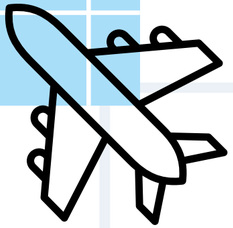
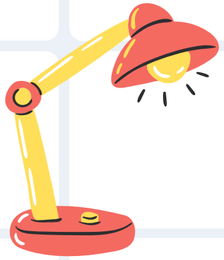
PREDICCIÓN DE PRECIO DE VUELOS



MICAELA MUZI
CODERHOUSE - DATA SCIENCE
COMISION 32755

INDICE



INTRODUCCIÓN

En la actualidad, las **empresas que venden boletos de avión online** necesitan conocer el **valor de los boletos** para elaborar promociones en su sitio web, y así poder atraer más clientes.
Pero se encuentran con un problema, y es que el valor de los tickets **depende de muchas variables** y **cambia con el correr de los días**.

Este proyecto pretende, mediante el uso de diferentes herramientas, facilitar el **análisis de la fluctuación de precios** y permitir el uso de un **modelo fiable para la predicción** de los mismos.

Se utilizará un dataset de compra de tickets de avión obtenido de la página 'EaseMyTrip'. Dicho dataset contiene 300k registros de compras en un rango de 50 días, desde Febrero a Marzo del 2022. Los vuelos son entre las 6 ciudades más grandes de India.

EaseMyTrip.com

HIPÓTESIS

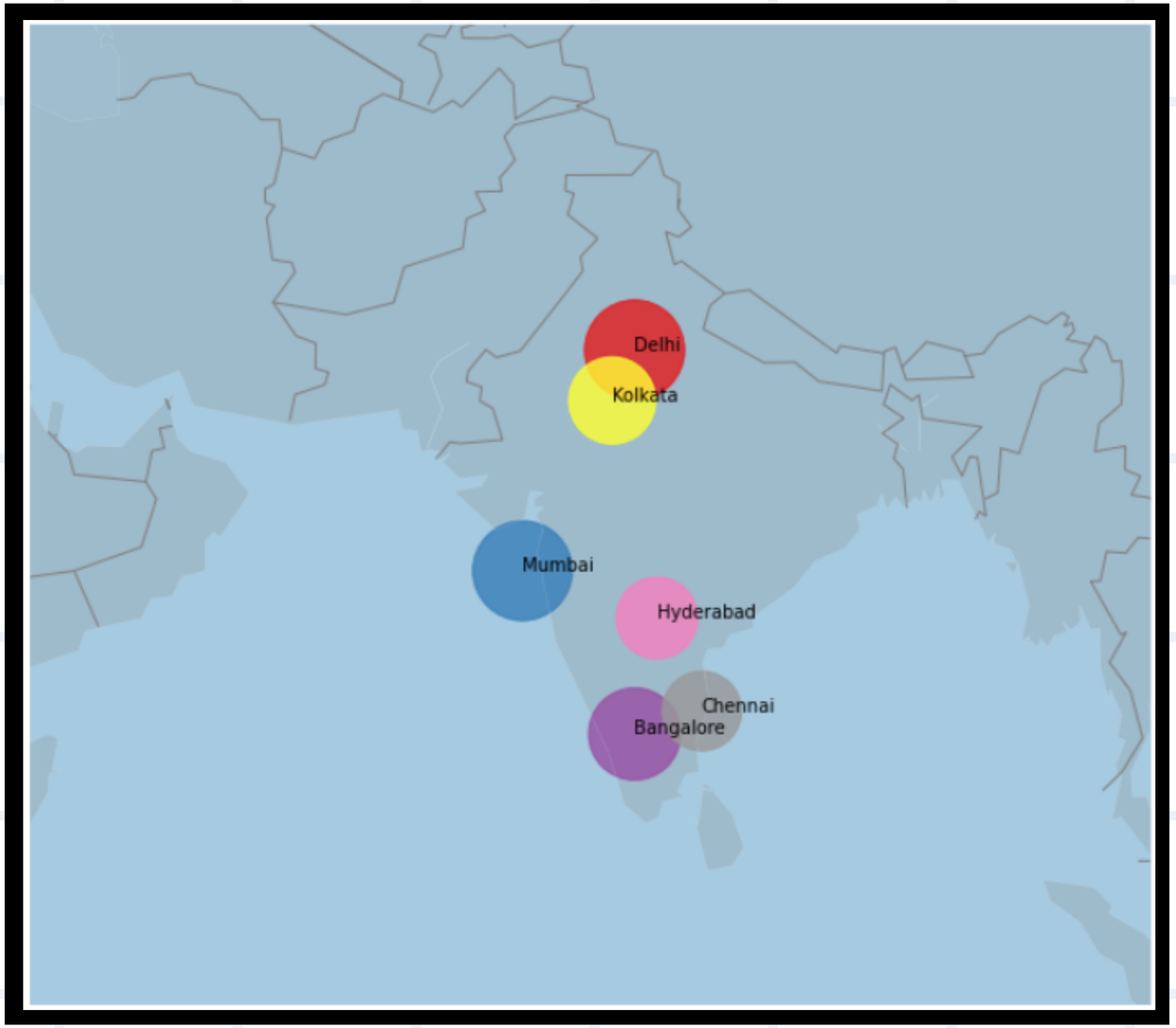
- A **mayor** duración de vuelo, **mayor** precio
- A **mayor** cantidad de días restantes para el vuelo, **menor** precio
- Aerolínea con **precio más elevado** es Vistara y el **menor** es AirAsia
- Las siguientes combinaciones son las **más baratas** (por proximidad):

Ciudad origen	Ciudad destino
Hyderabad	Mumbai
Chennai	Hyderabad
Bangalore	Chennai

Ciudad origen	Ciudad destino
Delhi	Kolkata
Kolkata	Delhi
Chennai	Bangalore

- Los vuelos directos (cero escalas) presentan **menor precio**
- Los vuelos que salen muy temprano (Early morning) son los **más elegidos**

