Sprint 10

Capítulo 1/8

Introducción al curso

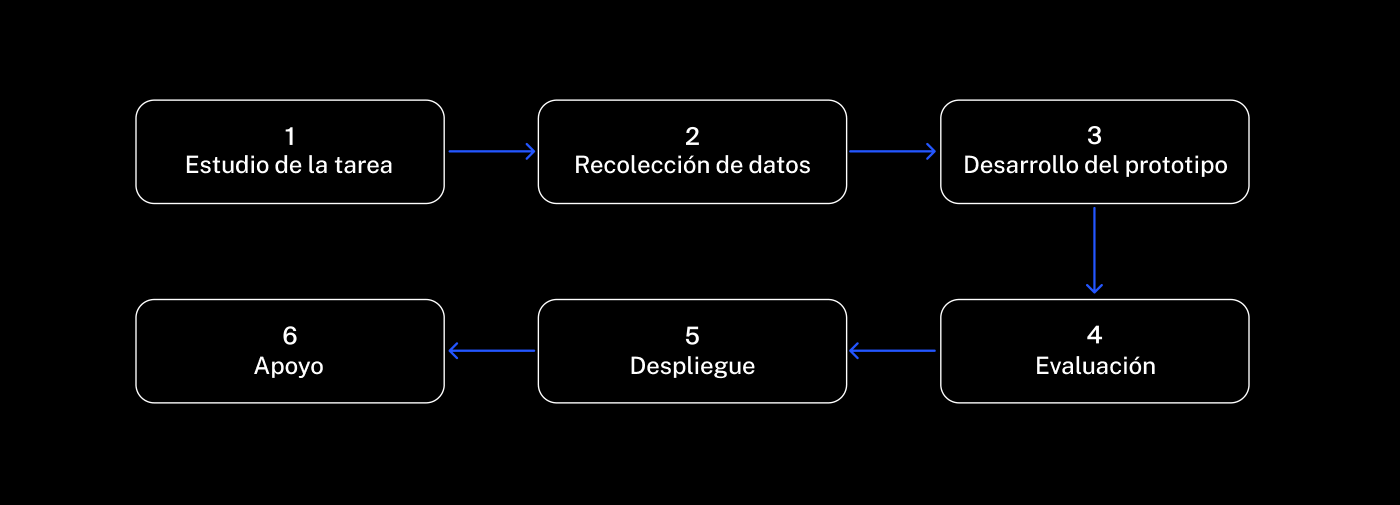
**Machine Learning en los negocios**

**En este curso descubriremos cómo se puede aplicar el machine learning a los negocios.**

Ya sabes cómo se implementan nuevas funcionalidades en un servicio mediante ML. Ahora es el momento de profundizar en la recopilación de datos y de explorar la conexión entre las métricas de producto y las métricas de ML.

**Estructura del curso**

Las lecciones anteriores estaban dedicadas al desarrollo de prototipos. Has aprendido a estudiar los datos y a construir modelos basados en ellos.



Empezaremos por el análisis de tareas y las métricas de negocio. Ya conoces algunas métricas de machine learning, pero no siempre se ajustan a los objetivos de negocios. Por ejemplo, ¿cómo podemos relacionar el abandono de clientes con la exactitud del modelo? ¿Cómo se correlaciona el error cuadrático medio del modelo con el tiempo medio de resolución de problemas por parte del servicio de asistencia? Primero exploraremos estas cuestiones.

Luego pasaremos a la evaluación del modelo y al soporte. Aprenderemos a determinar si el modelo es rentable o no, así como a señalar el momento en que el modelo deja de ser eficaz.

Terminaremos nuestro estudio del ciclo de vida del modelo con la recopilación de datos. Descubrirás dónde obtener los datos y cómo conseguir un dataset de mayor calidad.

Como siempre, al final tendrás tu proyecto de este curso. Haz clic [aquí](https://practicum-content.s3.us-west-1.amazonaws.com/new-markets/DS_sprint_9/moved_Descripcin_del_proyecto.pdf) si quieres ver la descripción del proyecto.

**Tus objetivos:**

* Aprender acerca de las métricas de negocio y su relación con las métricas de ML.
* Dominar la técnica de *bootstrapping* y aprender a analizar los resultados de las pruebas A/B.
* Entender qué es el etiquetado de datos y quiénes son los asesores.

En este sprint desarrollarás aún más tus habilidades de:

Código QR

Descripción generada automáticamenteGráfico

Descripción generada automáticamenteImagen que contiene Código QR

Descripción generada automáticamenteImagen que contiene Gráfico

Descripción generada automáticamente

**¿Cuánto tiempo llevará?**

Este es un curso de dificultad moderada. Todo lo que aprendiste en el curso de Análisis Estadístico de Datos te servirá de gran ayuda para dominar este material. Espera pasar entre 30 y 50 horas para completar este material, dependiendo de tus conocimientos previos y hábitos de estudio. Si sientes que te estás quedando atrás, no dudes en ponerte en contacto con nuestro equipo de orientación para comunicárselo. Como siempre, nos comprometemos a ayudarte en cada paso del camino.

Capítulo 2/8

Métricas de negocio

**Introducción**

**¿Qué necesita una empresa y cómo podemos medirlo?**

**Aprenderás a:**

* Calcular el beneficio y el margen
* Trabajar con conversiones y embudos
* Distinguir las métricas online de las métricas offline

**¿Cuánto tiempo tomará?**

10 lecciones de 5-10 minutos cada una.

**Descripción del ejercicio**

Examinar los indicadores de negocio a partir de diferentes casos de estudio.

Capítulo 2/8

Métricas de negocio

**Ingresos, coste de los bienes vendidos y margen**

**La tarea principal de cualquier empresa es ganar dinero.**

La mayoría de los planes comerciales pueden describirse con esta frase: comprar barato y vender caro. Las tiendas compran y venden productos. Las agencias asesoras compran el trabajo realizado por sus asesores y venden soluciones a los clientes. Los medios de comunicación compran la atención del público y la venden a los anunciantes. Una empresa sana siempre tiene costes e ingresos.

Estudiémoslos más detenidamente y calculemos los indicadores de negocio.

Por ejemplo, la empresa Juan e Hijos vende auriculares inalámbricos. En 2018, los ingresos de la empresa ascendieron a 1 000 000 de dólares. El coste de los bienes fue de 300 000 dólares. Gastaron $150 000 en salarios y $100 000 en alquiler y mantenimiento de la oficina. Juan e Hijos opera bajo un sistema fiscal simplificado y paga el 6 % de sus ingresos en impuestos.

Analicemos el problema:

* *Los ingresos de la empresa ascendieron a 1 000 000 de dólares.*

Los ingresos son la cantidad de dinero que los clientes pagaron a la empresa. Por ejemplo, un solo artículo tiene un precio de 10 dólares. Si se venden 10 unidades, los ingresos serán de 10×10=100 dólares.

* \*El coste de los bienes es de $300 000.

El coste de los bienes vendidos es el dinero pagado por la empresa por la compra del producto. Este valor no solo se refiere a los bienes, sino también a los servicios. Por ejemplo, para un taxista, el coste de la gasolina está incluido en el coste del viaje de un pasajero.

Conociendo los ingresos y el coste de los bienes vendidos, se puede calcular el primer indicador de negocio: el beneficio bruto.

Beneficio bruto=Ingresos−Coste de los bienesBeneficio bruto=Ingresos−Coste de los bienes

La empresa compró poco y vendió mucho, por lo que el beneficio bruto resultó positivo. El beneficio bruto es el primer indicador de la salud de la empresa. Es bastante fácil de calcular.

