ITBA
Instituto Tecnológicc
de Buenos Aires

82.05 - Análisis Predictivo

Examen final

SOL ALEJANDRA WINKEL

AGENDA

01 **INTRODUCCIÓN**

Planteo de objetivos. Planteo de la hipótesis

02 **ANÁLISIS EXPLORATORIO**

Variables. Variable target. Missing. Outliers. Correlación. Análisis gráfico

03 **MODELOS PREDICTIVOS**

Partición de la base. Creación de modelos predictivos. Comparación de métricas

04 **CONCLUSIONES**

Modelo ganador. Conclusiones. Recomendaciones

HOTEL RESERVATIONS

Reservas de alojamientos

OBJETIVO

El objetivo es comprender las características de las personas que cancelan reservas hoteleras para proporcionarle a los hoteles un aumento significativo en el rendimiento del hotel al tener la ocupación al máximo y aumentar las ganancias mediante el uso de un modelo predictivo.

HIPÓTESIS

¿Se pueden predecir las cancelaciones en las reservas hoteleras?

¿Influye la anticipación o el medio por el cual se realizan las reservas?

ITBA
Instituto Tecnológicc
de Buenos Aires

ANÁLISIS EXPLORATORIO

ANÁLISIS EXPLORATORIO

Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	booking_status	total_huespedes
1	2	0	1	2	1	0	0	2
2	2	0	2	3	0	0	0	2
3	1	0	2	1	1	0	1	1
4	2	0	0	2	1	0	1	2
5	2	0	1	1	0	0	1	2
6	2	0	0	2	2	0	1	2
7	2	0	1	3	1	0	0	2
8	2	0	1	3	1	0	0	2
9	3	0	0	4	1	0	0	3
10	2	0	0	5	1	0	0	2

La base cuenta con **36275 registros y 19 variables.**

Las reservas son del año **2017 y 2018**

VARIABLES

Booking_ID: identificador único de cada reserva
no_of_adults: cantidad de adultos
no_of_children: cantidad de niños
no_of_weekend_nights: cantidad de noches de fin de semana (Sábado o Domingo) que el cliente reservó para quedarse.
no_of_week_nights: cantidad de noches de la semana (Lunes a Viernes) que el cliente reservó para quedarse.
type_of_meal_plan: tipo de menú que el cliente reservó
Required_car_parking_space: El cliente necesita una cochera? (0 en caso de que No, 1 en caso de que Si)
room_type_reserved: Tipo de habitación reservada por el cliente. Los valores son codificados por INN Hotels.
lead_time: Número de días entre la fecha de reserva y la fecha de llegada
arrival_year: año de la fecha de llegada

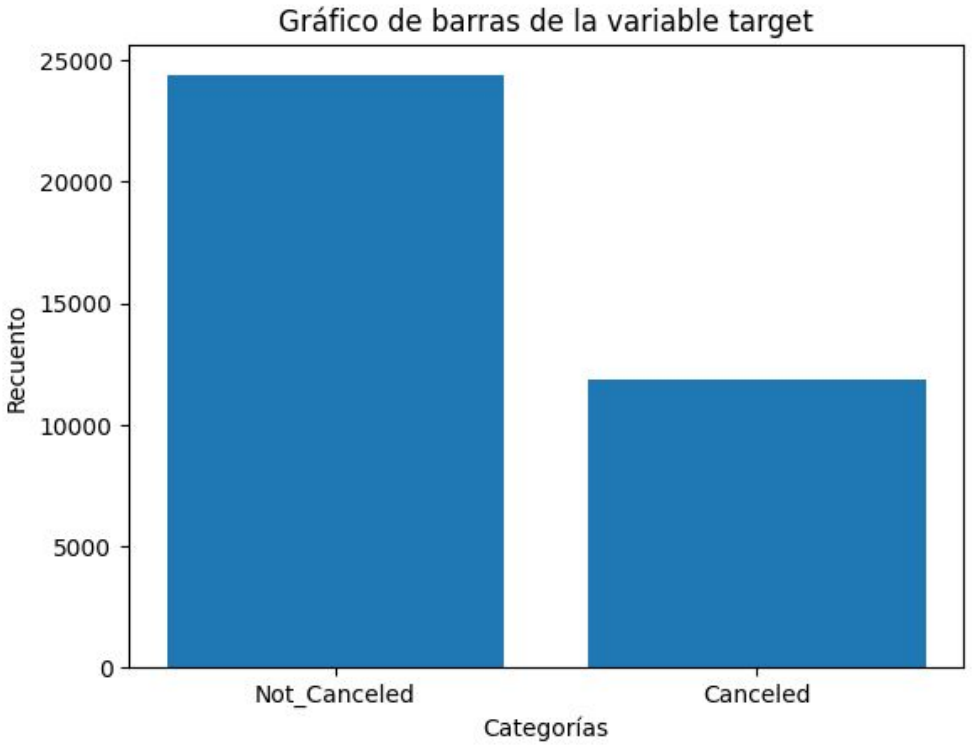
arrival_month: mes de la fecha de llegada
arrival_date: día del mes de la fecha de llegada
market_segment_type: designación del segmento de mercado.
repeated_guest: es el cliente un huésped repetido? (0 en caso de que No, 1 en caso de que Si))
no_of_previous_cancellations: Número de reservas anteriores canceladas por el cliente antes de la reserva actual
no_of_previous_bookings_not_canceled: Número de reservas anteriores no canceladas por el cliente antes de la reserva actual
avg_price_per_room: Precio medio por día de la reserva (en euros)
no_of_special_requests: Número total de solicitudes especiales realizadas por el cliente (por ejemplo, piso alto, vista desde la habitación, etc.)

VARIABLE TARGET

Objetivo: predecir la cancelación de una reserva hotelera

Variable Target: Booking_status

Modelo: clasificación




```
Not_Canceled    24390
Canceled        11885
```

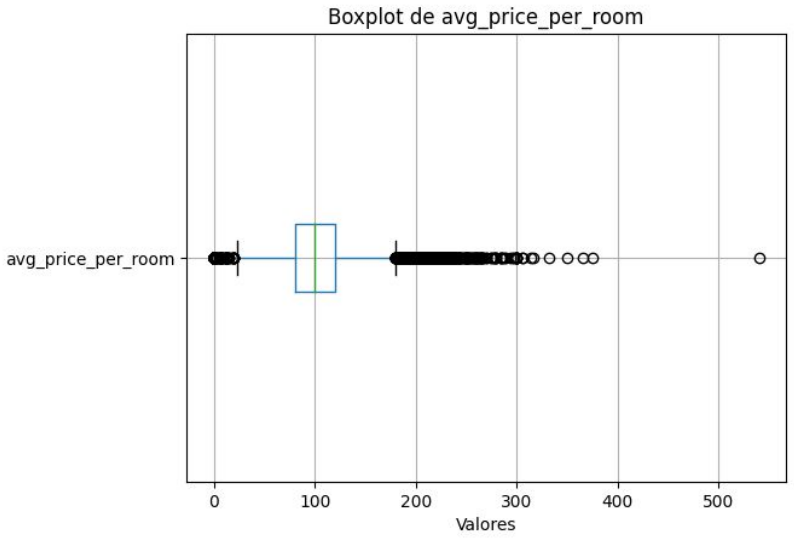
MISSINGS

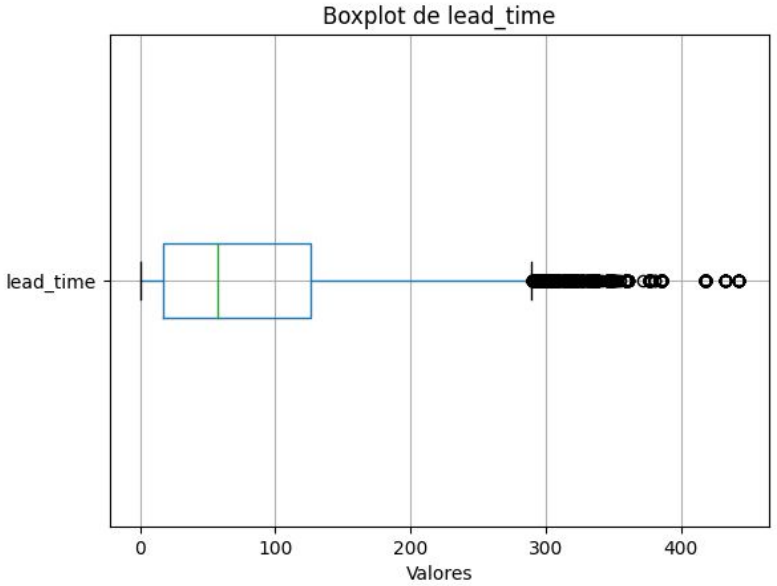
```
Columna: Booking_ID, Nulos: 0
Columna: no_of_adults, Nulos: 0
Columna: no_of_children, Nulos: 0
Columna: no_of_weekend_nights, Nulos: 0
Columna: no_of_week_nights, Nulos: 0
Columna: type_of_meal_plan, Nulos: 0
Columna: required_car_parking_space, Nulos: 0
Columna: room_type_reserved, Nulos: 0
Columna: lead_time, Nulos: 0
Columna: arrival_year, Nulos: 0
Columna: arrival_month, Nulos: 0
Columna: arrival_date, Nulos: 0
Columna: market_segment_type, Nulos: 0
Columna: repeated_guest, Nulos: 0
Columna: no_of_previous_cancellations, Nulos: 0
Columna: no_of_previous_bookings_not_canceled, Nulos: 0
Columna: avg_price_per_room, Nulos: 0
Columna: no_of_special_requests, Nulos: 0
Columna: booking_status, Nulos: 0
```

