CA3: Diseño de una solución integral

1. Introducción y Objetivos

¡Bienvenidos! La sesión de hoy está diseñada para ir más allá de la simple implementación de algoritmos. El objetivo es que se enfrenten a problemas de machine learning realistas, complejos y abiertos, donde el verdadero desafío no es solo ejecutar el código, sino **diseñar, justificar y defender una estrategia de solución coherente**.

En la práctica profesional, la capacidad de analizar un problema, identificar sus dificultades inherentes, proponer un plan de ataque y comunicar la lógica detrás de sus decisiones es tan importante como la precisión final del modelo. Esta actividad se centra en desarrollar precisamente esa habilidad estratégica.

Dinámica de la Actividad:

- **Hoy (Jueves):** Se organizarán en equipos y se les asignará uno de los cuatro retos propuestos. El objetivo principal de esta sesión es realizar un análisis exploratorio del problema, definir una estrategia de solución y documentarla.
- **Lunes (Presentaciones):** Cada equipo presentará su propuesta de solución en una exposición breve y concisa.

2. Challenges

A continuación se presentan cuatro escenarios. Cada uno aborda un desafío fundamental y común en la ciencia de datos. Lean cada uno con atención y elijan el que más les interese como equipo.

Challenge 0: Navegando el Laberinto de Características

Escenario

Ustedes son científicos de datos en una firma de análisis de medios. Su equipo busca comprender los factores que impulsan la popularidad de las publicaciones en blogs. Su tarea es construir un modelo predictivo que pueda estimar con precisión el número de comentarios que una publicación de blog recibirá en las siguientes 24 horas. Este modelo ayudará a los creadores de contenido a optimizar sus publicaciones para maximizar la interacción de la audiencia.

Conjunto de Datos

Se les proporciona el conjunto de datos **UCI BlogFeedback**. Este conjunto de datos contiene 60,021 instancias, cada una representando una publicación de blog. Para cada publicación, se dispone de 280 características predictoras y una variable objetivo. Las características incluyen metadatos sobre el blog, características basadas en el texto e información temporal. La variable objetivo es el número de comentarios que la publicación recibió en las 24 horas posteriores a un "tiempo base" de referencia.

El Reto Central

La dificultad principal de esta tarea no reside en la elección del algoritmo de regresión, sino en la gestión del vasto y heterogéneo espacio de características. Con 280 predictores, es muy probable que muchos sean redundantes, irrelevantes o estén altamente correlacionados. Un modelo que utilice ingenuamente todas las características será ineficiente, difícil de interpretar y propenso al sobreajuste.

Su misión es desarrollar y justificar una estrategia completa para manejar este espacio de características de alta dimensionalidad. No hay un único camino correcto; el valor de su trabajo radicará en la lógica y la justificación de sus decisiones.

Challenge 1: Encontrando la Aguja en el Pajar

Escenario

Ustedes trabajan para una importante institución financiera en el departamento de prevención de fraudes. Su principal responsabilidad es proteger a los clientes y a la empresa de transacciones fraudulentas con tarjetas de crédito. Se les ha proporcionado un conjunto de datos de transacciones donde, como es de esperar, los casos de fraude son extremadamente raros.

Conjunto de Datos

Se les proporciona el conjunto de datos de **Detección de Fraude con Tarjetas de Crédito**. Este conjunto de datos contiene más de 550,000 transacciones, de las cuales solo 492 son fraudulentas. La clase positiva (fraude) constituye aproximadamente el 0.172% del total. Las características predictoras (V1 a V28) son componentes principales anónimos para proteger la privacidad, junto con las características 'Time' y 'Amount'.

El Reto Central

Un modelo que simplemente prediga "no es fraude" para cada transacción logrará una precisión superior al 99.8%, pero sería completamente inútil. Este es el problema central de la clasificación con datos desequilibrados.

Su tarea es desarrollar un sistema de machine learning que sea genuinamente efectivo para identificar transacciones fraudulentas. Esto requiere un enfoque de dos pasos:

1. Definir y justificar qué significa "efectivo" en este contexto, seleccionando métricas de evaluación apropiadas que capturen el rendimiento en la clase minoritaria.
2. Diseñar e implementar un pipeline de modelado que aborde explícitamente el severo desequilibrio de clases.

