Análisis de cancelaciones y ocupación hotelera

Ciencia de Datos Aplicada 2025-2

Juan David Valencia - 201728857
Juan Esteban Cuellar - 202014258

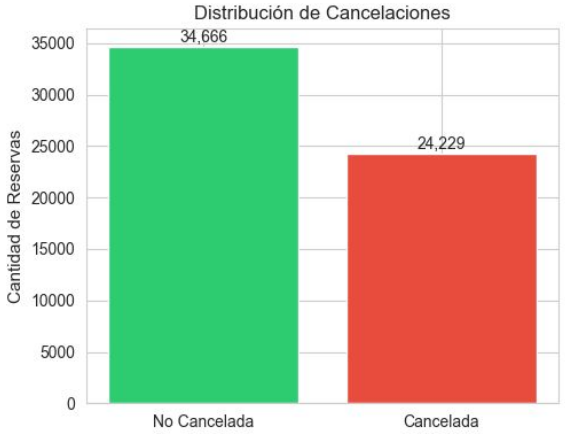
Agenda

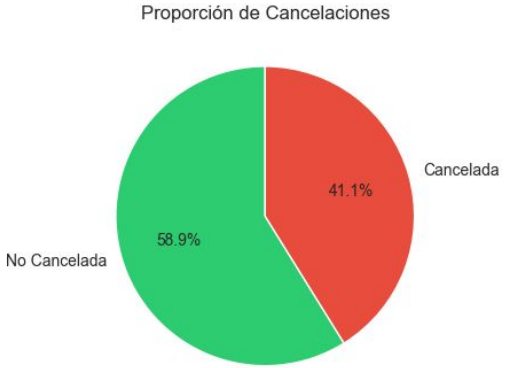
1. Contexto y objetivos
2. Datos y variables clave
3. Estrategia de análisis
4. Hallazgos principales
5. Segmentos de alto riesgo
6. Modelo predictivo (resumen)
7. Recomendaciones con evidencia
8. Próximos pasos / Q&A

Contexto y objetivo

La cadena enfrenta cancelaciones elevadas que afectan ingresos y planeación.

Objetivo: identificar factores de cancelación y proponer acciones para mejorar la ocupación.





Entendimiento de datos

Datos y variables clave

is_canceled-Variable objetivo

lead_time-Días entre la reserva y la llegada

adr-Tarifa diaria promedio

deposit_type-Tipo de depósito solicitado

hotel-Tipo de hotel (urbano o resort)

La tasa de cancelación global alcanza el 41.1%, un desafío relevante para la cadena.

La mediana del lead time es de 69 días, lo que indica que la mitad de los clientes reserva con poca antelación.

El ADR promedio es de $97.84, con una dispersión amplia y algunos valores extremos.

La mayoría de reservas (≈88.9%) se hacen sin depósito, aumentando el riesgo de cancelación.

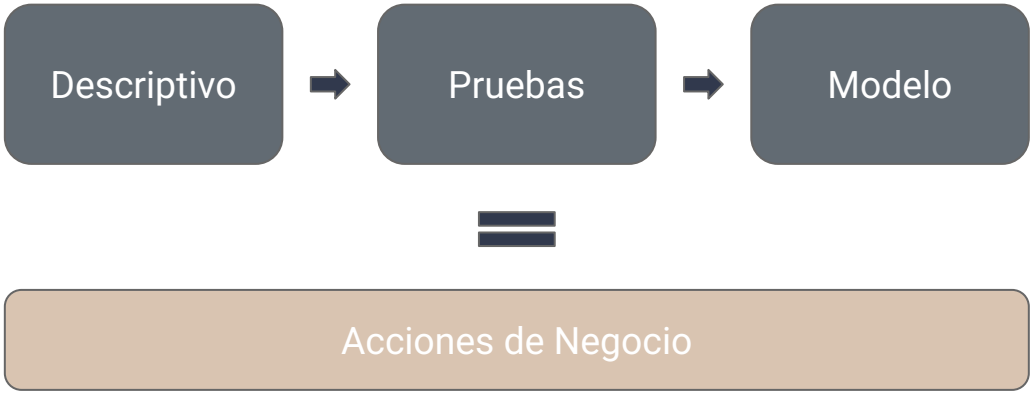
El Resort Hotel concentra la mayoría de las reservas, mientras que el City Hotel representa cerca de un tercio.

Estrategia de análisis (resumen)

Descriptivo segmentado (hotel, canal, temporada).

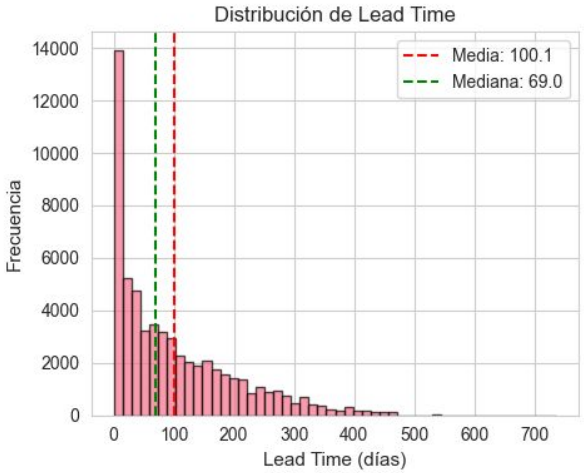
Validación estadística de diferencias.