BUBBLE MAP



METODOLOGÍA

DATASET INICIAL

DATA WRANGLING

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

SPLIT TRAIN/TEST

ENTRENAMIENTO Y PREDICCIÓN

EVALUACIÓN DE PERFORMANCE

CONCLUSIONES

DATA WRANGLING

Primera visualización del dataset:

	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

Primer análisis de las variables numéricas:

	count	mean	std	min	25%	50%	75%	max
duration	300153.0	12.221021	7.191997	0.83	6.83	11.25	16.17	49.83
days_left	300153.0	26.004751	13.561004	1.00	15.00	26.00	38.00	49.00
price	300153.0	20889.660523	22697.767366	1105.00	4783.00	7425.00	42521.00	123071.00

De las variables categóricas:

	count	unique	top	freq
airline	300153	6	Vistara	127859
flight	300153	1561	UK-706	3235
source_city	300153	6	Delhi	61343
departure_time	300153	6	Morning	71146
stops	300153	3	one	250863
arrival_time	300153	6	Night	91538
destination_city	300153	6	Mumbai	59097
class	300153	2	Economy	206666

DATA WRANGLING

INSIGHTS INICIALES:

- El dataset cuenta con 300.000 rows y 12 columnas
- 8 variables categóricas (string) y 3 numéricas
- No se observan valores nulos
- No se observan valores duplicados
- Variable 'flight' es prescindible
- El promedio de la variable 'price' es mucho mayor que la mediana (2do cuartil). Esto indica que está sesgada a la derecha
- Presencia de outliers
- Casi la mitad de los vuelos se realizaron con aerolínea Vistara, la gran mayoría en clase Economy y de 1 escala
- Se convertirá la variable 'price' de rupias a dólares

DATA WRANGLING

Se aplica '**Label Encoding**' para asignar valores numéricos a las variables categóricas y así poder utilizarlas en un modelo de predicción:

Aerolínea
'SpiceJet': **1**
'AirAsia': **2**
'Vistara': **3**
'GO_FIRST': **4**
'Indigo': **5**
'Air_India': **6**

Ciudad origen
Ciudad destino
'Delhi': **1**
'Mumbai': **2**
'Bangalore': **3**
'Kolkata': **4**
'Hyderabad': **5**
'Chennai': **6**

Hora de partida
Hora de llegada
'Early_Morning': **1**
'Morning': **2**
'Afternoon': **3**
'Evening': **4**
'Night': **5**
'Late_Night': **6**

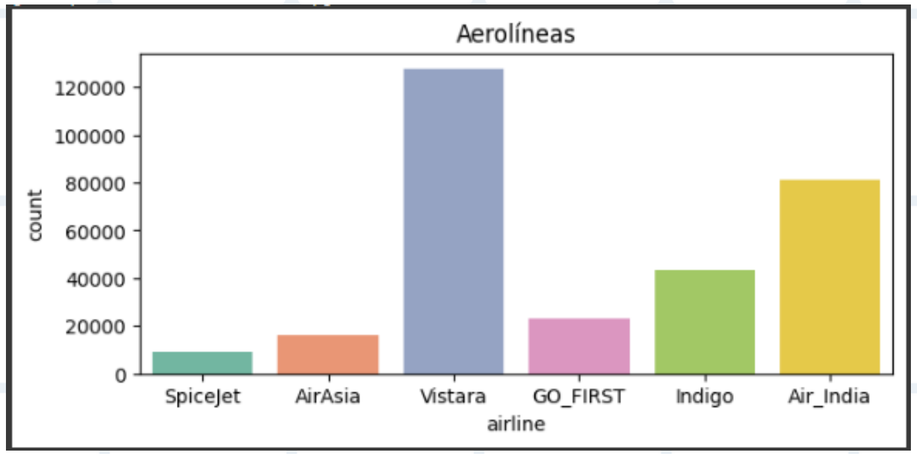
Escalas
'zero': **0**
'one': **1**,
'two_or_more': **2**

Clase
'Economy': **0**
'Business': **1**

	duration	days_left	airline_id	source_city_id	destination_city_id	departure_time_id	arrival_time_id	stops_id	class_id	price_usd
0	2.17	1	1	1	2	4	5	0	0	71
1	2.33	1	1	1	2	1	2	0	0	71
2	2.17	1	2	1	2	1	1	0	0	71
3	2.25	1	3	1	2	2	3	0	0	71
4	2.33	1	3	1	2	2	2	0	0	71

