Data Science Proyect: Hotel Booking Cancellations

Trabajo Final

Evelyn Delacoste • Agosto 2023

Temas

Temática de Negocio

Data Analizada

Preguntas de Análisis

EDA: Análisis Exploratorio de Datos

Insigths

Modelado ML Clasificación

Conclusiones Finales

Temática de Negocio

Hotel Booking Cancellations: ¿Cuáles son las causas por las cuales un cliente cancela una reservación?

El presente proyecto apunta a analizar una base de datos de las reservas realizadas a lo largo de 2 años en dos hoteles de una misma cadena, con la finalidad de detectar patrones de comportamiento o acciones del hotel que pudieran generar la cancelación de una reserva.

Este análisis puede ser de interés para cualquier propietario de la industria hotelera que desee mejorar el rendimiento de su alojamiento y así afianzar clientes.

Data Analizada

El dataset analizado contiene información de las reservas de 2 tipos de hotel de una misma cadena: **City Hotel y Resort Hotel**.

Cada registro tiene detalles sobre la solicitud y si la reserva fue efectiva o no.

Ubicación

Portugal

Período

Jul 2015 - Ago 2017

Variables disponibles

35

Reservas analizadas

119.390 registros

Algunas variables disponibles:

- Is canceled: 0 no / 1 si
- Tipo de hotel: City o Resort
- lead time: tiempo reserva/arribo
- arrival_date...: fecha de arribo
- stay: estadías
- days_in_waiting_list:
- ...

Preguntas de Análisis

1. ¿Cuál es el hotel más demandado y cuál posee mayor % de cancelaciones?
2. ¿Cuáles son las épocas del año más demandadas? ¿Qué sucede con las estadías?
3. ¿Cómo se comporta el cliente con respecto al tiempo previo de arribo al hotel? ¿ Tiene que ver con la Nacionalidad o Composición Familiar?
4. ¿ El hotel posee clientela fija? ¿Cómo se comportan aquellos clientes frecuentes?
5. ¿Cómo gestiona el hotel las reservas?
6. ¿Cómo son las tarifas de los hoteles?

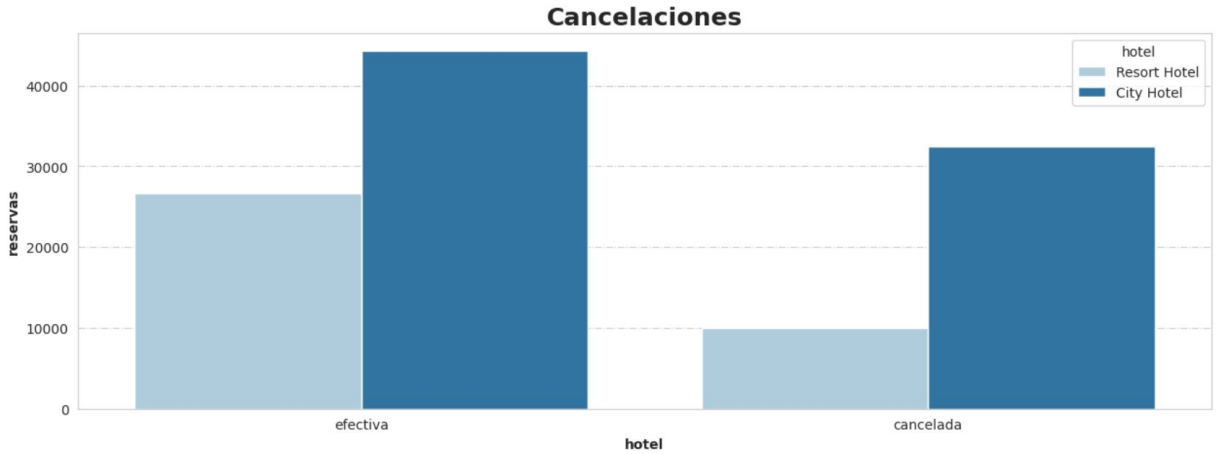
EDA: Análisis Exploratorio de Datos

¿Cuál es el hotel más demandado y cuál posee mayor % de cancelaciones?

