

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

# SATISFACCIÓN DE VUELOS EN AEROLÍNEAS

DATA SCIENCE



# ¿POR QUÉ NECESITA UNA AEROLÍNEA CONOCER LOS NIVELES DE SATISFACCIÓN DE SUS PASAJEROS?



MÉTRICA CLAVE

¿POR QUÉ ESTÁN O NO SATISFECHOS?  
SABERLO AYUDA A COMPRENDER PUNTOS  
FUERTES Y DÉBILES DE LA AEROLÍNEA DESDE LA  
PERSPECTIVA DEL CLIENTE



# ENTENDER LOS DATOS



Visualizar los datos nos puede evitar hacer supuestos incorrectos.

Se usan estadísticas de resumen y herramientas gráficas para llegar a conocer los datos y comprender lo que se puede averiguar de ellos.



# VENTAJAS

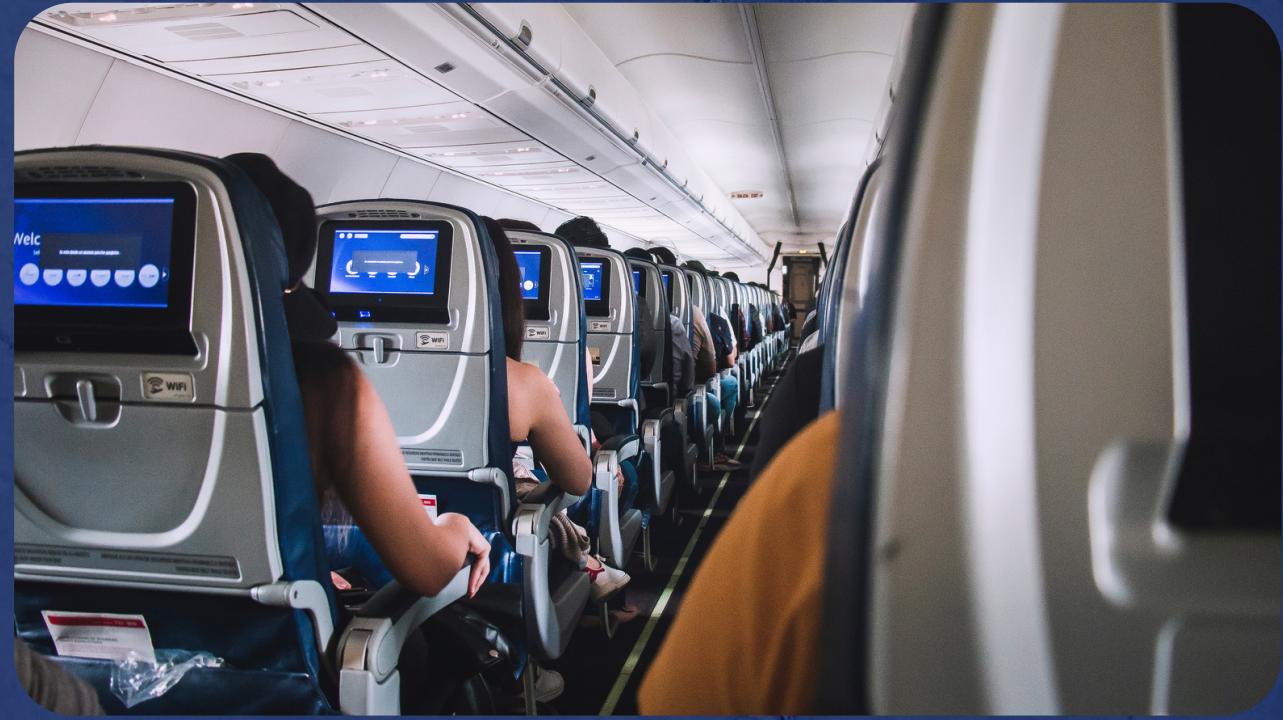
- Toma de decisiones
- Análisis de atributos específicos
- Metodología imparcial y científica
- Potencial ilimitado



# Entendimiento del Negocio/Problema

Planear un vuelo en la aerolínea conlleva:

- Comprar el ticket
- Abordar el avión
- Transitar el vuelo
- Desembarcar
- Recibir el equipaje –si corresponde–.



El conjunto de datos muestra si el transcurso fue satisfactorio o no. Se tienen en cuenta:  
23 variables  
129880 observaciones



# EDA: ANÁLISIS EXPLORATORIO



# VARIABLES

cualitativa categórica

- Satisfaction (Satisfacción)
- Customer Type (Tipo de cliente)
- Gender (Género)
- Class (Clase)
- Type of Travel (Tipo de viaje)



# VARIABLES

cuantitativa discreta

- Flight Distance (Distancia de vuelo)
- Departure Delay in Minutes (Retraso de salida en minutos)
- Arrival Delay in Minutes (Retraso de llegada en minutos)
- Age (Edad)

