Roger D. Peng y Elizabeth Matsui

The art of data science

Apuntes por FODE

Índice general

1

Data Analysis es arte

Imagina que le preguntas a un compositor cómo escribe sus canciones. Hay muchas herramientas a las que puede recurrir. Tenemos una comprensión general de cómo debe estructurarse una buena canción: cuánto tiempo debe ser, cuántos versos, tal vez haya un verso seguido de un coro, etc. En otras palabras, existe un marco abstracto para las canciones en general. De manera similar, tenemos la teoría musical que nos dice que ciertas combinaciones de notas y acordes funcionan bien juntas y otras combinaciones no suenan bien. Por muy buenas que puedan ser estas herramientas, en última instancia, el conocimiento de la estructura de la canción y la teoría musical por sí solo no es una buena canción. Se necesita algo más.

Todo es arte, por ende es importante darse cuenta de que el análisis de datos es un arte.

Los analistas de datos tienen muchas herramientas a su disposición, desde regresión lineal hasta árboles de clasificación e incluso aprendizaje profundo, y todas estas herramientas se han enseñado cuidadosamente a las computadoras. Pero, en última instancia, un analista de datos debe encontrar una manera de reunir todas las herramientas y aplicarlas a los datos para responder una pregunta relevante, una pregunta de interés para las personas. En 1991, Daryl Pregibon, un destacado estadístico anteriormente de AT& T Research y ahora de Google, dijo en referencia al proceso de análisis de datos2 que ılos estadísticos tienen un proceso que adoptan pero que no comprenden completamenteȷ.

Lo que nos hemos propuesto hacer en este libro es escribir el proceso de análisis de datos. Lo que describimos no es una "fórmula.^{es}pecífica para el análisis de datos, algo como .^{a}plicar este método y luego ejecutar esa prueba", sino que es un proceso general que se puede aplicar en una variedad de situaciones.

2

Epiciclos del análisis

En realidad, el análisis de datos es un proceso altamente iterativo y no lineal, mejor reflejado por una serie de epiciclos, en los cuales se aprende información en cada paso, que luego informa si (y cómo) refinar y rehacer, el paso que se acaba de realizar, o si (y cómo) continuar con el siguiente paso. Un epiciclo es un círculo pequeño cuyo centro se mueve alrededor de la circunferencia de un círculo más grande.

1. Las expectativas se desarrollan.

2. Recolectar Datos.

3. expectativas matemáticas con datos

Estos "pasos" se engranan con este otro epiciclo siguiente

1.- Planteando la pregunta.

2.- Análisis exploratorio de datos.

3.- Análisis exploratorio de datos.

4.- Interpretar.

5.- Comunicar.

2.1. Preparando la escena

Dado que un análisis de datos supone que los datos ya se han recopilado, incluye el desarrollo y el refinamiento de una pregunta y el proceso de análisis e interpretación de los datos. Es importante señalar que, aunque un análisis de datos a menudo se realiza sin realizar un estudio, también se puede realizar como un componente de un estudio.

2.2. epiciclo de análisis

Hay 5 actividades centrales del análisis de datos:

1. Formular y refinar la pregunta.

2. Explorar los datos.

3. Construir modelos estadísticos formales.

4. Interpretar los resultados.

5. Comunicar los resultados.

Estas 5 actividades pueden ocurrir en diferentes escalas de tiempo: por ejemplo, puede pasar por los 5 en el transcurso de un día, pero también tratar con cada uno, para un proyecto grande, en el transcurso de muchos meses.

Para cada una de las cinco actividades principales, es fundamental que participe en los siguientes pasos:

a) Establecer expectativas.

b) Recopilar información (datos), comparar los datos con sus expectativas y si las expectativas no coinciden.

c) Revisar sus expectativas o corregir los datos para que sus datos y sus expectativas coincidan.

La iteración a través de este proceso de 3 pasos es lo que llamamos el "epiciclo del análisis de datos".

	Establecer expectativas	**Recopilar información**	**Revisar expectativas**
Pregunta	pregunta de interés para la audiencia	búsqueda de literatura / expertos	agudizar la pregunta
Explorar datos	los datos son apropiados para preguntas	hacer gráficos exploratorios de datos	refinar la pregunta o llamar más datos
Modelo formal	modelo primario responde a la pregunta	ajustar modelos secundarios, análisis de sensibilidad	revisar el modelo formal para incluir más predictores
Interpretación	La interpretación de los análisis proporciona una respuesta específica y significativa a la pregunta.	interpretar la totalidad de los análisis centrándose en los tamaños del efecto y la incertidumbre	revisar EDA y / o modelos para proporcionar una respuesta específica e interpretable
Comunicación	El proceso y los resultados del análisis son entendidos, completos y significativos para la audiencia.	Buscar retroalimentación	revisar análisis o enfoque de presentación

2.3. Estableciendo expectativas

Desarrollar expectativas es el proceso de pensar deliberadamente en lo que espera antes de hacer algo, como inspeccionar sus datos, realizar un procedimiento o ingresar un comando. Por ejemplo averiguar el costo de una comida en un restaurant de lujo puede ser una expectativa.

