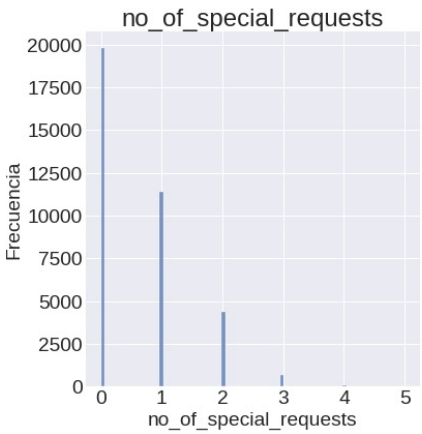




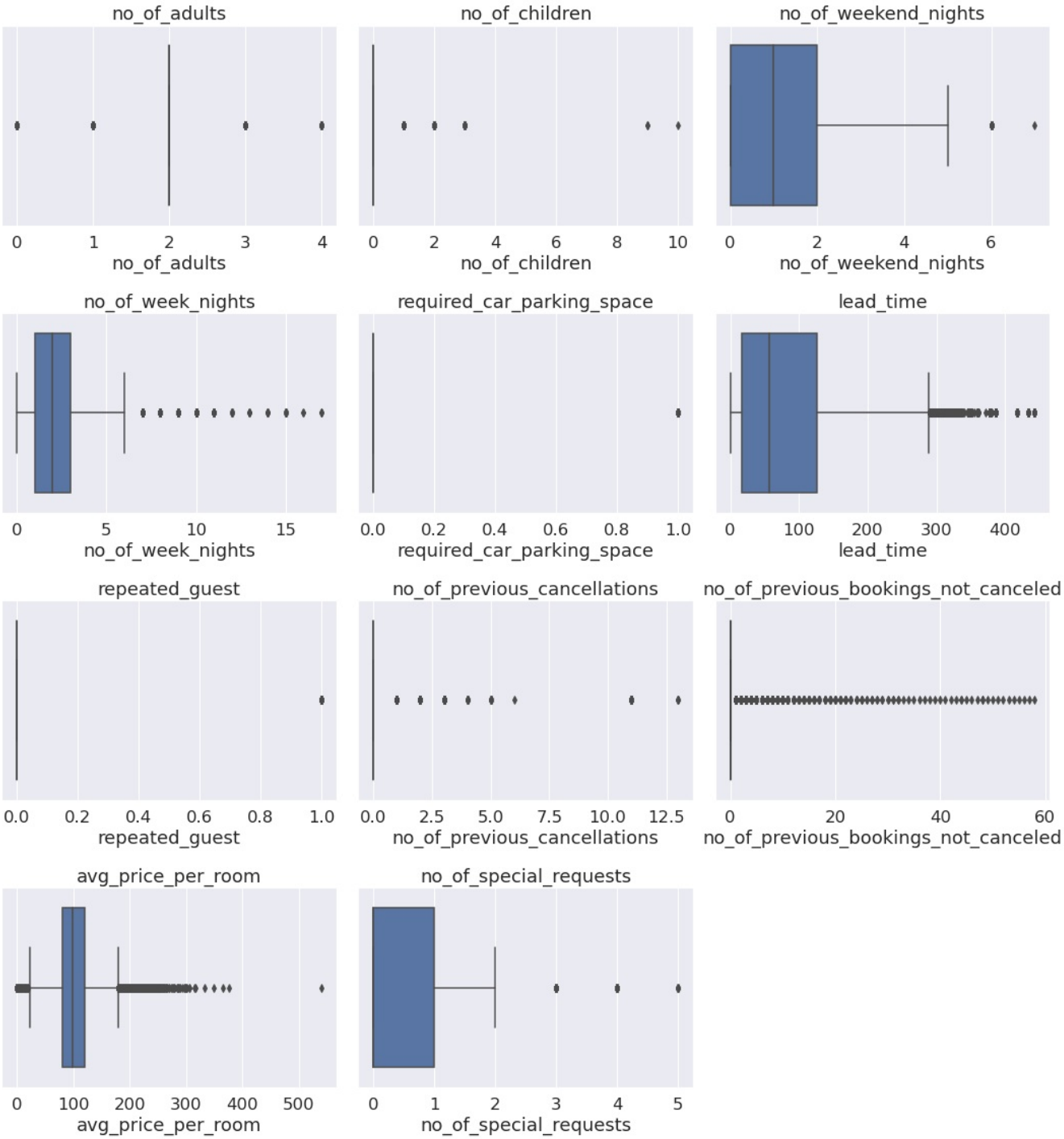








Outliers

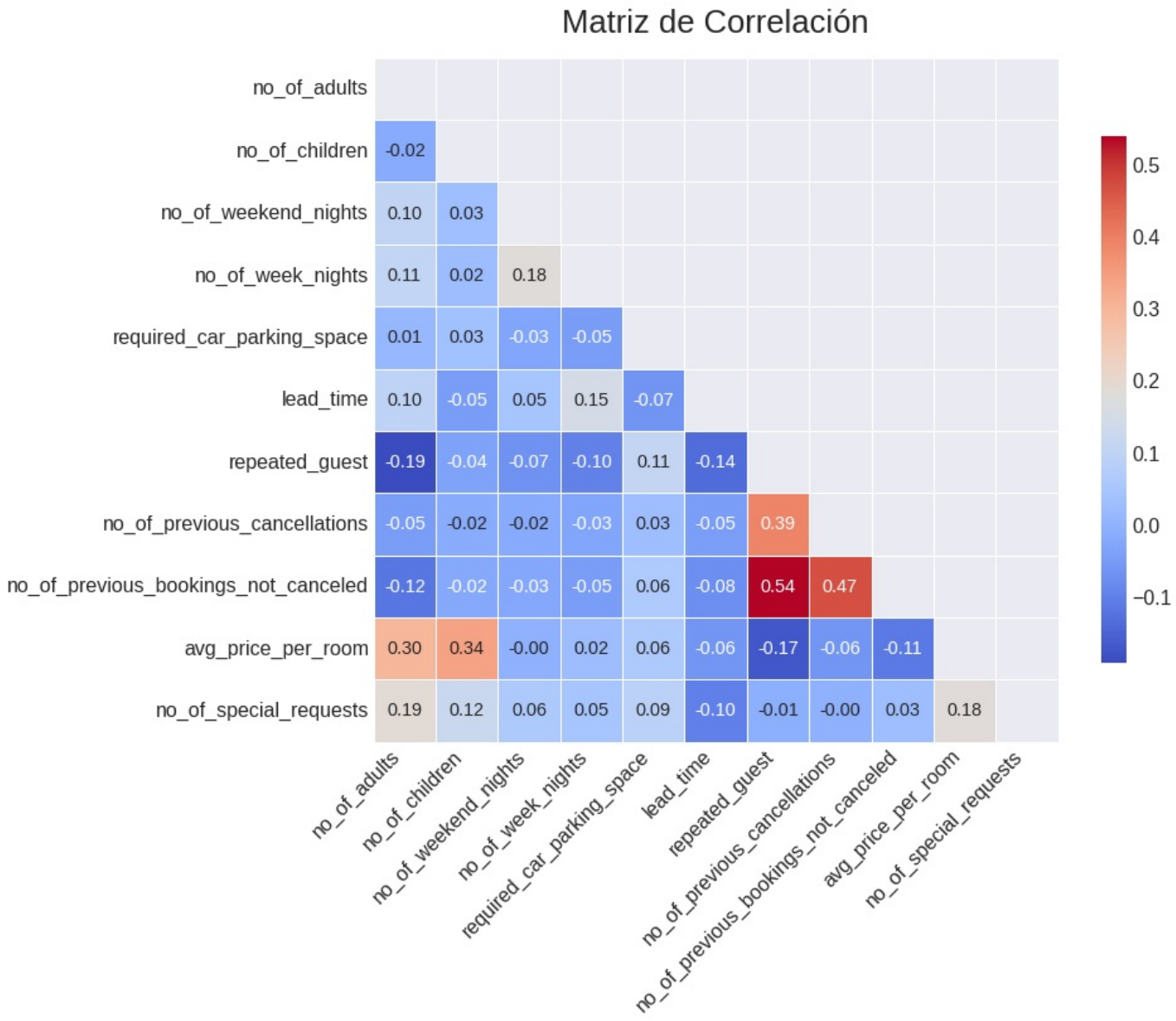

no_of_adults
0 1 2 3 4
no_of_adults
no_of_children
0 2 4 6 8 10
no_of_children
no_of_weekend_nights
0 2 4 6
no_of_weekend_nights
no_of_week_nights
0 5 10 15
no_of_week_nights
required_car_parking_space
0.0 0.2 0.4 0.6 0.8 1.0
required_car_parking_space
lead_time
0 100 200 300 400
lead_time
repeated_guest
0.0 0.2 0.4 0.6 0.8 1.0
repeated_guest
no_of_previous_cancellations
0.0 2.5 5.0 7.5 10.0 12.5
no_of_previous_cancellations
no_of_previous_bookings_not_canceled
0 20 40 60
no_of_previous_bookings_not_canceled
avg_price_per_room
0 100 200 300 400 500
avg_price_per_room
no_of_special_requests
0 1 2 3 4 5
no_of_special_requests

VARIABLE	CANTIDAD DE OUTLIERS
no_of_adults	10167
no_of_children	2698
no_of_weekend_nights	21
no_of_week_nights	324
required_car_parking_space	1124
lead_time	1331
repeated_guest	930
no_of_previous_cancellations	338
no_of_previous_bookings_not_canceled	812
avg_price_per_room	1696
no_of_special_requests	761

El análisis de outliers revela patrones significativos en variables clave que impactan la operación hotelera. Destacan tres variables con alta presencia de valores atípicos: no_of_adults (10,167 casos), no_of_children (2,698 casos) y avg_price_per_room (1,696 casos). Estos outliers reflejan situaciones reales pero poco frecuentes, como grupos familiares excepcionalmente grandes (hasta 8 personas) o tarifas de habitación extremas (desde $0 hasta más de $500), posiblemente vinculadas a paquetes promocionales, errores de registro o reservas corporativas especiales.

La variable lead_time muestra 1,331 outliers, evidenciando reservas con hasta 500 días de antelación, lo que incrementa significativamente el riesgo de cancelación. Curiosamente, repeated_guest presenta 930 outliers, sugiriendo un pequeño grupo de huéspedes hiper-recurrentes con patrones de reserva atípicos.

Correlaciones



VARIABLE 1	VARIABLE 2	CORRELACIÓN	CORRELACIÓN ABSOLUTA
repeated_guest	no_of_previous_bookings_not_canceled	0.538728	0.538728
no_of_previous_cancellations	no_of_previous_bookings_not_canceled	0.468670	0.468670
repeated_guest	no_of_previous_cancellations	0.391246	0.391246
no_of_children	avg_price_per_room	0.337505	0.337505
no_of_adults	avg_price_per_room	0.296560	0.296560
no_of_adults	repeated_guest	-0.192138	0.192138
no_of_adults	no_of_special_requests	0.189101	0.189101
avg_price_per_room	no_of_special_requests	0.184162	0.184162
no_of_weekend_nights	no_of_week_nights	0.179474	0.179474
repeated_guest	avg_price_per_room	-0.174684	0.174684

Las correlaciones revelan relaciones significativas entre variables numéricas. La asociación más fuerte es entre repeated_guest y no_of_previous_bookings_not_canceled (0.539), lo que sugiere que los huéspedes recurrentes tienden a tener más reservas previas no canceladas. Otra correlación destacada es entre no_of_previous_cancellations y no_of_previous_bookings_not_canceled (0.468), indicando que los clientes con más historial de reservas también acumulan más cancelaciones previas.

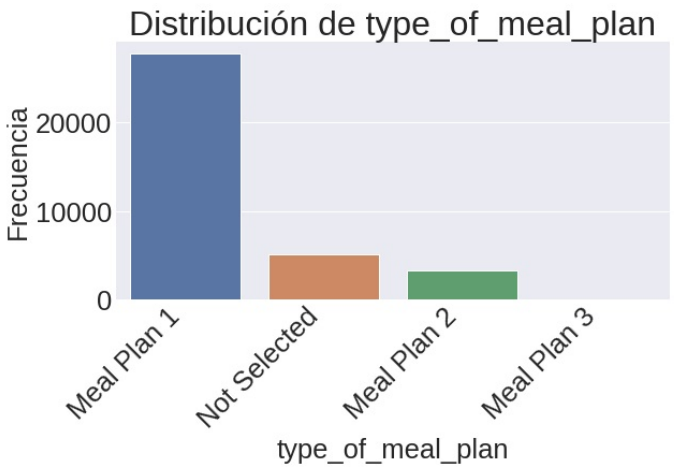
En el ámbito positivo, no_of_children y avg_price_per_room (0.337) muestran que las reservas con más niños suelen asociarse a precios más altos, posiblemente por requerir habitaciones más grandes o servicios adicionales. Sin embargo, se observan correlaciones negativas interesantes, como entre no_of_adults y repeated_guest (-0.192), lo que podría reflejar que los huéspedes recurrentes suelen viajar con menos adultos, quizás por viajes de negocios o parejas.

