Project
DATA SCIENCE
“Hotel Reservation”

Alumno: *JUAN CORTEZ ZAMAR*
Profesor: *Jonathan Domínguez*
Tutora: *Noelia Ferrero*
Academia CODERHOUSE

PRESENTACIÓN

Juan Cortez Zamar

Soy de la Provincia de Jujuy - ARGENTINA, actualmente resido en la Provincia de Tucumán - ARGENTINA

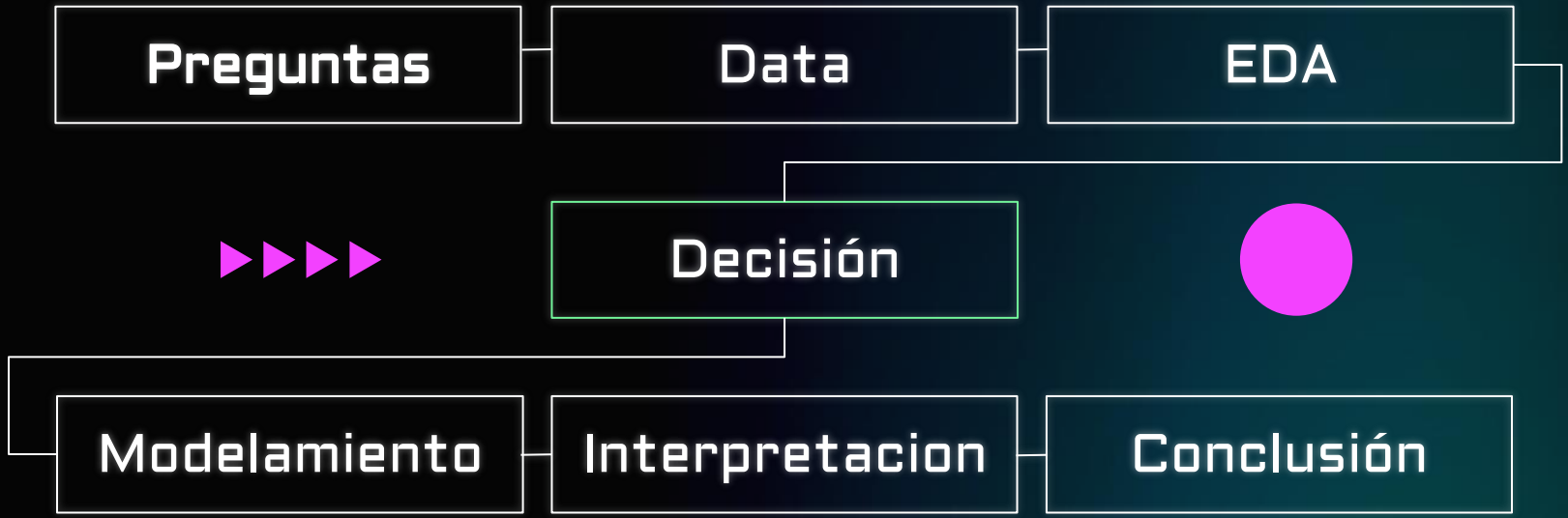
Me recibí de Licenciado en Biotecnología en la Universidad Nacional de Tucumán, hice hasta 2 año de Tec. en Programación de la UTN – FRT y actualmente estoy cursando la Tecnicatura en ciencias de datos e IA en Instituto Superior Politécnico de Córdoba

Hasta fin de agosto he trabajado como pasante de Analista de Datos en la planta industrial de Scania que se encuentra en la Provincia de Tucumán.

TABLA DE CONTENIDOS

TABLA DE CONTENIDOS

ARQUITECTURA DE PROYECTO DE DATA SCIENCE



INTRODUCCION

"En el mundo de la industria hotelera, la toma de decisiones informadas es esencial para el éxito. En este contexto, nuestro proyecto de Data Science se enfoca en analizar y optimizar las reservas de hoteles. Utilizando técnicas avanzadas de análisis de datos, machine learning y visualización, hemos desarrollado un enfoque que permite a los hoteles comprender mejor el comportamiento de los clientes, predecir la demanda y mejorar la eficiencia operativa. A través de este proyecto, exploraremos cómo los datos pueden impulsar estrategias más efectivas para la gestión de reservas, maximizando la satisfacción del cliente y los ingresos."

01

Abstract

- La industria hotelera es una aventura empresarial para el propietario y una odisea para el viajero y/o turista.
- En nuestro estudio, hemos identificado un desafío fundamental en la reserva de habitaciones de hotel: la necesidad de que los clientes se presenten físicamente en el hotel para garantizar una habitación, ya que los recepcionistas tienden a prestar más atención a esta modalidad.
- Los canales de reserva de hoteles en línea han transformado radicalmente las posibilidades de reserva y han influido en el comportamiento de los clientes.

- Además, hemos observado que un número significativo de reservas de hotel se cancelan debido a cambios de planes, conflictos de horarios y otros motivos comunes.
- Nuestro proyecto se enfoca en abordar estos desafíos y en mejorar la experiencia de reserva para los viajeros, al mismo tiempo que proporciona a los hoteles herramientas para gestionar de manera más eficiente la demanda y reducir las cancelaciones.

Problemas/ Contextos

02

PLANTEO DE PROBLEMAS

En este estudio, abordamos una serie de desafíos y cuestiones cruciales en la gestión de reservas hoteleras en los años 2017 y 2018:

- Comprobamos la cantidad de reservas que se concretaron durante estos dos años y verificamos la situación de las mismas.
- Analizamos el impacto de los factores monetarios, como el precio medio por habitación, en la concreción de reservas.
- Evaluamos la influencia del **lead_time** en la tasa de cancelación de reservas.
- Identificamos los meses en los que se registra la mayor tasa de reservas canceladas.
- Investigamos los canales y segmentos utilizados por los clientes para realizar sus reservas.
- Estudiamos si las solicitudes especiales por parte de los clientes tienen un efecto en la concreción de las reservas.

CONTEXTO COMERCIAL

En el contexto del sector hotelero, hemos emprendido un estudio exhaustivo de las reservas en el Hotel X, el cual ofrece una amplia variedad de tipos de habitaciones tanto para adultos como para familias, opciones de estacionamiento, diversos planes de comida y la capacidad de satisfacer requerimientos especiales de los huéspedes.

Nuestra investigación se basa en una base de datos rica en información que abarca tanto las reservas confirmadas como las canceladas, con detalles que incluyen las fechas de entrada y salida, la cantidad de adultos y niños en cada reserva y los gastos asociados a las estadías.

CONTEXTO COMERCIAL

En este proyecto, abordamos una serie de desafíos comerciales clave, como la optimización de la gestión de reservas, la reducción de cancelaciones y la personalización de la experiencia del cliente.

A través del análisis de esta valiosa información, buscamos mejorar tanto la eficiencia operativa del Hotel X como la satisfacción y fidelización de los huéspedes, con un enfoque en la creación de estrategias efectivas y la implementación de un sistema de reservas en línea que garantice la elección de habitaciones deseada por parte de los clientes tras una visita virtual.

