Instituto Tecnológico Autónomo de México

Diplomado: Ciencia de Datos y Machine Learning Aplicado a Finanzas

Módulo 3: Ciencia de Datos

Tema

SMOTE: Synthetic Minority Over-sampling Technique

Nombre: Denzel Arik Martinez Maldonado

En la práctica real hay ocasiones donde la información puede parecer suficiente para entender los diferentes fenómenos que los datos quieren explicar, pero muchas veces esta información, si bien explica algunos fenómenos muy bien, su contraparte puede resultar difusa o difícil de explicar. Por lo que, para esta contraparte, lo ideal sería obtener más datos que puedan explicar este otro fenómeno contenido en los datos, pero muchas veces esto no es alcanzable por varias razones, como pueden ser: alto costo de recursos para obtener nueva información o falta de tiempo para tomar acciones.

A esta falta de información se le conoce como desbalance de clases, donde cada clase es propia de un fenómeno o escenario diferente. Al traducir la información, estas clases pueden tener diferentes relevancias en el mundo real: no es el mismo coste humano intentar predecir quién no tiene cáncer que quién sí lo tiene; para una empresa no es el mismo coste de recursos otorgar dinero a personas que sí pagan que a personas con comportamiento moroso o fraudulento. Entonces, dentro de las clases habrá una de mayor relevancia dependiendo de su coste e importancia y, muchas, si no es que todas las veces, la clase que más coste tiene es la que menos información tiene y, por ende, la más difícil de explicar.

Para esto se han experimentado diversas formas para hacer que la información "sea suficiente" y poder explicar todos los escenarios que la información tiene contenida.

- **Under-sampling**
  El under-sampling (sub-muestreo) es una técnica en la que a la clase con más información, también llamada clase mayoritaria, se le "reduce" su información para así intentar discernir mejor entre los diferentes fenómenos que se presenten usando muestreo aleatorio para reducirla. Sin embargo, justamente al reducir su información, resulta en un menor poder explicativo de la clase mayoritaria; esta pérdida de poder explicativo estará determinada en qué tanto se sub-muestree la clase.

- **Over-sampling**
  El over-sampling (sobre-muestreo) es una técnica en la que a la clase con menor información, también llamada clase minoritaria, se le "aumenta" su información, pero aquí, a diferencia de under-sampling, no necesariamente el "aumentar" la información aumenta su poder explicativo, y esto se debe a la manera en la que se aumenta. En el over-sampling, del mismo modo que el under-sampling, se hace un muestreo aleatorio duplicando casos, lo que lleva a un sobreajuste: al usar clasificadores, estos, en lugar de "entender" el fenómeno, lo memorizan y, cuando llegan datos nuevos, no los clasifican bien.

Estas dos técnicas son las pioneras en la facilitación para poder explicar diversos fenómenos dentro de un cúmulo de datos, pero, como dijimos, aumentar o reducir la información no necesariamente aumenta el poder explicativo y, por ende, la capacidad de discernir los distintos escenarios.

Ante esto, se han implementado nuevas metodologías que intentan, mediante el aumento de información, aumentar también su poder explicativo aplicando estrategias sobre cómo se aumenta esta información. Es aquí donde surge SMOTE (Synthetic Minority Over-sampling Technique).

- **SMOTE**
  SMOTE es una técnica que, a diferencia del muestreo aleatorio simple que duplica casos y puede inducir sobreajuste, toma vecinos cercanos y los une mediante rectas; estas rectas se usan para generar la nueva información. En lugar de duplicar casos, genera "nuevos" casos en la dirección de las rectas creadas por los vecinos cercanos

de la clase minoritaria. La generación de estos sobre esta recta se hace mediante una función de probabilidad, asegurando que estos nuevos casos no se salgan de esta recta.

Ahora bien, una parte muy importante en este tipo de problemas no solo es "balancear" la información, sino también medir correctamente si lo que hicimos realmente ayudó. En datos desbalanceados es muy común que un modelo parezca bueno con una métrica, pero en la práctica sea malo para lo que nos interesa. Para esto, se usan métricas que se basan en la matriz de confusión, que separa lo que el modelo predice en cuatro casos: verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN).