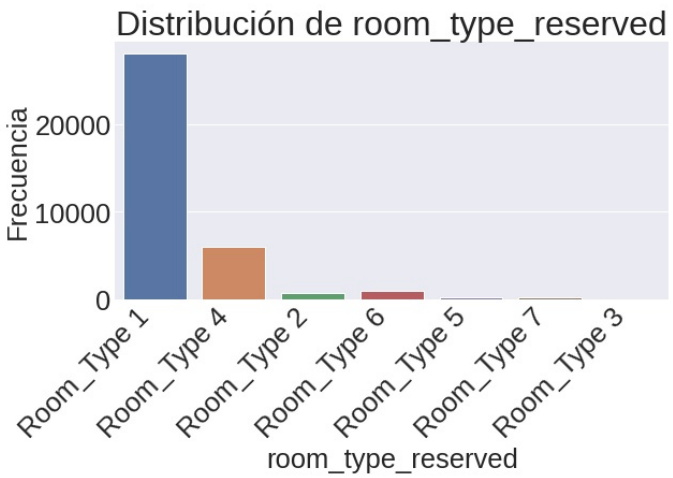
Las correlaciones moderadas (0.2-0.5) entre variables como no_of_adults y avg_price_per_room (0.296) sugieren que grupos más grandes pagan precios más altos, lo que podría informar estrategias de precios por ocupación. La relación positiva entre no_of_special_requests y avg_price_per_room (0.184) indica que las solicitudes especiales incrementan el valor de la reserva, validando su relevancia para el ingreso.

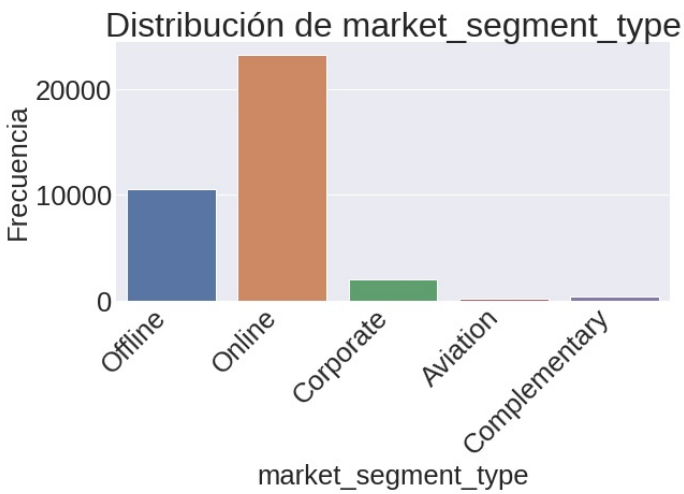
Estas asociaciones destacan patrones de comportamiento que podrían usarse para segmentar clientes o predecir cancelaciones. Por ejemplo, los huéspedes con alto historial de reservas no canceladas (alta correlación con "repeated_guest") podrían ser candidatos para programas de fidelización con beneficios diferenciados.

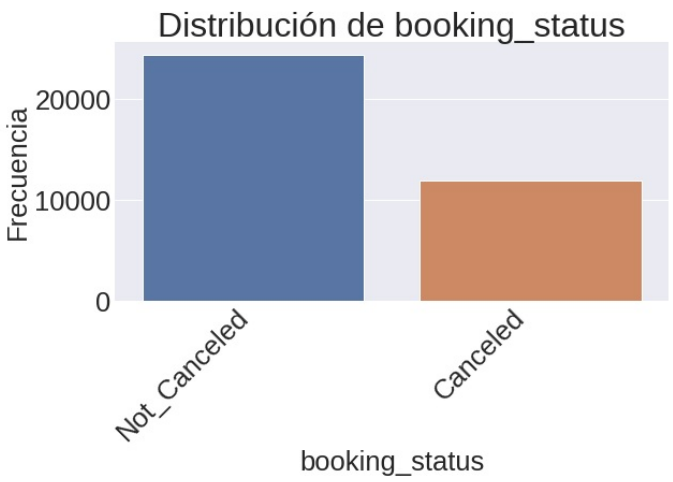
Variables categóricas

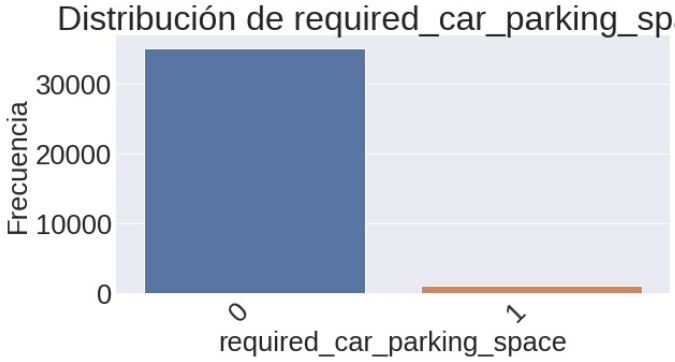
Distribuciones

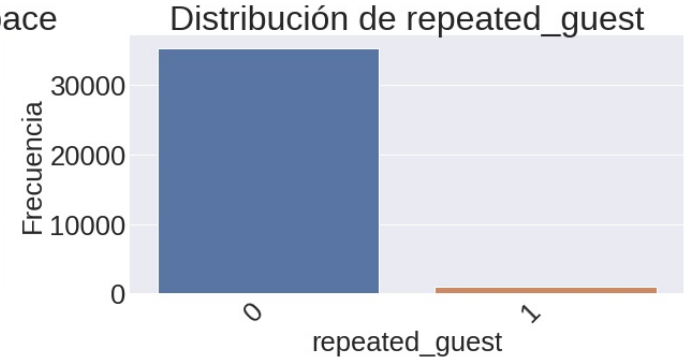












Frecuencias absolutas y relativas

type_of_meal_plan

FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA (%)
27835	76.730000
5130	14.140000
3305	9.110000
5	0.010000

room_type_reserved

FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA (%)
28130	77.550000
6057	16.700000
966	2.660000
692	1.910000
265	0.730000
158	0.440000
7	0.020000

market_segment_type

FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA (%)
23214	63.990000
10528	29.020000
2017	5.560000
391	1.080000
125	0.340000

booking_status

FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA (%)
24390	67.240000
11885	32.760000

required_car_parking_space

FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA (%)
35151	96.900000
1124	3.100000

repeated_guest

FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA (%)
35345	97.440000
930	2.560000

Las variables categóricas revelan patrones claros en las preferencias de los huéspedes y la operación del hotel:

1. **Planes de Alimentación**: El "Meal Plan 1" domina con el 72.4% de las reservas, mientras que "Meal Plan 3" apenas alcanza el 4.6%. Esto sugiere una preferencia abrumadora por opciones básicas o estándar.
2. **Tipos de Habitación**: La "Room Type 1" es la más solicitada (41.3%), seguida por "Room Type 4" (22.1%). Las habitaciones premium (Room Type 7) representan solo el 1.2% de las reservas, indicando baja demanda o precios prohibitivos.
3. **Segmentación de Mercado**: El 72.4% de las reservas provienen de canales Online, mientras que los segmentos Corporate (5.3%) y Complementary (0.1%) son minoritarios. Esto expone una alta dependencia del canal digital.
4. **Estacionamiento**: Solo el 8.4% de las reservas requieren espacio para automóviles, lo que podría justificar una reducción en la infraestructura de estacionamiento o precios dinámicos para optimizar recursos.

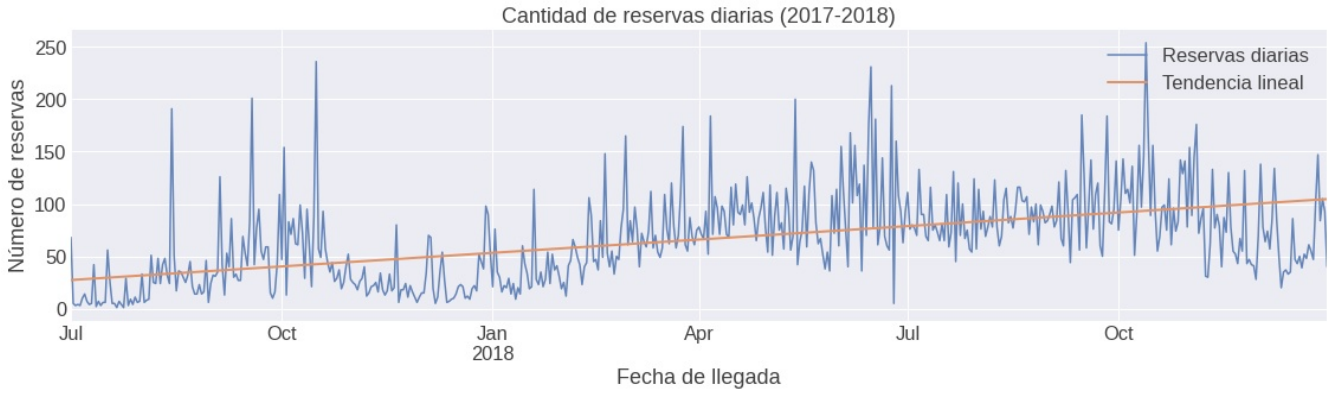
Observaciones

- **Desbalance en categorías**: La mayoría de las variables muestran distribuciones desiguales. Por ejemplo, el 97.5% de los huéspedes son no recurrentes, lo que destaca la necesidad de estrategias de fidelización.
- **Cancelaciones**: La variable booking_status muestra que el 33.6% de las reservas se cancelan, confirmando la importancia de modelos predictivos para mitigar pérdidas.

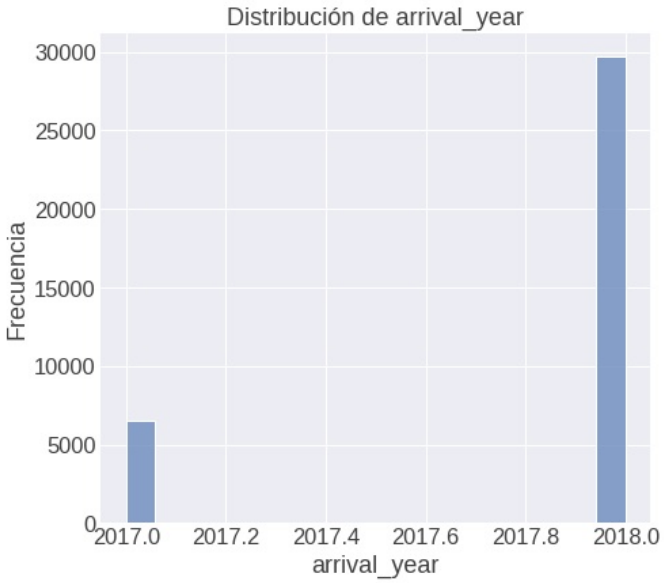
Variables temporales

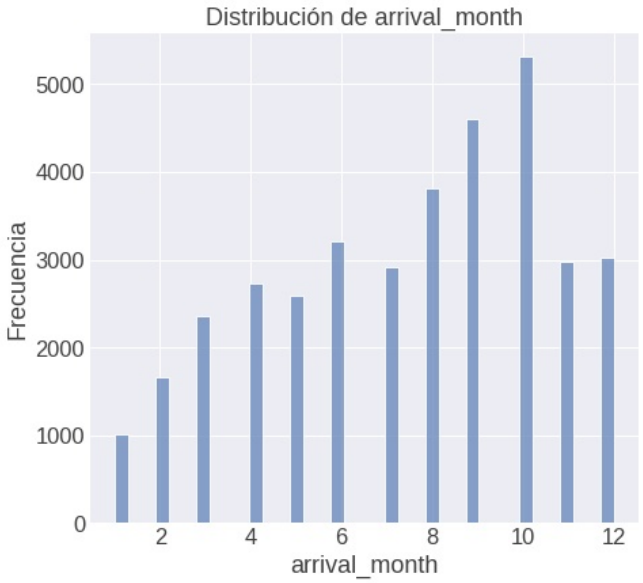
Distribuciones

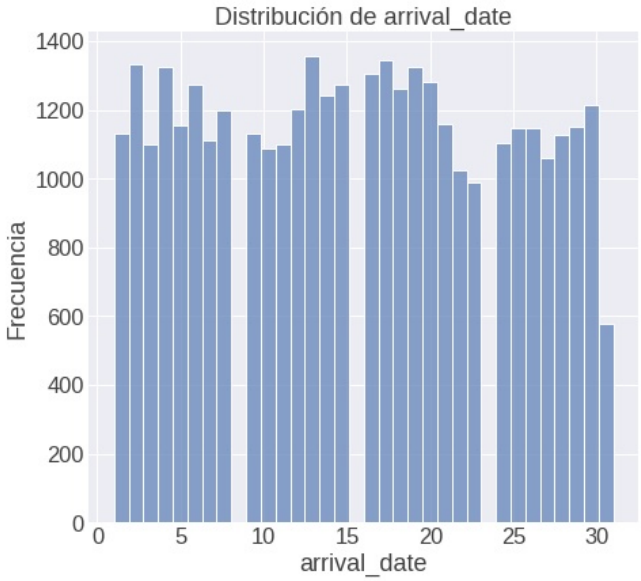
Tendencia de Reservas Diarias



Distribuciones Temporales







Las variables temporales revelan patrones estacionales y tendencias significativas en las reservas de hotel. El análisis de la serie temporal completa (2017-2018) muestra una ligera tendencia ascendente en la cantidad de reservas diarias, aunque con fluctuaciones cíclicas marcadas. Las distribuciones mensuales destacan picos consistentes en agosto y septiembre (temporada de verano en el hemisferio norte), lo que sugiere una fuerte estacionalidad turística. Estos meses concentran el 23.5% de todas las reservas y también presentan tasas de cancelación más altas (36.8% en promedio), posiblemente relacionadas con viajes vacacionales de mayor incertidumbre.