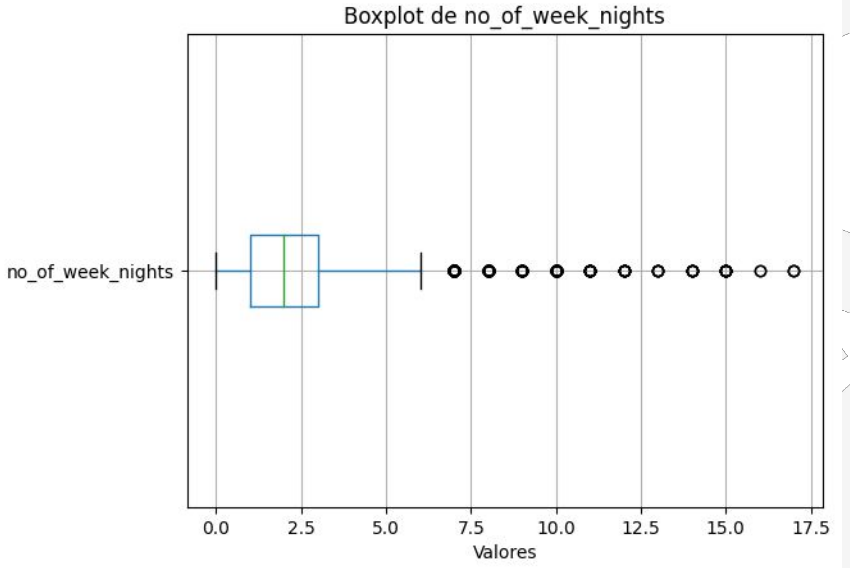
```
[11] nulos_por_columna = dfOriginal.isnull().sum()

     for columna, cantidad_nulos in nulos_por_columna.items():
         print(f'Columna: {columna}, Nulos: {cantidad_nulos}')
```

OUTLIERS








```
Columna: avg_price_per_room
Máximo: 540.0
Mínimo: 0.0
Media: 103.42353907649897
Mediana: 99.45
```

En la variable precio promedio por noche, los valores son muy variados. Los valores van desde 0 a 540. La mediana es 99,45,

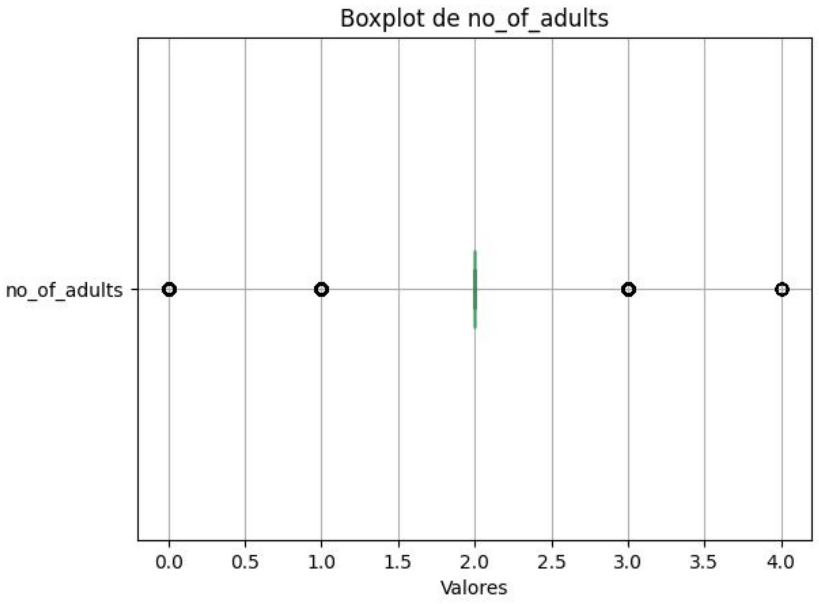
```
Columna: lead_time
Máximo: 443
Mínimo: 0
Media: 85.23255685733976
Mediana: 57.0
```

Valor más alto es 443, siendo un año y dos meses aproximadamente. Hay hoteles que permiten esa anticipación

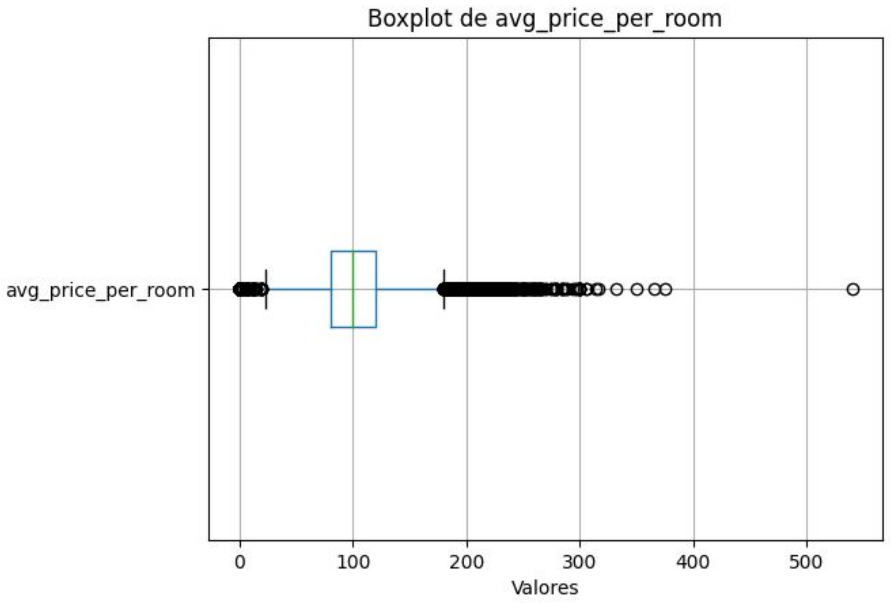
```
Columna: no_of_week_nights
Máximo: 17
Mínimo: 0
Media: 2.2043004824259134
Mediana: 2.0
```

Hay un solo registro de 17 noches de de la semana. (24 días en total)

OUTLIERS



Hay 139 registros que no tienen adultos en sus reservas pero sí chicos, lo que llama la atención.
Se puede tratar de reservas en las que los menores son de entre 16 y 18 años (pueden alojarse solos en muchos países pero cuentan como menores).



Es llamativo que hay 545 registros que el precio promedio por noche es 0.

