I+D+i y Fashion Tech: Revelando una curiosa relación estadística con los sistemas planetarios

Roberto Sanchis Ojeda
Data Scientist en Stitch Fix

En el mundo de las *startups*, y particularmente en el Sillicon Valley donde se encuentra la empresa para la que trabajo, el I+D+i es una necesidad vital en una continua competición por financiamiento y talento humano donde solo sobrevive el más fuerte. Y dentro de las diferentes vertientes del I+D+i, el estudio estadístico del comportamiento de nuestros clientes y la transferencia tecnológica desde otros campos científicos son dos pilares fundamentales. Por eso, quiero aprovechar la oportunidad de contribuir a esta edición de la revista Índice dedicada al I+D+i para que podáis descubrir como dos campos tan dispares como la moda y los sistemas planetarios pueden tener expresiones matemáticas muy parecidas.

BUSCANDO EXOPLANETAS MEDIANTE LA DETECCIÓN DE SUS TRÁNSITOS

En las últimas dos décadas científicos de todo el mundo han descubierto miles de planetas fuera de nuestro sistema solar. Una de las técnicas más exitosas para descubrirlos es la de los tránsitos, que se basa en detectar el descenso en la cantidad de luz que nos llega de una estrella cuando uno de sus planetas la oculta.

La técnica hace uso de que la órbita de cada planeta tiene una escala temporal asociada (el periodo orbital P), de tal manera que el planeta pasa por delante de la estrella cada P días. De esa manera, los tránsitos del planeta generan una señal periódica detectable cuya forma depende del tamaño relativo entre el planeta y la estrella, así como de la velocidad de la órbita y su orientación con respecto del ecuador de la estrella.

LA CONEXIÓN CON EL MUNDO DE LA MODA

Actualmente un gran número de industrias están siendo conquistadas por compañías que han sido capaces de obtener y estudiar la gran cantidad de datos que sus clientes generan. Este tipo de análisis está llegando al mundo de la moda, y en particular son esenciales en la compañía en la que trabajo, Stitch Fix. Nuestra compañía provee un servicio personalizado de ropa (actualmente solo disponible en EEUU), mediante el cual nuestros clientes reciben una serie de piezas de ropa en su casa recomendadas por un algoritmo que tiene en cuenta las preferencias de cada cliente. Las recomendaciones del algoritmo son revisadas y actualizadas por un ser humano (el estilista), que se encarga de ponerlas en contexto en una nota personalizada para que el cliente sepa cómo combinar la ropa con otras piezas que haya podido recibir en el pasado.

La conexión con los exoplanetas surge cuando uno quiere entender cómo las preferencias de los clientes cambian con el tiempo. Puesto que algunas de ellas se repiten cada año, estas pueden ser estudiadas de una manera análoga a los exoplanetas, para poder anticiparnos a ellas y planear nuestras decisiones de la manera más eficiente.

CONSTRUYENDO EL ESPECTRO DE FOURIER: EL CASO DE LOS EXOPLANETAS

Para estudiar la forma de una señal periódica no hay manera más elegante que construir el espectro de Fourier de dicha señal, que nos ayuda a identificar visualmente cuáles son las periodicida-

Actualmente un gran número de industrias están siendo conquistadas por compañías que han sido capaces de obtener y estudiar la gran cantidad de datos que sus clientes generan

des más importantes que la describen. Esto puede ser usado para detectar un planeta que transita su estrella, pero primero necesitamos obtener la cantidad de luz que nos llega desde la estrella cada cierto tiempo t. Estas medidas tendrán una componente temporal y un error de medida asociado, que le añade una componente estocástica que se asume distribuida de forma normal debido a la gran cantidad de luz que se recibe de la estrella en cada observación.

Mundos tan distintos como los sistemas planetarios y la moda tienen una conexión bastante directa en lo que se refiere a la manera en la que los estudiamos usando datos y técnicas estadísticas

En el caso en el que las observaciones se tomen en una serie de tiempos t_i equidistantes, se pueden aplicar ecuaciones muy sencillas para obtener el espectro de Fourier. En astronomía, el principio de equidistancia no se cumple casi nunca, así que el espectro de Fourier que se usa, conocido como periodograma de Lomb-Scargle, se construye ajustando los datos a una función sinusoidal general de periodo P, con la siguiente forma

$$f(t) = a * \sin(2 \pi t / P) + b * \cos(2 \pi t / P)$$

Para cada valor de t_i podemos calcular los términos sinusoidales y usar un modelo lineal de mínimos cuadrados para obtener a y b. Un gráfico que muestre $a^2 + b^2$ en función de la periodicidad P, o la frecuencia 1/P, nos permitirá entender la periodicidad de nuestras observaciones de una manera visual y directa.