Versión final del dataset

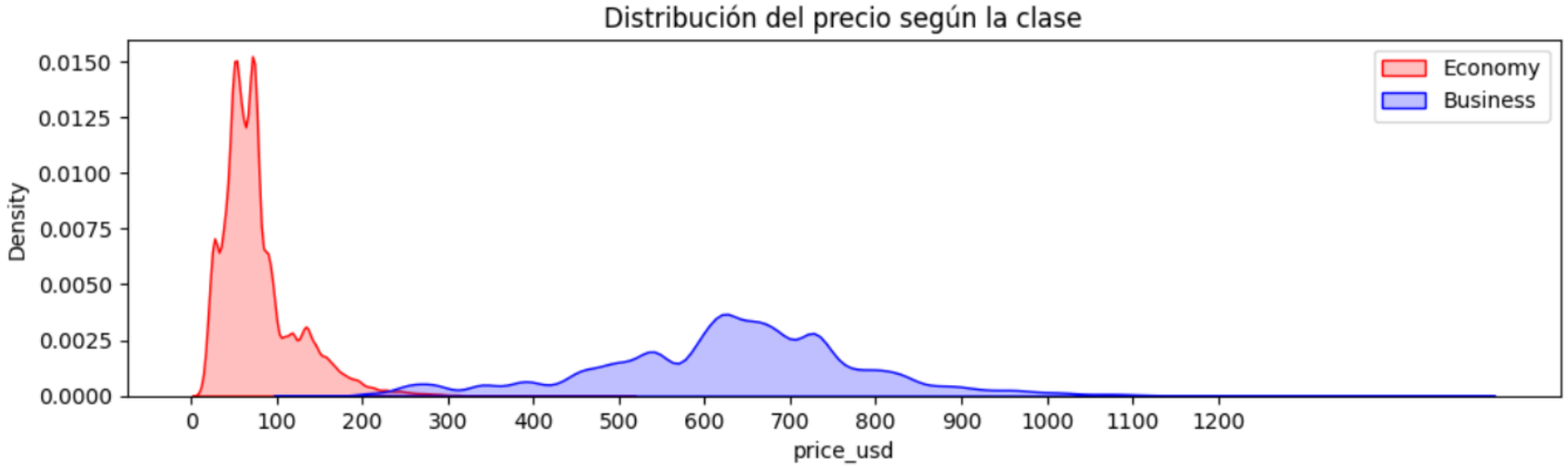
EDA



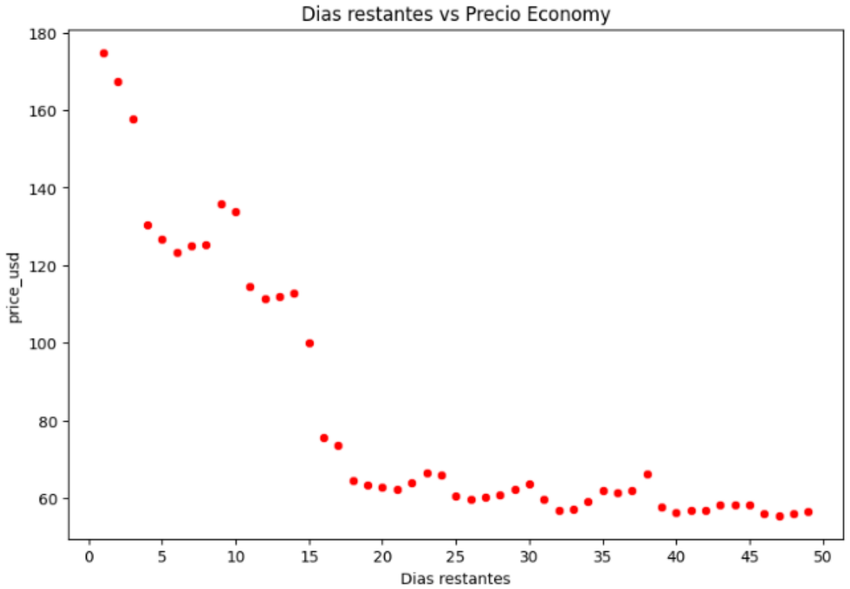
	airline	count	percent	cumulative_count	cumulative_percent
0	Vistara	127859	42.597942	127859	42.597942
1	Air_India	80892	26.950255	208751	69.548197
2	Indigo	43120	14.366007	251871	83.914204
3	GO_FIRST	23173	7.720396	275044	91.634600
4	AirAsia	16098	5.363265	291142	96.997864
5	SpiceJet	9011	3.002136	300153	100.000000

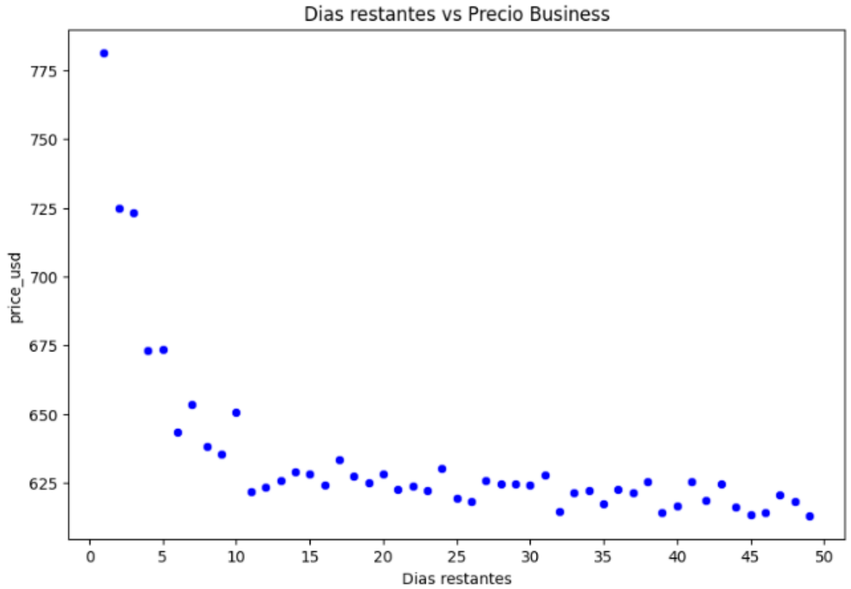
- Las aerolíneas Vistara y AirIndia concentran casi el 70% de los vuelos. SpiceJet posee la menor cantidad de vuelos.
- Los horarios más populares para partir son Morning, Early morning y Evening
- Los horarios más populares para llegar son Night, Evening y Morning
- La mayoría de los vuelos poseen 1 escala (83%)
- El 68% de los vuelos se hicieron en clase Economy
- Delhi y Mumbai son las ciudades con mayor cantidad de vuelos

