HospeDate S.A.

Proyecto Final de Data Science para CoderHouse – Marzo 2023

Autores: Kevin Cruz – Mauro Nicolás Rivero

Email: mauro1992351916@gmail.com

GitHub: https://github.com/MauroNicolasRivero/Data-Science

LinkedIn: www.linkedin.com/in/mauro-nicolas-rivero

Contexto Comercial

HospeDate SA es una cadena portuguesa que administra hoteles, a través de su área de contabilidad pidió recabar información sobre las reservas hechas en sus establecimientos, para llevar a cabo un análisis completo con el fin de detectar oportunidades para aumentar sus ingresos y conocer las preferencias de sus clientes.

Problema Comercial

El rubro hotelero está sujeto a varios eventos externos que hacen complejas las tomas de decisiones como la competencia, clima meteorológico y la estacionalidad, pero hay otros eventos que podemos llegar a prever como las **Cancelaciones de reservas** que afectan la explotación de las habitaciones, es por eso que vamos a concentrarnos en este fenómeno.

Definición del Objetivo

Como objetivo principal nos proponemos desarrollar un modelo de Machine Learning que pueda predecir qué porcentaje aproximado de reservas serán canceladas.

Como objetivo secundario y en base al porcentaje arrojado por el modelo, decidir de qué forma se va a trabajar ese resultado, para maximizar las ganancias al sobrevender habitaciones susceptibles a ser canceladas en una primera instancia o minimizar las pérdidas tratando de conservar al cliente ofreciéndole mejoras en los servicios

Contexto Analítico

Disponemos de un dataset con las reservas hechas entre los meses de Julio del 2015 y Agosto del 2017, son unos 120.000 datos aproximadamente divididos en 32 columnas con datos cuantitativos y cualitativos. Los datos fueron obtenidos de la plataforma *Kaggle* y subidos a Github para que sea más fácil y rápido su acceso a todas las personas que estén interesadas, no obstante también dejamos aquí el link

Fuente: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Breve Descripción de las Variables que vamos a utilizar:

1. **Hotel**: *Tipos de hoteles*
2. **Is_canceled**: *Muestra si fue o no cancelada la reserva*
3. **Arrival_date_week_number**: *Muestra la semana en número en la cual llegarán los huéspedes*
4. **Meal**: *Comidas ofrecidas por el hotel*
5. **Country**: *País procedente de los huéspedes*
6. **Market_Segment**: *Segmento del mercado que operó en la reserva*
7. **Is_repeated_guest**: *Si el cliente ya ha reservado en alguna otra ocasión*
8. **Previuos_cancellations**: *Si el cliente ha incurrido en una cancelación anteriormente*
9. **Reserved_room_type**: *Tipo de habitación reservada*
10. **Assigned_room_type**: *Tipo de habitación asignada*
11. **Deposit_type**: *Muestra si han hecho un depósito previo o no*
12. **Customer_type**: *Muestra que tipos de clientes son*
13. **Adr**: *Es una métrica que muestra la tarifa promedio por habitación*
14. **Required_car_parking_spaces**: *Indica si el huésped solicitó lugar en el estacionamiento*
15. **Total_of_special_requests**: *Indica la cantidad total de solicitudes que hizo el huésped*
16. **Reservation_status_date**: *Fecha de la última actualización de la reserva*
17. **People_sum**: *Sumamos las distintas variables que representan los huéspedes en una sola columna*
18. **Days_Sum**: *Sumamos los días reservados ya sean días entre semana o fin de semana*
19. **Date_arrive**: *Fecha de llegada declarada en la reserva*
20. **Day_dif**: *Representa cuantos días antes de la fecha de arribo se produjo la cancelación*

Preprocesamiento de los datos:

El dataset contaba con demasiados datos nulos en 2 variables las cuales decidimos cortarlas por completo ya que para nosotros no eran de relevancia para el modelo y otras 2 variables con pocos datos nulos pero que al ser difícil su estimación decidimos eliminar dichas filas. También creamos nuevas variables a partir de la información brindada por el archivo original con el fin de concentrarla en una sola columna y hacer más eficiente su análisis.