# VARIABLES

cualitativa ordinal

- Seat comfort (Comodidad del asiento),
- Departure/Arrival time convenient (Hora de salida/llegada conveniente),
- Food and drink (Comida y bebida),
- Gate location (Ubicación de la puerta)
- Inflight wifi service (Servicio de wifi a bordo),
- Inflight entertainment (Entretenimiento a bordo)
- Online support (Soporte en línea),
- Ease of Online booking (Facilidad de reserva en línea)
- On-board service (Servicio a bordo),
- Leg room service (Servicio de sala de piernas)
- Baggage handling (Gestión de equipaje),
- Checkin service (Servicio de facturación),
- Cleanliness (Limpieza)
- Online boarding (Embarque en línea)



# 1. Exploración:



- Ver primeros registros
- Ver columnas
- Dimensiones dataset : (129880, 23)
- ¿Hay valores nulos?
- Distribución variables numéricas (medidas de t.central, dispersión, etc)
- Comportamiento variables categóricas (clases, frecuencia)





- Reemplazo columnas categóricas por valores numéricos:

Ej: **satisfaction**

0=dissatisfied, 1=satisfied

- Las variables dummy ("ficticias") se utilizan para explicar valores cualitativos en un modelo de regresión

\*\*La función get\_dummies permite eliminar la primera de las columnas generadas para cada característica codificada para evitar la denominada colinealidad (que una de las características sea una combinación lineal de las otras), lo que dificulta el correcto funcionamiento de los algoritmos.





# VARIABLE OBJETIVO:

Expresa el grado de satisfacción del cliente (Satisfecho/Insatisfecho).

**SATISFACCIÓN**

Cualitativa categórica – valores posibles:  
"satisfied" : 71087  
"dissatisfied" : 58793



# satisfecho

55%



La media de 0.547328 y la mediana nos indican que más de la mitad está satisfecho



# Ejemplo: variable objetivo agrupada por género



Insatisfecho :

Female 22971

Male 35822

Satisficho:

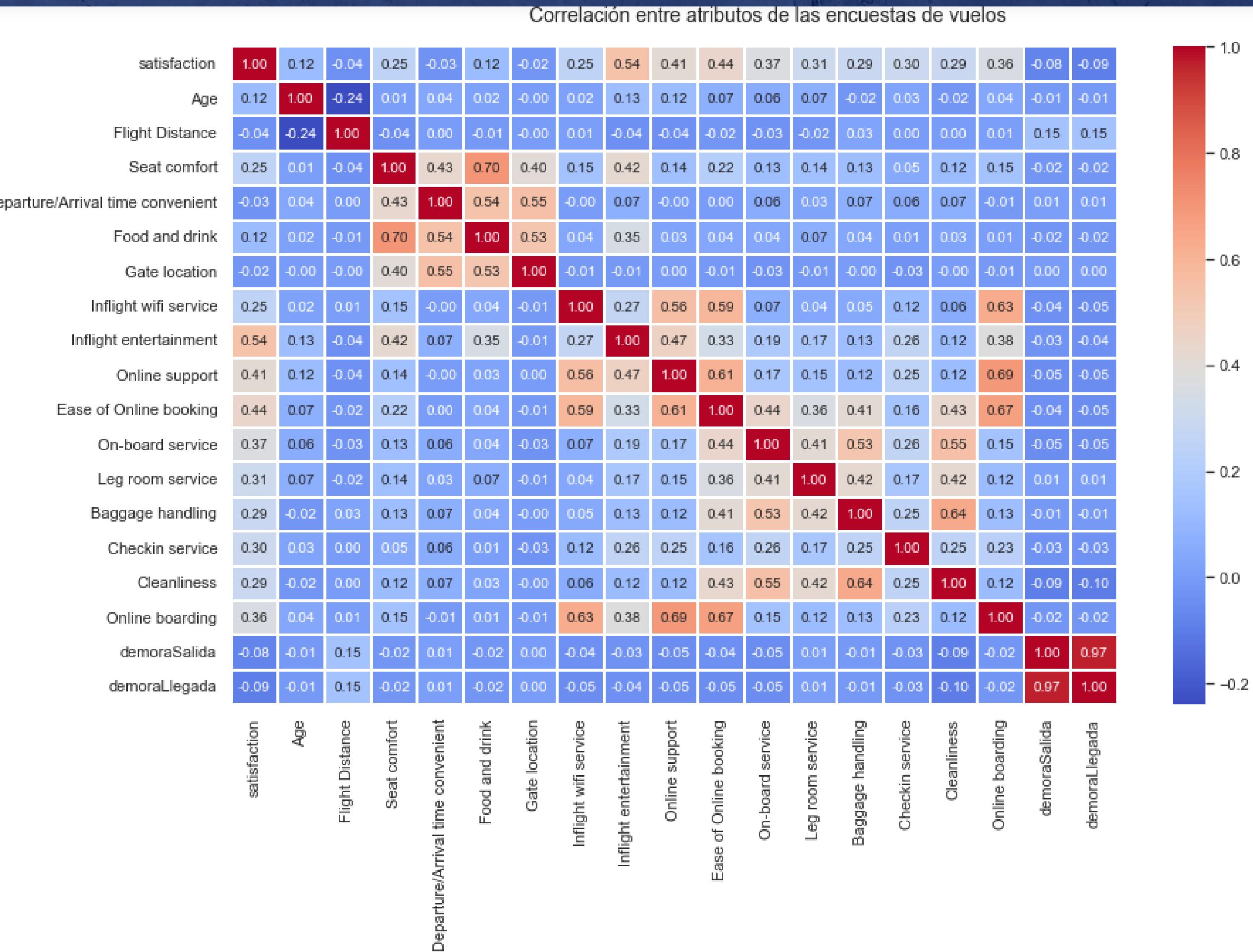
Female 42928

Male 28159

# HEAT MAP

## Correlación

Correlación entre atributos de las encuestas de vuelos



# BASELINE PREDICTION ALGORITHM



$$\hat{Y} = f(\hat{X})$$



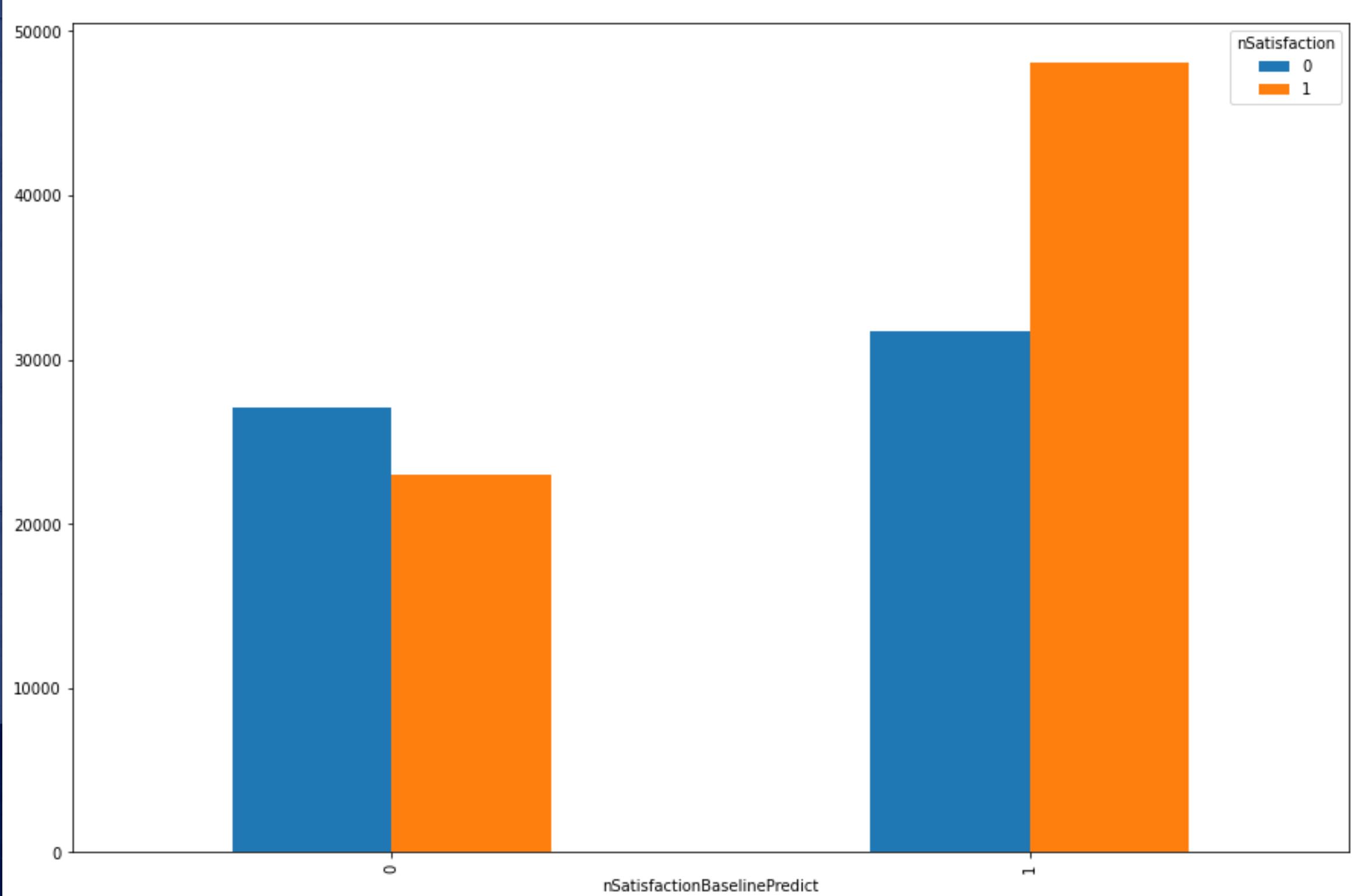
Estimación de  
la realidad

Función de  $X$   
para estimar la  
realidad.