EDA



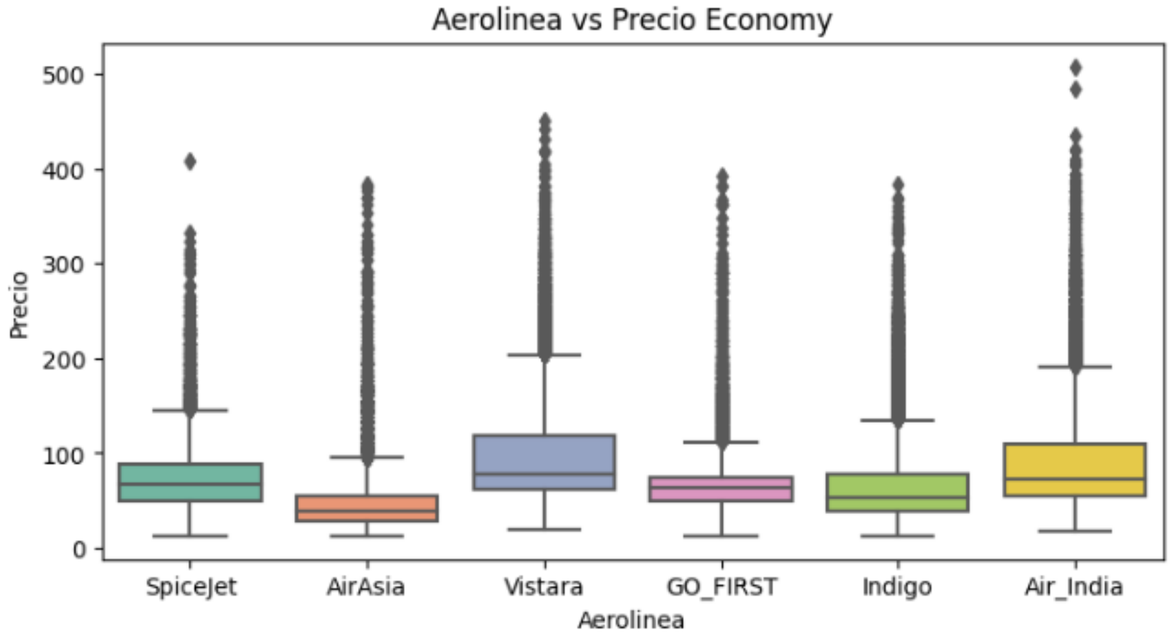
Si calculamos la media del precio agrupando por cantidad de días restantes al vuelo, para clase **Economy** se observa un **precio mayor** claramente definido en el rango de los **últimos 15 días** antes de partir. Mientras que para **Business** el rango es en los **últimos 5 días**.

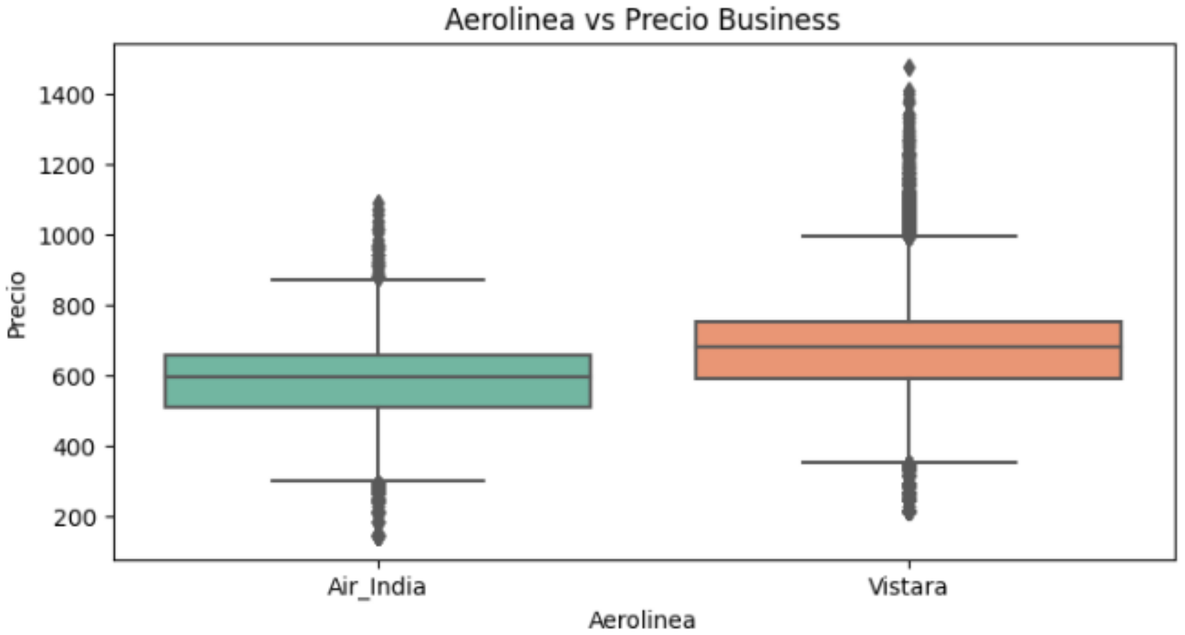




EDA

Vistara y Air India presentan los **mayores precios**, con un máximo de Air India cerca de los 500 usd en Economy, mientras que AirAsia tiene los menores precios. En cuanto a clase Business, Vistara tiene mayores precios que Air India. Se presentan outliers.





EDA

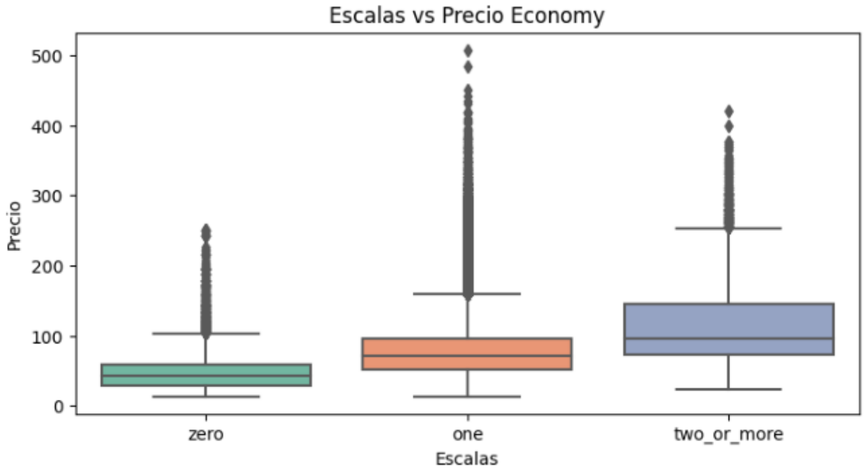
destination_city source_city	Bangalore	Chennai	Delhi	Hyderabad	Kolkata	Mumbai
Bangalore	0	6410	13756	8928	10028	12939
Chennai	6493	0	9783	6103	6983	9338
Delhi	14012	10780	0	9328	11934	15289
Hyderabad	7854	6395	8506	0	7987	10064
Kolkata	9824	6653	10506	7897	0	11467
Mumbai	12885	10130	14809	10470	12602	0

