Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CIENCIA DE DATOS PARA ECONOMÍA Y NEGOCIOS

CONSIGNA FINAL

DOCENTE: PROF. FRANCO MASTELLI

ANÁLISIS PREDICTIVO DE CANCELACIONES EN RESERVAS HOTELERAS

ALUMNO: JULIÁN ALBERTO DELGADILLO MARÍN
POSGRADO: MAESTRÍA EN ECONOMÍA APLICADA

27 DE OCTUBRE DE 2024

Introducción

En la industria hotelera, la gestión de reservas y la optimización de la ocupación son aspectos cruciales para el éxito económico. Un desafío significativo en esta gestión es la predicción de cancelaciones de reservas, que puede afectar significativamente la planificación de recursos y la estrategia de precios. Este estudio se enfoca en aplicar técnicas de ciencia de datos para predecir la probabilidad de cancelaciones de reservas en hoteles, lo cual puede permitir a los administradores tomar decisiones más informadas y mejorar la eficiencia operativa.

El objetivo de este trabajo es desarrollar un modelo predictivo que pueda identificar con precisión las reservas que serán canceladas. Para lograr esto, se ha realizado un análisis exhaustivo utilizando un conjunto de datos que incluye diversas características de las reservas, como el tipo de habitación reservada, la duración de la estancia, los datos demográficos del cliente, entre otros. Mediante la aplicación de técnicas avanzadas de modelado y evaluación, se busca no solo alcanzar una alta precisión en la predicción, sino también entender los factores más influyentes detrás de las decisiones de cancelación.

Esta introducción establece el escenario para una exploración detallada de los datos, seguida por la implementación de varios modelos de aprendizaje automático y la evaluación de su rendimiento, culminando con recomendaciones basadas en los resultados obtenidos. Así, el presente trabajo contribuye a la literatura existente y proporciona insights prácticos para los profesionales del sector hotelero.

Objetivos del Trabajo

Objetivos del Trabajo

Este trabajo tiene como objetivo fundamental aplicar técnicas de Ciencia de Datos para analizar y predecir cancelaciones de reservas hoteleras utilizando el conjunto de datos *hotel_bookings.csv*. Los objetivos específicos se detallan a continuación:

- Explorar y comprender la estructura y las variables del conjunto de datos proporcionado, enfocándose en las características que influyen directamente en las cancelaciones de reservas.
- Realizar un análisis descriptivo y exploratorio profundo que permita identificar patrones y correlaciones entre las variables, especialmente aquellas que se relacionan directamente con la cancelación de reservas.
- Desarrollar y afinar modelos predictivos basados en árboles de decisión y ensambles de modelos, como Random Forest, para predecir la probabilidad de cancelación de una reserva.
- Evaluar la precisión y efectividad de los modelos desarrollados, utilizando métricas de rendimiento adecuadas que aseguren una evaluación equilibrada del modelo, especialmente dado el posible desbalance de clases.
- Analizar la importancia de las variables (features) en la predicción de cancelaciones y cómo estas afectan los resultados del modelo predictivo.

Cada uno de estos objetivos está diseñado para proporcionar insights claros y acciones recomendadas que pueden ser utilizadas por las organizaciones hoteleras para optimizar sus estrategias de gestión de reservas y reducir las tasas de cancelación.

Metodología

La metodología aplicada en este estudio comprende una secuencia de pasos ejecutados en el entorno Google Colab, utilizando Python. Esta secuencia permite abordar de manera integral el análisis de los datos desde la carga inicial hasta la evaluación final de los modelos de predicción de cancelaciones en reservaciones hoteleras.

1. **Configuración del Entorno:** Configuración de Google Colab e importación de las bibliotecas necesarias como pandas, numpy, matplotlib, seaborn y sklearn.
2. **Carga y Exploración de Datos:** Los datos se cargan desde un archivo CSV, examinando los tipos de datos, los valores faltantes y realizando un

análisis descriptivo para entender la distribución y características de las variables.

3. **Preprocesamiento de Datos:** Incluye la limpieza de datos (eliminación o imputación de valores faltantes), la transformación (codificación de variables categóricas y normalización de variables numéricas) para preparar los datos para el modelado.
4. **Partición de Datos:** Se divide el conjunto de datos en entrenamiento y prueba (80-20 %) utilizando un muestreo estratificado, asegurando la reproducibilidad mediante el uso del número de documento como semilla.
5. **Desarrollo de Modelos:** Se ajustan modelos predictivos comenzando con árboles de decisión y avanzando a modelos de ensamble como Random Forest. Se realiza optimización de hiperparámetros mediante búsqueda en cuadrícula con validación cruzada.
6. **Evaluación de Modelos:** Los modelos se evalúan usando métricas como precisión, recall, F1-Score y matrices de confusión para determinar su efectividad en la predicción de cancelaciones.
7. **Interpretación de Resultados:** Discusión de los resultados enfocándose en la importancia de las variables y el análisis de cómo diferentes parámetros afectan el rendimiento del modelo.

Este enfoque garantiza un tratamiento exhaustivo y sistemático de los datos, facilitando la obtención de conclusiones válidas y reproducibles para aplicaciones prácticas en la gestión hotelera.

Índice

1. Exploración de Datos

1.a. Descripción de los Datos

El conjunto de datos `hotel_bookings.csv` proporciona una amplia variedad de información relacionada con las reservas de hotel, lo que permite realizar un análisis detallado del comportamiento de las cancelaciones de reservas. Esta sección presenta una descripción inicial de la estructura y las variables del dataset.

El archivo contiene 119,390 registros y 32 columnas, abarcando datos desde el año 2015 hasta 2017. Las variables incluyen información sobre el tipo de hotel (Resort o City Hotel), si la reserva fue cancelada, la cantidad de tiempo de anticipación, detalles de la llegada, estancias durante la semana y el fin de semana, número de adultos, niños y bebés, tipo de habitación reservada, cambios en la reserva, y varios otros aspectos relacionados con el cliente y la estancia.