Cambios aplicados al Dataset Original:

1° Usando la librería Missingno graficamos los valores nulos del dataset original en formato Matrix

2° Creamos la variable People_Sum sumando las columnas adults,children y babies

3° Creamos la variable Days_Sum sumando las columnas stays_in_weekend_nights y stays_in_week_nights.

4° Creamos la variable Date_arrive sumando las columnas arrival_date_year, arrival_date_month y arrival_date_day_of_month las cuales con anterioridad fueron cambiadas los tipos de datos para permitir la concatenacion como Strings.

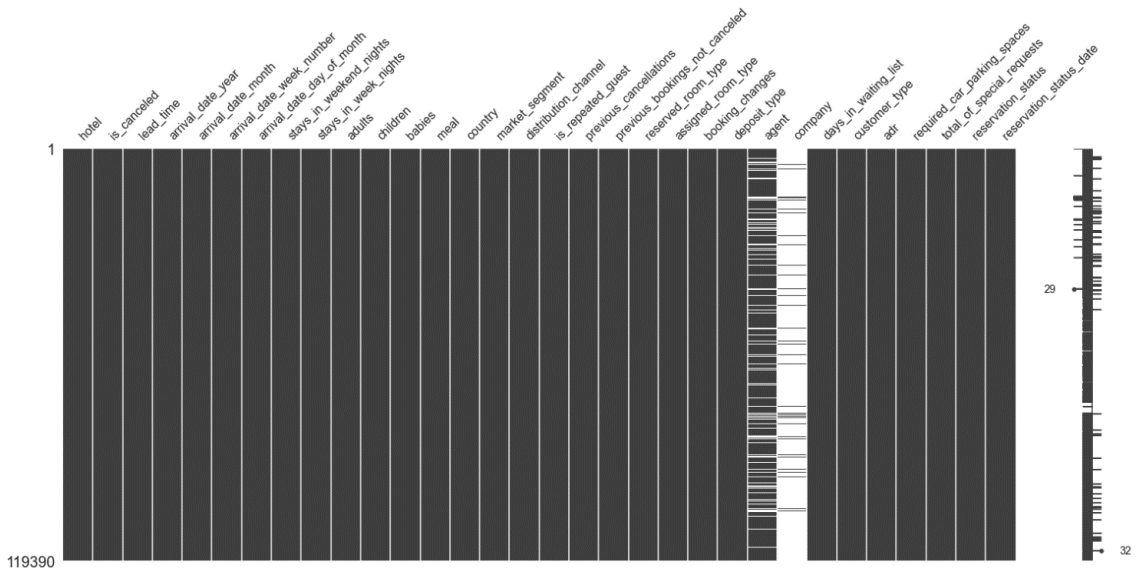
5° Creamos la variable Day_dif restando las fechas Date_arrive menos reservation_status_date para conocer cuantos días antes a la fecha de arribo se producia la cancelación, también se le cambio el tipo de dato para luego incluirla como variable numérica en el modelo.

6° Cambiamos el tipo de dato a las columnas que tenian información sobre fechas.

7° Eliminamos las variables que no vamos a tener en cuenta para el modelo

8° Eliminamos los valores en cero que no nos interesa tener en el dataset

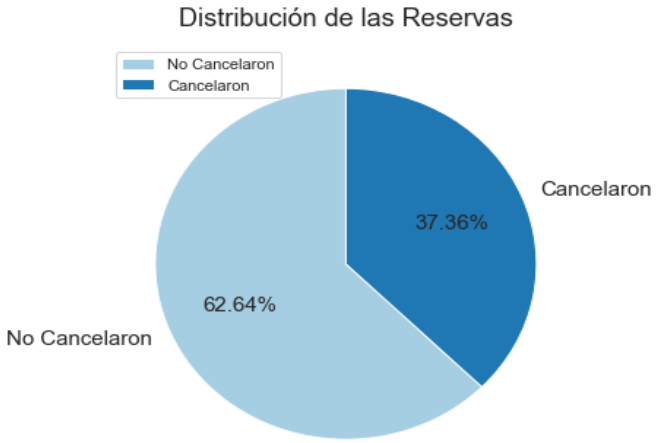
Gráfico de datos ausentes en el data set original usando “Missingno”

