Un modelo CART para la predicción de cancelaciones hoteleras

Trabajo Práctico I - Aprendizaje Automático

Junio 2020

GRUPO N° 6: MORA, SEBASTIÁN[1]; PEREZ, PABLO [2],VÁZQUEZ BROQUÁ, JUAN[3]

1:moradiazsebastian@gmail.com

2: pablofp92@gmail.com

3: juanivazquez@gmail.com

Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

Resumen: El presente trabajo se enmarca dentro del programa de estudios de la materia Aprendizaje Automático de la Especialización de Explotación de Datos y Descubrimiento del Conocimiento de la Universidad de Buenos Aires. El objetivo principal es el análisis de un set de datos de reservas hoteleras y el desarrollo de algoritmos de árboles de decisiones para predecir cancelaciones. Las etapas de preprocesamiento, elección de parámetros, entrenamiento y evaluación de la performance fueron partes de un proceso iterativo en el que se ensayaron diferentes formas de preprocesamiento, métricas de performance y cortes de poda. Dado el criterio de predecir todas las cancelaciones y la consideración de alto costo de un falso negativo, se puso énfasis en la medida de performance de $f_2 - score$. Se estudió la variación de esta métrica tanto en particiones aleatorias como con validación cruzada de 50 *cross-folds*, obteniendo mejores resultados aunque más dispersos para el último caso. Se evaluaron distintos valores de poda, se observó que valores mínimos de poda (cercanos a 0.000085) maximizaban el $f_2 - score$. a costa de una profundidad desproporcionada que conllevaba *overfitting*. El modelo final opta por un valor menor de poda, reportando así un $f_2 - score$. de 0.82 para el set de prueba, la matriz de confusión (ver anexo) muestra que se trata de un modelo con pocos falsos negativos pero con gran ocurrencia de falsos positivos, siendo inadecuado para un uso final en el mercado. Las variables de días de estadía, tipo de depósito y costo promedio diario (adr) fueron los selectores más importantes del árbol, no obstante, el entrenamiento del árbol utilizando únicamente estos atributos dio pobres resultados.

1. Introducción

La cancelación de reservas tiene un alto impacto en la industria de la hospitalidad porque limita la capacidad de predecir la demanda de habitaciones y las decisiones de gestión de ingresos (RM) asociadas. Para sortear este inconveniente los hoteles suelen apelar a políticas de cancelación con castigos económicos, o a sobrevender habitaciones (Chen, Schwartz y Vargas 2011; Chen y Xie 2013; Talluri y Van Ryzin 2006) esperando compensar su efecto o inhibir el comportamiento. Esto, sin embargo, suele perjudicar tanto la reputación del hotel como su rentabilidad (Mehrotra y Ruttley 2006; Smith *et al*. 2015); más aún, en los últimos años se observa un alza en la proporción de cancelaciones (Chen y Xie 2013) por cambios en el comportamiento de los consumidores. A pesar de que Morales y Wang (2010) afirmaron que es imposible predecir con precisión las cancelaciones [1], Antonio, Almeida y Nunes (2017) presentó un modelo que demuestra que es posible hacerlo obteniendo muy buenas métricas de evaluación. A continuación se construye un modelo predictivo CART. En las sección 2 se presenta el dataset y sus principales características. En las secciones 3.1 y 3.2 se desarrolla una breve exploración de los datos y se detalla el trabajo de preprocesamiento y preparación de la información para el modelado. En 3.3 se presenta el modelo CART y la métrica de performance elegida junto con los valores obtenidos. En 4 se comentan los resultados obtenidos y 5 recoge las conclusiones del trabajo.

2. Datos utilizados

El dataset *booking.csv* corresponde una la información provista por una cadena hotelera en Portugal para 2 de sus hoteles: uno en la ciudad y un resort. La información proviene de la consolidación de sistemas de información gerencial y la integridad de los datos es buena en líneas generales. El set de datos contiene 32 variables y 119.390 entradas. Laa mayoría de los alojados son locales de Portugal aunque hay registro de reservas 128 países. Algunas variables de tipo booleanas son consideradas numéricas entre 0 y 1, como es el caso de la cancelación de la reserva o si la persona ya visitó el hotel previamente.Respecto de la presencia de valores atípicos, se encontró un registro en la variable adr fuera de escala con el registro tarifario. Otras variables que registraron outliers están asociadas al número de personas por habitación. Esto es así porque operativamente algunas reservas asignan todas las personas a una misma reserva para ser corregidas en múltiples habitaciones al momento del *check-in*. Se trata de un set de datos no balanceado en cuanto a la proporción de cancelados (37 %) y no cancelados (62,7 %). Si bien la mayoría de los atributos están completos, los campos de compañía y representante (*agent*) están principalmente vacíos (NULL). Esto es así en los casos en que las reservas fueron realizadas por particulares.
En el anexo se informa una tabla que resume los principales comentarios de los atributos.

3. Metodología

3.1. Análisis exploratorio

El análisis exploratorio se basó en la realización de gráficas de dispersión entre las diferentes variables numéricas y el porcentaje de reservas canceladas (Figura 4). Se observó que mayores porcentajes de cancelación correlacionan positivamente con los días de reserva hasta la llegada (*lead_time*), las noches reservadas y el número de cancelaciones previas, mientras que con la cantidad de pedidos especiales que realiza el cliente se observó una correlación negativa. Los coeficientes de correlación junto al

[1] "it is hard to imagine that one can predict whether a booking will be canceled or not with high accuracy simply by looking at PNR information"

information gain asociado a cada variable dan un indicio *a priori* de qué criterios son más importantes para la separación de clases.

3.2. Preprocesamiento

El preprocesamiento del dataset para alimentar un modelo de árboles tiene menos requisitos que otras técnicas. Respecto a los datos, las ventajas de los árboles de decisión son múltiples: (1) pueden reconocer y descartar variables irrelevante, (2) no requieren mayores transformaciones ni escalado de variables (las transformaciones monotónicas no afectan los cortes -splits- porque el orden de la información se preserva,(3) tiene una buena tolerancia a los datos faltantes (en el caso de CART se realizan cortes sustitutos) (Seni y Elder 2010). Se siguieron los siguientes criterios generales:

1. Variables Numéricas: como los árboles no son afectados por transformaciones monotónicas y escalado, no se hicieron cambios a las variables numéricas a pesar de observar una fuerte asimetría . *(aunque durante el análisis exploratorio se evaluó la correlación entre sus valores transformados para detectar la existencia de correlaciones no lineales). Un paso en favor de la reducción de la dimensionalidad fue realizado al combinar las variables stays_in_weekend_nights' y 'stays_in_week_nights'.

2. Variables Categóricas: como el método de trabajo elegido generaría una variable nueva para cada alternativa en cada variable categórica se buscó reducir la cantidad de alternativas en cada una. El principal criterio de agrupación fue la frecuencia relativa de cada caso, para después concentrarlas en grupos de cancelación alta y baja. Para en un paso posterior transformarlas en variables binarias mediante el método de *OneHotEncoding*. Por ejemplo, en el caso de la variable países primero se agruparon los países de baja frecuencia y después construyó una variable binaria para cada caso, obteniendo un total de 60 variables finales.

3. Variables Ordinales:

   a) No Temporales: el único caso con propiedad ordinal es el asociado a la habitación reservada. Por cumplir con la anonimidad de los datos, las categorías no están identificadas. Una alternativa para reconstruir el orden de precio de las habitaciones hubiera sido ordenarlas en función de la tarifa promedio diaria (adr) del canal directo, asumiendo una relación lineal entre la categoría y el precio de la habitación.

   b) Temporales las únicas variables ordinales disponibles están asociadas a eventos temporales (fecha de reserva, año, mes, etc.). Se optó por mantener la variable semana como punto intermedio entre colapsar toda la información en un año y desagregarla en información diaria. De esta manera se mantuvo información temporal y con estacionalidad.

En base a esas consideraciones se realizó una limpieza de datos y la imputación de valores vacíos. El campo de niños (*children*) posee 4 valores faltantes y se decidió realizar una imputación por la media, mientras que para país, las entradas con 488 valores ausentes fueron eliminadas debido a la dificultad de hallar un método de imputación adecuado. Para la variable adr (*Average Daily Rate*) se eliminaron los valores inconsistentes.

El filtrado de información futura o *data leakage* fue un problema a considerar al momento de desarrollar un modelo predictivo. Las variables reservation_status, reservation_status_date, assigned_room_type incluyen información sobre si el cliente concretó la reserva y por lo tanto fueron eliminadas.

3.3. Selección del algoritmo

El modelo se construyó usando diversas herramientas del paquete *sklearn* (Pedregosa *et al*. 2011). El algoritmo CART de Scikit-learn es un árbol de decisión de clasificación y regresión[2]. CART construye árboles binarios usando el atributo y definiendo el umbral que permite maximizar la ganancia de información en cada nodo.

Un 20 % de los datos fueron separados como conjunto de prueba utilizando de forma aleatoria (50 semillas) y estratificada para compensar el desbalance de clases . El restante 80 % constituyó el set de desarrollo.
Teniendo en cuenta la premisa que el árbol de decisión deba predecir todas las cancelaciones, se consideró de mayor importancia evitar los falsos negativos (Cuadro 1). Si bien *Recall* es la mejor métrica de performance cuando es necesario identificar todos

[2]Por las siglas en inglés para Classification and Regression Trees

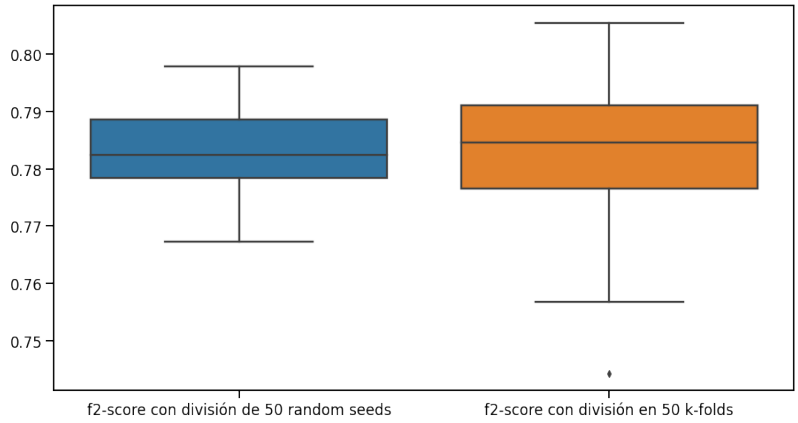


Figura 1. Variación de $f_2 score$ según particionado aleatorio y 50 k-folds

los positivos (*is_canceled* = 1), evaluar el algoritmo únicamente por su *recall* introduce un gran riesgo de obtener demasiados falsos positivos. La métrica $f_1 - score$ evalúa tanto *recall* como precisión, por lo que se considera una mejor medida para set de datos no balanceados, como es el caso de este trabajo (Müller, Guido *et al*. 2016). Se consideró, en cambio, la métrica $f_\beta - score$ donde un valor de β mayor a 1 asigna mayor peso al *recall*[3].

$$F_\beta = (1 + \beta^2) \frac{precision * recall}{(\beta^2 * precision) + recall)}$$

	Predice no cancelación (0)	**Predice cancelación (1)**
El cliente no cancela (0)	Verdadero negativo (TN)	Falso positivo (FP)
El cliente cancela (1)	Falso negativo (FN)	Verdadero positivo (TP)

Cuadro 1. Matriz de confusión aplicada a la predicción de cancelaciones de reservas

En la figura 1, se pueden apreciar los distintos valores de performance obtenidos particionando el set de desarrollo de dos maneras distintas, usando la métrica $f_2 - score$. En primera instancia, se separaron los sets de entrenamiento y validación de manera aleatoria con 50 semillas. De manera similar, se realizaron particiones con validación cruzada de 50 *k-folds*. Ambas metodologías de particionado arrojaron resultados similares, con un $f_2 - score$ medio de 0,785.
Se evaluaron los mejores valores posibles de α en función del performance (Figura 2) con validación cruzada de estratificada de 10 *k-folds* y *Grid Search* (evaluación de todos los posibles valores). La performance en la evaluación del set de desarrollo decrece escalonadamente en función de alpha y bruzcamente a partir de un valor de poda de 0,000085. Mientras que para el set de prueba, la métrica se mantiene relativamente constante hasta el mismo valor.

De manera similar pero con una búsqueda aleatoria (*Randomized Search*) se hallaron los mejores hiperparámetros de profundidad y criterio de pérdida de información. La combinación de un parámetro de poda adecuado y el resto de los hiperparámetros resultó un proceso iterativo. Se consideró fijo el parámetro *'class_weight' = {1:1.7}* para asignar un mayor peso a las cancelaciones.

Como última etapa se reentrenó el árbol con los descriptores más importantes utilizando la herramienta selector para la técnica de eliminación recursiva.

[3]En rigor, $f_1 - score$ no es otra cosa que un caso especial de $f_\beta - score$ cuando β es igual a 1.

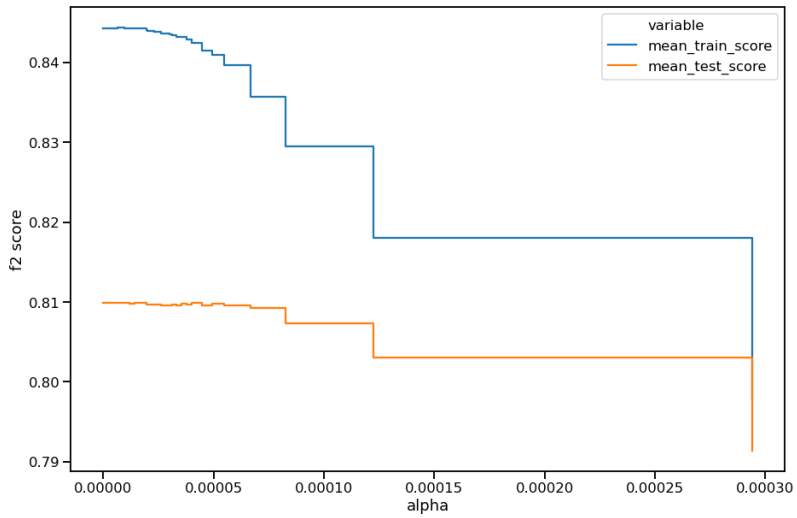


Figura 2. $f_2 score$ en función del alpha

Figura 3. Plot del arbol con mejor performance: Los descriptores más importantes son el tipo de depósito, adr y lead$_t ime$

4. Resultados

El proceso de modelado fue un trabajo iterativo, explorando con diferentes variables, revisando los hiperparámetros obtenidos en procesos de validaciones e interpretando los resultados obtenidos. Adicionalmente al flujo de trabajo del modelo final, se hicieron ejercicios de modelado con intención de explorar y tener una mayor comprensión de las cancelaciones. Los casos que valen la pena mencionar son 2: (1) se probó un caso reducido a los 3 parámetros con mayor ganancia de información. La performance arrojó una precisión baja y un recall aún menor. Esto lo interpretamos como una confirmación de que la variable objetivo depende de más de 3 descriptores. (2) En otro ejercicio se mantuvieron solamente las variables numéricas, obteniéndose resultados mejores a los esperados y confirmando la relevancia de estas variables[4].

[4]Durante el trabajo la visualización de scatter plots clasificando la variable objetivo entre distintas variables categóricas (deposit_type, customer_type y hotel) generaron dudas sobre la relevancia de algunas variables numéricas

5. Conclusiones

- Se pudo evidenciar que el rendimiento del modelo es sensible al balance de los datos. Otros factores que pueden afectar el performance es la introducción de variables que no son tan relevantes en el proceso de decisión, y que terminan causando ruido en el entrenamiento del modelo. De esta manera, resalta la importancia de un preprocesamiento de los datos antes de generar un modelo de predicción.[5]
- Los valores de performance (f_2score) son similares en la predicción del set de validación y prueba.
- Se obtuvo un modelo de predicción (Figura 3) adecuado para evitar falsos negativos, es decir que puede predecir correctamente la mayoría de las cancelaciones, no obstante su performance en cuanto a los falsos positivos es pobre. La aplicación de este modelo en la realidad podría llevar a la sobreventa de plazas por sobrestimar la cantidad de cancelaciones.
- Es posible compensar los falsos negativos y falsos positivos mediante la eliminación del parámetro de peso (class_weight) o el aumento de la poda
- Un aspecto que define la importancia de predecir las cancelaciones es evitar tomar medidas que inhiban la toma de reservas. Si bien el modelo obtenido alcanza una performance de interés, el principal nodo del árbol está asociado al cobro no reembolsable de la reserva. Esto quiere decir que la capacidad predictiva depende fuertemente de que se mantenga la política de reservas no reembolsables.

[5]En el caso de del algoritmo CART, la calidad del modelo depende no sólo de la cantidad de variables sino también de los casos registrados por cada una.

Referencias

Mehrotra, Ravi y James Ruttley (2006). *Revenue management*. American Hotel y Lodging Association.

Talluri, Kalyan T y Garrett J Van Ryzin (2006). *The theory and practice of revenue management*. Vol. 68. Springer Science & Business Media.

Morales, Dolores Romero y Jingbo Wang (2010). «Forecasting cancellation rates for services booking revenue management using data mining». En: *European Journal of Operational Research* 202.2, págs. 554-562.

Seni, Giovanni y John F Elder (2010). «Ensemble methods in data mining: improving accuracy through combining predictions». En: *Synthesis lectures on data mining and knowledge discovery* 2.1, págs. 1-126.

Chen, Chih-Chien, Zvi Schwartz y Patrick Vargas (2011). «The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers». En: *International Journal of Hospitality Management* 30.1, págs. 129-135.

Pedregosa, F. *et al*. (2011). «Scikit-learn: Machine Learning in Python». En: *Journal of Machine Learning Research* 12, págs. 2825-2830.

Buitinck, Lars *et al*. (2013). «API design for machine learning software: experiences from the scikit-learn project». En: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, págs. 108-122.

Chen, Chih-Chien y Karen Lijia Xie (2013). «Differentiation of cancellation policies in the US hotel industry». En: *International Journal of Hospitality Management* 34, págs. 66-72.

Smith, Scott J *et al*. (2015). «Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry». En: *Journal of Travel & Tourism Marketing* 32.7, págs. 886-906.

Müller, Andreas C, Sarah Guido *et al*. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc."

Antonio, Nuno, Ana de Almeida y Luis Nunes (2017). «Predicting hotel booking cancellations to decrease uncertainty and increase revenue». En: *Tourism & Management Studies* 13.2, págs. 25-39.

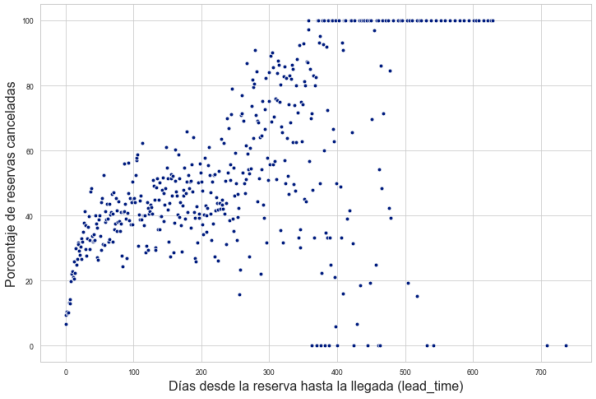
Antonio, Nuno *et al*. (2018). «Hotel online reviews: different languages, different opinions». En: *Information Technology & Tourism* 18.1-4, págs. 157-185.

Antonio, Nuno (2019). «Predictive models of hotel booking cancellation: a semi-automated analysis of the literature». En: *Tourism & Management Studies* 15.1, págs. 7-21.

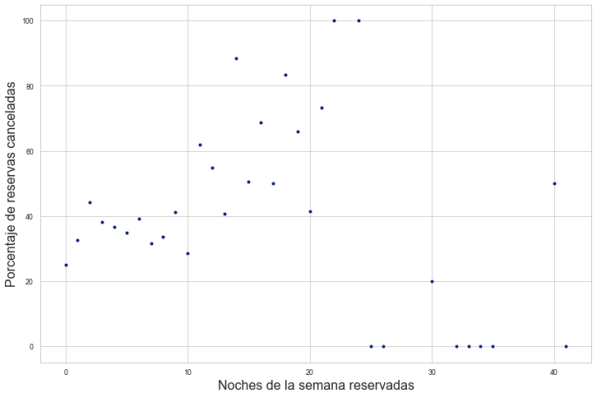
Antonio, Nuno, Ana de Almeida y Luis Nunes (2019). «Hotel booking demand datasets». En: *Data in brief* 22, págs. 41-49.

Anexo

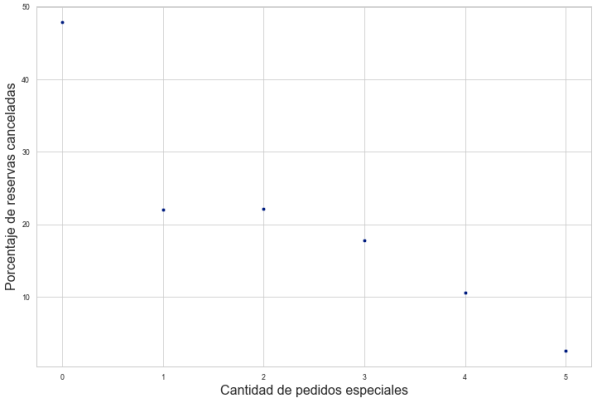
Variable	Tipo	Definición	Categorías/Valores	Comentarios	Cambios realizados
hotel	Categórica	Indica el hotel asociado a la observación	Resort - Hotel		
-					
is_canceled	Categórica	Es la variable objetivo e indica si la reserva fue tomada (check out) o si se canceló (tanto para cancelaciones informadas como cuando el cliente no se presenta (no show))	0-1	Construida a partir de la variable Reservation_status y consolida como cancelaciones tanto cancelaciones explícitas como los casos "no show"de clientes que no cumplimentan la reserva. El valor 0 se asigna a reservas cumplidas (Check Out).	
lead_time	Numérica	Cuenta los días entre la fecha de creación de la reserva y la fecha de check in asociada.	num.	Fuerte asimetría positiva.	
arrival_date_year	Categórica (Fecha)	Año de check in asociado a la reserva.	años	Todas las variables temporales las integramos en una variable binaria: mes de alta cancelación /mes de baja cancelación	Eliminación
arrival_date_month	Categórica (Fecha)	Mes de check-in asociado a la reserva.	meses		Se redujo a día del año
arrival_date_week_number	Categórica (Fecha)	Semana de check-in asociado a la reserva.	fecha		Se redujo a día del año
arrival_date_day_of_month	Categórica	Día del mes de check-in asociado a la reserva.	1-31		Se redujo a día del año
stays_in_weekend_nights	Numérica	Indica si alguna noche de la estadía ocurre durante el fin de semana.	numerica	Fuerte asimetría positiva.	
stays_in_week_nights	Numérica	Indica si alguna noche de la estadía ocurre durante la semana hábil.	numérica	Fuerte asimetría positiva.	
adults	Numérica	Cantidad de adultos indicados en la reserva	numérica	Fuerte asimetría positiva.	
children	Numérica	Cantidad de niños indicados en la reserva	numérica	Fuerte asimetría positiva.	
babies	Numérica	Cantidad de bebes indicados en la reserva	numérica	Fuerte asimetría positiva.	
meal	Categórica	Tipo de servicio solicitado: sólo desayuno, 1 comida, menú completo.	Undefined (sin datos) SC (sin comidas) BB – Bed &Breakfast HB (desayuno y una comida) FB – Full board(3 comidas)	No se interpretan los casos sin definir como valores faltantes.	
country	Categórica	Indica el país de origen de la reserva.	Información de 175 países siguiendo la clasificación alpha-3.	Se detectaron 2 países con errores de nomenclatura: China con dos nomenclaturas y la nomenclatura de Timor del Este está desactualizada.	Agrupamiento por frecuencia relativa en reservas: 60 valores finales.
market_segment	Categórica	Indica el segmento de mercado al que se asignó la reserva.	Direct, Corporate, Online/Offline TA (Agentes de Viajes) TO (Operadoras de Viajes) Groups, Aviation, Complementarity		2 Undefined eliminados
distribution_channel	Categórica	Indica el canal a través del cual se hizo la reserva.	Direct, Corporate, TA/TO, GDS		5 Undefined eliminados
is_repeated_guest	Categórica	Controla si el cliente ya se alojó en otra oportunidad en el hotel.	0-1		
previous_cancellations	Numérica	Controla si el cliente canceló alguna reserva previa.	num.	Fuerte asimetría positiva.	
-					
previous_bookings_not_canceled	Numérica	Previas reservas cumplimentadas.	num.	Fuerte asimetría positiva.	
reserved_room_type	Categórica	Tipo de habitación reservado	Codigo de tipo de habitación		
assigned_room_type	Categórica	Tipo de habitación asignado	data leakage		
booking_changes	Numérica	BL	numerica	Fuerte asimetría positiva.	
deposit_type	Categórica	Tipo de depósito asociado a la reserva	Reembolsable, No Reembolsable Sin Depósito		
agent	Categórica	Indica el código de la agencia en que se hizo la reserva	N° de agente		Los NULL hacen referencia a clientes particulares
company	Categórica	Indica la compañía asociada a la reserva realizada.	N° de Compañía		Se consolida en una variable tricotómica: Compañía con alta proporción de cancelaciones / c. c/ baja ratio de cancelaciones / particulares. Los NULL hacen referencia a clientes particulares
days_in_waiting_list	Numérica	Particulares / Agentes con ratio canc. Alto / Resto de Agentes	num.	Fuerte asimetría positiva.	
customer_type	Categórica	Indica el tipo de cliente:	Contract (reservas con contrato asociado) Group (reserva grupal) Transient (individual) Transient-party (múltiples reservas individuales asociadas entre sí)	.	
adr	Numérica	Tarifa Diaria Promedio (Average Daily Rate)	num.		Se detectaron dos valores incompatibles con la tarifa promedio diaria, se reemplazó por la media
required_car_parking_spaces	Numérica	Indica cuántas cocheras se reservaron.	Los clientes que solicitan parking tienen menor tasa de cancelación	Fuerte asimetría positiva.	
total_of_special_requests	Numérica	Cantidad de solicitudes asociadas a la reserva	num.	Fuerte asimetría positiva.	
reservation_status	Categórica	Registra el estado final de la reserva. Sirve de input para la creación de la variable objetivo.	fechas		Eliminada por data leakage
reservation_status_date	Categórica (Fecha)	Fecha en que se realizó la actualización del estado de la reserva.	fechas		Eliminada por data leakage



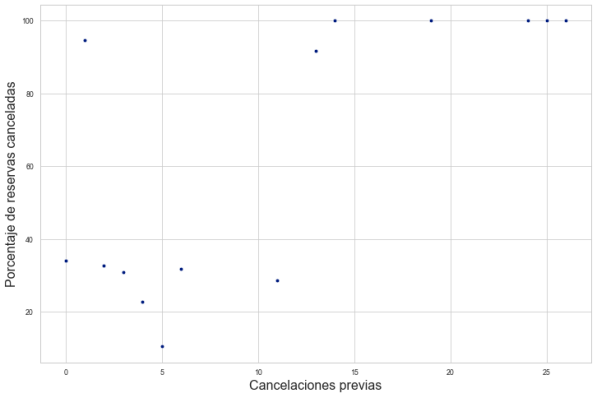
(a) Influencia del tiempo de arribo en la cancelación



(b) Influencia de las noches reservadas en la cancelación



(c) Pedidos especiales en función de proporción de cancelación



(d) Cancelaciones previas del cliente en función de la cancelación de la reserva

Figura 4. Scatter plots de las variables más correlacionadas con la cancelación

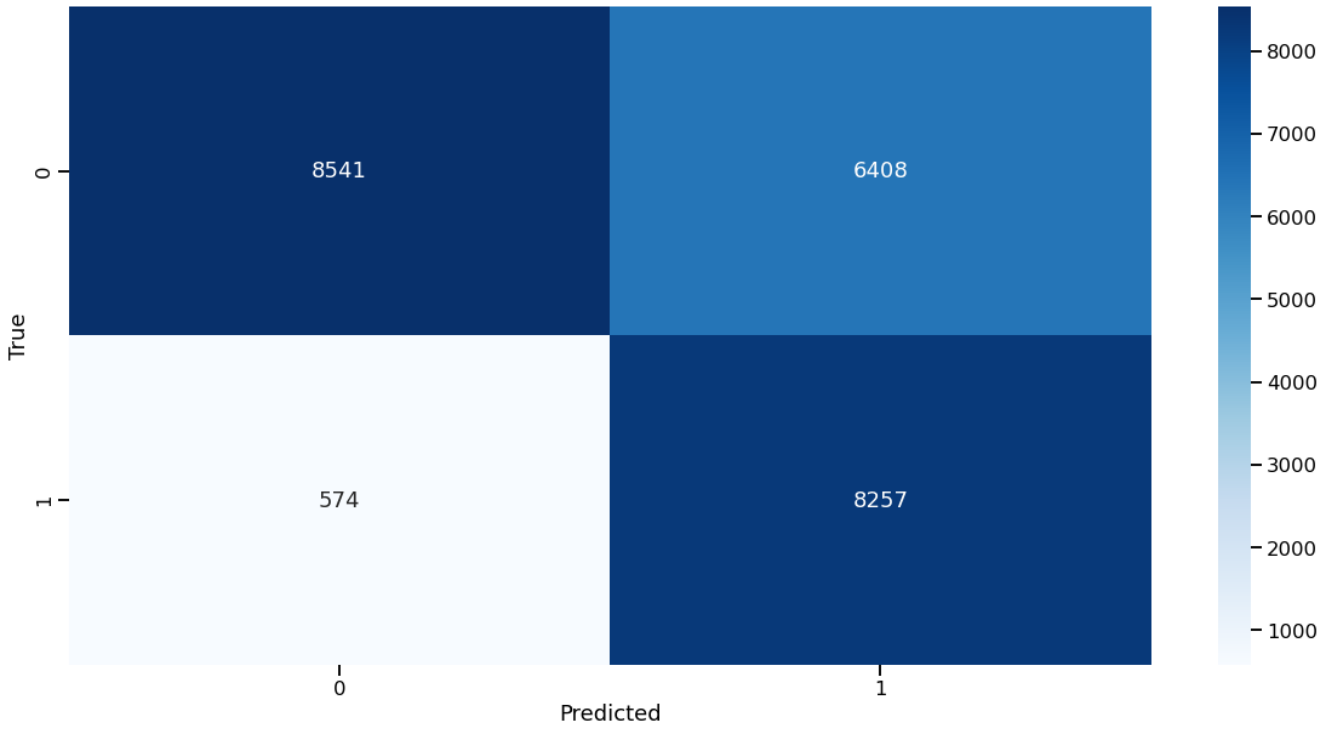


Figura 5. Matriz de confusión

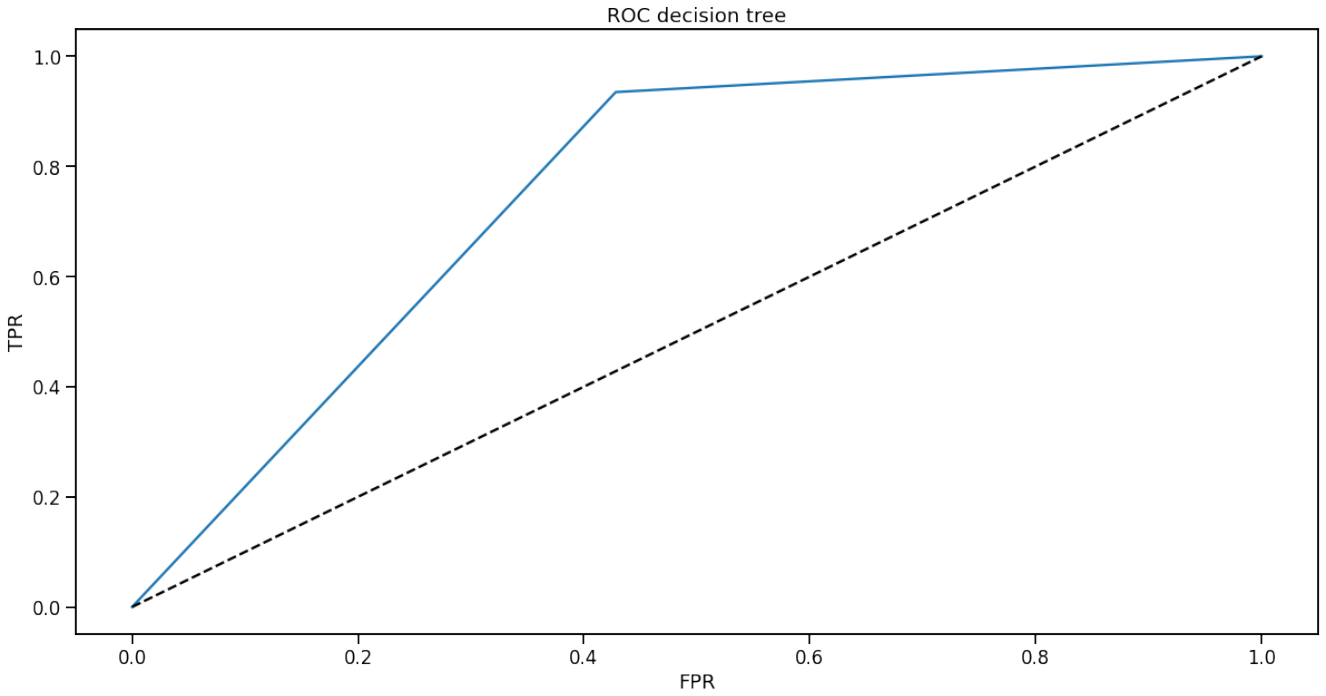


Figura 6. ÇurvaROC