2019 Airline Delays

Predicción de retrasos
{Alumno}

Introducción

Importancia del análisis de retrasos

- Impacta experiencia del pasajero, operaciones y costos aerolíneas.
- Permite tomar medidas proactivas y optimizar recursos.

Objetivo del proyecto

- Desarrollar un modelo predictivo para anticipar retrasos en vuelos.
- Basado en datos reales de vuelos de 2019.

Enfoque

- Clasificación binaria (retraso vs. no retraso).
- Análisis de factores clave: clima, congestión, antigüedad de aviones, etc.

Motivación y audiencia

La predicción de retrasos aéreos es clave para **mejorar la eficiencia operativa** de las aerolíneas, reducir costos asociados a impuntualidades y **optimizar la experiencia de los pasajeros**. Este proyecto busca desarrollar un modelo de machine learning que anticipe retrasos, permitiendo una gestión proactiva de recursos. Está dirigido a:

- Equipos operativos (control de vuelos, planificación)
- Analistas de datos (enfoque técnico en modelado predictivo)
- Tomadores de decisiones (gerentes de operaciones, ejecutivos)

Resumen de los datos

Volumen: 6.5+ millones de registros de vuelos correspondientes al año 2019

Variables clave

- Operacionales: aerolínea, aeropuerto, hora de salida
- Climáticas: precipitación (PRCP), velocidad del viento (AWND), nieve (SNOW)
- Técnicas: antigüedad del avión, número de asientos
- Temporales: mes, día de la semana

Hallazgos relevantes

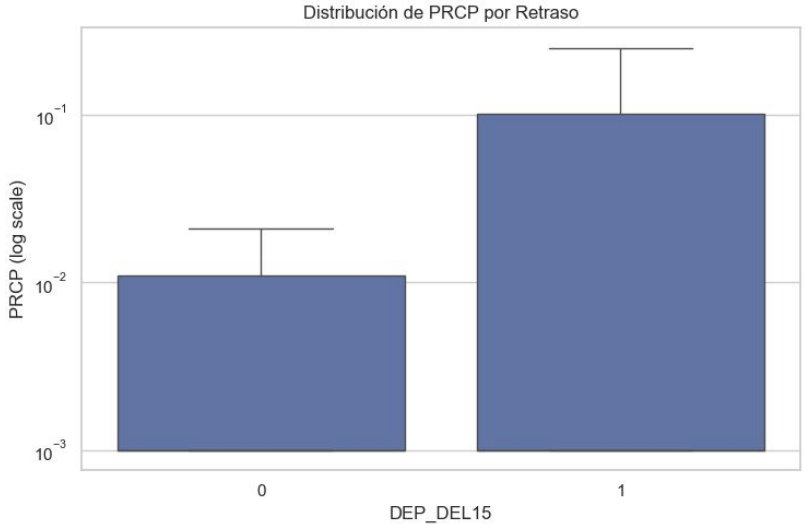
- 19% de los vuelos presentaron retrasos (>15 min)
- Variables climáticas muestran fuerte correlación con retrasos
- Patrones estacionales claros (picos en diciembre/julio)
- Datos desbalanceados (81% no retrasos vs 19% retrasos)

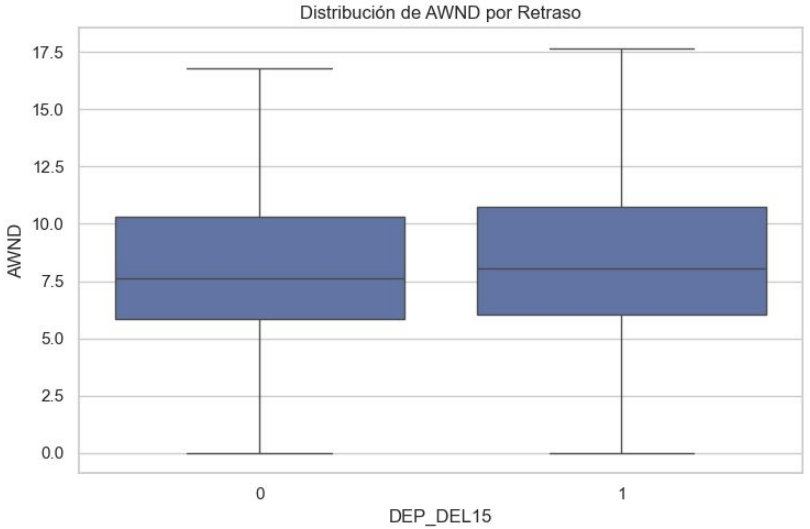
Estructura

- 22 variables numéricas / 4 categóricas
- Datos meteorológicos con alta frecuencia de ceros (sin eventos) pero valores extremos predictivos