El beneficio bruto de Juan e Hijos es de:

1000000−300000=7000001000000−300000=700000

Si el margen bruto es negativo, suele ser una mala señal, ya que indica que la empresa vende menos de lo que compra. Habría que revisar urgentemente los precios o el conjunto de productos y servicios. Sin embargo, existen algunas excepciones a esta regla. A veces, las empresas venden productos por debajo del coste para "enganchar" al cliente a sus servicios o a la asistencia continuada.

La relación entre el beneficio bruto y los ingresos se denomina margen bruto.

Margen de beneficio bruto=Beneficio brutoIngresos×100%Margen de beneficio bruto=IngresosBeneficio bruto​×100%

Por ejemplo, el margen bruto de Juan e Hijos es de:

7000001000000=70%1000000700000​=70%

Esto significa que solo el 70 % de los ingresos se queda en la empresa. Para los diferentes productos y categorías, el margen se calcula por separado. De esta manera, veremos con más claridad en la venta de qué productos debe centrarse la empresa.

Pregunta

En 2015, la empresa Where Did Everybody Go obtuvo unos beneficios de 345.023 dólares gracias a la venta de matamoscas eléctricos. El coste de los bienes fue de 147 604 dólares. Se gastaron 53 210 dólares en salarios y 32 945 dólares en gastos de oficina. La empresa Where Did Everybody Go opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el margen bruto de la empresa?

42%

26%

32%

74%

57%

¡Eso es! La parte del beneficio bruto en los ingresos asciende aproximadamente al 57 %.

No hay respuesta correcta

¡Tu comprensión del material es impresionante!

Pregunta

En el año 2018, la empresa Kevin Miller generó unos beneficios de 117 132 dólares por la venta de calcetines. El coste de los bienes fue de 152 189 dólares. Los salarios fueron de 35 103 dólares. Y el alquiler de la oficina fue de 54 979 dólares. La propiedad Kevin Miller opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el beneficio bruto de la empresa? Estas son las opciones:

-$24 540

-$35 057

Esto es exactamente lo que ocurre cuando un autónomo, por alguna razón, compra un producto por un precio superior al precio de venta. ¡Es peligroso!

-$125 139

-$132 167

¡Buen trabajo!

Sprint 10

Capítulo 2/8

Métricas de negocio

**Beneficios y gastos operativos**

**Los salarios y los gastos de alquiler no están incluidos en el coste de los bienes vendidos. Tienen una categoría aparte, llamada gastos operativos.**

Aquí es donde se incluyen todos los gastos relacionados con las actividades comerciales de la empresa, por ejemplo, los gastos de electricidad e Internet.

Juan e Hijos alquila una oficina por 100 000 dólares y gasta 100 000 dólares en salarios. El total de los gastos operativos de la empresa es de 200 000 dólares al mes.

Si restamos los gastos operativos del beneficio bruto, obtendremos el beneficio operativo.

Beneficio operativo=Beneficio bruto−Gastos operativosBeneficio operativo=Beneficio bruto−Gastos operativos

El beneficio operativo de Juan e Hijos es de:

700000−200000=500000700000−200000=500000

El beneficio operativo ayuda a averiguar cuánto gana una empresa gracias a sus actividades comerciales. Salvo contadas excepciones, el beneficio operativo se correlaciona con el beneficio neto: cuanto mayor sea el beneficio operativo, mayor será el beneficio neto. Además, el beneficio operativo es más rápido y fácil de calcular.

Si el beneficio operativo es negativo, esto se denomina pérdida operativa. Muestra que los propietarios de la empresa aún no pueden ganar dinero con su negocio. Pero al igual que un beneficio bruto negativo, una pérdida operativa no significa necesariamente que todo esté perdido. A menudo la empresa tiene pérdidas planificadas porque invierte todos los beneficios en un rápido crecimiento.

Si dividimos el beneficio operativo entre los ingresos, obtenemos el margen de beneficio operativo. Se trata de la parte de los ingresos que queda en la empresa tras deducir el coste de los bienes vendidos, los salarios, el alquiler, el marketing y otros gastos relacionados con las actividades principales. A los inversores les suelen interesar los márgenes operativos porque permiten comparar distintas empresas.

Margen de beneficio operativo=Beneficio operativoIngresos×100%Margen de beneficio operativo=IngresosBeneficio operativo​×100%

Por ejemplo, el margen de beneficio operativo de Juan e Hijos es de:

5000001000000×100%=50%1000000500000​×100%=50%

Pregunta

En 2015, la empresa Where Did Everybody Go obtuvo unos beneficios de 115 417 dólares por la venta de matamoscas eléctricos. El coste de los bienes fue de 343 031 dólares. 59 107 dólares se gastaron en salarios, y 62 170 dólares en alquiler de la oficina. La empresa Where Did Everybody Go opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el valor de los gastos operativos de la empresa?

69 095

121 277

En este caso, los gastos operativos están compuestos por los salarios y el alquiler.

464 308

62 170

No hay respuesta correcta

¡Tu comprensión del material es impresionante!

Pregunta

En 2015, la empresa Key for Intercom generó unos beneficios de 323 864 dólares gracias a la venta de llaves de repuesto. El coste de los bienes vendidos fue de 156 731 dólares. Se gastaron 50 632 dólares en salarios y 54 595 dólares en alquiler de oficinas. Key for Intercom opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el beneficio operativo de la empresa?

106 009

61 906

¡Correcto!

167 133

42 474

No hay respuesta correcta

¡Buen trabajo!

Sprint 10

Capítulo 2/8

Métricas de negocio

**Beneficios y gastos operativos**

**Los salarios y los gastos de alquiler no están incluidos en el coste de los bienes vendidos. Tienen una categoría aparte, llamada gastos operativos.**

Aquí es donde se incluyen todos los gastos relacionados con las actividades comerciales de la empresa, por ejemplo, los gastos de electricidad e Internet.

Juan e Hijos alquila una oficina por 100 000 dólares y gasta 100 000 dólares en salarios. El total de los gastos operativos de la empresa es de 200 000 dólares al mes.

Si restamos los gastos operativos del beneficio bruto, obtendremos el beneficio operativo.

Beneficio operativo=Beneficio bruto−Gastos operativosBeneficio operativo=Beneficio bruto−Gastos operativos

El beneficio operativo de Juan e Hijos es de:

700000−200000=500000700000−200000=500000

El beneficio operativo ayuda a averiguar cuánto gana una empresa gracias a sus actividades comerciales. Salvo contadas excepciones, el beneficio operativo se correlaciona con el beneficio neto: cuanto mayor sea el beneficio operativo, mayor será el beneficio neto. Además, el beneficio operativo es más rápido y fácil de calcular.

Si el beneficio operativo es negativo, esto se denomina pérdida operativa. Muestra que los propietarios de la empresa aún no pueden ganar dinero con su negocio. Pero al igual que un beneficio bruto negativo, una pérdida operativa no significa necesariamente que todo esté perdido. A menudo la empresa tiene pérdidas planificadas porque invierte todos los beneficios en un rápido crecimiento.

Si dividimos el beneficio operativo entre los ingresos, obtenemos el margen de beneficio operativo. Se trata de la parte de los ingresos que queda en la empresa tras deducir el coste de los bienes vendidos, los salarios, el alquiler, el marketing y otros gastos relacionados con las actividades principales. A los inversores les suelen interesar los márgenes operativos porque permiten comparar distintas empresas.