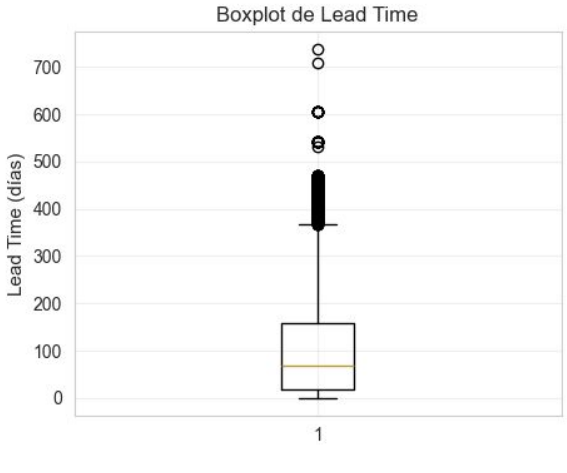
Regresión logística interpretable para estimar riesgo de cancelación.

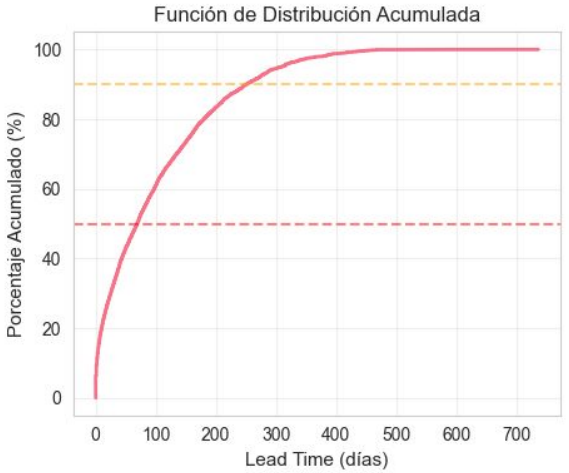


Hallazgo 1: Lead time

Reservas con mucha anticipación cancelan más.





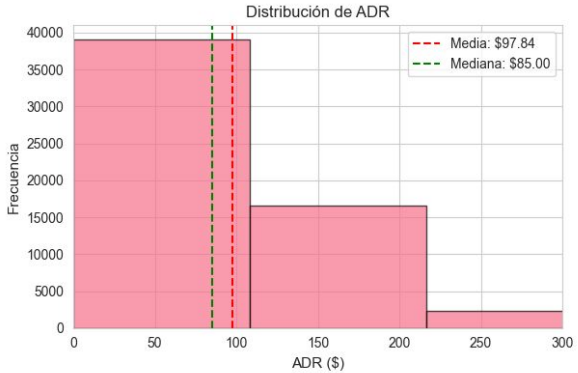


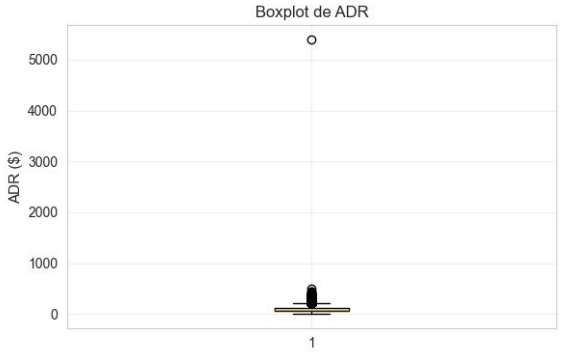
Hallazgo 2: ADR (precio)

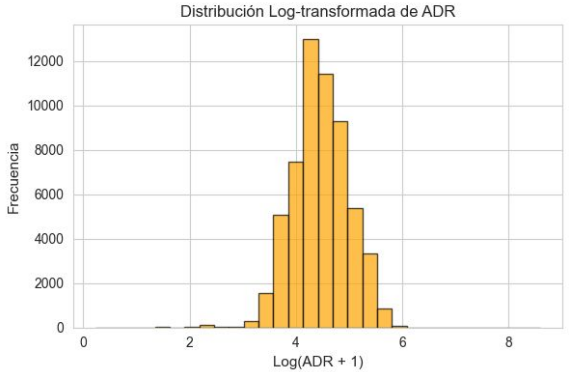
Tarifas típicas entre $61 y $121.

