Autor: **Martin Celi**

Demoras en Aerolíneas

Tabla de Contenidos

01. INTRODUCCIÓN

Para un pasajero no debe haber mayor alegría que llegar y más aún, a tiempo al destino que eligió para sus vacaciones, visitar a un familiar o amigo, o para realizar un viaje de negocio. Esa sensación de que todo salió bien, genera una felicidad inmensa, y esto se convierte en un cliente que volverá a elegir a la compañía para volver a volar al destino que quiera.

Pero, el secreto para llegar a tiempo, es **salir a tiempo**. Para lograr esto existen muchos factores internos como la gestión de la tripulación y el mantenimiento de aeronaves, o factores externos como tormentas inesperadas y congestión del tráfico aéreo que afectan la salida de un vuelo según su itinerario. Por lo cual, es crucial desarrollar un itinerario que sea rentable y que logre la satisfacción de los clientes.

02. CONTEXTO Y AUDIENCIA

Contexto

Identificar los factores que influyen en la **satisfacción del cliente** es de suma importancia en un mercado sumamente competitivo. Dentro las métricas de satisfacción del clientes, el cumplimiento del **horario de llegada** al destino es **crucial**, por lo cual, es de suma importancia no solo desarrollar **itinerarios eficientes** en materia económica y operativa, sino también, en *experiencia al cliente*. El estudio de los datos para **identificar patrones** y **predecir posibles demoras** en el desarrollo de la operación no solo puede mejorar la experiencia sino optimizar los recursos y enfocar los esfuerzos donde

Audiencia

Esta presentación está dirigida a **directores** y **gerentes** de **áreas comerciales** y **operativas** de aerolíneas, así como a cualquier persona interesada en comprender el impacto del análisis de datos en la eficiencia y la satisfacción del cliente en la industria aérea. *Los resultados de este estudio proporcionará información valiosa para la toma de decisiones estratégicas, la optimización de recursos y la mejora continua de la experiencia del pasajero.*

03. HIPÓTESIS

En una operación aérea existen muchos protagonistas que influyen en la salida a tiempo de un vuelo. Analizando arduamente la estructura de una operación podemos identificar tres principales actores. Para cada uno de estos, vamos a tratar de identificar qué influencia tiene en la salida a tiempo de un vuelo.

Operador

Con operador nos referimos a la aerolínea que realiza el vuelo, esta no solo diseña el itinerario, sino que es responsable en la ejecución del mismo.

Aeropuerto

El lugar donde parte el vuelo es crucial, ya que no solo es el punto de partida, sino que la infraestructura del mismo determina la atención que se puede brindar a la operación de diferentes compañías.

Climatología

Este factor aunque, no es controlable, si es posible determinar ciertos patrones en lugares y estaciones que permiten anticipar posibles que pueden llegar a ocurrir.

Ante este análisis de factores el estudio se basará en entender si:

- ¿Es posible predecir si un vuelo se demora?
- ¿Existen factores internos y/o externos que influyen en la predicción?
- ¿Existe estacionalidad en las demoras y sus posibles predicciones?
- ¿Qué factores internos son los más preponderantes a influir en la demora?

OBJETIVO → *PREDECIR SI UN VUELO PROGRAMADO SUFRIRÁ UNA DEMORA MAYOR A 15 MIN.*

04. EXPLORATORIA DE DATOS

Para el estudio se tomó la información de los vuelos operados en el año **2019** en los **Estados Unidos de América.** De esta fuente se obtiene:

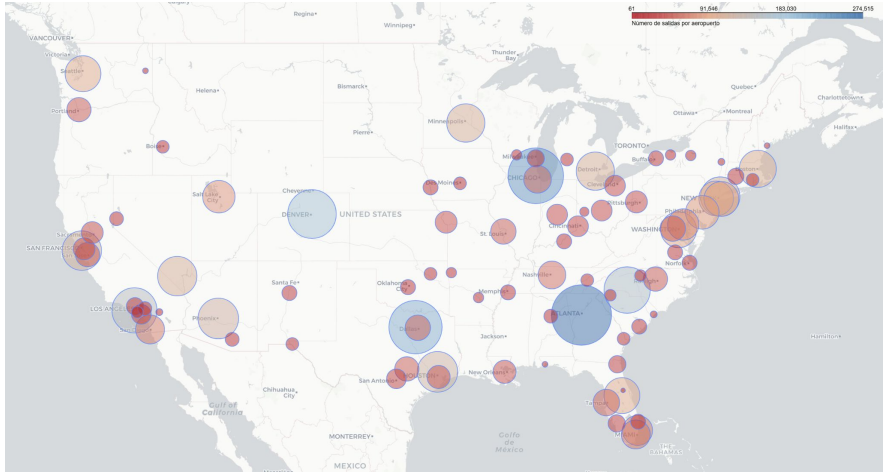
+4,5M de registros

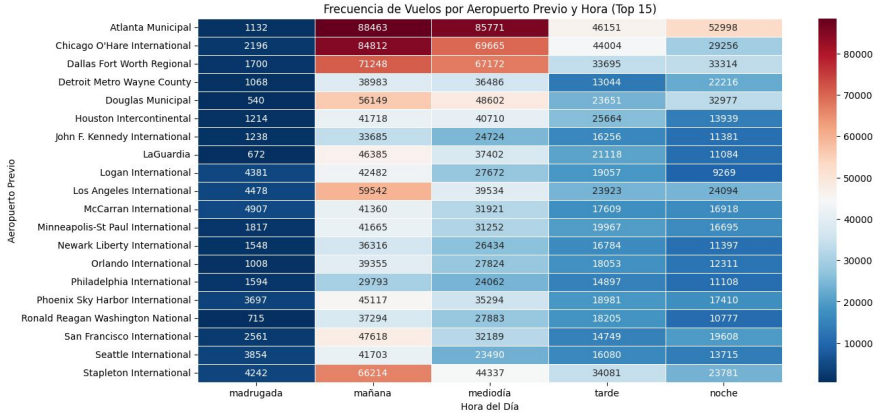
A lo largo del año 2019 se operaron más de **4,5 millones de vuelos** en los Estados Unidos.

26 variables

Es el número total de **variables registradas para cada vuelo**, teniendo datos como el operador, el aeropuerto de salida y destino, el horario, y factores climatológicos.

AEROPUERTOS CON MAYOR TRÁNSITO





Con el gráfico podemos observar que el aeropuerto con mayor tránsito de vuelos es el de **Atlanta**.

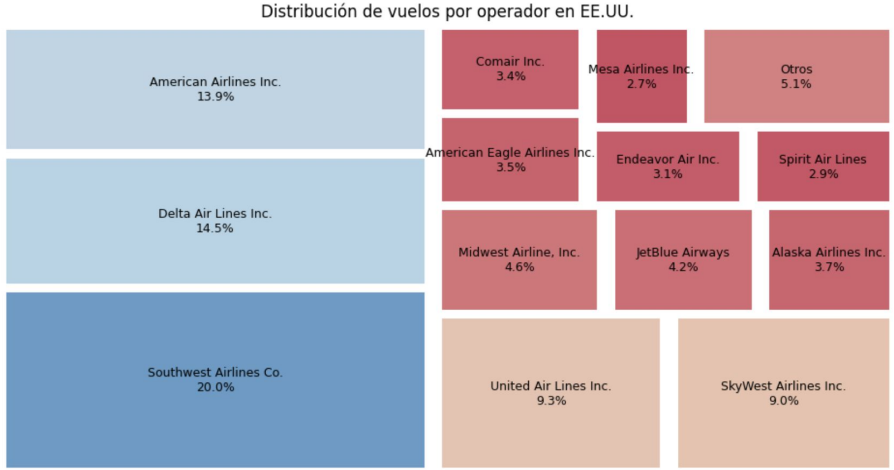
Además, el **este** se tiene *mayor tránsito* que en el **oeste**, y prácticamente en el centro del país no hay tráfico aéreo.

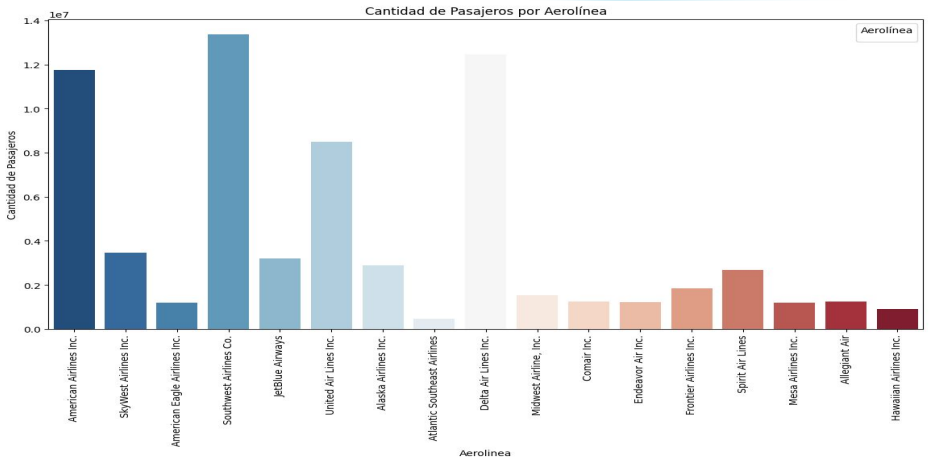
Respecto al **oeste** se puede observar una mayor concentración en dos puntos siendo esto los aeropuertos de *Los Ángeles* y de *San Francisco.*

Por otro lado, el **este** posee mayor cantidad de aeropuertos distribuidos a lo largo de la costa atlántica, pero los aeropuertos con mayor flujo se encuentran más oeste de la costa.