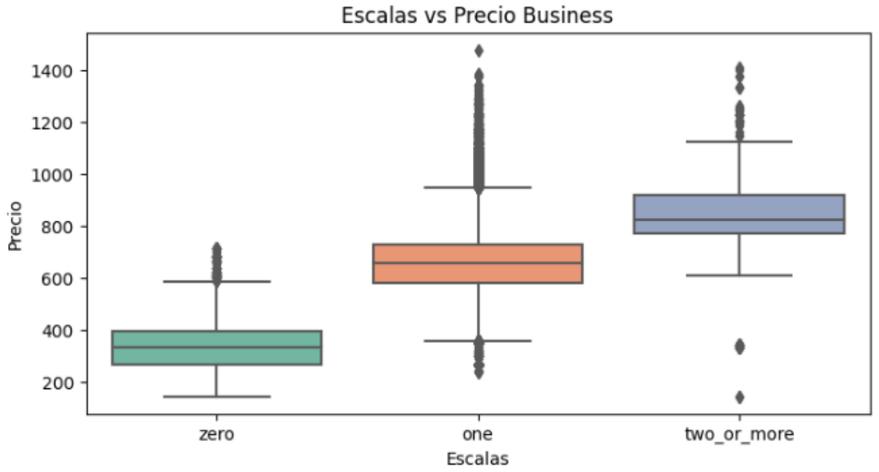
Analizando la cantidad de vuelos, vemos que:
La combinación más frecuente es **Delhi** -> **Mumbai**
La menos frecuente es **Chennai** -> **Hyderabad**

Del análisis de precios, se observa:

Ciudad de destino	Ciudad de origen con menor precio
Delhi	Hyderabad
Mumbai	Delhi
Bangalore	Delhi
Kolkata	Delhi
Hyderabad	Delhi
Chennai	Hyderabad

EDA



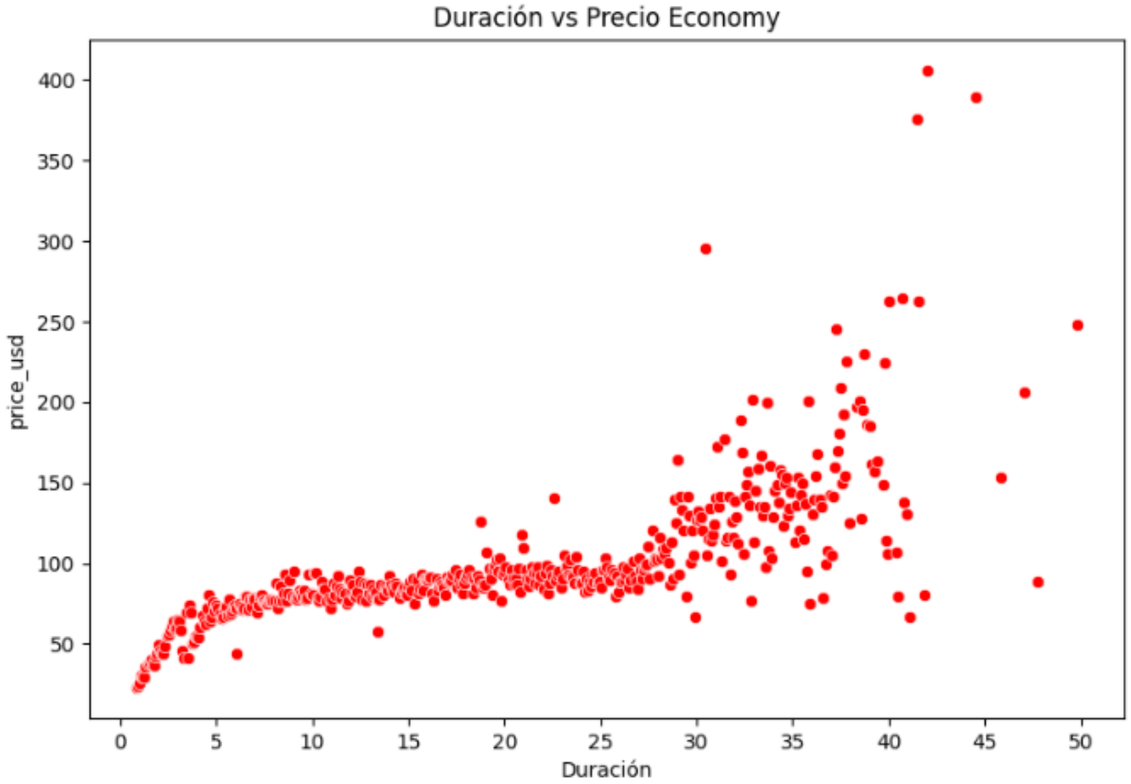


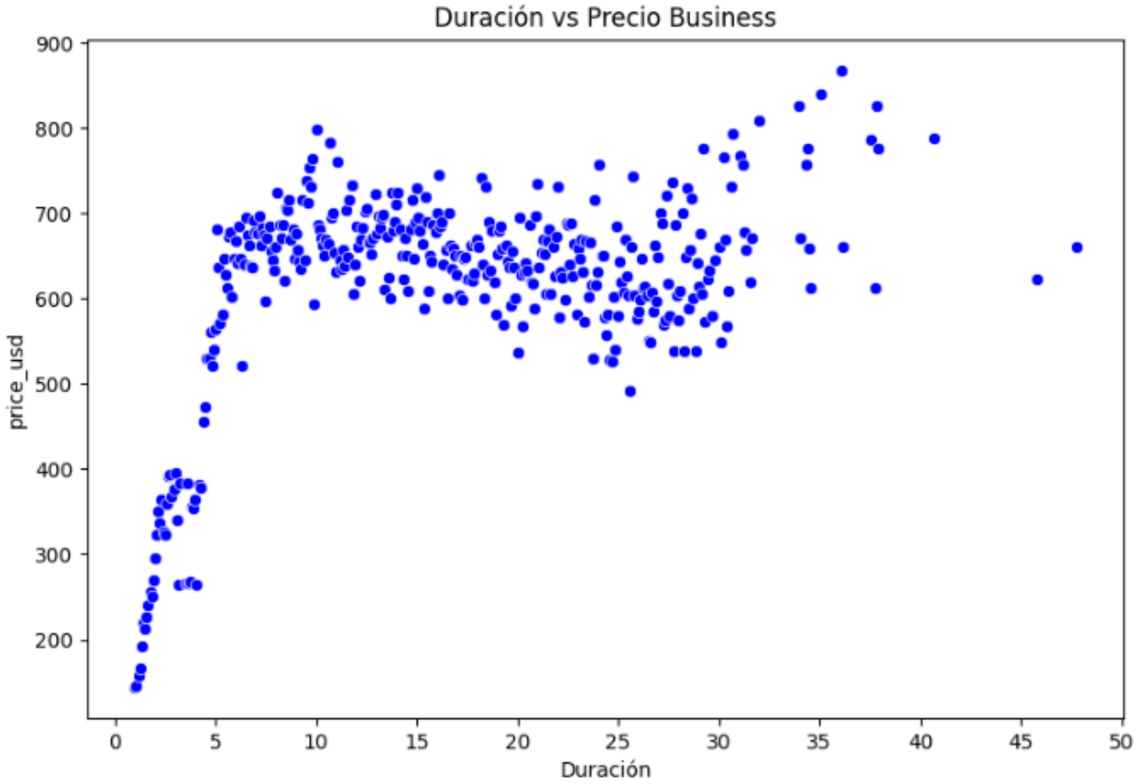
Si diferenciamos por cantidad de escalas, se observan **menores precios** en **vuelos directos**, mientras que los vuelos de 1 escala presentan el precio máximo.