Margen de beneficio operativo=Beneficio operativoIngresos×100%Margen de beneficio operativo=IngresosBeneficio operativo​×100%

Por ejemplo, el margen de beneficio operativo de Juan e Hijos es de:

5000001000000×100%=50%1000000500000​×100%=50%

Pregunta

En 2015, la empresa Where Did Everybody Go obtuvo unos beneficios de 115 417 dólares por la venta de matamoscas eléctricos. El coste de los bienes fue de 343 031 dólares. 59 107 dólares se gastaron en salarios, y 62 170 dólares en alquiler de la oficina. La empresa Where Did Everybody Go opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el valor de los gastos operativos de la empresa?

69 095

121 277

En este caso, los gastos operativos están compuestos por los salarios y el alquiler.

464 308

62 170

No hay respuesta correcta

¡Tu comprensión del material es impresionante!

Pregunta

En 2015, la empresa Key for Intercom generó unos beneficios de 323 864 dólares gracias a la venta de llaves de repuesto. El coste de los bienes vendidos fue de 156 731 dólares. Se gastaron 50 632 dólares en salarios y 54 595 dólares en alquiler de oficinas. Key for Intercom opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el beneficio operativo de la empresa?

106 009

61 906

¡Correcto!

167 133

42 474

No hay respuesta correcta

¡Buen trabajo!

Sprint 10

Capítulo 2/8

Métricas de negocio

**Beneficio neto**

**Ni los ingresos ni el beneficio operativo muestran cuánto dinero gana la empresa en realidad.**

El indicador de beneficio neto es el que tiene en cuenta el coste de los bienes vendidos, los gastos operativos y las obligaciones ante el Estado y los acreedores. Esta es la cantidad de dinero que los propietarios pueden tener a su disposición o reinvertir en el desarrollo de la empresa.

Beneficio neto=Beneficio operativo−Impuestos y preˊstamosBeneficio neto=Beneficio operativo−Impuestos y preˊstamos

La empresa Juan e Hijos no tiene ningún préstamo. Esto significa que las obligaciones solo incluyen los impuestos, de modo que la empresa paga el 6 % de sus ingresos (1 000 000 de dólares) en impuestos. Al calcular los impuestos de la empresa, obtenemos 1,000,000×0.06=60,0001,000,000×0.06=60,000. El beneficio neto es de: 500,000−60,000=440,000500,000−60,000=440,000.

El beneficio neto no puede calcularse hasta el final del año, cuando se determinen todas las obligaciones fiscales y crediticias. Por lo tanto, los beneficios brutos y operativos se utilizan para la gestión diaria de la empresa, mientras que el beneficio neto se calcula para la junta anual de accionistas.

Un beneficio neto negativo se denomina pérdida neta. Muestra que la empresa no consiguió ganar dinero.

Pregunta

En 2011, la empresa Where Did Everybody Go ganó 261 812 dólares tras la venta de matamoscas eléctricos. El coste de los bienes vendidos fue de 221 643 dólares. Los salarios ascendieron a 69 021 dólares y el alquiler de la oficina a 55 519 dólares. La empresa Where Did Everybody Go opera bajo un sistema fiscal simplificado y paga el 6 % de los ingresos. ¿Cuál es el beneficio neto de la empresa?

40 169

-84 371

-100 080

Correcto, Where Did Everybody Go sigue sin ser rentable. Mejor releer esa biografía de Jeff Bezos.

No hay respuesta correcta

¡Buen trabajo!

Pregunta

En 2011, la empresa ColorForest ganó 604 000 dólares gracias a la venta de juguetes navideños de madera. El coste de los bienes vendidos fue de 34 500 dólares. Se gastaron 63 201 dólares en salarios y 32 945 dólares en alquiler de la oficina. ColorForest opera bajo un sistema fiscal simplificado y paga el 6 % de sus ingresos. La empresa también ha pedido un préstamo este año y ha gastado 230 550 dólares en concepto de aval. ¿Cuál es el beneficio neto de la empresa?

569 500

473 345

206 564

¡Correcto!

96 146

No hay respuesta correcta

**Rendimiento sobre la inversión**

**Al principio de su ciclo de vida, una empresa suele entrar en números rojos, para luego salir poco a poco de esta situación. ¿Cómo puedo saber si la inversión será rentable o no?**

Para poner en marcha un negocio, hay que invertir dinero en él. Por ejemplo, para lanzar una tienda online, hay que elaborarle un diseño, configurarla, programar su funcionamiento, establecer un sistema de seguimiento de los productos, alquilar un almacén y contratar personal. Todas estas acciones requieren dinero, y las ventas no empezarán inmediatamente.

Los inversionistas quieren saber cuándo las inversiones darán sus frutos. Para ellos, la métrica principal es el ROI (return on investment o rendimiento sobre la inversión).

ROI=Beneficio neto−InversioˊnInversioˊnROI=InversioˊnBeneficio neto−Inversioˊn​

Por ejemplo, el inversor Simon Bishop invirtió 100 millones de dólares en la construcción de una fábrica de equipamiento para patinaje artístico. La fábrica ha abierto sus puertas y hasta ahora ha obtenido 20 millones de dólares de beneficio neto.

El ROI actual de la inversión de Simon Bishop se calcula de la siguiente manera:

20 millones−100 millones100 millones=−80%100 millones20 millones−100 millones​=−80%

Es decir, el 80 % de sus inversiones aún no se han amortizado.

Pregunta

El empresario Aaron vendió su apartamento e invirtió 10 millones de dólares en el lanzamiento de un estudio de diseño web. Durante el primer año, el estudio obtuvo 2.5 millones de dólares de beneficio neto. ¿A qué equivale el ROI?

25%

-25%

-75%

Así es, el 75 % de la inversión aún no se ha amortizado.

75%

No hay respuesta correcta

¡Buen trabajo!