Respecto a la cantidad de vuelos operados en franjas horarias, la mayoría de los vuelos salen durante la **mañana y el mediodía**, mientras que durante la **noche y la madrugada** la actividad se reduce de manera significativa. (En el gráfico se muestran los 20 aeropuertos con mayor cantidad de salidas).

MERCADO AEROCOMERCIAL DE EE.UU

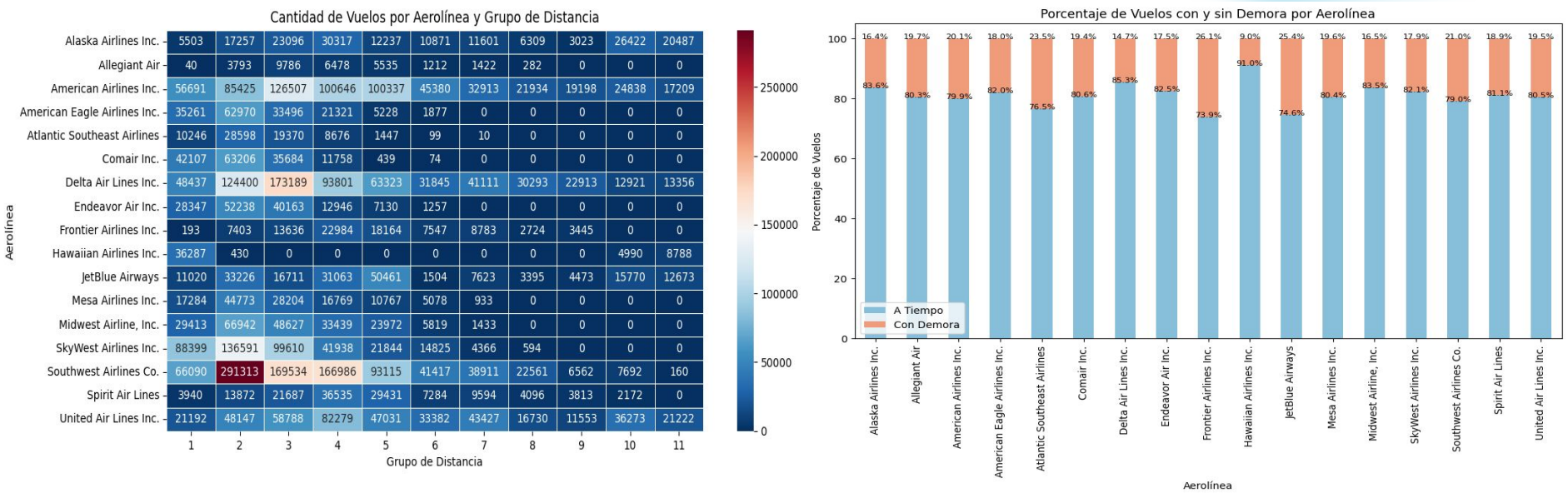




Se puede observar con el primer gráfico de la izquierda que más del **60%** de la operación se encuentra concentrada en 5 aerolíneas principalmente, siendo que 3 aerolíneas controlan casi el **50%** del mercado aerocomercial.

A su vez, si comparamos la cantidad de pasajeros transportados nos encontramos con para las primeras 4 principales líneas aéreas los valores condicen con lo mostrado anteriormente, pero en el caso de Skywest donde es la quinta aerolínea con mayor cantidad de vuelos registrados, al cantidad de pasajeros es similar la de aerolíneas como JetBlue, Alaska Airlines o Spirit Airlines.