El horario de salida más elegido por los usuarios varía con la ciudad. Por ejemplo, en el caso de Bangalore es Evening, mientras que para Chennai es Morning.

departure_time source_city	Afternoon	Early_Morning	Evening	Late_Night	Morning	Night
Bangalore	5183	13611	14243	457	12323	6244
Chennai	5807	9319	5402	72	10550	7550
Delhi	11234	12248	16790	357	13679	7035
Hyderabad	7221	8524	5991	38	9923	9109
Kolkata	7863	8133	9594	114	12065	8578
Mumbai	10486	14955	13082	268	12606	9499

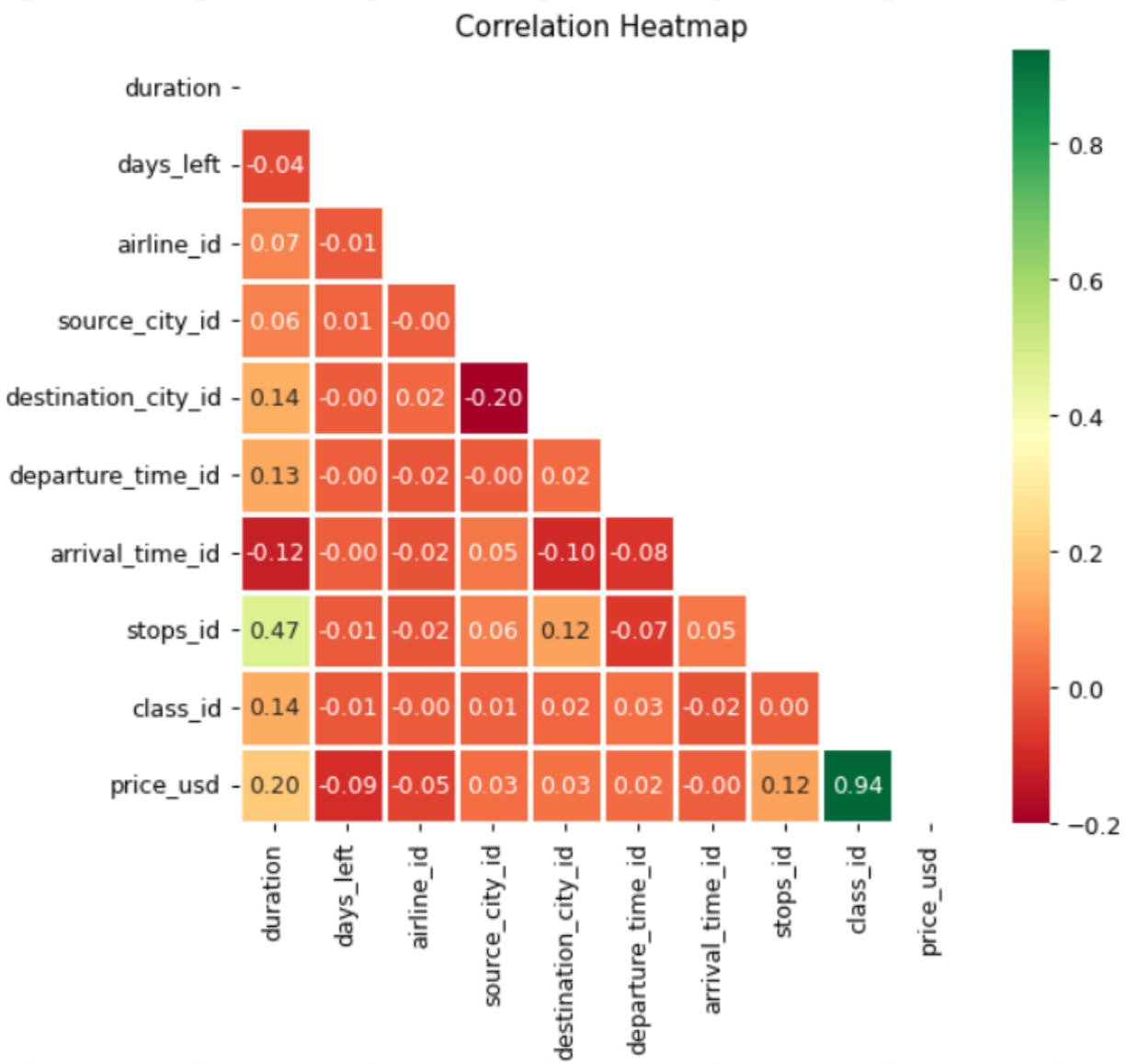
EDA





Si consideramos la media del precio agrupando por duración del vuelo, para clase **Economy** se observa un **aumento del precio a partir de las 25 hs**. En el caso de **Business**, el precio va **aumentando hasta las 10 hs**, a partir del cual se mantiene estable.