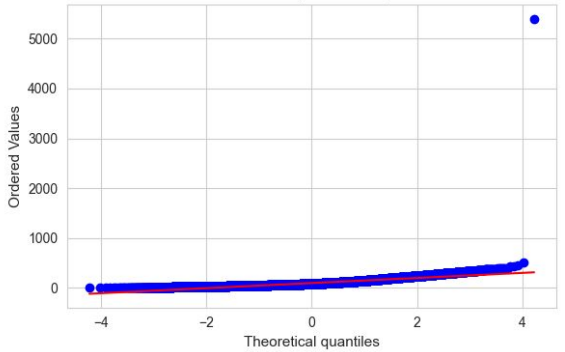
Casos extremos hasta $5,400.

Oportunidad: ajustar precios y controlar outliers.





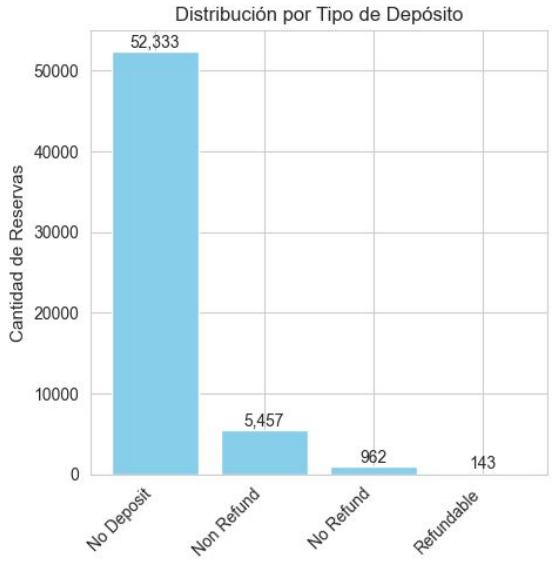


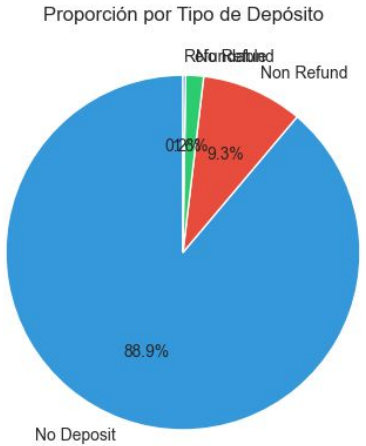


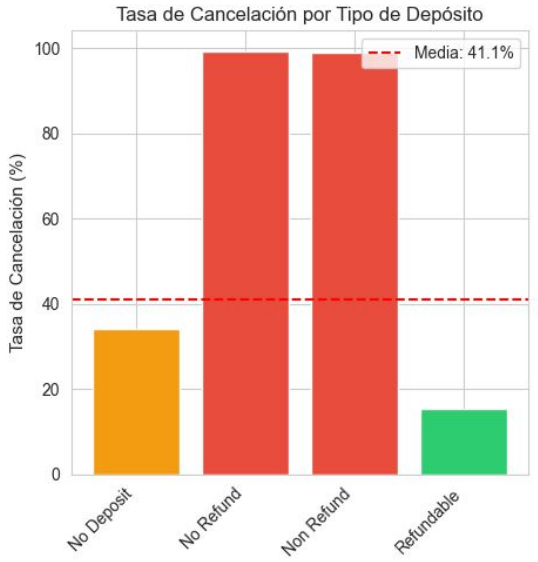
Hallazgo 3: Depósito

No-deposit concentra cancelaciones; no-refundable reduce fuerte el riesgo.

Punto medio: refundable.



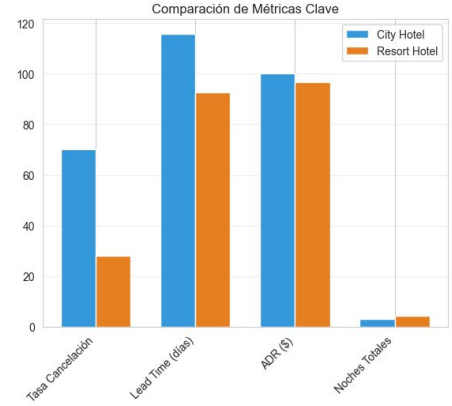


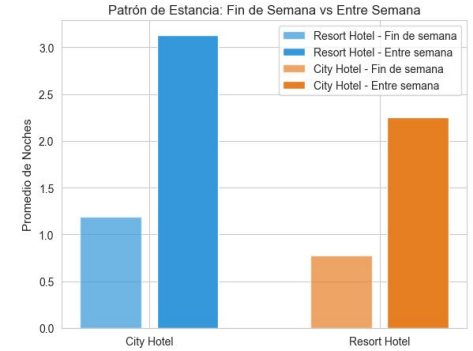


Hallazgo 4: Tipo de hotel

Diferencias claras entre City y Resort (perfil y estancia).

Estrategias deberían ser distintas.