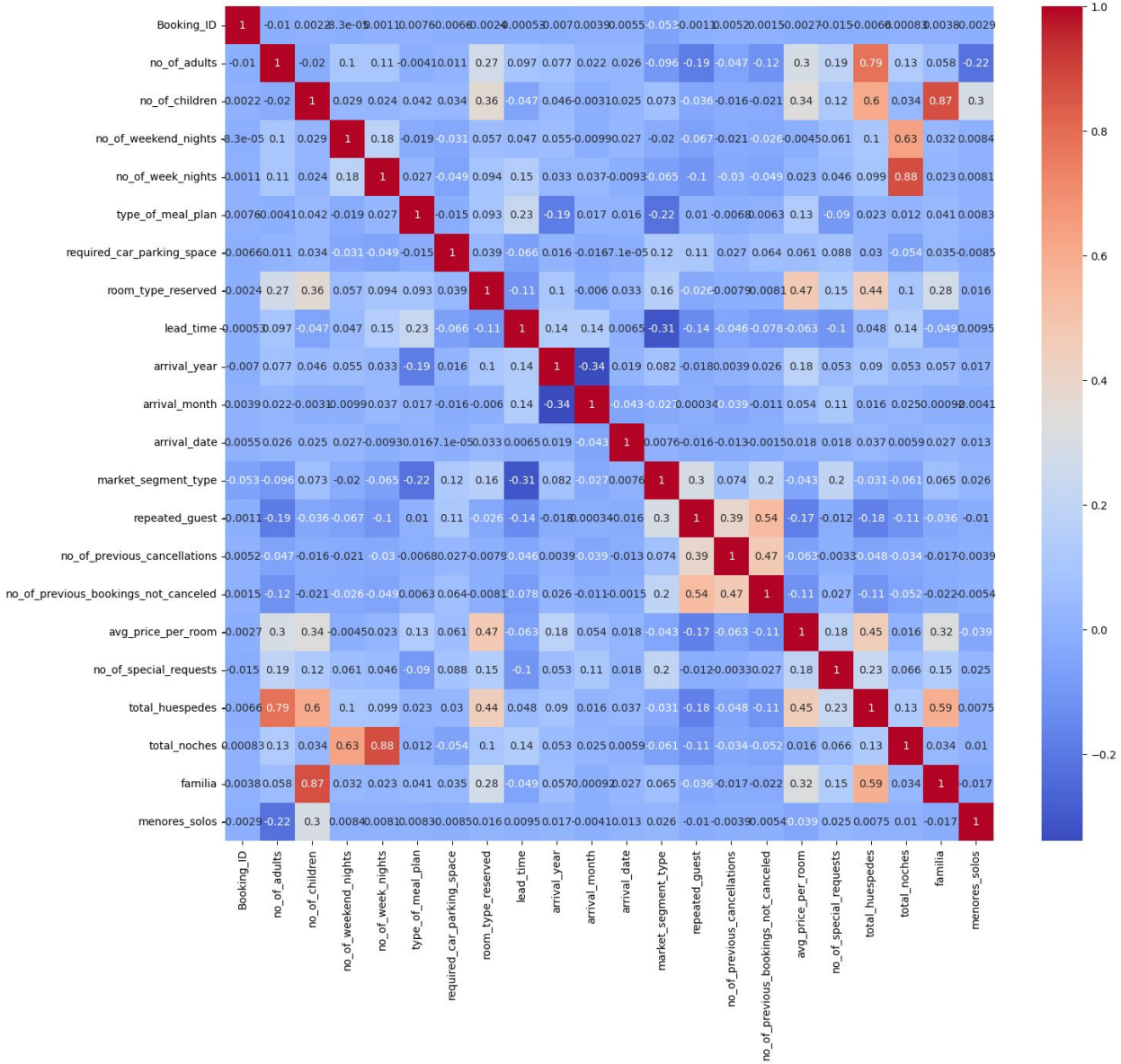
- En el caso de que se traten de Complementary puede que sean 0 debido a que son compensaciones.
- Los de Online deben imputarse. Se imputan los valores en el que el costo promedio por noche es 0, por la media de los costos de las reservas "Online"

CREACIÓN DE NUEVAS VARIABLES

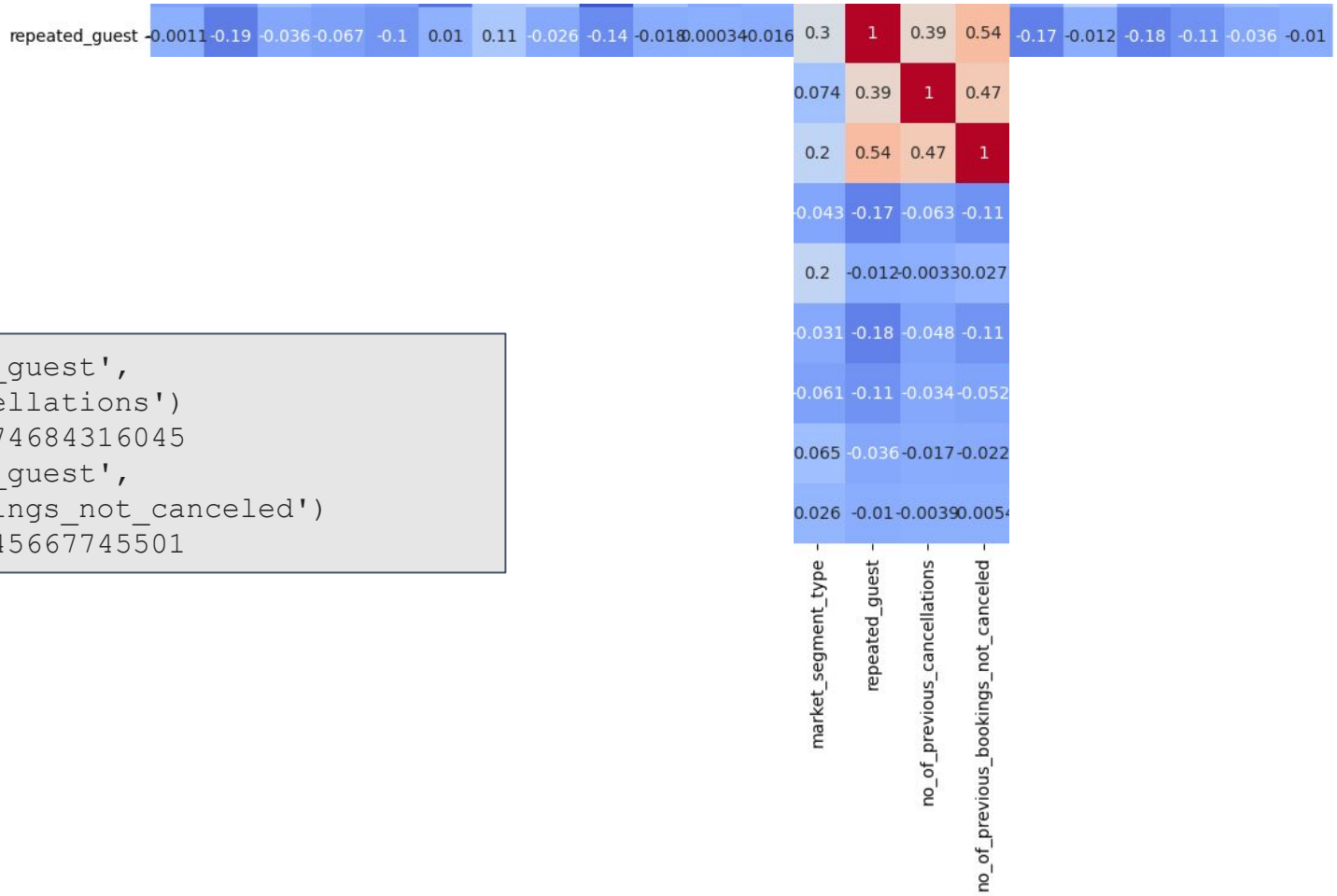
01	TOTAL_HUESPEDES	Se calcula la cantidad total de huéspedes, sumando la cantidad de adultos y la cantidad de chicos.
02	TOTAL_NOCHES	Se calcula la cantidad total de noches, sumando las noches de semana y de fin de semana
03	FECHA_LLEGADA	Se utiliza el día, mes y año para formar la fecha de llegada del huésped
04	FAMILIA	Se analiza si es una familia o no. Se considera familia en caso de tener hijos
05	MENORES_SOLOS	Se analiza si hay menores solos o no

CORRELACIÓN

PEARSON



PEARSON + SPEARMAN

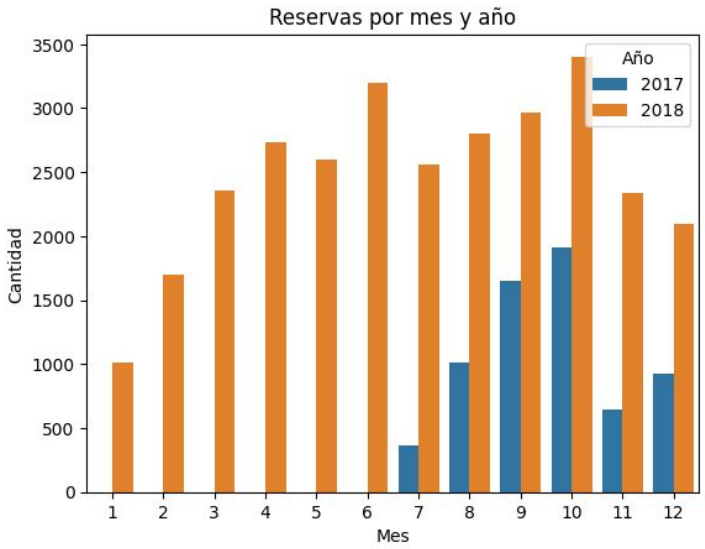


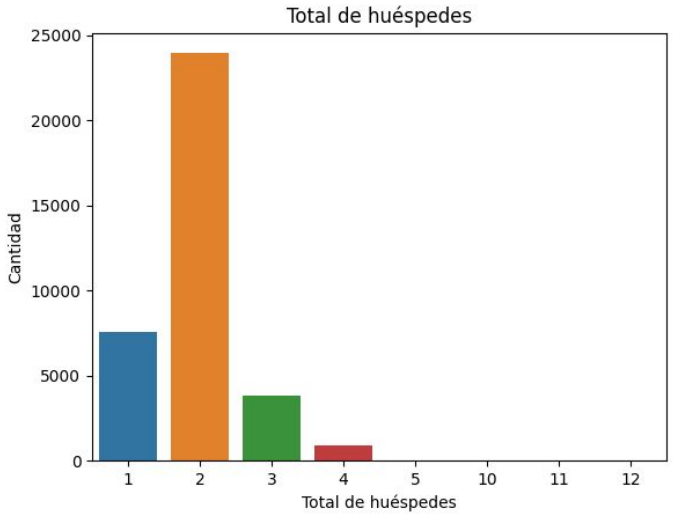

```
Columnas: ('repeated_guest',
'no_of_previous_cancellations')
Correlación: 0.5978674684316045
Columnas: ('repeated_guest',
'no_of_previous_bookings_not_canceled')
Correlación: 0.9327745667745501
```

ANÁLISIS GRÁFICO

RESERVAS

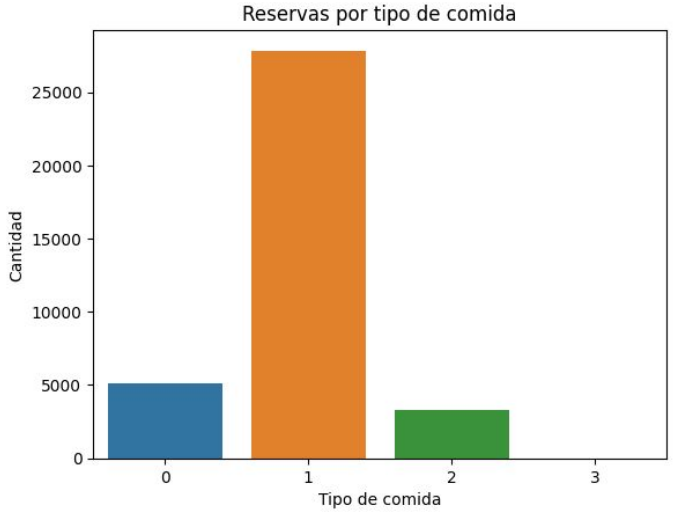
Cantidad de reservas por mes y por año

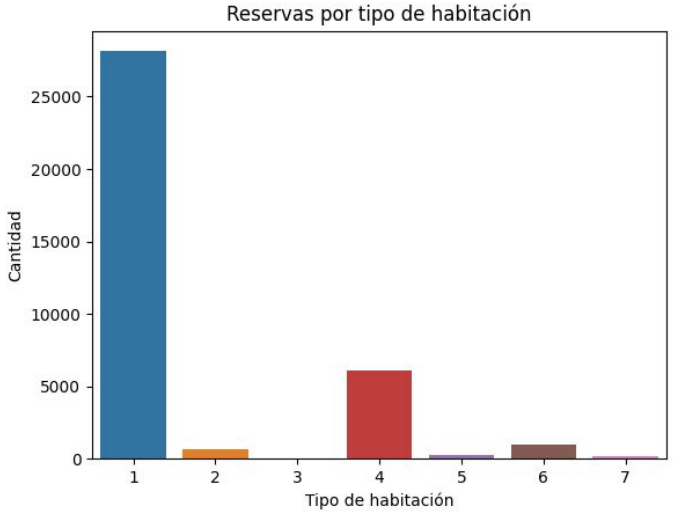




Cantidad de reservas por cantidad total de huéspedes

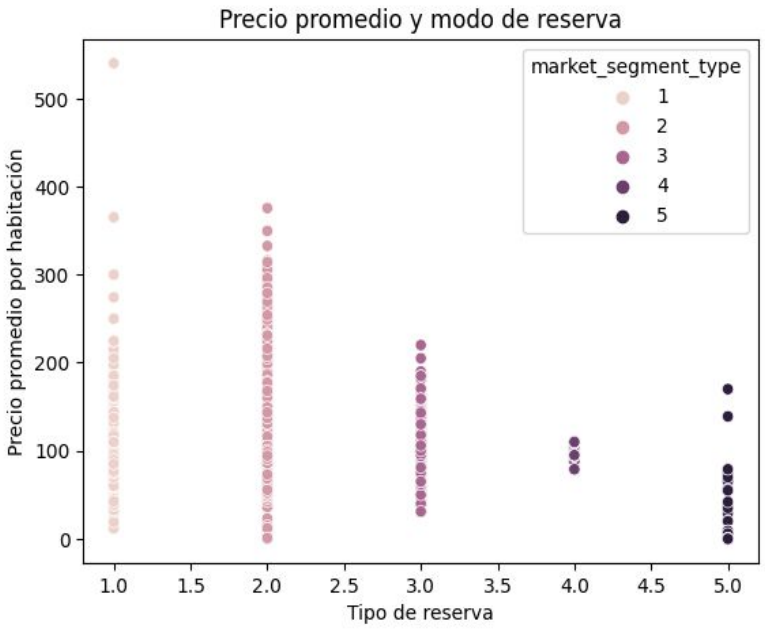
Cantidad de reservas por tipo de plan de comidas