- a. Definimos una función que acepta una lista de features y devuelve la prueba RMSE
- b. Creamos lista de features
- c. Creamos nuevamente X and Y
- d. Definimos X e y para el set de entrenamiento y el de test
- e. Creamos un array numpy con la misma forma que y\_test
- f. Rellenamos el array con el valor medio de y\_test



# Baseline Predict



- Los pasajeros frecuentes en viaje personal que utilizan clase económica masculinos están insatisfechos.
- Los pasajeros que viajan por negocios en clase económica están insatisfechos.
- El resto esta satisfecho

# SEPARACIÓN EN TRAIN Y TEST



# YA TENEMOS NUESTROS CONJUNTOS TRAIN Y TEST

- X = Se utilizan todas las características del dataset (las columnas categóricas se reemplazan por nuevas columnas numéricas: Columna --> nColumna. (22 variables)).
  - y = Se utiliza la característica nSatisfaction que reemplaza a satisfaction que también es categórica (Variable objetivo o **Target**)
  - Se observa que el tamaño del conjunto de test es 0.33
- 
- De los 3 enfoques estudiados:
    1. Validation Set (Hold-Out): En este enfoque mezclamos el dataset y utilizamos la mitad para entrenar y la otra mitad como set de validación
    2. Leave One Out Cross Validation (LOOCV): Este enfoque consiste en dejar fuera una observación la cual se usará para testear y utilizar el resto para poder entrenar
    3. K-Fold Cross Validation: Este enfoque consiste en separar el dataset en K-Folds y entrenar  $K$  veces el modelo utilizando  $K_i$  como set de validación

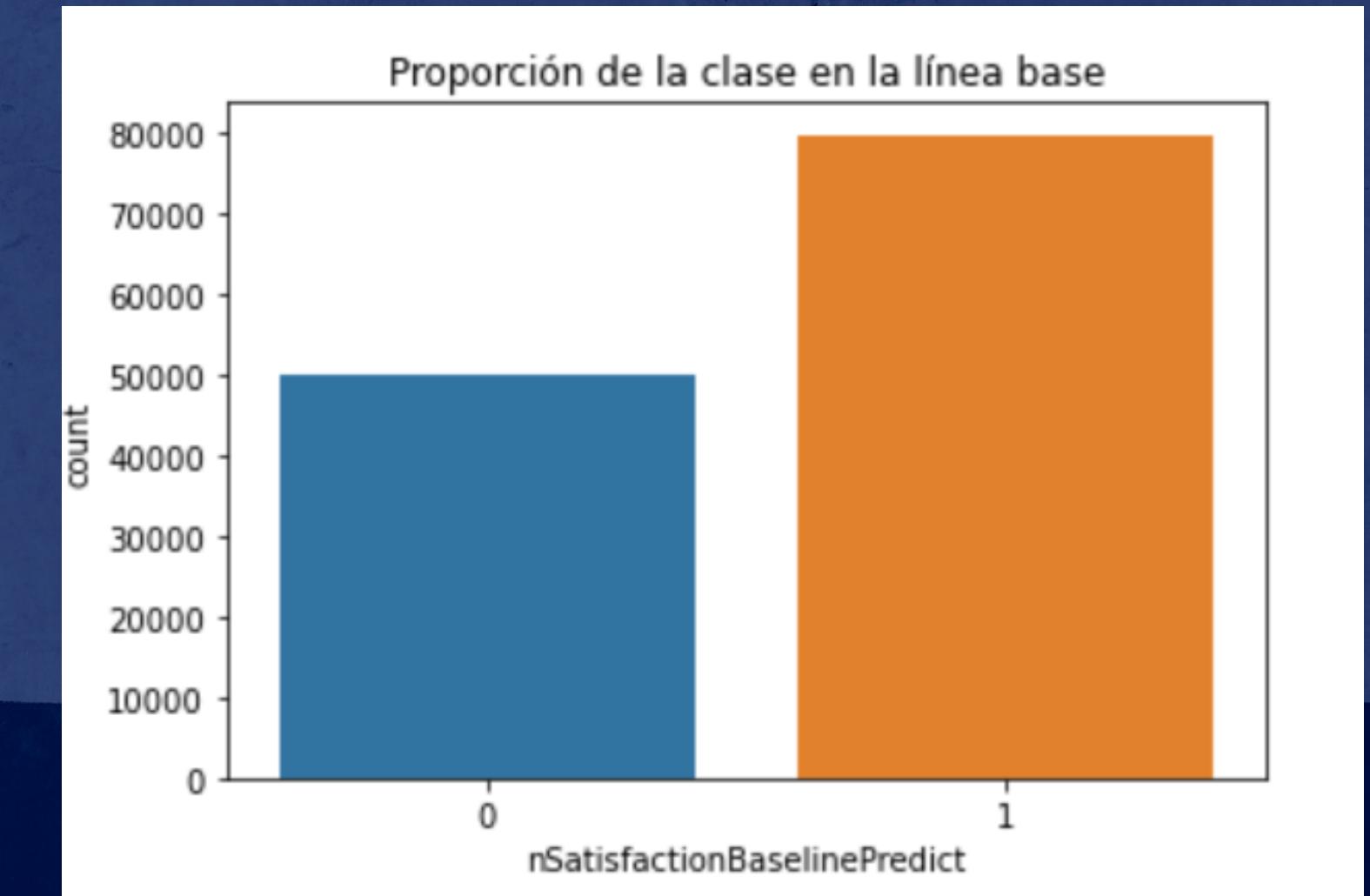
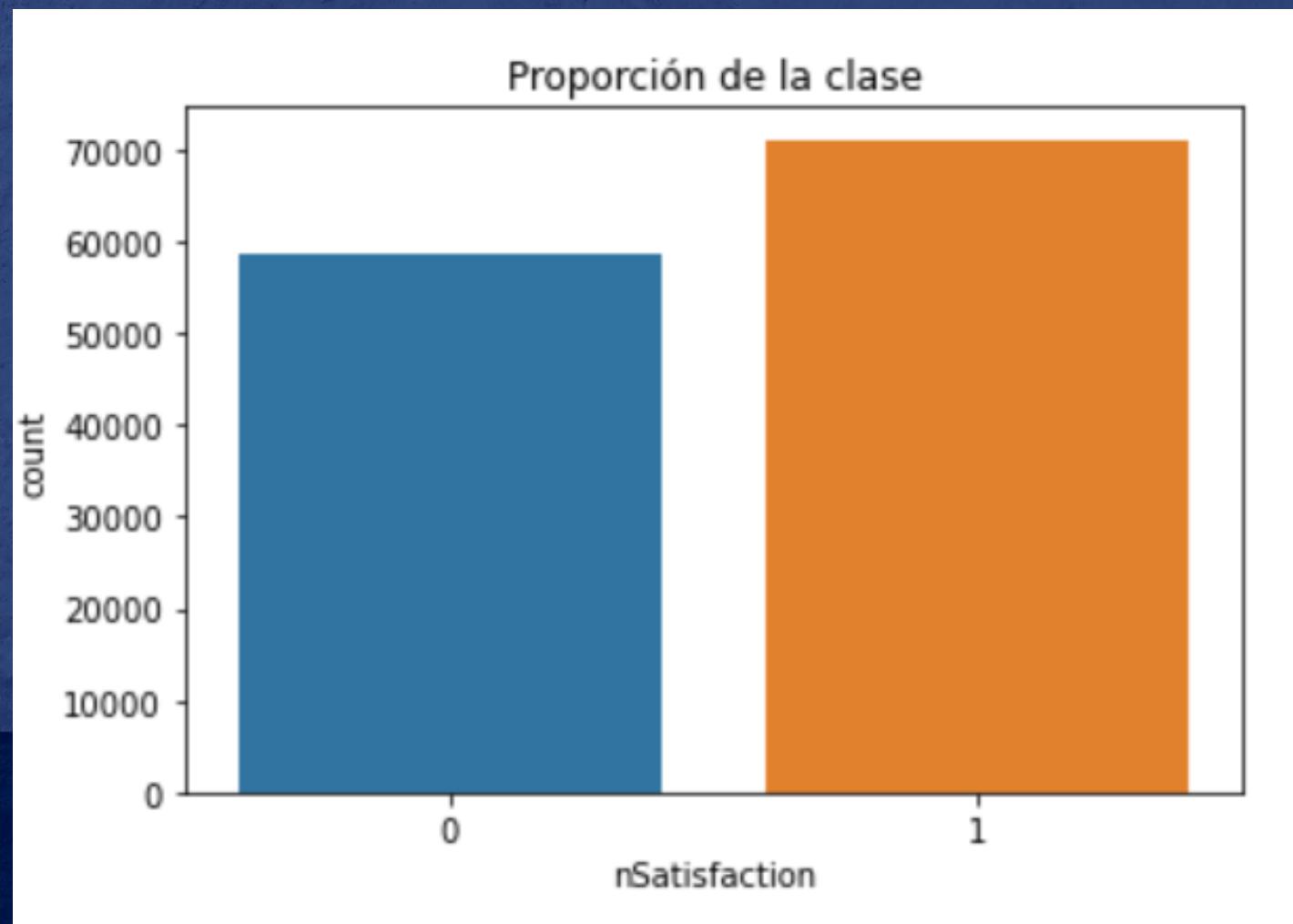
# EVALUACIÓN DE RENDIMIENTO

