RESERVA DE HOTELES 2015-2017

PROYECTO FINAL

Autor: Omar Fernández | Comisión data science 46275 | CoderHouse | 2023-24

“No sabemos cuándo lloverá ni cuanto caerá, pero cuando lo haga, si sabremos por donde pasará”

TABLA DE CONTENIDOS

01 INTRODUCCIÓN

Enfoque | Alcance | Audiencia | Límites

INTRODUCCIÓN

ENFOQUE

El proyecto se enfoca en el análisis de reservas de hoteles, del tipo resorts y hoteles de ciudad para lograr entrenar un modelo de machine learning que pueda predecir si las reservas serán canceladas.

ALCANCE

El conjunto de datos incluye reservas realizadas a través de diversas plataformas electrónicas entre el 1 de enero 2015 y 31 de diciembre de 2017. Cada entrada representa una reserva o cancelación.

AUDIENCIA

El análisis está destinado a las gerencias estratégicas de hoteles, con el fin de aumentar los niveles de reservas confirmadas, evitando las pérdidas ocasionadas por las cancelaciones.

LÍMITES

Se desconoce el nombre y lugar de los hoteles incluidos en el conjunto de datos. Se detalla el tipo de habitación, el número de adultos, niños y bebés, el país de origen, la fecha de llegada y la duración de la reserva, entre otros datos.

02 PROBLEMÁTICA COMERCIAL

Contexto comercial | Objetivo | Analítica

PROBLEMÁTICA COMERCIAL

¿SE PUEDEN IDENTIFICAR PATRONES ESPECÍFICOS EN LAS RESERVAS QUE PODRÍAN SERVIR COMO INDICADORES PARA PREDECIR FUTURAS CANCELACIONES?

37% CANCELACIONES

RESERVAS

NUESTRO OBJETIVO: ENTRENAR UN MODELO DE MACHINE LEARNING CAPAZ DE PODER PREDECIR SI UNA RESERVA SERÁ CANCELADA Y LOGRAR ASÍ MEJORAR LA OCUPACIÓN DE LOS HOTELES.

Análisis realizados

En el marco de este proyecto de ciencia de datos, se llevaron a cabo una serie de pruebas y análisis para profundizar en el comportamiento de los huéspedes en relación con sus reservas.

A continuación, se detallan los puntos clave de interés que se abordaron mediante el análisis exploratorio de datos:

1. Identificar el país con el mayor volumen de reservas confirmadas y elaborar un ranking de los 10 principales países de origen de dichas reservas.