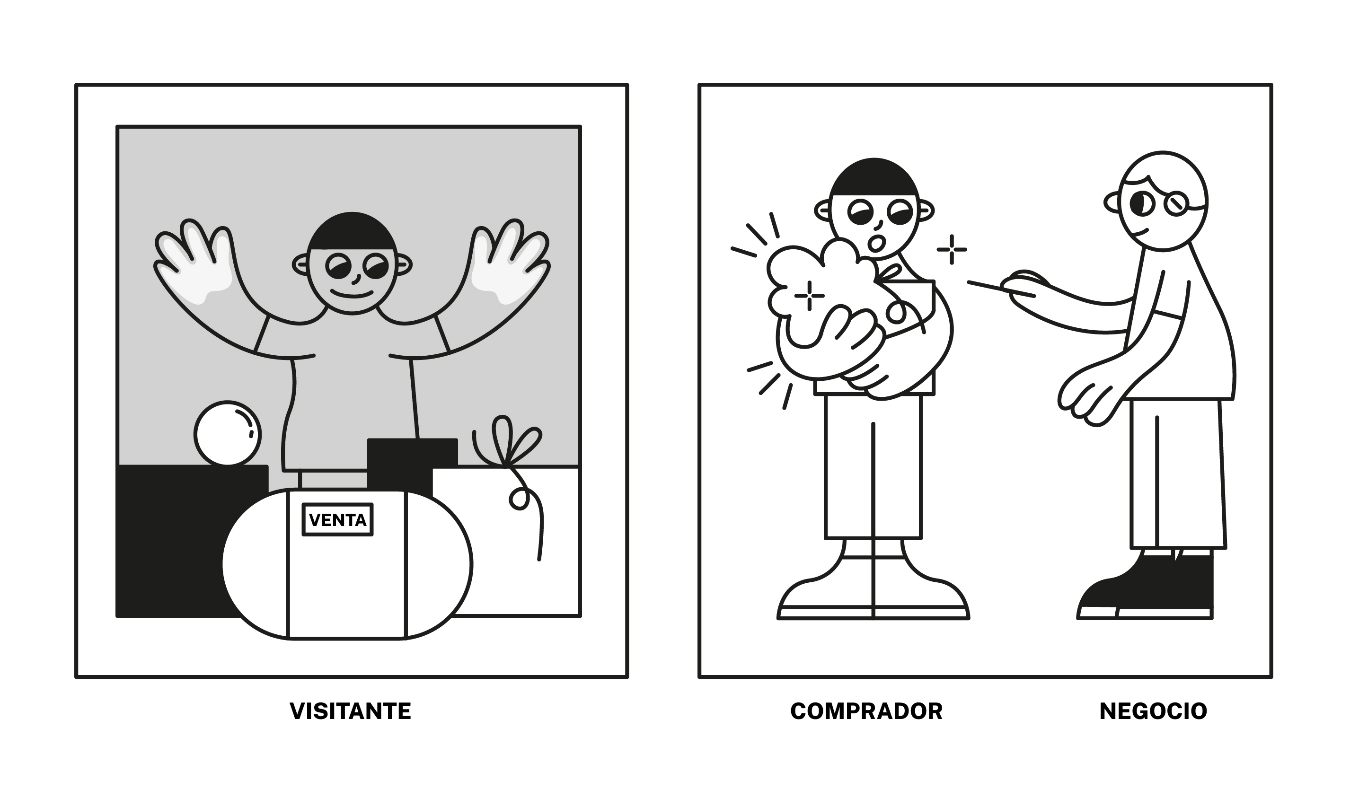
Capítulo 2/8

Métricas de negocio

**Conversión**

**El proceso de venta puede describirse así: visitante → venta → comprador.**

Todo comienza por el hecho de que un usuario se ponga en contacto con la empresa. No importa si se trata de una llamada telefónica, de la instalación de una aplicación móvil o de una visita a un sitio web o a una tienda real. Lo principal es que la persona esté interesada en los productos o servicios de la empresa.



A continuación, se realiza la venta: o bien el usuario añade un artículo al carrito, o bien supera los niveles de entrenamiento en el juego, o bien se comunica con el vendedor. Aunque estos procesos son diferentes, el objetivo es el mismo, a saber, convertir a un usuario en cliente. Cuanto más eficaz sea la empresa en este sentido, mejor le irán las cosas. Por lo tanto, una de las métricas más importantes es la conversión, que muestra el porcentaje de usuarios que completaron una acción determinada (por ejemplo, descargaron una aplicación o hicieron una suscripción).

Si 1000 personas visitaron la tienda y 10 de ellas compraron algo, la tasa de conversión de usuarios a clientes es de 10/1000 = 0.01 o 1 %.

Pregunta

La aplicación móvil Download-Launch fue instalada por 22 755 personas, de las cuales 555 compraron la versión de pago. ¿Cuál es la tasa de conversión de la aplicación móvil?

1.342%

2.439%

¡Correcto!

3.112%

0.424%

2.001%

No hay respuesta correcta

¡Bien hecho!

Pregunta

La conversión de la tienda online de smartphones Ph0nz.com es de 2.154%. La tienda recibió 21 824 visitantes. ¿Cuántos clientes hubo?

611

235

470

¡Correcto!

0

329.8

No hay respuesta correcta

¡Excelente trabajo!

Pregunta

La tasa de conversión del juego online World of Dump Trucks es del 1.821 %, y el número de compradores es de 712. ¿Cuántos visitantes hubo?

50 842

0

39 099

¡Correcto!

19 554

27 376

No hay respuesta correcta

¡Buen trabajo!

**Embudos**

**Para aumentar las ventas, no basta con conocer la tasa de conversión. ¡Necesitamos embudos!**

Los embudos representan una forma de mostrar:

* El camino que sigue el usuario para comprar un producto
* La proporción de personas que pasan a la siguiente etapa

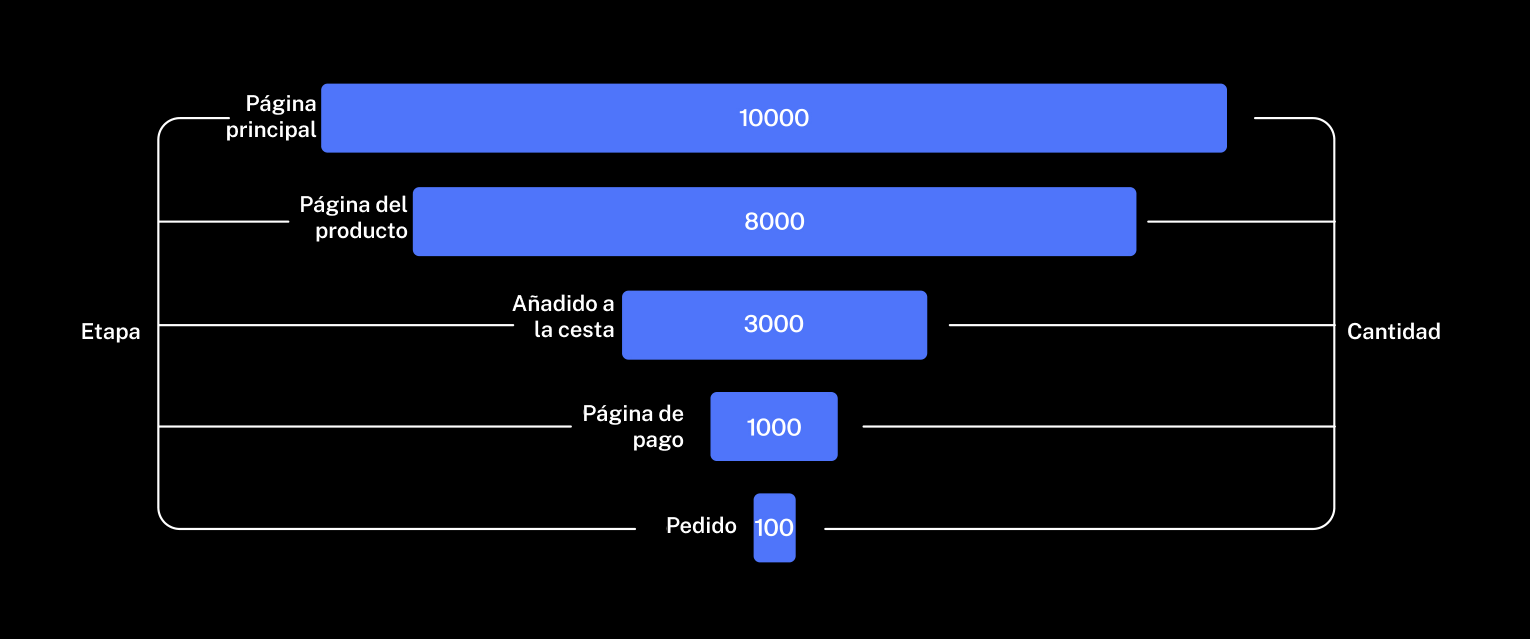
Por ejemplo, en una tienda online, el usuario realizó los siguientes pasos:

1. Accedió a la página de inicio de la tienda
2. Entró en la página del producto
3. Añadió un artículo al carrito
4. Pasó a la página de pago
5. Pagó el pedido

Para construir un embudo, es necesario determinar cuántas personas llegaron a cada etapa. Digamos que 10 000 usuarios abrieron la página principal del sitio, de los cuales 8 000 accedieron a la página del producto. 3 000 usuarios añadieron el producto al carrito de compras, pero solo mil de ellos llegaron a la página de pago. Cien completaron todo el trayecto del pedido:

EtapaCantidadPaˊgina principal10000Paˊgina del producto8000An˜adido a la cesta3000Paˊgina de pago1000Pedido100EtapaPaˊgina principalPaˊgina del productoAn˜adido a la cestaPaˊgina de pagoPedido​Cantidad10000800030001000100​​

El gráfico que representa el número de personas en cada etapa se asemeja a un embudo:



Al conocer el número de personas en cada etapa, se puede calcular el porcentaje de personas que llegaron a un determinado paso, así como el porcentaje de los que dieron cada paso posterior:

EtapaCantidadConversioˊn totalConversioˊn a la siguiente etapaPaˊgina principal10000100%100%Paˊgina del producto800080%80%An˜adido a la cesta300030%38%Paˊgina de pago100010%33%Pedido1001%10%EtapaPaˊgina principalPaˊgina del productoAn˜adido a la cestaPaˊgina de pagoPedido​Cantidad10000800030001000100​Conversioˊn total100%80%30%10%1%​Conversioˊn a la siguiente etapa100%80%38%33%10%​​

Esta pequeña tabla es la fuente de muchas conclusiones e hipótesis. Por ejemplo, solo una de cada diez personas que empezaron a hacer un pedido lo completó. ¿Hay algún problema con la página de pago?

Podemos ver que 2 900 personas añadieron un artículo al carrito, pero no lo compraron. ¿Deberíamos enviarles correos electrónicos para recordarles que su compra no se ha completado?

El análisis de embudo te permite formular hipótesis, ponerlas a prueba y hacer un seguimiento de los cambios.

Pregunta

Selecciona el mejor embudo de conversión para el juego de móvil Attack of Clowns. Es gratuito, pero con dinero se pueden comprar armas más potentes y maquillaje.

Superó el entrenamiento → creó un personaje → instaló el juego → pagó

Instaló el juego → creó un personaje → superó el entrenamiento → pagó

Así es, los usuarios siguen este camino.

Pagó → creó un personaje → superó el entrenamiento → instaló el juego

No hay respuesta correcta

¡Perfecto!

Capítulo 2/8

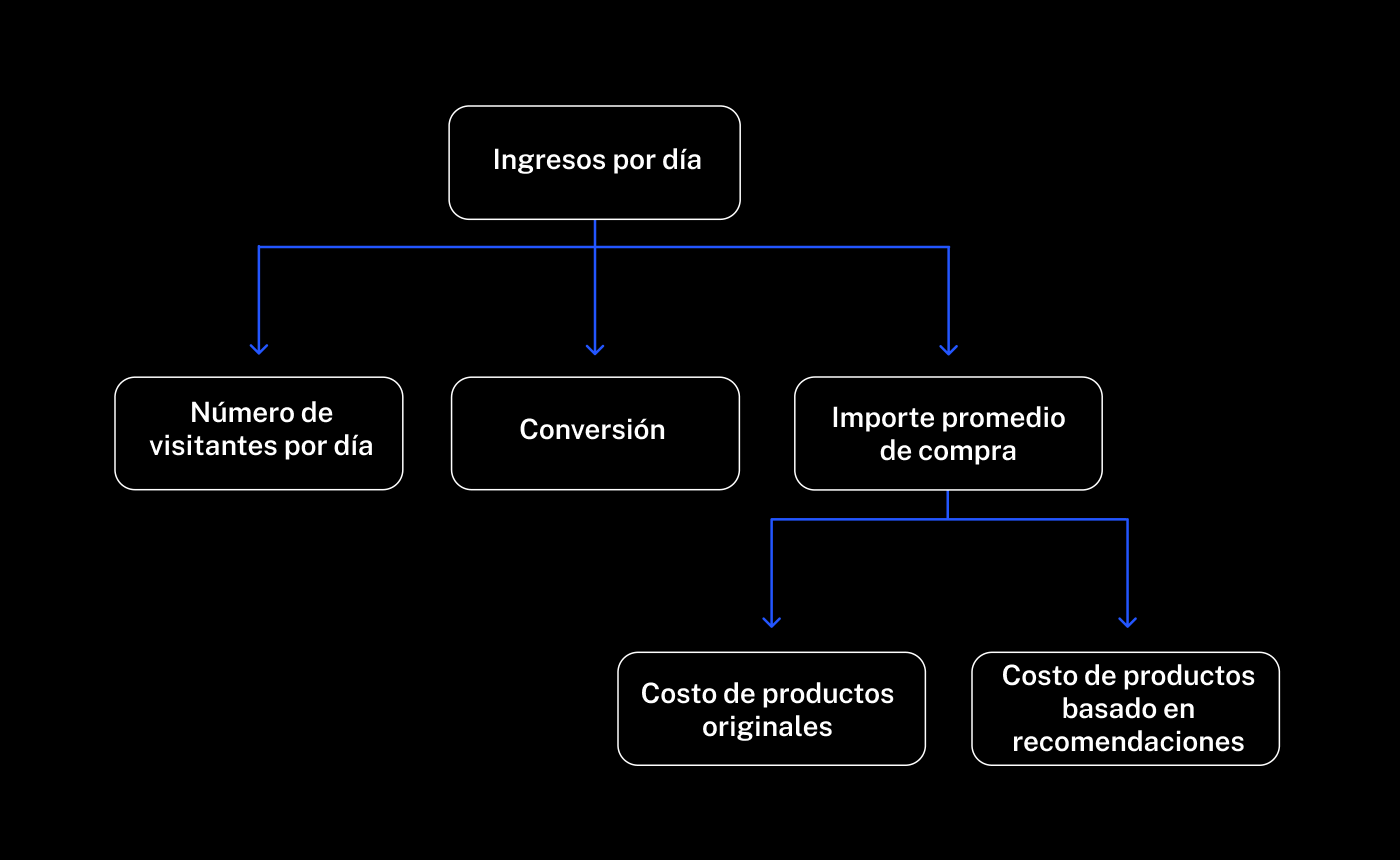
Métricas de negocio

**Métricas online y offline**

**Estás desarrollando un sistema de recomendaciones para una tienda online. Tienes que decidir qué productos ofrecer a los compradores en función de su carrito. ¿Cómo puede afectar la clasificación a los ingresos?**

Para responder a esta pregunta, haremos una descomposición de las métricas, o sea, dividir las métricas en componentes.

Los ingresos representan la métrica principal. Para calcularla, tenemos que multiplicar el número de usuarios por día, su conversión y el recibo promedio, que es la suma de los productos añadidos inicialmente por el usuario y de los recomendados por el modelo.



La métrica objetivo de evaluación es el precio medio de los productos que se añaden con base en las recomendaciones. Es la métrica online calculada en un sistema en funcionamiento con usuarios reales. No puedes calcularla para otros modelos a partir de datos del historial.

Para construir nuevos modelos, también necesitarás métricas offline. Se calculan en función de los datos del historial. Ya conoces dos métricas offline, que son las métricas *exactitud* y *ECM* de machine learning.

El coste de los bienes añadidos con base en las recomendaciones se ve afectado por:

1. ¿Cuántos productos de la lista de recomendaciones son interesantes para el usuario? Esto se mide con la métrica *precisión*.
2. ¿El modelo añade todos los productos interesantes para el usuario a la lista? Se mide con la métrica *recall*.

Recuerda que el *valor F1* combina las métricas *precisión* y *sensibilidad*.

Pregunta

El servicio de entrega de FoodDude está desarrollando un modelo que predice el tiempo de entrega en función del tiempo atmosférico y el tráfico. Selecciona las métricas offline que necesites de la lista: opciones:

Elige tantas como quieras

El número de quejas de que la entrega tardó más de lo previsto por el modelo.