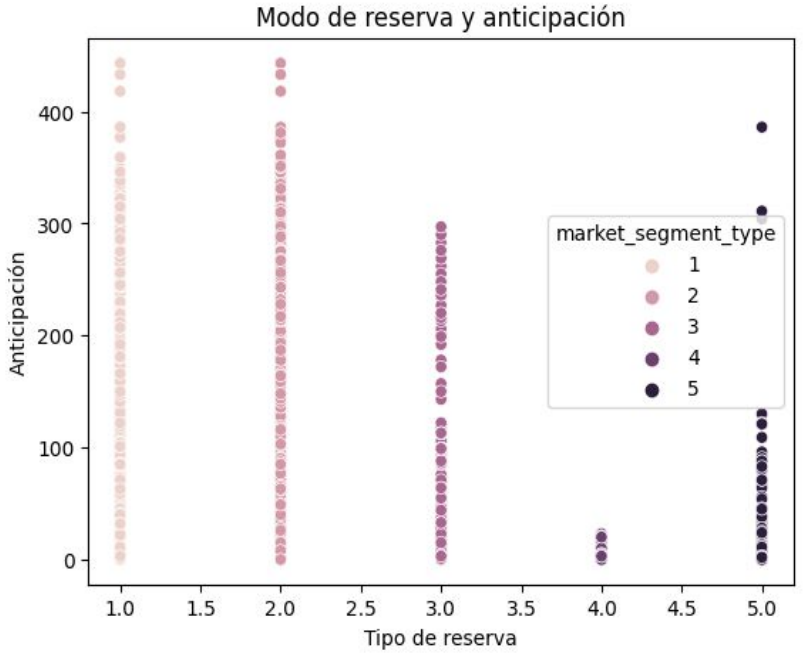


Cantidad de reservas por tipo de habitación

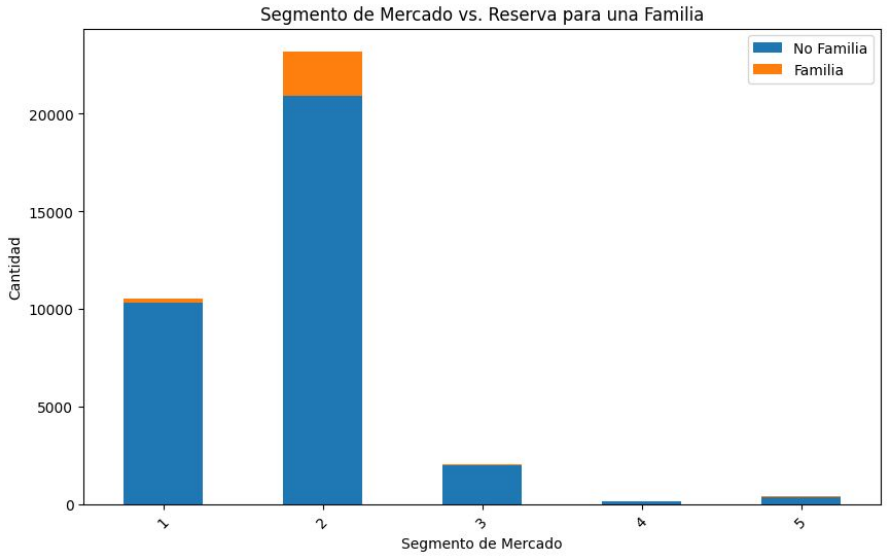
SEGMENTOS DEL MERCADO



Precio promedio de la reserva por segmento de mercado



Anticipación de la reserva por segmento de mercado



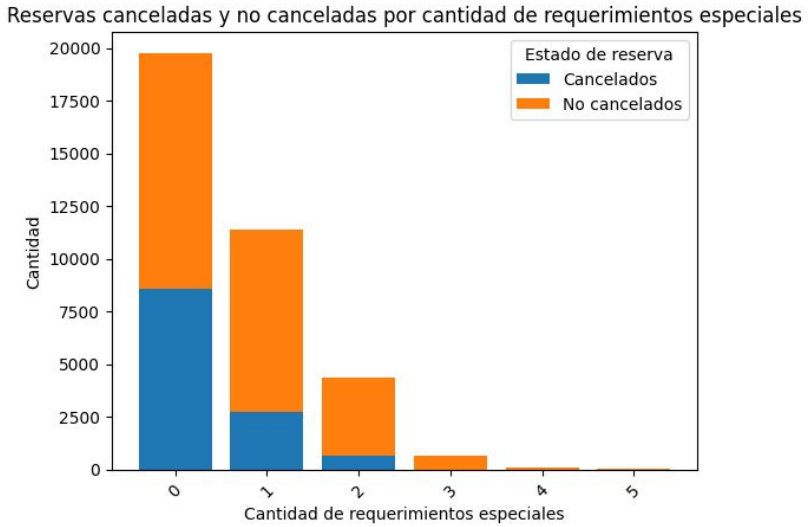
Cantidad de huéspedes especificando las familias por segmento de mercado

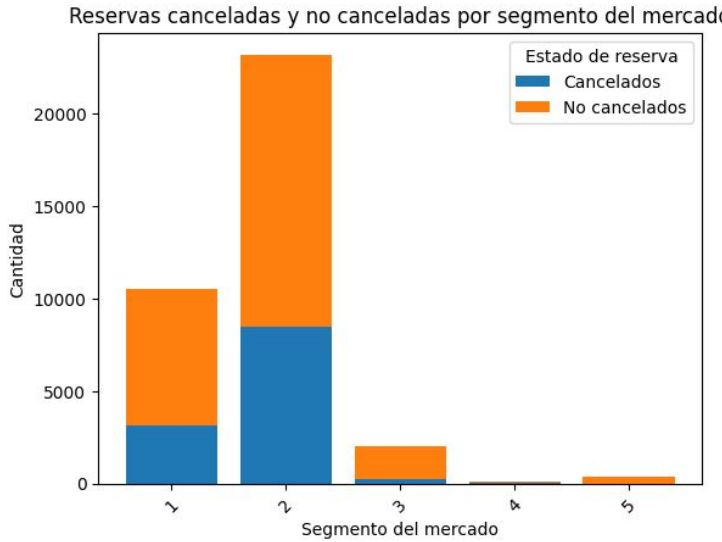
'Offline' = 1, 'Online' = 2, 'Corporate' = 3, 'Aviation' = 4,'Complementary' = 5

VISUALIZACIONES

CANCELACIONES

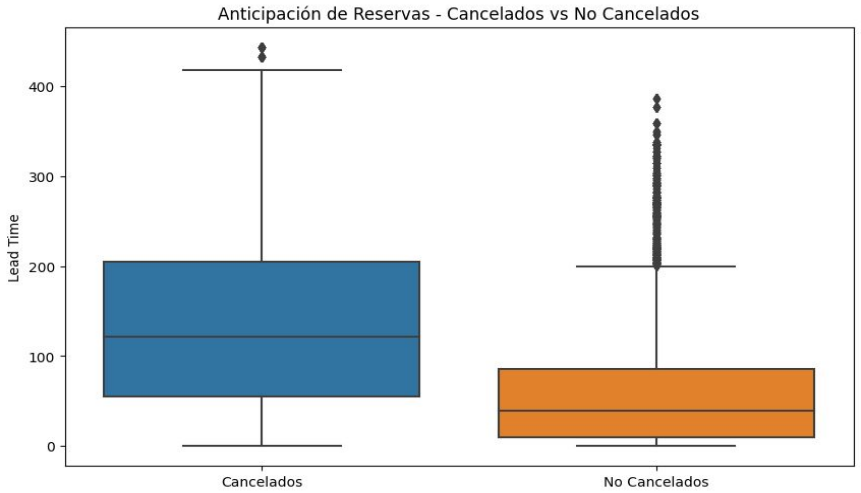
Reservas canceladas y no canceladas por la cantidad de requerimientos especiales





Reservas canceladas y no canceladas or segmento de mercado

Reservas canceladas y no canceladas junto con la anticipación de las reservas



Reservas canceladas y no canceladas por mes

MODIFICACIÓN DE VARIABLES NO NUMÉRICAS

01	BOOKING_ID	Se eliminan los primeros 3 caracteres del ID debido a que en todos los registros es igual (INN)
02	TYPE_OF_MEAL_PLAN	Se crea un diccionario para pasar a numéricas las categorías (Meal Plan 1, Not Selected, Meal Plan 2, Meal Plan 3)
03	ROOM_TYPE_RESERVED	Se crea un diccionario para pasar a numéricas las categorías (Room_Type 1, Room_Type 4, Room_Type 2, Room_Type 6, Room_Type 5, Room_Type 7, Room_Type 3)
04	MARKET_SEGMENT_TYPE	Se crea un diccionario para pasar a numéricas las categorías (Offline, Online, Corporate, Aviation, Complementary)

PARTICIÓN DE LA BASE

Realizaré una partición en tres partes.

- La **base train** será la que voy a usar para entrenar los modelos
- La **base validación** será la que utilizare para verificar el funcionamiento de los modelos (testear los modelos luego de entrenar)
- La base **test** será la que guardaré hasta el final para probar el funcionamiento de los modelos, con el fin de no comparar con esos resultados previamente y hacer el análisis más serio

La partición será estratificada porque al tener una diferente cantidad de reservas canceladas y no canceladas, considero importante que estén equilibradas en ambas bases.

- Se separa un 80% para el conjunto de val y el 20% para el conjunto de test
- Se separa un 80% del conjunto val para train y el 20% para test

```
[ ]  X = dfOriginal[columnas]
     y = dfOriginal['booking_status']

     X_val_train, X_test, y_val_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y)

     # X_val_train: Conjunto de características para validación y entrenamiento
     # y_val_train: Conjunto de variable objetivo para validación y entrenamiento
     # X_test: Conjunto de características para prueba
     # y_test: Conjunto de variable objetivo para prueba

     X_train, X_test_train, y_train, y_test_train = train_test_split(X_val_train, y_val_train, test_size=0.20, stratify=y_val_train)

     # X_train: Conjunto de características para entrenamiento
     # y_train: Conjunto de variable objetivo para entrenamiento
     # X_test_train: Conjunto de características para prueba en el conjunto train
     # y_test_train: Conjunto de variable objetivo para prueba en el conjunto train
```

