Nequi

Prueba Técnica Machine Learning Engineer - Nequi

Consideraciones iniciales

- Entregue toda la solución en un repositorio de GitHub (con un archivo README que permite seguir el paso a paso de toda la solución propuesta).
- Debe contener mínimo un (1) dataset con datos públicos.
- El dataset debe contener mínimo 1 millón de registros (filas).

Recursos

En Ciencia de Datos, los pasos de identificación y extracción de datos son los más importantes. ya que de allí dependerá mucho los resultados y estrategias que tome el negocio. A continuación, le listamos algunos lugares donde puede tomar los datos:

- Google:DatasetSearch
- KaggleDatasets
- Github:AwesomePublicDatasets
- Github:PublicAPIs
- Data.gov
- Dataquest:18placestofinddatasetsfordatascienceprojects
- KDnuggets:DatasetsforDataMiningandDataScience
- UCIMachineLearningRepository
- Reddit:rdatasets
- LastCall:Top50MostPopularAPIsonRapidAPI(2018)
- Facebook:GraphAPI

1. Propuesta de Arquitectura en la Nube

- Proponga una arquitectura basada en la nube para desplegar un modelo de ML en Batch (*Utilice preferentemente la nube de AWS*).
- Explique cómo la arquitectura facilita escalabilidad y confiabilidad con grandes conjuntos de datos.
- Explique la elección de los componentes para la preparación de datos, trabajos ETL, implementación de modelos y su papel en la arquitectura.

2. Step-by-Step

- Diagrame la secuencia de pasos desde la ingestión de datos hasta la monitorización y el entrenamiento continuo, incluyendo la orquestación.
- Explique cómo contribuye el componente de orquestación a la ejecución del pipeline de ML.

3. Estructura de Directorios

- Proponga una estructura de directorios para el proyecto que mejore la organización y mantenimiento del código.
- ¿Cómo manejaría la versión del pipeline de preprocesamiento y los modelos entrenados en el directorio de modelos?

4. **Datos**

    - Explore los datos para hacer un paso a paso de la limpieza y calidad que tienen, donde muestre qué estrategia utilizó para sacar el máximo provecho.
    - Incluya un diccionario de datos.
    - Trace el modelo de datos conceptual y explique la selección del mismo.
    - **(BONUS TRACK - ETL)** – Este ítem da puntos extras en caso tal de resolverlo
        1. Cree el pipelline de los datos
        2. Ejecute controles de calidad
        3. Pruebas de unidad en los scripts para validad que lo que funciona, funcione como debe ser

5. **Entrenar un Modelo de ML**

    - Con el conjunto de datos previamente procesado, entrene un modelo de ML siguiendo las mejores prácticas de Ciencia de Datos y Machine Learning (no debe ser el mejor modelo, pero sí cumplir con unas métricas aceptables).

6. **Despliegue del Modelo**

    - Despliegue el modelo entrenado en la arquitectura propuesta.

7. **Pipeline de CI/CD/CT**

    - Construya un pipeline de CI/CD/CT (*usando GitHub Actions preferiblemente*).
    - Explique cómo se ejecutaría el pipeline de entrenamiento continuo.

8. **Propuesta de Monitoreo**

    - Hable sobre las métricas propuestas para monitorear el pipeline de ML y por qué son importantes.
    - ¿Cómo configuraría alertas para cambios y problemas en calidad de datos en el modelo implementado?
    - ¿Cómo configuraría alertas y acciones para una posible degradación del modelo?

Nombre del candidato a MLE:	*Por favor coloque su nombre completo*
URL del repositorio en GitHub:	*Por favor coloque la URL del repositorio de Github*
Comentarios adicionales:	*Acá le agradecemos los comentarios adicionales que pueda tener y/o el feedback que nos quiera dar con respecto a esta prueba.*

¡ Mucho ánimo !