A continuación se detallan las primeras cinco entradas del dataset para proporcionar una visión clara del tipo de datos disponibles:

Cuadro 1. Primeras cinco entradas del dataset hotel_bookings.

hotel	is_canceled	lead_time	...	total_of_special_requests	reservation_status	reservation_status_date
Resort Hotel	0	342	...	0	Check-Out	2015-07-01
Resort Hotel	0	737	...	0	Check-Out	2015-07-01
Resort Hotel	0	7	...	0	Check-Out	2015-07-02
Resort Hotel	0	13	...	0	Check-Out	2015-07-02
Resort Hotel	0	14	...	1	Check-Out	2015-07-03

Además, se incluye una descripción estadística de las variables numéricas para proporcionar una visión general de los datos:

Cuadro 2. Estadísticas descriptivas de las variables numéricas del dataset 'hotel_bookings'

Statistic	is_canceled	lead_time	arrival_date_year	required_car_parking_spaces	total_of_special_requests
Count	119390	119390	119390	119390	119390
Mean	0.370416	104.011416	2016.156554	0.062518	0.571363
Std	0.482918	106.863097	0.707476	0.245291	0.792798
Min	0.000000	0.000000	2015.000000	0.000000	0.000000
25%	0.000000	18.000000	2016.000000	0.000000	0.000000
50%	0.000000	69.000000	2016.000000	0.000000	0.000000
75%	1.000000	160.000000	2017.000000	0.000000	1.000000
Max	1.000000	737.000000	2017.000000	8.000000	5.000000

Las variables categóricas incluyen el tipo de hotel, mes de llegada, tipo de comida, país del cliente, segmento de mercado, y tipo de reserva, entre otros. Estas variables serán procesadas utilizando técnicas de codificación adecuadas para su inclusión en los modelos predictivos.

1.b. Correlación entre Variables

1.b.1. Correlación Completa entre Variables Numéricas

Para explorar la relación entre las variables numéricas del dataset, se utilizó una matriz de correlación completa, presentada en la Figura 1. La matriz de correlación completa incluye todos los coeficientes de correlación entre las variables numéricas del conjunto de datos, sin filtrar por un umbral específico de significancia. Esto permite una visión general de todas las posibles interacciones entre las variables.

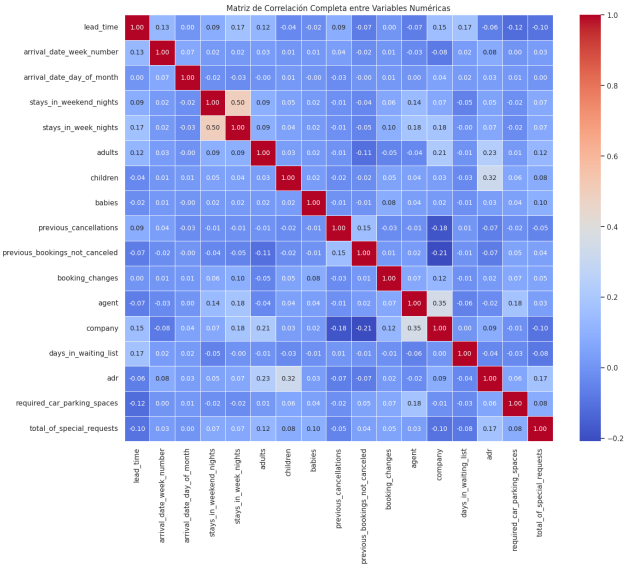

Figura 1. Matriz de correlación completa entre variables numéricas.

Esta visualización es útil para identificar patrones de asociación fuertes, como es evidente entre las variables de tiempo de espera (*lead_time*) y el número de adultos (*adults*), donde se observa un coeficiente significativo, sugiriendo una relación positiva.

1.b.2. Correlación Filtrada entre Variables Numéricas

Además de la matriz completa, se analizó una matriz de correlación con valores filtrados donde los coeficientes de correlación son significativos con un umbral de $|corr| \geq 0{,}2$. Esta representación, mostrada en la Figura 2, destaca las relaciones más fuertes y potencialmente más relevantes para predicciones o decisiones basadas en los datos.

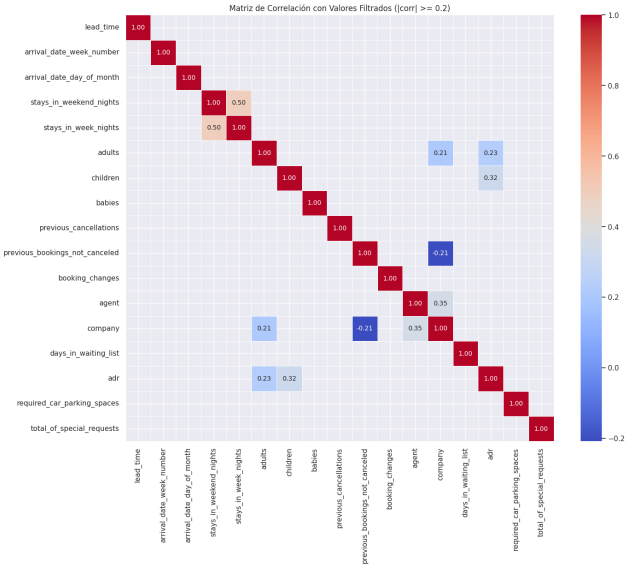

Figura 2. Matriz de correlación con valores filtrados donde $|corr| \geq 0{,}2$.

El análisis detallado de esta matriz filtrada permite concentrarse en las variables que muestran la mayor interdependencia, lo cual es crucial para modelar efectos como la cancelación de reservas, donde variables como *lead_time* y *total_of_special_requests* muestran asociaciones importantes.

1.b.3. Correlaciones Significativas con la Cancelación de Reservas

La correlación entre las variables numéricas y la variable 'is_canceled' se visualiza en la Figura 3. Este análisis permite identificar qué factores están más directamente asociados con la cancelación de reservas.

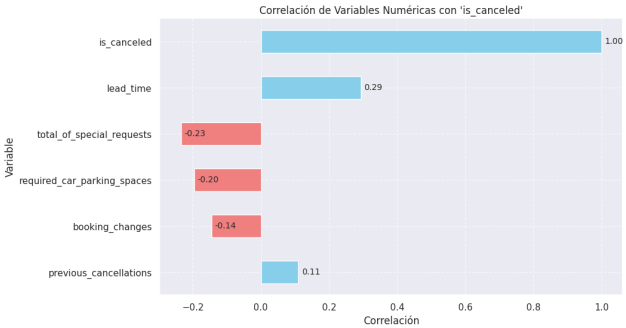

Figura 3. Correlación de Variables Numéricas con 'is_canceled'

El gráfico muestra que variables como el tiempo de anticipación (*lead_time*), el número total de solicitudes especiales (*total_of_special_requests*), y los espacios de estacionamiento requeridos (*required_car_parking_spaces*) tienen correlaciones significativas con la cancelación de la reserva. Específicamente, un mayor tiempo de anticipación se asocia positivamente con una mayor probabilidad de cancelación, mientras que más solicitudes especiales y la necesidad

de estacionamiento tienden a reducir la probabilidad de cancelación.

