Bootcamp Data Science 3 - Ejercicio semana 8 – Gerardo Rodríguez

Situación:

Eres un científico de datos en una empresa de comercio electrónico. La empresa ha recolectado una gran cantidad de datos sobre las interacciones de los usuarios con el sitio web durante los últimos años, generando terabytes de datos. Tu tarea es analizar estos datos para identificar patrones de compra y, en base a ello, mejorar las recomendaciones de productos a los usuarios.

1. Big Data: Dada la gran cantidad de datos recolectados, ¿cuál considerarías una herramienta adecuada para almacenar y procesar estos datos: una base de datos relacional tradicional o una plataforma como Hadoop o Spark? Justifica tu elección.

En este escenario son varias las razones para recomendar el uso de estas herramientas cuyo su foco justamente es el Big Data, la base de datos tradicional tendría diferentes desventajas:

a) Escalabilidad, las tecnologías big data permiten crecer la capacidad de procesamiento utilizando múltiples servidores interconectados.
b) Arquitectura, permite acceder a datos no estructurados como aquellos que vienen de texto, imágenes, video, o datos de sensores. O incluso también con datos estructurados.
c) Procesamiento en paralelo y distribuido, al poder acceder a múltiples nodos, las tecnologías como Haddop y Spark pueden aumentar su capacidad de procesamiento y la velocidad en que procesan información ya que pueden operar de manera independiente al no depender de una unidad central que coordine las operaciones.
d) Costo, las bases tradicionales pueden requerir de hardware y software costoso, las nuevas herramientas de big data pueden utilizar hardware menos especializado y por tanto de menor costo.
e) Tolerancia a fallos, las nuevas herramientas pueden encapsular el problema con alguno de los clusters y reasignar la tarea a otro nodo, esto permite que el proceso no se detenga.

2. Reducción de la Dimensionalidad: Una vez que tienes los datos listos para el análisis, te das cuenta de que hay cientos de características para cada usuario (edad, género, historial de navegación, historial de compras, etc.). ¿Qué técnica usarías para reducir la dimensionalidad de estos datos y por qué?

En especifico para este escenario, la mejora forma de reducir la dimensionalidad sería PCA (Análisis de componentes principales), algunas de las causas son las siguientes:

- El proceso permite reducir las variables o dimensiones que se utilizan, conservando la mayor parte de la información original, identificando los componentes de los datos que generan mayor variabilidad.
- Mantiene la mayor cantidad de datos diferentes para mantener la relevancia de la información.
- Elimina las variables que generan correlaciones, lo que limita el exceso de redundancia, esto se logra utilizando nuevas variables, también llamadas componentes principales. Parte del éxito de este proceso, es que las nuevas variables son ortogonales entre si.

- Simplificación del modelo, al reducir las variables redundantes se tiene una mayor simplicidad de la información, esto ayuda a mejorar la interpretación y a mejorar la predictividad.

3. Modelado: Quieres construir un modelo que prediga si un usuario compraría un producto basado en sus características. ¿Qué tipo de modelo usarías: un modelo de regresión o de clasificación? Justifica tu elección.

Lo ideal sería usar Clasificación, ya que buscamos un valor, en este caso binario si compra o no compra.

De elegir una regresión, estaríamos buscando un valor numérico.

4. Ajuste del Modelo: Una vez que hayas decidido el tipo de modelo, ¿qué técnica utilizarías para asegurarte de que tu modelo está bien ajustado y no está sobreajustado o subajustado?

Es posible usar varias técnicas de manera conjunta para evitar que el modelo este bien ajustado y aplique de manera general a datos no vistos.

Algunas recomendaciones son dividir los datos entre entrenamiento y prueba para analizar como se comporta con datos no vistos. Y utilizar la validación cruzada para entrenar el modelo en subconjuntos, esto reduce el riesgo de sobreajuste.

5. Evaluación del Modelo: Finalmente, antes de implementar tu modelo en producción, necesitas evaluar su rendimiento. ¿Qué métricas usarías para evaluar un modelo de clasificación y por qué?

Algunos ejemplos son: Precisión, Sensibilidad, Especificidad, y Área bajo la curva.

Estos indicadores permiten evaluar el modelo en relación con los casos verdaderos (positivos y negativos) para tener una ide más clara de su eficacia al momento de clasificar.