MERCADO AEROCOMERCIAL DE EE.UU

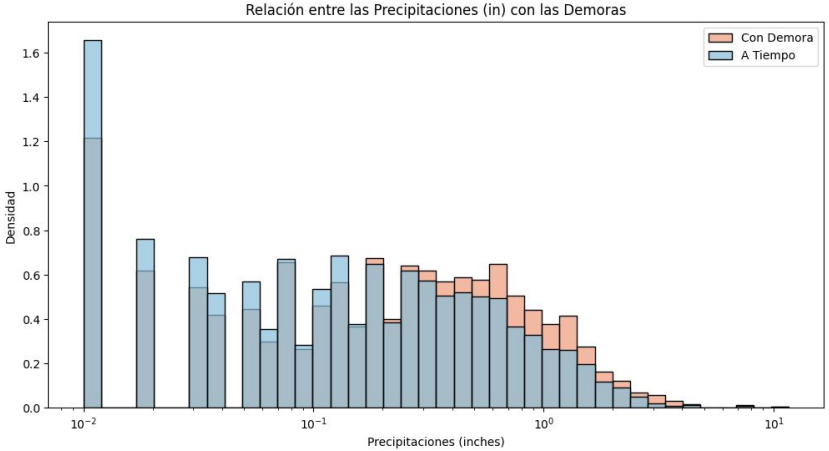


Con respecto a la operación se puede observar que la mayor concentración se encuentra en los vuelos de corta y mediana distancia. Siendo siete aerolíneas que ofrecen vuelos en todos los grupos de distancia. El caso de Hawaiin Airlines es particular y se puede explicar que los grupos 1 y 2 pertenecen a vuelos dentro de las islas Hawaianas, mientras que los grupos 10 y 11 corresponden a los vuelos con destino a Estados Unidos.

La puntualidad posee una media del 80%. Siendo la línea aérea **más puntual** es **Hawaiian Airlines**, mientras que la línea aérea con la **menor puntualidad** es **Frontier Airlines**, seguida por **JetBlue Airways**.

En relación con las líneas aéreas con mayor participación en el mercado ofrecen una *media del 81% de puntualidad*, siendo **Delta Airlines**, mientras que **Southwest** es la aerolínea es la que menor puntualidad.

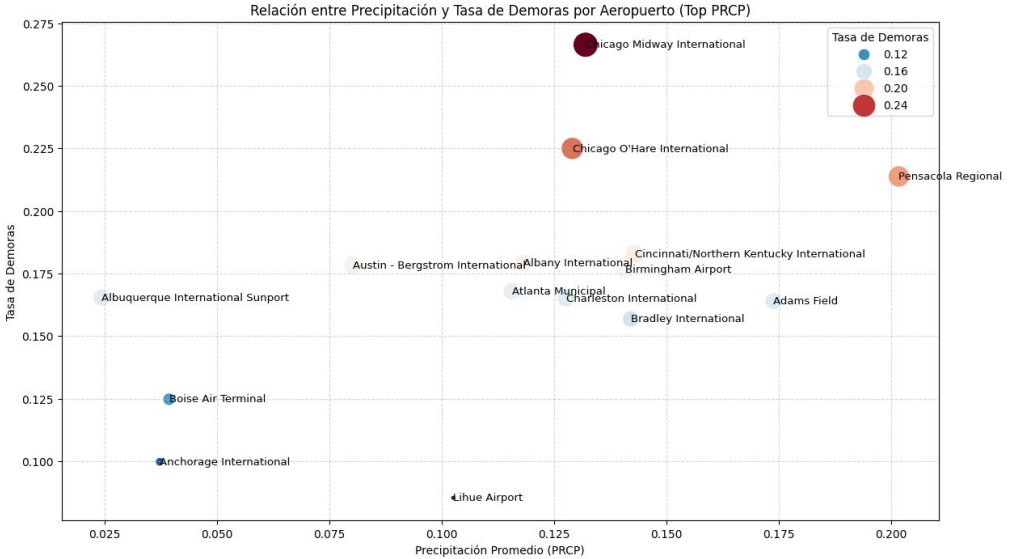
INFLUENCIA DEL CLIMA



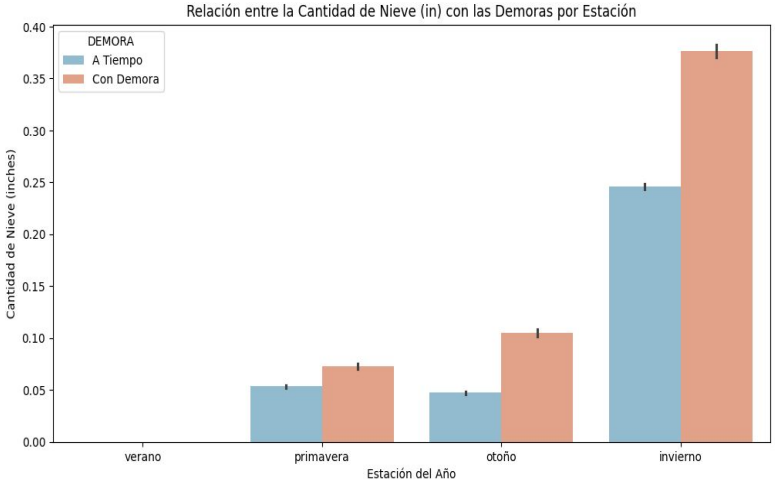
Analizando la información, se puede observar que el clima tiene una influencia en las demoras.