ITBA
Instituto Tecnológicc
de Buenos Aires

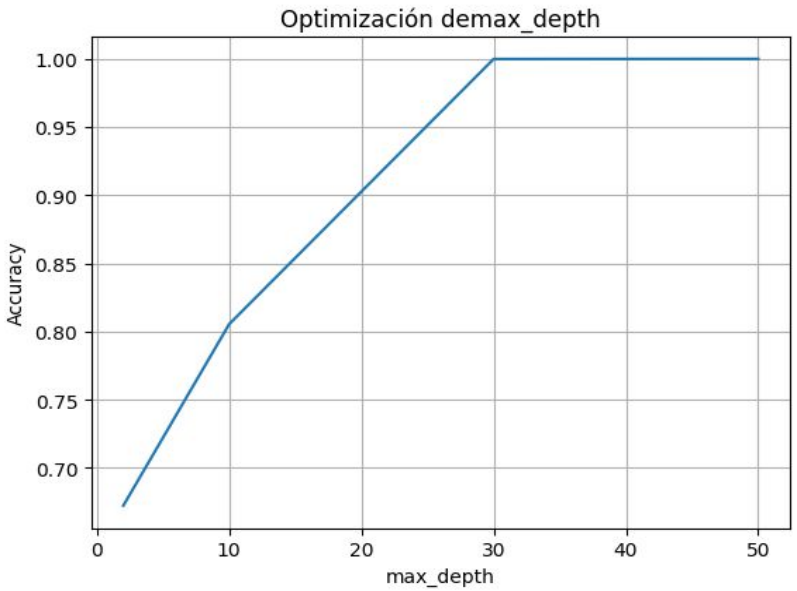
MODELOS PREDICTIVOS

MODELOS

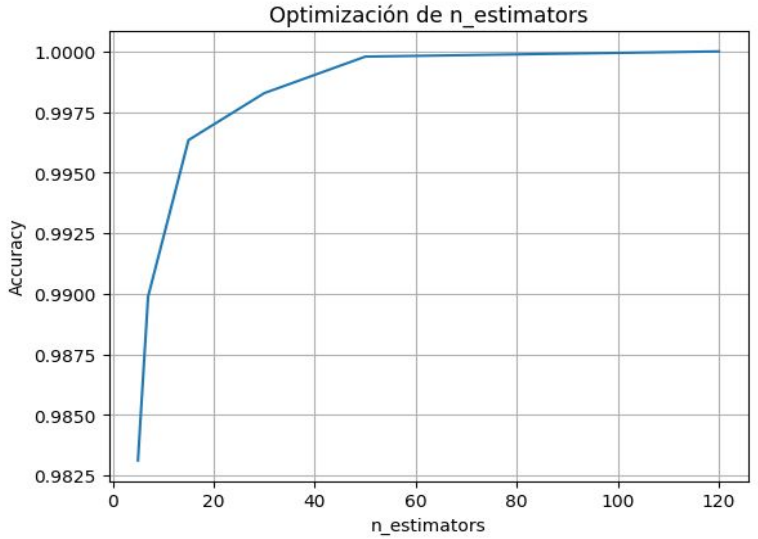
Los modelos utilizados son:

1. DecissionTreeClassifier()
2. RandomForestClassifier()
3. ExtraTreeClassifier()
4. CatBoostClassifier()
5. AdaBoostClassifier()
6. LightGBMClassifier()

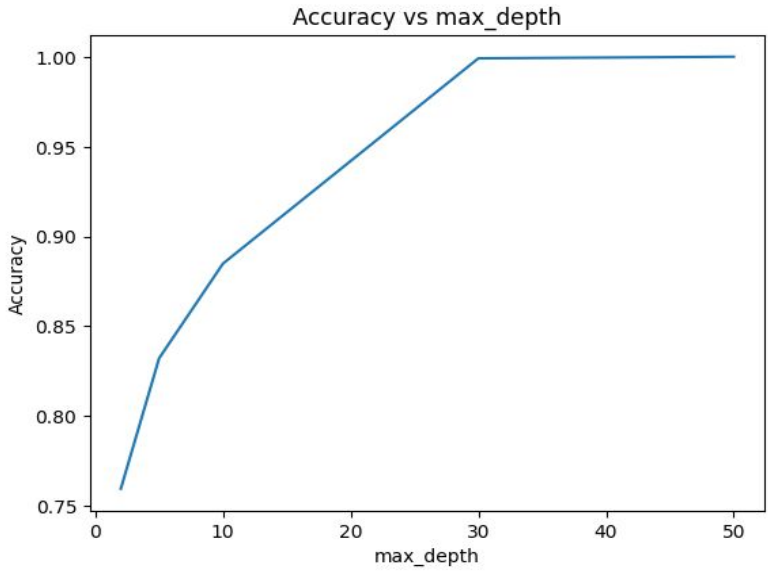
AJUSTE DE HIPER PARÁMETROS



EXTRA TREE CLASSIFIER



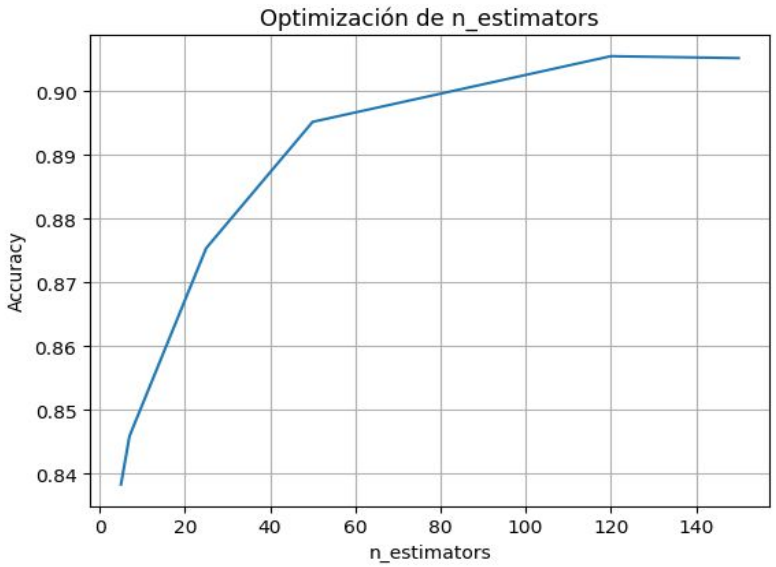
RANDOM FOREST CLASSIFIER



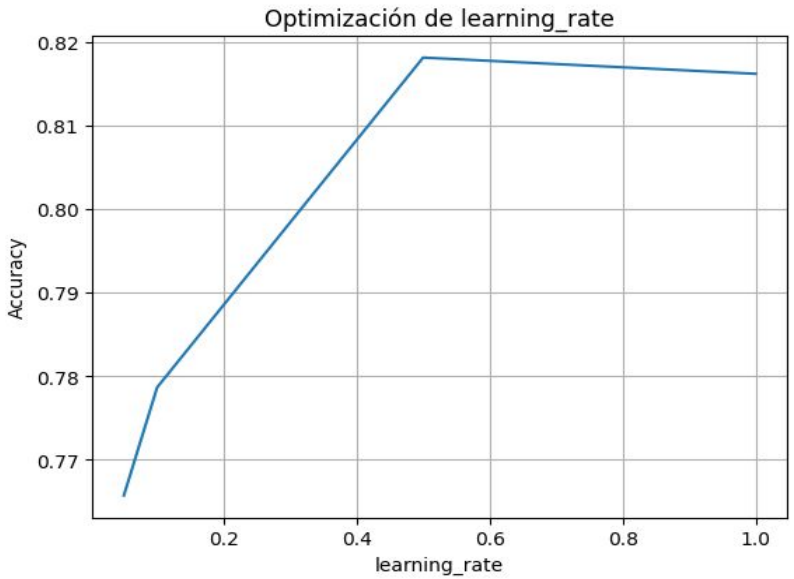
DECISION TREE CLASSIFIER

Se **evaluaron** algunos de los hiper parámetros **respecto al accuracy** del modelo para analizar cómo el mismo afecta el rendimiento del modelo. **Ayuda a identificar el valor óptimo del hiper parámetro**

