Segunda entrega de proyecto

Cristian Camilo Serna Betancur[1] Diego Alonso Herrera Ramírez[2] Sharid Samantha Madrid Ospina[3]

I. Introducción

ESTE informe se centra en la exploración y análisis de un conjunto de datos con el propósito de abordar un desafío crítico en la industria hotelera: la predicción de cancelaciones de reservas. Además, se abordará una dificultad significativa que surgió durante el proceso de selección de la base de datos inicial, junto con los argumentos que respaldan esta elección y la solución encontrada, lo que llevó a la adopción de un nuevo conjunto de datos. Se presentará una descripción detallada de esta nueva base de datos y se expondrán las decisiones tomadas durante el proceso de procesamiento de datos, incluyendo la eliminación de filas y columnas para garantizar resultados más precisos. Además, se introducirá el modelo que se implementará y que servirá como núcleo de este análisis.

II. Problemas Detectados en el Proceso de Manipulación de Datos

Durante el proceso de manipulación del conjunto de datos, específicamente al dividir los datos en conjuntos de entrenamiento y prueba, se identificaron dos problemas significativos que impactan negativamente la calidad de los datos y la integridad de cualquier análisis o modelo que se pueda desarrollar.

A. Problema 1: Desequilibrio en las Clases de las Variables Objetivo

Uno de los problemas centrales detectados se relaciona con el desequilibrio notable en las clases de las variables objetivo. Este desequilibrio afecta a varias variables, lo que significa que la distribución de categorías en estas variables es altamente irregular. Las variables afectadas incluyen:

P6210 Nivel educativo más alto alcanzado
P7454 Ha buscado trabajo alguna vez
P7456 ¿Cuánto hace que buscó trabajo por última vez?
P7472 Recibió o ganó el mes pasado ingresos por concepto de trabajo?

Este desequilibrio de clases representa un desafío significativo en la construcción de modelos precisos y en la interpretación de resultados confiables. En situaciones de desequilibrio, los modelos pueden sesgar hacia la clase mayoritaria, lo que resulta en predicciones incorrectas y una falta de sensibilidad para detectar eventos o resultados en la clase minoritaria.

B. Problema 2: Falta de una Variable de Salida Claramente Definida

Otro desafío importante que se identificó es que la base de datos carece de una variable de salida claramente definida que se pueda utilizar para clasificar si un hogar es vulnerable o no. Como resultado, no hay una etiqueta de clase preexistente que indique la situación de pobreza o vulnerabilidad económica de los hogares.

Para abordar esta carencia, se consideró utilizar las variables anteriormente mencionadas (P6210, P7454, P7456, P7472) como variables objetivo potenciales para construir un modelo que clasifique los hogares en función de su situación económica. Sin embargo, se encontró que estas variables tenían un gran porcentaje de datos faltantes, lo que compromete la viabilidad de utilizarlas como variables de salida en un modelo predictivo.

III. Solución propuesta

Debido a las dificultades mencionadas, hemos tomado la decisión de cambiar al ”Hotel Booking Dataset”. A pesar de que este conjunto de datos es de naturaleza diferente y no se relaciona directamente con la pobreza, ofrece datos completos y limpios que son fundamentales para llevar a cabo un análisis preciso y confiable.

Este cambio se realizó para garantizar que los resultados y conclusiones de nuestro proyecto sean precisos y útiles para los fines previstos. Las acciones que un hotel puede considerar para gestionar cancelaciones, basadas en la probabilidad de que una reserva sea cancelada, se han discutido en detalle y resaltan la importancia de utilizar datos confiables en cualquier estrategia de toma de decisiones.

IV. Problema predictivo

El problema predictivo a resolver se relaciona con la capacidad de predecir la probabilidad de que una reserva de hotel sea cancelada. Este problema es fundamental en la industria hotelera y permite a los hoteles tomar medidas proactivas para gestionar sus reservas de manera más eficiente. La cancelación de reservas puede llevar a una pérdida de ingresos y, por lo tanto, es esencial identificar con anticipación aquellas reservas con una alta probabilidad de ser canceladas.

A. Enlace del dataset

Hotel Booking Dataset

B. Número de filas

$119,390$

C. Número de columnas

32

D. Porcentaje de columnas categóricas

37.5%

E. Porcentaje de Datos Faltantes

1) agent: 13.69%
2) company: 94.31%
3) required_car_parking_spaces: 5%

F. Columnas relevantes seleccionadas

1) hotel: El tipo de hotel, ya sea "Hotel de ciudad" o "Hotel Resort".
2) is_canceled: Valor binario que indica si la reserva fue cancelada (1) o no (0).
3) lead_time: Número de días entre la reserva y la llegada.
4) arrival_date_year: Año de fecha de llegada.
5) arrival_date_month: Mes de la fecha de llegada.
6) arrival_date_week_number: Número de semana de la fecha de llegada.
7) arrival_date_day_of_month: Día del mes de fecha de llegada.
8) stays_in_weekend_nights: Número de noches de fin de semana (sábado o domingo) que se hospeda el huésped.
9) stays_in_week_nights: Número de noches entre semana (de lunes a viernes) que se hospeda el huésped.
10) adults: Número de adultos.
11) children: Número de hijos.
12) bebies: Número de bebés.
13) meal: Tipo de comida reservada.
14) country: País de origen.
15) market_segment: Designación del segmento de mercado.
16) distribution_channel: Canal de distribución de reservas.
17) is_repeated_guest: Valor binario que indica si el invitado es un invitado repetido (1) o no (0).
18) previous_cancellations: Número de cancelaciones de reservas anteriores.
19) previous_bookings_not_canceled: Número de reservas anteriores no canceladas.
20) reserve_room_type: Código del tipo de habitación reservada.
21) assigned_room_type: Código del tipo de habitación asignado en el check-in.
22) booking_changes: Número de cambios/modificaciones realizadas en la reserva.
23) deposit_type: Tipo de depósito realizado.
24) agent: DNI de la agencia de viajes.
25) company: ID de la empresa.
26) days_in_waiting_list: Número de días en lista de espera antes de reservar.
27) customer_type: Tipo de reserva.
28) adr: Tarifa promedio diaria.
29) require_car_parking_spaces: Número de plazas de aparcamiento necesarias.
30) total_of_special_requests: Número de solicitudes especiales realizadas.
31) reserve_status: último estado de la reserva.
32) reserve_status_date : Fecha del último estado.
33) name: nombre del huésped. (Irreal)
34) email: dirección de correo electrónico del huésped. (No es real)
35) phone-number: número de teléfono del huésped. (Irreal)
36) credit_card: datos de la tarjeta de crédito del huésped. (Irreal)

G. Columnas Eliminadas

Durante la exploración del dataset, se efectuaron decisiones cruciales de eliminación de columnas y filas para optimizar la calidad y pertinencia de nuestros datos. Estas decisiones se tomaron tras consideraciones específicas:

Se excluyeron las columnas que contenían información personal de los clientes, a saber, 'name', 'email', 'phone-number' y 'credit_card'. Esta acción no solo protege la privacidad de los clientes, sino que también elimina datos irrelevantes para nuestra tarea de predicción. Dado que aproximadamente el 94.31% de los datos en la columna 'company' estaban ausentes (valores nulos), esta columna se eliminó. La alta tasa de datos faltantes hacía que la columna resultara poco confiable para nuestro modelo, y su eliminación no compromete la integridad de los datos.

H. Eliminación de Valores Nulos

'Agent': Se optó por eliminar las filas que contienen valores nulos en la columna. A pesar de su relevancia para nuestro análisis, los datos faltantes en esta columna podrían haber introducido complejidades innecesarias en la fase de preprocesamiento. La eliminación de estas filas aseguró la consistencia y calidad de los datos restantes. 'required_car_parking_spaces': Dado que esta columna contiene alrededor del 5% de datos faltantes, se optó por eliminar las filas que presentaban valores nulos en esta columna. Esta medida busca asegurar que los datos restantes sean coherentes y estén listos para su procesamiento y análisis.

I. Métricas de desempeño

Como métrica de machine learning, utilizaremos principalmente la Función de Costo Logística o Log Loss para evaluar el rendimiento de nuestro modelo.

J. Función de Costo Logística o Log Loss

Esta métrica permite evaluar con precisión la eficiencia de nuestros modelos de clasificación en la predicción de cancelaciones de reservas de hotel. El objetivo principal es minimizar el valor de Log Loss, dado que un valor más bajo indica un mejor rendimiento. En un entorno en el que la anticipación de cancelaciones puede influir significativamente en la gestión de reservas y la optimización de los ingresos, esta métrica se convierte en un elemento crítico. Log Loss proporciona una medida cuantitativa que cuantifica la calidad de nuestras predicciones, permitiéndonos tomar decisiones informadas basadas en el rendimiento de los modelos.

Además, exploramos el uso de modelos de clasificación que emplean funciones de densidad gaussianas con el criterio de máxima verosimilitud. Estos modelos nos permitirán abordar

eficazmente la tarea de predicción de cancelaciones, mejorando así nuestra capacidad de gestionar reservas y optimizar los ingresos de manera proactiva.

K. Métricas de negocio

El proyecto se considerará El proyecto se considerará exitoso en términos de métricas de negocio si el modelo proporciona información relevante y útil para los hoteles en la gestión de sus reservas. Así pues, buscamos minimizar la pérdida de ingresos por cancelaciones; Cuanto más efectivo sea el modelo en la predicción de cancelaciones, menor será la pérdida de ingresos por reservas no cumplidas. El desempeño deseable en términos de negocio implica una reducción significativa de esta pérdida.

L. Criterio de Desempeño Deseable en Producción

El desempeño deseable en producción se medirá por la capacidad del modelo para predecir con alta precisión las cancelaciones de reservas. El objetivo es lograr un modelo que minimice Log Loss y, al mismo tiempo, reduzca la pérdida de ingresos y optimice la ocupación de habitaciones. Un desempeño deseable implicaría un Log Loss bajo y una pérdida de ingresos significativamente reducida. En términos cuantitativos, esto podría traducirse en Log Loss cercano a cero y una disminución considerable en las cancelaciones no anticipadas.