2.4. Recopilando información

Los resultados de esa operación son los datos que necesita recopilar y luego determina si los datos que recopiló coinciden con sus expectativas. Para extender la metáfora del restaurante, cuando vas al restaurante, obtener el cheque es recopilar los datos.

2.5. omparación de expectativas con datos

Un indicador clave de qué tan bien va su análisis de datos es lo fácil o difícil que es hacer coincidir los datos que recopiló con sus expectativas originales.

2.6. Aplicación del proceso de Epicyle of Analysis

Antes de analizar un par de ejemplos, repasemos los tres pasos que se deben utilizar para cada actividad de análisis de datos básicos. Estos son:

1. Establecer expectativas.
2. Recopilar información (datos), comparar los datos con sus expectativas y, si las expectativas no coinciden.
3. Revisar sus expectativas o corregir los datos para que sus expectativas y los datos coincidan.

Los modelos estadísticos sirven para producir una formulación precisa de su pregunta para que pueda ver exactamente cómo desea usar sus datos, ya sea para estimar un parámetro específico o para hacer una predicción.
argumentaríamos que un buen análisis de datos requiere comunicación, retroalimentación y luego acciones en respuesta. Su análisis de datos trajo preguntas adicionales al frente, ya que esta es una característica de un análisis de datos exitoso.

3

Formular y perfeccionar la pregunta

Hacer análisis de datos requiere pensar bastante y creemos que cuando ha completado un buen análisis de datos, ha pasado más tiempo pensando que haciendo.

3.1. Tipos de preguntas

Los seis tipos de preguntas son:

1. Descriptivo.

2. Exploratorio.

3. Inferencial.

4. Predictivo.

5. Causal.

6. Mecanismo.

Una pregunta descriptiva es aquella que busca resumir una característica de un conjunto de datos. Los ejemplos incluyen determinar la proporción de hombres, el número medio de porciones de frutas y verduras frescas por día o la frecuencia de enfermedades virales en un conjunto de datos recopilados de un grupo de personas. No hay interpretación del resultado en sí, ya que el resultado es un hecho, un atributo del conjunto de datos con el que está trabajando.
Una pregunta exploratoria es aquella en la que analiza los datos para ver si existen patrones, tendencias o relaciones entre las variables. Estos tipos de análisis también se denominan análisis de ıgeneración de hipótesisȷ porque en lugar de probar una hipótesis como se haría con una pregunta inferencial, causal o mecanicista, se buscan patrones que respalden la propuesta de una hipótesis. Si tuviera la idea general de que la dieta estaba relacionada de alguna manera con enfermedades virales, podría explorar esta idea examinando las relaciones entre una variedad de factores dietéticos y enfermedades virales. Usted encuentra en su análisis exploratorio que los individuos que consumían una dieta alta en ciertos alimentos tenían menos enfermedades virales que aquellos cuya dieta no estaba enriquecida con estos alimentos, por lo que propone la hipótesis de que entre los adultos, comer al menos 5 porciones al día de fruta fresca y las verduras se asocia con menos enfermedades virales por año.
Una pregunta inferencial sería una reafirmación de esta hipótesis propuesta como una pregunta y se respondería analizando un conjunto diferente de datos, que en este ejemplo, es una muestra representativa de adultos en los EE. UU. Al analizar este conjunto

diferente de datos, ambos están determinando si la asociación que observó en su análisis exploratorio se mantiene en una muestra diferente y si se mantiene en una muestra que es representativa de la población adulta de EE. UU., Lo que sugeriría que la asociación es aplicable a todos los adultos en los Estados Unidos. En otras palabras, podrá inferir lo que es cierto, en promedio, para la población adulta en los EE. UU. A partir del análisis que realice en la muestra representativa.