Este proyecto se enmarca en el contexto comercial altamente competitivo del sector hotelero, donde la toma de decisiones basada en datos es esencial para el éxito continuo del Hotel X y la mejora de la experiencia de los huéspedes.

PROBLEMA COMERCIAL

En el contexto del Hotel X, nos enfrentamos a un problema comercial crítico: la tasa de cancelación de reservas. Nuestro objetivo principal es reducir esta tasa identificando los factores que influyen en la toma de decisiones de nuestros clientes al realizar reservas, utilizando la rica información de la que disponemos.

Este proyecto busca comprender a fondo los patrones que llevan a las cancelaciones de reservas y, a partir de este conocimiento, diseñar estrategias efectivas para reducir las cancelaciones. Nuestra estrategia se basa en la aplicación de análisis de datos avanzados y técnicas de machine learning para identificar los factores clave que inciden en la decisión de los clientes de cancelar sus reservas.

PROBLEMA COMERCIAL

Una vez identificados estos factores, planeamos implementar campañas de marketing personalizadas respaldadas por datos sólidos. Estas campañas estarán diseñadas para abordar las preocupaciones específicas de los clientes, mejorar la satisfacción y fomentar la finalización de las reservas.

Este proyecto se enmarca en el contexto altamente competitivo de la industria hotelera, donde la retención de clientes y la reducción de las cancelaciones son esenciales para el éxito a largo plazo del Hotel X. A través del uso estratégico de los datos, buscamos no solo resolver nuestro problema comercial sino también mejorar la experiencia de reserva de nuestros huéspedes y fortalecer nuestra posición en el mercado.

CONTEXTO ANALÍTICO

En nuestra búsqueda por abordar el problema comercial de reducir la tasa de cancelaciones en el Hotel X, hemos adoptado un enfoque analítico integral. Reconocemos que múltiples factores influyen en el comportamiento de nuestros clientes al realizar reservas, y nuestra estrategia se centra en aprovechar la información disponible para identificar patrones de comportamiento clave y reorientar nuestras estrategias comerciales de manera efectiva.

Para lograrlo, hemos empleado técnicas de Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) que nos permiten explorar y comprender las relaciones complejas entre diversos factores, como el tiempo de anticipación de la reserva, la duración de la estadía y otros. Estos análisis nos brindan una visión más profunda de cómo estas variables se relacionan con las reservas y, en última instancia, con la tasa de cancelación.

CONTEXTO ANALÍTICO

Nuestra meta es desarrollar modelos predictivos sólidos que nos permitan anticipar la tasa de cancelación de reservas. Esto nos proporcionará una herramienta valiosa para evaluar el impacto de las estrategias comerciales y publicitarias en la reducción de las cancelaciones.

Objetivos principales

03

Objetivo 1

Reducir las Cancelaciones de Reservas: El primer objetivo principal es implementar estrategias efectivas para reducir significativamente la tasa de cancelaciones de reservas en el Hotel X. Esto implicará identificar los factores clave que llevan a las cancelaciones y desarrollar medidas preventivas para abordarlos.

Objetivo 2

Predecir Tendencias de Nuevas Reservas: El segundo objetivo central es utilizar análisis predictivos para anticipar las tendencias de nuevas reservas en respuesta a campañas de marketing específicas. Esto nos permitirá tomar decisiones estratégicas más informadas y optimizar nuestras iniciativas de marketing para atraer y retener a nuestros clientes de manera más efectiva.

Dataset / Metadata

04

Dataset: “Hotel Reservation”

Fue extraído de la página Kaggle: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset

COLUMN	DESCRIPTION	DATATYPE
Booking_ID	Unique identifier of each booking	String
no_of_adults	Number of adults	Integer
no_of_children	Number of Children	Integer
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	Integer
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	Integer
type_of_meal_plan	Type of meal plan booked by the customer	String
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)	Integer
room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.	String
lead_time	Number of days between the date of booking and the arrival date	Integer
arrival_year	Year of arrival date	Integer
arrival_month	Month of arrival date	Integer
arrival_date	Date of the month	Integer
market_segment_type	Market segment designation.	String
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)	Integer
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking	Integer
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking	Integer
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)	Decimal
no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)	Integer
booking_status	Flag indicating if the booking was canceled or not.	String

Tiene 36275 rows × 19 columns

Key columns

- no_of_adults
- no_of_children
- no_of_week_nights
- no_of_weekend_nights
- room_type_reserved
- lead_time
- arrival_month
- arrival_date
- repeated_guest
- avg_price_per_room
- no_of_special_requests

Booking_status es la variable objetivo del análisis de estudio

Información del dataset

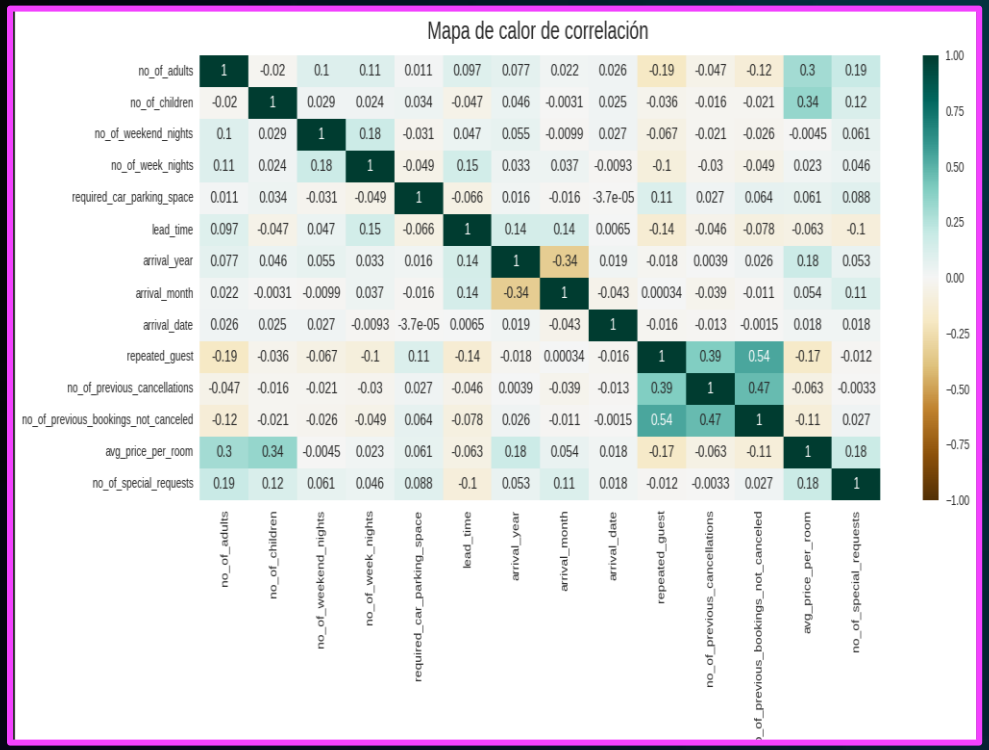
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