EDA



Del análisis de correlación entre variables, se puede observar una **alta correlación** entre el **precio** y la **clase** (Economy o Business). Una menor correlación encontramos entre el **precio** y la **duración del vuelo**.

Por otro lado, hay una alta correlación entre la duración y la cantidad de escalas.

MODELOS PREDICTIVOS

- LINEAR REGRESSION (OLS)
- LASSO
- RIDGE
- DUMMY
- DECISION TREE
- EXTREME GRADIENT BOOSTING
- K NEAREST NEIGHBORS
- RANDOM FOREST

MODELOS PREDICTIVOS

¿CÓMO EVALUAREMOS LOS MODELOS?

R2 Devuelve el porcentaje de la variación observada en el precio que se explica por las variables independientes. Aumenta a medida que se agregan más variables independientes en los modelos.

MAE Se calcula como la media de la diferencia absoluta entre los valores reales y predichos.

RMSE Raíz del error cuadrático medio. El objetivo es que sea lo menor posible, ya que mide las diferencias entre los valores reales y los pronosticados, penalizando más las grandes diferencias que las pequeñas.

MAPE Mide el error en términos de porcentaje.

MODELOS PREDICTIVOS

LINEAR REGRESSION (OLS)

Es un modelo sencillo. Asume que existe una relación lineal entre las variables independientes y la variable dependiente (el precio). Encuentra la línea que minimiza la suma de los errores cuadrados entre las predicciones del modelo y los valores reales

LASSO y RIDGE

Son variaciones del modelo de regresión lineal. Intentan evitar el sobreajuste, agregando un término a la función de coste del modelo que penaliza los coeficientes de las variables independientes.

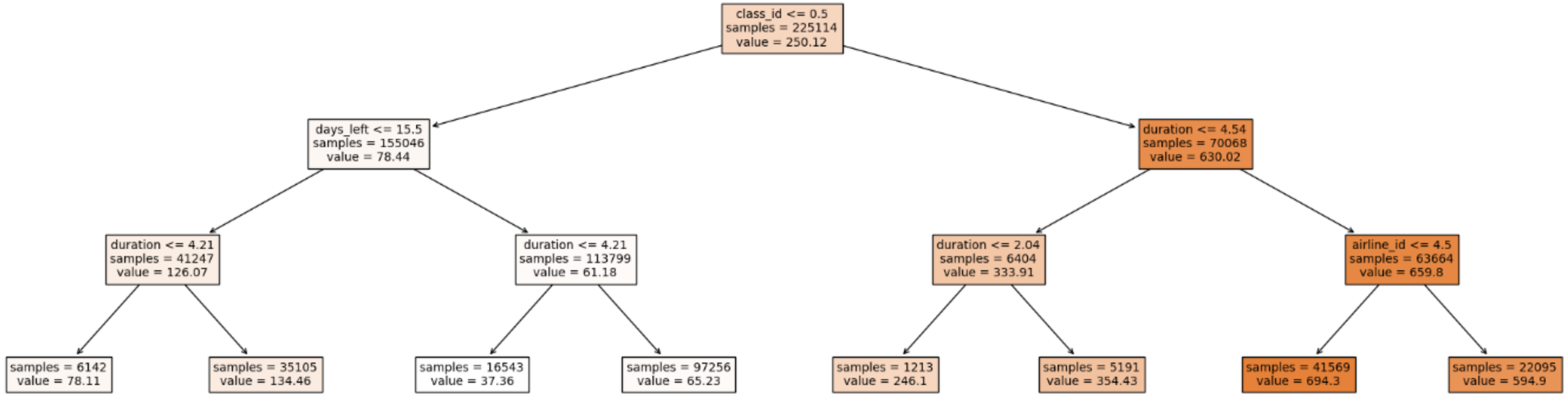
DUMMY

Es el modelo más simple de todos. Calcula el promedio de todos los valores del target y devuelve este como la predicción. Es por esto que es esperable una pésima performance del modelo. Lo utilizamos para comparar las métricas con otros modelos.

MODELOS PREDICTIVOS

DECISION TREE

Este modelo crea una estructura en forma de árbol que consta de nodos que representan una pregunta o una decisión y ramas que representan las posibles respuestas.



MODELO 1 (simple): Profundidad del árbol = 3

Si se cumple que: 'class_id' <= 0.5 & 'days_left' <= 15.5 & 'duration' <= 4.21 entonces el modelo predice un valor del boleto de $78.11

MODELOS PREDICTIVOS

DECISION TREE

MODELO 2

Mediante una técnica de validación cruzada llamada GridSearchCV se obtienen los hiperparámetros ideales (es decir, los que minimizan el error) para este tipo de modelo.

Los hiperparámetros son Profundidad = 50, Features = 10
Obtenemos:

R2 0,97
RMSE 42,67
MAE 14,21
MAPE 0,08

Comparando con los modelos anteriormente mencionados, este árbol de décision es el más eficiente.

MODELOS PREDICTIVOS

XGBOOST

Extreme Gradient Boosting genera múltiples modelos de predicción “pobres” iterativamente, donde cada uno de estos toma los resultados del modelo anterior, para generar un modelo con mejor poder predictivo. Los tiempos de procesamiento son elevados.

K NEAREST NEIGHBORS

Este algoritmo toma la media de los valores de los K vecinos más cercanos para predecir el valor del target. A medida que aumenta K, aumentan los tiempos.

RANDOM FOREST

El modelo crea múltiples árboles de decisión, cada uno a partir de una muestra aleatoria de los datos. La predicción la obtiene promediando las salidas de cada árbol individual.