EL ESPECTRO DE FOURIER ADAPTADO AL COMPORTAMIENTO HUMANO

Una de las maneras más directas que tenemos en nuestra compañía de saber qué quieren nuestros clientes en su próximo envío es escucharles, puesto que durante el proceso de pedido de un nuevo "fix" al cliente se le da la oportunidad de describir en sus propias palabras que anda buscando en ese momento.

Para hacer uso de estos datos, primero utilizamos diversas técnicas de procesado de lenguaje natural (NLP por sus siglas en inglés) que nos permiten identificar diversos temas que se repiten a menudo en las notas escritas por los clientes, sobre todo cuando son mencionadas de una manera positiva.

El siguiente paso es construir series de datos históricos en las que para cada día contemos la cantidad de veces N(t) que nuestros clientes nos han escrito, y dentro de esos pedidos, cuantas veces n(t) nos han pedido un tipo de ropa particular. La división n(t)/N(t) nos permite visualizar como la probabilidad de que un cliente pida un tipo de ropa cambia con el tiempo.

Estadísticamente hablando, cada día una cantidad de clientes N(t) se enfrentan a la decisión individual de decidir si quieren pedir un tipo de ropa particular, con una cierta probabilidad p(t) de que eso suceda. Esto implica que la cantidad n(t) se distribuye de una manera binomial en vez de normal, como lo hacían los datos en el caso de los exoplanetas. Pero aun así podemos usar el mismo tipo de ajuste lineal, usando lo que se conoce como modelo lineal generalizado (GLM por sus siglas en inglés), mucho más flexible a la hora de aceptar diferentes tipos de datos. Con este modelo, el proceso para generar el espectro de Fourier es exactamente el mismo, y el espectro refleja la periodicidad de la probabilidad p(t).

LA COMPARACIÓN FINAL

Pongamos en práctica las técnicas descritas con datos reales. El planeta escogido es Kepler-78b, el primer planeta que descubrí durante mi doctorado precisamente gracias a que su señal en el espectro de Fourier es muy fácil de identificar. Con un periodo orbital de 8,5 horas, el planeta completó unas 3000 órbitas durante los cuatro años de vida de la misión Kepler. En la parte superior izquierda de la Figura 1 de la página siguiente se puede ver el espectro de Fourier de los datos descargados de la base de datos de la NASA. El pico más prominente en este espectro se encuentra alrededor de 3 ciclos por día (las órbitas que hay en un día). El resto de picos aparecen en múltiplos exactos de esta frecuencia principal, y aparecerán siempre que la parte periódica de nuestra señal no pueda ser descrita con una sola función sinusoidal. Una vez hemos identificado el periodo de la señal podemos