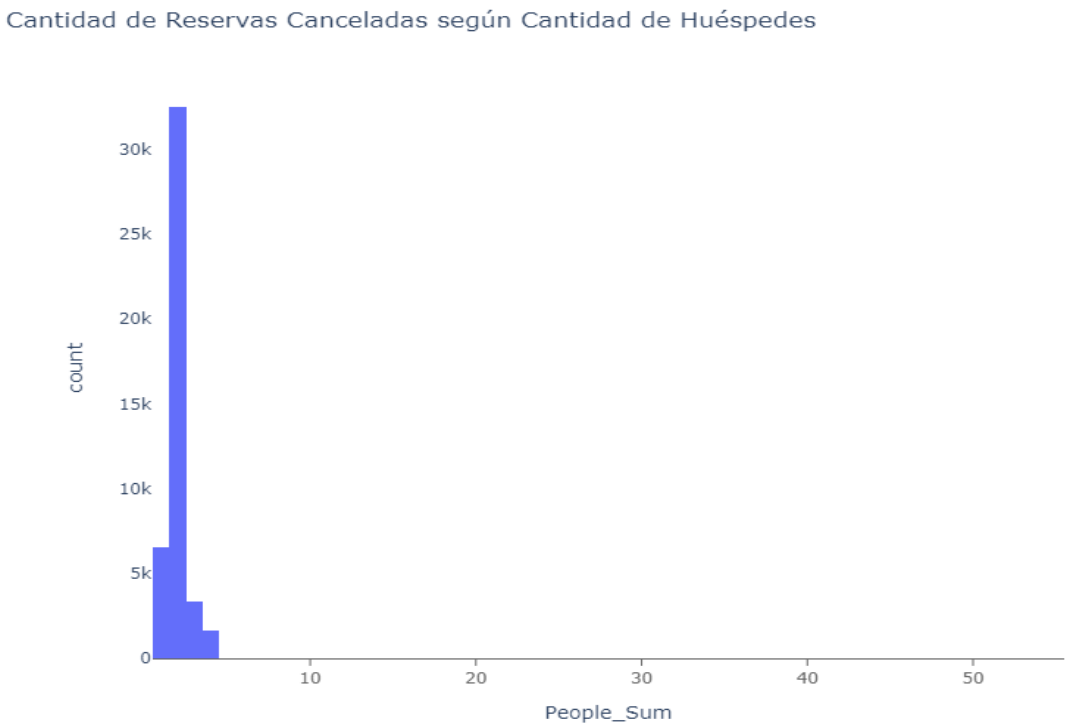


Analisis Exploratorio de Datos:

Aquí mostraremos algunos graficos elaborados a partir de la informacion brindada por el dataset junto con una breve descripción como una forma mas clara de analizar los datos y al final haremos una conclusión general