Estrategia de analisis

Framework de Análisis

Fase 1: Preparación:

1. Limpieza de datos (outliers, valores faltantes)
2. Ingeniería de variables (total_guests, lead_time_bucket, season)
3. Validación de calidad de datos

Fase 2: Análisis Exploratorio:

1. Distribuciones univariadas de variables clave
2. Análisis bivariado (cancelación vs predictores)
3. Matriz de correlación y detección de multicolinealidad
4. Segmentación inicial por dimensiones de negocio

Fase 3: Análisis Estadístico:

1. Tablas de contingencia y tasas por segmento
2. Tests de hipótesis (chi-cuadrado, Mann-Whitney)
3. Cálculo de tamaños de efecto
4. Intervalos de confianza para métricas clave

Fase 4: Modelado:

1. Selección de variables por importancia
2. Entrenamiento de modelo logístico
3. Validación y métricas de desempeño
4. Interpretación de coeficientes y odds ratios

Estrategia propuesta:

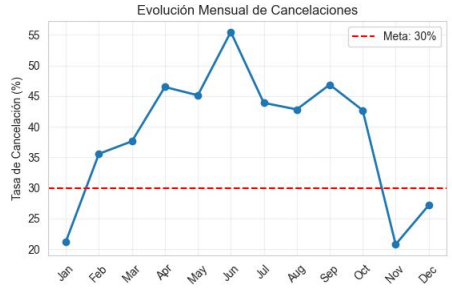
1) Análisis descriptivo segmentado por tipo de hotel, temporada y cliente para detectar patrones de cancelación.

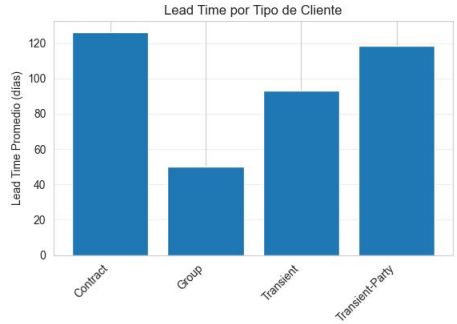
2) Pruebas estadísticas simples para validar que las diferencias encontradas sean reales y no ruido.

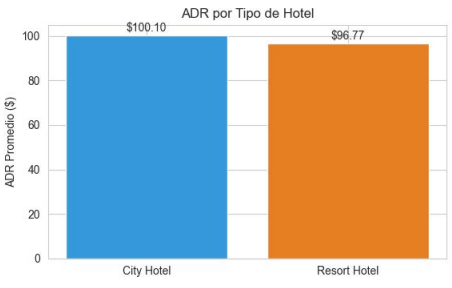
3) Un modelo de regresión logística interpretable que permita estimar la probabilidad de cancelación según factores como lead_time, ADR y políticas de depósito.

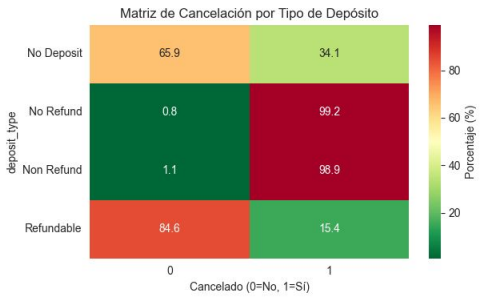
KPI’s

Kpi	Actual	Meta	Impacto
Tasa de Cancelación	41.14%	<30%	Alto
Lead Time Promedio	100 días	Pendiente	Medio
ADR	$97.84	+10% próximo año	Alto
Ocupacion efectiva	58.9%	>75%	Alto
Revenue x habitacion	Pendiente	+10% próximo año	Alto
Conversion por deposito	11.1%	>30%	Medio