De toda la data analizada, sólo el **37 %** de las reservas son **canceladas**. Mientras que el **67 %** de todas las reservas disponibles corresponden al **City Hotel.**

Las cancelaciones en base al tipo de hotel:

- **City Hotel:** 40 %
- **Resort Hotel:** 27 %



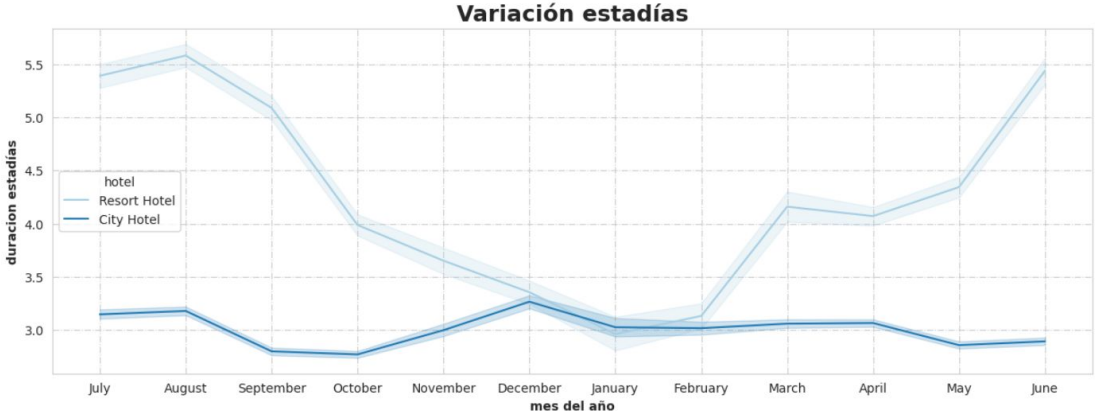
¿Cuáles son las épocas del año más demandadas? ¿Qué sucede con las estadías?

Mayor demanda: Agosto (verano)

Menor demanda: Enero (invierno)

Ambos hoteles tienen los mismos meses de mayor y menor demanda

No hay ningún mes donde cambie el comportamiento o relación de las cancelaciones con respecto a las reservas



Estadías Promedio

General: 3 días

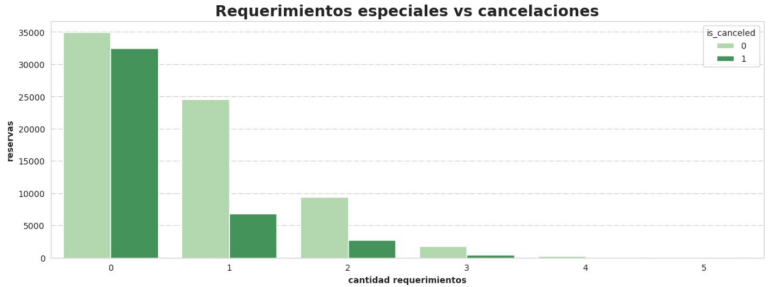
Resort Hotel: 4 días

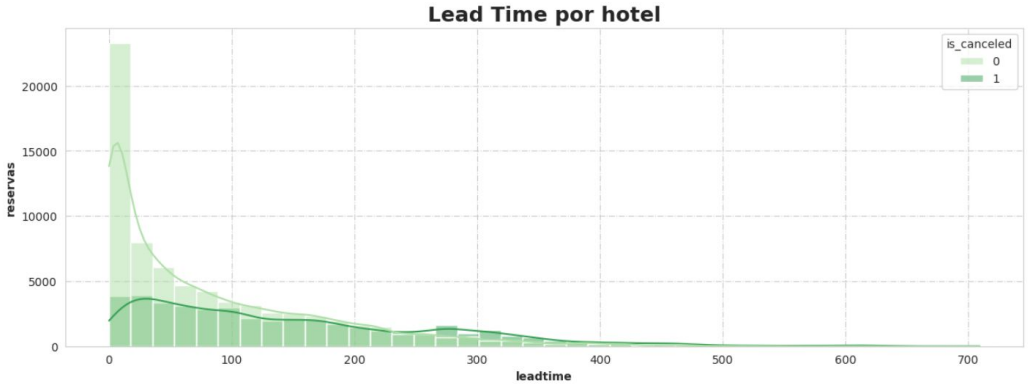
City Hotel: 3 días

Las estadías del Resort se ven afectadas por el mes del año

¿Cómo se comporta el cliente con respecto al tiempo previo de arribo al hotel? ¿ Tiene que ver con la Nacionalidad o Composición Familiar?