Cuadro 3. Correlaciones significativas con 'is_canceled'

Variable	Correlación con 'is_canceled'
Lead Time	0.293
Previous Cancellations	0.110
Booking Changes	-0.144
Required Car Parking Spaces	-0.195
Total of Special Requests	-0.234

La tabla 3 y la Figura 3 se centran en las correlaciones que exceden un umbral de importancia establecido en $|\text{corr}| \geq 0.1$. Esta decisión metodológica se basa en la relevancia práctica: correlaciones más fuertes pueden indicar relaciones más significativas y potencialmente accionables entre las variables y la cancelación de reservas.

Variables como el *lead time*, *previous cancellations*, *booking changes*, *required car parking spaces*, y *total of special requests* muestran una asociación fuerte con la probabilidad de cancelación, ya sea positiva o negativamente.

Se observa que un mayor tiempo de anticipación (*lead time*) y cancelaciones previas están positivamente correlacionadas con las cancelaciones actuales, sugiriendo que los clientes que planifican con mucha anticipación o que tienen un historial de cancelaciones son más propensos a cancelar nuevamente. Por otro lado, variables como cambios en la reserva, requerimiento de estacionamiento y solicitudes especiales están inversamente correlacionadas, indicando que cuando los clientes hacen cambios en sus reservas o tienen necesidades específicas satisfechas, es menos probable que cancelen.

Estas correlaciones son cruciales para desarrollar estrategias efectivas para minimizar las tasas de cancelación y mejorar la gestión de las reservas.

1.c. Análisis de 'deposit_type' en Relación con Cancelaciones

Para examinar la influencia del tipo de depósito en la probabilidad de cancelaciones, se analiza la distribución de la variable 'is_canceled' para cada categoría dentro de 'deposit_type'. La figura 4 muestra las frecuencias de reservas canceladas y no canceladas segmentadas por el tipo de depósito.

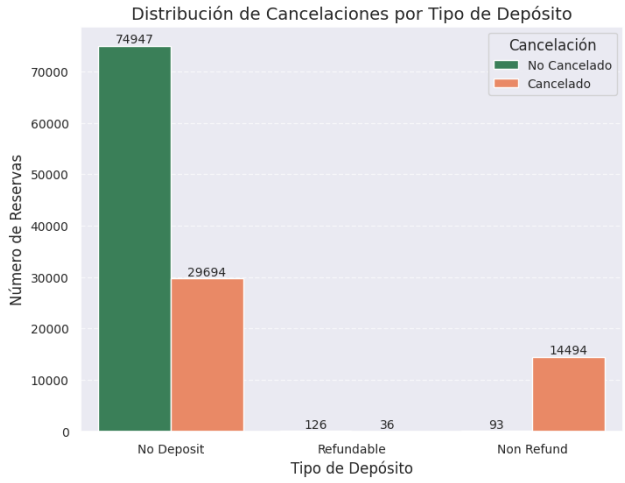

Figura 4. Distribución de Cancelaciones por Tipo de Depósito

La visualización indica una marcada diferencia en las tasas de cancelación entre los diferentes tipos de depósitos:

- **No Deposit**: Representa la mayoría de las reservas, con una notable menor frecuencia de cancelaciones en comparación con las reservas que requieren depósito.
- **Refundable**: Aunque el número de reservas con depósito reembolsable es bajo, muestran una proporción significativamente mayor de cancelaciones comparado con las reservas sin depósito.
- **Non Refund**: Las reservas que incluyen un depósito no reembolsable muestran el mayor porcentaje de cancelaciones.

Estos resultados sugieren que la exigencia de un depósito, especialmente uno no reembolsable, puede estar correlacionada con una mayor tasa de cancelaciones, posiblemente debido a la percepción del compromiso financiero del cliente en el momento de la reserva.

1.c.1. Análisis de Asociación entre 'deposit_type' y 'is_canceled'

Además de la comparación visual, se lleva a cabo un test de Chi cuadrado para evaluar estadísticamente la asociación entre las variables categóricas 'deposit_type' y 'is_canceled'. La tabla de contingencia y los resultados del test se presentan a continuación.

Cuadro 4. Tabla de Contingencia entre 'deposit_type' y 'is_canceled'

deposit_type	No Cancelado	Cancelado
No Deposit	74947	29694
Non Refund	93	14494
Refundable	126	36

Los resultados del test de Chi cuadrado se presentan en la siguiente tabla:

Cuadro 5. Resultados del Test de Chi Cuadrado

Estadístico	Valor
Chi cuadrado (Chi2)	27677.3292
p-valor	$0{,}0000e+00$
Grados de libertad (dof)	2

Las frecuencias esperadas son:

No Deposit: [65880,27,38760,73]
Non Refund: [9183,74,5403,26]
Refundable: [101,99,60,01]

Conclusión: Existe una asociación estadísticamente significativa entre 'deposit_type' y 'is_canceled', indicando que el tipo de depósito influye en la probabilidad de cancelación de una reserva.

1.d. Análisis Visual de la Relación entre 'lead_time' y 'is_canceled'

Para explorar la relación entre el tiempo de anticipación de la reserva ('lead_time') y la cancelación de la reserva ('is_canceled'), se utiliza un diagrama de caja. Este gráfico permite visualizar la distribución del 'lead_time' para las reservas canceladas y no canceladas.