Una pregunta predictiva sería aquella en la que se pregunta qué tipos de personas consumirán una dieta rica en frutas y verduras frescas durante el próximo año. En este tipo de preguntas, usted está menos interesado en lo que hace que alguien coma una dieta determinada, solo en lo que predice si alguien comerá esta dieta determinada. Por ejemplo, un ingreso más alto puede ser uno de los últimos factores de predicción, y es posible que no sepa (o ni siquiera le importe) por qué las personas con ingresos más altos tienen más probabilidades de comer una dieta rica en frutas y verduras frescas, pero lo más importante es que los ingresos son un factor que predice este comportamiento. Aunque una pregunta inferencial podría decirnos que las personas que consumen cierto tipo de alimentos tienden a tener menos enfermedades virales, la respuesta a esta pregunta no nos dice si comer estos alimentos provoca una reducción en el número de enfermedades virales, que sería la caso de una pregunta causal.

Una pregunta causal se refiere a si cambiar un factor cambiará otro factor, en promedio, en una población. A veces, el diseño subyacente de la recopilación de datos, de forma predeterminada, permite que la pregunta que hace sea causal. Un ejemplo de esto serían los datos recopilados en el contexto de un ensayo aleatorizado, en el que las personas fueron asignadas al azar a comer una dieta rica en frutas y verduras frescas o una que era bajo en frutas y verduras frescas. En otros casos, incluso si sus datos no son de un ensayo aleatorio, puede adoptar un enfoque analítico diseñado para responder una pregunta causal.

Finalmente, ninguna de las preguntas descritas hasta ahora conducirá a una respuesta que nos diga, si la dieta, efectivamente, causa una reducción en el número de enfermedades virales, cómo la dieta conduce a una reducción en el número de enfermedades virales. Una pregunta que pregunta cómo una dieta rica en frutas y verduras frescas conduce a una reducción en el número de enfermedades virales sería **una pregunta mecanicista**.

si un análisis de datos tiene como objetivo responder una pregunta inferencial, las preguntas descriptivas y exploratorias también deben responderse durante el proceso de respuesta a la pregunta inferencial. Para continuar con nuestro ejemplo de dieta y enfermedades virales, no saltaría directamente a un modelo estadístico de la relación entre una dieta alta en frutas y verduras frescas y el número de enfermedades virales sin haber determinado la frecuencia de este tipo de dieta y enfermedades virales. y su relación entre sí en esta muestra. Un segundo punto es que el tipo de pregunta que hace está determinado en parte por los datos disponibles (a menos que planee realizar un estudio y recopilar los datos necesarios para realizar el análisis). Por ejemplo, es posible que desee hacer una pregunta causal sobre la dieta y las enfermedades virales para saber si una dieta rica en frutas y verduras frescas provoca una disminución en el número de enfermedades virales, y el mejor tipo de datos para responder a esta pregunta causal es una en la que las dietas de las personas cambian de una rica en frutas y verduras frescas a una que no lo es, o viceversa. Si este tipo de conjunto de datos no existe, lo mejor que puede hacer es aplicar métodos de análisis a los datos de observación o, en cambio, responder a una pregunta inferencial sobre la dieta y las enfermedades virales.

3.2. Aplicar el epiciclo para formular y perfeccionar su pregunta

Ahora puede usar la información sobre los tipos de preguntas y las características de las buenas preguntas como guía para refinar su pregunta. Para lograr esto, puede iterar a

través de los 3 pasos de:

1. Establecer expectativas.
2. Recopilar información (datos), comparar los datos con sus expectativas y, si las expectativas no coinciden.
3. Revisar sus expectativas o corregir los datos para que sus expectativas y los datos coincidan.

Los modelos estadísticos sirven para producir una formulación precisa de su pregunta para que pueda ver exactamente cómo desea usar sus datos, ya sea para estimar un parámetro específico o para hacer una predicción.
argumentaríamos que un buen análisis de datos requiere comunicación, retroalimentación y luego acciones en respuesta. Su análisis de datos trajo preguntas adicionales al frente, ya que esta es una característica de un análisis de datos exitoso.

3.3. Caracteristicas de una buena pregunta

Para empezar, la pregunta debe ser de interés para su audiencia, cuya identidad dependerá del contexto y el entorno en el que esté trabajando con los datos. Si está en el mundo académico, la audiencia puede ser sus colaboradores, la comunidad científica, los reguladores gubernamentales, sus patrocinadores de Establecimiento y perfeccionamiento de la Pregunta 21 y / o el público. Si está trabajando en una startup, su audiencia es su jefe, el liderazgo de la empresa y los inversores.
Puede asegurarse de que su pregunta se basa en un marco plausible utilizando su propio conocimiento del área temática y haciendo un poco de investigación, que juntos pueden ser de gran ayuda en términos de ayudarlo a resolver si su pregunta se basa en un marco plausible .
La especificidad también es una característica importante de una buena pregunta. Un ejemplo de una pregunta general es: ¿Es mejor para usted llevar una dieta más saludable? Trabajar hacia la especificidad refinará su pregunta e informará directamente qué pasos tomar cuando comience a buscar datos. El proceso de aumento de la especificidad debería conducir a una pregunta final y refinada como: "¿Comer al menos 5 porciones al día de frutas y verduras frescas provoca menos infecciones del tracto respiratorio superior (resfriados)?"