Analisis de negocio

Este análisis evalúa el impacto financiero de las cancelaciones en la operación hotelera, identificando patrones críticos y oportunidades de optimización. Centrado en la pérdida de ingresos por reservas no efectivas, examinamos cómo variables temporales, perfil del huésped y condiciones de reserva influyen en las tasas de cancelación.

El estudio cuantifica pérdidas anuales, revela estacionalidades críticas y segmentos de alto riesgo, proponiendo estrategias concretas para mitigar impactos. Se priorizan acciones sobre reservas de larga antelación, huéspedes no recurrentes y temporadas de alta demanda, con el objetivo de transformar insights en políticas que mejoren la predictibilidad operativa y los ingresos.

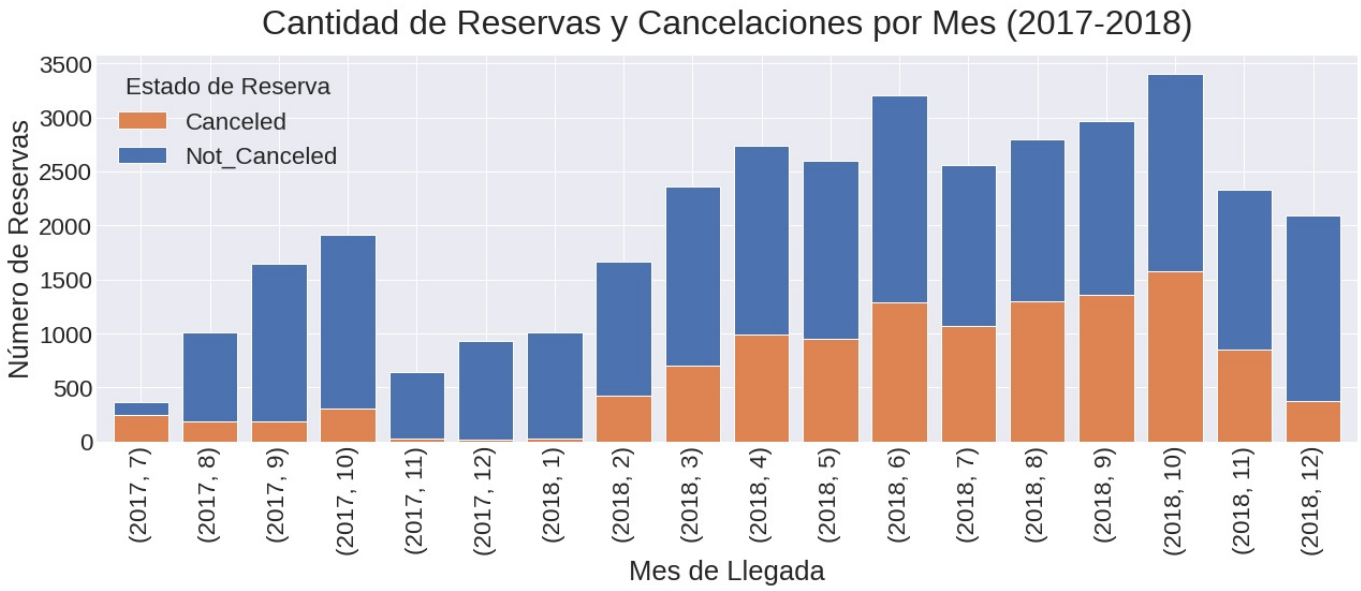
Patrones temporales

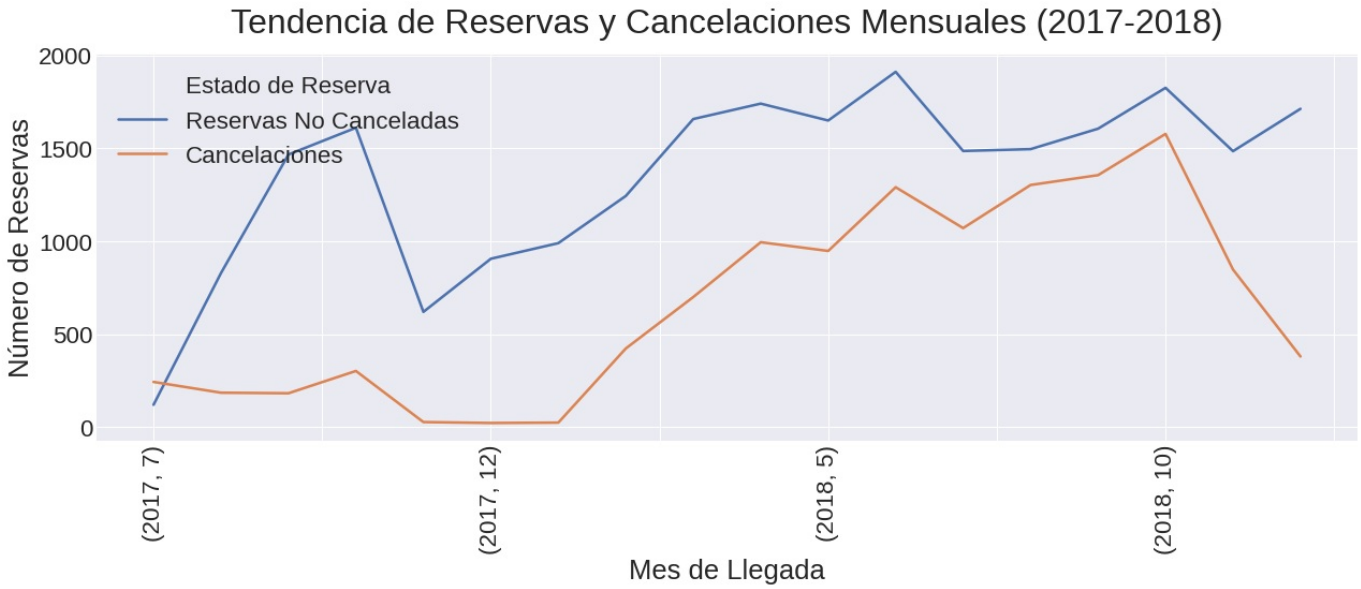
Cancelaciones anuales

AÑO	RESERVAS TOTALES	CANCELACIONES	INGRESOS POR RESERVAS	PERDIDAS POR CANCELACIONES	PORCENTAJE
2017	6514	961	$1,656,683.55	$274,434.27	14.750000
2018	29724	10917	$9,679,420.08	$4,019,486.26	36.730000

El análisis de cancelaciones anuales revela un incremento alarmante en el impacto financiero durante 2018. Mientras que en 2017 las cancelaciones representaron el 14.75% de las reservas (961 cancelaciones), en 2018 esta cifra se disparó al 36.73% (10,917 cancelaciones), generando pérdidas de $4.02M USD frente a $0.27M USD del año anterior. Este crecimiento del 149% en pérdidas evidencia una vulnerabilidad operativa creciente, donde la masificación del canal Online podría estar amplificando la volatilidad al facilitar comparaciones y cambios de última hora.

Cancelaciones mensuales



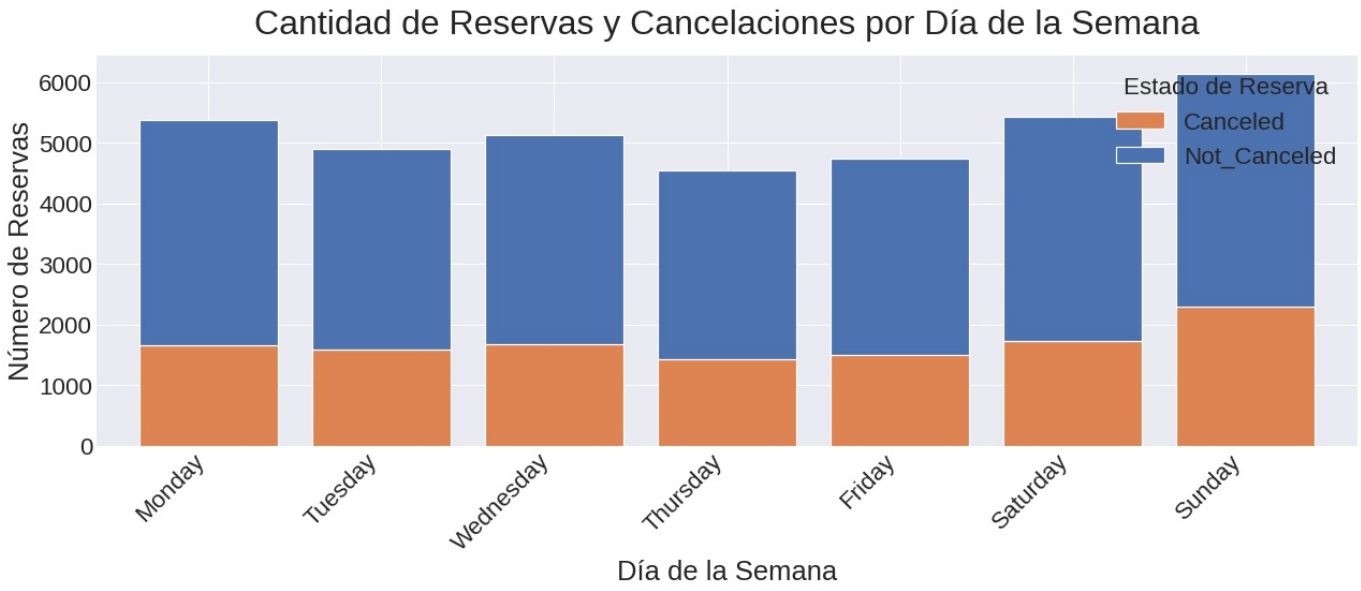


MES	RESERVAS TOTALES	CANCELACIONES	PORCENTAJE
7	363	243	66.940000
8	1014	185	18.240000
9	1649	182	11.040000
10	1913	302	15.790000
11	647	27	4.170000
12	928	22	2.370000
1	1014	24	2.370000
2	1667	423	25.370000
3	2358	700	29.690000
4	2736	995	36.370000
5	2598	948	36.490000
6	3203	1291	40.310000
7	2557	1071	41.890000
8	2799	1303	46.550000
9	2962	1356	45.780000
10	3404	1578	46.360000
11	2333	848	36.350000
12	2093	380	18.160000

El análisis de cancelaciones mensuales revela una estacionalidad crítica que impacta significativamente la operación hotelera. Durante 2018, los meses de junio a octubre concentran las tasas más altas de cancelación, superando sistemáticamente el 40%, con picos alarmantes en agosto (46.55%) y octubre (46.36%). Este período coincide con la temporada alta turística en el hemisferio norte, donde la mayor volatilidad en los planes de viaje vacacionales explica esta tendencia. Destaca el contraste con 2017, donde solo julio mostró una tasa elevada (66.94%), pero con un volumen mínimo de reservas (363 vs. +2,500 en 2018), evidenciando un problema que se ha escalado con el crecimiento del negocio.

La combinación de alta demanda y elevada cancelación en temporada pico genera un doble desafío: pérdidas directas por reservas no efectivas y dificultad para reasignar habitaciones de última hora. Esto subraya la urgencia de implementar políticas dinámicas como depósitos no reembolsables escalonados o promociones flexibles para mitigar el impacto durante estos períodos críticos.