- Principalmente adultos
- Reservas de pocas personas (1-2)
- Las reservas se suelen realizar mediante agentes turísticos
- Es mayor el % de visitantes extranjeros

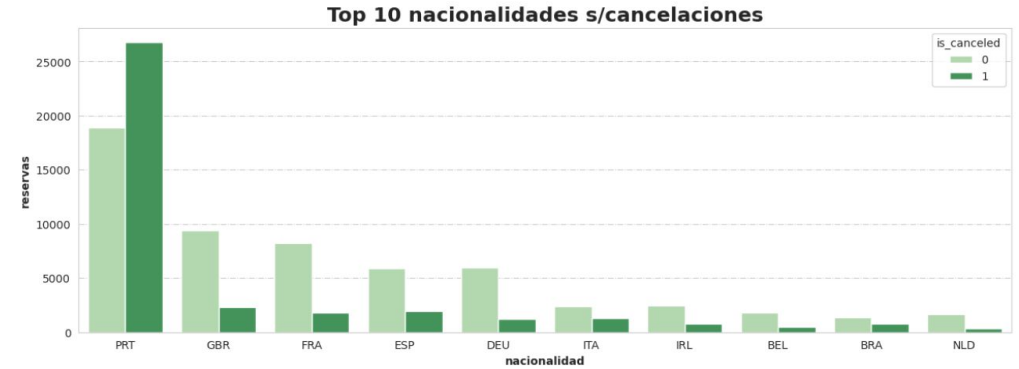


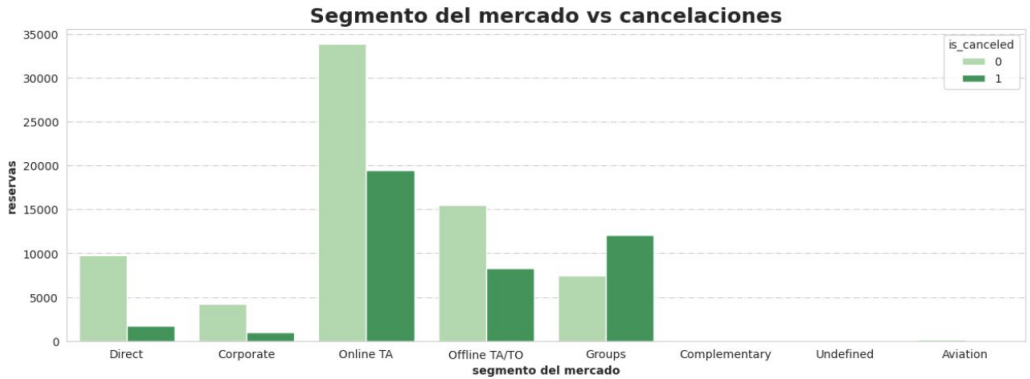


Si el cliente pide requerimientos especiales, disminuyen las cancelaciones

A medida que aumenta el tiempo de espera, aumenta el % de cancelaciones sobre las reservas

¿Cómo se comporta el cliente con respecto al tiempo previo de arribo al hotel? ¿ Tiene que ver con la Nacionalidad o Composición Familiar?