2. Investigar la relación entre el promedio mensual de cancelaciones y el mes en que se realiza la reserva.

3. Analizar la relación entre el tiempo de espera y la cantidad de reservas confirmadas.

4. Determinar la distribución de la cantidad total de adultos según el mes en que se realiza la reserva.

5. Explorar la relación entre las tarifas promedio y el mes y año de llegada al hotel.

6. Identificar el tipo de servicio de comida contratado y las noches de estadía por tipo de hotel.

7. Analizar la composición de las reservas según el tipo de hotel y el perfil de los pasajeros que las realizan.

03 RESUMEN DE META DATA

Un breve resumen de los datos del Proyecto

PRINCIPALES DATOS ANÁLIZADOS

DATA SET

Observaciones
~119 K

Características
36

Datos externos
API Amadeus

SEGMENTACIÓN

8 SEGMENTOS DE MERCADO

5 CANALES DE DISTRIBUCIÓN

12 TIPOS DE HABITACIONES

148 PAISES DE ORIGEN

260 AGENCIAS

4 TIPOS DE CLIENTES

LÍMITES

Desde/Hasta
1-ene-2015 / 31-dic-2017

Cancelaciones
44 k

Reservas ejecutadas
75 k

Fuente de datos: Kaggle

DATOS EN NÚMEROS

61%

CITY HOTEL

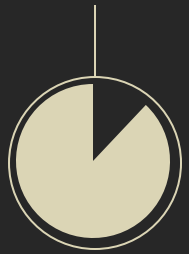
Hotel más elegido

65%

PAREJAS

Grupo predominante

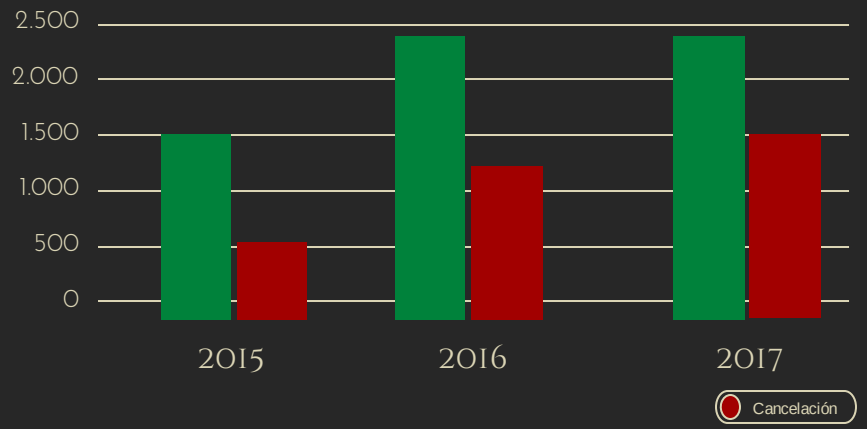
9%



GRUPO FAMILIAR

% de grupo familiar

KPI TOTAL DE RESERVAS POR AÑO* Y ESTADO



USD 103,73	3 NOCHES	2 HUÉSPEDES	USD 301
Gasto por noche	Estadía promedio	Promedio por reserva	Gasto total promedio

04 Análisis exploratorio de datos

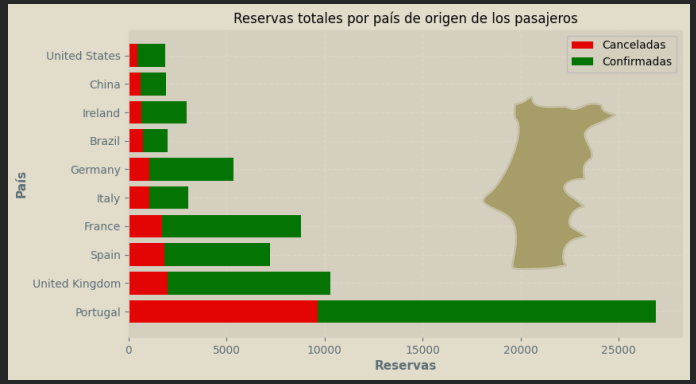
Reservas | Huéspedes | Países | Habitaciones

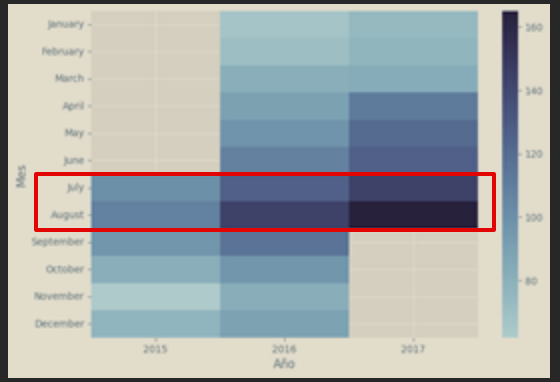
Nuevas características

Estas son las nuevas características que se generaron para comprender mejor los datos, ayudar en el modelaje y entrenamiento supervisado:

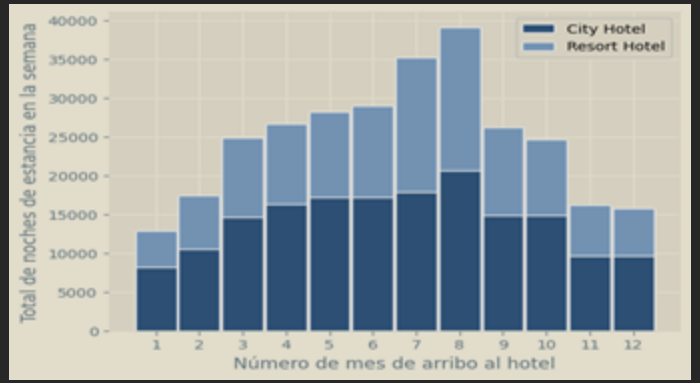
Grupo de pasajeros	Se logró segmentar a los húespedes por segmentos según cantidad y vinculo familiar.
Es grupo familiar	Identificando los grupos que tienen hijos, se logro determinar cuando es una familia.
Fecha de check-out	Se pudo calcular la fecha de check-out combinando nuevas características no existentes previamente.
Fecha arribo al hotel	Se pudo identificar la fecha de arribo al hotel combinando datos existentes en la reserva.
Estadía total	Calculada como el total de noches en el hotel al sumar dos características existentes en el set de datos.
Importe total	Se calculó multiplicando la nueva característica 'Estadía total' por el valor del ADR existente.
Total huéspedes	Se calculó el total de personas que integran una reserva.

Dentro del Top 10, los países de Europa occidental domina los 7 primeros puestos. Sorprende que EEUU no sea uno de los líderes.

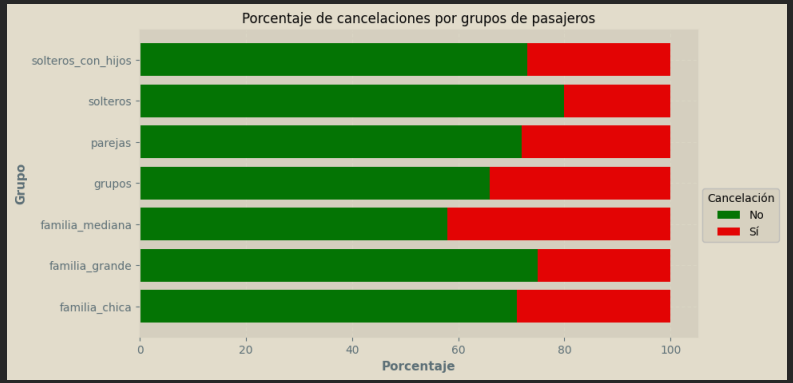




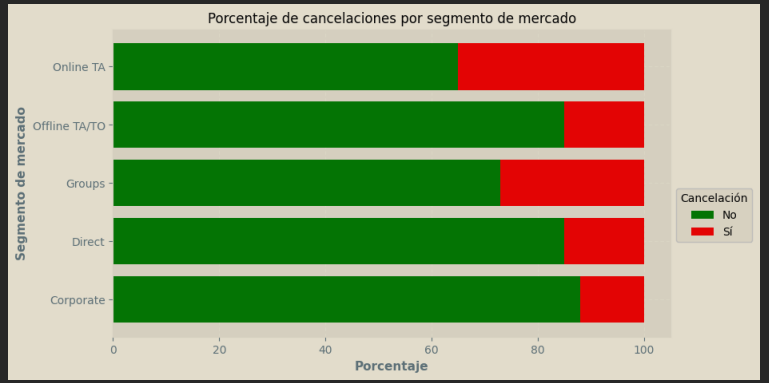
Las estadías en semana en hoteles de ciudad dominan cada mes, pero se comprueba una paridad durante los meses de julio y agosto, cuando los resorts tienen picos de demanda.



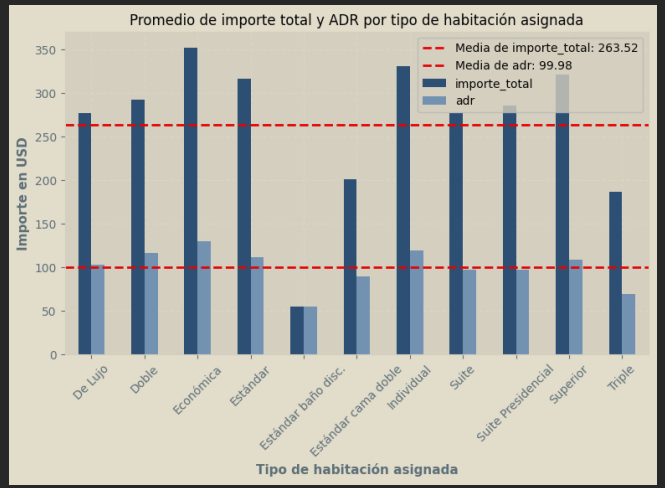
Se establece un claro incremento en los valores promedio año a año. Al comienzo del verano europeo, los valores evidentemente se incrementan para luego volver a bajar con el fin de este



<< Familia mediana es el grupo con mayor porcentaje de cancelaciones. Online TA >> lo es dentro de segmentos de mercado.