Respecto a las precipitaciones se observa que al aumentar la cantidad de las precipitaciones, la cantidad de demoras comienzan a ser más significativas. Esto implica que al **aumentar las precipitaciones** las probabilidades de que el vuelo programado tenga *demoras es más alta*.

Respecto al impacto, se puede observar que algunos aeropuertos se pueden ver afectados en las tasas de demoras que poseen por este factor climático, donde encontramos que los aeropuertos de Chicago tienen una al tasa de precipitaciones y de demora, sin embargo, el aeropuerto de Pensacola es el aeropuerto que mayor cantidad de precipitaciones promedio posee y una alta tasa de demoras.

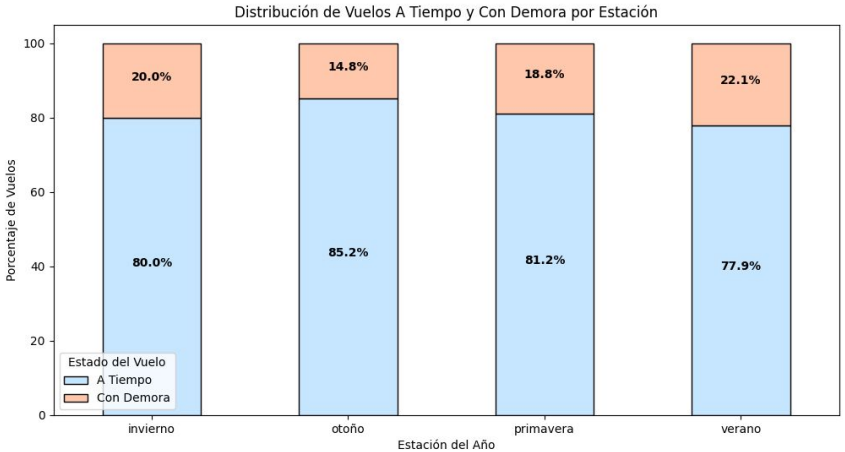


INFLUENCIA DEL CLIMA

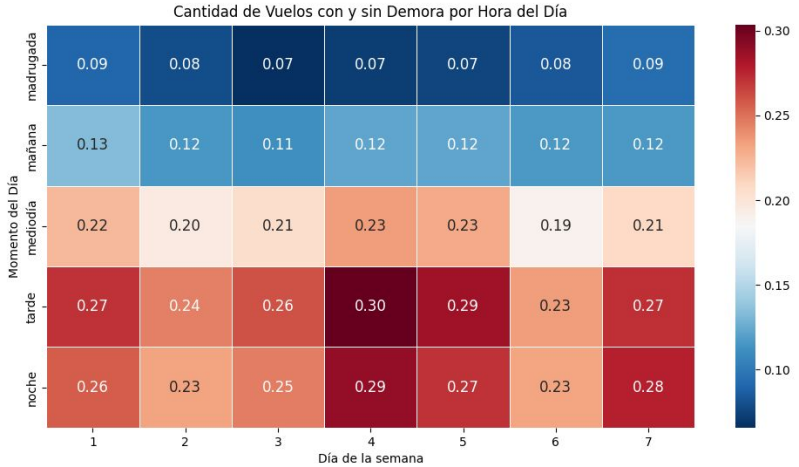


Respecto a la estacionalidad, aunque el invierno y las estaciones de transición tienen climatologías que en el verano no se presentan, en el verano se observa una mayor cantidad de demoras que en el resto del año, aunque hay mayor cantidad de vuelos, el porcentaje sobre el total es mayor.

Continuando con el análisis de la influencia del clima en las demoras de los vuelos, se observa que en tres estaciones del año la nieve en pista es un factor determinante en sí un vuelo será demorado o no.
En el mes de **invierno** se observa que la cantidad de nieve en pista debe ser mayor para generar un impacto en el vuelo programados, mientras que **primavera** y **otoño** se observa que el comportamiento es bastante similar, siendo que en primavera la cantidad de nieve en pista que no genera impacto es mayor que en otoño, y en otoño se observa que la cantidad de nieve en pista es mayor que en primavera.