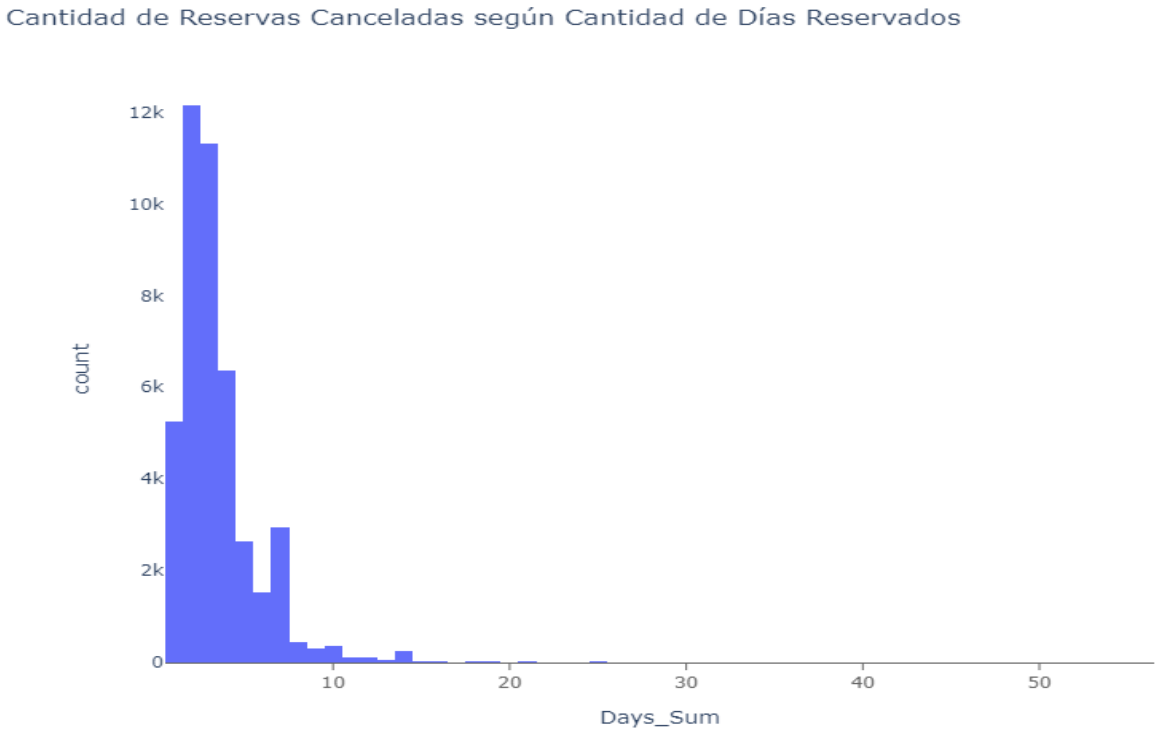
El 37 % de las reservas hechas fueron canceladas posteriormente, en los siguientes gráficos buscaremos detectar un patrón entre los clientes que cancelaron.