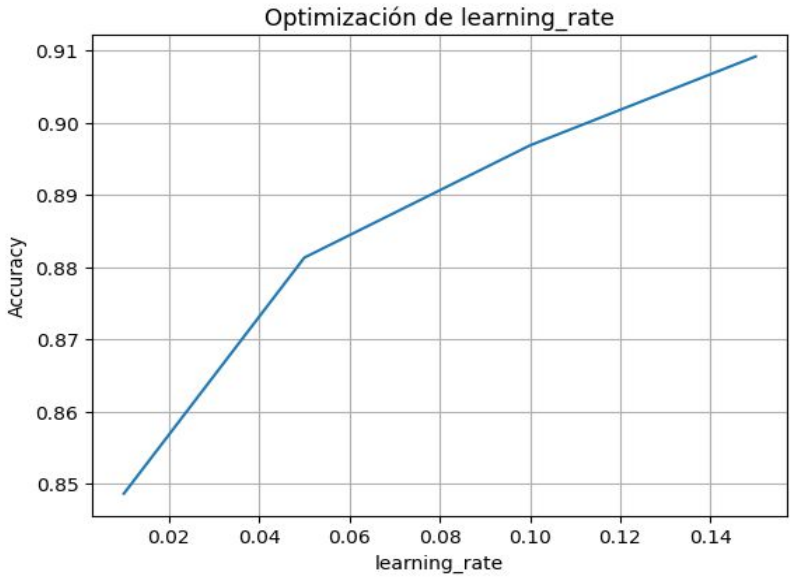
AJUSTE DE HIPER PARÁMETROS



CATBOOST CLASSIFIER



ADABOOST CLASSIFIER



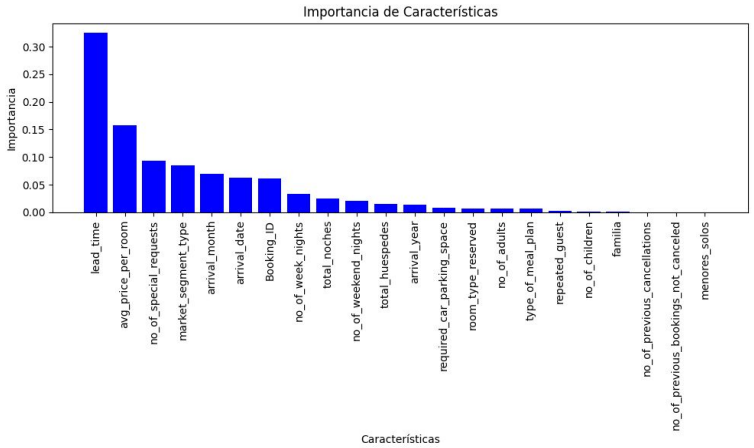
LIGHT GBM CLASSIFIER

Luego, se utilizó **GridSearch** para obtener la mejor combinación de ellos.

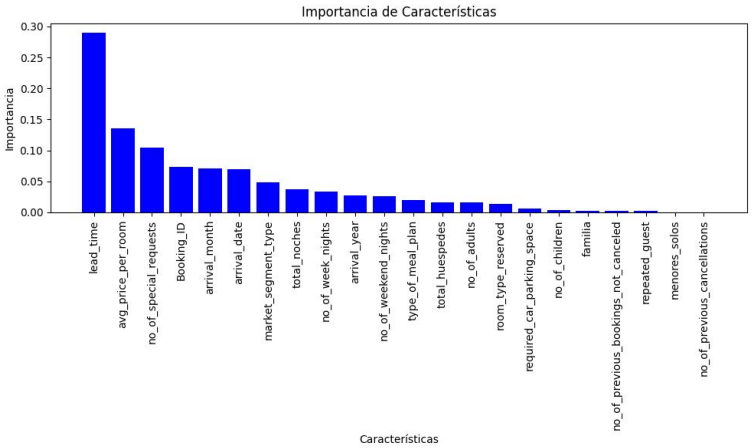
A su vez, se aplicó **K Fold Cross-Validation (cv = 5)** para evitar el overfitting en el modelo

IMPORTANT FEATURES

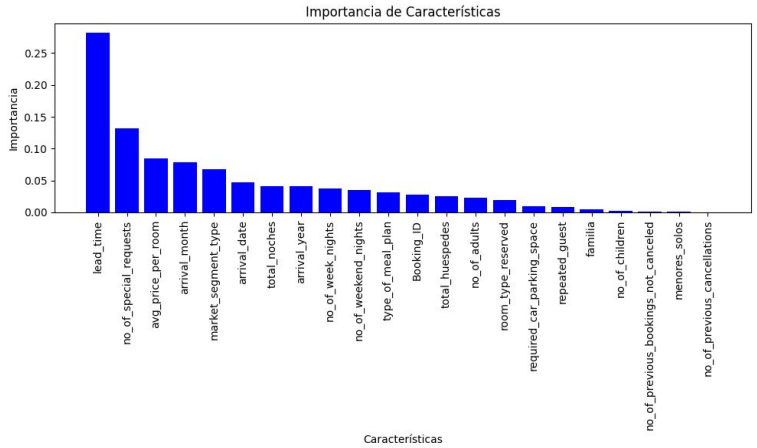
DECISION TREE CLASSIFIER



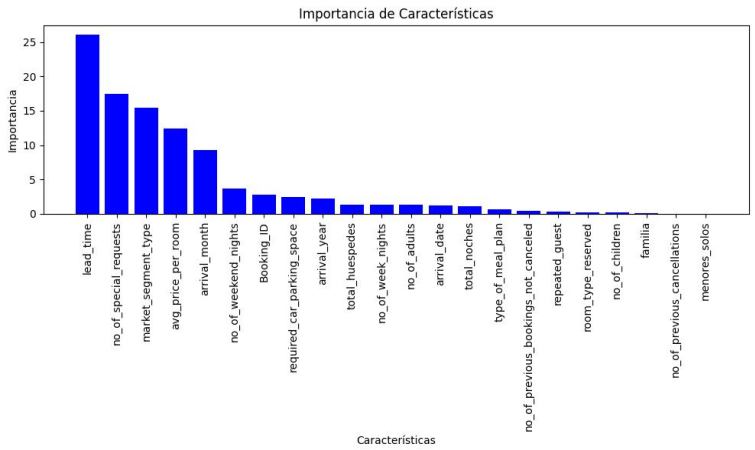
RANDOM FOREST CLASSIFIER



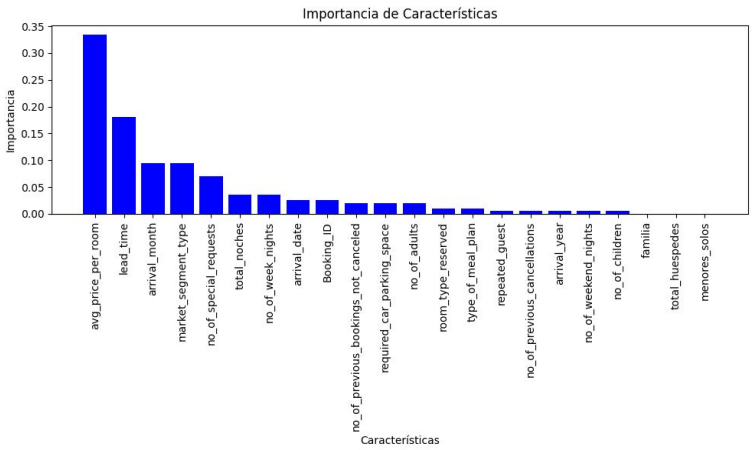
EXTRA TREE CLASSIFIER



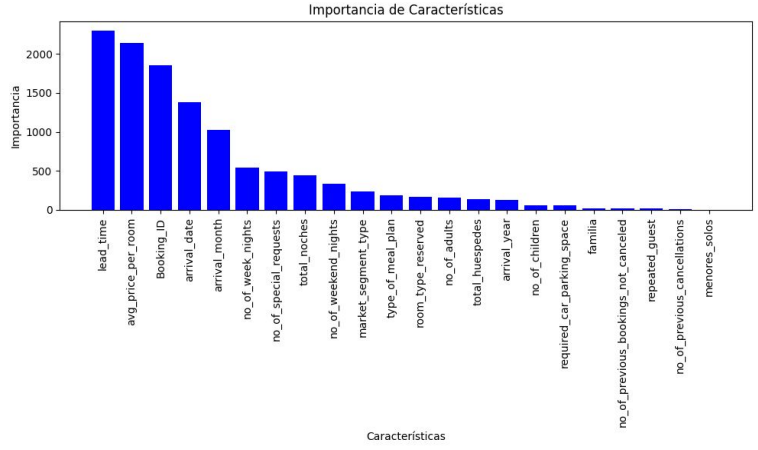
CATBOOST CLASSIFIER



ADABOOST CLASSIFIER

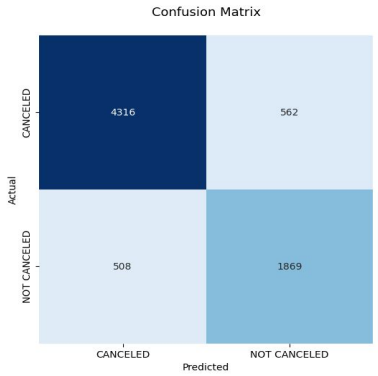


LIGHT GBM CLASSIFIER



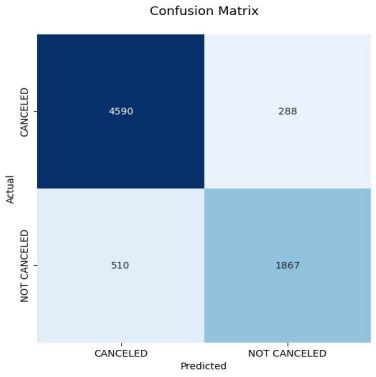
COMPARACIÓN DE MÉTRICAS

DECISION TREE CLASSIFIER



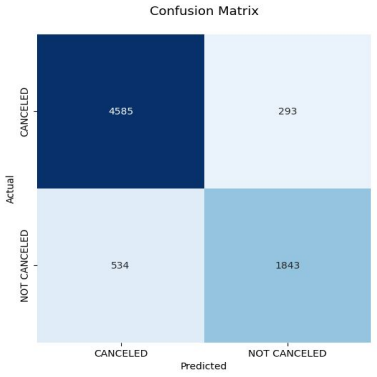
Accuracy: 0.8525155065472089
Precision: 0.7688194158782394
Recall: 0.7862852334875894
F1-Score:: 0.7774542429284526
AUC-ROC: 0.8355370406880341
Log Loss: 5.315879962281923