El valor de una habitación económica se ubica por encima de la media, pero resulta ser la que genera mayores ingresos.



Se pudo geolocalizar a los huéspedes según su país de origen. Vemos una concentración muy marcada en Europa como principal emisor de turismo.

05 Insights

Principales datos obtenidos

Los clientes solicitan reservar una suite presidencial en un ~65% de los casos, pero solo la obtienen en el ~53% de las reservas. La habitación más rentable es la económica. Pudiéndose realizar la reserva durante el mes de septiembre, se logra un ahorro promedio de 35 dólares versus el mes de agosto.

HABITACIÓN

assigned_room_type	adr
Económica	130.436456
Individual	119.154147
Doble	116.604846
Estándar	111.903104
Superior	108.884970
De Lujo	103.348634
Suite	97.859859
Suite Presidencial	97.428126
Estándar cama doble	89.656825
Triple	69.648275
Estándar baño disc.	54.891628

SERVICIO

El 77% de las reservas se efectúan con "cama y desayuno".

AGENCIAS

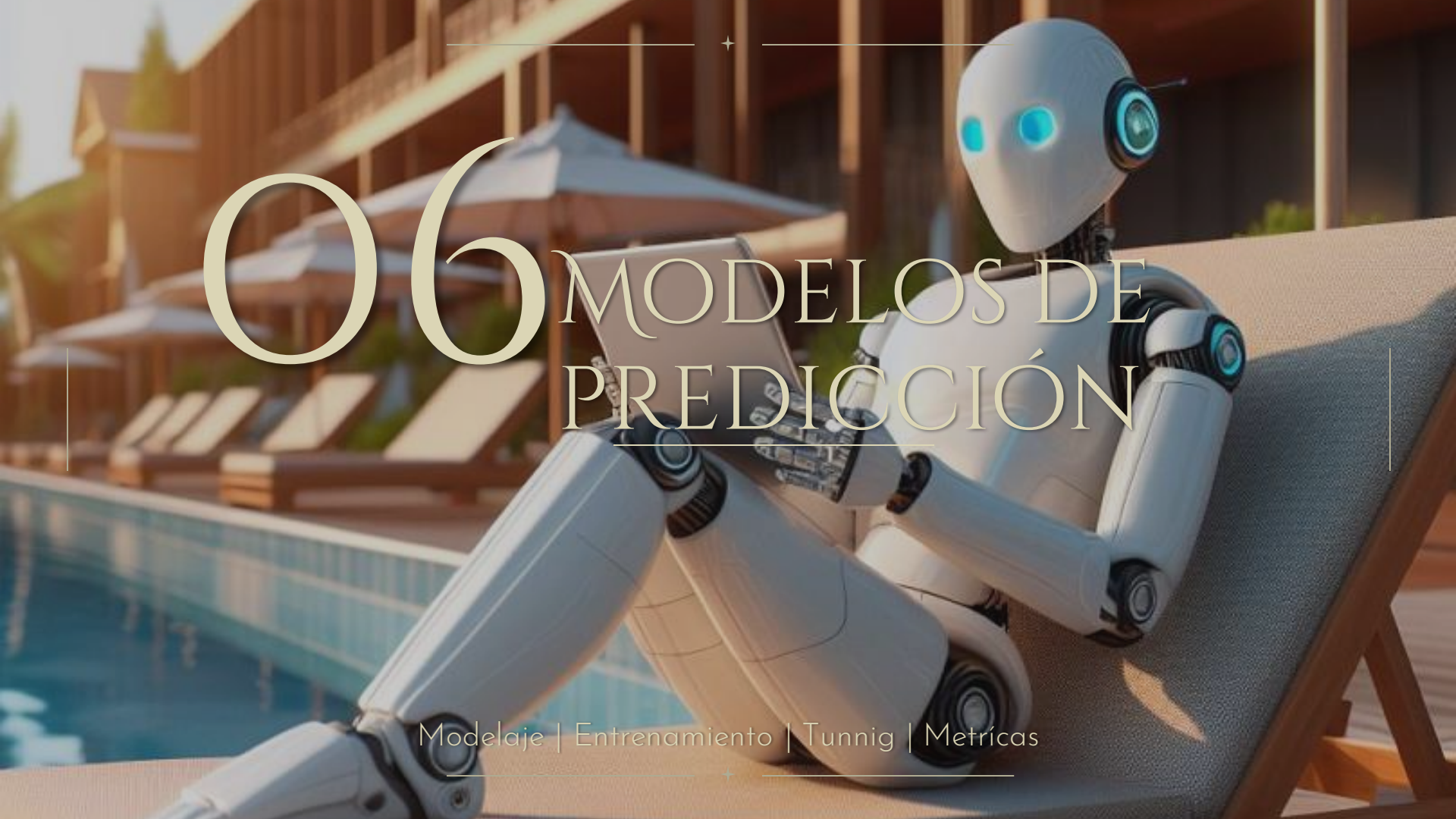
La tasa de cancelaciones en TA Online es del 35%, la mayor en los segmentos analizados.