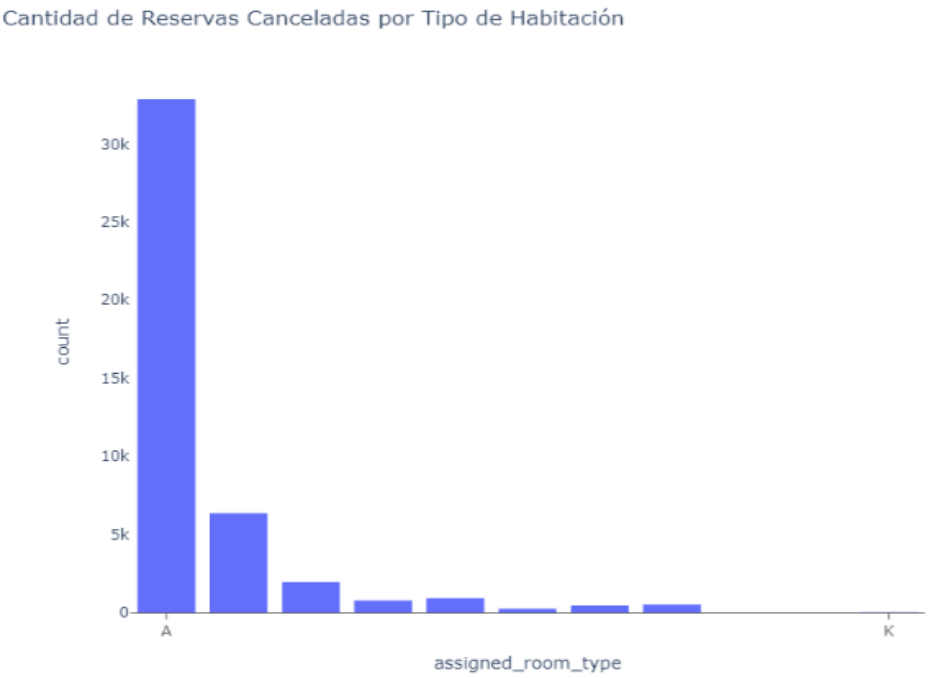


La cantidad de huéspedes que más frecuentemente se dio en las reservas que luego fueron canceladas corresponden a 2 personas adultas

Como podemos observar en el gráfico inmediatamente superior Portugal es el país con mayor cantidad de reservas canceladas, esto se debe a que los datos de este estudio provienen de hoteles ubicados en ese país por lo que hay una clara tendencia del turismo local.

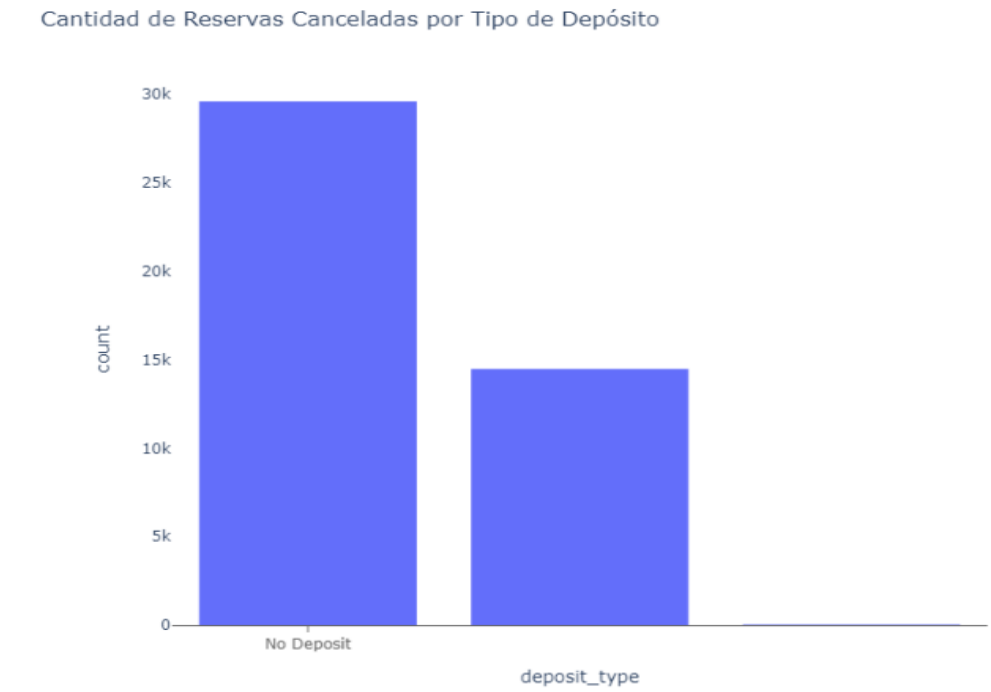


La cantidad de días elegidos con mayor frecuencia fueron entre 2 y 3 días, lo que marca una preferencia por los viajes cortos usando los fines de semana sumado a algún día festivo que permita extender 1 día más la estadía