El coste total de los pedidos cancelados: el tiempo de entrega previsto por el modelo era demasiado largo.

El tiempo promedio de retraso: la entrega tardó más de lo previsto por el modelo.

Correcto, esta es una métrica offline.

El número de pedidos que fueron reprogramados a una hora anterior. El modelo predijo correctamente que la entrega sería rápida.

Número de pedidos cuyo plazo de entrega previsto es tres días superior al real.

Correcto, es una métrica offline.

¡Excelente trabajo!

**Establecer objetivos**

**Al seleccionar una métrica, el equipo de producto determina cómo aumentar su valor en función del objetivo comercial.**

Los ingresos se determinan según las diferentes circunstancias. No solo según la calidad del modelo de machine learning, sino también, por ejemplo, según la interfaz de una tienda online.

A veces, una tarea de machine learning es fácil de resolver, pero el responsable de producto deja su resolución para más adelante. ¿Por qué? Porque, por ejemplo, hay un problema con la interfaz. No importa si las recomendaciones son buenas o incluso excelentes si el usuario ni siquiera ve el botón de pago.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Debemos tener en cuenta que los productos de una compañía cambian, al igual que sus objetivos.

Consideremos un escenario en el que las métricas offline crecen, pero el beneficio no aumenta. ¿Dónde está el problema? Entender el comportamiento del usuario es crucial, así que analiza dicho comportamiento. La forma más fácil es hacer una encuesta. Así, sabrás qué puede estar obstaculizando que los usuarios realicen compras.

Otro ejemplo sería el siguiente. Imagina que las recomendaciones no están siendo tan efectivas como se desea, ¿no significa eso que la métrica de *precisión* es más importante que *recall*? En lugar de mejorar el modelo con el valor *F1*, selecciona el umbral de clasificación en el que la *precisión* se incrementará mediante la reducción de la *recall*.

¡Métrica seleccionada, objetivos fijados! Pasemos a desarrollar y a mejorar el modelo. Esto ayudará a:

* iterar sobre diferentes hiperparámetros,
* cambiar el modelo,
* mejorar la calidad y la cantidad de los datos.

Capítulo 2/8 · Última lección

Métricas de negocio

**Conclusión**

**Has aprendido a:**

* Calcular el beneficio neto y el rendimiento sobre las inversiones
* Explicar qué son las conversiones y los embudos
* Distinguir las métricas online y offline

En el próximo capítulo, sabrás determinar si la implementación de una nueva funcionalidad tendrá éxito.

**Material adicional**

[Guía de KPIs para Google Play Apps y Juegos: Introducción a las métricas](https://medium.com/googleplaydev/a-kpis-guide-for-google-play-apps-and-games-intro-to-metrics-214da979f5a2)

**Llévate esto contigo**

Descárgate el resumen del capítulo para poder consultarlos cuando los necesites.

* [Resumen del capítulo: Métricas de negocios](https://practicum-content.s3.us-west-1.amazonaws.com/new-markets/DS_sprint_9/ES_Resumen_del_captulo_Mtricas_de_negocios.pdf?etag=1ac18885982e33b7d2235486b7599bda)

Capítulo 3/8

Implementar una nueva funcionalidad

**Introducción**

**¿Cómo podemos saber si la implementación de una nueva funcionalidad mejorará las métricas?**

**Aprenderás:**

* qué son las pruebas A/B y cómo analizar sus resultados;
* a encontrar el intervalo de confianza;
* a dominar la técnica de bootstrapping.

**¿Cuánto tiempo tomará?**

11 lecciones de 10 minutos cada una.

**Descripción del ejercicio**

Examinar las técnicas de evaluación de métricas mediante diferentes casos.

Sprint 10

Capítulo 3/8

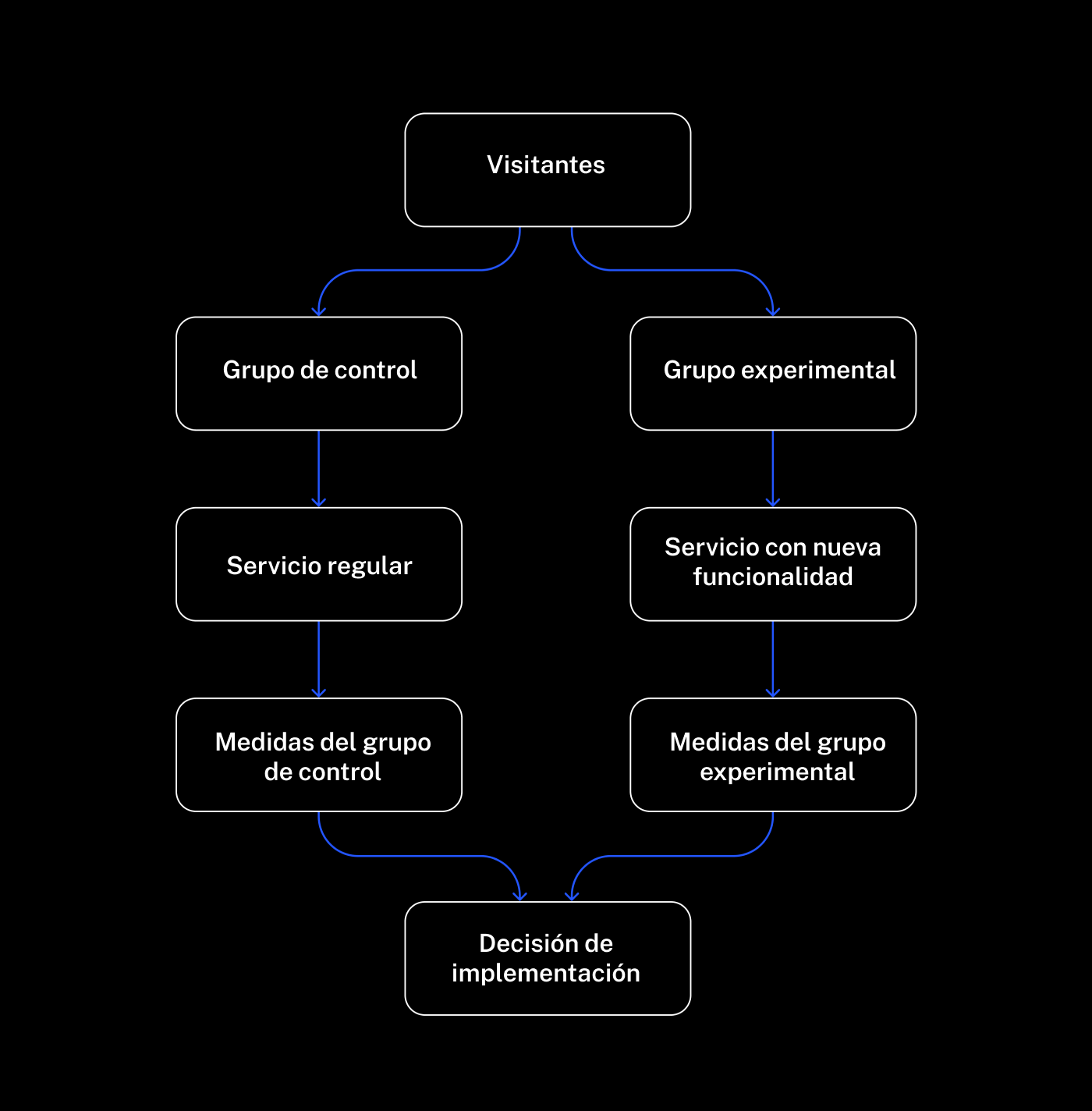
Implementar una nueva funcionalidad

**Planificación de la implementación**

**¿Debería la empresa implementar un nuevo servicio o no? ¿Cómo podemos tomar la decisión correcta? La respuesta es mediante las pruebas A/B (A/B testing).**

Las pruebas A/B o split testing son una técnica de comprobación de hipótesis que ayuda a controlar el impacto que provocan los cambios de un servicio o producto sobre los usuarios. Se realiza de la siguiente manera: la población se divide en grupo de control (*A*) y grupo experimental (*B*). El grupo *A* utiliza el servicio habitual sin cambios. El grupo *B* utiliza la nueva versión, que es la que tenemos que probar.