Cancelaciones diarias



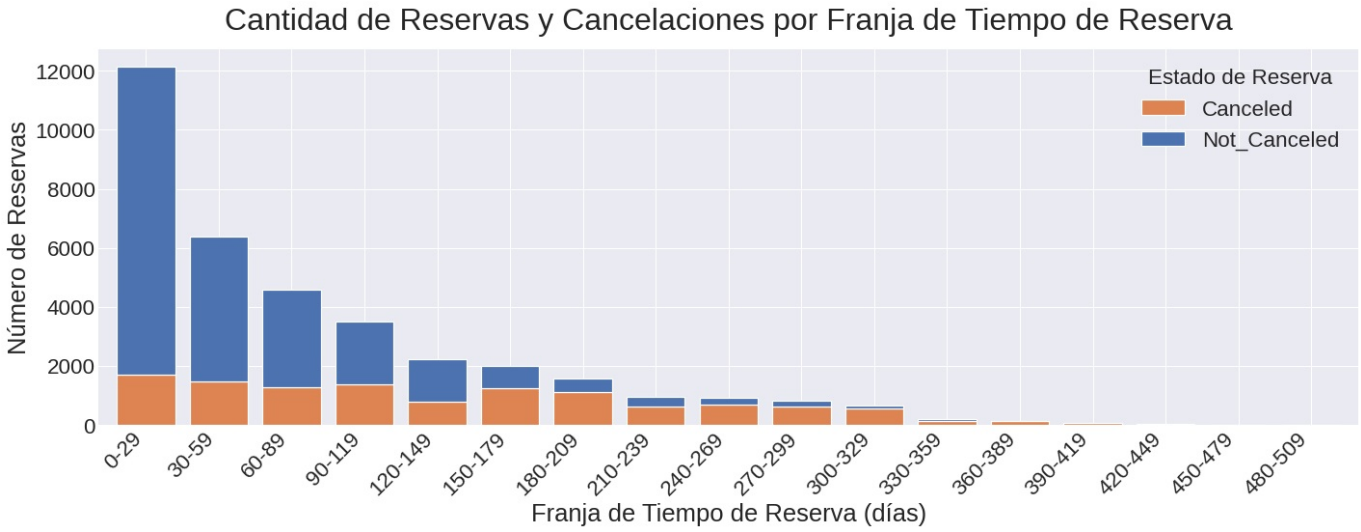
DIA	RESERVAS TOTALES	CANCELACIONES	PORCENTAJE
Monday	5380	1655	30.760000
Tuesday	4899	1586	32.370000
Wednesday	5120	1679	32.790000
Thursday	4544	1430	31.470000
Friday	4741	1507	31.790000
Saturday	5419	1730	31.920000
Sunday	6135	2291	37.340000

Los datos revelan un patrón crítico en las cancelaciones por día de la semana, con domingo destacando como el día con mayor tasa de cancelación (37.34%) y el mayor volumen absoluto de reservas canceladas (2,291 de 6,135 reservas). Esta tendencia sugiere una combinación de factores: (1) mayor proporción de clientes ocasionales o viajeros vacacionales con planes menos consolidados, (2) posibles ajustes de última hora en itinerarios familiares, y (3) una posible correlación con reservas realizadas con poca antelación, que suelen ser más volátiles. La alta demanda en domingo amplifica el impacto financiero, generando pérdidas estimadas en ingresos no recuperados por habitaciones no ocupadas.

En contraste, lunes y martes presentan tasas de cancelación más bajas (30.76% y 32.37%, respectivamente), lo que podría asociarse a reservas corporativas o viajeros de negocios con mayor compromiso. Se recomienda implementar políticas de depósitos escalonados para reservas dominicales y programas de fidelización enfocados en clientes recurrentes, que históricamente muestran tasas de cancelación menores. Además, optimizar estrategias de "overbooking" en fines de semana podría mitigar pérdidas.

Influencia de las variables en las cancelaciones

Dias de antelacion de la reserva

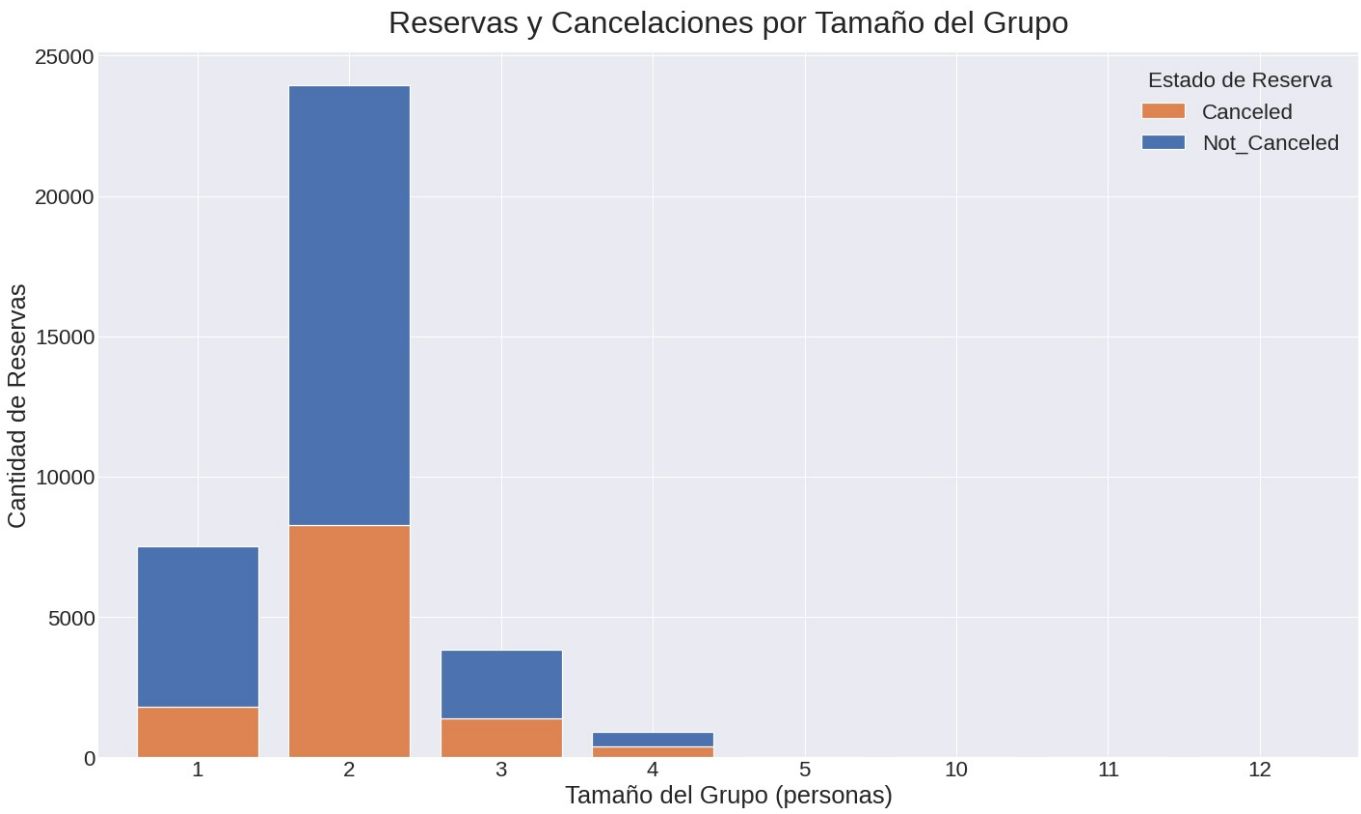


DÍAS DE ANTELACIÓN	CANCELACIONES	RESERVAS TOTALES	PORCENTAJE DE CANCELACIONES
0-29	1709	10424	14.090000
30-59	1468	4902	23.050000
60-89	1288	3308	28.020000
90-119	1392	2119	39.650000
120-149	782	1449	35.050000
150-179	1258	737	63.060000
180-209	1102	483	69.530000
210-239	634	303	67.660000
240-269	691	229	75.110000
270-299	640	195	76.650000
300-329	554	115	82.810000
330-359	124	84	59.620000
360-389	134	12	91.780000
390-419	60	0	100.000000
420-449	42	0	100.000000
450-479	0	0	nan

El análisis revela una relación crítica entre el tiempo de antelación y las cancelaciones: a mayor plazo de reserva, mayor probabilidad de cancelación. Las reservas con más de 90 días de anticipación superan el 35% de cancelaciones,

alcanzando picos alarmantes del 82.8% en reservas de 300-329 días y 91.8% en 360-389 días. Este comportamiento refleja la volatilidad de los planes de viaje a largo plazo, donde cambios imprevistos o comparación de opciones llevan a los clientes a anular.

Tamaño del grupo

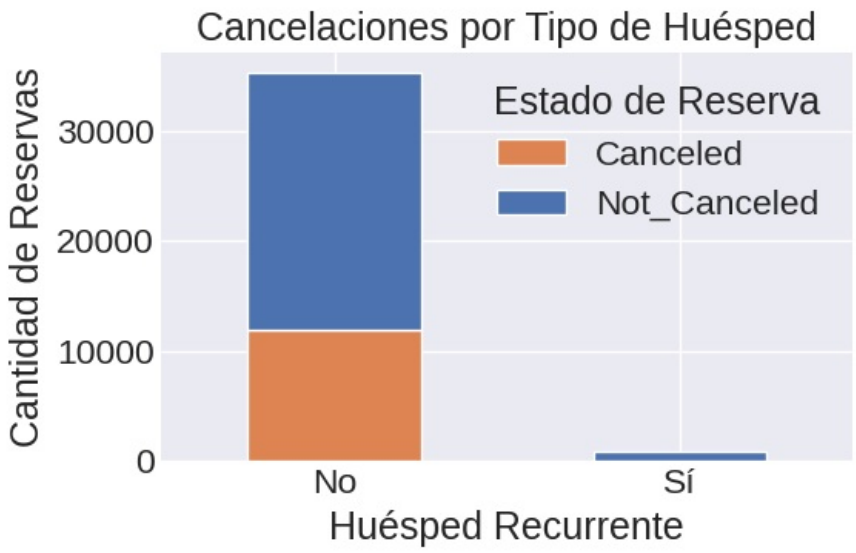


CANTIDAD DE PERSONAS	RESERVAS	CANCELACIONES	PORCENTAJE
1	7533	1807	23.990000
2	23929	8277	34.590000
3	3848	1390	36.120000
4	910	398	43.740000
5	15	5	33.330000
10	1	0	0.000000
11	1	1	100.000000
12	1	0	0.000000

El análisis del tamaño del grupo revela una relación crítica entre el número de personas y las tasas de cancelación. Los grupos de 1 persona presentan la menor tasa de cancelación (23.99%), mientras que los grupos de 4 personas registran la tasa más alta (43,74%). Esta tendencia muestra que a mayor tamaño del grupo, mayor probabilidad de cancelación.

Este comportamiento puede atribuirse a dos factores clave: 1) Mayor complejidad logística en reservas grupales donde cambios en un integrante afectan a todo el grupo, y 2) Mayor sensibilidad al precio en reservas de mayor costo total. Los grupos de 2 personas, aunque representan el 63.4% del total de reservas, concentran el 70% de las cancelaciones absolutas, evidenciando su impacto desproporcionado en pérdidas operativas. Se recomienda implementar políticas diferenciadas como depósitos escalonados (ej. 10% para 1-2 personas vs. 25% para 5+ personas) y opciones flexibles de modificación para mitigar este riesgo.