SEGMENTO

El segmento corporativo representa el 12% de cancelación de reservas, siendo el más bajo.

ADR

Los precios más altos se abonan durante los meses de julio y agosto, mientras que entre noviembre y febrero se encuentran las bandas de precios más bajos



06 MODELOS DE PREDICCIÓN

Modelaje | Entrenamiento | Tunnig | Metrícas

PERFORMANCE DE LOS MODELOS PROBADOS

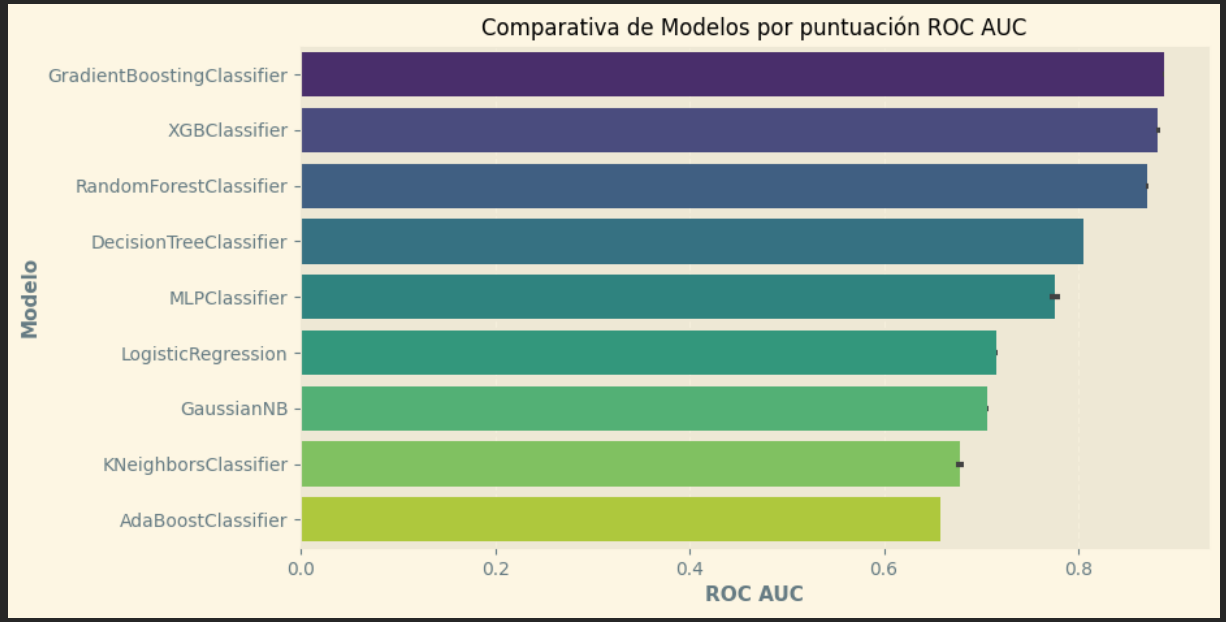
Evaluación robusta del rendimiento de modelos con curva ROC y validación cruzada K-Fold
En este análisis:
Se empleó la curva ROC como métrica para validar el rendimiento de los modelos entrenados.