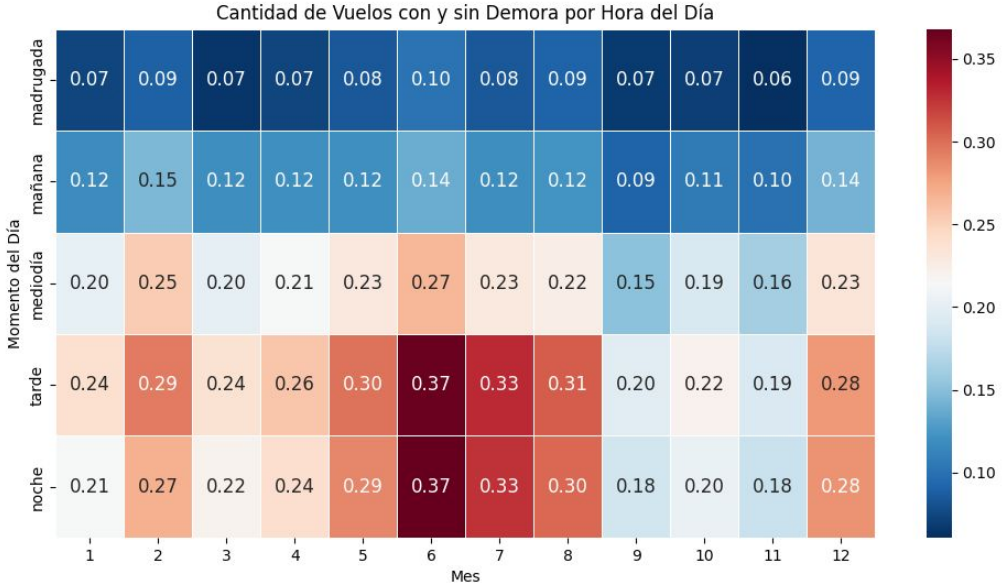


INFLUENCIA DEL DÍA



Se puede observar que el momento del día que mayor demora presenta es entre la tarde y noche, siendo el día miércoles, jueves, sábados y domingos los días donde mayor posibilidad de demora es posible.

Haciendo el mismo análisis sobre los meses del año, se puede observar que tiene un comportamiento similar, siendo los meses más afectados los meses invernales (mayo, junio, julio y agosto).



05. CONCLUSIONES

¿Es posible predecir si un vuelo se demora?

Con el análisis realizado no existe una relación directa entre una variable y la demora, sin embargo, las combinaciones de factores seguramente permitirán predecir si un vuelo se demora o no.

¿Existe estacionalidad en las demoras y sus posibles predicciones?

Realmente si se puede observar una estacionalidad siendo los meses de verano e invierno donde se producen las mayores demoras, junto con los horarios de la noche y madrugada.

¿Existen factores internos y/o externos que influyen en la predicción?

Si, ya que la predicción se basará en una combinación de factores internos como es el horario programado de un vuelo y el aeropuerto de donde sale ese vuelo, que a su vez, dependerá de la estación del año y las condiciones meteorológicas.

¿Qué factores internos son los más preponderantes a influir en la demora?

Se ha observado que prácticamente todas las aerolíneas poseen el mismo rate de demora, sin embargo, se observa que pueden haber factores como el número de segmento.

06. MODELOS DE MACHINE LEARNING

El problema planteado es determinar si un vuelo se demora o no, es decir, estando en un problema de ***categorización***. Para esto utilizaremos los siguientes modelos para intentar resolver el problema planteado:

- Regresión Logística
- KNN
- Random Forest
- XGBoost

Antes de realizar el entrenamiento, el **dataset** se transformaron las variables en categorías para poder trabajar con un modelo numérico y que permita resumir el tratamiento en los modelos.

Además, implementando los diferentes técnicas, dado que nos encontramos con un modelo que no existe una correlación directa con las variables predictoras, se eligieron las siguientes variables para lograr el entrenamiento:

- MOMENT_OF_THE_DAY_ENCODED
- PREVIOUS_AIRPORT_ENCODED
- DEPARTING_AIRPORT_ENCODED
- SEGMENT_NUMBER
- SEASON_ENCODED
- PRCP_ENCODED
- SNOW_ENCODED
-

REGRESIÓN LOGÍSTICA

Este modelo permite definir por medio de las variables predictoras la probabilidad de que un vuelo se demore o no. Teniendo en cuenta las variables utilizadas como predictoras en la primera aproximación se logran los siguientes resultados:

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0        12163           18
Actual 1         2807           12