El experimento dura un periodo determinado (por ejemplo, dos semanas). El objetivo es recopilar datos sobre el comportamiento de los visitantes en ambos grupos. Si la métrica clave en el grupo experimental mejora en comparación con el grupo de control, entonces se implementará la nueva funcionalidad.



Antes de las pruebas A/B, a menudo se utiliza la prueba A/A, o comprobación de validez, en la que los visitantes se dividen en dos grupos de control que se exponen a la misma versión del servicio. La métrica clave debe coincidir en ambos grupos, en caso contrario, hay que buscar un error.

Pregunta

La tienda online de mascotas Purr Se ha cambiado la interfaz de usuario. ¿Qué métrica puede utilizarse para evaluar la viabilidad de los cambios?

Vistas a la página

El número de compras

Una nueva interfaz de usuario puede incitar a los visitantes a comprar (los gatos felices se suman a tu karma) o, por el contrario, ahuyentarlos.

Velocidad de carga de la página

¡Buen trabajo!

Estos son algunos de los factores que debemos tener en cuenta al comparar las métricas:

* Si la diferencia entre las métricas es insignificante, ¿cómo podemos identificar los casos que se deban a la casualidad? En próximas lecciones veremos algunas herramientas estadísticas que pueden ser útiles en este asunto.
* Si el experimento es demasiado largo, ralentiza la implementación de la nueva funcionalidad. En cambio, es posible que un experimento demasiado corto no proporcione resultados fiables. Aprenderás a calcular el número de observaciones suficientes para tomar una decisión.
* La comparación de los valores medios de las métricas no tiene en cuenta la dispersión. Aprenderás a medir la magnitud de la desviación de las métricas basándote en los resultados de las pruebas *A/B*.

En las siguientes lecciones, analizaremos los factores anteriores con más detalle.

Pregunta

Aquí están los gráficos de distribución de métricas correspondientes a dos experimentos. La diferencia entre los valores medios es la misma. ¿Qué experimento es más fiable?

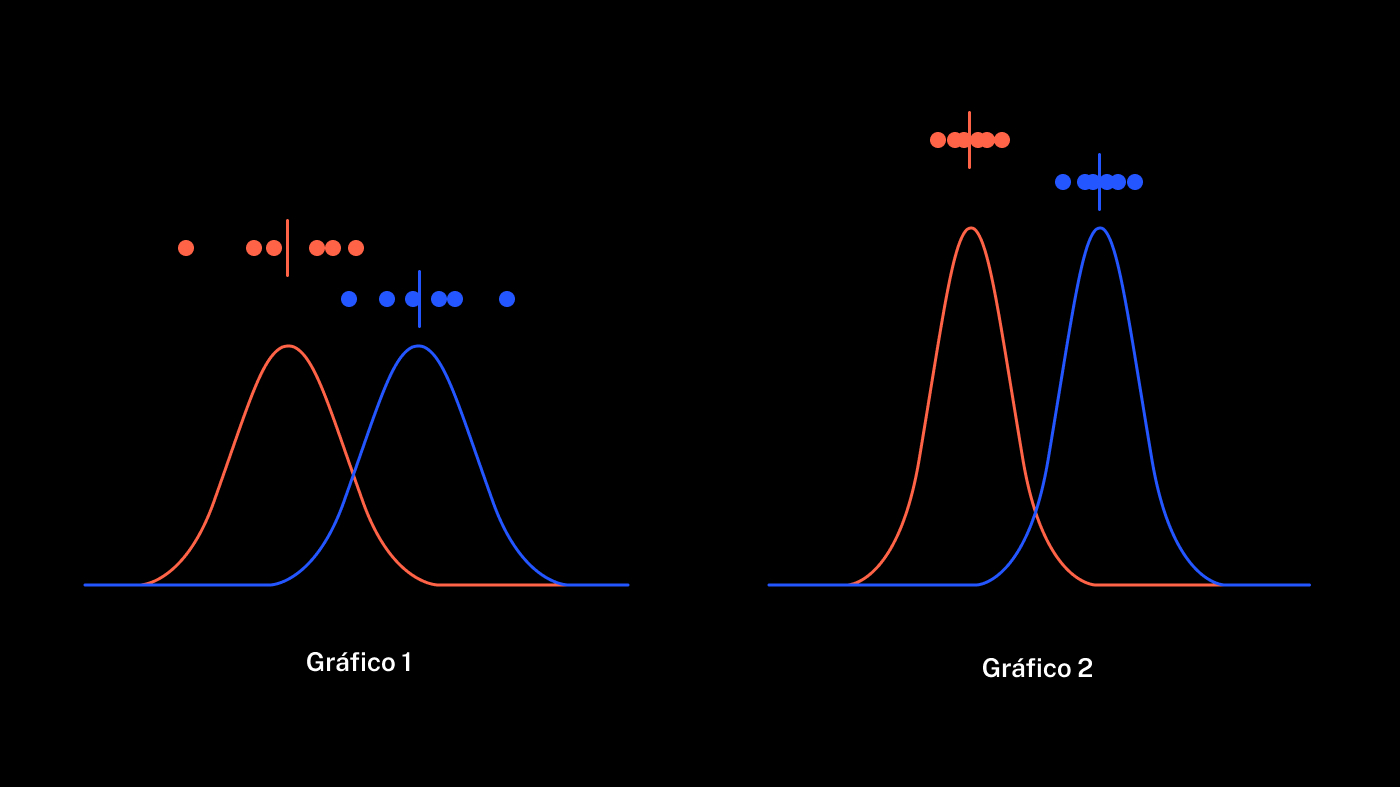


Gráfico 1

Gráfico 2

La dispersión de los datos es baja. Los resultados del experimento son fiables.

No hay diferencia

¡Buen trabajo!