- Rendimiento de Regresión Lógistica:  $0.776 \pm 0.006$
- Rendimiento de Árbol de decisión:  $0.937 \pm 0.003$
- Rendimiento de Bagging AD:  $0.965 \pm 0.002$
- Rendimiento de Random Forest:  $0.965 \pm 0.001$
- Rendimiento de Extra Trees:  $0.964 \pm 0.001$
- Rendimiento de AdaBoostClassifier:  $0.91 \pm 0.002$
- Rendimiento de GradientBoostingClassifier:  $0.93 \pm 0.001$

Como tipo de separador de los conjuntos de train y test usamos: K-Fold Cross Validation para la evaluación de los modelos.

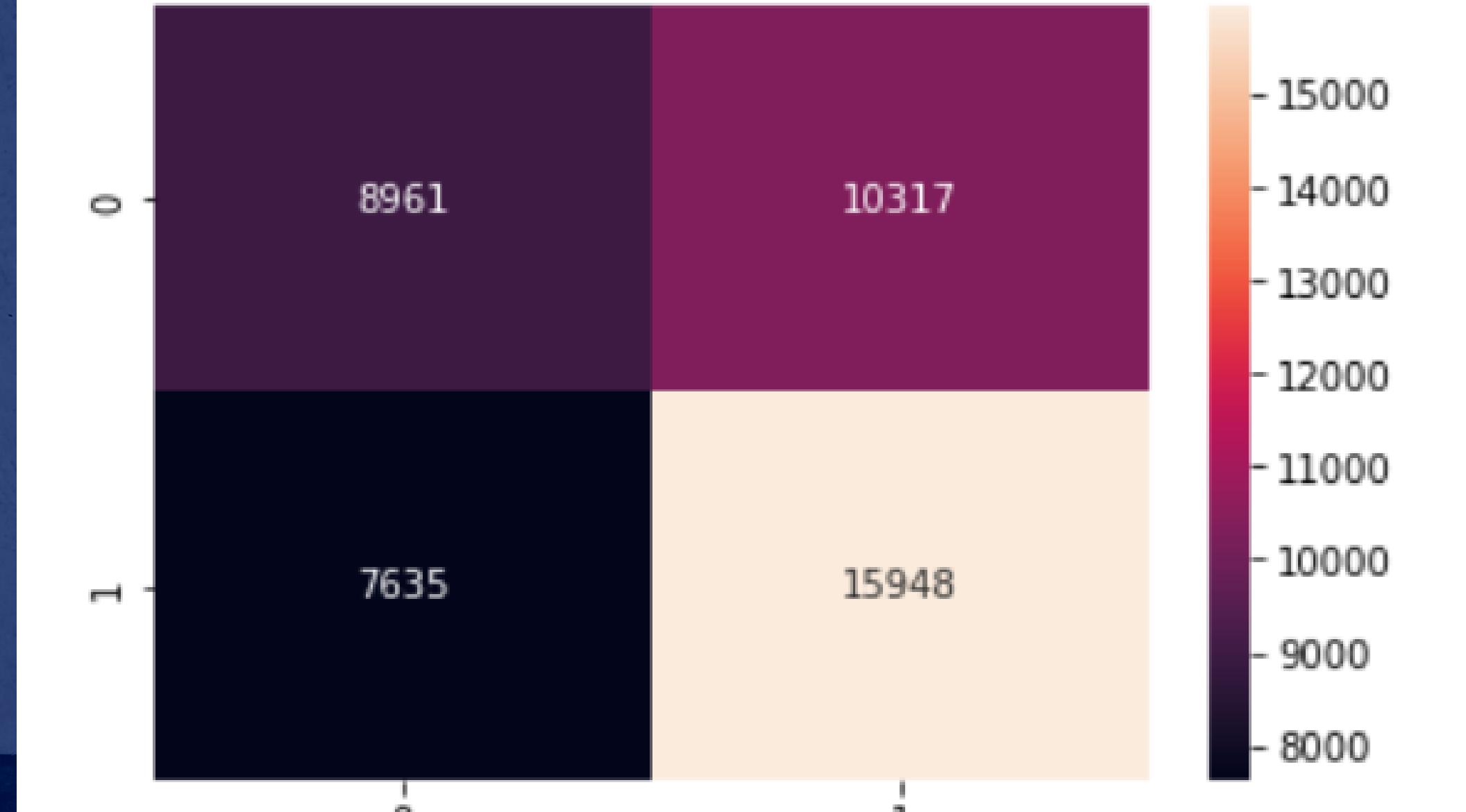
# ELECCIÓN: RANDOM FOREST

# PROPORCIÓN DE LA CLASE/PROPORCIÓN EN BASELINE



# Matriz de Confusión (RandomForestClassifier)

Matriz de confusión entre y\_test e y\_baseline predict



| Metrics   | dissatisfied | satisfied    | Classifiers | accuracy | macro avg    | weighted avg |
|-----------|--------------|--------------|-------------|----------|--------------|--------------|
| f1-score  | 0.499582     | 0.639865     | Línea Base  | 0.581158 | 0.569724     | 0.576769     |
| precision | 0.539949     | 0.607196     | Línea Base  | 0.581158 | 0.573573     | 0.576950     |
| recall    | 0.464830     | 0.676250     | Línea Base  | 0.581158 | 0.570540     | 0.581158     |
| support   | 19278.000000 | 23583.000000 | Línea Base  | 0.581158 | 42861.000000 | 42861.000000 |

