Manual de Estudio: AWS Certified AI Practitioner

Dominio 4.2: Modelos Transparentes y Explicables

Material de Preparación Detallado

1. Transparencia: Interpretabilidad vs. Explicabilidad

La **transparencia** mide el grado en que las partes interesadas pueden entender el funcionamiento de un modelo y el porqué de sus resultados. Se divide en dos conceptos clave:

Característica	Interpretabilidad	Explicabilidad
Definición	Capacidad de entender los mecanismos internos y fórmulas del modelo.	Capacidad de describir qué hace el modelo sin conocer su funcionamiento interno exacto.
Enfoque	El modelo es "transparente"por diseño.	Trata al modelo como una **çaja negra"**.
Ejemplos	Regresión lineal ($y = mx + b$), árboles de decisión.	Redes neuronales profundas, LLMs.
Uso	Requisitos reglamentarios estrictos (ej. banca).	Objetivos empresariales generales (ej. detección de spam).

2. Compromisos de Diseño (Trade-offs)

Al seleccionar un modelo, el arquitecto de IA debe equilibrar tres factores críticos:

2.1. 1. Rendimiento vs. Transparencia

Existe una correlación inversa entre la complejidad de un modelo y su transparencia.

- **Modelos Simples:** Alta transparencia pero rendimiento limitado en tareas complejas (ej. traducción de idiomas fluida).
- **Modelos Complejos:** Bajo nivel de transparencia pero alto rendimiento (capturan interacciones no lineales).

[Image of the trade-off between AI model performance and interpretability]

2.2. 2. Seguridad vs. Transparencia

- **Vulnerabilidad:** Los modelos muy transparentes son más susceptibles a ataques, ya que los hackers pueden encontrar debilidades al conocer los mecanismos internos.
- **Ingeniería Inversa:** La exposición de algoritmos propietarios a través de explicaciones detalladas puede comprometer la propiedad intelectual.

3. Herramientas de AWS para la Transparencia

AWS proporciona documentación y servicios para auditar y explicar modelos:

3.1. Documentación y Auditoría

- **Tarjetas de Servicio de IA (AI Service Cards):** Documentación oficial de AWS para sus servicios pre-entrenados (Rekognition, Textract, Comprehend). Detalla casos de uso, limitaciones y diseño responsable.
- **Fichas del Modelo de SageMaker (Model Cards):** Documentación generada automáticamente que registra el ciclo de vida del modelo: entrenamiento, datasets y evaluación.

3.2. Análisis Técnico con SageMaker Clarify

- **Valores de Shapley:** Método para determinar la contribución de cada característica a una predicción específica (atribución de características).
- **Gráfico de Dependencia Parcial:** Muestra cómo cambia la predicción del modelo al variar un solo atributo (ej. cómo afecta la edad a la aprobación de un crédito).

[Image of Shapley values for model explainability]

4. IA Centrada en el Ser Humano

El diseño centrado en el humano prioriza las necesidades y valores de las personas, mejorando las capacidades humanas en lugar de reemplazarlas.

4.1. Revisión Humana con Amazon A2I

Amazon Augmented AI (A2I) incorpora el *"Human-in-the-loop"*:

- Envía inferencias con **puntuaciones de confianza bajas** a revisores humanos.
- Los comentarios humanos se usan para corregir resultados y re-entrenar modelos.
- Permite auditorías aleatorias para verificar la imparcialidad.

4.2. RLHF (Aprendizaje por Refuerzo con Feedback Humano)

Técnica para alinear LLMs con valores humanos:

1. Los humanos clasifican múltiples respuestas del modelo según su preferencia.
2. Se entrena un **modelo de recompensa** basado en estas preferencias.
3. El LLM se refina usando el modelo de recompensa para producir contenido veraz e inofensivo.

[Image of RLHF reinforcement learning from human feedback workflow]