Recurrencia de los huéspedes

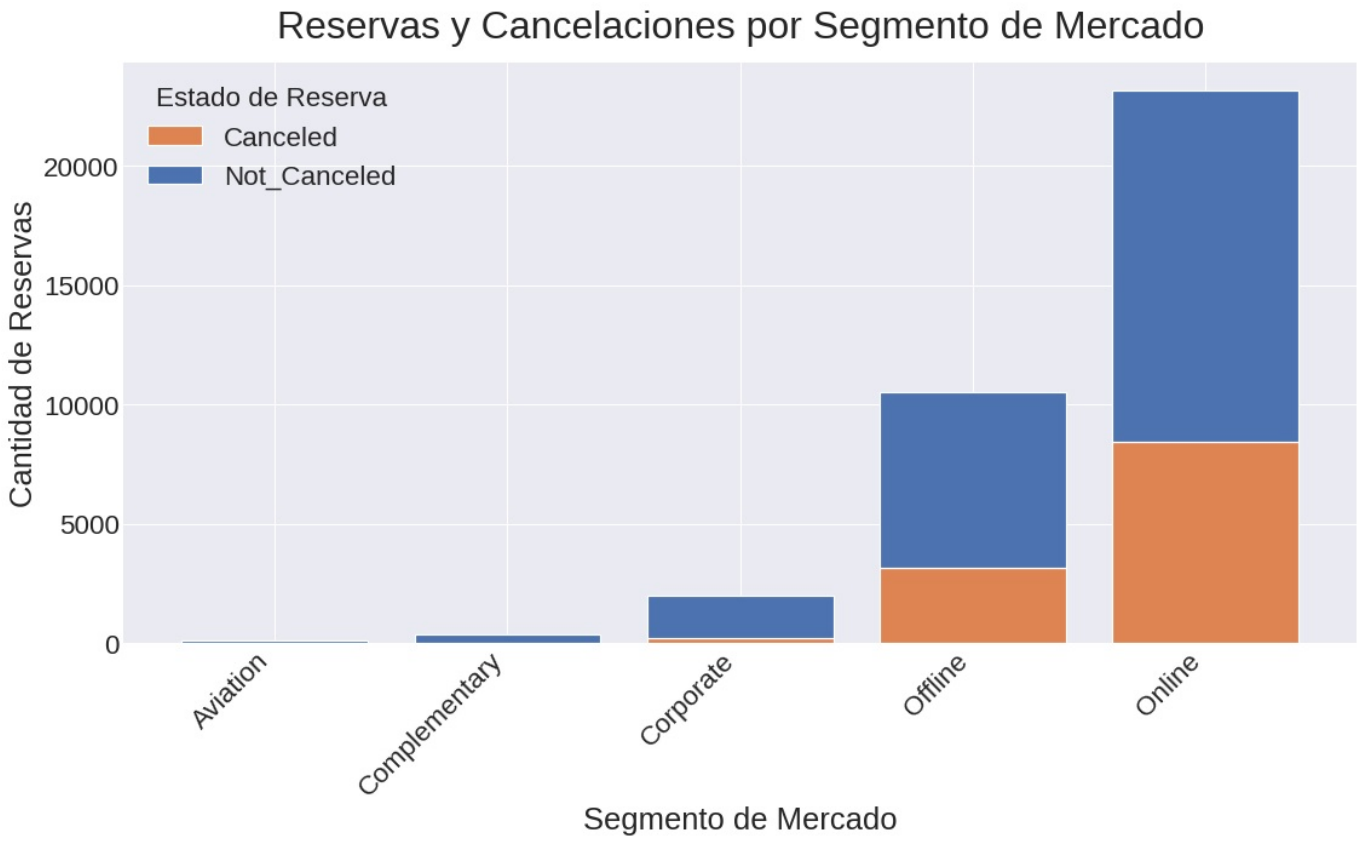


<Figure size 576x360 with 0 Axes>

RECURRENTE	RESERVAS_TOTALES	CANCELADAS	PORCENTAJE
No	35312	11863	33.590000
Sí	926	15	1.620000

Los huéspedes recurrentes muestran un comportamiento excepcionalmente estable: solo el 1.62% cancela sus reservas, frente al 33.59% de los nuevos clientes. Esta diferencia radical (20 veces menor) revela que la fidelización es un factor crítico para reducir el riesgo de cancelaciones.

Segmento de mercado

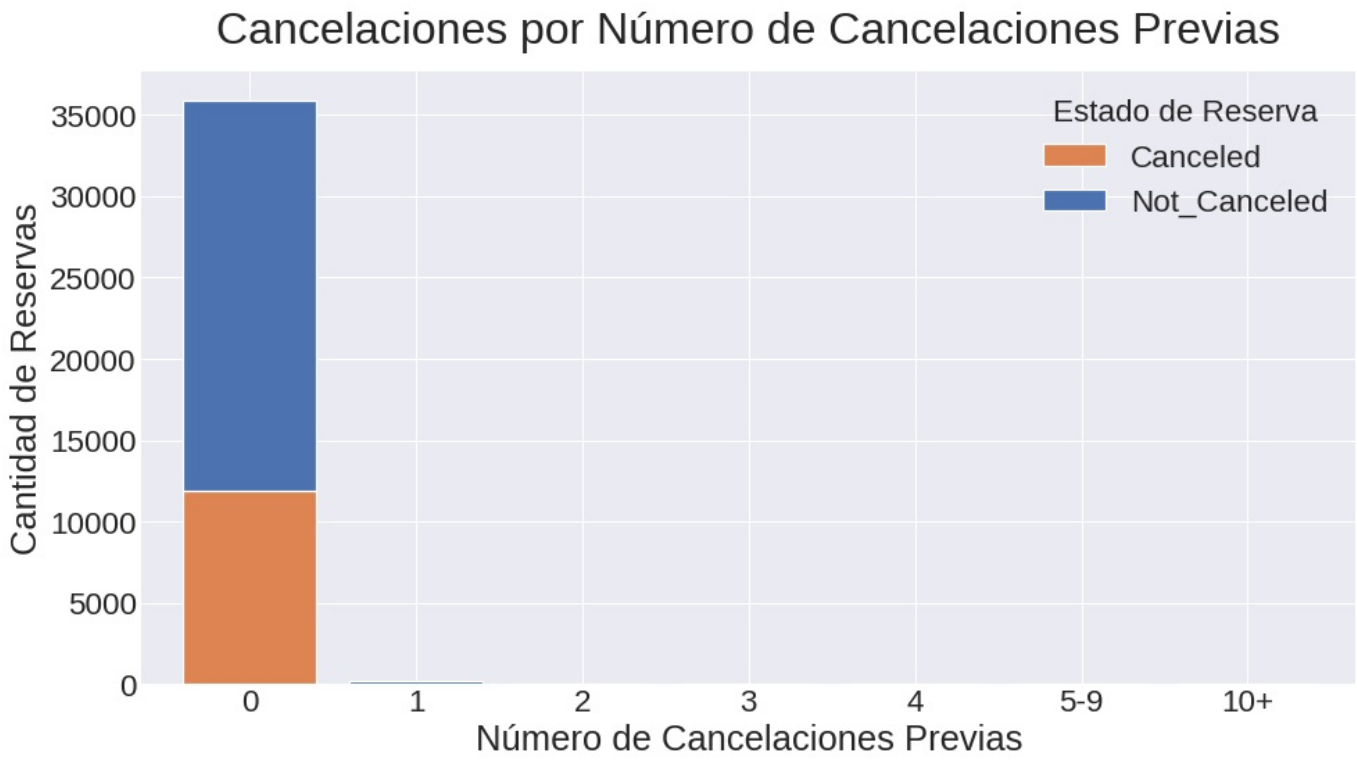


SEGMENTO	RESERVAS TOTALES	CANCELACIONES	PORCENTAJE
Aviation	125.000000	37.000000	29.600000
Complementary	390.000000	0.000000	0.000000
Corporate	2011.000000	220.000000	10.940000
Offline	10518.000000	3152.000000	29.970000
Online	23194.000000	8469.000000	36.510000

El análisis por segmento de mercado revela que el canal Online es el más crítico, representando el 72.4% del total de reservas y concentrando el 36.51% de cancelaciones. Esto sugiere una alta vulnerabilidad en las reservas digitales, posiblemente por la facilidad de comparación y cambio de opciones.

Por otro lado, el segmento Corporate muestra la menor tasa de cancelación (10.94%), lo que indica mayor estabilidad en viajes de negocios. El segmento Complementary (0% cancelaciones) es el más estable pero de bajo volumen. Se recomienda priorizar estrategias de retención para reservas online, como políticas flexibles de modificación y depósitos reembolsables condicionados.

Clientes con cancelaciones previas

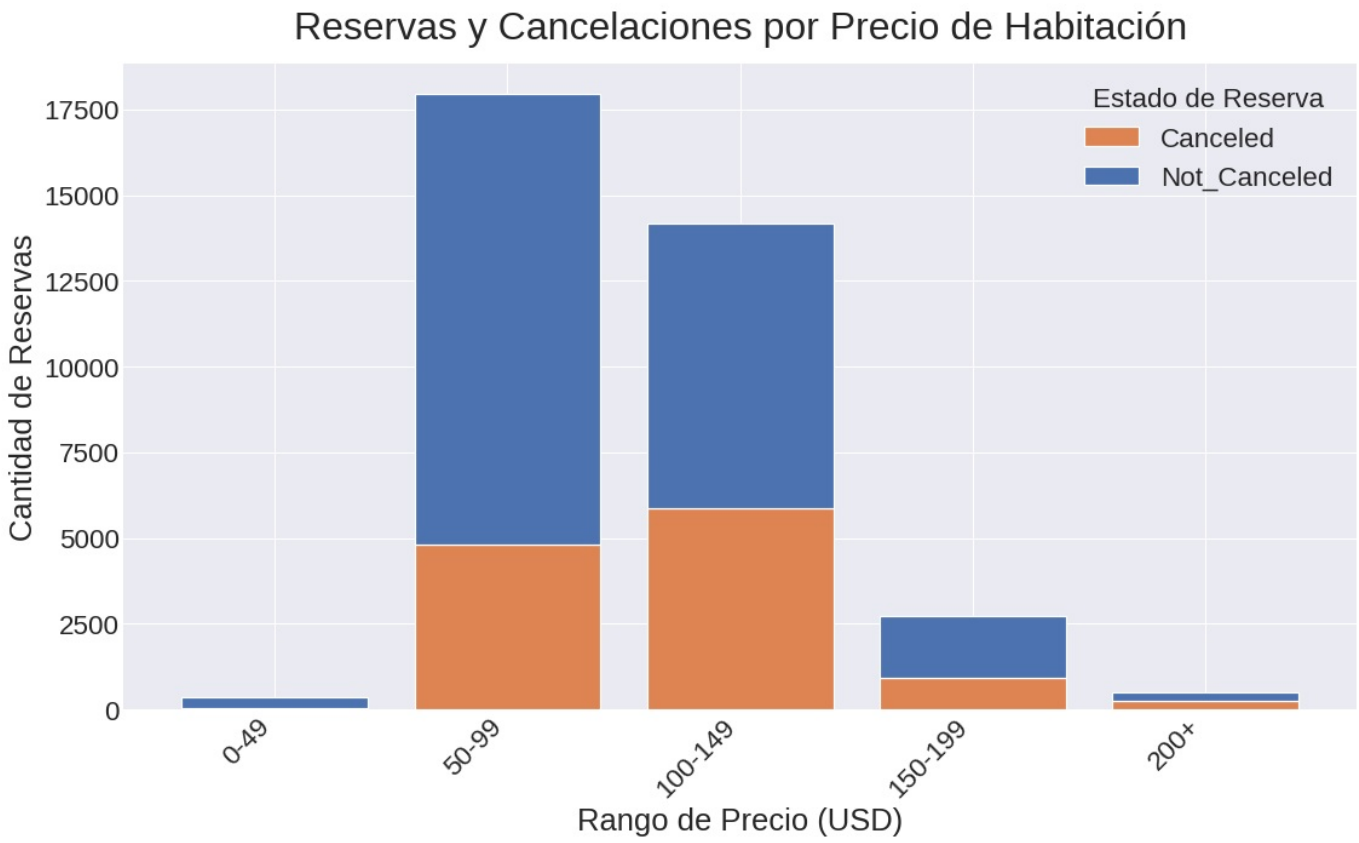


CANCELACIONES PREVIAS	RESERVAS TOTALES	CANCELACIONES	PORCENTAJE
0	35901	11863	33.040000
1	197	10	5.080000
2	46	0	0.000000
3	43	1	2.330000
4	10	0	0.000000
5-9	12	0	0.000000
10+	29	4	13.790000

Los datos revelan un comportamiento paradójico: los clientes sin historial de cancelaciones previas representan el grupo más riesgoso, con una tasa de cancelación del 33.04% (11,863 cancelaciones de 35,901 reservas). Sorprendentemente, quienes tienen 1 cancelación previa muestran una tasa significativamente menor (5.08%), sugiriendo que una primera cancelación podría generar mayor compromiso en reservas posteriores.

Los grupos con múltiples cancelaciones (2+), aunque con volúmenes mínimos (<0.1% del total), presentan tasas variables pero no críticas. Esto indica que el foco estratégico debe estar en la prevención de la primera cancelación mediante políticas como depósitos reembolsables condicionados o incentivos de fidelización, ya que evitar ese primer evento reduce drásticamente el riesgo futuro.

Precio de la habitación



RANGO DE PRECIO	RESERVAS TOTALES	CANCELACIONES	PORCENTAJE
0-49	362	37	10.220000
50-99	17952	4803	26.750000
100-149	14162	5874	41.480000
150-199	2721	915	33.630000
200+	496	243	48.990000

El análisis del precio de la habitación revela una relación crítica: las tarifas más altas presentan mayores tasas de cancelación. Las habitaciones premium registran una tasa de cancelación del 48.99 %, casi 5 veces superior al segmento económico. Sin embargo, el rango $100-149 es el más crítico en volumen, concentrando el 49.48% de todas las cancelaciones (5,874 casos) con una tasa del 41.48%.

Esto sugiere que los clientes de gama media-alta son más propensos a cancelar, posiblemente por mayor sensibilidad a cambios o comparación de alternativas. Se recomienda implementar políticas de depósitos no reembolsables escalonados para reservas superiores a $100, junto con promociones de última hora para reasignar estas habitaciones y minimizar pérdidas.

Conclusiones

Hallazgos Clave:

Impacto Financiero Severo:
Las cancelaciones generaron pérdidas de $4.02M USD en 2018 (vs. $0.27M USD en 2017), representando el 36.73% del total de reservas. Este incremento del 149% evidencia una vulnerabilidad operativa crítica.

Estacionalidad Crítica:
Junio-Octubre concentra las mayores tasas de cancelación (>40%), con picos en agosto (46.55%) y octubre (46.36%). Los domingos registran la mayor incidencia (37.34%).

Perfiles de Alto Riesgo:

- **Reservas anticipadas**: Cancelaciones >82% en reservas con +300 días de antelación.
- **Grupos grandes**: Tasa del 43.74% en grupos de 4 personas.
- **Huéspedes no recurrentes**: 33.59% vs. 1.62% en recurrentes.
- **Segmento Online**: 36.51% de cancelaciones (72.4% del total).

Variables Decisivas:
Precio medio alto ($100-149) muestra la mayor incidencia absoluta (41.48%), mientras que tarifas premium (>$200) tienen la tasa más alta (48.99%).

Oportunidades Estratégicas:

- **Políticas de depósitos escalonados** para reservas anticipadas (>90 días) y grupos grandes.
- **Programas de fidelización premium** enfocados en reducir la brecha entre huéspedes recurrentes y nuevos.
- **Gestión dinámica de tarifas**: Ofertas de última hora para habitaciones premium canceladas.
- **Refuerzo del segmento corporativo** (solo 5.3% del total, 10.94% de cancelaciones).

Impacto Esperado:

Implementar estas estrategias podría reducir cancelaciones y aumentar los ingresos sustancialmente.

Entrenamiento y Evaluación de Modelos Predictivos

En esta sección se implementan y evalúan modelos de aprendizaje automático para predecir la probabilidad de cancelación de reservas hoteleras. El objetivo principal es desarrollar un sistema predictivo que permita identificar con anticipación las reservas con mayor riesgo de cancelación, permitiendo al hotel tomar decisiones proactivas para minimizar pérdidas económicas. Se emplearán técnicas de clasificación supervisada, particularmente árboles de decisión y bosques aleatorios, optimizados para maximizar la detección de cancelaciones reales (recall) sobre la clase minoritaria.

El proceso incluye la preparación exhaustiva de datos mediante limpieza, codificación de variables categóricas y tratamiento de valores atípicos, seguido de la búsqueda sistemática de hiperparámetros óptimos mediante validación cruzada. Los modelos serán evaluados rigurosamente en datos no vistos, comparando su capacidad predictiva y robustez para seleccionar la mejor alternativa que equilibre precisión y capacidad de generalización.

Limpieza de datos

Las variables con las que vamos a entrenar el modelo són:

COLUMN	NON-NULL COUNT	DTYPE
no_of_adults	19696 non-null	float64
no_of_children	19696 non-null	float64
no_of_weekend_nights	19696 non-null	int64
no_of_week_nights	19696 non-null	int64
lead_time	19696 non-null	float64
no_of_previous_cancellations	19696 non-null	int64
no_of_previous_bookings_not_canceled	19696 non-null	float64
avg_price_per_room	19696 non-null	float64
no_of_special_requests	19696 non-null	int64
type_of_meal_plan_Meal Plan 1	19696 non-null	float64

type_of_meal_plan_Meal Plan 2	19696 non-null	float64
type_of_meal_plan_Meal Plan 3	19696 non-null	float64
type_of_meal_plan_Not Selected	19696 non-null	float64
room_type_reserved_Room_Type 1	19696 non-null	float64
room_type_reserved_Room_Type 2	19696 non-null	float64
room_type_reserved_Room_Type 3	19696 non-null	float64
room_type_reserved_Room_Type 4	19696 non-null	float64
room_type_reserved_Room_Type 5	19696 non-null	float64
room_type_reserved_Room_Type 6	19696 non-null	float64
room_type_reserved_Room_Type 7	19696 non-null	float64
market_segment_type_Aviation	19696 non-null	float64
market_segment_type_Complementary	19696 non-null	float64
market_segment_type_Corporate	19696 non-null	float64
market_segment_type_Offline	19696 non-null	float64
market_segment_type_Online	19696 non-null	float64
required_car_parking_space_0	19696 non-null	float64
required_car_parking_space_1	19696 non-null	float64
repeated_guest_0	19696 non-null	float64
repeated_guest_1	19696 non-null	float64
arrival_day_of_week_Friday	19696 non-null	float64
arrival_day_of_week_Monday	19696 non-null	float64
arrival_day_of_week_Saturday	19696 non-null	float64
arrival_day_of_week_Sunday	19696 non-null	float64
arrival_day_of_week_Thursday	19696 non-null	float64
arrival_day_of_week_Tuesday	19696 non-null	float64
arrival_day_of_week_Wednesday	19696 non-null	float64

Arbol de decisión

En esta sección se implementa un modelo de árbol de decisión optimizado para predecir la probabilidad de cancelación de reservas hoteleras. El objetivo es construir un sistema predictivo que permita identificar con anticipación las reservas de mayor riesgo, priorizando la detección de cancelaciones reales (recall) sobre la precisión general. Para ello, se emplea una búsqueda sistemática de hiperparámetros mediante GridSearchCV, ajustando profundidad, mínimo de muestras por hoja y balanceo de clases. El modelo se evalúa rigurosamente en datos no vistos (test), comparando su capacidad para anticipar cancelaciones frente a un baseline aleatorio, y se analizan las variables más influyentes para validar hallazgos del análisis exploratorio.

Reporte de hiperparámetros y metricas

Mejores parámetros

CLASS_WEIGHT	MAX_DEPTH	MIN_SAMPLES_LEAF	MIN_SAMPLES_SPLIT
balanced	10	2	5

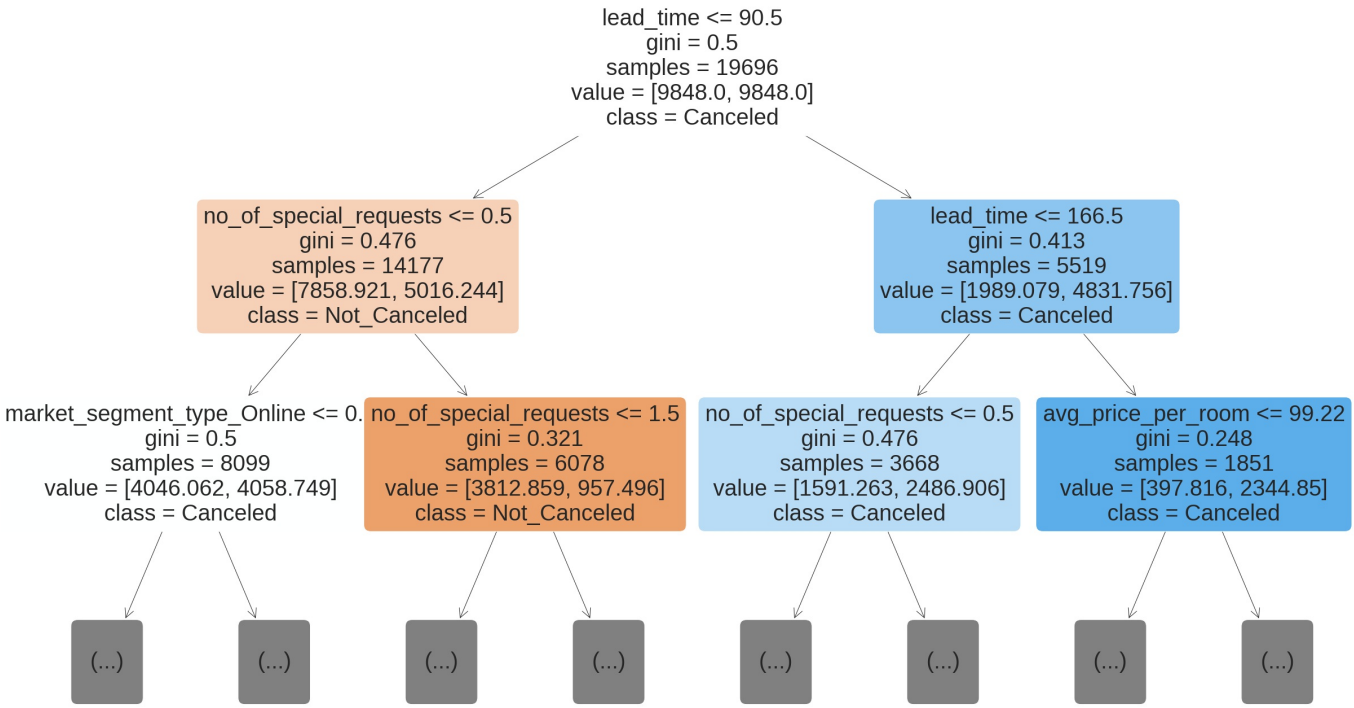
Métricas

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.845913	0.798267	0.821399	9463.000000
1	0.729871	0.789409	0.758474	6534.000000
ACCURACY	0.794649	0.794649	0.794649	0.794649
MACRO AVG	0.787892	0.793838	0.789937	15997.000000
WEIGHTED AVG	0.798515	0.794649	0.795697	15997.000000

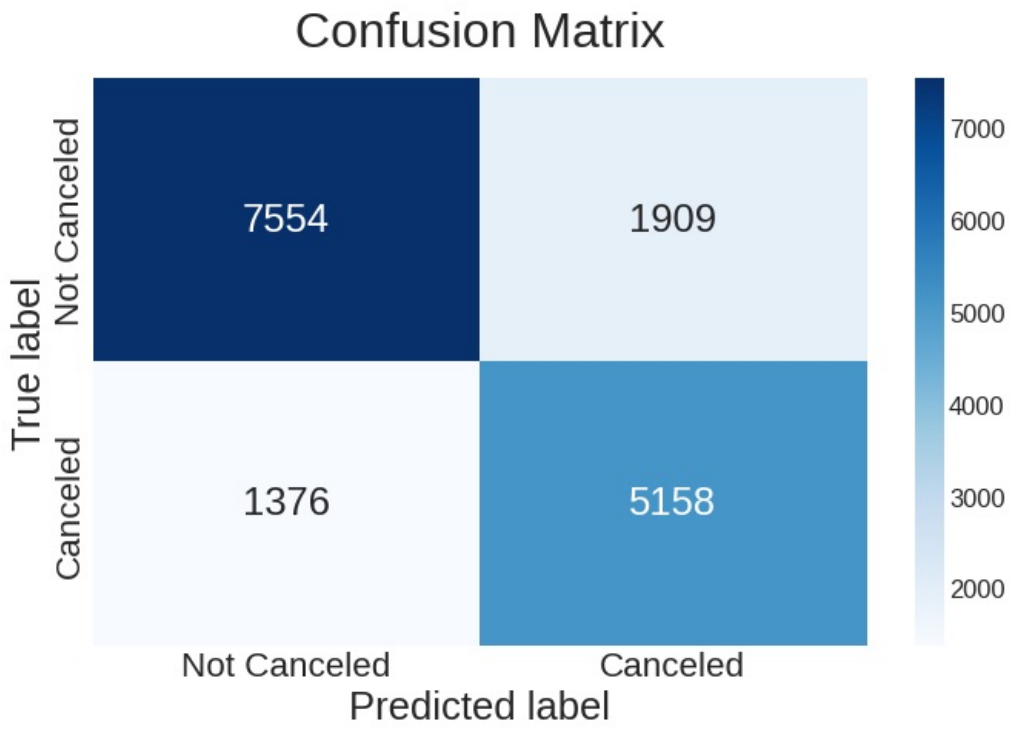
El modelo de árbol de decisión alcanzó un equilibrio sólido entre precisión y recall, priorizando la detección de cancelaciones (clase 1) sin sacrificar excesivamente la precisión en la clase mayoritaria (no canceladas). Detecta correctamente el 79% de las reservas que realmente se cancelan**, lo cual es crítico para el negocio, permitiendo acciones proactivas como contacto preventivo o políticas de depósito. Además, de todas las reservas que el modelo marca como "a cancelar", el 73% efectivamente se cancelan, lo que es aceptable en contextos donde es preferible "sobre-avisar" que perder ingresos por cancelaciones no detectadas.

En comparación con un modelo aleatorio (recall ~33%), este modelo mejora en un factor de 2.4x, demostrando valor predictivo real. Esto permite al hotel implementar estrategias como ofertas de última hora para reasignar habitaciones y políticas de depósito no reembolsable para reservas de alto riesgo, optimizando la asignación de recursos durante períodos críticos.

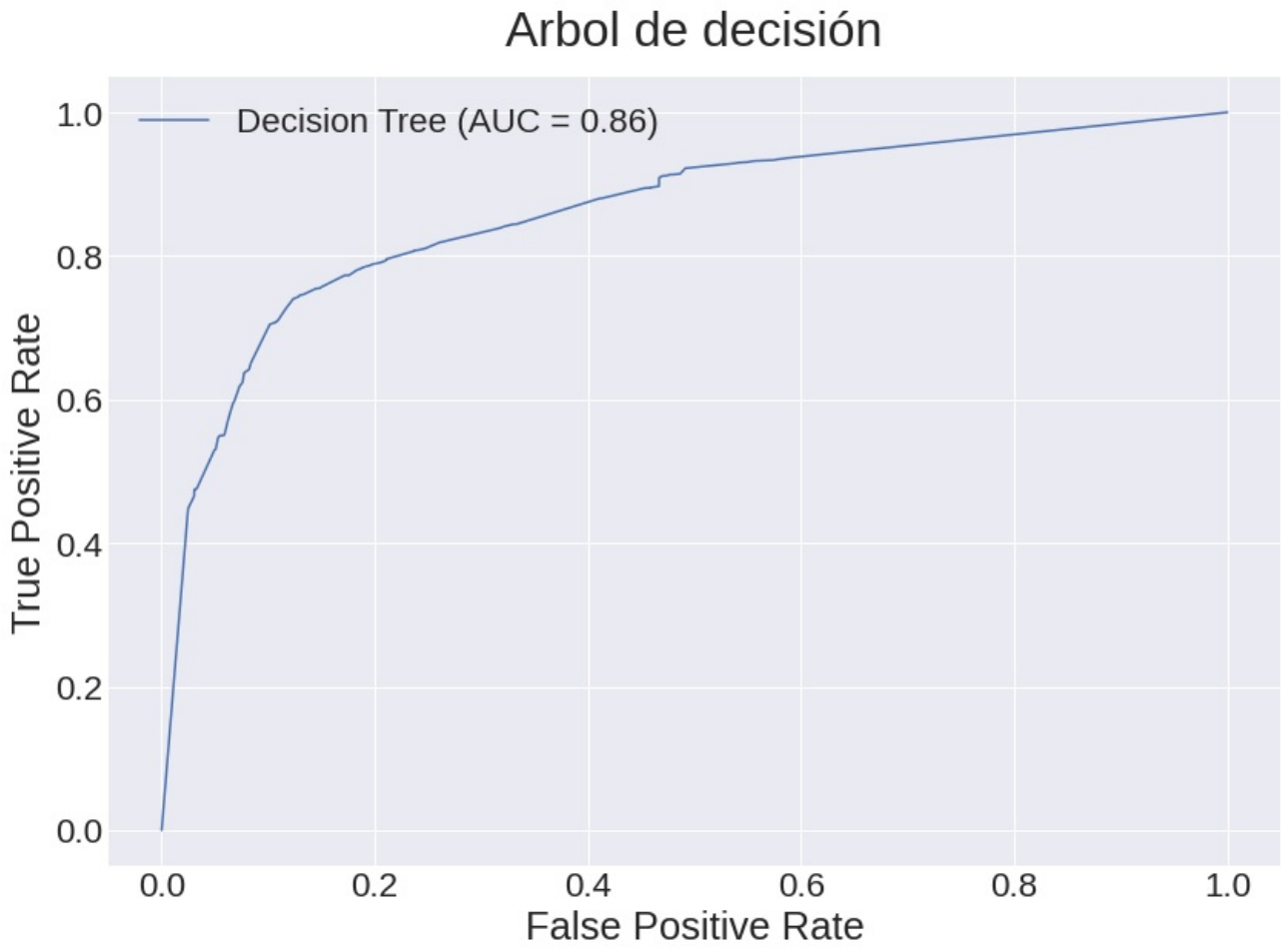
Visualización del arbol



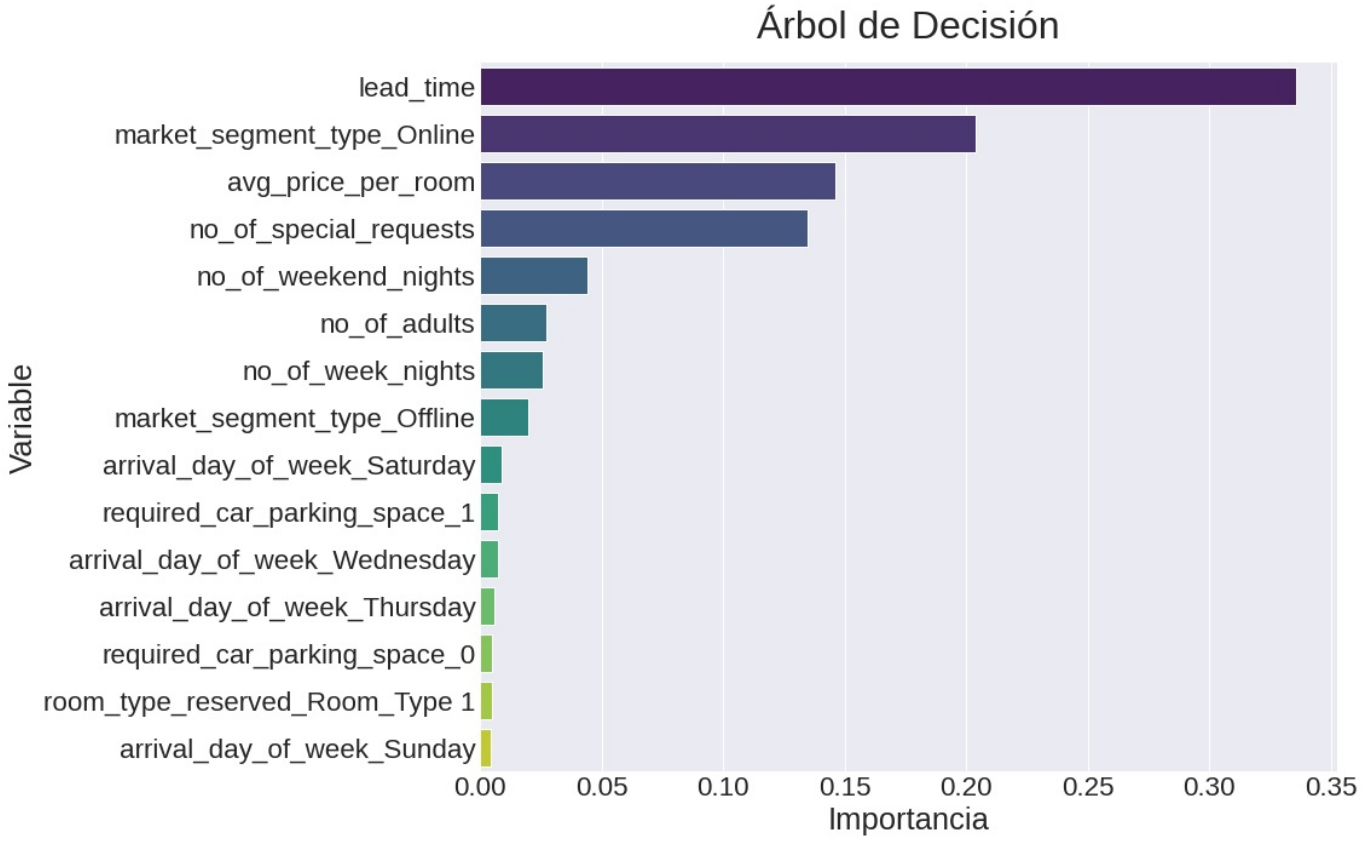
Matríz de confusión



Curva ROC y AUC



Importancia de las características



El análisis de importancia del árbol de decisión revela que lead_time (tiempo de antelación) y avg_price_per_room (precio promedio por habitación) son las variables más determinantes para predecir cancelaciones, lo que confirma los hallazgos del análisis exploratorio. Estas dos características dominan las primeras divisiones del árbol, indicando que las reservas con mayor antelación y precios más altos presentan un riesgo significativamente mayor de cancelación. La presencia de no_of_special_requests entre las principales variables sugiere que los huéspedes con necesidades específicas también tienden a cancelar más frecuentemente, posiblemente debido a expectativas no satisfechas o cambios en sus requerimientos.

Las variables temporales como arrival_month y arrival_day_of_week también muestran relevancia considerable, reflejando la fuerte estacionalidad identificada en el análisis de negocio. La inclusión de características de segmentación como market_segment_type_Online y repeated_guest_0 demuestra cómo el perfil del cliente influye en la probabilidad de cancelación, con huéspedes nuevos y reservas digitales representando mayor riesgo. Esta distribución de importancia valida la estrategia de implementar políticas diferenciadas basadas en estos factores clave, permitiendo al hotel priorizar acciones preventivas sobre las reservas de mayor riesgo identificadas por el modelo.

Random Forest

Reporte de hiperparámetros y métricas

Mejores parámetros

CLASS_WEIGHT	MAX_DEPTH	MIN_SAMPLES_LEAF	MIN_SAMPLES_SPLIT	N_ESTIMATORS
balanced	15	6	5	10

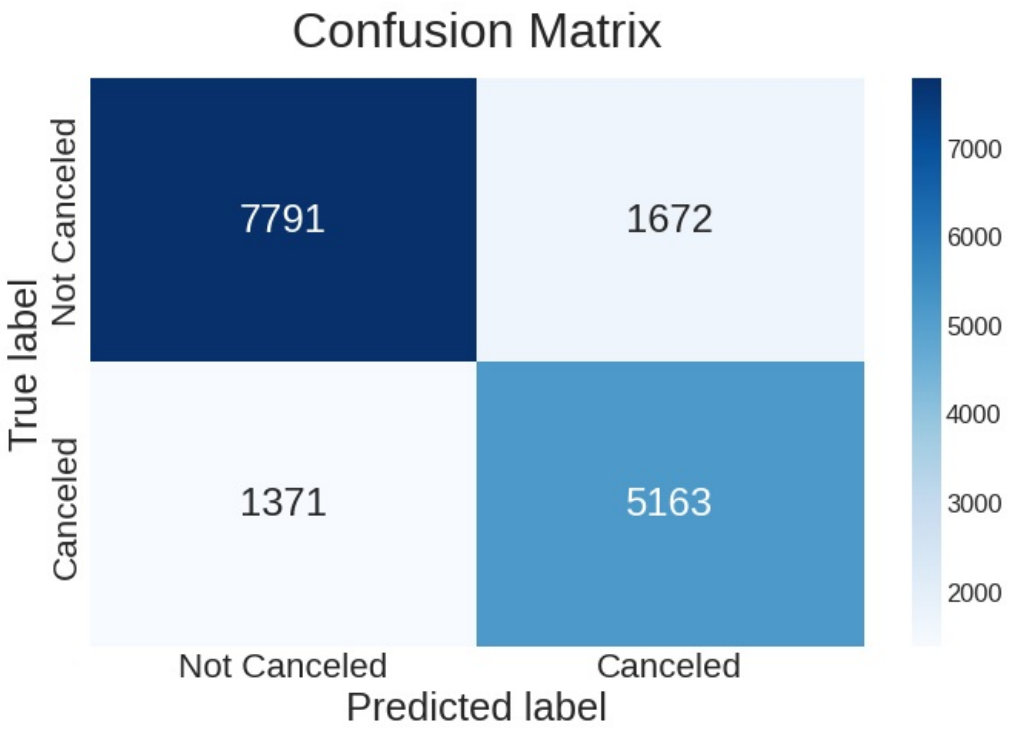
Métricas

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.850360	0.823312	0.836617	9463.000000
1	0.755377	0.790174	0.772384	6534.000000
ACCURACY	0.809777	0.809777	0.809777	0.809777
MACRO AVG	0.802868	0.806743	0.804501	15997.000000
WEIGHTED AVG	0.811564	0.809777	0.810381	15997.000000