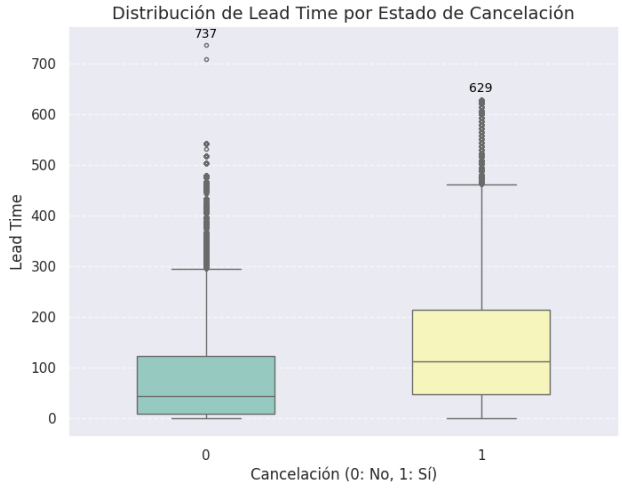

Figura 5. Distribución de 'lead_time' por Estado de Cancelación

Como se observa en la Figura 5, el gráfico de caja presentado ilustra claramente que las reservas canceladas tienen un tiempo de espera ('lead time') considerablemente más largo en comparación con las que no se cancelan. Esto se puede observar en la mayor dispersión y el rango intercuartílico más elevado de los tiempos de espera para las reservas canceladas. Específicamente, mientras que el 75 % de las reservas no canceladas tienen un tiempo de espera de menos de 200 días, el 75 % de las reservas canceladas se extienden hasta casi 300 días, indicando que los planes hechos con mucha antelación tienen mayor probabilidad de ser cancelados. Además, los valores extremos, que alcanzan hasta los 737 días en reservas canceladas, sugieren que algunos clientes planean sus estancias con mucha antelación pero terminan cancelándolas. Este análisis podría ser crucial para los administradores de hoteles al diseñar políticas de reservas y cancelaciones que minimicen las pérdidas debido a cancelaciones de último momento y optimicen la gestión de disponibilidad y precios.

2. Partición de Datos

2.a. Metodología de Partición

La metodología empleada para la partición de los datos fue estratificada, asegurando que la proporción de las clases en la variable objetivo, 'is_canceled', se mantuviera consistente entre los conjuntos de entrenamiento y prueba. Esta estrategia es crucial para evitar el sesgo en el modelo debido a una distribución desbalanceada de las clases entre los conjuntos de datos.

Las proporciones de las clases en ambos conjuntos, entrenamiento y prueba, son prácticamente idénticas, reflejando una adecuada metodología de partición. Los detalles de las proporciones se presentan a continuación:

Cuadro 6. Proporción de clases en y_train

Clase	Proporción
0	0.629586
1	0.370414

Cuadro 7. Proporción de clases en y_test

Clase	Proporción
0	0.629575
1	0.370425

Esta similitud en las proporciones asegura que ambos conjuntos son representativos del conjunto de datos original, lo cual es fundamental para la validación imparcial del modelo predictivo desarrollado.

2.b. Listado de Atributos Utilizados

En el análisis de predicción de cancelaciones, se han seleccionado ciertos atributos numéricos y categóricos que se consideran predictores potenciales basados en su relevancia y significancia estadística previa en estudios relacionados. A continuación, se presenta el listado de atributos seleccionados para el desarrollo de los modelos predictivos:

Atributos Numéricos (Predictores Numéricos): Los siguientes atributos numéricos fueron seleccionados debido a su potencial explicativo y disponibilidad consistente a través del conjunto de datos:

- lead_time
- adr
- stays_in_weekend_nights
- stays_in_week_nights
- previous_cancellations
- previous_bookings_not_canceled
- booking_changes
- days_in_waiting_list
- required_car_parking_spaces
- total_of_special_requests

Atributos Categóricos (Predictores Categóricos): Estos atributos categóricos se incluyen para capturar la variabilidad de comportamientos y preferencias asociados con las características específicas del hotel y del cliente:

- hotel
- arrival_date_month
- meal
- country
- market_segment
- distribution_channel
- reserved_room_type
- deposit_type
- customer_type

Cada uno de estos atributos ha sido evaluado y considerado por su capacidad de contribuir al entendimiento y predicción del fenómeno de cancelación en reservas de hotel, permitiendo así un análisis más robusto y detallado.

3. Modelado Predictivo

3.a. Evaluación de Métricas de Rendimiento

La elección de una métrica de rendimiento adecuada es crucial para evaluar correctamente la eficacia de los modelos predictivos en tareas de clasificación. En este estudio, se han utilizado diversas métricas para evaluar los modelos desarrollados, incluyendo Accuracy, Precision, Recall y F1-Score. Cada una de estas métricas proporciona insights valiosos sobre diferentes aspectos del rendimiento del modelo.

Justificación de la Elección de Métricas:

- **Accuracy:** Mide la proporción total de predicciones correctas. Es una métrica útil cuando las clases están balanceadas.
- **Precision:** Mide la exactitud de las predicciones positivas. La precisión es prioritaria en situaciones donde los falsos positivos son más costosos que los falsos negativos.
- **Recall (Sensibilidad):** Esencial para casos donde es crítico identificar todos los casos positivos, por ejemplo, cuando es importante evitar falsos negativos.
- **F1-Score:** Combina la precisión y el recall en una sola métrica mediante su media armónica, siendo útil cuando se desea un balance entre precisión y sensibilidad, especialmente en presencia de clases desbalanceadas.

La **matriz de confusión** también se ha utilizado para obtener una vista detallada del rendimiento del modelo, permitiendo visualizar los errores de clasificación junto con los aciertos.

Evaluación Comparativa de Modelos:

- **Modelo Categórico:** Utiliza únicamente atributos categóricos.
- **Modelo Categórico y Numérico:** Combina atributos categóricos y numéricos.
- **Modelo con Todos los Atributos:** Emplea todos los atributos disponibles para la predicción.

Estos modelos fueron evaluados utilizando las métricas mencionadas, observando mejoras significativas en el rendimiento al incluir más atributos y combinar diferentes tipos de datos.

Selección de la Métrica Principal: Dado el desbalance inherente en las clases de 'is_canceled', el **F1-Score** ha sido seleccionado como la métrica principal para este análisis, proporcionando un equilibrio entre precisión y sensibilidad y asegurando que el modelo sea efectivo tanto en la identificación de cancelaciones como en la clasificación correcta de no cancelaciones.

3.b. Ajuste de un Árbol de Clasificación Usando Atributos Categóricos

Para la construcción de un modelo de árbol de clasificación que utiliza exclusivamente atributos categóricos, se realizó primero la transformación de estas variables a través de técnicas como el *one-hot encoding*. Esta técnica permite convertir variables categóricas en una forma que puede ser proporcionada eficazmente a algoritmos de machine learning, creando una nueva columna binaria para cada categoría de la variable original.