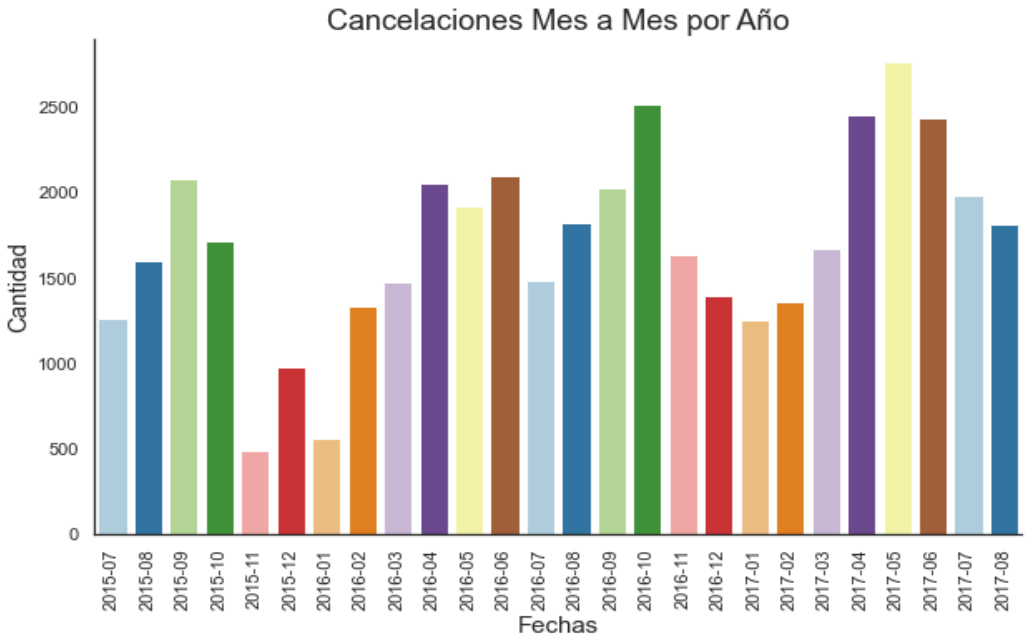


A la hora de elegir en que clase de habitación hospedarse la mayoría eligió la clase “A” como preferida siguiéndole la clase “D” y “E”

En este gráfico podemos ver como se distribuyeron las cantidades de reservas canceladas en relación a los tipos de hoteles, siendo el tipo “City” el tipo de hotel con más cancelaciones versus el tipo “Resort”.



El gráfico nos muestra la distribución de las reservas que fueron canceladas según si hicieron o no depósito previo, en el caso de que si lo hayan hecho el mismo es no reembolsable. Como vemos el 66 % de las reservas canceladas no hicieron depósito previo y el restante 34 % si lo habían hecho.



Este gráfico nos permite reconocer la estacionalidad del turismo, indicandonos los meses de mayor afluencia de turistas en los meses calidos del verano europeo, donde efectivamente se incrementan también las cancelaciones de reservas

Perfil de Reserva Susceptible de Cancelación:

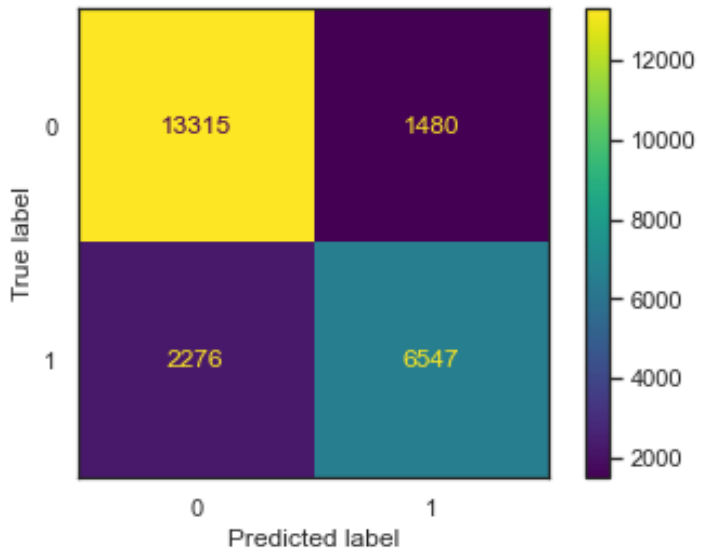
Reservas de 2 personas, con residencia en el mismo país, de 2 a 3 noches, en habitaciones clase "A" en hoteles tipo City y no suelen hacer depósitos previos. Todo esto ocurre con más frecuencia en temporada alta de reservas que va desde abril hasta septiembre

Por último, sabemos que el promedio de cantidad de días que transcurren desde la futura fecha de arribo elegida por el huésped hasta la fecha de cancelación es de 86 días, lo que nos otorga un rango estimado desde el cuál debemos estar atentos a una posible cancelación.