Hipótesis 1

Los vuelos con condiciones meteorológicas adversas tienen una probabilidad significativamente mayor de retraso

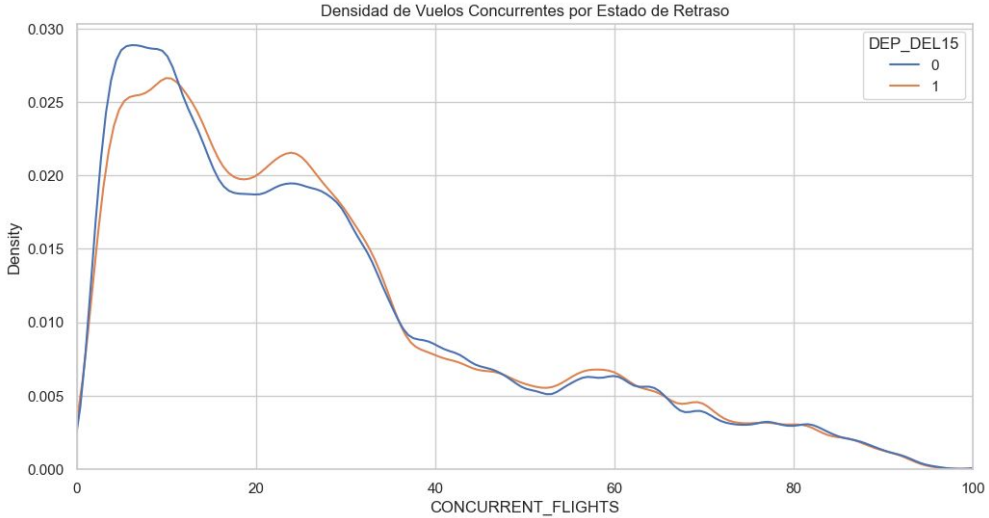




El efecto climático es de baja frecuencia pero alta severidad: aunque pocos vuelos experimentan estos eventos, cuando ocurren, multiplican el riesgo de retraso (OR > 1.8 para PRCP/SNOW > 0)

Hipótesis 2

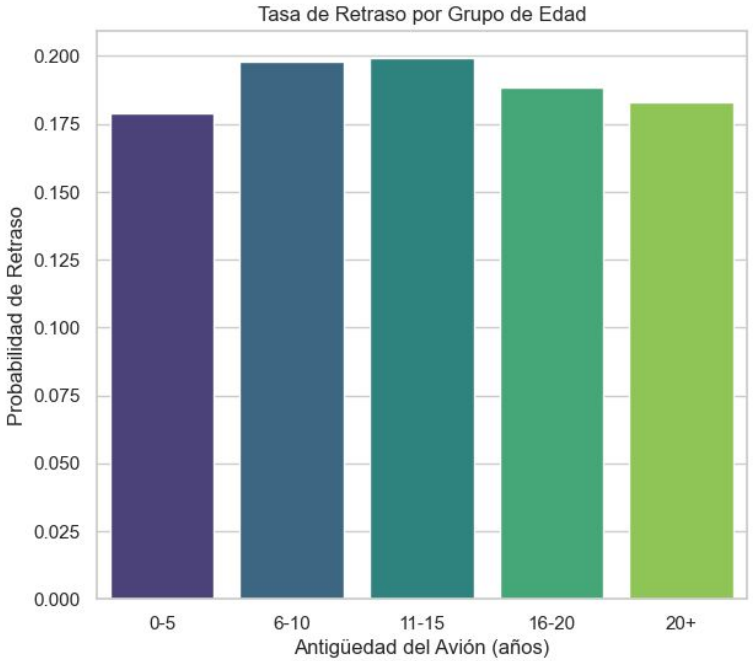
Los aeropuertos con mayor número de vuelos concurrentes presentan tasas de retraso superiores a la media



Los resultados sobre la hipótesis de congestión revelan hallazgos contraintuitivos: no parecen correlacionar con la demora del vuelo.