# Importancia de features

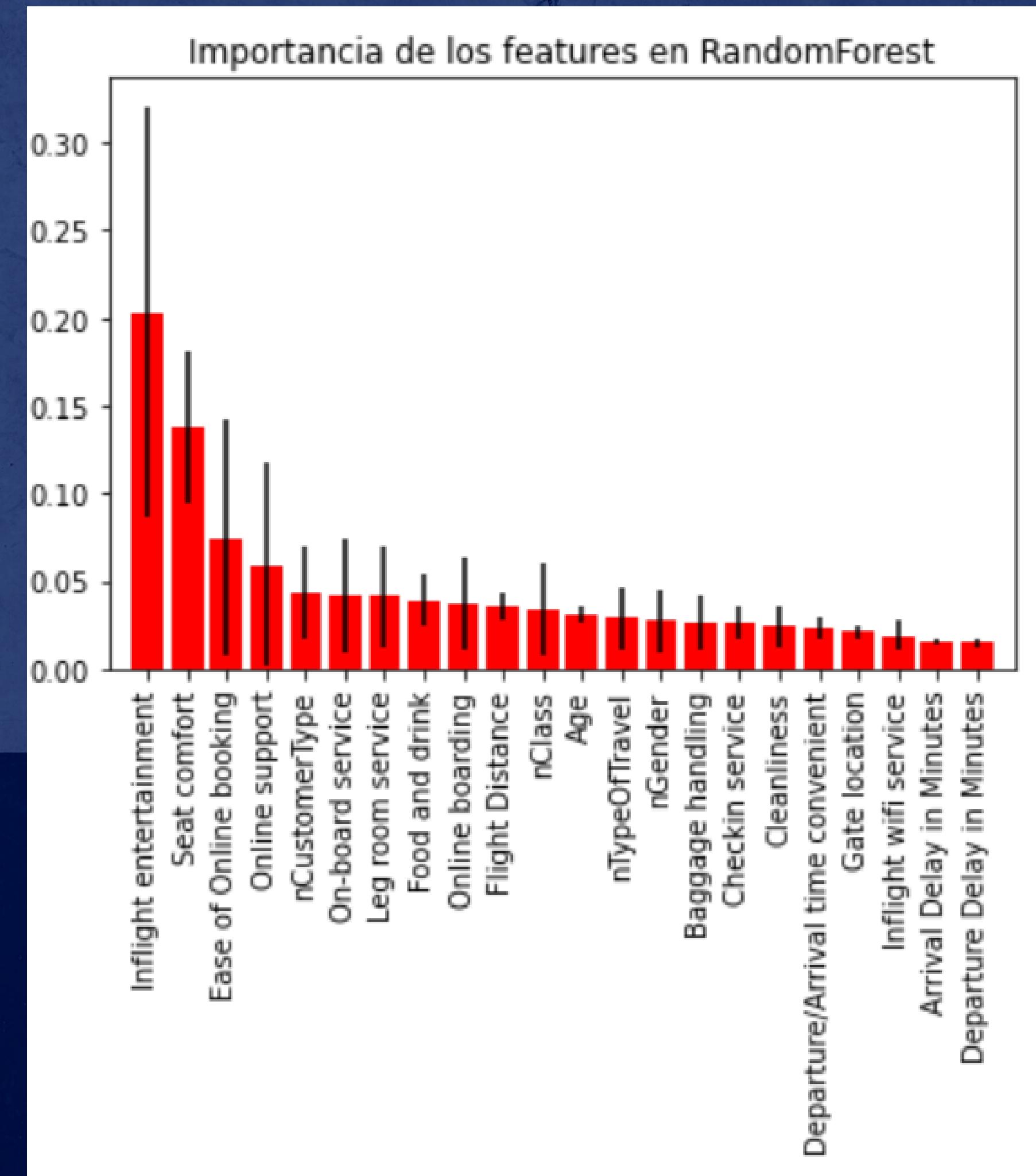
## (RandomForestClassifier)

1. Entretenimiento a bordo
2. Comodidad del asiento
3. Facilidad de reserva online

Pese a resultar contraintuitivo, la demora tiene baja incidencia en la satisfacción

Posibilidad:

- Son muchos las observaciones con demora muy baja o nula
- Las demoras grandes pueden tener motivos por los cuales el pasajero no hace responsable a la aerolínea



# RECALL

Del total de personas satisfechas,  
¿Cuántas logra clasificar  
correctamente el modelo?

# GridSearchCV: Validación cruzada

Mejor clasificador para nuestro modelo:  
ExtraTreesClassifier

|   | model               | best_score | best_params                                |
|---|---------------------|------------|--|
| 0 | logistic_regression | 0.847508   | {'C': 5}                                   |
| 1 | decision_tree       | 0.940910   | {'criterion': 'gini', 'max_depth': 20}     |
| 2 | random_forest       | 0.946194   | {'n_estimators': 20}                       |
| 3 | extra_tree          | 0.946931   | {'n_estimators': 20}                       |
| 4 | ada_boost           | 0.894177   | {'n_estimators': 20}                       |
| 5 | gradient_boosting   | 0.934385   | {'learning_rate': 0.8, 'n_estimators': 20} |

¿Qué puede ofrecer nuestro modelo a  
su aerolínea?