Classification Report:
              precision    recall  f1-score       support
0              0.812492  0.998522  0.895952  12181.000000
1              0.400000  0.004257  0.008424   2819.000000
accuracy       0.811667  0.811667  0.811667      0.811667
macro avg      0.606246  0.501390  0.452188  15000.000000
weighted avg   0.734971  0.811667  0.729156  15000.000000
```

Con estos resultados, podemos observar que el modelo permite predecir con una buena precisión los vuelos que no se demorarán y que realmente no se demoraron, sin embargo, el modelo no es bueno para predecir si un vuelo puede llegar a demorarse o que realmente se demoró. Por lo tanto, aunque tenga una buena precisión, este modelo no logra el objetivo planteado.

REGRESIÓN LOGÍSTICA

Si realizamos ajustes en el modelo por medio del uso de hiperparámetros y GridSearch se obtiene el siguiente resultado:

- *Mejores hiperparámetros encontrados: {'C': np.float64(1e-05), 'penalty': 'l2', 'solver': 'liblinear'}*

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0        11354          827
Actual 1         2413          406

Classification Report:
              precision    recall  f1-score   support
0              0.824726  0.932107  0.875135  12181.000
1              0.329278  0.144023  0.200395   2819.000
accuracy       0.784000  0.784000  0.784000      0.784
macro avg      0.577002  0.538065  0.537765  15000.000
weighted avg   0.731615  0.784000  0.748329  15000.000
```

Se puede observar que, aunque se busca mejorar el modelo de regresión logística, este no ofrece una mejora en los resultados.

Por lo tanto, el modelo de regresión logística no cumple con los requisitos del modelo.

KNN

KNN permite predecir valores basándose en la suposición de que datos similares tienden a tener valores similares. Para un nuevo dato, identifica los 'k' datos más similares según sus variables independientes y usa los valores de la variable dependiente de esos vecinos para realizar la predicción. Teniendo en cuenta esto, el resultado del primer modelo utilizado nos brinda el siguiente resultado.

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0        10961         1220
Actual 1         2391          428

Classification Report:
              precision    recall  f1-score       support
0              0.820926  0.899844  0.858575  12181.000000
1              0.259709  0.151827  0.191627   2819.000000
accuracy       0.759267  0.759267  0.759267      0.759267
macro avg      0.540317  0.525835  0.525101  15000.000000
weighted avg   0.715454  0.759267  0.733233  15000.000000
```

Nos encontramos que el modelo mejora la performance en definir si un vuelo se demora, sin embargo, sigue teniendo una performance muy baja para el objetivo del proyecto.

KNN

Lo mismo hecho anteriormente, se utilizan hiperparámetros y GridSearch para intentar mejorar la performance del modelo.

- *Mejores hiperparámetros encontrados: {'algorithm': 'brute', 'n_neighbors': 3, 'p': 2, 'weights': 'distance'}*

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0        10743         1438
Actual 1         2324          495

Classification Report:
              precision    recall  f1-score     support
0              0.822147  0.881947  0.850998  12181.0000
1              0.256079  0.175594  0.208333   2819.0000
accuracy       0.749200  0.749200  0.749200      0.7492
macro avg      0.539113  0.528771  0.529666  15000.0000
weighted avg   0.715764  0.749200  0.730220  15000.0000
```

Se observa una leve mejora en la determinación de los vuelos que se demoran, aunque disminuye la precisión, ya que detecta mayor cantidad de falsas demoras.

Por lo cual, aunque el modelo mejora podemos evidenciar que no cumple con el objetivo del problema

RANDOM FOREST

Este modelo utiliza múltiples árboles de decisión, cada uno de los cuales se comporta como un flujograma. Cada árbol toma decisiones basadas en un subconjunto aleatorio de los datos y las características, y el modelo final predice el resultado agregando las predicciones de todos los árboles.

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0        11151         1030
Actual 1         2409          410

Classification Report:
              precision    recall  f1-score       support
0              0.822345  0.915442  0.866400  12181.000000
1              0.284722  0.145442  0.192533   2819.000000
accuracy       0.770733  0.770733  0.770733      0.770733
macro avg      0.553534  0.530442  0.529467  15000.000000
weighted avg   0.721308  0.770733  0.739758  15000.000000
```

Podemos observar que es primer modelo de Random Forest, posee el mismo comportamiento que el de KNN, sin embargo, tiene mejor rendimiento en las clase de detectar que un vuelo no se demora.

RANDOM FOREST

Similar a lo realizado en los otros dos modelos, se utilizarán hiper parámetros y el método de Grid Search para mejorar la precisión del modelo:

- Mejores hiper parámetros: {'class_weight': 'balanced', 'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 100}

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0         7851         4330
Actual 1         1170         1649

Classification Report:
              precision    recall  f1-score       support
0              0.870303  0.644528  0.740591  12181.000000
1              0.275799  0.584959  0.374858   2819.000000
accuracy       0.633333  0.633333  0.633333      0.633333
macro avg      0.573051  0.614744  0.557724  15000.000000
weighted avg   0.758576  0.633333  0.671857  15000.000000
```

Ahora el modelo es más *"sensible"* a las demoras, pero esto lo hace a costa de predecir muchas demoras que no ocurren y de clasificar incorrectamente vuelos puntuales como demorados.