Conclusión: Se puede observar que en el dataset las columnas poseen las mismas filas (36275) y tiene 3 tipos de datos: float, int y object

EDA (exploratory data analysis)

05

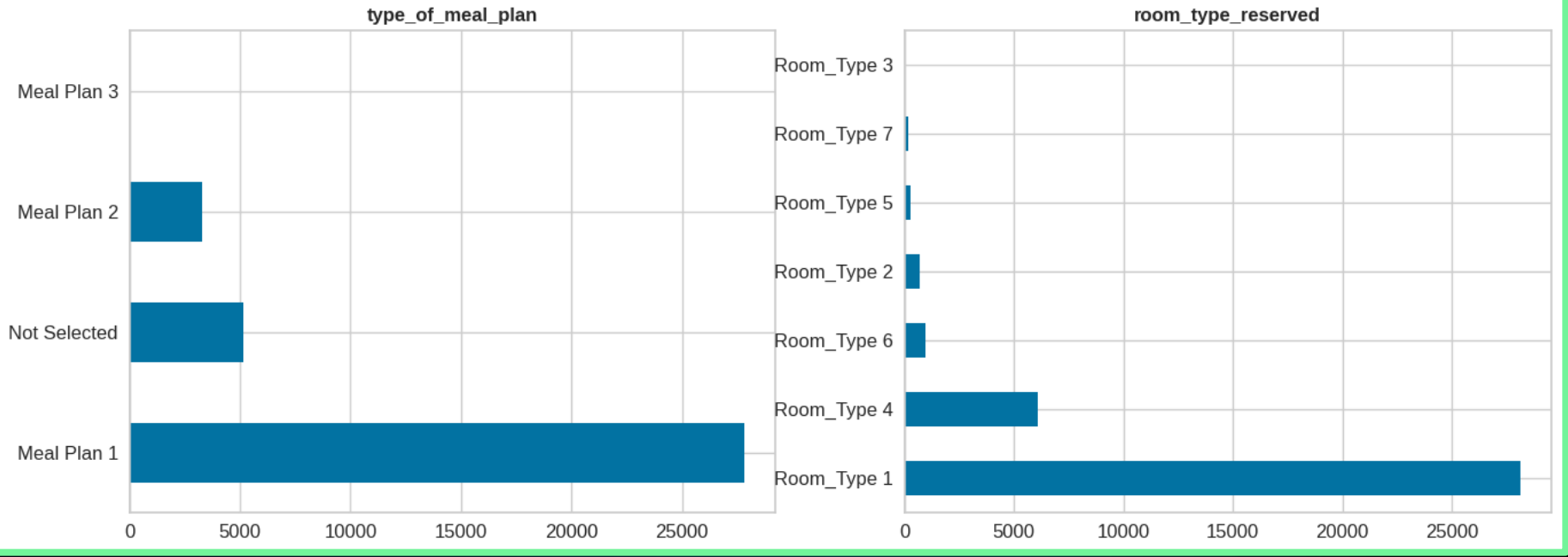
Análisis Multivariado: Gráfico de mapa de calor de correlación del dataset



Coef crr	Variable 1	Variable 2
0,47	no_of_previous_bookings_not_canceled	no_of_previous_cancellations
0,54	repeated_guest	no_of_previous_bookings_not_canceled

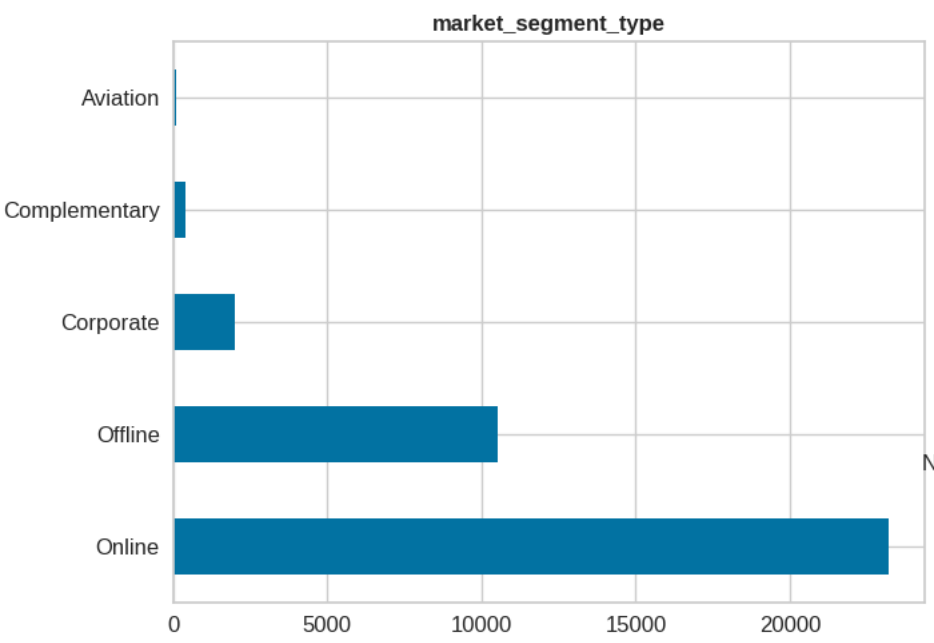
Las variables con mayor correlación

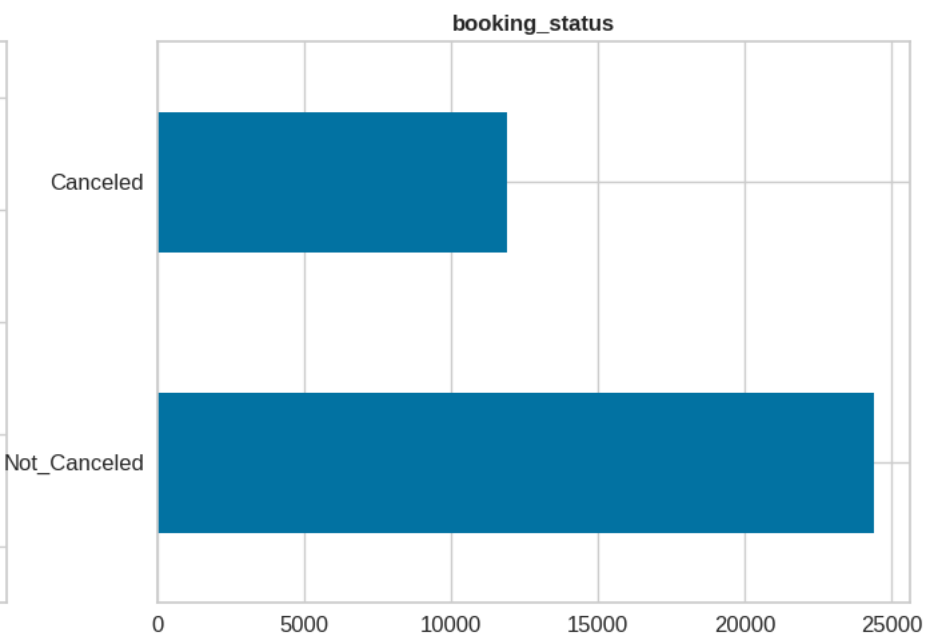
EDA: Variables Categóricas



Conclusión: **Room Type Reserved:** La mayoría de los clientes reservan las habitaciones de tipo 1

Type of meal plan: La mayoría de los clientes reservan el plan de comida tipo 1





Conclusión: **Booking Status:** Se observa una mayor proporción de reservas no canceladas con respecto a la proporción de reservas canceladas

Market Segment Type: Se observa que la mayoría de las reservas se realizan de manera online.