Elección de Modelos de Machine Learning:

Random Forest Classifier optimizado con Grid SearchCV:

La matriz de confusión nos permite visualizar de forma simple los resultados obtenidos luego de correr el algoritmo optimizado



Conclusiones - Random Forest Classfier:

1° La **Exactitud** del 84 % es una buena estimación, nos dice que el porcentaje de elementos clasificados correctamente es alto. En nuestro caso en particular nos indica que sobre 23618 reservas, 19862 fueron correctamente clasificadas ya sea que efectivamente fueron canceladas (6547 VP) como las que no lo fueron (13318 VN).

2° La **Precisión** del 82 % indica que el modelo se equivoca al clasificar instancias generando Falso Positivos, en nuestro caso en particular, No iban a cancelar y el modelo predijo que Si, error tipo I. 6547 (VP) / (6547 (VP) + 1480 (FP)) = 0.8156

3° El **Recall** del 74 % indica que el modelo comete errores al clasificar instancias generando Falsos Negativos, en nuestro caso en particular, Si iban a cancelar y el modelo predijo que NO, error tipo II. 6547 (VP) / (6547 (VP) + 2276 (FN)) = 0.7420

Como nuestro objetivo es predecir con el mayor porcentaje posible las reservas que van a ser canceladas, nuestra métrica a mirar y mejorar de aquí en adelante es el **Recall**.

Otros Modelos probados:

Con el fin de encontrar el mejor modelo posible decidimos correr los datos en 3 modelo más, Regresión Logística, Gradient Boosting y XGBoost y comparar así los distintos rendimientos obtenidos agrupándolos todos en la siguiente grilla comparativa:

	Model	Train Accuracy	Train Precision Pos	Train Precision Neg	Test Recall Sens	Test Recall Esp	F1 Score Pos	F1 Score Neg
0	Modelo 1: Random Forest Classifier	0.840969	0.815622	0.854018	0.742038	0.899966	0.777092	0.876390
1	Modelo 2: Regresión Logistica	0.792828	0.816731	0.784311	0.574294	0.923150	0.674386	0.848086
2	Modelo 3: Gradient Boosting	0.800406	0.817886	0.793821	0.599116	0.920446	0.691613	0.852457
3	Modelo 4: XGBoost	0.807858	0.833775	0.798186	0.606596	0.927881	0.702270	0.858161
4	Modelo 5: XGBoost Op	0.808112	0.865191	0.789213	0.576108	0.946468	0.691659	0.860717

Conclusión Final:

El modelo de boosting XGBoost optimizado en base a la búsqueda de hiperparámetros con Halving Randomized SearchCV fue el que mejor performance tuvo a la hora de clasificar las reservas que NO iban a ser canceladas con una Precisión del 87 % pero también obtuvo el Recall más bajo, 55 %, lo que indica que no identifica de manera fiable las reservas pasibles de cancelación, nuestro principal objetivo. Por eso mismo nos inclinamos a elegir como modelo al Random Forest Classifier optimizado a través del Grid SearchCV que obtuvo el 74 % de aciertos en clasificar las reservas susceptibles de cancelación sin penalizar tanto la Precisión con un 82 % de aciertos.

Sabiendo ya las métricas de nuestro modelo estamos conformes con los resultados obtenidos cumpliendo así el primer objetivo propuesto, para el segundo objetivo y en base a las métricas obtenidas, **Accuracy** 84 %, **Precision** 82% y **Recall** 74 % recomendamos la opción de sobreventa de habitaciones con el fin de maximizar el uso de las instalaciones.