Challenge 2: Abriendo la Caja Negra

Escenario

Ustedes son científicos de datos en una empresa de telecomunicaciones. La dirección está preocupada por la tasa de abandono de clientes (churn) y les ha encargado no solo predecir qué clientes tienen más probabilidades de irse, sino, más importante, entender *por qué*. El objetivo es diseñar campañas de retención efectivas basadas en estos conocimientos.

Conjunto de Datos

Se les proporciona el conjunto de datos **Telco Customer Churn**. Este conjunto de datos incluye información sobre alrededor de 7,000 clientes, con 21 características que describen sus datos demográficos, servicios contratados, información de la cuenta y si finalmente abandonaron la empresa.

El Reto Central

Existe una tensión inherente entre la precisión de un modelo y su interpretabilidad. Los modelos más complejos (cajas negras) suelen ser más precisos, pero difíciles de entender. Los modelos simples (cajas blancas) son fáciles de interpretar, pero pueden ser menos precisos.

Su misión es navegar esta tensión. Deben construir un modelo de clasificación de alto rendimiento y luego utilizar técnicas de Explicabilidad de la IA (XAI) para "abrir la caja negra" y extraer información de negocio procesable. Su éxito se medirá tanto por la precisión del modelo como por la profundidad y claridad de las explicaciones que puedan proporcionar.

3. Entregables y Cronograma

La evaluación de esta actividad se divide en dos partes principales, complementadas por una coevaluación entre equipos.

Entregable 1 (Jueves): Propuesta de Solución

Este es el principal resultado de la sesión de 3 horas. Deben entregar un único documento que contenga su plan de solución.

- **Formato:** Libre elección (Jupyter Notebook, Google Colab, Deepnote, Markdown, Google Docs, etc.).
- **Contenido Esencial:**
  1. **Equipo:** Nombres de los integrantes y el título del reto asignado.
  2. **Análisis Exploratorio de Datos (EDA) Inicial:** Presenten sus hallazgos iniciales sobre el conjunto de datos. ¿Cuáles son las características principales del problema? (ej. distribución de la variable objetivo, correlaciones, valores faltantes, dimensionalidad, desequilibrio, etc.). Incluyan visualizaciones clave que respalden su análisis.
  3. **Definición de la Estrategia:** Describan paso a paso el plan que proponen para abordar el reto. Sean específicos.
     - ¿Qué métricas de evaluación utilizarán y por qué son las más adecuadas para este problema?
     - ¿Qué pasos de preprocesamiento son necesarios?
     - ¿Qué técnicas o modelos planean comparar? (ej. Lasso vs. PCA, SMOTE vs. `class_weight`, t-SNE vs. UMAP, Regresión Logística vs. XGBoost, SHAP?).
  4. **Justificación de la Estrategia:** Esta es la parte más importante. Expliquen *por qué* han elegido esa estrategia. ¿Qué ventajas esperan obtener? ¿Qué desventajas o compensaciones (trade-offs) han considerado? ¿Por qué su enfoque es adecuado para el problema central del reto?

Entregable 2 (Lunes): Presentación de la Solución

El lunes, cada equipo realizará una presentación oral de su propuesta y los resultados preliminares obtenidos.

- **Formato:** Presentación de diapositivas (Google Slides, PowerPoint, etc.).
- **Duración:** 7 minutos de exposición + 2 minutos para preguntas.
- **Contenido:** La presentación debe ser un resumen ejecutivo de su documento del jueves. Debe explicar de forma clara y concisa el problema, la estrategia propuesta, la justificación de la misma y los resultados (aunque sean preliminares) que hayan podido obtener.
- **Regla Fundamental:** No se puede añadir contenido, análisis o propuestas significativamente diferentes a las que se documentaron en el entregable del jueves. La presentación es para exponer el plan que ya diseñaron.

4. Sistema de Evaluación

La nota final se compondrá de la siguiente manera:

- **Propuesta de Solución (Jueves):** 40% (8 puntos)
- **Presentación y Resultados (Lunes):** 40% (8 puntos)
- **Coevaluación Recibida de sus Compañeros:** 15% (3 puntos)
- **Participación en la Coevaluación a otros equipos:** 5% (1 punto)

A continuación se detallan las rúbricas que se utilizarán para cada componente.

