ENTREGA FINAL DEL PROYECTO

Cristian Camilo Serna Betancur
Diego Alonso Herrera Ramírez
Sharid Samantha Madrid Ospina

Resumen__ Este informe se enfoca en abordar el desafío clave en la industria hotelera: predecir cancelaciones de reservas. Se detalla el análisis exploratorio y las decisiones de procesamiento de datos para garantizar la precisión del modelo. Se presentan dos modelos: uno basado en densidad gaussiana y otro en Máquinas de Vectores de Soporte (SVM). Se describen iteraciones, ajustes de parámetros, problemas de sobreajuste y la implementación final del SVM, que demostró alta precisión.

Palabras Clave: Predicción de cancelaciones de reservas, Análisis exploratorio de datos, Preprocesamiento de datos, Modelos de clasificación: Densidad gaussiana y SVM

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. 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. Cantidad de datos por cada clase

- Case 0: 75166
- Clase 1:44224

E. Porcentaje de columnas categóricas

37.5%

F. Porcentaje de Datos Faltantes

- **agent**: 13.69%
- **Company**: 94.31%

G. Simulación de datos faltantes

- **required_car_parking_spaces**:5.00%

H. Descripciones de las Columnas:

- **hotel:** El tipo de hotel, ya sea "Hotel de ciudad" o "Hotel Resort".
- **is_canceled:** Valor binario que indica si la reserva fue cancelada (1) o no (0).
- **lead_time**: Número de días entre la reserva y la llegada.
- **Arrival_date_year**: Año de fecha de llegada.
- **Arrival_date_month**: Mes de la fecha de llegada.
- **Arrival_date_week_number**: Número de semana de la fecha de llegada.
- **Arrival_date_day_of_month:** Día del mes de fecha de llegada.
- **Stays_in_weekend_nights:** Número de noches de fin de semana (sábado o domingo) que se hospeda el huésped.
- **Stays_in_week_nights:** Número de noches entre semana (de lunes a viernes) que se hospeda el huésped.
- **adults**: Número de adultos.
- **children**: Número de hijos.
- **bebies**: Número de bebés.
- **meal:** Tipo de comida reservada.
- **country:** País de origen.
- **market_segment:** Designación del segmento de mercado.
- **Distribution_channel:** Canal de distribución de reservas.
- **is_repeated_guest:** Valor binario que indica si el invitado es un invitado repetido (1) o no (0).
- **previous_cancellations:** Número de cancelaciones de reservas anteriores.
- **previous_bookings_not_canceled:** Número de reservas anteriores no canceladas.
- **reserve_room_type:** Código del tipo de habitación reservada.
- **assigned_room_type:** Código del tipo de habitación asignado en el check-in.
- **booking_changes:** Número de cambios/modificaciones realizadas en la reserva.
- **deposit_type:** Tipo de depósito realizado.
- **agent:** DNI de la agencia de viajes.
- **company:** ID de la empresa.

- **days_in_waiting_list**: Número de días en lista de espera antes de reservar.
- **customer_type**: Tipo de reserva.
- **adr**: Tarifa promedio diaria.
- **require_car_parking_spaces** : Número de plazas de aparcamiento necesarias.
- **total_of_special_requests** : Número de solicitudes especiales realizadas.
- **reserve_status** : último estado de la reserva.
- **reserve_status_date** : Fecha del último estado.
- **name** : nombre del huésped. (Irreal)
- **email** : dirección de correo electrónico del huésped. (No es real)
- **phone-number** : número de teléfono del huésped. (Irreal)
- **credit_card**: datos de la tarjeta de crédito del huésped. (Irreal)

I. 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:

1. Eliminación de Datos Personales

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.

2. Columna 'Company'

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.

3. Columna 'required_car_parking_spaces'

Debido a que la columna fue clasificada como una columna constante, se optó por su completa eliminación.

4. Eliminación de Valores Nulos

- **'agent', 'children':** Se optó por eliminar las filas que contienen valores nulos en las columnas. 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.

5. Balanceo de muestras

Al realizar nuestras primeras iteraciones notamos que la cantidad de datos por clase posee un desfase mayor a 20000 muestra entre ellas.

- Case 0: 75166
- Clase 1:44224

Así pues, realizó un submuestreo de los datos. Partimos desde la base en la que ya tenemos la contabilidad de datos por cada clase, luego obtenemos la cantidad mínima de muestras entre las clases, realizamos el submuestreo en ambas clases para igualar la cantidad mínima, combinamos lo datos del sub-muestreados y finalmente mezclamos los datos para que no estén ordenamos por clase.

Así pues, obtenemos que la cantidad de datos por cada clase es:

- Case 0: 44224
- Clase 1: 44224

III. EVALUACIÓN DE MODELOS DE CLASIFICACIÓN

A. Modelo con Función de Densidad Gaussiana

Se entrenan dos modelos con la función de densidad Gaussiana, uno por cada clase, con el 80% de muestras del conjunto de datos. El 20% restante se usa para la validación y el cálculo de métricas. Para decidir entre una clase y la otra, se toma la probabilidad de la clase más alta con respecto a la otra.

1. Métricas

La eficiencia del modelo con datos de validación (datos no vistos en el entrenamiento) fue del 99.34%. Con una matriz de confusión:

```
[[7367   99]
 [   2 7943]]
```

El modelo ha hecho un buen trabajo clasificando los casos negativos (7367 verdaderos negativos) y positivos (7943 verdaderos positivos). Ha cometido errores al clasificar 99 casos negativos como positivos (falsos positivos) y 2 casos positivos como negativos (falsos negativos). En resumen, la mayoría de las predicciones son correctas.