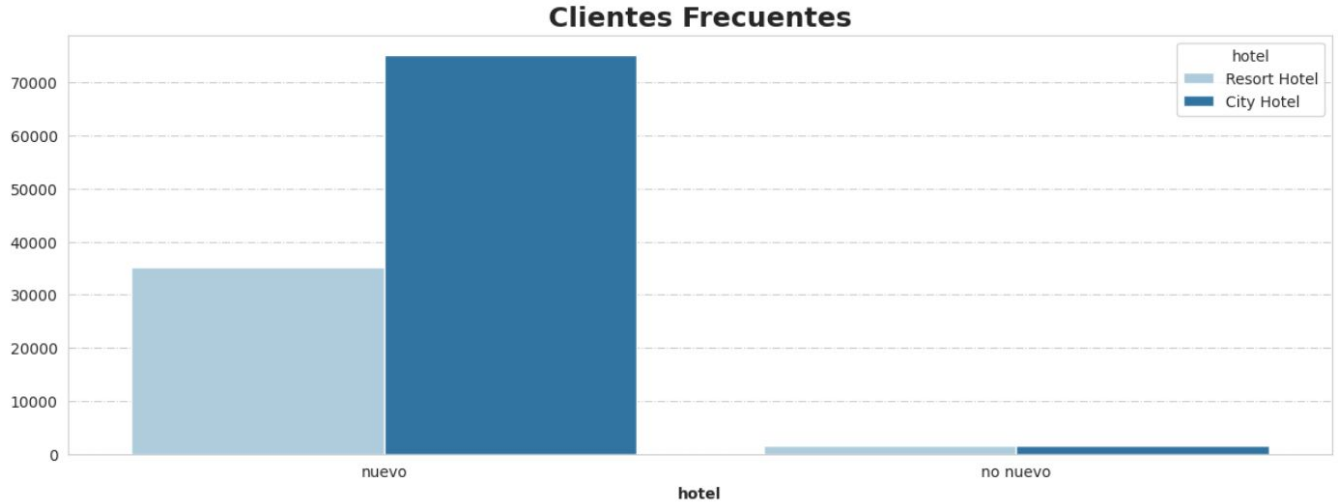




Cuando las reservas son de clientes internacionales o de grupos de personas; las cancelaciones aumentan con respecto a las reservas efectivas y cambia el comportamiento que se observa en los otros parámetros

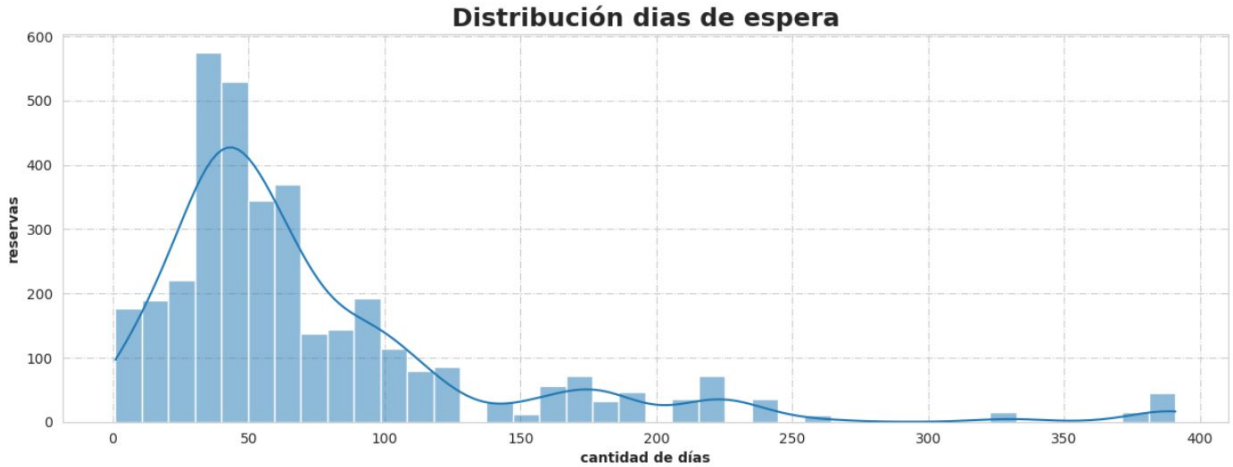
¿ El hotel posee clientela fija? ¿Cómo se comportan aquellos clientes frecuentes?

El **96 %** de las reservas corresponden a **nuevos clientes**.
Teniendo en cuenta que tiene un público nacional, sería necesario verificar la satisfacción de los clientes y porque no regresan al mismo.



¿Cómo gestiona el hotel las reservas?

El **80 %** de las reservas **no presentan cambios** en base a lo solicitado por el cliente, incluyendo requerimientos especiales, tipo de habitación y duración de la estadía.



En cuanto al tiempo de espera de confirmación de la reserva, se observan valores muy disparejos lo que da indicio de desorganización en la gestión de las reservas.

¿Cómo son las tarifas de los hoteles?

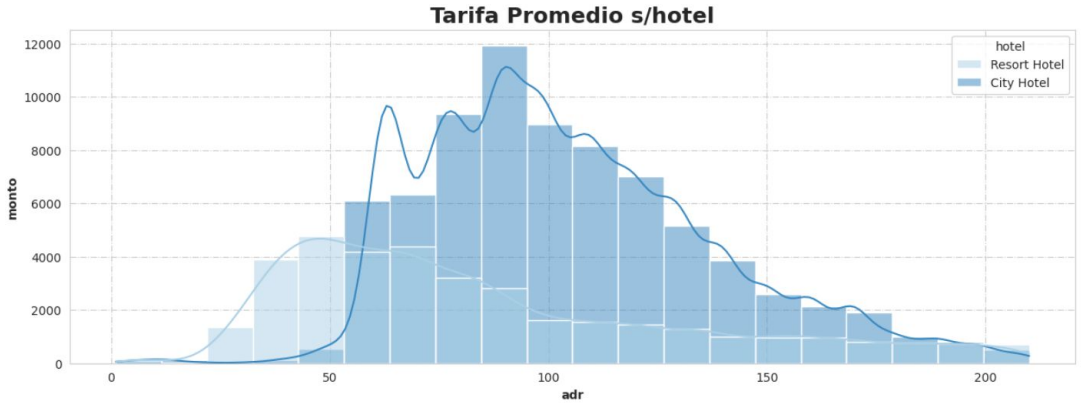
La tarifa promedio es de 100 USD por noche.