EDA: Variables numéricas

Data cleaning: Análisis de datos nulos

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

dtype: int64

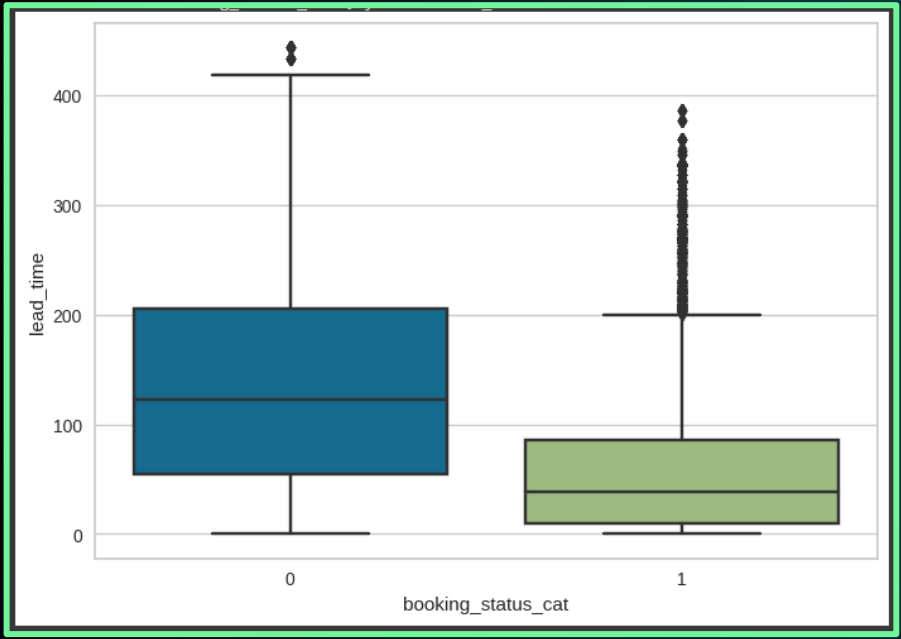
Conclusión: De acuerdo a los resultados no se observan datos nulos en ninguna de las columnas

Datección de outliers: valores extremos

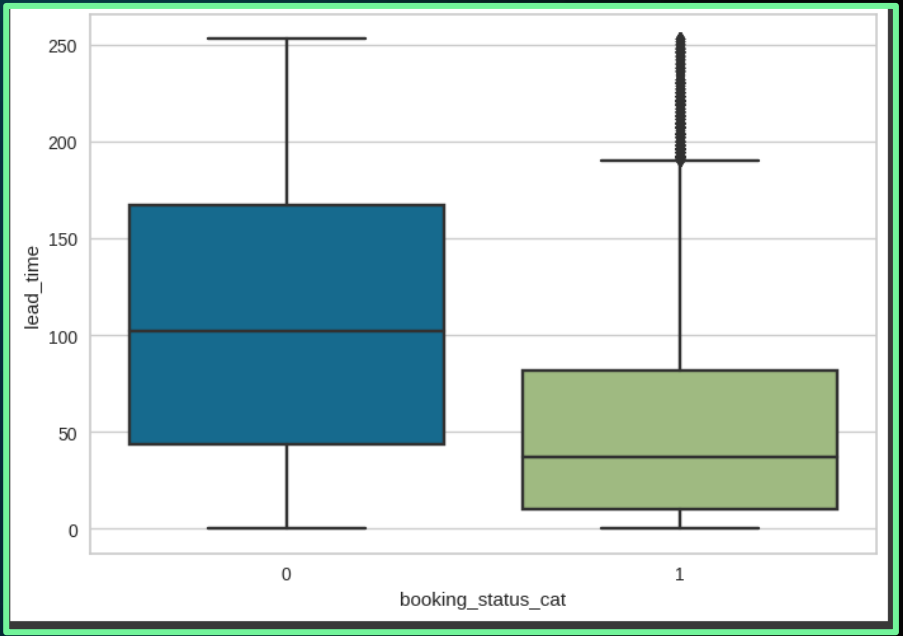
	count	mean	median	std	min	25%	50%	75%	max
no_of_adults	36275.0	1.844962	2.00	0.518715	0.0	2.0	2.00	2.0	4.0
no_of_children	36275.0	0.105279	0.00	0.402648	0.0	0.0	0.00	0.0	10.0
no_of_weekend_nights	36275.0	0.810724	1.00	0.870644	0.0	0.0	1.00	2.0	7.0
no_of_week_nights	36275.0	2.204300	2.00	1.410905	0.0	1.0	2.00	3.0	17.0
required_car_parking_space	36275.0	0.030986	0.00	0.173281	0.0	0.0	0.00	0.0	1.0
lead_time	36275.0	85.232557	57.00	85.930817	0.0	17.0	57.00	126.0	443.0
arrival_year	36275.0	2017.820427	2018.00	0.383836	2017.0	2018.0	2018.00	2018.0	2018.0
arrival_month	36275.0	7.423653	8.00	3.069894	1.0	5.0	8.00	10.0	12.0
arrival_date	36275.0	15.596995	16.00	8.740447	1.0	8.0	16.00	23.0	31.0
repeated_guest	36275.0	0.025637	0.00	0.158053	0.0	0.0	0.00	0.0	1.0
no_of_previous_cancellations	36275.0	0.023349	0.00	0.368331	0.0	0.0	0.00	0.0	13.0
no_of_previous_bookings_not_canceled	36275.0	0.153411	0.00	1.754171	0.0	0.0	0.00	0.0	58.0
avg_price_per_room	36275.0	103.423539	99.45	35.089424	0.0	80.3	99.45	120.0	540.0
no_of_special_requests	36275.0	0.619655	0.00	0.786236	0.0	0.0	0.00	1.0	5.0

Conclusión: De acuerdo a la observación de la tabla estadística, se puede ver que la variable **lead_time** posee valores extremos

- Realizar la detección de ouliers en "lead_time“:



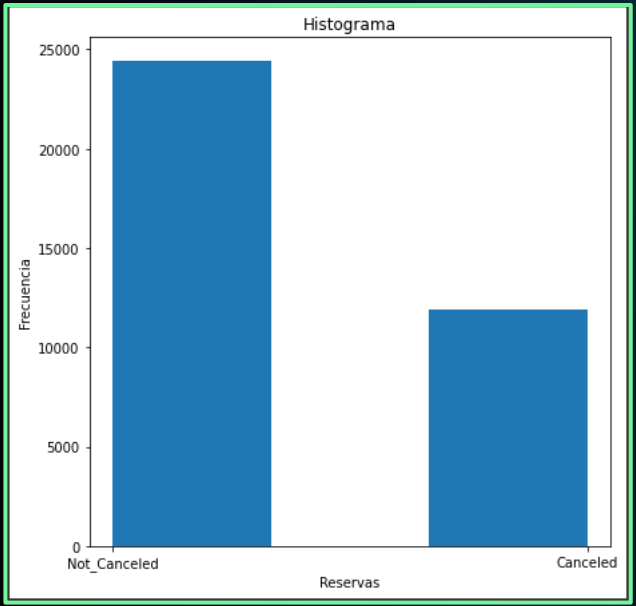
Comprobaremos que hemos eliminado los datos atípicos:

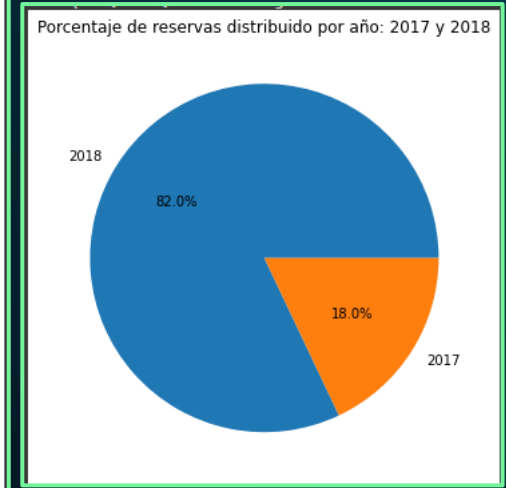


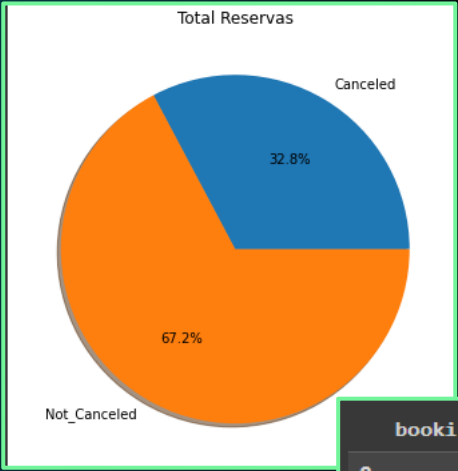
Conclusión: al comparar el boxplot de dos gráficos, la cual el de antes de realizar la detección de outliers, se observó muchos valores atípicos y el segundo gráfico donde se observa que ha sido eliminado la mayor parte de los valores extremos, mejorando los datos para realizar el siguiente paso que sería Machine Learning, ya que, si poseían datos atípicos, no iba a dar buenos resultados para modelo de entrenamiento y test (**malos datos generan malos modelos**)

Insights 06

¿Cuántas reservas no canceladas y canceladas se hicieron en los años 2017-2018?







	booking_status	Booking_ID
0	Canceled	11885
1	Not_Canceled	24390

Conclusión:

El total de las reservas de los años 2017 y 2018 fue: 36.275

En el año **2017** se hicieron 6514 reservas y en el año **2018** se hicieron 29761 reservas.

De acuerdo al total de reservas del período 2017-2018: el 18% de las reservas corresponde al año 2017 y el 82% de las reservas corresponde al año 2018.

De acuerdo al total de reservas del período 2017-2018: el 67,2% (24390) de las reservas fue confirmadas y 32,8% (11885) de las reservas fue canceladas.

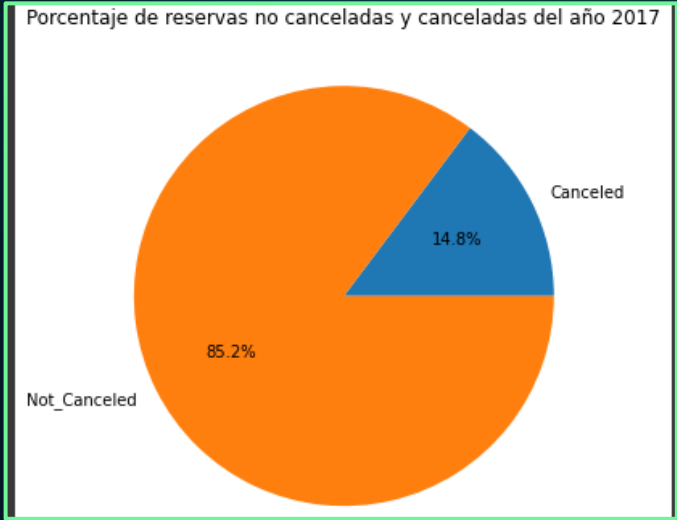
La tasa de cancelación es bastante alta. Se cancelan 11.885 de 36.275 reservas, es decir, el 32,8% de las reservas. **Desafortunadamente, no hay datos que puedan explicar por qué los clientes cancelan sus reservas.**

¿Cómo fue la situación de las reservas en el año 2017 y 2018?

Estudio de reservas de Hotel del año 2017

booking_status	arrival_year
Canceled	961
Not_Canceled	5553

arrival_month	arrival_year
7	363
8	1014
9	1649
10	1913
11	647
12	928



Conclusiones sobre la situación de las reservas del año 2017:

En el año 2017 hubo **5553 reservas confirmadas** la cual corresponde el 85,2% y **961 reservas canceladas** la cual corresponde el 14,8%

Se observa que los clientes realizaron reservas mayormente en los meses Septiembre y Octubre

Estudio de reservas de Hotel del año 2018

booking_status	arrival_year
Canceled	10924
Not_Canceled	18837

arrival_month	arrival_year
1	1014
2	1704
3	2358
4	2736
5	2598
6	3203
7	2557
8	2799
9	2962
10	3404
11	2333
12	2093

Conclusiones sobre la situación de las reservas del año 2018:
En el año 2018 hubo **18837 reservas confirmadas** la cual corresponde el 63,3% y **10924 reservas canceladas** la cual corresponde el 36,7%