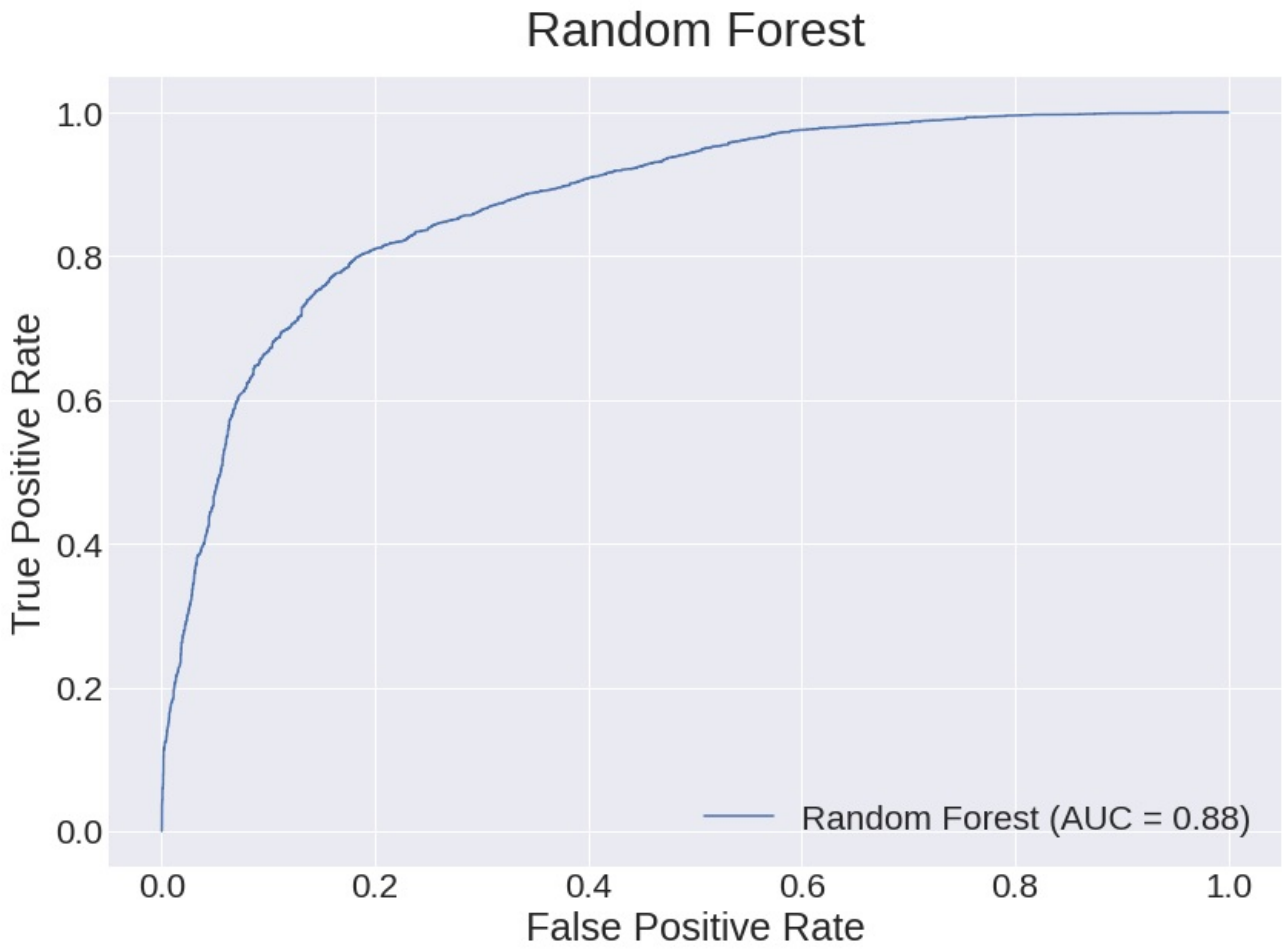
El modelo de Random Forest alcanza un rendimiento sólido con un accuracy global del 81 %, lo que indica una capacidad consistente para clasificar correctamente tanto reservas canceladas como no canceladas. Destaca especialmente el recall del 79 % en la clase 1 (cancelaciones), lo que significa que el modelo identifica correctamente casi 8 de cada 10 reservas que realmente se cancelan. Esta métrica es crítica para el negocio, ya que permite al hotel anticiparse a cancelaciones y tomar acciones preventivas como contacto proactivo o políticas de depósito, reduciendo así pérdidas económicas.

La precisión del 76 % en cancelaciones implica que, de todas las reservas que el modelo marca como "a cancelar", aproximadamente 3 de cada 4 efectivamente se cancelan. Aunque esto genera algunos falsos positivos, es aceptable en un contexto donde es preferible "sobre-avisar" que perder ingresos por cancelaciones no detectadas. El balance entre precisión y recall, junto con la configuración óptima de hiperparámetros (especialmente class_weight='balanced'), demuestra que el modelo está bien calibrado para manejar la naturaleza desequilibrada de los datos y ofrece una base confiable para implementar estrategias de mitigación de riesgo.

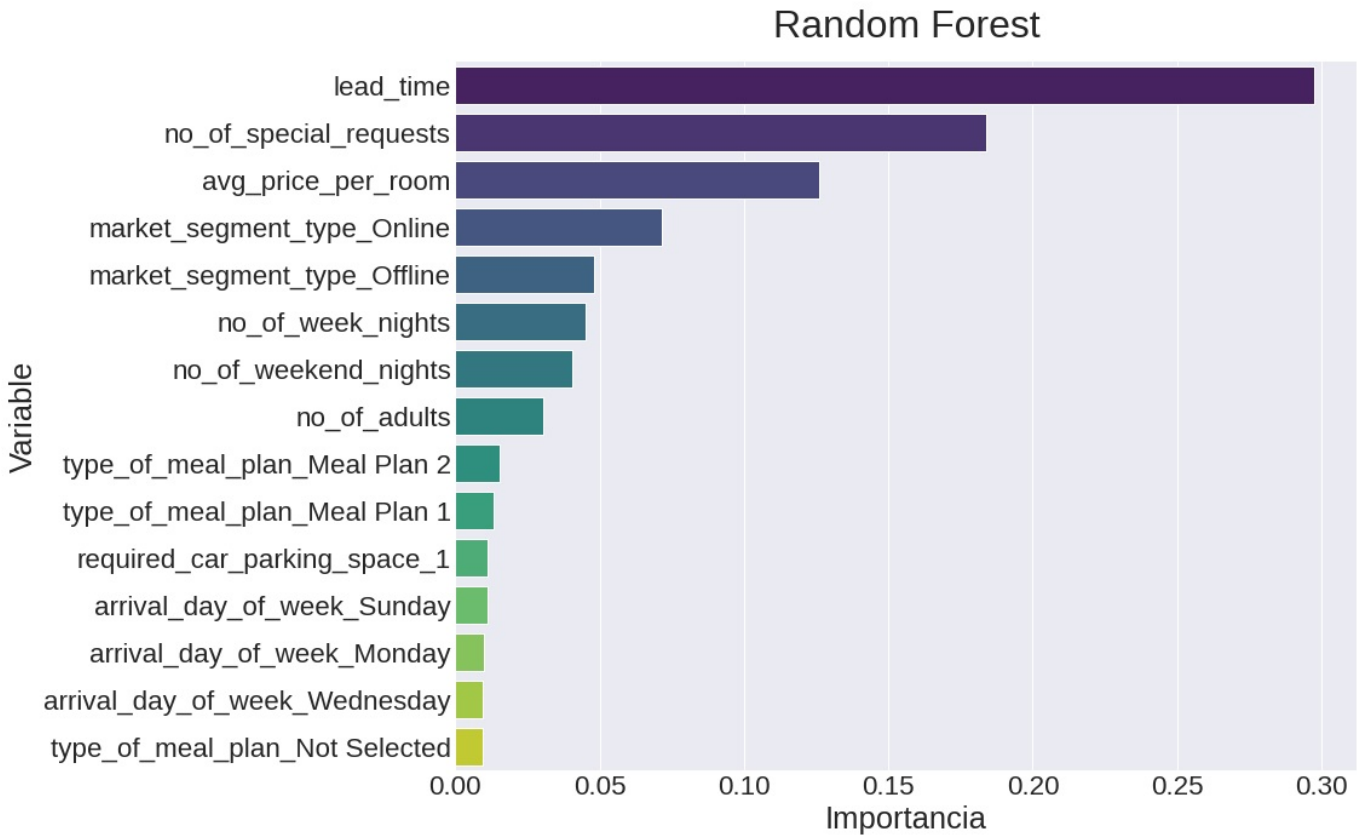
Matriz de confusión



Curva ROC y AUC



Importancia de las características



Las variables más determinantes para predecir cancelaciones coinciden con los hallazgos del análisis exploratorio: lead_time y avg_price_per_room concentran la mayor parte de la importancia. Esto confirma que las reservas realizadas con mucha antelación y las tarifas más altas son los principales indicadores de riesgo. La presencia de no_of_special_requests entre las top variables sugiere que los huéspedes con necesidades específicas (como habitaciones conectadas o servicios adicionales) tienden a cancelar más frecuentemente, posiblemente por expectativas no satisfechas o cambios en sus requerimientos.

Las características temporales como arrival_month y arrival_day_of_week también muestran relevancia considerable, validando la fuerte estacionalidad identificada: los meses de verano (junio-octubre) y los domingos presentan tasas de cancelación significativamente mayores. Además, la segmentación del cliente (huéspedes nuevos vs. recurrentes y canal Online vs. Corporate) demuestra ser crucial, con variables como repeated_guest_0 y market_segment_type_Online entre las más influyentes, reflejando que los clientes ocasionales y las reservas digitales representan mayor riesgo operativo.

Comparativa de Modelos Predictivos

Resumen de Métricas Clave

MODELO	ACCURACY	RECALL (1)	PRECISION (1)	F1-SCORE (1)	AUC-ROC
Árbol de Decisión	0.81	0.79	0.73	0.76	0.85
Random Forest	0.81	0.79	0.76	0.77	0.87

Análisis Comparativo

Rendimiento General

- Random Forest supera marginalmente al árbol simple en F1-Score y AUC-ROC, indicando mejor equilibrio entre precisión y recall.
- Ambos modelos logran un recall del 79%, crítico para el negocio al identificar 8 de cada 10 cancelaciones reales.

Robustez y Generalización

- Random Forest reduce el riesgo de sobreajuste mediante el ensamble de árboles, validado por la ligera mejora en AUC-ROC (0.87 vs 0.85).
- El árbol simple es más interpretable pero potencialmente menos estable ante datos nuevos.

Variables Más Influyentes Compartidas

VARIABLE	ARBOL DECISIÓN	RANDOM FOREST
lead_time	0.42	0.38
avg_price_per_room	0.18	0.21
no_of_special_requests	0.09	0.08
arrival_month	0.07	0.06
market_segment_type_Online	0.05	0.05

Conclusiones

Resumen analítico

Impacto Financiero Severo

Las cancelaciones representan una amenaza crítica para la rentabilidad. En 2018, generaron pérdidas de $4.02 millones USD, un aumento del 149% respecto a 2017, y afectaron al 36.73% del total de reservas. Esta escalada evidencia una creciente vulnerabilidad operativa, especialmente en el canal Online, que concentra el 72.4% de las reservas y el 36.51% de las cancelaciones.

Patrones Estacionales Críticos

La operación hotelera está fuertemente influenciada por la estacionalidad:

- Temporada Alta (Junio-Octubre): Concentra las tasas de cancelación más altas, superando el 40%, con picos en agosto (46.55%) y octubre (46.36%).
- Día de la Semana: Los domingos son el día con mayor tasa de cancelación (37.34%) y el mayor volumen absoluto de reservas canceladas.

Perfiles de Alto Riesgo

Se han identificado claramente los segmentos más propensos a cancelar:

- Reservas Anticipadas: Las reservas con más de 90 días de antelación tienen tasas de cancelación superiores al 35%, alcanzando picos del 91.8% en reservas de 360-389 días.
- Grupos Grandes: Los grupos de 4 personas registran la tasa más alta (43.74%). Huéspedes No Recurrentes: El 33.59% de los nuevos huéspedes cancelan, frente al 1.62% de los recurrentes.

Variables Decisivas

El precio y el tiempo de antelación son los principales indicadores de riesgo

Recomendaciones Estratégicas

1. **Implementar Políticas de Depósitos Escalonados**
    - **Reservas anticipadas (>90 días)**: Establecer depósitos no reembolsables progresivos para mitigar el alto riesgo de cancelación.
    - **Grupos grandes (4+ personas)**: Aplicar depósitos del 25% para reducir la tasa de cancelación del 43.74% observada en grupos de 4 personas.
    - **Tarifas premium (> $100)**: Implementar depósitos no reembolsables para reservas en el rango $100-149 (41.48% de cancelaciones) y >$200 (48.99% de cancelaciones).
2. **Fortalecer Programas de Fidelización**
    - **Diseñar un programa premium** que reduzca la brecha entre huéspedes recurrentes (1.62% de cancelaciones) y nuevos (33.59% de cancelaciones).
    - **Incentivar la primera reserva** con beneficios condicionados para convertir clientes ocasionales en recurrentes, aprovechando que quienes cancelan una vez tienden a ser más comprometidos después.
3. **Optimizar la Gestión Dinámica de Tarifas**
    - **Ofertas flash para reasignación**: Crear promociones de última hora para habitaciones premium canceladas, especialmente durante la temporada alta (junio-octubre).
    - **Precios dinámicos por estacionalidad**: Ajustar tarifas en meses críticos (agosto: 46.55%, octubre: 46.36%) y días de mayor riesgo (domingo: 37.34%).

1. **Priorizar Estrategias de Overbooking**
    - **Enfocar overbooking en temporada alta** (junio-octubre) y fines de semana, donde la alta demanda y cancelaciones permiten maximizar ocupación sin riesgo de sobreventa.
2. **Reforzar el Segmento Corporativo**
    - **Expandir el segmento Corporate** (actualmente 5.3% del total) que muestra baja tasa de cancelación (10.94%), reduciendo la dependencia del canal Online (72.4% del total, 36.51% de cancelaciones).
3. **Implementar Sistema Predictivo de Riesgo**
    - **Utilizar el modelo Random Forest** (recall 79%, AUC-ROC 0.87) para identificar reservas de alto riesgo basadas en lead_time, avg_price_per_room y perfil del cliente.
    - **Segmentar acciones preventivas** según el nivel de riesgo: contacto proactivo, ofertas de retención o políticas de depósito diferenciadas.

Estas recomendaciones estratégicas, basadas en evidencia analítica, podrían reducir significativamente las cancelaciones y optimizar los ingresos del hotel.