La misma aumenta a medida que el cliente indica requerimientos especiales en la reserva.

City Hotel tiene tarifas más elevadas que el Resort.

Las habitaciones más caras son la G y H y a su vez son las que presentan mayor aumento en relación a los requerimientos especiales.

La habitación más económica y la más reservada es la A que posee muchos valores fuera del rango promedio; se le puede atribuir a la solicitud de requerimientos especiales.



Insights

A lo largo de 2 años, el City Hotel es el más demandado; se reserva por estadías cortas y el precio es más elevado.

El Resort Hotel tiene menor demanda pero posee estadías más largas en los meses de verano.

En promedio general, el 60-70 % de las reservas no se cancelan.

Cuanto menor es la anticipación con la que el cliente reserva y mayor el número de requerimientos especiales, es menos probable que se cancele dicha reserva.

El comportamiento de las cancelaciones se mantiene estable a través de las distintas variables, exceptuando la nacionalidad y los viajes grupales.

Los hoteles poseen un % muy bajo de clientes frecuentes, por lo que es importante destacar que se debe verificar el nivel de satisfacción de los mismos.

Modelado Machine Learning Clasificación

Selección de variables

Se aplicaron diferentes modelos de machine learning supervisados de clasificación; con la intención de predecir la variable **“is_canceled”** que indica si la reserva fue efectiva o cancelada.

Las variables de entrada al modelo fueron 17 en total ; dentro de las más importantes, se encuentran:

- lead time : tiempo de arribo
- country : nacionalidad del cliente
- market segment : segmento del mercado que reserva
- total of special requests : cantidad de requerimientos especiales

Modelos considerados y métricas de evaluación

A partir del EDA, se detectó que el dataset analizado tenía un desbalance/ sesgo con respecto a la cantidad de reservas canceladas: sólo el 37 % del total analizado correspondía a cancelaciones.

Por ello, para evaluar el rendimiento se estudió en detalle el desempeño de las métricas: **precision y recall**, y por defecto **f1-score** apuntando específicamente en la prediccion de cancelaciones.

Se realizó el modelado de los etapas:

- 1er etapa: se probaron modelos de árboles de decisión, regresión logística y vecinos más cercanos (KNN) ponderando el desbalance de los datos.
- 2da etapa: a partir de la selección de dos modelos de la primer etapa, se optimizaron hiper parámetros y se probó un modelo de ensamble con Random Forest

Resultados de la modelación

En total, se probaron **8 modelos de clasificación** supervisados.

Ningún modelo indicó un accuracy mayor a 0,80 ni un f1-score mayor a 0,65

El mejor modelo fue Random Forest con 56 predictores. Sus métricas fueron las siguientes:

Classification Report (Test Set):

	precision	recall	f1-score	support
Efectiva	0.81	0.94	0.87	17218
Cancelada	0.74	0.42	0.54	6690
accuracy			0.80	23908
macro avg	0.77	0.68	0.70	23908
weighted avg	0.79	0.80	0.78	23908

Con una precisión de 73 % y un **recall de 44 %** indica que si bien predice con un buen desempeño aquellas reservas canceladas; hay un **porcentaje considerable de reservas que serán canceladas y que escapan de la detección del modelo.**

Conclusiones Finales

A partir del análisis de hiper parámetros y aplicación de modelos de ensamble; se concluye que el mejor modelo es un **Random Forest con 56 árboles de decisión como estimadores.**

Si es importante recalcar que no es un modelo de presente alta precisión a la hora de detectar que reservas son canceladas, las causas de este comportamiento son a raíz del desbalance que se posee en los datos originales.

Como soluciones para mejora, se plantea:

- Aplicar al modelo existente planteado, mayor peso a la clase 1: reservas canceladas.
- Aumentar el volumen de datos con el fin de minimizar el desbalance de los mismos.