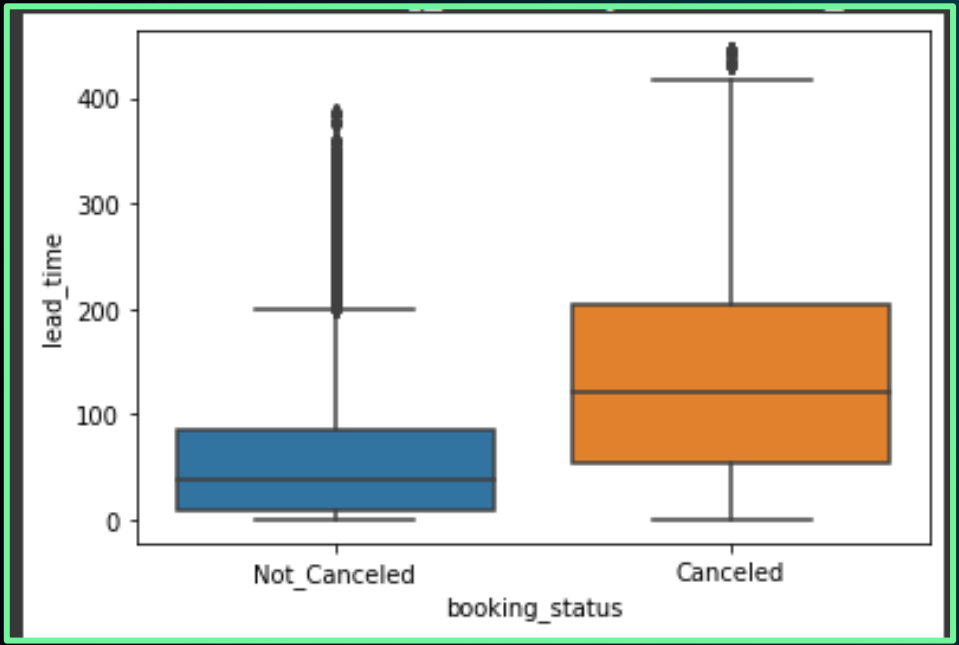
Comparando la situación de los años mencionados, se puede observar que en el año 2018 tiene mayor tasa de reservas canceladas con respecto al año 2017.

¿El precio medio por habitación es un factor que incide en las reservas?

	booking_status	avg_price_per_room
0	Canceled	110.589966
1	Not_Canceled	99.931412

Conclusión: Se observa que el precio medio por habitación cancelada es ligeramente superior al precio medio por habitación no cancelada

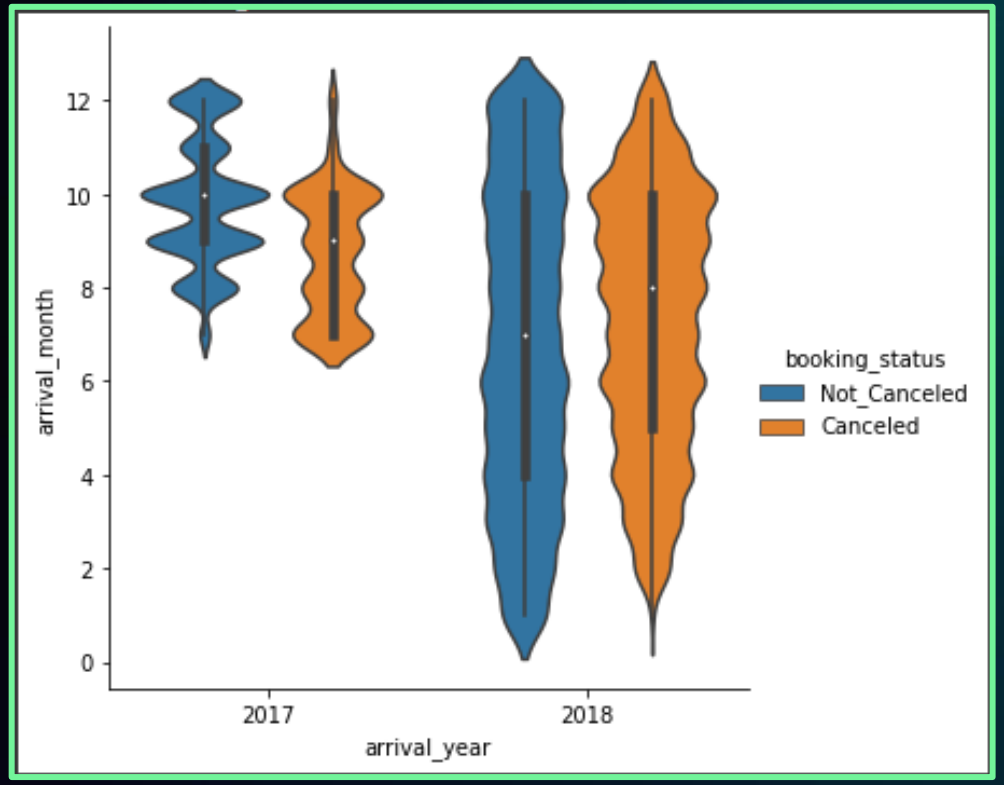
¿El plazo de reserva influye en la tasa de cancelación?



lead_time es la variable que se expresa en días, la cual se puede observar en el gráfico en función de la variable **booking_status**

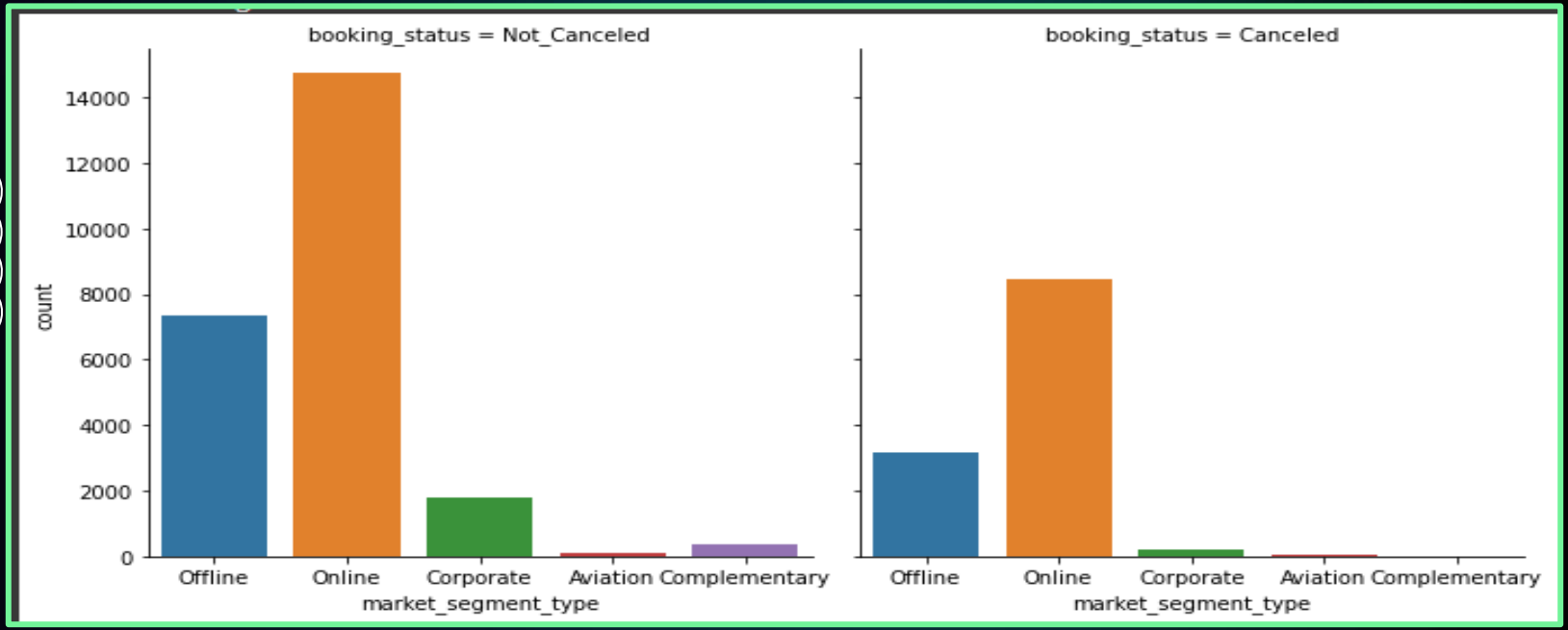
Conclusión: A mayor plazo, mayor tasa de cancelación

