- YEAR, MONTH, DAY, DAY_OF_WEEK: dates of the flight
- AIRLINE: An identification number assigned by US DOT to identify a unique airline
- ORIGIN_AIRPORT and DESTINATION_AIRPORT: code attributed by IATA to identify the airports
- SCHEDULED_DEPARTURE and SCHEDULED_ARRIVAL : scheduled times of take-off and landing
- DEPARTURE_TIME and ARRIVAL_TIME: real times at which take-off and landing took place
- DEPARTURE_DELAY and ARRIVAL_DELAY: difference (in minutes) between planned and real times
- DISTANCE: distance (in miles)

Preguntas sobre el modelo:

• ¿Cuál es el problema de regresión que están tratando de resolver?

El objetivo es usar el mejor modelo de regresión para predecir el tiempo de retraso en la salida de los vuelos

• ¿Qué columna de los datos de retrasos de aerolíneas especificaron como la columna objetivo?

La columna de datos que es la columna objetivo es “`DepDelay`”

• ¿Qué modelo de regresión seleccionaron como el mejor y por qué?

Es el mejor debido a que da los mejores resultados en las pruebas de MSE RMSE y R2

```
MejorModelo = reg.compare_models()
```

38m 1.8s — Python

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	11.9746	819.3498	28.6148	0.0140	1.3491	1.6874	4.6100
gbr	Gradient Boosting Regressor	11.9973	824.8562	28.7109	0.0073	1.3463	1.6652	68.6810
ridge	Ridge Regression	12.1525	835.2987	28.8923	-0.0052	1.3387	1.6841	2.3650
lar	Least Angle Regression	12.1525	835.2987	28.8923	-0.0052	1.3387	1.6841	2.3710
br	Bayesian Ridge	12.1523	835.3035	28.8923	-0.0052	1.3387	1.6840	2.4800
lr	Linear Regression	12.1525	835.2987	28.8923	-0.0052	1.3387	1.6841	2.5010
rf	Random Forest Regressor	12.8700	835.4719	28.8956	-0.0055	1.3860	2.1601	51.0080
omp	Orthogonal Matching Pursuit	12.1719	841.4083	28.9978	-0.0126	1.3210	1.6431	2.3610
en	Elastic Net	12.1033	842.3483	29.0140	-0.0137	1.2841	1.5939	2.3950
et	Extra Trees Regressor	12.9433	843.6247	29.0362	-0.0153	1.3909	2.1785	29.5670
lasso	Lasso Regression	12.1447	845.6097	29.0702	-0.0177	1.2461	1.5915	2.4140
llar	Lasso Least Angle Regression	12.1447	845.6098	29.0702	-0.0177	1.2461	1.5915	2.3620
dummy	Dummy Regressor	12.3197	854.5593	29.2238	-0.0285	1.2355	1.6174	2.1380
ada	AdaBoost Regressor	11.9191	855.7389	29.2430	-0.0299	1.3587	1.4268	28.6350
huber	Huber Regressor	11.3924	879.6739	29.6506	-0.0588	1.5084	1.0483	3.1020
knn	K Neighbors Regressor	13.6892	892.9932	29.8748	-0.0749	1.4587	2.3565	9.4800
par	Passive Aggressive Regressor	14.1287	906.9612	30.1092	-0.0925	1.4585	2.4206	3.2660
dt	Decision Tree Regressor	19.6630	1854.3654	43.0540	-1.2341	1.6852	4.3887	7.3420

• ¿Cómo evaluaron el rendimiento del modelo en el conjunto de pruebas? ¿Cuáles fueron las métricas de rendimiento que utilizaron para evaluar el modelo? ¿Cuáles fueron los resultados?

