Instituto Tecnológico Autónomo de México

Diplomado: Ciencia de Datos y Machine Learning Aplicado a Finanzas

Módulo 3: Ciencia de Datos

Tema

Data Analysis

Nombre: Denzel Arik Martinez Maldonado

Protocolo de exploración de datos

En la práctica real, muchas veces uno se quiere ir directo a entrenar un modelo o correr una regresión, pero el problema es que si los datos vienen "sucios", el resultado también sale "sucio". Y no porque el método sea malo sino porque los supuestos que ese método necesita no se cumplen, o porque los datos tienen detalles escondidos como outliers colinealidad, dependencia temporal, etc., que distorsionan la historia.

La idea principal es que antes de ajustar cualquier modelo, hay que hacer una exploración, no solo ver la tabla. En el articulo proponen un protocolo tipo checklist para detectar problemas comunes y tomar decisiones prácticas.

- **1) ¿Hay outliers en Y o en las X?**
  Un outlier no es automáticamente malo. Puede ser:

  - Un error de captura, unidad equivocada o que algun sensor falló.
  - Un caso real extremo y entonces es información valiosa.

  Lo importante es que un solo outlier puede dominar una regresión o "empujar" un modelo hacia conclusiones falsas. Lo recomendable es graficar: Y y cada X con un grafico de dispersión, y también usar gráficas sencillas tipo boxplots/dotplots. Si el outlier es error, se corrige o se elimina justificando. Si es real, se considera robustez: transformación, modelos robustos, o revisar si falta una variable que explique ese extremo.

- **2) Relación entre Y y X**
  Mucha gente confunde regresión lineal con una relación explicada con una línea recta. Pero en la vida real, la relación puede ser curva por tramos o con saturación. Si uno mete una relación curvada como si fuera lineal, el modelo va a "inventar" conclusiones. Por eso antes se grafican Y vs X y se busca el patrón: ¿crece?, ¿disminuye?, ¿se curva?, ¿hay umbrales?, ¿hay forma de campana?

- **3) ¿La varianza es más o menos constante o cambia? (heterogeneidad)**
  A veces el promedio se comporta bien, pero la dispersión no. Ejemplo: para valores bajos de Y todo está juntito, pero para valores altos hay mucha variación. Esto rompe supuestos típicos y también rompe interpretaciones. La forma más práctica de verlo es con gráficos de residuales y comparar dispersión en distintos rangos. Si la varianza cambia mucho, se suelen considerar transformaciones, pesos, o modelos que permitan varianza no constante.

- **4) Normalidad**
  Mucha gente se obsesiona con "que todo sea normal", pero en realidad la normalidad no es el objetivo, es un supuesto en ciertos contextos. Además, los tests de normalidad pueden fallar por dos lados:

  - Con pocas observaciones, no detectan nada.
  - Con muchas observaciones, detectan "cualquier cosita" aunque no sea relevante.

  Por eso lo sano es usar gráficas como histogramas o QQ-plot y sentido práctico: ¿lo que estás viendo afecta de verdad la validez del modelo que quieres usar?

- **5) Colinealidad: cuando dos o más X te cuentan la misma historia**
  Este es de los problemas más traicioneros: cuando tienes variables explicativas que están muy correlacionadas entre sí, el modelo se confunde. No necesariamente predice peor,

pero las varianzas de los estimadores suben, los errores estándar se inflan, y se vuelve más difícil detectar efectos reales subiendo los p-values y aumentando errores tipo II.
Una forma clásica de detectarlo es con el **VIF** (Variance Inflation Factor). La estrategia práctica es identificar las variables más colineales (VIF alto) y eliminar o replantear, no solo por número sino también con criterio (cuál tiene más sentido, cuál es más fácil de medir, cuál es más interpretable).

- **6) Dependencia entre observaciones (temporal o espacial)**
  Un supuesto común es que las observaciones sean independientes. Si hay datos por tiempo como series, mediciones mensuales, etc. o por espacio como sitios cercanos, lo más probable es que haya dependencia.
  Si ignoras esto puedes terminar encontrando significancia donde no la hay, porque el modelo cree que tiene más información independiente de la que realmente tiene. Lo mínimo es graficar variables contra tiempo/espacio y buscar patrones, y si hay dependencia, considerar modelos que lo incorporen por ejemplo, efectos aleatorios o estructuras de correlación.

- **7) Ceros en multivariado y ceros inflados en conteos**
  En datos multivariados, los "dobles ceros"(dos especies ausentes al mismo tiempo, por ejemplo) pueden engañar ciertas medidas de similitud. Y en modelos de conteo como GLM, puede existir **zero inflation**: demasiados ceros que no encajan con una Poisson normal.
  Si eso pasa los parámetros se sesgan. En esos casos se considera un modelo más adecuado por ejemplo modelos con inflación de ceros o técnicas que sí tengan sentido con esa estructura.

El artículo insiste en separar dos cosas:

- **Explorar** para detectar problemas de datos y supuestos.
- **Buscar patrones** para sacar hipótesis y luego venderlas como confirmación.

Explorar puede ayudarte a generar ideas, pero si usas la misma exploración para afirmar conclusiones fuertes, te puedes engañar fácilmente. Lo correcto es: explorar, proponer hipótesis, y luego validar con un enfoque formal o con datos nuevos.

La principal enseñanza es que antes de ajustar modelos debes validar la distinta informacion provista para no sesgar al modelo porque si no revisas outliers, colinealidad, dependencia, forma de relación y estructura de ceros, el modelo puede darte una respuesta que pareciera correcta pero falsa.