Memoria

Predicción Cancelaciones de Hotel

Laura Ledo
8 de julio de 2022

Las cancelaciones de habitaciones tienen un impacto muy negativo en los hoteles: si la habitación cancelada no se consigue ocupar -algo muy posible si son cancelaciones con poca antelación- el establecimiento pierde ese ingreso, pero mantiene los costes asociados a dicha habitación. A su vez, esa habitación reservada, y posteriormente cancelada, tiene unos costes añadidos como el de operación o distribución (gasto en marketing, por ejemplo).

Desde la aparición y expansión de las OTAs (‘*agencias de viaje’* online) en el sector, la posibilidad de realizar reservas online con cancelación gratuita aumentó considerablemente, provocando un aumento de las cancelaciones en los hoteles debido a las reservas ‘*por si acaso’* que al final no se ejecutan.

Objetivo

El objetivo de este proyecto es crear un modelo predictivo que, dada una reserva, consiga predecir con alta fiabilidad si esa reserva va a ser cancelada o no. Para ello, crearemos un modelo de clasificación binario que clasifique cada reserva introducida en el modelo como:

1: La reserva va a ser cancelada.

0: La reserva no va a ser cancelada.

Análisis

Para poder crear dicho modelo, primero tenemos que entender nuestro dataset. Comenzamos analizando cada variable (notebook 1_resumen_variables) y su relación con el target (notebook 2_EDA), para poder preparar los datos a introducir en el modelo.

El dataset original, descargado de Kaggle, contiene las reservas de un hotel resort, situado en la zona del Algarve, y un hotel de ciudad, situado en la ciudad de Lisboa. Para este análisis, nos centraremos en las reservas del hotel de ciudad.

Las variables proporcionadas sobre las reservas son:

Variable	Tipo	Descripción
is_canceled	TARGET	1: Reserva Cancelada / 0: Reserva No Cancelada
lead_time	Numérica	N días entre que se hizo la reserva y la fecha de llegada
stay_in_week_nights	Numérica	N de noches entre semana (L-V) de la reserva
stays_in_weekend_	Numérica	N de noches de fin de semana (S-D) de la reserva
adult	Numérica	N adultos por reserva
children	Numérica	N niños por reserva
babies	Numérica	N bebés por reserva
previous cancellations	Numérica	N de cancelaciones, anteriores a la reserva, realizadas por el huésped
Previous bookings not	Numérica	N de reservas, anteriores a la reserva de la observación, no canceladas
booking changes	Numérica	N cambios realizados en la reserva
days_in_waiting_list	Numérica	N de días que la reserva estuvo en lista de espera
adr	Numérica	Tarifa Media Diaria de la hab
required_car_parking_s	Numérica	N de plazas de parking reservadas
total of special requests	Numérica	N de peticiones especiales
company	Categórica	ID de la agencia o entidad que realizó la reserva
agent	Categórica	ID del agente o agencia
country	Categórica	País de procedencia del huésped

market segment	Categórica	Modo de reserva
distribution channel	Categórica	Canal de distribución
reserved room type	Categórica	Tipo de habitación reservada
assigned room type	Categórica	Tipo de habitación asignada
reservation status	Categórica	Último estado de la reserva
meal	Categórica	Paquete de comidas reservado
customer type	Categórica	Tipo de huésped
deposit type	Categórica	Indica si el huésped dejo un depósito para garantizar la reserva
arrival date year	Fecha	Año de la fecha de llegada: 2015, 2016 o 2017
arrival date month	Fecha	Mes de la fecha de llegada: de enero a diciembre
arrival date week	Fecha	Semana del año de la fecha de llegada
arrival date day of	Fecha	Día del mes de la fecha de llegada: 1 - 31
reservation status date	Fecha	Fecha de la última modificación hecha en la reserva

**Las variables categóricas con 4 o menos categorías (la mayoría) las transformamos a numéricas con dummies, y las de más de 4, con WOE.*

Para realizar el análisis previo a la creación del modelo, creamos estas dos variables -que al final decidimos no incluir en el modelo por correlación:

- **total n noches**: N noches de duración de la estancia
- **días modificación llegada**: días desde la fecha de la última modificación hecha en la reserva, hasta la fecha de llegada.

