TRABAJO FINAL

DATA SCIENCE

MATEO KLOSZ

CODERHOUSE


5- Algoritmos
4- Exploración
1- Presentación
3- Objetivos
2- Situación
6- Conclusiones

ÍNDICE

1- PRESENTACIÓN

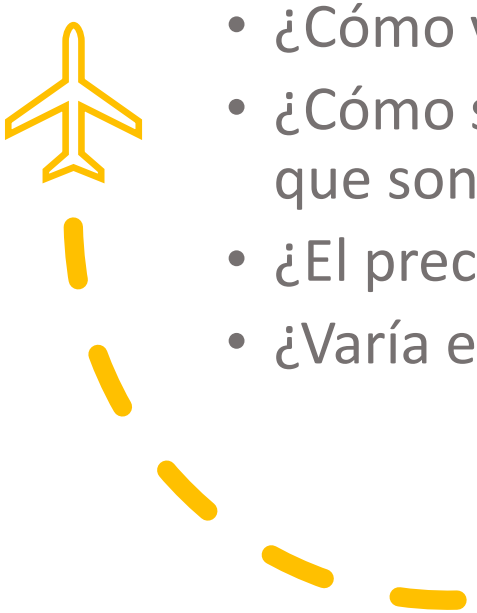
- Se obtuvo del sitio web Kaggle el dataset de reservas de vuelos obtenidos del sitio *Ease My Trip*.
- Se analizarán los datos aplicando diferentes métodos de exploración de datos y pruebas estadísticas.
- Con la información procesada se utilizarán algoritmos de regresión para predecir el precio de los vuelos con la mayor precisión posible.
- Se pretende ayudar a las personas a elegir el pasaje de avión que más le convenga entendiendo la dinámica del mercado y anticipando el precio de compra.

2- SITUACIÓN

- El dataset contiene 300.261 reservas de vuelos y contiene 11 variables.
- Los datos fueron recopilados entre el 11 de Febrero al 31 de Marzo de 2022 y se utilizó la herramienta Octoparse para el scrapping.
- El dataset no contaba con datos faltantes y 8 de las 11 variables son categóricas.
- Se relevaron datos de 6 aerolíneas y de 6 ciudades (Origen y Destino)
- Los horarios de arribo y llegada de dividieron en 6 bloques de tiempo.
- Se definió *Business* y *Economy* para los tipos de clase de vuelos.

3- OBJETIVOS

- Poder determinar con algoritmos de regresión una predicción confiable de los precios de los vuelos.
- Responder a las siguiente preguntas:
  - ¿Varía el precio según la aerolínea?
  - ¿Cómo varía el precio del boleto entre las clases Economy y Business?
  - ¿Cómo se relaciona el precio de los boletos con el tiempo de anticipación en que son comprados?
  - ¿El precio del boleto cambia según la cantidad de paradas que tenga el vuelo?
  - ¿Varía el precio del boleto en función de la hora de salida y la hora de llegada?



4- EXPLORACIÓN

- El precio mínimo para un pasaje es 1.105 Rp y el máximo 123.071 Rp
- La ciudad más frecuente como origen es *Delhi* y como destino *Mumbai.*
- *Vistara* es la aerolínea más popular con el 43% de las reservas a pesar de ser la más cara en las dos clases (Min. *Business* 17.604 Rp y Min. *Economy* 1714 Rp).
- Comprando con 20 días de anticipación se obtiene un precio bajo, después comienzan a subir los precios a excepción de el día del vuelo que los precios vuelven a bajar.
- La cantidad de paradas de un viaje condiciona al precio del pasaje en forma directa, aunque tiene menos influencia en la compañía *AirAsia.*
- Los vuelos que salen o llegan en la madrugada son más económicos.

5- ALGORITMOS

- Se evaluaron 4 algoritmos distintos:
  - KNeighborsRegressor
  - LinearRegression
  - XGBRegressor
  - RandomForestRegressor
- Las métricas utilizadas para la evaluación fueron RMSE (Root Mean Square Error) y R^2 (Coeficiente de determinación)*.
- RandomForestRegressor fue el de mejor rendimiento con un RMSE=33 y R2= 98.5%
- Se buscó optimizar el rendimiento realizando una selección de características relevantes para XGBR y RFR obteniendo que: *class', 'duration', 'days_left, 'airline_Air_India', 'source_city_Delhi'* permiten predecir con mucha aproximación el precio vs utilizar todas las características.