Desarrollo Estrategia

Test Estadísticos

Chi cuadrados

TESTS CHI-CUADRADO: ASOCIACIÓN CON CANCELACIONES

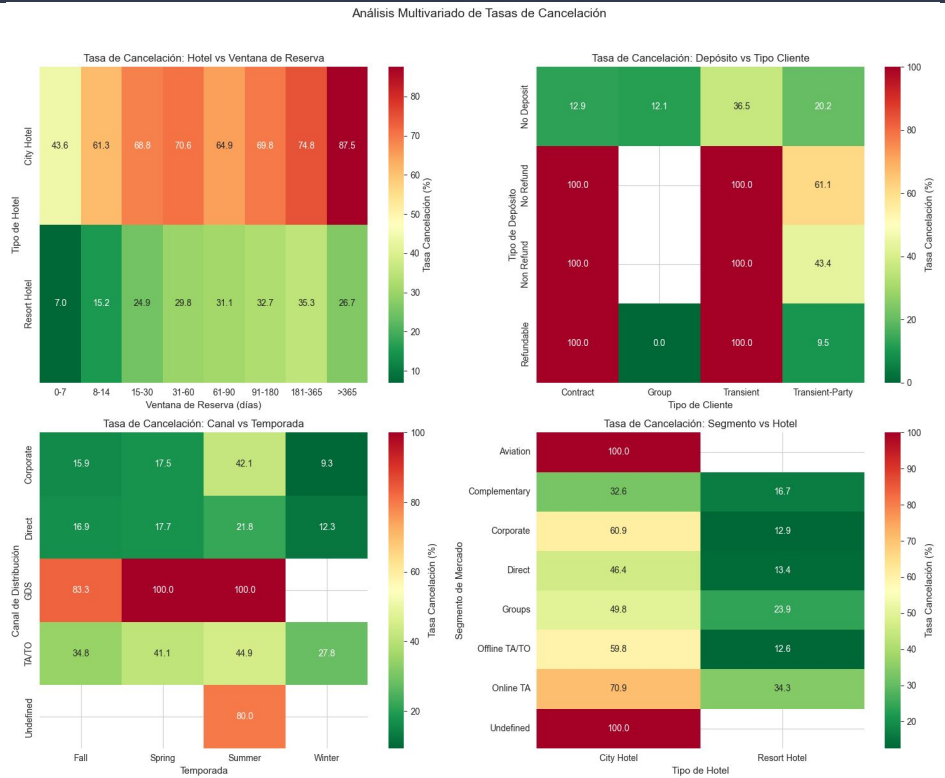
Variable	Chi2	p-value	Cramers V	Efecto	Significativo
hotel	6161.63	0.0000	0.371	Efecto mediano	Sí
deposit_type	1828.38	0.0000	0.202	Efecto pequeño	Sí
customer_type	1347.33	0.0000	0.174	Efecto pequeño	Sí
distribution_channel	1620.57	0.0000	0.190	Efecto pequeño	Sí
market_segment	3213.62	0.0000	0.268	Efecto pequeño	Sí
meal	1209.81	0.0000	0.164	Efecto pequeño	Sí
reserved_room_type	194.93	0.0000	0.066	Efecto despreciable	Sí

MANN-WHITNEY

TESTS MANN-WHITNEY: DIFERENCIAS ENTRE CANCELADOS Y NO CANCELADOS

Variable	Media No Canc.	Media Canc.	U-stat	p-value	Cohen's d	Significativo
lead_time	73.81	104.36	162890663	0.0000	0.344	Sí
adr	93.22	115.15	163838074	0.0000	0.386	Sí
total_stay_nights	4.11	4.25	211689492	0.0000	0.045	Sí
total_guests	2.24	2.53	199885346	0.0000	0.058	Sí
days_in_waiting_list	1.25	1.31	225061426	0.0031	0.005	Sí
total_of_special_requests	0.69	0.47	259179660	0.0000	-0.279	Sí

Análisis Multivariado



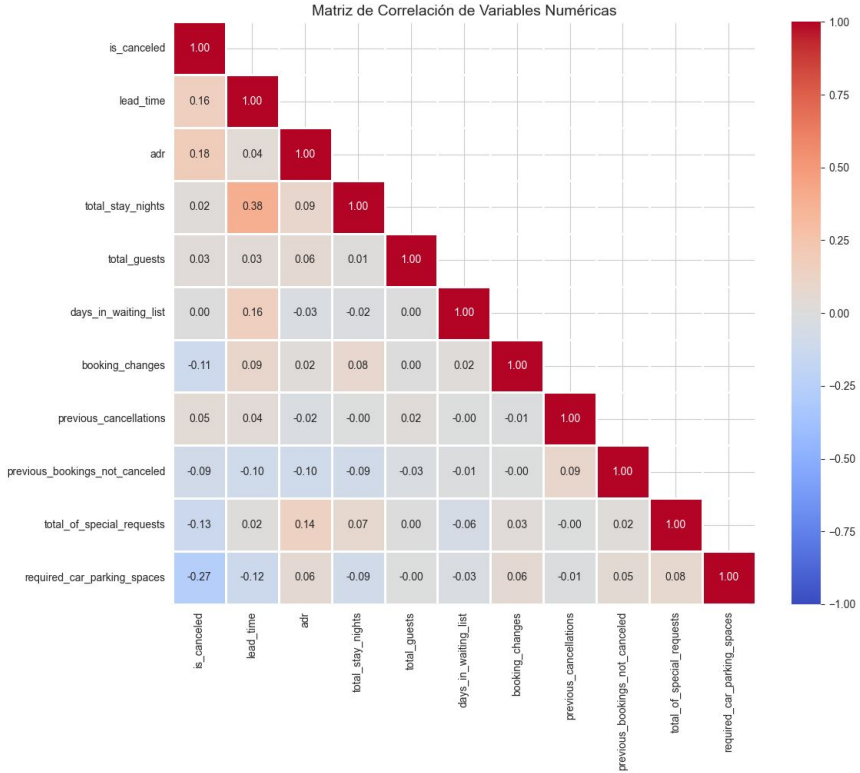
City Hotel + lead time largo → muy alta cancelación (hasta 88%).

Depósitos no reembolsables → casi 100% cancelación; reembolsables bajan el riesgo.

TA/TO y verano → mayores tasas de cancelación.

City Hotel cancela más en todos los segmentos, Resort Hotel es más estable.

Análisis de Correlaciones



is_canceled se asocia débilmente con lead_time (0.16) y adr (0.18).

total_stay_nights se relaciona más con lead_time (0.38) que con cancelación.

required_car_parking_spaces muestra la correlación negativa más fuerte con cancelación (-0.27).