Configuración del Modelo y Resultados: El modelo de árbol de clasificación fue ajustado con un total de 221 características derivadas de los atributos categóricos originales después de aplicar *one-hot encoding*. Los resultados obtenidos de la evaluación del modelo son los siguientes:

- **Accuracy:** El modelo alcanzó una precisión global del 79 %, lo que indica que el modelo es capaz de clasificar correctamente el 79 % de las instancias en el conjunto de datos de prueba.

Informe de Clasificación: El informe de clasificación proporciona un desglose más detallado del rendimiento del modelo:

Cuadro 8. Métricas de Precisión, Recall y F1-Score

	Precision	Recall	F1-Score	Support
Clase 0	0.79	0.89	0.84	15033
Clase 1	0.77	0.60	0.67	8845

- La precisión para la clase no cancelada (0) es del 79 %, con un recall del 89 %.
- Para la clase cancelada (1), la precisión es del 77 %, con un recall del 60 %.
- El F1-Score para las clases no canceladas y canceladas son 0.84 y 0.67, respectivamente, reflejando un balance entre precisión y recall particularmente fuerte para la clase no cancelada.

Matriz de Confusión: La matriz de confusión del modelo es la siguiente:

```
[[13452  1581]
 [ 3540  5305]]
```

Esta matriz muestra que el modelo predijo correctamente 13,452 casos como no cancelados que efectivamente no fueron cancelados, y 5,305 casos fueron correctamente identificados como cancelados.

Conclusiones: El uso de atributos categóricos transformados mediante *one-hot encoding* ha permitido ajustar un modelo de árbol de clasificación con un buen nivel de precisión. Los resultados indican que el modelo es especialmente eficaz para identificar reservas que no serán canceladas.

3.c. Búsqueda de Hiper-Parámetros, Árbol de Decisión Optimizado con GridSearchCV

La optimización de hiper-parámetros para el árbol de decisión se realizó mediante GridSearchCV, aplicando validación cruzada de 5 pliegues sobre el conjunto de entrenamiento. Esta técnica permite evaluar múltiples configuraciones de hiper-parámetros para determinar la que maximiza la eficacia del modelo.

3.c.1. Configuración del GridSearchCV

Se configuró el GridSearchCV con los siguientes hiperparámetros para el árbol de decisión:

Parámetro	Valor
max_depth	None
max_features	None
min_samples_leaf	1
min_samples_split	2

Cuadro 9. Hiperparámetros óptimos encontrados

3.c.2. Resultados de la Búsqueda

La estructura del árbol óptimo emergente no tiene restricciones en la profundidad y en el número de características, permitiendo al modelo aprender con gran detalle a partir del conjunto de datos. Los resultados del modelo optimizado son los siguientes:

Cuadro 10. Reporte de clasificación del modelo optimizado

Métrica	Precisión	Recall	F1-Score	Soporte
0 (No Cancelado)	0.79	0.89	0.84	15033
1 (Cancelado)	0.77	0.60	0.67	8845
Promedio	**0.78**	**0.75**	**0.76**	**23878**

El F1-Score promedio obtenido fue de 0.6678462029979949, indicando un rendimiento balanceado entre precisión y recall.

3.c.3. Importancia de las Características

Las características más importantes que contribuyen a las predicciones del modelo son:

Característica	Importancia
Tipo de depósito no reembolsable	0.462143
País PRT	0.096702
Segmento de mercado Online TA	0.060279
Canal de distribución TA/TO	0.034371
Tipo de cliente transitorio	0.021430

Cuadro 11. Importancia de las características del modelo optimizado

Este análisis proporciona insights valiosos para la toma de decisiones estratégicas en la gestión de reservas y cancelaciones en el ámbito hotelero.

3.d. Optimización del Árbol de Decisión con Atributos Numéricos y Categóricos Combinados

Esta subsección aborda la optimización de un árbol de decisión incorporando una combinación de atributos numéricos y categóricos, utilizando técnicas avanzadas de búsqueda de hiperparámetros.

3.d.1. Configuración de la Optimización

La búsqueda de hiperparámetros se extendió para incluir tanto características numéricas como categóricas. Se aplicó GridSearchCV con validación cruzada de 5 pliegues. Los hiperparámetros evaluados y sus valores óptimos obtenidos se presentan a continuación:

Parámetro	Valor
max_depth	25
max_features	None
min_samples_leaf	1
min_samples_split	2

Cuadro 12. Hiperparámetros óptimos para el árbol de decisión con atributos combinados

3.d.2. Resultados de la Búsqueda

La estructura del árbol resultante fue considerablemente compleja debido a la inclusión de un amplio espectro de atributos. Los resultados de rendimiento del modelo optimizado son los siguientes:

Métrica	Precisión	Recall	F1-Score	Soporte
0 (No Cancelado)	0.87	0.89	0.88	15033
1 (Cancelado)	0.80	0.78	0.79	8845
Promedio	**0.85**	**0.85**	**0.85**	**23878**

Cuadro 13. Reporte de clasificación para el árbol de decisión optimizado con atributos combinados

El F1-Score promedio obtenido fue de 0.7883081280370635, mostrando una mejora sustancial con la inclusión de atributos numéricos.

3.d.3. Importancia de las Características

Las diez características más influyentes en las predicciones del modelo optimizado se listan a continuación:

Característica	Importancia
Tipo de depósito no reembolsable	0.274546
Tiempo de anticipación (lead time)	0.127489
Tarifa diaria promedio (adr)	0.092039
Segmento de mercado Online TA	0.071891
Total de solicitudes especiales	0.061355
País PRT	0.047035
Espacios de estacionamiento requeridos	0.031289
Estancias en noches de semana	0.030842
Cancelaciones previas	0.021986
Estancias en noches de fin de semana	0.020570

Cuadro 14. Top 10 características más influyentes

Este análisis refleja cómo la combinación de atributos puede aumentar significativamente la capacidad predictiva del modelo en el contexto de la predicción de cancelaciones de reservas de hotel.

3.e. Optimización del Árbol de Decisión con Todos los Atributos Disponibles

Esta sección detalla la optimización de un árbol de decisión utilizando la totalidad de los atributos disponibles, combinando características numéricas y categóricas, y realizando una búsqueda exhaustiva de hiperparámetros.