- **Accuracy (exactitud)**
  La accuracy se define como
  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$
  Esta métrica funciona bien cuando las clases están balanceadas y cuando equivocarse en una u otra clase cuesta más o menos lo mismo. Sin embargo, cuando hay desbalance, la accuracy puede engañar fácilmente porque un modelo puede "acertar" mucho prediciendo casi siempre la clase mayoritaria, pero fallando justo en la clase importante.

- **Recall (sensibilidad)**
  El recall se define como
  $$\text{Recall} = \frac{TP}{TP + FN}.$$
  Esta métrica mide, de todos los casos que sí eran de la clase minoritaria, cuántos detectó el modelo. Es especialmente útil cuando lo más importante es no dejar pasar casos relevantes, por ejemplo: no dejar pasar fraude o no dejar pasar un diagnóstico positivo.

- **Precision (precisión)**
  La precisión se define como
  $$\text{Precision} = \frac{TP}{TP + FP}.$$
  Esta métrica mide qué tan "confiable" es una alerta del modelo: de todo lo que el modelo dijo que era minoritario, qué porcentaje realmente lo era. Es útil cuando lo costoso es generar falsas alarmas, por ejemplo: bloquear clientes buenos, mandar a revisión demasiadas transacciones, o saturar un área de control.

- **F1-score**
  En muchas ocasiones no quieres enfocarte solo en recall o solo en precision, sino un balance. El F1-score es una forma de combinar ambos:
  $$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$
  Funciona bien cuando quieres un equilibrio entre detectar casos importantes y no generar demasiados falsos positivos.

- **ROC y AUC**
  La curva ROC se construye con la tasa de verdaderos positivos y la tasa de falsos positivos:
  $$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}.$$

La idea es que no siempre existe un solo punto "correcto" de operación: depende de qué tanto estés dispuesto a aceptar falsos positivos a cambio de detectar más verdaderos positivos. El AUC resume esa curva en un solo número y suele ser útil cuando quieres comparar modelos de manera general.

- **Cuándo una métrica puede no servir tanto**
  La accuracy es la más problemática en desbalance. En el caso de ROC/AUC, aunque es una herramienta muy usada, puede dar una impresión optimista cuando la clase minoritaria es extremadamente rara, porque un FPR pequeño puede significar muchos falsos positivos en términos absolutos. Por eso, en casos de eventos muy raros, suele ser importante revisar también precision y recall, ya que directamente reflejan si las alertas del modelo son útiles o si solo generan ruido.

- **Under-Sampling + SMOTE**
  Puede mejorar el modelo bajo ciertas condiciones. La intuición es la siguiente: el under-sampling reduce el dominio numérico de la clase mayoritaria y SMOTE "fortalece" la clase minoritaria generando más estructura en su región. Cuando ambas clases quedan más comparables, el modelo deja de estar tan sesgado hacia la clase mayoritaria y puede aprender mejor una frontera de decisión que sí separe fenómenos.

Sin embargo, este beneficio depende de condiciones prácticas:

- **Cuando tiende a ayudar**
  Suele ayudar cuando la clase minoritaria tiene cierta estructura, cuando hay vecinos cercanos razonables y cuando las variables permiten medir cercanía de manera coherente. También ayuda cuando la clase mayoritaria es enorme y redundante, porque al sub-muestrear no se pierde tanto, y SMOTE puede aportar diversidad en la minoritaria.

- **Cuando puede empeorar**
  Puede empeorar cuando hay mucho **solapamiento** entre clases, porque al generar nuevos puntos sintéticos podrías crear ejemplos que caen en regiones más parecidas a la clase mayoritaria, aumentando falsos positivos. También puede fallar si la clase minoritaria tiene mucho ruido o outliers, ya que SMOTE podría generar ejemplos sintéticos alrededor de observaciones que en realidad no representan bien el fenómeno.

- **Qué se debe cuidar en la práctica**
  No existe un porcentaje "mágico" de under-sampling ni de SMOTE. La mejora suele aparecer cuando se ajustan ambos de forma gradual y se valida con las métricas correctas. Además, es muy importante aplicar estas técnicas solo en el conjunto de entrenamiento y evaluar en datos no modificados, porque si se mezcla esta generación en el conjunto de prueba se puede obtener una evaluación engañosa.

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.