Challenge AI Engineer

Instrucciones

- Debes entregar tu solución en un repositorio GitHub
- En el repositorio deben estar todos los archivos utilizados para la resolución de tu desafío. - La solución debe estar implementada utilizando python 3, indicando claramente la pregunta/funcionalidad qué estás resolviendo. No serán revisados otros lenguajes como R o similar.
- Recuerda que no estamos en tu cabeza! Escribe los supuestos que estás asumiendo.
- Para este desafío te recomendamos que describas claramente cómo mejorar cada parte de tu ejercicio en caso de que tenga opción de mejora.
- Debes subir el link al repositorio en el formulario enviado, máximo 5 días de corrido después de haberlo recibido.

Problema

El problema consiste en predecir la probabilidad de atraso de los vuelos que aterrizan o despegan del aeropuerto de Santiago de Chile (SCL). Para eso les entregamos un dataset usando datos públicos y reales donde cada fila corresponde a un vuelo que aterrizó o despegó de SCL. Para cada vuelo se cuenta con la siguiente información:

1. **Fecha-I**: Fecha y hora programada del vuelo.
2. **Vlo-I**: Número de vuelo programado.
3. **Ori-I**: Código de ciudad de origen programado.
4. **Des-I**: Código de ciudad de destino programado.
5. **Emp-I**: Código aerolínea de vuelo programado.
6. **Fecha-O**: Fecha y hora de operación del vuelo.
7. **Vlo-O**: Número de vuelo de operación del vuelo.
8. **Ori-O**: Código de ciudad de origen de operación
9. **Des-O**: Código de ciudad de destino de operación.
10. **Emp-O**: Código aerolínea de vuelo operado.
11. **DIA**: Día del mes de operación del vuelo.
12. **MES**: Número de mes de operación del vuelo.
13. **AÑO**: Año de operación del vuelo.
14. **DIANOM**: Día de la semana de operación del vuelo.
15. **TIPOVUELO** : Tipo de vuelo, I =Internacional, N =Nacional.
16. **OPERA**: Nombre de aerolínea que opera.
17. **SIGLAORI**: Nombre ciudad origen.
18. **SIGLADES**: Nombre ciudad destino.

Tu desafío consiste en crear un servicio que entregue la probabilidad de atraso para los próximos vuelos y exponerlo para que sea utilizado por el resto de la compañía, cumpliendo los siguientes requisitos:

1. ¿Cómo se distribuyen los datos? ¿Qué te llama la atención o cuál es tu conclusión sobreesto?
2. Genera las columnas adicionales y luego expórtelas en un archivo synthetic_features.csv:
    a. **Temporada alta**: la temporada alta se considera si *Fecha-I* está entre 15 Diciembre y 3 Marzo, o 15 Julio y 31 Julio, o 11 Septiembre y 30 Septiembre.
    b. **Diferencia en minutos** : diferencia en minutos entre Fecha-O y Fecha-I .
    c. **Atraso menor**: atrasos menores son aquellos donde el tiempo total de atraso es más de 0 minutos y menos de 15 minutos.
    d. **Periodo día**: donde los periodos del día se consideran como *mañana* (entre 5:00 y 11:59), *tarde* (entre 12:00 y 18:59) y *noche* (entre 19:00 y 4:59), en base a Fecha-I.
3. Entrena uno o varios modelos usando los algoritmos que prefieras para estimar la probabilidad de atraso de un vuelo. Siéntete libre de generar variables adicionales y/o complementar con variables externas.
4. Escoge el modelo que a tu criterio tenga una mejor performance, argumentando tu decisión.
5. Serializa el mejor modelo seleccionado e implementa una API REST para poder predecir atrasos de nuevos vuelos.
6. Automatiza el proceso de build y deploy de la API, utilizando uno o varios servicios cloud. Argumenta tu decisión sobre los servicios utilizados.
7. Realiza pruebas de estrés a la API con el modelo expuesto con al menos 50.000 requests durante 45 segundos. Para esto debes utilizar esta herramienta y presentar las métricas obtenidas. ¿Cómo podrías mejorar la performance de las pruebas anteriores?

Consideraciones

- Recuerda escribir los supuestos que tomaste y argumentar tus decisiones.
- Documentar MUY bien tu trabajo. Recomendamos utilizar un README o markdown donde puedas contar y dar a entender tus decisiones y supuestos. Recuerda que no estamos en tu cabeza!
- Criterios a considerar:
    - Creatividad en las técnicas y/o herramientas utilizadas.
    - Simplicidad y eficiencia.
    - Performance.
    - Calidad de conclusiones.
    - Orden y documentación.