Preguntas Clave de la Sesión 8

Bloque 1: El Ciclo MLOps y Desafíos Clave

Pregunta de Diferencia de Conceptos: Drift

Dentro de la fase de Monitorización, ¿cuál es la diferencia crítica entre Data Drift y Concept Drift, y qué implicación tiene un caso como el *Concept Drift* para el modelo en producción?

- Solución:
    - Data Drift: Es el cambio en la distribución de los datos de entrada.
    - Concept Drift: Es el cambio en la relación entre las variables de entrada y la variable objetivo.
    - Implicación: Ambos requieren Reentrenamiento y Mantenimiento, ya que el rendimiento del modelo se degrada con el tiempo.

Pregunta de MLOps DIY: El Desafío más Costoso

Una organización debe elegir entre resolver los desafíos de Escalabilidad o Reproducibilidad en un enfoque DIY. ¿Cuál de estos desafíos principales del MLOps "Do It Yourself" se considera el más costoso de resolver en términos de infraestructura y gestión de cómputo, y por qué?

- Solución: La Escalabilidad. Es el más costoso, ya que exige gestionar clústeres de GPUs y la capacidad de cómputo para escalar de GB a TB, lo cual desvía el foco de la creación de valor.

Pregunta de Gobernanza y Sesgo

¿Por qué la medición y el Análisis de Sesgos (*Bias*) son esenciales para la Gobernanza del modelo, y qué herramienta de SageMaker proporciona esta funcionalidad de equidad (*fairness*)?

- Solución: El análisis de sesgos es vital para asegurar la equidad (*fairness*) del sistema, verificando que las predicciones no sean injustas para subgrupos. SageMaker Clarify proporciona esta funcionalidad, midiendo el

sesgo en el conjunto de datos inicial (Pre-entrenamiento) y evaluando las predicciones (Post-entrenamiento).

Bloque 2: Optimización y Orquestación de Plataformas

Pregunta de Estrategia de Costes

Para un trabajo de entrenamiento de Deep Learning largo y costoso, ¿qué estrategia de SageMaker permite reducir los costes hasta en un 90% y cómo gestiona la plataforma la interrupción de la capacidad?

- Solución: Se utiliza Managed Spot Training. Este mecanismo integra el uso de Instancias Spot. SageMaker gestiona las interrupciones guardando *checkpoints* periódicamente en S3 y reanuda el entrenamiento desde el último punto de control.

Pregunta de MLOps y Abstracción

Las plataformas gestionadas ofrecen la Democratización del MLOps. ¿Cuál es la principal ventaja que aporta un servicio como SageMaker Pipelines para la colaboración entre equipos, y qué tipo de estructura utiliza para automatizar el flujo?

- Solución: La ventaja es la Automatización y la Reproducibilidad. SageMaker Pipelines permite definir todo el flujo de trabajo de ML como un Grafo Acíclico Dirigido (DAG), lo que garantiza que todos los modelos se construyan siguiendo los mismos pasos, mejorando la fiabilidad[31].

Pregunta de Comparativa Cloud y Ecosistema

Tu empresa está considerando migrar su *stack* de datos a GCP, ya que su estrategia de datos gira en torno a BigQuery. ¿Qué plataforma de ML Cloud sería la opción estratégica más natural para integrar el entrenamiento de modelos directamente desde esa fuente de datos?

- Solución: La opción más natural es GCP Vertex AI. Su principal diferenciador es la fuerte integración con BigQuery, que permite a los usuarios entrenar modelos directamente utilizando SQL (BigQuery ML).

Escenario A: Despliegue de Alto Tráfico y Baja Latencia

Descripción: Eres una gran plataforma de *e-commerce* y necesitas un modelo de recomendación que responda en **milisegundos** a cientos de miles de **peticiones constantes** por hora, 24/7.

- **Opciones de Despliegue:** 1. Inferencia Asíncrona, 2. Real-time Endpoint (Aprovis.), 3. Batch Transform.
- **Respuesta Correcta:** 2. Real-time Endpoint (Aprovis.).