Rúbrica 1: Propuesta de Solución (Jueves - 40%)

Criterio 1.1: Análisis Exploratorio de Datos (EDA)

Nivel	Descripción	Puntos
Excelente	Identifica y visualiza claramente las características críticas del dataset (ej. desequilibrio, multicolinealidad, dispersión) y conecta directamente estos hallazgos con la estrategia de solución propuesta.	3
Bueno	Identifica las características principales del dataset con visualizaciones adecuadas, pero la conexión con la estrategia es superficial o implícita.	2
Aceptable	Realiza un análisis descriptivo básico del dataset, pero no profundiza en las características que definen el reto central.	1
Deficiente	El análisis es muy limitado, con visualizaciones poco informativas, o no se presenta análisis.	0

Criterio 1.2: Diseño y Justificación de la Estrategia

Nivel	Descripción	Puntos
Excelente	Propone una estrategia clara, detallada y bien estructurada que aborda directamente el núcleo del reto. Justifica cada decisión con argumentos sólidos, considerando explícitamente las compensaciones (trade-offs).	3
Bueno	Propone una estrategia coherente que aborda el reto. La justificación es correcta, pero no explora en profundidad las alternativas o las compensaciones.	2
Aceptable	La estrategia propuesta es válida pero genérica, y no se adapta completamente a las particularidades del reto. La justificación es débil.	1
Deficiente	La estrategia es confusa, incompleta, inapropiada para el problema, o no se presenta.	0

Criterio 1.3: Claridad y Estructura del Documento

Nivel	Descripción	Puntos
Excelente	El documento está excepcionalmente bien organizado, es fácil de seguir y utiliza un lenguaje técnico preciso. El código y las visualizaciones están limpios y bien comentados.	2
Bueno	El documento está bien estructurado y es comprensible, pero puede tener áreas menores de mejora en la organización o en la claridad de las explicaciones.	1
Deficiente	El documento es muy difícil de entender, carece de estructura, presenta la información de manera confusa, o no se entrega.	0

Rúbrica 2: Presentación de la Solución (Lunes - 40%)

Criterio 2.1: Claridad y Coherencia de la Exposición

Nivel	Descripción	Puntos
Excelente	Expone el problema, la estrategia y los resultados de forma extremadamente clara, lógica y concisa. La narrativa es fluida y fácil de seguir para toda la audiencia.	3
Bueno	Expone los puntos clave de manera clara. La presentación es coherente en su mayor parte, aunque puede haber pequeñas transiciones confusas.	2
Aceptable	La exposición es comprensible, pero carece de una estructura clara o una narrativa conductora. Algunos puntos son ambiguos.	1
Deficiente	La presentación es desorganizada, difícil de seguir, o no se realiza.	0

Criterio 2.2: Profundidad del Análisis y Resultados

Nivel	Descripción	Puntos
Excelente	Demuestra un dominio completo del problema, justificando la estrategia con confianza y presentando resultados (incluso preliminares) que están bien interpretados en el contexto del reto.	3
Bueno	Presenta la estrategia y los resultados de forma correcta. La interpretación de los hallazgos es adecuada pero podría ser más profunda.	2
Aceptable	Menciona la estrategia y los resultados, pero la justificación o la interpretación son superficiales o carecen de rigor.	1
Deficiente	Los resultados presentados son irrelevantes, incorrectos, no se conectan con la estrategia, o no se presentan.	0

Criterio 2.3: Gestión del Tiempo y Calidad de las Respuestas

Nivel	Descripción	Puntos
Excelente	Se ajusta perfectamente al tiempo asignado (7 min). Las respuestas a las preguntas son directas, precisas y demuestran una comprensión profunda del tema.	2
Bueno	Se ajusta razonablemente al tiempo. Las respuestas a las preguntas son correctas, aunque podrían ser más concisas o detalladas.	1
Deficiente	Excede o no utiliza significativamente el tiempo asignado, o las respuestas a las preguntas son evasivas, superficiales o incorrectas.	0

Rúbrica 3: Coevaluación (20%)

Criterio 3.1: Coevaluación Recibida

Este componente se calculará a partir del promedio de las puntuaciones otorgadas por los otros equipos durante la fase de coevaluación. Se utilizará una escala simple para la evaluación entre pares y el resultado se normalizará a un máximo de 3 puntos.

Criterio 3.2: Participación en la Coevaluación (Máximo 1 punto)

Nivel	Descripción	Puntos
Completo	Proporciona feedback constructivo, específico y respetuoso a todos los equipos asignados, demostrando una escucha activa y un análisis crítico.	1
Incompleto	Proporciona feedback genérico, superficial, o no evalúa a todos los equipos asignados.	0.5
Nulo	No participa en el proceso de coevaluación.	0