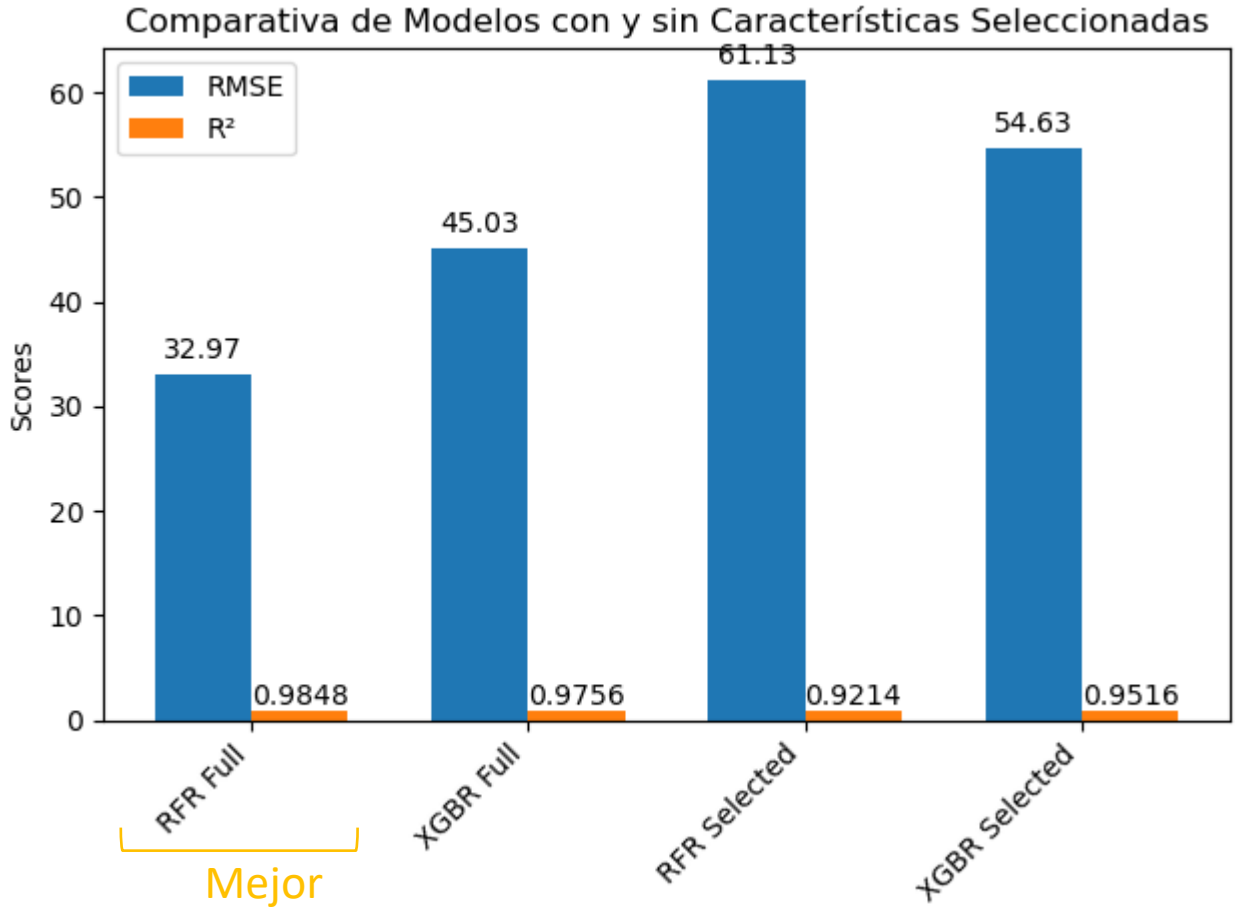
* RMSE se centra en medir el tamaño de los errores de predicción, mientras que R^2 se enfoca en la capacidad del modelo para explicar la variabilidad en los datos.

6- CONCLUSIONES

- Los vuelos con salida durante la madrugada, tarde y noche son más económicos.
- Los vuelos desde y hacia Delhi son los más económicos.
- Los vuelos son más caros con más escalas.
- *Vistara* más caro del mercado para (*Economy* y *Business*) y la más popular con el 43% de las reservas.
- Los precios suben gradualmente hasta 20 días antes; un día antes, bajan significativamente.
- RandomForestRegressor logra un RSME = 33 y un R^2 = 98,5 con un gran nivel de

6- CONCLUSIONES

- Con la reducción de *features* la performance continúa siendo muy buena y con un alto grado de predicción del precio.
- Características clave: *class, duration, days_left, airline_Air_India, source_city_Delhi*.



¡ GRACIAS !