¿En qué mes o meses ocurrió la mayor tasa de cancelación de los años 2017-2018?



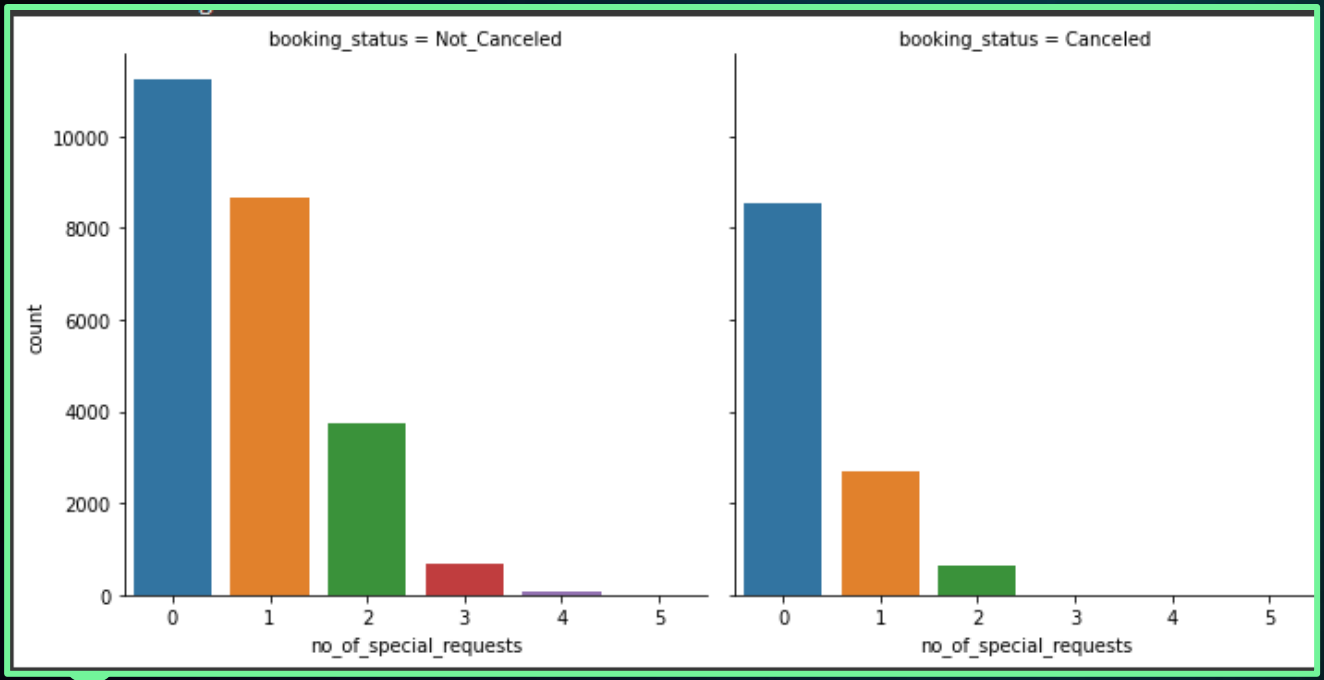
Conclusión: En los meses 7-8 de **2017** hay mayor tasa de cancelación mientras que a principios de **2018** hay menor tasa de cancelación

¿En qué segmento o por dónde realizan los clientes las reservas?



Conclusión: Los clientes realizan las reservas mayormente por online, seguido Offline, también lo hacen de la misma manera cuando cancelan las reservas

¿Qué tipo de solicitud especial fue mayor en las reservas confirmadas y canceladas de los años 2017-2018?



Conclusión:
Se puede apreciar que cuando los clientes tienen peticiones especiales hay menor tasa de cancelación de reservas. Los clientes que cancelan las reservan son mayormente los que no piden solicitudes especiales

Machine Learning 07

PASO 01

Modelamiento con 70/30 (train/test) con equilibrio de clases (sin OverSampling)

PASO 02

Modelamiento con optimización de hiperparámetros

PASO 03

Comparación de resultados

Tabla de comparación de modelos Machine Learning (con OverSampling)

Al aplicar OverSampling y luego modelamiento con diferentes modelos se obtuvo los siguientes resultados:

	Modelo	Precisión	Recall	F1-score	ROC-AUC
0	Naive Bayes	0.563812	0.144	0.56	0.800
1	Logistic Regression	0.766093	0.760	0.77	0.845
2	Random Forest classifier	0.926532	0.910	0.92	0.978
3	Extra Trees Classifier	0.933879	0.920	0.92	0.989

	Cross-Validation	K-folds
0	0.557	0.55
1	0.765	0.77
2	0.936	0.95
3	0.947	0.97

Conclusión: el mejor modelo para realizar la predicción es **"Extra Trees Classifier"**, también Random Forest dio excelente resultados.