3.4. Traducir una pregunta en un problema de datos

A medida que refina su pregunta, dedique algún tiempo a identificar los posibles factores de confusión y a pensar si su conjunto de datos incluye información sobre estos posibles factores de confusión.
Otro tipo de problema que puede ocurrir cuando se utilizan datos inapropiados es que el resultado no es interpretable porque la forma subyacente en la que se recopilaron los datos conduce a un resultado sesgado.

Las dos tareas principales que se debe abordar son:

1. pensar en cómo su pregunta cumple o no con las características de una buena pregunta y
2. determinar qué tipo de pregunta está haciendo para que tenga una buena idea.

buena comprensión de qué tipo de conclusiones se pueden (y no se pueden) sacar una vez finalizado el análisis de datos.

4

Análisis exploratorio de datos

El análisis de datos exploratorio más confiable consiste en visualizar datos utilizando una representación gráfica de los datos.
Hay varios objetivos del análisis de datos exploratorios, que son:

1. Para determinar si hay algún problema con su conjunto de datos.

2. Para determinar si la pregunta que está haciendo puede ser respondida por los datos que tiene.

3. Desarrollar un bosquejo de la respuesta a su pregunta.

Explorará los datos para determinar si hay problemas con el conjunto de datos y para determinar si puede responder a su pregunta con este conjunto de datos.
Es importante notar que aquí, nuevamente, se aplica el concepto de epiciclo de análisis. Debe tener una expectativa de cómo se verá su conjunto de datos y si su pregunta puede ser respondida por los datos que tiene. Si el contenido y la estructura del conjunto de datos no coinciden con sus expectativas, entonces deberá volver atrás y averiguar si sus expectativas eran correctas (pero hubo un problema con los datos) o, alternativamente, sus expectativas eran incorrectas, por lo que no puede usar el conjunto de datos para responder la pregunta y necesitará encontrar otro conjunto de datos. También debe tener alguna expectativa de cuáles serán los niveles de ozono, así como si el ozono de una región debe ser más alto (o más bajo) que el de otra.

4.1. Lista de verificación de análisis de datos exploratorios: un estudio de caso

En esta sección repasaremos una "lista de verificación" informal de cosas que hacer al embarcarse en un análisis de datos exploratorio. Como ejemplo continuo, usaré un conjunto de datos sobre los niveles de ozono por hora en los Estados Unidos para el año 2014. Los elementos de la lista de verificación son:

1. Formule su pregunta

2. Lea sus datos.

3. Verifique el empaquetado.

4. Mire la parte superior e inferior de sus datos.

5. Verifique sus “n” s.

6. Valide con al menos una fuente de datos externa.

7. Haga una gráfica.

8. Pruebe primero la solución fácil.

9. Haga un seguimiento.

4.1.1. Formule su pregunta

En particular, una pregunta o hipótesis aguda puede servir como una herramienta de reducción de dimensión que puede eliminar variables que no son inmediatamente relevantes para la pregunta.
Por lo general, es una buena idea dedicar unos minutos a averiguar cuál es la pregunta que realmente le interesa y reducirla para que sea lo más específica posible
una de las preguntas más importantes que puede responder con un análisis exploratorio de datos es "¿Tengo los datos correctos para responder esta pregunta?.[A] menudo, esta pregunta es difícil de responder al principio, pero puede volverse más clara a medida que revisamos y examinamos los datos.

4.1.2. Leer en sus datos

¿Alguna vez recibió un regalo antes del momento en que se le permitió abrirlo? Seguro, todos lo hemos hecho. El problema es que el presente está envuelto, pero deseas desesperadamente saber qué hay dentro. ¿Qué debe hacer una persona en esas circunstancias? Bueno, puede agitar un poco la caja, tal vez golpearla con los nudillos para ver si hace un sonido hueco, o incluso pesarla para ver qué tan pesado es. Así es como debe pensar en su conjunto de datos antes de comenzar a analizarlo de verdad.
Más importante aún, puede examinar las clases de cada una de las columnas para asegurarse de que estén correctamente especificadas (es decir, las letras numéricas son numéricas y las cadenas de caracteres, etc.)