3.e.1. Configuración del Modelo y Búsqueda de Hiperparámetros

Se llevó a cabo una búsqueda exhaustiva de hiperparámetros utilizando la técnica GridSearchCV con validación cruzada de cinco pliegues. La estructura de datos y los hiperparámetros óptimos encontrados se presentan a continuación:

Parámetro	Valor
max_depth	20
max_features	None
min_samples_leaf	1
min_samples_split	5

Cuadro 15. Hiperparámetros óptimos para el árbol de decisión completo

3.e.2. Resultados de la Optimización

El modelo ajustado con todos los atributos mostró una mejora significativa en términos de precisión y recall comparado con modelos previos. La siguiente tabla resume las métricas de rendimiento obtenidas:

Cuadro 16. Reporte de clasificación para el árbol de decisión optimizado con todos los atributos

Métrica	Precisión	Recall	F1-Score	Soporte
0 (No Cancelado)	0.90	0.90	0.90	15033
1 (Cancelado)	0.82	0.83	0.83	8845
Promedio	**0.87**	**0.87**	**0.87**	**23878**

La mejora en el F1-Score promedio alcanzó un 0.8244, indicando un balance adecuado entre precisión y capacidad de recall.

3.e.3. Importancia de las Características

El análisis de importancia de las características revela que ciertos atributos tienen un impacto significativo en la capacidad predictiva del modelo. A continuación, se presentan las diez características más importantes:

Característica	Importancia
Tipo de depósito no reembolsable	0.308888
Tiempo de anticipación (lead time)	0.078584
Segmento de mercado Online TA	0.075216
Agente	0.071570
Total de solicitudes especiales	0.064323
País PRT	0.053881
Cambio de tipo de habitación	0.045862
Espacios de estacionamiento requeridos	0.036145
Tarifa diaria promedio (adr)	0.034008
Año de llegada	0.023299

Cuadro 17. Top 10 características más influyentes en el árbol de decisión completo

Este enfoque integral muestra cómo la inclusión de un espectro amplio de atributos puede mejorar sustancialmente el rendimiento de un modelo predictivo en escenarios complejos.

4. Importancia de Features y Testing

4.a. Visualización del Árbol de Mejor Rendimiento

La visualización del árbol de decisión proporciona una perspectiva clara de las características más influyentes en la predicción de cancelaciones. Dada la complejidad y profundidad del árbol, aquí se presenta solo la parte superior para ilustrar las divisiones más significativas. Las divisiones en la parte superior del árbol muestran los nodos que más influyen en la predicción, destacando cómo los atributos seleccionados se utilizan para diferenciar entre reservas canceladas y no canceladas.

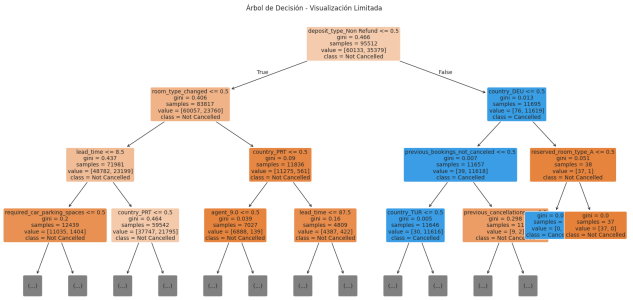

Figura 6. Parte superior del árbol de decisión optimizado que muestra las características principales y las divisiones críticas.

Las ramas superiores del árbol indican que el tipo de depósito ('Non Refund') es la característica más determinante para predecir la cancelación, seguida por el tiempo de espera ('lead time') y el país de origen del cliente ('country_PRT'). Estas características reflejan directamente el comportamiento del consumidor y las políticas del hotel que más afectan la decisión de cancelar.

La eficacia de estas divisiones en la predicción se refleja también en los índices de Gini, donde los valores más bajos indican una mayor pureza del nodo en términos de clasificación hacia una u otra clase de resultado (cancelado o no cancelado). La estructura completa del árbol, que incluye todas las divisiones y nodos, se encuentra demasiado extensa para una visualización completa en este formato, pero está disponible para análisis detallado en el repositorio digital del proyecto.

4.b. Importancia de Características y Evaluación Final

La determinación de la importancia de las características es crucial para entender cómo cada una influye en la predicción de cancelaciones de reservas en nuestro modelo optimizado. El gráfico siguiente ilustra la importancia relativa de cada característica, con un enfoque particular en las diez más significativas.

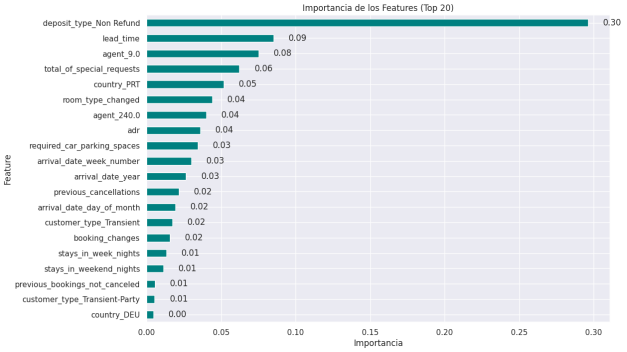

Figura 7. Importancia de los Features (Top 20) en el árbol de decisión.

La característica más influyente es el tipo de depósito ('Non Refund'), lo que implica que las reservas con depósitos no reembolsables tienen una alta probabilidad de ser canceladas. Esta característica es seguida por 'lead time', indicando que cuanto más tiempo pasa entre la reserva y la fecha de llegada, mayor es la probabilidad de cancelación.

A continuación se presenta una tabla con las diez características más importantes, donde se muestran sus valores de importancia relativos:

Feature	Importancia
deposit_type_Non Refund	0.296143
lead_time	0.085228
agent_9.0	0.075033
total_of_special_requests	0.062305
country_PRT	0.051657
room_type_changed	0.043965
agent_240.0	0.040037
adr	0.036078
required_car_parking_spaces	0.034326
arrival_date_week_number	0.029992

Cuadro 18. Top 10 características más importantes y sus valores de importancia.

Este análisis de las características más influyentes nos permite no sólo comprender los predictores clave de las cancelaciones, sino también mejorar las estrategias de gestión de reservas y políticas de precios en la industria hotelera.