Tampoco incluimos:

- La variable **company**, por tener más de un 90% de outliers.
- La variable **reservation status**, ya que es equivalente al target.
- La variable **reserved room type**, porque tiene mucha correlación lineal con assigned room type.
- La variable **agent**: tras un primer análisis nos dimos cuenta que +90% de todas las cancelaciones estaban clasificadas en el agente **9**. Como esto contaminaría el modelo, y no tenemos información de qué es el agente 9, eliminamos la columna.
- La variable **deposit_type**: tras un primer análisis nos dimos cuenta que todas las reservas con depósito 'Non Refund' son cancelaciones. Esto parece un error, pero como no tenemos más información, también eliminamos esta columna para evitar contaminación.

De las variables de tipo fecha, como están todas muy correlacionadas entre sí, seleccionamos para el análisis **arrival date month** y **arrival date day of month**, eliminando el resto.

Durante el análisis previo, también observamos algunas reservas con erratas que decidimos eliminar:

- Reservas que no tienen ni noches entre semana, ni noches en fin de semana.
- Reservas sin huésped (ni adultos, ni niños ni bebés).
- Reservas con bebés pero sin adultos y reservas con más de 8 bebés por habitación.
- Reservas que tienen la misma fecha de llegada y última modificación, pero como estado 'Check-Out'.

Finalmente, sustituimos los missings:

- **children**: sustituimos por 0 entendiendo que no había niños en la reserva.
- **agent**: sustituimos por 0, entendiendo que ningún agente realizó la reserva.
- c**ountry**: sustituimos por la moda.

Al tratarse de un problema de clasificación binaria, decidimos aplicar estos algoritmos al dataset para conseguir el modelo de predicción:

- Random Forest
- Support Vector Machine
- Bagging Random Forest
- Gradient Boosting Classifier
- XGBoost
- MultiLayer Classifier (DL)

Métricas

Precisión:

TP / (TP + FP)

De los que ha predicho como 1, cuántos ha acertado. Minimiza los falsos positivos (FP).

Nos interesa predecir bien los 1 (Cancela), ya que penaliza más al hotel decir que no cancela (0) y que al final cancela: la habitación se queda libre. En el caso de que predigamos que cancele, pero al final no cancele, el hotel no se queda con una habitación libre.

F1-Score

2*Precision*Recall / (Precision + Recall)

Combinación de las métricas Precision y Recall. Se utiliza para comparar clasificadores.

Accuracy

(TP + TN) / Total

Calcula el % de acierto teniendo en cuenta todas las clases del algoritmo de clasificación.

AUC

Área bajo la 'Roc Curve'. Cuanto más cercano a 1, mejor será el clasificador.

Conclusiones

De todas las métricas seleccionadas para evaluar los modelos, la que más nos interesa es **Precisión**, ya que nos interesa predecir bien las cancelaciones. Basándonos en esta métrica, seleccionamos el algoritmo SVM, aplicando al dataset un *feature selection* (aunque baja un poco el scoring, al aplicar *feature selection* evitamos overfitting:

Train:

Accuracy 0.8286310686149081
Precision 0.8526031731261341
F1 0.782617979927397
AUC 0.8151286814540766

Test:

Accuracy 0.8279692970643684
Precision 0.8462973325872039
F1 0.7802906526786483
AUC 0.8139214049557777
Las variables seleccionadas fueron:

- lead time: a mayor antelación en la reserva, mayor posibilidad de cancelación
- arrival date month
- arrival date day of month
- country: Portugal es el país de procedencia de los huéspedes que más cancelan (60% de las reservas hechas por huéspedes portugueses se cancelan.
- previous cancellations: si se ha cancelado previamente, mayor probabilidad de volver a cancelar.
- assigned room type
- booking changes y total os special requests: no suelen realizarse cambios o peticiones especiales en las reservas, pero sí se hacen, menor probabilidad de cancelación.
- adr
- market segment: el modo de reserva ‘Groups’ tiene un porcentaje mayor de cancelaciones que de reservas sin cancelar.
- customer type

Herramientas utilizadas

- Google Colab
- Python
- Numpy
- Pandas
- Plotly
- Category Enconders
- Sklearn
- XGBoost
- Keras

Fuentes

Datos: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand