

# Cancelación hotelera

Un modelo de predicción para la cancelación  
de reservas

Antonio Garcia Valverde

# Planteamiento inicial





# Planteamiento inicial

## **Problema de negocio**

Una cadena hotelera solicita el desarrollo de un modelo clasificador capaz de anticiparse a una posible cancelación de las reservas efectuadas en sus alojamientos.

Los motivos por los que considera relevante el desarrollo son los siguientes:

**Anticipación a la pérdida de negocio.**

**Mejora en las políticas de cancelación.**

**Elaboración de otros informes.**



# Planteamiento inicial

## Acceso a datos

La compañía facilita dos conjuntos de datos recopilados en sus espacios entre los años **2015-2018**.

Los dataset cuentan con **155.665** registros ordenados en un número diferente entre sí de columnas, aunque son coincidentes en las variables principales.

En ambos conjuntos se cuenta con un campo específico que indica si la reserva fue confirmada o cancelada y que se empleará como target.





# Planteamiento inicial

## Problema técnico

Para dar respuesta al problema de negocio planteado se propone el desarrollo de un modelo que reconozca las características que comparten las reservas canceladas.

Se trata por tanto de un

**problema supervisado de clasificación.**

Para evaluar el modelo se empleará una métrica que tenga en cuenta el número global de registros mal clasificados, independientemente de si trata de falsos positivos o falsos negativos, porque ambos casos pueden ser perjudiciales para el rendimiento del negocio.

La métrica escogida es **balanced\_accuracy.**

# Análisis exploratorio de datos - EDA



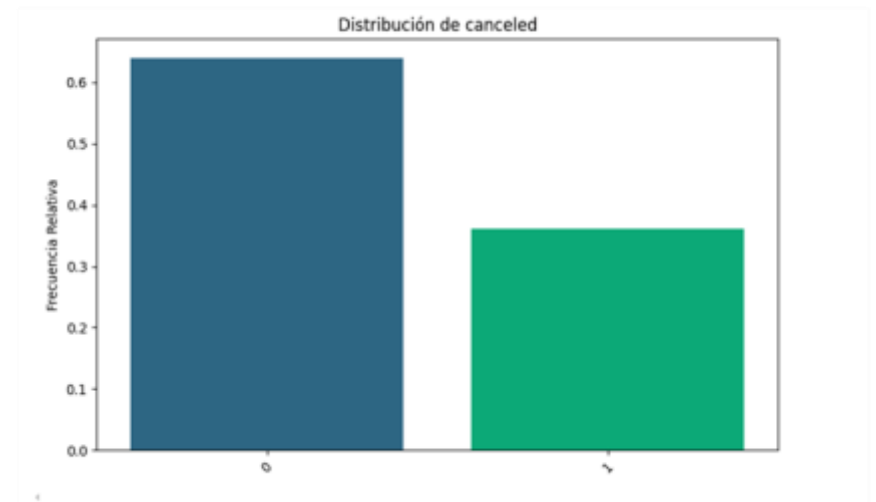


# EDA

## Variable target

La primera observación es que la variable directora, que nos indica si la reserva está confirmada o cancelada, tiene muchos más registros en una de las opciones.

Se trata, por tanto, de una variable **desbalanceada**, como puede comprobarse en el gráfico de frecuencia para ambos casos.





# EDA

## Features

Tras el análisis visual del conjunto de variables de forma individual, combinadas de dos en dos y combinadas con la variable directora, se aplican los siguientes cambios:

**Eliminación de nulos y valores inconsistentes**

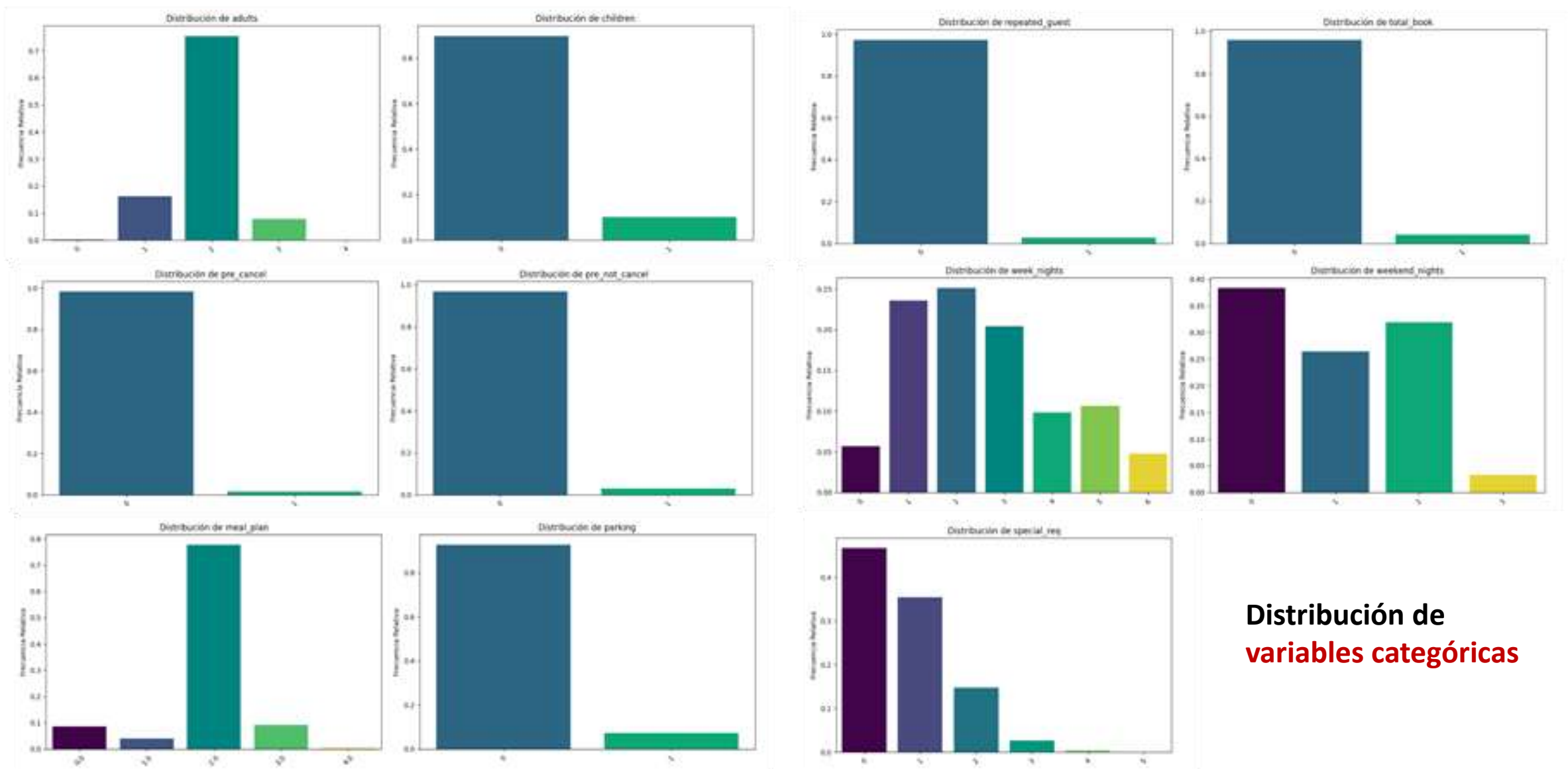
**Aplicación de función logística  
a las variables 'daily\_price' y 'lead\_time'**

**Categorización binaria de las variables categóricas  
'children', 'total\_book', 'pre\_cancel',  
'pre\_not\_cancel' y 'parking'**

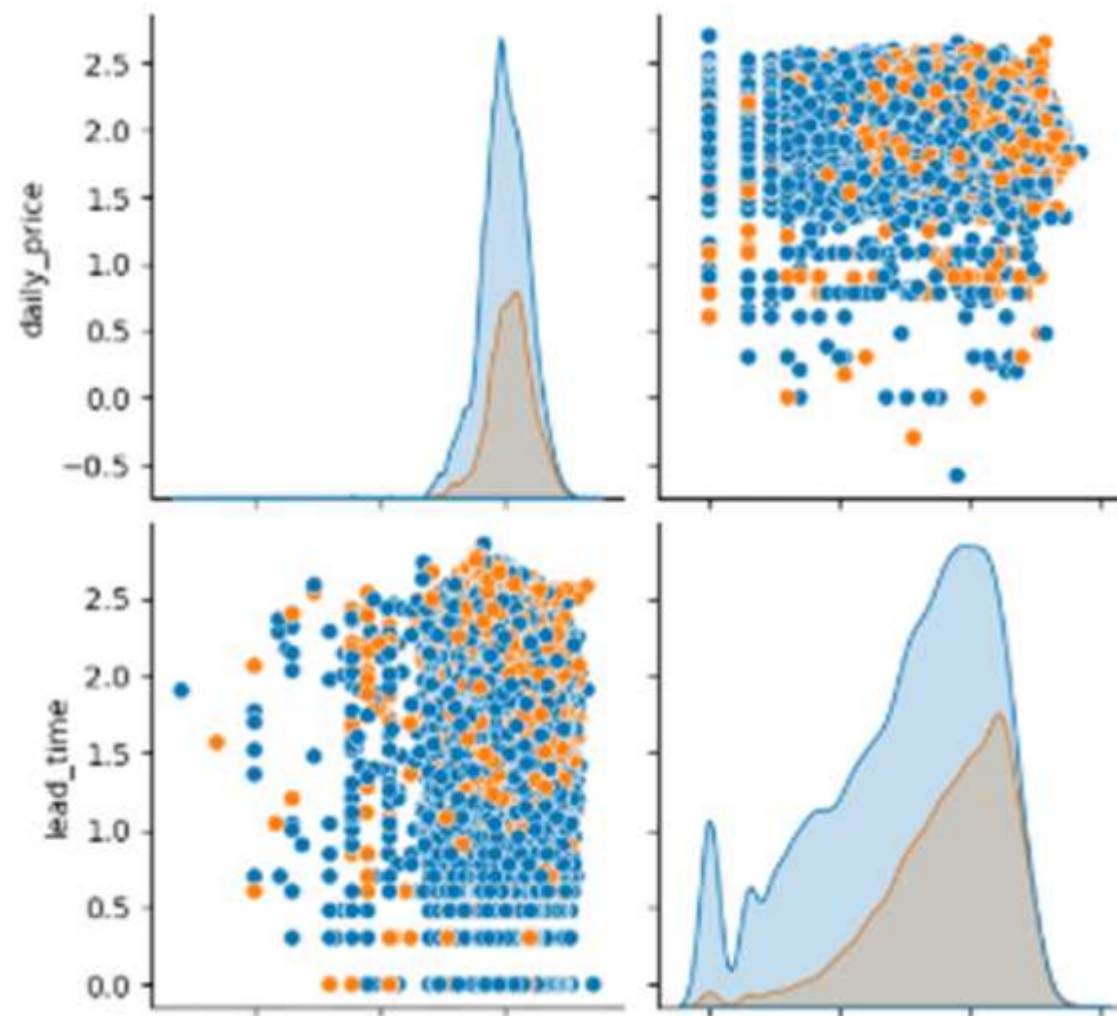
**Codificación ordinal de la variable 'meal\_plan'**

**Eliminación de la variable 'arr\_date'**





**Distribución de  
variables categóricas**



Distribución de **variables numéricas**



A vibrant tropical beach scene. In the foreground, the fronds of palm trees hang down from the top of the frame. The ocean is a brilliant turquoise color with white foam from waves breaking onto a sandy beach. The sky is a clear, bright blue with a few wispy white clouds. In the distance, a small island is visible on the horizon.

# **Selección de modelo**





# Selección de modelo

## Paso 1

Desarrollo de **Baseline** como punto de partida para la selección del modelo a optimizar y mejorar.

## Paso 2

**Optimización** de hiperparámetros.

## Paso 3

Entreno y **evaluación**.



# Selección de modelo

## Paso 1

Desarrollo de **Baseline** como punto de partida para la selección del modelo a optimizar y mejorar.

## Paso 2

Optimización de hiperparámetros.

## Paso 3

Entreno y **evaluación**.

```
DecisionTreeClassifier(random_state=42)
RandomForestClassifier(random_state=42)
XGBClassifier(base_score=None, booster=None, callbacks=None,
               colsample_bylevel=None, colsample_bynode=None,
               colsample_bytree=None, device=None, early_stopping_rounds=None,
               enable_categorical=False, eval_metric=None, feature_types=None,
               feature_weights=None, gamma=None, grow_policy=None,
               importance_type=None, interaction_constraints=None,
               learning_rate=None, max_bin=None, max_cat_threshold=None,
               max_cat_to_onehot=None, max_delta_step=None, max_depth=None,
               max_leaves=None, min_child_weight=None, missing=nan,
               monotone_constraints=None, multi_strategy=None, n_estimators=None,
               n_jobs=None, num_parallel_tree=None, ...)
LGBMClassifier(random_state=42, verbose=-100)
<catboost.core.CatBoostClassifier object at 0x000001C784A56EA0>
Model <DecisionTree>, Accuracy_CV: 0.804466399005277
Model <Random Forest>, Accuracy_CV: 0.826950019786295
Model <XGBoost>, Accuracy_CV: 0.7733859840491862
Model <LightGBM>, Accuracy_CV: 0.758100957794794
Model <CatBoost>, Accuracy_CV: 0.774147052608057
```

El modelo con mejor métrica es **Random Forest**

# Selección de modelo

## Paso 1

Desarrollo de **Baseline** como punto de partida para la selección del modelo a optimizar y mejorar.

## Paso 2

**Optimización** de hiperparámetros.

## Paso 3

Entreno y **evaluación**.

Valores de hiperparámetros seleccionados por **Optuna** según los criterios de ajuste establecidos.

```
criterion = "log_loss"  
n_estimators = 290  
max_depth = 30  
min_samples_split = 2  
min_samples_leaf = 1  
max_features = None  
class_weight = "balanced"
```

# Selección de modelo

## Paso 1

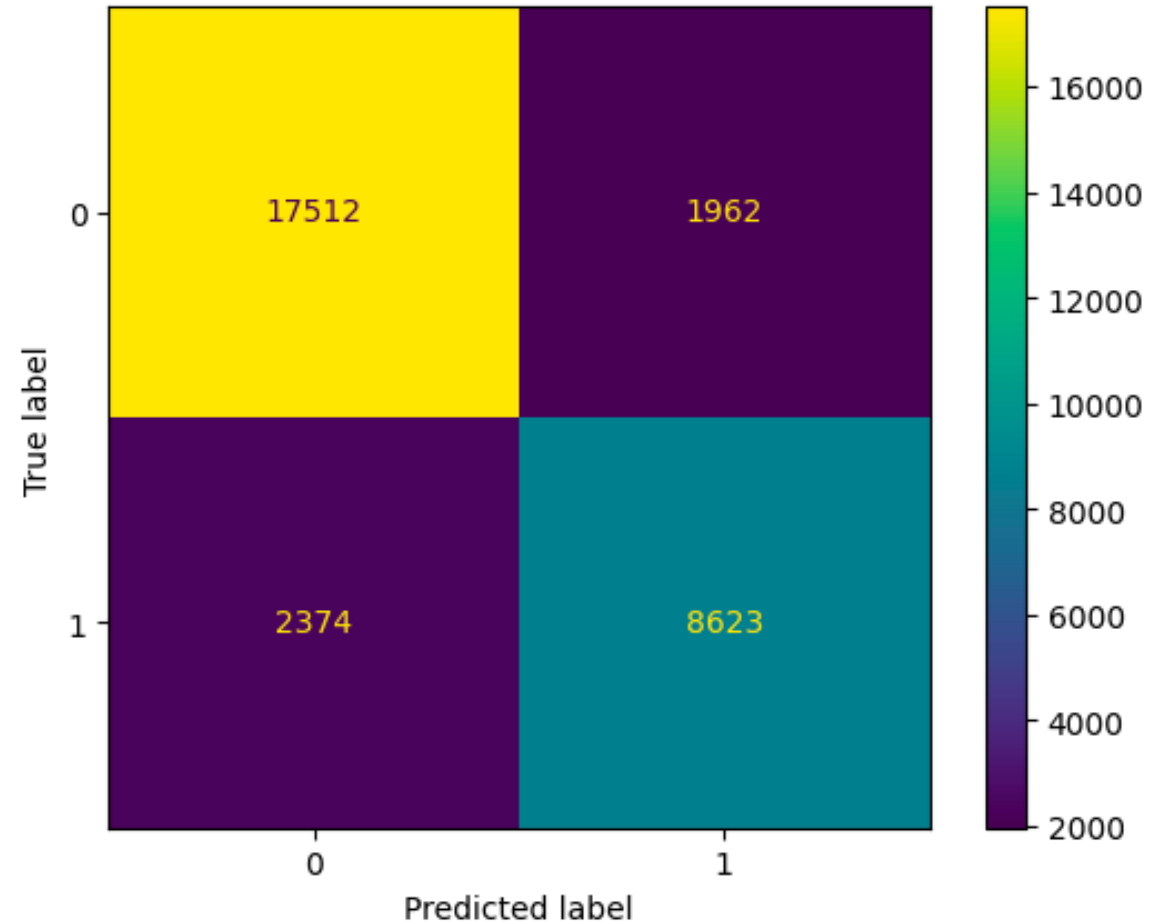
Desarrollo de **Baseline** como punto de partida para la selección del modelo a optimizar y mejorar.

## Paso 2

Optimización de hiperparámetros.

## Paso 3

Entreno y **evaluación**.





# Resultado

## **Problema de negocio**

Una cadena hotelera solicita el desarrollo de un modelo clasificador de las reservas efectuadas en sus alojamientos capaz de anticiparse a una probable cancelación de la reserva.





# Resultado

Se desarrolla un **modelo clasificador** basado en **árboles de decisión**.

El porcentaje de precisión general en la clasificación de las futuras reservas logra alcanzar el

**86%**

# Cancelación hotelera

Un modelo de predicción para la cancelación  
de reservas

Antonio Garcia Valverde