Con F1-score igual a 0.9936823669231251. Este valor sugiere que el modelo está logrando un equilibrio efectivo entre la precisión y la exhaustividad.

B. Máquinas de Vectores de Soporte (Support Vector Machines - S. V. M.)

El modelo S. V. M. Utiliza vectores de soporte para clasificación. Se realizaron ajustes mediante búsqueda de cuadrícula.

1. Iteraciones

C	Gamma	Kernel	F1-score	Precisión
0.1	0.001	Poly	1.0	**1.0**

Fig 1. Tabla de resultado de la primera iteración del modelo SVM

Al iniciar las iteraciones nos dimos cuenta que la precisión y el F1-score, nos daban igual a 1.0 y la matriz de confusión es [[7417 0][0 07994]], esto nos indica que el modelo está sobre ajustado.

Para solucionar este problema realizamos:

- **Estandarización de los datos**

utilizando la clase 'StandardScaler' de Scikit-learn; con el objetivo de normalizar los datos.

- **Validación cruzada**

Se utiliza K-folds con K=5 para evaluar el rendimiento del modelo utilizamos la función 'cross_val_score' de Scikit-learn.

- **Evaluación del rendimiento mediante validación cruzada:**
  - El parámetro scoring='accuracy' indica que se evalúa la precisión del modelo.

- cross_val_score calculará la precisión para cada pliegue y devolverá una lista de las precisiones obtenidas en cada uno de los pliegues.

Sin embargo, aunque realizamos varias iteraciones con respecto a los parámetros de SVM, por temas de costos computacionales no fue posible obtener los resultados a dichas iteraciones

C	Gamma	Kernel	Tiempo de ejecución
0.1	0.01	Poly	5721.25
0.1	0.1	Poly	3120.524
1	0.01	Poly	3254.002

Fig 2. Tabla de resultado iteración fallidas del modelo SVM

Una vez cambiamos los parámetros por C = 1, gamma = 0.001, kernel = 'linear' obtenemos que:

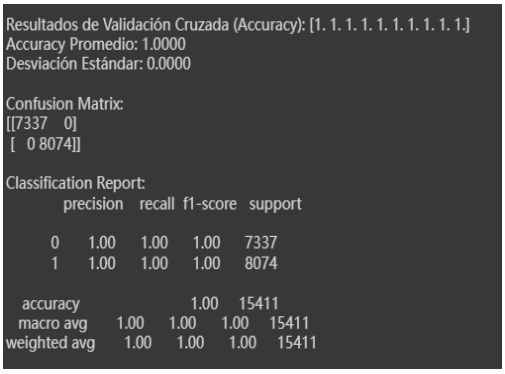

Fig 3. Resultados de validación cruzada, matriz de confusión y demás métricas para los mejores parámetros del modelo SVM

Finalmente, cuando validamos con datos que el modelo no conoce, obtenemos que:

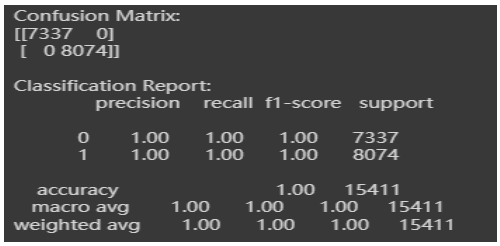

Fig 3. Resultados de validación cruzada, matriz de confusión y demás métricas para datos de test que no fueron vasitos previamente por el modelo SVM

Por lo tanto, podemos concluir que el modelo es muy preciso a la hora de clasificar si una reserva será o no cancelada.

IV. RETOS Y CONSIDERACIONES DE DESPLIEGUE

Como reto inicial se tiene la estandarización de las columnas del conjunto de datos, empezando por la división de algunas (como la columna que representan fechas) y la conversión de columnas alfanuméricas a numéricas. Continúa con la incertidumbre de saber si los modelos fueron sobre ajustados en su etapa de entrenamiento debido a los buenos resultados producidos.

Los modelos arrojaron resultados dignos de puesta en producción. Sin embargo, el modelo de función de densidad gaussiana, al haber sido codificado manualmente, no se tiene claro cómo puede ser exportado (h5, xml, yaml), para ser puesto en marcha en un servidor de producción, ya que no se tiene una función de utilidad para dicho escenario.

V. CONCLUSIONES

En el transcurso de este proyecto, se ha abordado de manera integral el desafío crítico en la industria hotelera relacionado con la predicción de cancelaciones de reservas. La capacidad para anticipar y gestionar eficientemente estas cancelaciones es esencial para optimizar la ocupación de habitaciones y mitigar las pérdidas de ingresos asociadas. A través de un enfoque estructurado y metodológico, se ha logrado avanzar significativamente en la comprensión y resolución de este problema.

El conjunto de datos "Hotel Booking Dataset" proporcionó la base fundamental para el desarrollo de modelos predictivos. El proceso de preprocesamiento de datos fue esencial para garantizar la calidad de los datos, involucrando decisiones cruciales como la eliminación de columnas con información personal y la gestión de datos faltantes mediante submuestreo para equilibrar las clases. Los resultados obtenidos son dignos de producción.

Los modelos presentaron una eficacia que los posiciona como herramientas valiosas para la toma de decisiones en la gestión de reservas hoteleras. En resumen, este proyecto ha logrado con éxito abordar el desafío de predicción de cancelaciones de reservas, proporcionando soluciones sólidas y sentando las bases para futuras mejoras y aplicaciones en el ámbito hotelero.