NeuralWorks

Challenge Machine Learning Engineer

Instrucciones

- Debes entregar tu solución en un repositorio GitHub
    - Crear un repositorio en la plataforma de git que más te acomode y que sea pública
    - Haber trabajado con una rama principal y otra de desarrollo. Opcional, ocupar alguna práctica de desarrollo de GitFlow.
- 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 que estás resolviendo.
- 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 4 días de corrido después de haberlo recibido.
- Tienes 4 días de corrido para realizar el desafío.

Problema

Te encuentras trabajando en un equipo conformado por varios Data Scientists. Uno de ellos, Juan, ha terminado sus experimentos y te ha pedido que habilites un servicio para que su modelo pueda ser consumido por varios actores fuera del equipo.

Juan escribió en un Jupyter Notebook una serie de modelos para 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.

Tu desafío consiste en tomar el trabajo de Juan y exponerlo para que sea utilizado por el resto de la compañia:

1. Escoge el modelo que a tu criterio tenga una mejor performance, argumentando tu decisión. 2. Implementa cambios sobre el modelo escogiendo la o las técnicas que prefieras buscando mejorar los resultados. Te recomendamos dejar los intentos que no lograron mejorar los resultados. 3. Serializa el modelo seleccionado (puede ser de los construidos en el punto 2) e implementa una API REST para poder predecir atrasos de nuevos vuelos.
4. Automatiza el proceso de build y deploy de la API, utilizando uno o varios servicios cloud. Argumenta tu decisión sobre los servicios utilizados.
5. 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: https://github.com/wg/wrk y presentar las métricas obtenidas. ¿Cómo podrías mejorar el performance de las pruebas anteriores?

Consideraciones

- 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.