Ficha Técnica: Proyecto 4 Datalab

Título del Proyecto
Proyecto 4 Datalab

Objetivo

Desarrollar un score de riesgo para los vuelos mediante el análisis de datos y la evaluación del riesgo relativo, que permita clasificar los vuelos en distintas categorías de riesgo basadas en su probabilidad de retraso. Esta clasificación ayudará a las aerolíneas a tomar decisiones informadas sobre cancelaciones y a reducir los riesgos de retraso en sus operaciones. Además, la integración de métricas existentes sobre el estado de los vuelos fortalecerá el modelo, mejorando su capacidad para identificar riesgos y, en última instancia, gestionar eficazmente los retrasos.

Equipo
Herramientas y Tecnologías
Listado de herramientas y tecnologías utilizadas en el proyecto:

1) Big Query (SQL).
2) Google Colab (Python).
3) Workspace Google (Presentaciones, Chat GPT y Documentos).
4) Videos Looms.
5) Git Hub.

Selecionamos el Proyecto Flight Delay and Cancellation: Se nos proporcionó el Conjunto de datos sobre cancelaciones y retrasos de vuelos de aerolíneas de enero de 2023, datos extraídos del Departamento de Transporte de EE.UU., Oficina de Estadísticas de Transporte (https://www.transtats.bts.gov) y disponibles en Kaggle.

Las variables incluyen rutas de vuelo (origen y destino), rangos de tiempo para eventos (minutos, hora en destino), motivos/atribuciones de retrasos y cancelaciones.

Descripción de las variables:
Tabla DOT CODE DICTIONARY

- Code: Identificador numérico del U.S. Department of Transportation (DOT) para aerolíneas
- Description: descripción de la aerolínea

Tabla AIRLINE CODE DICTIONARY

- Code: Código de operador único para agencias operadoras de aeronaves
- Description: descripción de la agencia operadora de aeronaves

Tabla flights_202301

- FL_DATE: Fecha de vuelo (yyyymmdd)
- AIRLINE_CODE: Código de operador único. Cuando varios operadores han utilizado el mismo código, se utiliza un sufijo numérico para usuarios anteriores, por ejemplo, PA, PA(1), PA(2).
- DOT_CODE: Un número de identificación asignado por el DOT de EE. UU. para identificar una aerolínea (transportista) única. Una aerolínea (transportista) única se define como aquella que posee y reporta bajo el mismo certificado DOT independientemente de su código, nombre o compañía/corporación holding.
- FL_NUMBER: Número de vuelo
- ORIGIN: Aeropuerto de origen
- ORIGIN_CITY: Aeropuerto de origen, nombre de la ciudad
- DEST: Aeropuerto de destino
- DEST_CITY: Aeropuerto de destino, nombre de la ciudad
- CRSDEPTIME: Hora de salida registrada CRS (Sistema de control de reservas) (hora local: hhmm)
- DEP_TIME: Hora de salida real (hora local: hhmm)
- DEP_DELAY: Diferencia en minutos entre la hora de salida prevista y la real. Las salidas anticipadas arrojan cifras negativas.
- TAXI_OUT: Tiempo de taxi en la salida en minutos (taxi es el proceso de mover un avión mientras se encuentra en la pista)
- WHEELS_OFF: hora exacta de despegue (hora local: hhmm)
- WHEELS_ON: hora exacta de aterrizaje (hora local: hhmm)
- TAXI_IN: tiempo de taxi en la llegada en minutos
- CRSARRTIME: Hora de llegada registrada en CRS (hora local: hhmm)
- ARR_TIME: Hora de llegada real (hora local: hhmm)
- ARR_DELAY: Diferencia en minutos entre la hora de llegada prevista y la real. Las llegadas anticipadas arrojan cifras negativas.
- CANCELLED: Indicador de vuelo cancelado (1=Sí)
- CANCELLATION_CODE: Especifica el motivo de la cancelación
- DIVERTED: Indicador de vuelo desviado (1=Sí)
- CRSELAPSEDTIME: Tiempo total de vuelo transcurrido en minutos registrado en CRS
- ELAPSED_TIME: Tiempo total de vuelo transcurrido en minutos real
- AIR_TIME: Tiempo de vuelo en el aire en minutos
- DISTANCE: Distancia entre aeropuertos (millas)
- DELAYDUECARRIER: Retraso del operador en minutos
- DELAYDUEWEATHER: Retraso meteorológico en minutos
- DELAYDUENAS: Retraso del Sistema Aéreo Nacional en Minutos
- DELAYDUESECURITY: Retraso de seguridad en minutos
- DELAYDUELATE_AIRCRAFT: Retraso de aeronaves tardías en minutos