En general, las correlaciones son bajas, lo que confirma que la cancelación depende de múltiples factores combinados más que de una sola variable.

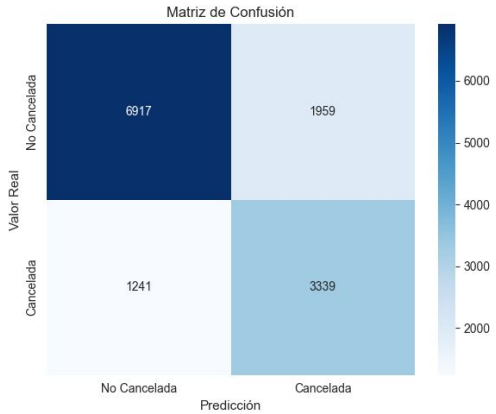
Modelo Predictivo: Regresión Logística

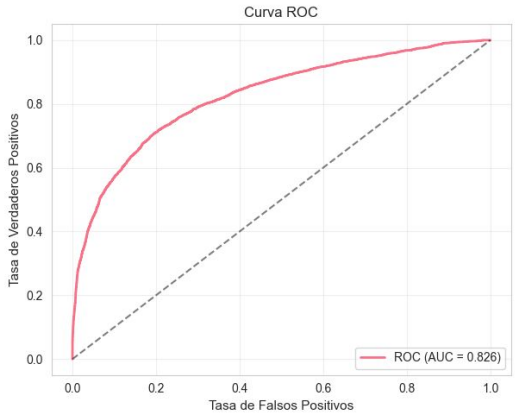
El modelo logra un AUC de 0.83, lo que indica buena capacidad discriminativa.

Predice correctamente 6,917 reservas no canceladas y 3,339 canceladas.

Los falsos positivos (1,959) y falsos negativos (1,241) son moderados, mostrando un balance aceptable entre precisión y recall.

Es una herramienta útil para anticipar cancelaciones, aunque aún puede mejorarse ajustando umbrales o incorporando más variables.





TOP 15 Factores

Mayor riesgo de cancelación: reservas con cancelaciones previas, lead time largo, clientes transients y depósitos “No Refund”.

Menor riesgo: huéspedes repetidos, reservas directas, más solicitudes especiales y Resort Hotel.

Feature	Coeficiente	Odds Ratio	Interpretación
previous_cancellations	+1.007	2.737	↑ 173.7% prob.
hotel_Resort Hotel	-0.777	0.460	↓ 54.0% prob.
lead_time	+0.507	1.660	↑ 66.0% prob.
customer_type_Transient	+0.464	1.590	↑ 59.0% prob.
total_of_special_requests	-0.449	0.639	↓ 36.1% prob.
market_segment_Offline TA/TO	-0.408	0.665	↓ 33.5% prob.
deposit_type_Non Refund	+0.398	1.489	↑ 48.9% prob.
is_repeated_guest	-0.386	0.680	↓ 32.0% prob.
deposit_type_No Refund	+0.353	1.424	↑ 42.4% prob.
market_segment_Direct	-0.347	0.707	↓ 29.3% prob.
adr	+0.312	1.366	↑ 36.6% prob.
booking_changes	-0.269	0.764	↓ 23.6% prob.
market_segment_Groups	-0.209	0.812	↓ 18.8% prob.
days_in_waiting_list	-0.126	0.882	↓ 11.8% prob.
market_segment_Online TA	+0.111	1.117	↑ 11.7% prob.