XGBoost

Este modelo es similar al Random Forest pero trabaja de una manera distinta, evitando el sobreajuste al realizar los árboles de manera secuencial, en vez de manera independiente, logrando reducir los errores de los árboles a medida que se va realizando la construcción. En la primera aproximación nos encontramos con el siguiente resultado:

```
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0        11951          230
Actual 1         2639          180

Classification Report:
              precision    recall  f1-score       support
0              0.819123  0.981118  0.892832  12181.000000
1              0.439024  0.063852  0.111490   2819.000000
accuracy       0.808733  0.808733  0.808733      0.808733
macro avg      0.629074  0.522485  0.502161  15000.000000
weighted avg   0.747690  0.808733  0.745992  15000.000000
```

Nuevamente, nos encontramos con un modelo que es muy bueno prediciendo vuelos que no sufren demoras, sin embargo, no es bueno para predecir aquellos que si se demoran.

XGBoost

Aplicando hiper parámetros y utilizando Grid Search tratamos de mejorar el resultado del primer modelo:

- Mejores hiper parámetros: {'class_weight': 'balanced', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}

```
Umbral: 0.6
Confusion Matrix:
          Predicted 0  Predicted 1
Actual 0         5147         7034
Actual 1          611         2208

Classification Report:
              precision    recall  f1-score       support
0              0.893887  0.422543  0.573834  12181.000000
1              0.238909  0.783256  0.366139   2819.000000
accuracy       0.490333  0.490333  0.490333      0.490333
macro avg      0.566398  0.602900  0.469986  15000.000000
weighted avg   0.770795  0.490333  0.534801  15000.000000
```

El resultado mostrado es el mejor umbral del modelo, el cual, permite determinar con una mayor precisión los vuelos que se demorarán pero esto sacrificando la precisión de los vuelos que no tienen demoras.

07. RESULTADOS

MODELO	RESULTADOS	CONCLUSIONES
Regresion Logística	Excelente para predecir vuelos que no se demoran (alta precisión y recall para la clase 0), pero muy deficiente para predecir vuelos que sí se demoran (baja precisión y recall para la clase 1).	El modelo está fuertemente sesgado hacia la clase mayoritaria (vuelos sin demora) y es prácticamente inútil para identificar demoras. El ajuste de hiper parámetros no mejoró significativamente este comportamiento.
KNN	Ligeramente mejor que la Regresión Logística en la detección de demoras, pero aún con un rendimiento deficiente en la clase 1. El ajuste de hiper parámetros no produjo una mejora sustancial.	KNN captura algo mejor la clase minoritaria que la Regresión Logística, pero sigue teniendo dificultades para equilibrar la precisión y el recall para las demoras.
Random Forest	Muestra un cambio importante en el equilibrio entre precisión y recall. Logra un mejor recall para la clase 1 (detecta más demoras reales) pero a costa de una menor precisión para la clase 0 (aumentan las falsas alarmas de demoras).	El modelo se vuelve más sensible a las demoras, pero predice muchas demoras que no ocurren y clasifica incorrectamente vuelos puntuales. El rendimiento general sigue siendo subóptimo.
XGBoost	Con el ajuste de hiper parámetros y el ajuste del umbral de probabilidad, XGBoost muestra la capacidad de lograr un mejor equilibrio entre precisión y recall. La elección del umbral permite priorizar la detección de demoras o la reducción de falsas alarmas según las necesidades.	XGBoost, con una cuidadosa optimización, demuestra ser el modelo más versátil y potencialmente más efectivo para el problema de predicción de demoras.

08. CONCLUSIONES

Luego de la utilización de diferentes modelos y métodos para lograr una mayor precisión de para determinar si un vuelo tendrá una demora o no, se pueden realizar las siguientes conclusiones:

- Existen múltiples variables en la operación que no se encuentran correlacionadas que pueden generar una demora.
- No existe un patrón o preponderancia que permitan realizar la predicción con una gran precisión.
- Al utilizar las variables más relevantes de la base de datos, cada modelo tiene comportamiento diferentes.
- El desbalance de las clases tiene un factor preponderante en la performance del modelo, ya que solo el 18% de la base de datos tienen registro de demoras.
- Los modelos de Random Forest o XGBoost son los que mejor se adaptan al requerimiento.

Por lo tanto, dada la naturaleza multifactorial y la baja correlación de las variables con las demoras, la predicción precisa resulta desafiante y compleja. Aún con los ajustes realizados, las precisiones e identificaciones no son las deseadas en el proyecto, pero son una gran primera aproximación para la resolución del problema.

FIN