Escenario B: MLOps y Despliegue Condicional

Descripción: Has entrenado un nuevo modelo y solo debe pasar a despliegue automático si su **F1-score es superior al 85%**. Si no lo es, el *pipeline* debe detenerse[3].

- **Opciones de Orquestación:** 1. ProcessingStep, 2. ConditionStep, 3. Model Monitor.
- **Respuesta Correcta:** 2. ConditionStep.

Escenario C: Eficiencia de Costes e Intermitencia

Descripción: Estás desarrollando un modelo interno para el departamento de RRHH que predice la rotación de personal, pero solo se ejecuta de forma **esporádica o intermitente** (una o dos veces al mes).

- **Opciones de Inferencia:** 1. Real-time Endpoint (Aprovis.), 2. Serverless Inference, 3. Batch Transform.
- **Respuesta Correcta:** 2. Serverless Inference.

Escenario D: Detección de Drift en Producción

Descripción: Tu modelo de clasificación de spam ha sido desplegado, pero las nuevas campañas de *phishing* han cambiado la distribución de los correos de entrada (**Data Drift**). ¿Qué servicio **compara automáticamente** el tráfico en vivo con la línea base de entrenamiento para generar una alerta?

- **Opciones de Monitorización:** 1. SageMaker Debugger, 2. SageMaker Model Monitor, 3. CloudWatch Logs.
- **Respuesta Correcta:** 2. SageMaker Model Monitor.

Escenario A: Contenedores y Algoritmos

Pregunta sobre Opciones de Entrenamiento

Tienes un modelo para clasificar comentarios de clientes utilizando un algoritmo de procesamiento de lenguaje natural llamado **BlazingText**. Quieres entrenarlo a gran escala con SageMaker, pero no tienes tiempo para escribir código de entrenamiento ni configurar un contenedor. ¿Cuál de las opciones de entrenamiento de SageMaker es la más adecuada para ti y por qué?

- **Solución:** Se utilizaría la opción de **Algoritmos Integrados (*Built-in Algorithms*)**.
    - **Justificación:** BlazingText es uno de los algoritmos integrados de SageMaker. Estos algoritmos están **ya implementados y optimizados** para AWS, por lo que no necesitas escribir el código de entrenamiento, solo configurar los hiperparámetros.

Pregunta sobre Contenedores y Flexibilidad

Tu equipo ha creado una librería de *Deep Learning* propietaria que no es ni TensorFlow ni PyTorch. ¿Cómo resuelves el desafío de la **Reproducibilidad** y la **Escalabilidad** para entrenar este modelo en la nube sin tener que reescribir todo el código?

- **Solución:** Se utiliza el principio **Bring Your Own Container (BYOC)**.
    - **Justificación:** BYOC te da la **máxima flexibilidad** al permitirte empaquetar tu entorno de entrenamiento completo (todas las librerías específicas) en un **contenedor Docker propio** y pedirle a SageMaker que lo ejecute. Esto garantiza que el entorno de entrenamiento sea idéntico cada vez.

Escenario B: MLOps y Desafíos Fundamentales

Pregunta sobre Data-Serving Skew

Tu modelo de predicción de *churn* (abandono de clientes) fue muy preciso durante las pruebas, pero empezó a fallar tan pronto como fue desplegado. El análisis indica que la *feature* "antigüedad del cliente" se calculó de forma diferente en el

entorno de entrenamiento y en el entorno de producción. ¿Qué desafío de MLOps es este y qué servicio de SageMaker se utiliza para prevenirlo?

- **Solución:** El desafío es el **Training-Serving Skew**.
    - **Prevención:** Se utiliza **SageMaker Feature Store**. Feature Store resuelve el problema al proporcionar un **repositorio centralizado** para almacenar y recuperar *features* de manera **consistente**, usando la **Online Store** para inferencia en producción y la **Offline Store** para el entrenamiento.