Primero modificamos el rendimiento del modelo en base a los datos obtenidos, tras esto introducimos las columnas que usaremos para trabajar los datos que son las siguientes:

```python
from sklearn.preprocessing import LabelEncoder

dataFrame_encoded = datos.copy()

# Create a label encoder object
label_encoder = LabelEncoder()

# List of columns to encode
columns_to_encode = ['UniqueCarrier', 'Origin', 'Dest']

# Apply the label encoder to each column
for column in columns_to_encode:
    dataFrame_encoded[column] = label_encoder.fit_transform(dataFrame_encoded[column])

X = dataFrame_encoded.drop(labels=["DepDelay"],axis=1)
Y = dataFrame_encoded["DepDelay"]

X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.7,random_state=101)
```

Una vez hecho esto vemos los resultados obtenidos usando un dataframe

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	11.9939	826.8066	28.7542	0.0085	1.3528	1.6639

• ¿Cómo visualizaron los resultados del modelo? ¿Qué información pueden extraer de la visualización?

Usamos nuestro modelo tuneado para trabajar el dataframe de los datos.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	11.9939	826.8066	28.7542	0.0085	1.3528	1.6639

```
[LightGBM] [Warning] feature_fraction is set=0.5, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=0.5
[LightGBM] [Warning] bagging_fraction is set=0.4, subsample=1.0 will be ignored. Current value: bagging_fraction=0.4
[LightGBM] [Warning] bagging_freq is set=0, subsample_freq=0 will be ignored. Current value: bagging_freq=0
```

	Month	DayofMonth	DayOfWeek	CRSDepTime	CRSArrTime	UniqueCarrier	Origin	Dest	Distance	DepDelay	prediction_label	Abs Error
0	10	11	7	1300.0	1535.0	1	193	154	2556.0	8.0	2.290135	5.709865
1	10	10	6	2035.0	2110.0	1	249	154	100.0	-3.0	-0.570523	2.429477
2	10	26	1	1200.0	1446.0	1	183	192	2475.0	6.0	2.553774	3.446226
3	10	9	5	1145.0	1512.0	1	183	307	2586.0	1.0	1.807758	0.807758
4	10	16	5	930.0	1149.0	1	309	154	2399.0	0.0	1.522772	1.522772
...	...	...	...	...	...	...	...	...	...	...	...	...
999995	7	30	2	835.0	940.0	27	327	223	317.0	-1.0	0.982156	1.982156
999996	7	29	1	1225.0	1633.0	29	266	297	843.0	1.0	4.373349	3.373349
999997	7	30	2	1515.0	1735.0	29	35	72	350.0	-2.0	7.226538	9.226538
999998	7	25	4	1335.0	1646.0	29	254	278	900.0	8.0	7.457405	0.542595
999999	7	31	3	530.0	645.0	29	14	251	723.0	-7.0	-0.626736	6.373264

999983 rows × 12 columns

Lo visualizamos a través de una tabla de la que podemos ver los resultados obtenidos a través de las columnas del final

Distance	DepDelay	prediction_label	Abs Error
2556.0	8.0	2.290135	5.709865
100.0	-3.0	-0.570523	2.429477
2475.0	6.0	2.553774	3.446226
2586.0	1.0	1.807758	0.807758
2399.0	0.0	1.522772	1.522772
...	...	...	...
317.0	-1.0	0.982156	1.982156
843.0	1.0	4.373349	3.373349
350.0	-2.0	7.226538	9.226538
900.0	8.0	7.457405	0.542595
723.0	-7.0	-0.626736	6.373264

• ¿Cuáles fueron los hiper parámetros que utilizaron para mejorar el rendimiento del modelo?

Los hiper parámetros usados son los de la columna “DepDelay” menos la columna “prediction_label”

• ¿Cuáles fueron las características más importantes del modelo?

Las características más importantes es que gracias a este modelo, tanto los clientes como las aerolíneas pueden ganar tiempo y dinero debido a que pueden saber si un vuelo tiene probabilidades de que se retrase y por tanto arreglar el retraso. Para ello, se debe conseguir predecir correctamente los datos que se le piden, para ello debemos de trabajar antes correctamente los datos y los tipos de datos.
Por ello, también se debe de tener en cuenta el error absoluto del modelo, para seguir trabajando con él para mejorarlo lo más posible.

• ¿Cómo desplegaron el modelo y cómo podría ser utilizado en una aplicación real?

Este modelo puede ser usado en una aplicación real para las propias aerolíneas puesto que esto puede ayudar tanto a los clientes como a las aerolíneas a la correcta organización de los vuelos y de los pasajeros.