CONCLUSIÓN

Se obtuvieron las siguientes métricas para los distintos modelos:

	Modelo	R2	RMSE	MAE	MAPE
0	Regresión lineal	0.903439	84.640636	55.197850	0.445568
1	Lasso	0.901692	85.402639	54.291457	0.404124
2	Ridge	0.903439	84.640636	55.197856	0.445568
3	Dummy Train	0.000000	272.343361	237.035521	2.405541
4	Dummy Test	-0.000001	272.495626	237.291385	2.409828
5	Decision Tree 1 Train	0.934869	69.503907	42.080349	0.277122
6	Decision Tree 1 Test	0.934081	69.962389	42.045192	0.276947
7	Decision Tree 2 Train	0.999314	7.132406	0.664081	0.002539
8	Decision Tree 2 Test	0.975475	42.673799	14.209766	0.077466
9	XGBoost Train	0.995746	17.762623	9.154434	0.063484
10	XGBoost Test	0.987955	29.905802	14.503519	0.091073
11	KNN Train	0.839756	109.020524	73.846736	0.565686
12	KNN Test	0.751876	135.735604	93.258551	0.087187
13	Random Forest Train	0.994123	20.877833	9.723267	0.059191
14	Random Forest Test	0.985047	33.321361	14.914099	0.087187

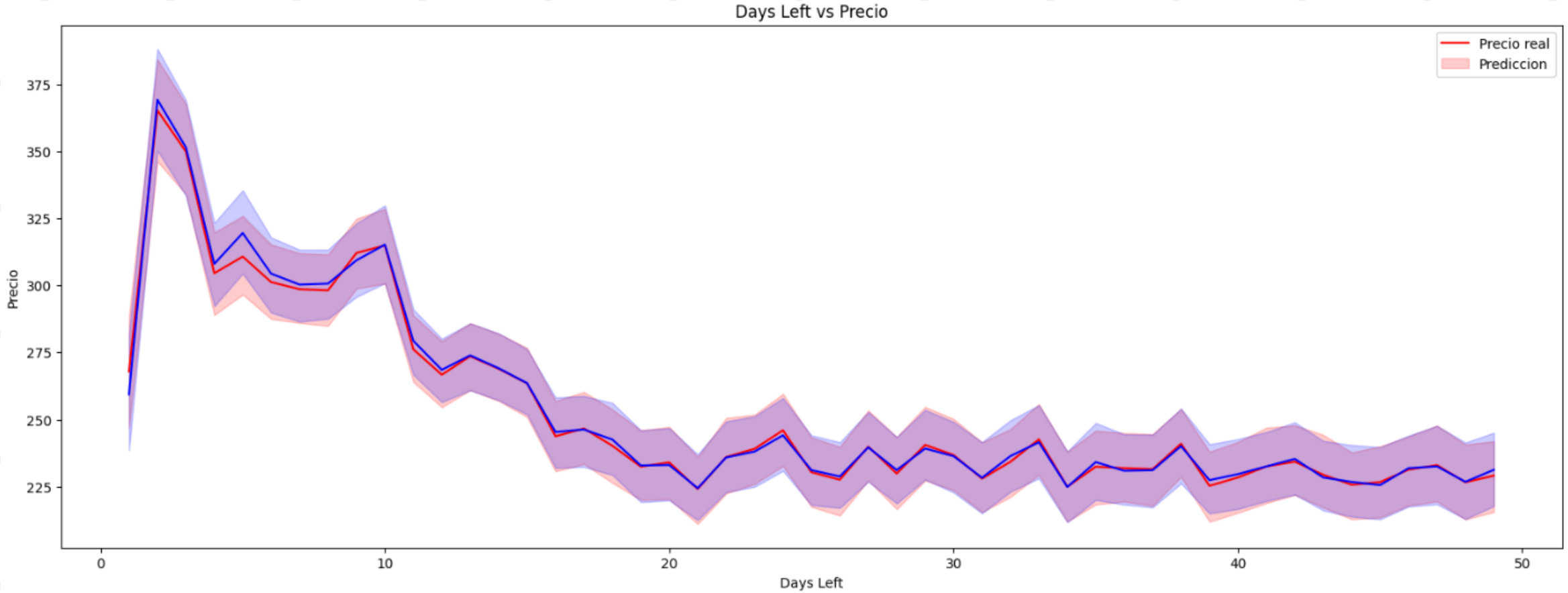
Si consideramos las métricas **R2** y **RMSE**, el modelo **XGBoost es el más eficiente** (se debe considerar los resultados para test)

Si consideramos **MAE** y **MAPE**, el modelo elegido es **Decision Tree 2**

La diferencia de las métricas RMSE y MAE entre los 2 subconjuntos train y test podría traducirse en una alta varianza, lo que implica un problema de overfitting. Sin embargo, un error absoluto medio de 14 dolares parece totalmente aceptable.

CONCLUSIÓN

Reflejamos en un gráfico la capacidad del modelo elegido para predecir valores:



CONCLUSIÓN

Hipótesis 1: VERDADERO
(↑ duración ↑ precio)

Para clase Economy se observa un aumento del precio a partir de las 25 hs de duración del vuelo. En el caso de Business, el precio va aumentando hasta las 10 hs, a partir del cual se mantiene estable.

Hipótesis 2: VERDADERO
(↑ días rest. ↓ precio)

Para clase Economy se observa un precio mayor claramente definido en el rango de los últimos 15 días antes de partir. Mientras que para Business el rango es en los últimos 5 días.

Hipótesis 3: VERDADERO
(↑$ Vistara ↓$ Air Asia)

Vistara y Air India presentan los mayores precios en Economy, mientras que AirAsia tiene los menores precios. En cuanto a clase Business, Vistara tiene mayores precios que Air India.

Hipótesis 4: FALSO

En realidad, estas son las más baratas combinaciones:

Ciudad origen	Ciudad destino
Hyderabad	Delhi
Delhi	Mumbai
Delhi	Bangalore

Ciudad origen	Ciudad destino
Delhi	Kolkata
Delhi	Hyderabad
Hyderabad	Chennai

Hipótesis 5: VERDADERO
(↓$ cero escalas)

Se observan menores precios en vuelos directos, mientras que los vuelos de 1 escala presentan el precio máximo.

Hipótesis 6: FALSO
(Early morning + elegido)

El horario de salida más elegido varía con la ciudad de origen. Por ejemplo, en el caso de Bangalore es Evening, mientras que para Chennai es Morning. El único caso donde Early Morning es el horario más elegido es desde Mumbai.

GRACIAS!
:)