Procesamiento y análisis

1) **Proceso de Análisis de Datos:**

1.1 Procesar y preparar base de datos

a) Identificar y manejar valores nulos
b) Identificar y manejar valores duplicados
c) Identificar y manejar datos fuera del alcance del análisis
d) Identificar y manejar datos discrepantes en variables categóricas
e) Identificar y manejar datos discrepantes en variables numéricas
f) Comprobar y cambiar tipo de dato
g) Unir tablas
h) Crear nuevas variables
i) Construir tablas auxiliares

1.2 Análisis exploratorio

a) Agrupar datos según variables categóricas.
b) Visualizar las variables categóricas.
c) Aplicar medidas de tendencia central.
d) Visualizar distribución.
e) Aplicar medidas de dispersión.
f) Calcular cuartiles, deciles o percentiles.
g) Calcular correlación entre variables.

1.3 Aplicar técnica de análisis

h) Aplicar segmentación.
i) Calcular riesgo relativo
j) Validar hipótesis.

Resultados y Conclusiones

Durante el desarrollo del proyecto se utilizó el proceso de Análisis de Datos, cálculo de riesgo relativo y la metodología de segmentación (Cuartiles) para validación de hipótesis. Se plantearon seis hipótesis, para saber qué hace que un vuelo presente retraso. Estas hipótesis se detallan a continuación y de esto se puede concluir lo siguiente:

I) Existen rutas que presentan una mayor frecuencia de retrasos en comparación con otras.

Se valida la hipótesis con los resultados obtenidos en Figura N° 1: Frecuencia de retrasos por ruta (arr_delay_winsorized) que muesta que la ruta LGA to ORD tiene una frecuencia de 147 retrasops en el periodo analizado.

II) Es posible calcular el tiempo promedio de retraso por ruta utilizando las variables disponibles.

Se valida la hipótesis con los resultados de la tabla N° 1 podemos indicar que existen 4 rutas que presentan el mayor valor promedio de retraso (AUS to SRQ, BOS to VPS, FLL to PIE y LGA to MTJ. y el valor promedio de retraso de la muestra es de 14.33 minutos.

III) Con los datos disponibles, se puede identificar el principal motivo que genera los retrasos en los vuelos.

Se valida la hipótesis con los resultados de la Figura N° 3: Motivos que generan retraso por ruta, siendo late_aircraft el principal motivo que genera los retrasos con un 39,21 %.

V) La Regresión Lineal permite predecir de manera efectiva el tiempo de retraso de un vuelo.

No es posible validar esta hipótesis dado que segùn la Figura N° 5: Correlación entre variables delay_due_late_aircraft y delay_due_late_aircraft, su correlaciòn es de 0, 6578.Sin embargo El modelo tiene un rendimiento deficiente en la clasificación de la Clase 0 (No Delay), pero un rendimiento muy bueno en la clasificación de la Clase 1 (Delay).

VI) La Regresión Logística es una herramienta adecuada para determinar el estado (a tiempo o retrasado) de un vuelo.

Se valida esta Hipótesis con los resultados de la Tabla N° 3: Métricas de matriz de confusión de regresión Logística, que nos indica que Las métricas proporcionadas muestran un modelo de regresión logística con un rendimiento muy alto en ambas clases. La Clase 0 (No Delay) y la Clase 1 (Delay).

Este modelo de regresión logística (RL) está logrando un rendimiento sobresaliente tanto en la clase "No Delay" como en la clase "Delay". Tiene una excelente capacidad para identificar correctamente ambas clases con muy pocos errores, como lo reflejan los altos valores de precisión, recall y F1 score. La exactitud del 98% es también muy indicativa de un modelo que generaliza bien y tiene un muy bajo número de errores. Sin embargo, es importante monitorear cualquier sesgo de clase, aunque aquí parece estar bien balanceado.

En general, el modelo de RL parece estar bien estructurado para reflejar los factores clave que influyen en los retrasos, y resalta cómo los problemas externos (como el clima y las operaciones del NAS) juegan un papel fundamental, se recomienda su uso.

Limitaciones/Próximos Pasos

Sin observaciones.

Enlaces de interés

Proeycto_4_Datalab.pdf

Proyecto 4 Datalab Flight Delay and Cancellation

Proyecto 4 Datalab Flight Delay and Cancellation

Presentación Video