4.1.3. Mire la parte superior e inferior de sus datos

A menudo, es útil mirar el ”principio” y el ”final” de un conjunto de datos inmediatamente después de comprobar el paquete. Esto le permite saber si los datos se leyeron correctamente, si las cosas están formateadas correctamente y si todo está ahí. Si sus datos son datos de series de tiempo, asegúrese de que las fechas al principio y al final del conjunto de datos coincidan con lo que espera que sean el período inicial y final.

4.1.4. ABC: Siempre revise sus "n"s

En general, contar cosas suele ser una buena forma de averiguar si algo está mal o no. En el caso más simple, si espera que haya 1,000 observaciones y resulta que solo hay 20, sabe que algo debe haber salido mal en alguna parte. Pero hay otras áreas que puede verificar según su aplicación.

4.1.5. Validar con al menos una fuente de datos externa

Es muy importante asegurarse de que sus datos coincidan con algo fuera del conjunto de datos. Le permite asegurarse de que las mediciones estén aproximadamente en línea con lo que deberían ser y sirve como una verificación de qué otras cosas podrían estar mal en su conjunto de datos.

4.1.6. Haga un gráfica

Hacer un diagrama para visualizar sus datos es una buena manera de comprender mejor su pregunta y sus datos. El trazado puede ocurrir en diferentes etapas de un análisis de datos. Para el trazado puede ocurrir en la fase exploratoria o más adelante en la fase de presentación / comunicación. Hay dos razones clave para realizar un gráfico de sus datos. Están creando expectativas y comprobando las desviaciones de las expectativas. En las primeras etapas del análisis, puede estar equipado con una pregunta / hipótesis, pero es posible que tenga poca idea de lo que está sucediendo en los datos. Es posible que haya echado un vistazo a algunos de ellos para hacer algunas comprobaciones de cordura, pero si su conjunto de datos es lo suficientemente grande, será difícil simplemente mirar todos los datos. Entonces, hacer algún tipo de gráfico, que sirva como resumen, será una herramienta útil para establecer expectativas sobre cómo deberían verse los datos. Una vez que tenga una buena comprensión de los datos, una buena pregunta / hipótesis y un conjunto de expectativas sobre lo que los datos deberían decir en relación con su pregunta, hacer un gráfico puede ser una herramienta útil para ver qué tan bien los datos coinciden con sus expectativas. . Los gráficos son particularmente buenos para permitirle ver desviaciones de lo que podría esperar. Por lo general, las tablas son buenas para resumir datos al presentar elementos como medias, medianas u otras estadísticas. Los gráficos, sin embargo, pueden mostrarle esas cosas, así como mostrarle cosas que están lejos de la media o la mediana, para que pueda verificar si se supone que algo está tan lejos. A menudo, lo que es obvio en una trama se puede ocultar en una tabla.

4.1.7. Pruebe primero la solución fácil

Es importante destacar que si no encuentra evidencia de una señal en los datos usando solo una gráfica o análisis simple, entonces a menudo es poco probable que encuentre algo usando un análisis más sofisticado.

Pon a prueba tu solución

Siempre debe pensar en formas de desafiar los resultados, especialmente si esos resultados concuerdan con sus expectativas anteriores. Recuerde que anteriormente notamos que tres estados tenían algunos valores inusualmente altos de ozono. No sabemos si estos valores son reales o no (por ahora, supongamos que son reales), pero podría ser interesante ver si el mismo patrón de este / oeste se mantiene si eliminamos estos estados que tienen actividad inusual.

4.1.8. Preguntas de seguimiento

En este punto, es útil considerar algunas preguntas de seguimiento.

1. **¿Tienes los datos correctos?** A veces, al final de un análisis de datos exploratorio, la conclusión es que el conjunto de datos no es realmente apropiado para esta pregunta de Análisis de datos exploratorios.

2. **¿Necesitas otros datos?** Si bien los datos parecían adecuados para responder la pregunta planteada, vale la pena señalar que el conjunto de datos solo cubrió un año (2014). Puede valer la pena examinar si el patrón este / oeste se mantiene durante otros años, en cuyo caso tendríamos que salir y obtener otros datos.

3. **¿Tienes la pregunta correcta?** En este caso, no está claro que la pregunta que intentamos responder tenga relevancia inmediata, y los datos realmente no indicaron nada para aumentar la relevancia de la pregunta.

El objetivo del análisis exploratorio de datos es hacer que piense en sus datos y razone sobre su pregunta. En este punto, podemos refinar nuestra pregunta o recopilar nuevos datos, todo en un proceso iterativo para llegar a la verdad

4.2. Uso de modelos para explorar sus datos