Tabla de comparación de modelo Extra Trees Classifier con OverSampling y optimización con hiperparámetro

	Modelo	Precisión	Recall	F1-score	ROC-AUC
0	Naive Bayes	0.56	0.144	0.56	0.800
1	Logistic Regression	0.77	0.760	0.77	0.845
2	Random Forest classifier	0.93	0.910	0.92	0.978
3	Extra Trees Classifier Optimizado	0.93	0.920	0.93	0.989

	Cross-Validation	K-folds
0	0.557	0.55
1	0.765	0.77
2	0.936	0.95
3	0.951	0.97

	Modelo	Precisión	Recall	F1-score	ROC-AUC
0	Naive Bayes	0.563812	0.144	0.56	0.800
1	Logistic Regression	0.766093	0.760	0.77	0.845
2	Random Forest classifier	0.926532	0.910	0.92	0.978
3	Extra Trees Classifier	0.933879	0.920	0.92	0.989

	Cross-Validation	K-folds
0	0.557	0.55
1	0.765	0.77
2	0.936	0.95
3	0.947	0.97

Conclusión: No se observó diferencias significativas entre modelo ExtraTreesClassifier (con oversampling) con respecto a ExtraTrees optimizado

A continuación, no usaremos todas las variables características en la que vamos a aplicar feature selection:

Test de chi square χ2

Utilizaremos la prueba estadística chi-cuadrado para elegir las dos mejores características con mayor correlación con la variable objetivo:

```
Las mejores carasterísticas de la prueba de Chi2 son:

1. no_of_previous_bookings_not_canceled
2. no_of_special_requests
```

Mutual information

Podemos hacer lo mismo, mirando la información mutua y eligiendo las variables de información mutua con el objetivo:

```
Las mejores características de los criterios de información mutua son:
avg_price_per_room
booking_status_cat
```

Entonces, ¿cuál es? **Entre estas dos, chi2 e mutual information, ¿cuál es la mejor para la clasificación? Creemos un modelo para cada una de ellas y veamos su poder predictivo:**

```
# Calcular la métrica de chi2
scores = cross_val_score(logistic, X_ch2, y, cv=4, scoring='accuracy')
print('Chi2 mean accuracy: {:.1f}%'.format(scores.mean()*100))
```

```
Chi2 mean accuracy: 67.6%
```

Conclusión: La precisión de chi2 es de 67,6%

```
# Calcular la métrica de mutual information
scores = cross_val_score(logistic, X_mi, y, cv=4, scoring='accuracy')
print('Mutual-information mean accuracy: {:.1f}%'.format(scores.mean()*100))
```

```
Mutual-information mean accuracy: 63.9%
```

Conclusión: La precisión de mutual information es de 63,9%

Así pues, el criterio de selección de Test chi2 (67,6%) es mejor que mutual information (63,9%) con una diferencia de casi 4% en este caso. Sin embargo, la diferencia es poca y no se puede llegar a una conclusión significativa.

Con respecto a las diferencias entre usar todas las variables carasterísticas y al aplicar feature selection, la cual para el caso de:

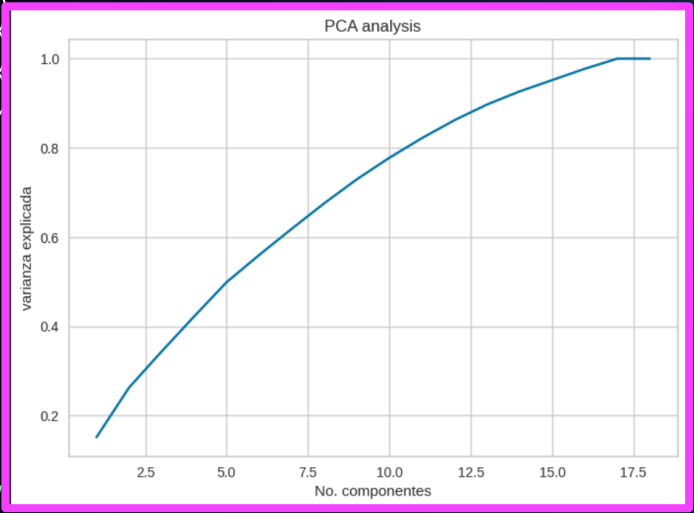
modelo logistic regression dió el 76,5%, es sólo casi un 10% menos que utilizar todas las características a la vez, lo que indica que no está muy bien y se recomienda usar todas las variables carasterísticas.

modelo Random forest dió el 92,47%, la cual es 30% menos que utilizar todas las características a la vez, lo que indica que no está muy bien y se recomienda usar todas las variables carasterísticas.

modelo Extra Trees Classifier dió el 93,38%, lo cual es 30% menos que utilizar todas las carasterísticas a la vez, lo que tampoco está muy bien y se recomienda a la vez usar todas las variables carasterísticas.

Para continuar nuestra búsqueda de un espacio de características más pequeño, apliquemos el análisis de componentes principales (PCA). El PCA requiere una estandarización de las características (la cual ya hemos estandarizado usando Standard Scaler).

- **Trazar la varianza explicada:** La "varianza explicada" cuantifica qué parte de la varianza de los datos explica cada "nuevo" espacio de características. En el siguiente gráfico, muestro esta cantidad en función del número de características PCA:



La varianza explicada alcanza el 90% con unos 14 componentes, pero sólo el 25% con dos componentes. Sin embargo, procedamos con sólo dos características. Una de las ventajas del PCA es que permite cuantificar la importancia relativa de cada una de las características transformadas en función de sus valores singulares.

- **La métrica de evaluación para PCA:**

```
scores = cross_val_score(logistic, X_pca_2, y, cv=4, scoring='accuracy')
print('PCA2 mean accuracy: {:.1f}%'.format(scores.mean()*100))
```

```
PCA2 mean accuracy: 63.8%
```

Conclusión: Con sólo dos características podemos obtener un 63,8% en comparación con el 67,6% seleccionado por el criterio de Test chi2. Sin embargo, no es una comparación justa porque el PCA crea nuevas características combinando datos de todas las características en cada una de las transformadas, de modo que las nuevas características contienen más información.

Conclusiones

08

Mirando la correlación, podemos ver que no hay una correlación fuerte entre nuestros datos. En nuestro EDA podemos ver que tenemos valores atípicos la cual en la sección de detección outliers se hizo tratamiento a la variable "lead_time" en la que se eliminó los datos atípicos muy extremos y quedaron algunos.

También podemos notar algunos patrones interesantes en nuestros datos en relación a las reservas, generalmente la mayoría de los datos tienen un patrón en las reservas, observando nuestra variable objetivo "booking_status" también podemos observar algunos patrones en las reservas que se cancelan o no, la más importante es la variable "lead_time", generalmente las reservas con un lead_time alto son más propensas a cancelarse, también tenemos otros datos interesantes, los huéspedes que se han alojado anteriormente y los huéspedes que solicitan una plaza de coche o hacen peticiones especiales también son menos propensos a cancelar.

Cuando nos fijamos en la variable Lead_Time, podemos ver un comportamiento similar a la variable Target, en general el mismo patrón de no cancelación tuvo un lead_time más corto, pero tenemos algunas especificidades, como en los 3 primeros y los 3 últimos meses del año el lead_time es menor.

Cuando lo comparamos con la variable coste medio de la habitación podemos ver que cuanto mayor es el precio de la habitación, y cuanto mayor es el lead_time, más probable es la cancelación de la habitación. Por último, también se observó que las reservas canceladas provinieron mayormente de las personas que reservan de manera online.

Por último, hemos realizado estudios con modelos de Machine Learning para poder realizar las predicciones con diferentes algoritmos, de acuerdo a los resultados se puede observar que el mejor modelo es ExtraTreesClassifier y Random Forest.

GRACIAS

Datos para mas info y consultas:

juancorzamar@gmail.com

github.com/juancorzamar93

linkedin.com/in/juanzamar