Pregunta sobre Plataformas Gestionadas

¿Cuál es la principal diferencia entre usar una **Instancia EC2** (IaaS) con una GPU para el entrenamiento manual y usar una plataforma **SageMaker Training Job** (PaaS) para el mismo entrenamiento, en términos de esfuerzo de Ingeniería?

- **Solución:** La principal diferencia es la **Abstracción de la Infraestructura**.
    - **IaaS (EC2):** Requiere configuración manual de sistema operativo, drivers CUDA, red, balanceo de carga, y gestión del *cluster*.
    - **PaaS (Training Job):** Abstrae completamente esa gestión. El desarrollador solo define el tipo de instancia y el contenedor, y SageMaker se encarga de provisionar, configurar y liberar el hardware automáticamente.

Preguntas Clave de la Sesión 9 (Despliegue, Monitorización y Orquestación)

Bloque 1: Despliegue e Inferencia (SageMaker Endpoints)

Pregunta de Componentes y Reproducibilidad

El artefacto **model.tar.gz** es el resultado de un *Training Job* exitoso y la unidad para el despliegue. ¿Cuáles son los tres archivos esenciales que contiene y por qué el **requirements.txt** es crucial para la **Reproducibilidad** en el *endpoint*?

- **Solución:** Los tres archivos esenciales son: 1. Archivos del modelo (pesos y arquitectura como .pkl, .pb, etc.), 2. Código de inferencia (inference.py), y 3. Requerimientos (requirements.txt). El requirements.txt es fundamental porque lista las dependencias exactas que necesita el código de inferencia para funcionar correctamente en el servidor.

Pregunta de Caso de Uso de Tareas Pesadas

Una firma de arquitectura necesita procesar **500 modelos 3D de alta resolución** cada noche. Cada uno tarda 15 minutos en analizarse. ¿Qué modo de inferencia es el más adecuado y por qué **no** usaría el *Real-time Endpoint*?

- **Solución:** El modo adecuado es la **Inferencia Asíncrona**. Es ideal para **payloads grandes** (hasta 1 GB) y **tiempos de procesamiento largos**. El *Real-time Endpoint* no es adecuado porque está optimizado para **baja latencia** y no para manejar tareas que tardan minutos.

Pregunta de Variantes de Modelos

¿Qué característica de los **SageMaker Real-time Endpoints** permite a las organizaciones realizar un **A/B Testing** de un nuevo modelo (Candidato B) contra la versión actual (Modelo A) sin cambiar la dirección de la API?

- **Solución:** Se utilizan las **Variantes de Modelos**. Esto permite desplegar **múltiples modelos** en el mismo *endpoint* y **ajustar automáticamente la asignación del tráfico** (ej. 90% a A, 10% a B).

Bloque 2: Monitorización, Orquestación y Comparativa Cloud

Pregunta de Model Monitor y Data Drift

¿Cómo utiliza **SageMaker Model Monitor** la **Línea Base (*Baseline*)** para detectar el **Data Drift** en producción?

- **Solución:** Durante el entrenamiento, se crea una **Línea Base** de estadísticas. Periódicamente, un trabajo de monitorización **compara** las estadísticas del **tráfico en vivo** capturado desde el *endpoint* con esa Línea Base. Si hay una desviación en la distribución de datos de entrada, lanza una **Alerta en CloudWatch**.

Pregunta de Orquestación y Gobernanza

¿Cuál es el rol del `ConditionStep` y el `RegisterModelStep` dentro de un **SageMaker Pipeline**, y qué desafío de MLOps resuelven juntos?

- **Solución:**

    - El `ConditionStep` se utiliza para **tomar decisiones basadas en métricas** (ej. si el F1-score es > 0.85).
    - El `RegisterModelStep` **registra la nueva versión** en el *Model Registry* **solo si** la condición es verdadera.
    - Resuelven el desafío de la **Gobernanza Automatizada**, asegurando que solo modelos aprobados pasen al despliegue.

Pregunta de Estrategia de Plataforma Cloud

Un equipo valora una interfaz de usuario **muy unificada** y necesita la infraestructura de **TPUs** para *deep learning* a gran escala. ¿Qué plataforma, alternativa a SageMaker, es la más adecuada?

- **Solución:** La plataforma más adecuada es **GCP Vertex AI**. Su filosofía se centra en ofrecer una **plataforma unificada** y da fácil acceso a las potentes **TPUs (Tensor Processing Units)**.

Actividad de Gamificación: Sesión 9 (Despliegue, Monitorización y Orquestación)

Escenario A: Despliegue de Alto Tráfico y Baja Latencia
Descripción: Eres una gran plataforma de *e-commerce* y necesitas un modelo de recomendación que responda en **milisegundos** a **cientos de miles de peticiones constantes** por hora, 24/7. Tu tráfico es alto y constante.

- **Opciones:** 1. Inferencia Asíncrona, 2. Real-time Endpoint (Aprovis.), 3. Batch Transform.
- **Respuesta Correcta:** 2. Real-time Endpoint (Aprovis.).
- **Justificación:** Se requiere **baja latencia** y tráfico constante, lo cual solo se logra con un *endpoint* **persistente** y aprovisionado 24/7.

Escenario B: MLOps y Despliegue Condicional Automatizado
Descripción: Has entrenado un nuevo modelo, y para el MLOps automatizado, solo debe pasar a despliegue si su **F1-score es superior al 85%**. Si no lo es, el *pipeline* debe detenerse. ¿Qué *Step* de **SageMaker Pipelines** se encarga de verificar esta métrica y controlar el flujo del DAG?

- **Opciones:** 1. EvaluationStep, 2. ConditionStep, 3. Model Monitor.
- **Respuesta Correcta:** 2. ConditionStep.
- **Justificación:** El **ConditionStep** se utiliza específicamente para **tomar decisiones basadas en métricas**. Si la métrica de rendimiento (del EvaluationStep) no se cumple, el *pipeline* no avanza al registro.

Escenario C: Eficiencia de Costes e Intermitencia
Descripción: Estás desarrollando un modelo interno para RRHH que predice la rotación de personal, pero solo se ejecuta de forma **esporádica o intermitente** (una o dos veces al mes). Necesitas pagar solo por el uso activo.

- **Opciones:** 1. Real-time Endpoint (Aprovis.), 2. Serverless Inference, 3. Batch Transform.
- **Respuesta Correcta:** 2. Serverless Inference.
- **Justificación:** Es la opción ideal para cargas de trabajo **intermitentes o esporádicas** porque **no tiene coste si no se usa** y se escala automáticamente de cero.

Escenario D: Detección de Drift en Producción
Descripción: Tu modelo de clasificación de spam ha sido desplegado, pero las nuevas campañas de *phishing* han cambiado la distribución de los correos de entrada (**Data Drift**). ¿Qué servicio **compara automáticamente** el tráfico en vivo capturado con la **línea base** de entrenamiento para generar una alerta en CloudWatch?

- **Opciones:** 1. SageMaker Debugger, 2. SageMaker Model Monitor, 3. CloudWatch Alarms.
- **Respuesta Correcta:** 2. SageMaker Model Monitor.
- **Justificación:** El **Model Monitor** es el servicio diseñado para **comparar** periódicamente el **tráfico en vivo** con la **línea base** de estadísticas de entrenamiento, detectando desviaciones de la calidad del modelo.

Escenario E: Procesamiento de Tareas Pesadas (Flujo Asíncrono)
Descripción: Una aplicación necesita analizar documentos extensos de hasta **1 GB de tamaño** y acepta que la predicción tarde unos minutos, siempre y cuando reciba una **notificación** cuando el resultado esté depositado en S3.

- **Opciones:** 1. Real-time Endpoint, 2. Batch Transform, 3. Inferencia Asíncrona.
- **Respuesta Correcta:** 3. Inferencia Asíncrona.
- **Justificación:** Es la única opción diseñada para manejar **payloads grandes (hasta 1 GB)** y tiempos de procesamiento largos a través de un flujo de **mensajería (SQS)** y **notificación (SNS)**.