- **Medir y poder predecir la satisfacción de los pasajeros**
- **Examinar las tendencias de la industria para facilitar tanto la evaluación de iniciativas pasadas como la configuración de estrategias futuras**
- **Comprender las necesidades y preferencias de sus clientes para identificar las áreas clave de mejora que tendrán el mayor impacto en la satisfacción de los pasajeros**
- **Analizar los atributos detallados para definir acciones de mejora claras y específicas**
- **Construir estrategia de experiencia del cliente con evidencia creíble**
- **Hacer un seguimiento del éxito de las mejoras implementadas y evaluar retorno de la inversión**

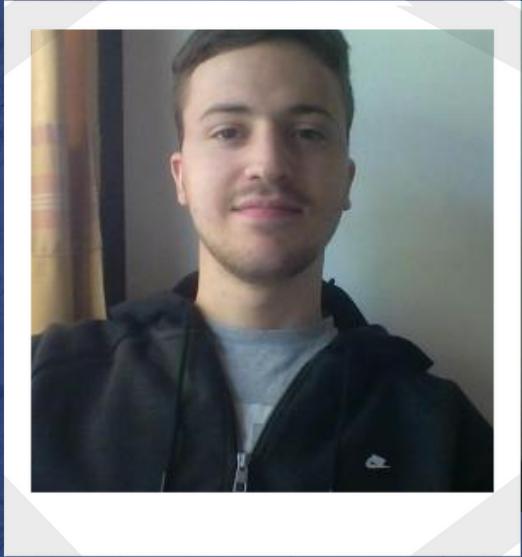
# Algunas preguntas y respuestas:

- Los pasajeros de género femenino presentan mayor % de satisfacción
- Los pasajeros mas insatisfechos son clientes leales masculinos que viajan por negocios en clase económica
- Los pasajeros más satisfechos son clientes leales femeninos que viajan por motivo personal en clase económica
- Las variables más correlacionadas entre sí, son demoras en partida y arribo. pero no muestran correlación con nuestra variable objetivo.
- Entretenimiento a bordo, Comodidad del asiento y Facilidad de reserva online son las features que más impactan en la

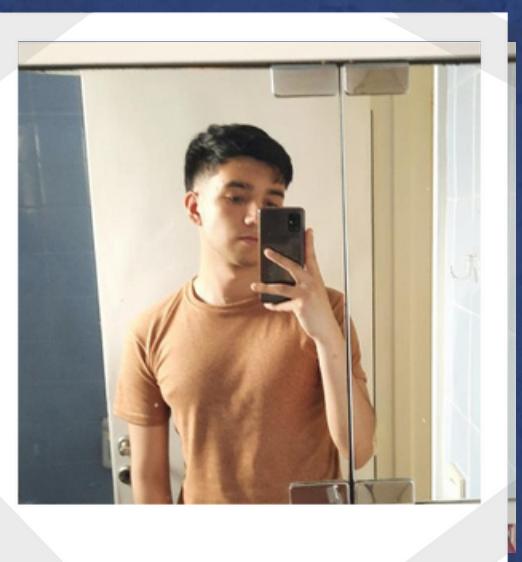
# Posibles estrategias:

- Diseñar estrategias divididas por género.
- Diseñar paquetes para vender a empresas de vuelos con upgrade
- Apuntar al público femenino con vuelos que ya tienen costo hundido ( vuelos en temporada baja, regresos vacíos, etc)
- Invertir en mejorar la variedad de entretenimiento a bordo
- Invertir en acondicionar los asientos. Los clientes pueden estar dispuestos a pagar más por un upgrade que los incluyan. También se pueden reacomodar pasajeros una vez que el vuelo cerró y tiene asientos disponibles en categoría superior.
- Invertir en mejorar el sitio web y la experiencia de reserva onlline en general.

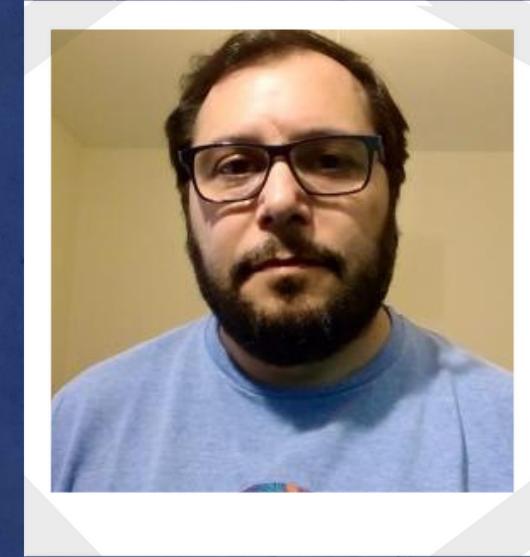
# Grupo 3



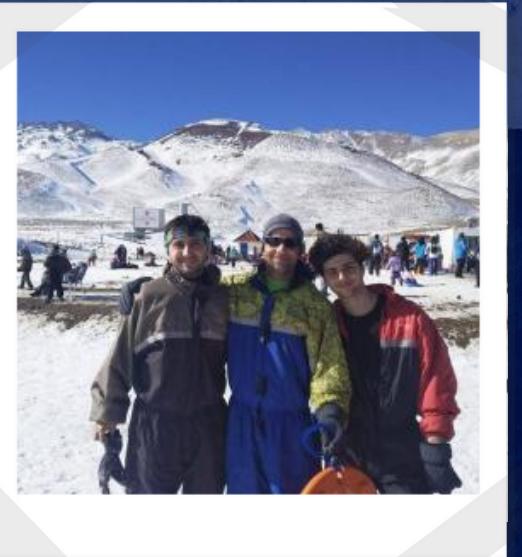
Matías Fasulino



Mateo Holzer



Fernando Llundai



Octavio Luna



Catalina Miganne



Martín Muñoz



❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖ ❖

Thank you.