Se evaluó el desempeño de cada modelo utilizando validación cruzada K-Fold, permitiendo una evaluación robusta y confiable.

85,900 REGISTROS

Set de entranamiento
70%

Set de pruebas
30%



También se midieron: ACCURACY | PRECISION | RECALL | F1-SCORE

MEJORES MODELOS ENTRENADOS

XGBCLASSIFIER

AUC 0,8854

colsample_bytree = 0.6
criterion= "entropy"
gamma= 0.5
max_depth= 5
min_child_weight= 1
subsample= 0.6

ACCURACY: 81,82%

GRADIENT BOOSTING CLASSIFIER

AUC 0,8888

Mejores parámetros:
cv folds= 10
max_depth= 7
n_estimators= 200

ACCURACY: 82,40%

RANDOM FOREST CLASSIFIER

AUC 0,8722

class_weight = "balanced"
max_features = "auto"
n_estimators= 40

ACCURACY: 81,38%

07 RECOMENDACIONES FINALES

Enfoque | Alcance | Audiencia | Límites

CONCLUSIÓN & RECOMEDANCIONES

Al concluir este proyecto de data science, podemos dar una serie de recomendaciones finales a fin de en un futuro lograr mejores resultados a los actualmente alcanzados.

- Los datos originales presentaban una serie de desafíos de calidad, incluyendo valores duplicados, ausencia de datos, errores de formato y discrepancias con la realidad. Estos obstáculos fueron abordados durante la fase de data wrangling, la cual consumió una parte significativa del tiempo del proyecto, pero resultó fundamental para mejorar la predicción del modelo.
- Durante este proceso, se identificaron y se crearon nuevas y valiosas características que enriquecieron la comprensión de las reservas. Esto permitió una visión más completa y detallada de los patrones y tendencias en los datos.
- Nuestro análisis sugiere que sería beneficioso optimizar la gestión de las campañas de marketing, anticipando las ofertas para los meses de diciembre y enero, períodos en los que los ingresos por habitación disponible (ADR) tienden a ser más bajos. Esto podría aumentar la efectividad de las campañas y mejorar la rentabilidad.
- A pesar de la muestra sesgada hacia ciertos países, se logró identificar aquellos en los que existe una brecha significativa entre las reservas realizadas y las cancelaciones totales. Este hallazgo ofrece oportunidades para mejorar las estrategias de retención y fidelización de clientes en dichos mercados.
- Se recomienda realizar una auditoría exhaustiva de los sistemas de captura de datos originales para garantizar la entrega de datos limpios y preformateados al modelo de machine learning. Esto contribuirá a mejorar la precisión y la confiabilidad del modelo, lo que resultará en predicciones más precisas y útiles.
- Mirando hacia el futuro, consideramos que sería altamente beneficioso contar con datos más detallados, como la ubicación del hotel, la edad y el género de los huéspedes, el propósito del viaje (negocios o placer) y si es la primera vez que visitan el hotel. Estos datos adicionales podrían enriquecer aún más nuestro análisis y mejorar la capacidad predictiva del modelo.
- Se podría realizar un ensamble con los modelos seleccionados, con el fin de mejorar aún más la predicción.

GRACIAS

POR SU ATENCIÓN

¿Tiene alguna consulta?

cvomarfernandez@gmail.com

Para seguir en contacto:

https://www.linkedin.com/in/omarfernandez/