Identificación segmentos de alto riesgo

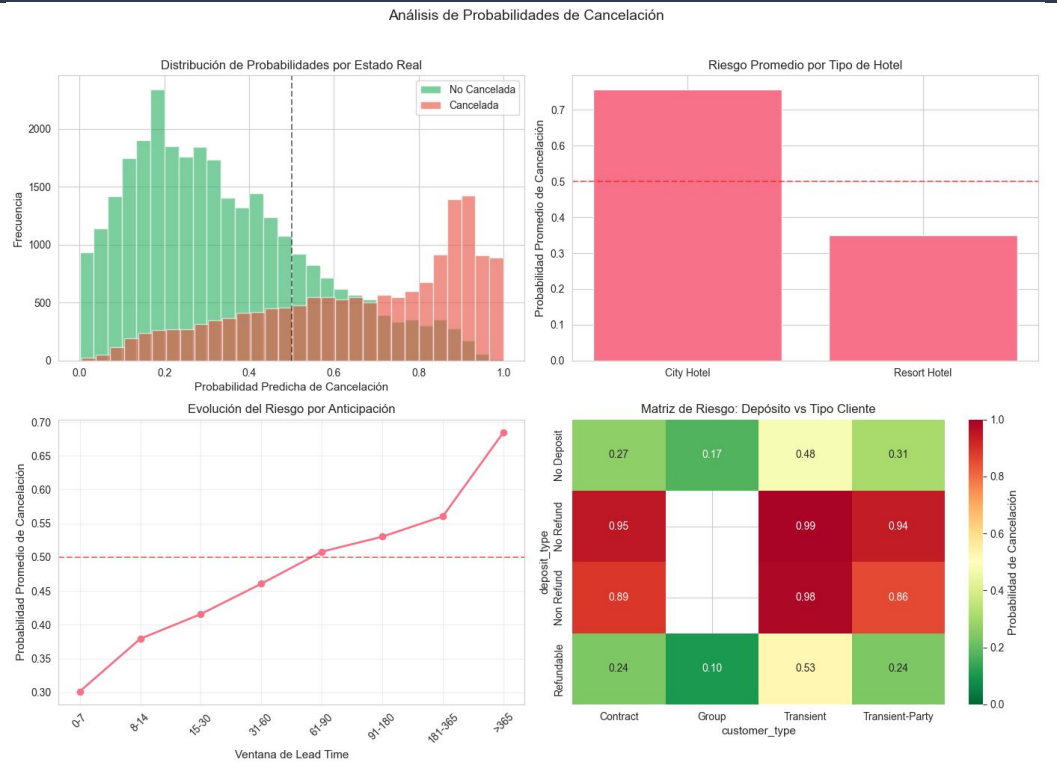
Reservas de alto riesgo: 13,511 (30.1% del total), con una tasa real de cancelación del 70.6%.

Hoteles: predominan en City Hotel (61.9%) frente a Resort (38.1%).

Depósitos: casi todas son sin depósito (92.4%).

Lead time: más frecuentes entre 91–180 días (28.7%) y 181–365 días (21.8%).

Tipo de cliente: mayoría Transient (91.9%).



Impacto económico

Impacto total

Ingresos potenciales: $19.5M

Pérdida por cancelaciones: $7.6M (39.1%)

Por tipo de hotel

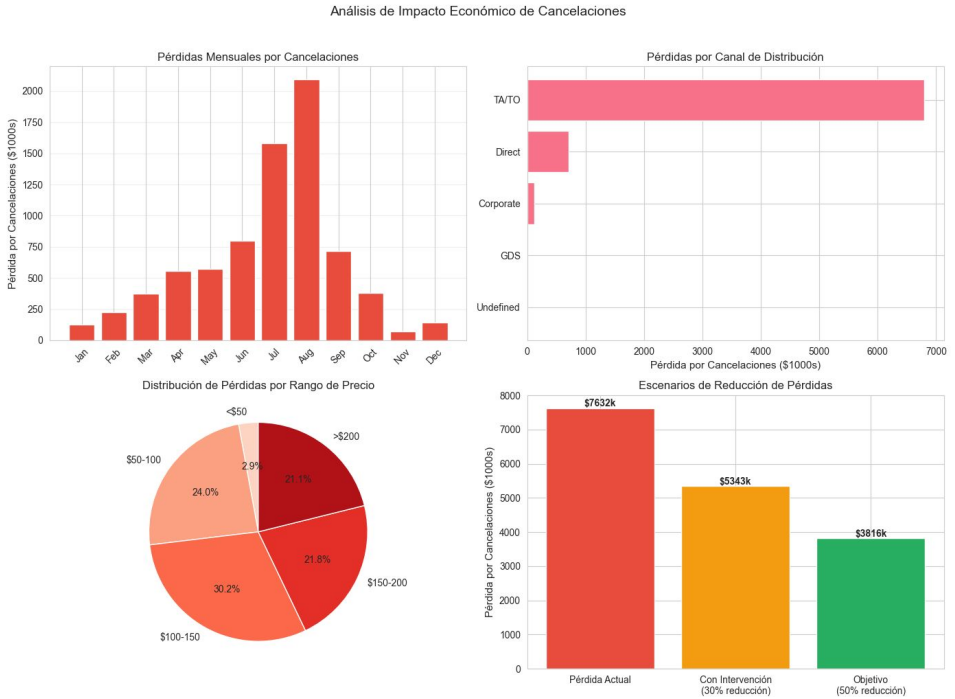
City Hotel: $2.7M perdidos → 72.2% de ingresos comprometidos

Resort Hotel: $5.0M perdidos → 31.3% de ingresos comprometidos

Oportunidad de recuperación

Con intervención en reservas de alto riesgo → ROI actual: 0x

Urge ajustar políticas, ya que sin cambios no se recupera valor



Cierre y recomendaciones

Recomendación 1

ALTA

Solicitar depósitos en reservas con más de 60 días de anticipación

Evidencia →	Las reservas con lead time >60 días muestran una tasa de cancelación ~15 pp más alta que la media.
Impacto →	Podría reducir cancelaciones en este segmento en 5-7 puntos porcentuales
Tiempo →	Implementación rápida (2-4 semanas)

Recomendación 2

ALTA

Implementar sistema de alertas para reservas con probabilidad de cancelación >60%