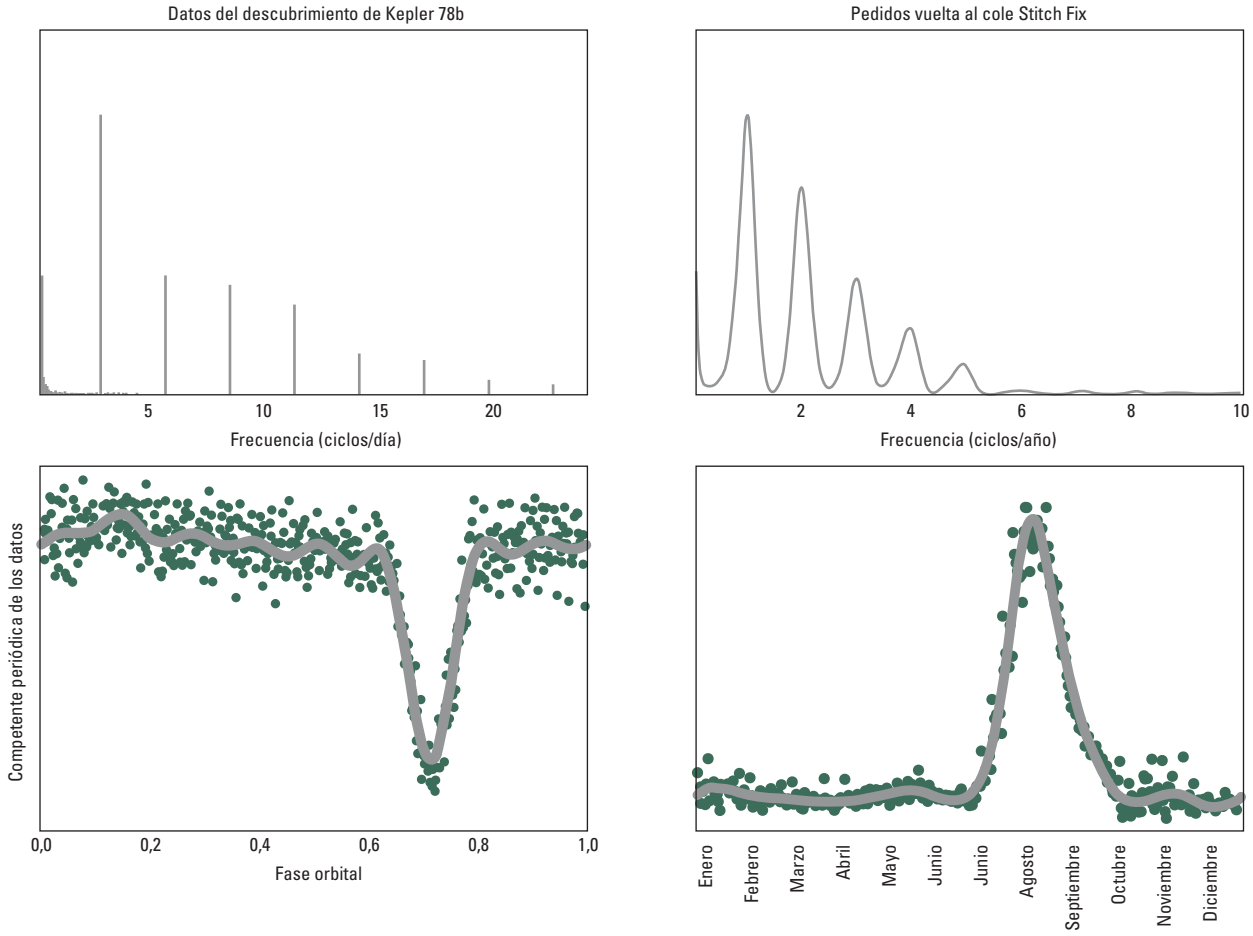
agrupar las 3000 órbitas en una señal única y de más alta precisión que nos permite ver claramente la presencia de los tránsitos del planeta. Podemos también ajustar un modelo que contenga la suma de todos los picos detectado, y con ello identificar el momento preciso en el que los tránsitos ocurren.

En el caso de la moda nos podemos centrar en un fenómeno muy conocido, el de la vuelta al cole. En la parte superior derecha de la figura podemos ver el espectro de Fourier de los datos asociados con este tema, con una estructura tremendamente parecida a la de los datos del exoplaneta. Podemos observar un pico principal que corresponde a una señal con periodo de un año. El mismo patrón de picos aparece en múltiplos de un ciclo por año, lo cual implica que la forma de la señal es parecida. Al agrupar los datos de los últimos años, podemos ver claramente una gran demanda durante el verano. El mismo modelo con todos los picos detectados nos permite distinguir que el pico de demanda se sitúa a mitad de agosto.

CONCLUSIONES

Como hemos podido ver, mundos tan distintos como los sistemas planetarios y la moda tienen una conexión bastante directa en lo que se refiere a la manera en la que los estudiamos usando datos y técnicas estadísticas. Esta conexión no es única, ya que una gran parte de las técnicas que se usan en la frontera entre la moda y el Big Data proceden de otros campos de la ciencia. Trazar este tipo de paralelismos es una de las tareas del Data Scientist, para sobre todo permitir que el campo siga evolucionando a la misma alta velocidad de la última década.

Figura 1. Comparación del espectro de Fourier de los datos del exoplaneta Kepler-78b y de las solicitudes de ropa para la vuelta al cole.



En los gráficos superiores, el valor de espectro de Fourier calculado. En la parte inferior, los puntos verdes representan los datos observados agregados para una mejor inspección visual. En gris, un modelo que contiene todas las periodicidades detectadas, y que permite predecir con precisión los tránsitos del planeta o el momento de mayor demanda de ropa para la vuelta al cole.