ML Engineer Challenge - LATAM Airlines

Instrucciones:

En Advanced Analytics valoramos muchísimo el trabajo en equipo y la constante interacción entre los distintos roles que trabajan en un producto basado en datos, como el Data Scientist, Machine Learning Engineering, Data Engineer, entre otros. Es por este motivo que una habilidad esencial que buscamos a la hora de buscar nuevos integrantes es el manejo adecuado de git. **Este desafío deberá ser entregado en la plataforma de git que más te acomode y que sea pública para que la podamos revisar.** Lo que buscamos con esto es poder entender de mejor manera el desarrollo que generaste con tu código, cómo lo fuiste mejorando en el tiempo y si tienes proyectos propios en este repositorio nos servirán para conocer mejor tu experiencia en base a tu propio portafolio.

Instrucciones Git:

1) Crear un repositorio en la plataforma de git que más te acomode y que sea pública
2) Haber trabajado con una rama principal y otra de desarrollo. Opcional, ocupar alguna práctica de desarrollo de GitFlow.

Instrucciones Desafío:

1. Debes enviar el link al repositorio al mail cristobal.guzman@latam.com con asunto `Machine Learning Engineer - [Nombre][Apellido]`, ejemplo `Machine Learning Engineer - Cristobal Guzman`.
2. Se aceptará los cambios en el repositorio hasta el **Domingo 13 de 2022 de 2022 a las 23:59 hrs**.
3. En la siguiente carpeta de Google Drive encontrarás las instrucciones del desafío, el archivo `dataset_SCL.csv` y un jupyter notebook `'to-expose.ipynb'` que utilizarás para el desafío.
4. En el repositorio deben estar todos los archivos utilizados para la resolución de tu desafío.
5. Puedes utilizar el lenguaje que prefieras, pero los desafíos con Python serán mejores rankeados.
6. Puedes utilizar las tecnologías y técnicas que prefieras para el procesamiento de datos. Incluso servicios cloud.
7. Recuerda que no estamos en tu cabeza! Escribe los supuestos que estás asumiendo.
8. Para este desafío te recomendamos que describas claramente cómo mejorar cada parte de tu ejercicio en caso de que tenga opción de mejora.
9. Una copia de tu CV (curriculum vitae) en formato .pdf
10. Éxito!

Problema:

- Se ha proporcionado un júpiter notebook que contiene el trabajo de un Data Scientist.
- El problema que trató de resolver el DS fue predecir la probabilidad de atraso de los vuelos que aterrizan o despegan del aeropuerto de Santiago de Chile (SCL). Para eso utilizó un dataset de 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:
    a. **Fecha-I**: Fecha y hora programada del vuelo.
    b. **Vlo-I**: Número de vuelo programado.
    c. **Ori-I**: Código de ciudad de origen programado.
    d. **Des-I**: Código de ciudad de destino programado.
    e. **Emp-I**: Código aerolínea de vuelo programado.
    f. **Fecha-O**: Fecha y hora de operación del vuelo.
    g. **Vlo-O**: Número de vuelo de operación del vuelo.
    h. **Ori-O**: Código de ciudad de origen de operación
    i. **Des-O**: Código de ciudad de destino de operación.
    j. **Emp-O**: Código aerolínea de vuelo operado.
    k. **DIA**: Día del mes de operación del vuelo.
    l. **MES**: Número de mes de operación del vuelo.
    m. **AÑO**: Año de operación del vuelo.
    n. **DIANOM**: Día de la semana de operación del vuelo.
    o. **TIPOVUELO** : Tipo de vuelo, I =Internacional, N =Nacional.
    p. **OPERA**: Nombre de aerolínea que opera.
    q. **SIGLAORI**: Nombre ciudad origen.
    r. **SIGLADES**: Nombre ciudad destino.

- En el jupyter vas a encontrar:
    1. ¿Cómo se distribuyen los datos?
    2. Generación de columnas adicionales:
        a. **Temporada_alta**: 1 si Fecha-I está entre 15-Dic y 3-Mar, o 15-Jul y 31-Jul, o 11-Sep y 30-Sep, 0 si no.
        b. **dif_min**: diferencia en minutos entre Fecha-O y Fecha-I.
        c. **atraso_15**: 1 si dif_min > 15, 0 si no.
        d. **periodo_dia:** 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. ¿Cómo se compone la tasa de atraso por destino, aerolínea, mes del año, día de la semana, temporada, tipo de vuelo?
    4. Entrenamiento de uno o varios modelos para estimar la probabilidad de atraso de un vuelo (target **atraso_15**).
    5. Evaluación del performance del o los modelos.

Desafío

Como ML Engineer, tu desafío consiste en tomar el trabajo de este Data Scientist y exponerlo para que sea explotado por un sistema:

1. Escoger el modelo que a tu criterio tenga un mejor performance, argumentando la decisión.
2. Implementar mejoras sobre el modelo escogiendo la o las técnicas que prefieras.
3. Exponer el modelo seleccionado como API REST para ser expuesto.
4. Hacer 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.
5. ¿Cómo podrías mejorar el performance de las pruebas anteriores?

Consideraciones

- Documentar MUY bien tu trabajo. Recomendamos utilizar UN jupyter notebook con todos los demás códigos necesarios donde puedas contar y dar a entender tus decisiones. Recuerda que no estamos en tu cabeza!
- Si la construcción y exposición del modelo es automatizada, sin importar la tecnología que estés utilizando, será un graaaan plus.
- Criterios a considerar:
    - Creatividad en las técnicas y/o herramientas utilizadas.
    - Elegancia en las soluciones.
    - Performance.
    - Orden y documentación.