LIGHT GBM CLASSIFIER



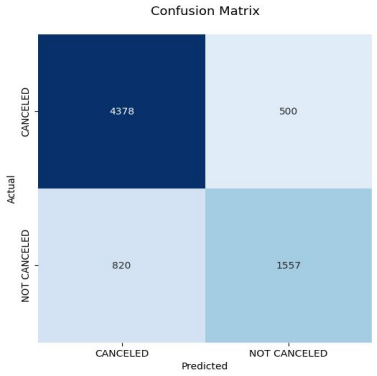
Accuracy: 0.8900068917987595
Precision: 0.8663573085846867
Recall: 0.7854438367690366
F1-Score: 0.823918799646955
AUC-ROC: 0.8632016231815662
Log Loss: 3.964553467197173

CATBOOST CLASSIFIER



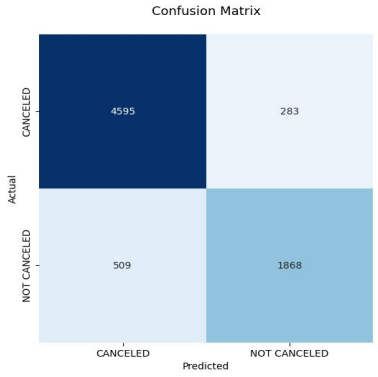
Accuracy: 0.8860096485182632
Precision: 0.8628277153558053
Recall: 0.7753470761464031
F1-Score: 0.8167516064701972
AUC-ROC: 0.8576407377451982
Log Loss: 4.108628718511356

ADABOOST CLASSIFIER



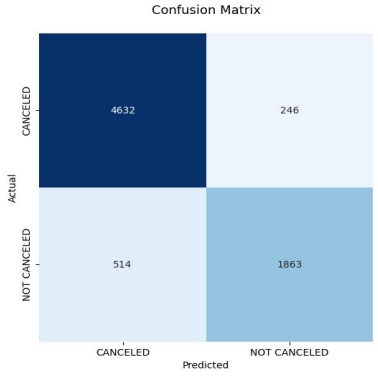
Accuracy: 0.8180565127498277
Precision: 0.7569275644141954
Recall: 0.6550273453933529
F1-Score: 0.7023004059539919
AUC-ROC: 0.7762631601915515
Log Loss: 6.557907990852467

EXTRA TREE CLASSIFIER



Accuracy: 0.8908339076498967
Precision: 0.8684332868433287
Recall: 0.785864535128313
F1-Score: 0.8250883392226148
AUC-ROC: 0.8639244774862557
Log Loss: 3.93474479451148

RANDOM FOREST CLASSIFIER



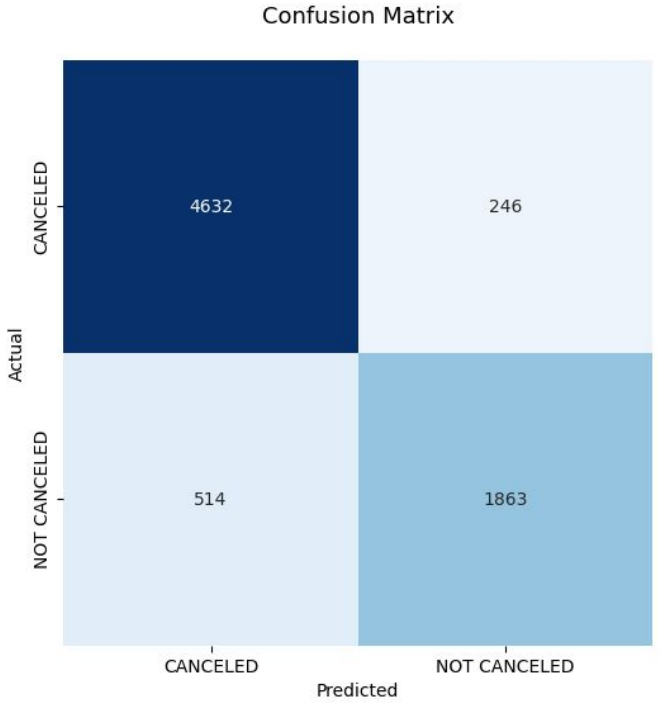
Accuracy: 0.8952446588559614
Precision: 0.883357041251778
Recall: 0.783761043331931
F1-Score: 0.8305840392331698
AUC-ROC: 0.866665269513444
Log Loss: 3.77576520685445

ITBA
Instituto Tecnológicc
de Buenos Aires

CONCLUSIONES

MODELO GANADOR - RANDOM FOREST CLASSIFIER

Utiliza un conjunto de árboles de decisión para realizar predicciones. Combina las predicciones de múltiples árboles de decisión para obtener una predicción final más robusta y precisa.



Classification report:

	precision	recall	f1-score	support
0	0.90	0.95	0.92	4878
1	0.88	0.78	0.83	2377
accuracy			0.90	7255
macro avg	0.89	0.87	0.88	7255
weighted avg	0.89	0.90	0.89	7255

Exactitud (Accuracy): 0.8952446588559614
Precisión (Precision): 0.883357041251778
Sensibilidad (Recall): 0.783761043331931
Valor F1 (F1-Score): 0.8305840392331698
AUC-ROC: 0.866665269513444
Log Loss: 3.77576520685445

rfc = RandomForestClassifier(n_estimators=50, max_depth=None,
min_samples_split=5,min_samples_leaf=1, criterion = gini, max_features = sqrt)

CONCLUSIONES

- **Random Forest Classifier** es el modelo ganador. Tiene el accuracy mas alto, asi como tambien el resto de las métricas. Este modelo dentro de sus ventajas tiene:
  - Capacidad para manejar conjuntos de datos grandes
  - Reducción del sobreajuste (overfitting)

HIPÓTESIS

Se pueden predecir las cancelaciones en las reservas hoteleras?

- **Sí**, el modelo predictivo **Random Forest Classifier** es un ejemplo

Influye la anticipación o el medio por el cual se realizan las reservas?

- **Lead_time** (anticipación de la reserva) es un **feature muy relevante** a la hora de predecir los valores
- **Market_segment_type** es un feature que aporta información pero **no es de los más significativos**

RECOMENDACIONES RESPECTO AL MODELO

- Utilizar el modelo que tenga mejor precisión para la clase positiva (en este caso “Canceled”)
- Considerando que el modelo es robusto, agregar nuevas variables que otorguen valor ayudará a que el score suba y así se pueda predecir de manera más acertada

RECOMENDACIONES RESPECTO AL NEGOCIO

- Ofrecer incentivos y beneficios adicionales para los clientes que no cancelan, fomentando que vuelvan a reservar y fidelizando a los clines
- Implementar políticas flexibles de cancelación. Con el objetivo de maximizar las ganancias, maximizando la ocupación del hotel, es importante que se disminuyan las cancelaciones con poca anticipación. Al tener políticas flexibles, fomenta que en caso de cancelar, lo realicen con anticipación para no pagar penalizaciones.
- Lanzar promociones y descuentos para las épocas en las cuales hay menos reservas, es decir durante Enero y Febrero.
- Armar paquetes para viajes de trabajo ya que las reservas son menos propensas a cancelarse
- Pedido de confirmación un tiempo anterior al viaje. En caso de cancelar en ese momento no tiene penalización, fomentando que ante la duda se cancele y se realice una nueva reserva.

ITBA
Instituto Tecnológicc
de Buenos Aires

Muchas gracias