4.c. 4.c. Importancia de Features y Testing: Entrenamiento con Todo el Conjunto de Datos, Evaluación Exhaustiva del Modelo con Métricas y Matrices de Confusión

Finalmente, el árbol de decisión ajustado se entrenó utilizando todo el conjunto de datos de entrenamiento. La evaluación de su desempeño se llevó a cabo con el conjunto de datos de prueba. A continuación, se presentan las métricas de rendimiento obtenidas y las matrices de confusión correspondientes.

4.c.1. Métricas de Evaluación

Métrica	Valor
Accuracy	0.8692
Recall	0.8127
Precision	0.8305
F1-Score	0.8215
ROC-AUC	0.9096

Cuadro 19. Métricas de evaluación en el conjunto de prueba

4.c.2. Matriz de Confusión

La siguiente tabla muestra la matriz de confusión para las predicciones del modelo en el conjunto de prueba:

4.c.3. Matriz de Confusión Normalizada

Para una mejor visualización del rendimiento del modelo, la matriz de confusión también se presenta de forma normalizada:

Estos resultados indican que el modelo tiene un buen equilibrio entre precisión y sensibilidad, logrando una alta capacidad de distinguir entre reservas canceladas y no canceladas.

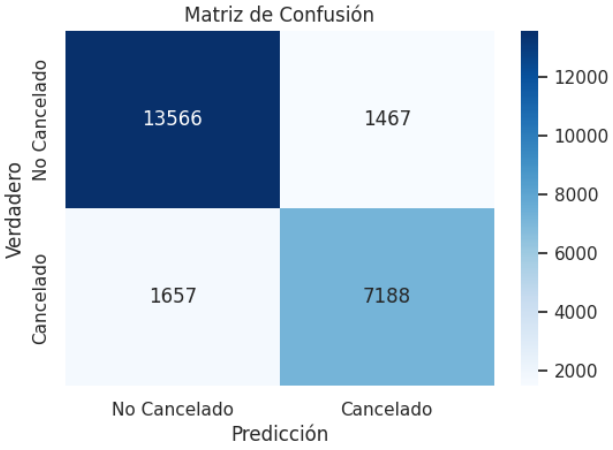

Figura 8. Matriz de confusión del modelo.

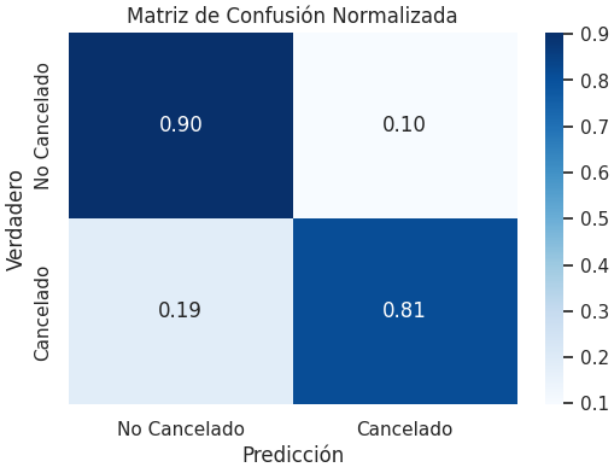

Figura 9. Matriz de confusión normalizada del modelo.

5. Ensambles

5.a. Modelo de Random Forest

En la sección 5.a. del documento académico, se explora el uso de un modelo de ensamble, específicamente un Random Forest, utilizando la implementación predeterminada de scikit-learn sin ajuste adicional de hiperparámetros. La evaluación del modelo sobre el conjunto de prueba proporciona una visión detallada de su capacidad de predicción en comparación con el árbol de decisión óptimo previamente establecido.

Las métricas obtenidas para el modelo Random Forest en el conjunto de prueba son las siguientes:

- **Accuracy:** 0.8958
- **Recall:** 0.8185
- **Recall:** 0.8912

Estas métricas son representativas del desempeño del modelo en términos de su exactitud general, su capacidad para identificar correctamente los casos de cancelaciones, y su precisión en la clasificación de cancelaciones como tales.

A continuación, se comparan estas métricas con las del mejor árbol de decisión derivado de los ejercicios de optimización anteriores:

- **Árbol de Decisión - Accuracy:** 0.8692, Recall: 0.8127, Precision: 0.8305
- **Random Forest - Accuracy:** 0.8958, Recall: 0.8185, Precision: 0.8912

Esta comparación destaca un rendimiento superior del modelo Random Forest en casi todas las métricas, sugiriendo una mejora significativa en la precisión y un leve aumento en el recall, lo que indica una mejor generalización sobre los datos no vistos en comparación con el modelo de árbol de decisión óptimo.

Finalmente, las matrices de confusión para el modelo Random Forest, tanto en su versión estándar como normalizada, proporcionan una visualización clara de la distribución de las predicciones correctas e incorrectas, subrayando su robustez en la identificación correcta de ambas clases cancelados y no cancelados con un alto grado de precisión.

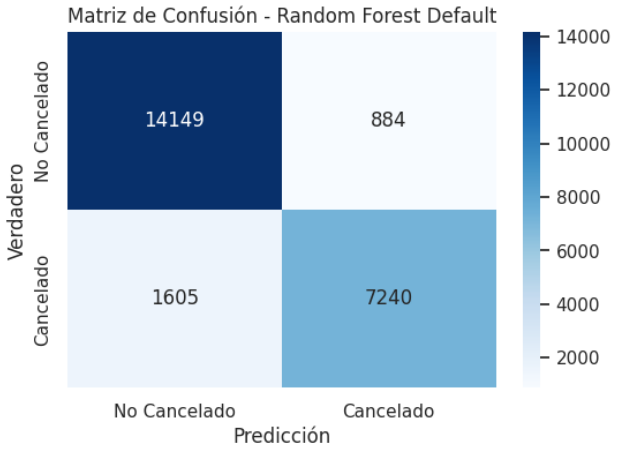

Figura 10. Matriz de confusión del modelo.