Evidencia →	El modelo predictivo identificó un grupo de alto riesgo con tasas de cancelación cercanas al 70%.
Impacto →	Permitiría intervenir de forma proactiva en ~30% de las cancelaciones
Tiempo →	Implementación rápida (4-6 semanas)

Recomendación 3

MEDIO

Aplicar overbooking controlado en segmentos de alto riesgo

Evidencia	En segmentos de alto riesgo, casi 1 de cada 3 reservas termina cancelándose, generando capacidad ociosa.
Impacto	Mejoraría la ocupación real en 2-4%, compensando cancelaciones esperadas
Tiempo	Implementación rápida (8-12 semanas)

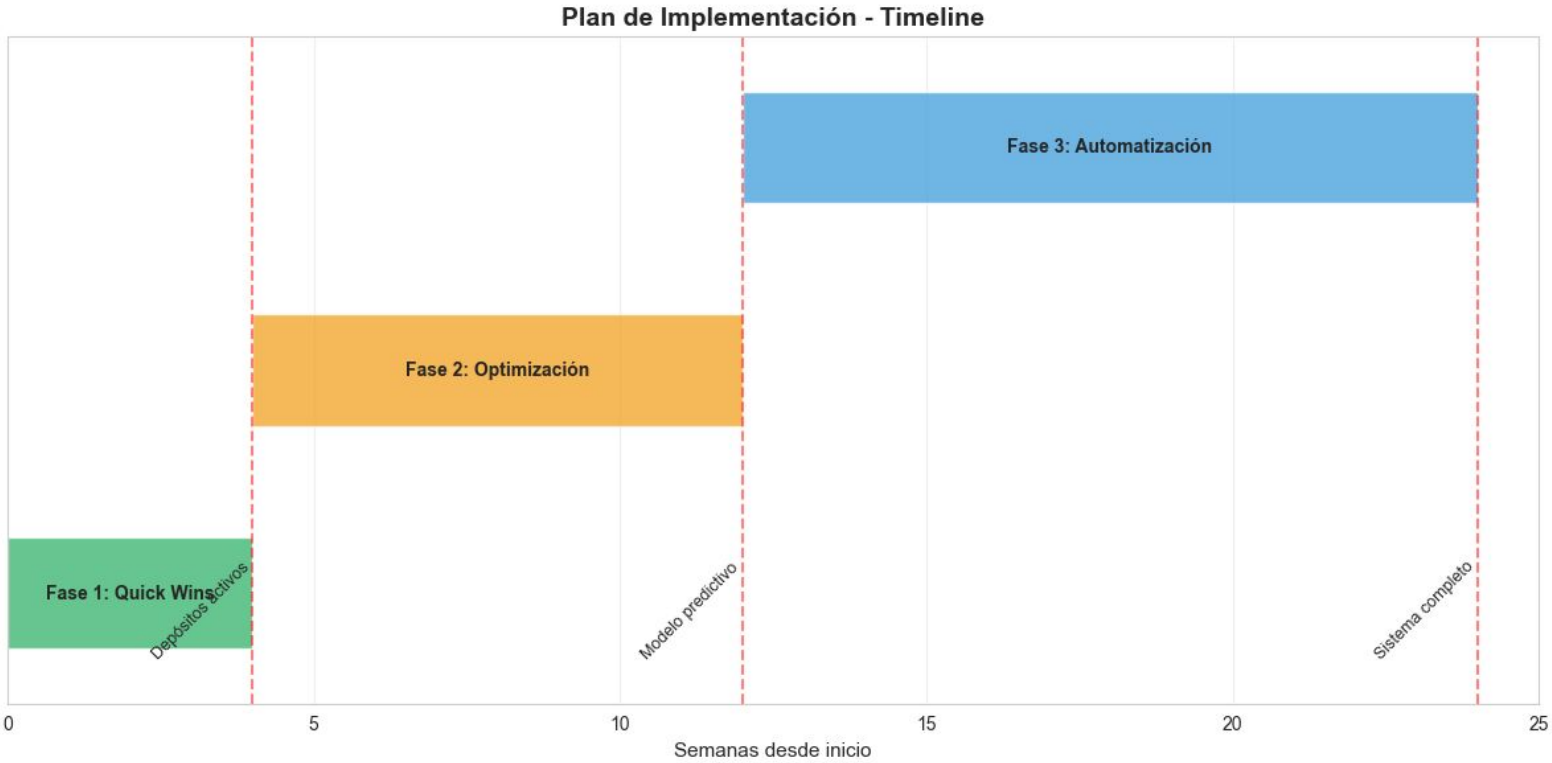
Recomendación 4

MEDIO

Ajustar precios dinámicos según lead time y canal de reserva

Evidencia →	Se observó que clientes que reservan a última hora pagan tarifas más bajas y cancelan con más frecuencia en ciertos canales.
Impacto →	Incremento de ADR estimado entre 3-5% al optimizar tarifas
Tiempo →	Implementación rápida (12-16 semanas)

Plan de Implementación



Proyección de impacto