Hipótesis 3

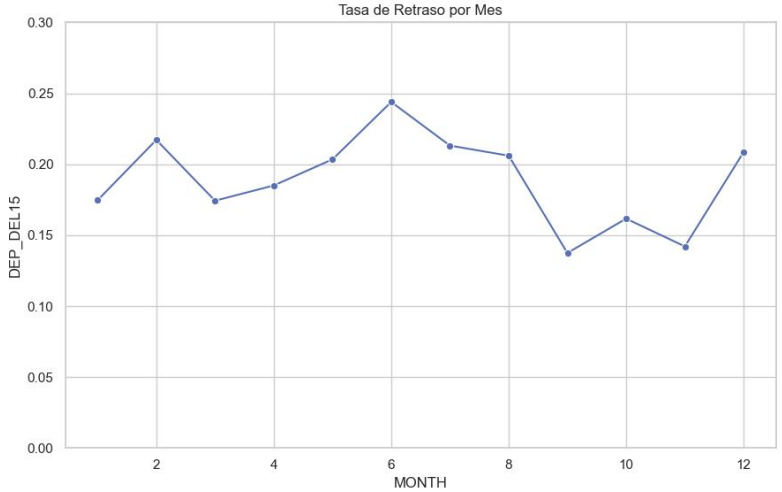
La edad de los aviones correlaciona positivamente con la frecuencia de retrasos por mantenimiento.

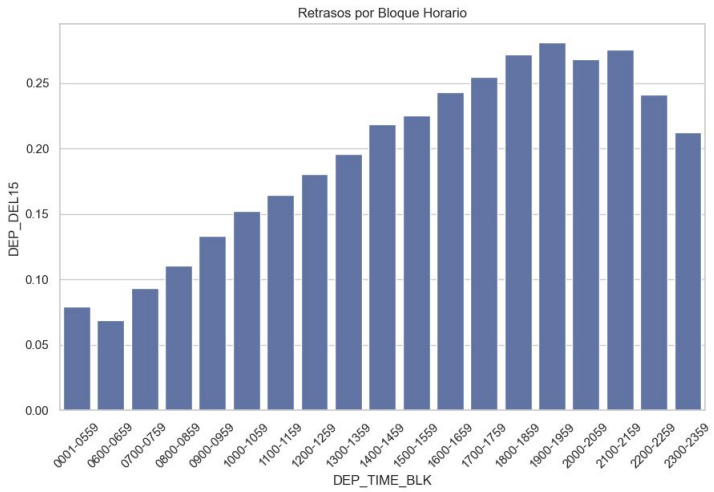


El pico en aviones de 11-15 años sugiere una ventana crítica donde el desgaste supera los protocolos de mantenimiento estándar, mientras que los aviones más antiguos podrían beneficiarse de inspecciones más frecuentes.

Hipótesis 4

Existen patrones estacionales (meses de invierno) y horarios (turno tarde-noche) con incidencia significativamente mayor de retrasos.





Se observan diferencias significativas en los distintos meses del año, así como una diferencia (esperable) en distintos momentos del día (dado que correlaciona muy fuertemente con volumen)

Procesamiento de los datos

[Detalle del pipeline de preprocesamiento de los datos realizado]

Entrenamiento de modelos

Modelo	Accuracy	Recall	Tiempo (s)
Random Forest	62.7%	66.0%	585.8
XGBoost	81.9%	8.2%	11.2
Regresión Logística	81.1%	1.3%	134.2
Árbol de Decisión	81.2%	0.9%	21.5

Hallazgos Clave

- Random Forest es el único modelo con Recall alto (66%), pero:
  - Baja precisión (28.8%): muchos falsos positivos
  - Tiempo de entrenamiento elevado
- XGBoost y modelos lineales:
  - Accuracy engañoso (>80%) por desbalance de clases
  - Recall inaceptable (<10%): fallan en detectar retrasos

Próximos Pasos

- Balancear dataset (SMOTE o class_weight)
- Optimizar hiper parámetros (foco en Recall)
- Probar LightGBM como alternativa rápida