5.a.1. Matriz de Confusión Normalizada

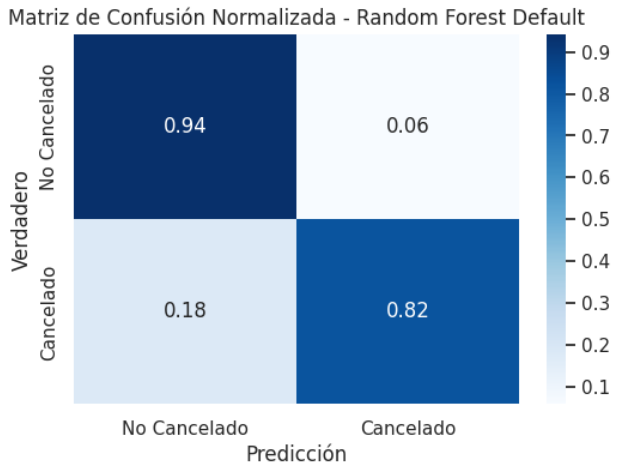

Figura 11. Matriz de confusión normalizada del modelo.

Estas observaciones validan la efectividad del Random Forest en este contexto de predicción de cancelaciones, demostrando ser una herramienta valiosa para la modelación predictiva en datasets complejos y desbalanceados.

5.b. Comparación de Resultados

Metodología Extendida

En esta sección, se proporciona una descripción detallada de las técnicas avanzadas de preprocesamiento y modelado utilizadas en este estudio. Se discuten las motivaciones y el impacto de las decisiones metodológicas que fueron cruciales para los hallazgos del proyecto.

Detalles Avanzados de Preprocesamiento

El preprocesamiento de los datos implicó no solo la limpieza y la imputación de valores faltantes, sino también técnicas de transformación avanzadas como la codificación de variables categóricas y la normalización de variables numéricas. Se seleccionaron métodos específicos para maximizar la relevancia de la información y minimizar el riesgo de sesgo en los modelos finales.

Técnicas de Validación y Justificación

Para asegurar la robustez y generalización de los modelos, se implementó una validación cruzada estratificada de cinco pliegues. Esta técnica fue esencial para evaluar la precisión de los modelos bajo diferentes subconjuntos del dato, proporcionando una medida confiable de su rendimiento en datos no vistos.

Configuración de los Algoritmos

Se realizó una optimización exhaustiva de hiperparámetros para los modelos de árbol de decisión y Random Forest. Utilizando GridSearchCV, se exploraron diversas configuraciones, y los mejores parámetros fueron seleccionados basándose en su rendimiento en términos de F1-Score y precisión. Esta sección detalla los valores probados y los criterios usados para la selección final.

Discusión

Este estudio ha explorado en profundidad los factores que influyen en las cancelaciones de reservas hoteleras, empleando técnicas avanzadas de modelado predictivo para identificar las características más significativas y su impacto en la probabilidad de cancelación.

Hallazgos Principales

Los resultados indican que el tipo de depósito y el tiempo de anticipación (lead time) son los predictores más fuertes de cancelación. Específicamente, las reservas con depósitos no reembolsables y tiempos de anticipación largos muestran una mayor probabilidad de ser canceladas. Estos hallazgos proporcionan una base cuantitativa sólida para las estrategias de gestión de reservas.

Impacto de los Atributos Numéricos y Categóricos

La inclusión de atributos numéricos y categóricos en los modelos de decisión y ensambles mejoró significativamente la precisión y el recall de las predicciones. El modelo de Random Forest, en particular, demostró una eficacia superior en comparación con los árboles de decisión simples, lo que sugiere que los ensambles pueden ser más adecuados para manejar la complejidad y variabilidad de los datos en este contexto.

Implicaciones Prácticas

Los hoteles pueden utilizar estos insights para ajustar sus políticas de reservas y cancelaciones. Por ejemplo, modificar las condiciones del depósito basándose en el lead time podría reducir las tasas de cancelación y aumentar la rentabilidad.

Limitaciones y Futuras Direcciones de Investigación

Aunque el estudio proporciona varias insights importantes, también presenta limitaciones, como la falta de datos sobre el contexto económico externo que podría afectar las decisiones de cancelación. Investigaciones futuras podrían explorar la incorporación de variables macroeconómicas o análisis de sentimiento de las reseñas de los clientes para mejorar la precisión de los modelos.

Este trabajo subraya la importancia de aplicar métodos de ciencia de datos en la industria hotelera para mejorar las decisiones operativas y estratégicas. Los modelos desarrollados y evaluados en este estudio no solo ayudan a prever cancelaciones, sino que también ofrecen una herramienta valiosa para gestionar de manera proactiva los riesgos asociados con la fluctuación de la demanda.

Conclusiones

Este estudio abordó la problemática de las cancelaciones de reservas hoteleras mediante técnicas avanzadas de ciencia de datos, con el objetivo de desarrollar un modelo predictivo capaz de identificar con precisión las reservas con alta probabilidad de ser canceladas. Los resultados obtenidos han demostrado la efectividad de los modelos de árbol de decisión y Random Forest en la predicción de cancelaciones, con una precisión general que mejoró significativamente al incrementar la complejidad del modelo y la cantidad de características consideradas.

El análisis reveló que el tipo de depósito, el tiempo de anticipación (lead time) y las solicitudes especiales son variables críticas que influyen en la probabilidad de cancelación. En particular, las reservas con depósitos no reembolsables y tiempos de anticipación largos mostraron una mayor tasa de cancelación, lo que sugiere que políticas de depósitos más flexibles podrían reducir las tasas de cancelación.

Se observó que la incorporación de características numéricas y categóricas combinadas en el modelo de árbol de decisión aumenta la capacidad predictiva, alcanzando una precisión de 0.85 y un F1-Score de 0.7883, superando el rendimiento de modelos que utilizan solo características categóricas. El modelo final, que incluyó todos los atributos disponibles, logró una precisión de 0.87 y un F1-Score de 0.8244, destacando la importancia de utilizar un enfoque integral que incorpore una amplia gama de características para mejorar la precisión predictiva.

Estos hallazgos no solo contribuyen a la literatura académica sobre la gestión de reservas hoteleras, sino que también ofrecen insights prácticos que pueden ser implementados por los administradores de hoteles para optimizar sus estrategias de gestión de reservas y reducir las tasas de cancelación.

Anexo

Datos Fuente

Datos fuente utilizados en este estudio se pueden encontrar disponibles públicamente para su descarga y revisión. Los datos corresponden al conjunto de datos de reservaciones de hotel y están disponibles en el siguiente enlace: Hotel Booking Demand on Kaggle.

Código Fuente

El código fuente utilizado para los análisis y modelados presentados en este trabajo está disponible en Google Colab y puede ser accesado a través del siguiente enlace: Google Colab Notebook.