Implementar una nueva funcionalidad

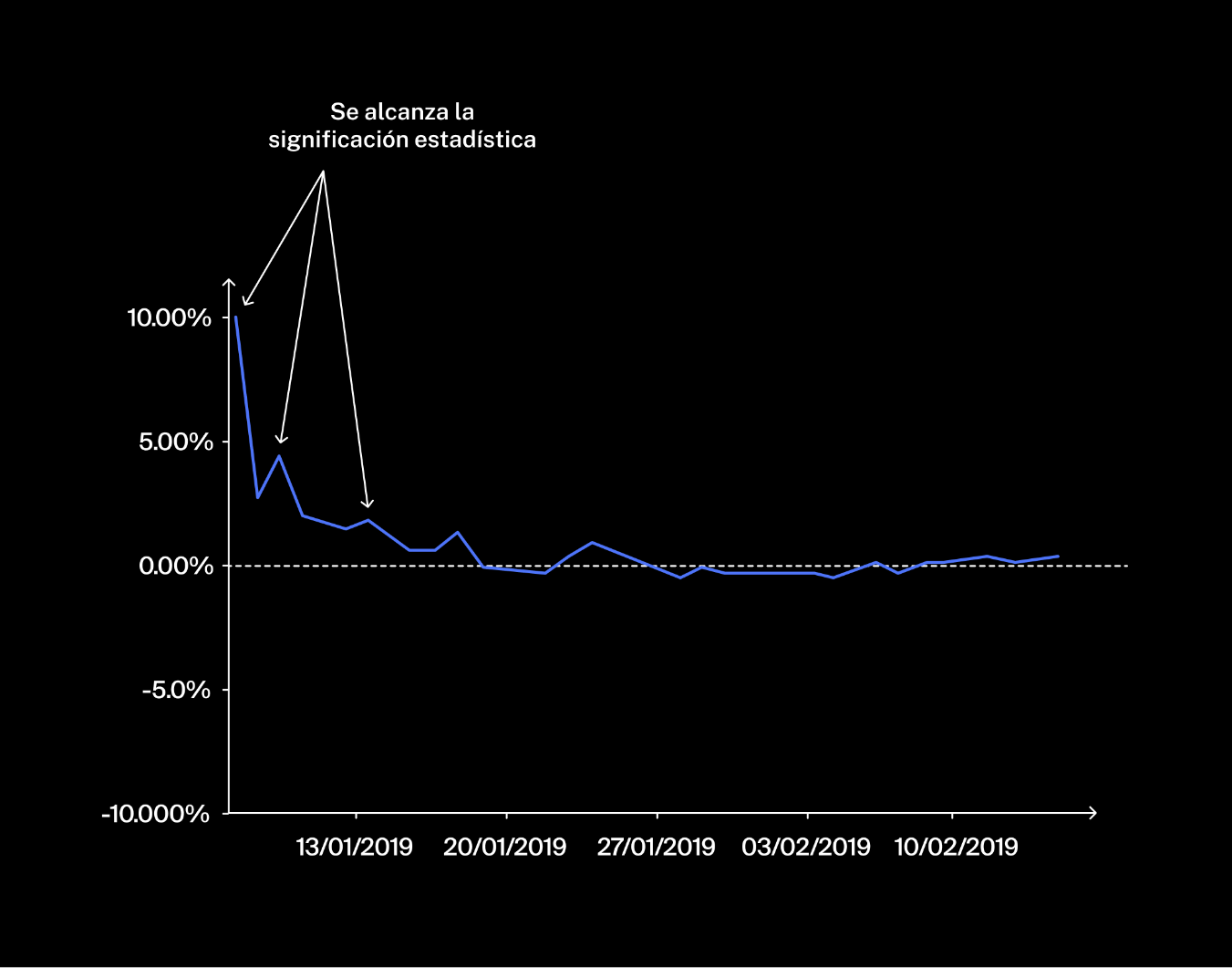
**Duración de las pruebas A/B**

**¿Cómo elegir la duración de las pruebas *A/B*?**

La implementación de una nueva funcionalidad cambia inevitablemente el comportamiento de los usuarios. Por lo general, los usuarios necesitan tiempo para acostumbrarse a los cambios. Una vez que se hayan acostumbrado por completo, podremos evaluar con certeza si el experimento ha sido un éxito. Cuantos más datos tengamos, menor será la probabilidad de error al comprobar las hipótesis estadísticas.

Las pruebas A/B sufren el llamado peeking problem (es decir, el problema de vislumbrar los resultados), que consiste en que el resultado final se distorsiona cuando se añaden nuevos datos al principio del experimento. Incluso un pequeño fragmento de datos nuevos es importante en relación con los datos ya acumulados y la significación estadística se alcanza en poco tiempo.

Esta es una de las manifestaciones de la ley de los grandes números. La dispersión tiende a ser mayor cuando el número de observaciones es escaso. En cambio, cuando tenemos un gran número de observaciones, el impacto de los valores atípicos se reduce. De este modo, los valores atípicos tienen una gran influencia en los resultados en caso de que la muestra sea demasiado pequeña. Cuando consideramos las pruebas estadísticas, este tipo de resultado significa que, sin una validación adecuada, podemos alcanzar nuestro umbral de valor p de forma incorrecta y, por tanto, rechazar erróneamente la hipótesis nula.

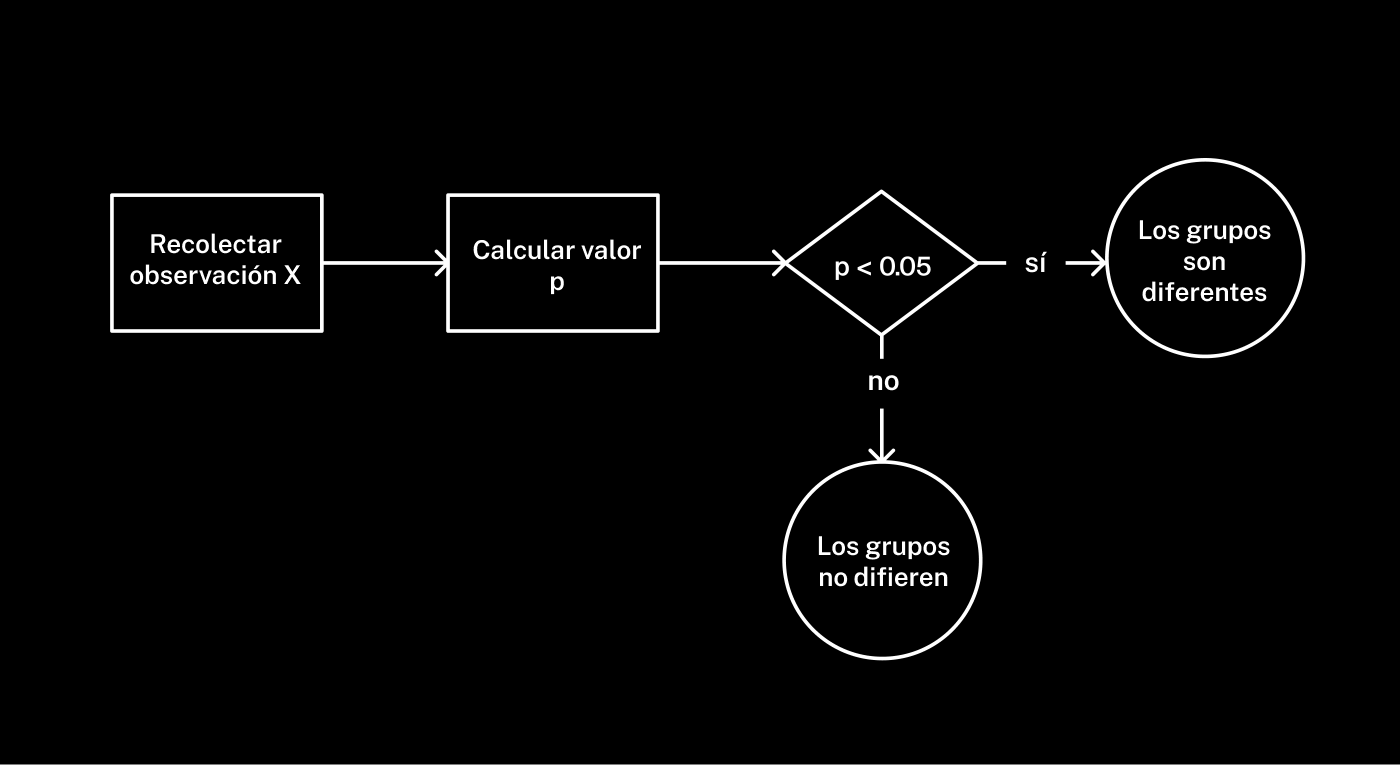


El gráfico anterior representa la diferencia de conversión entre segmentos obtenida mediante una prueba *A/B* simulada. Los datos fueron recogidos de la misma población, por lo que no debería haber diferencias en las medias de las muestras.

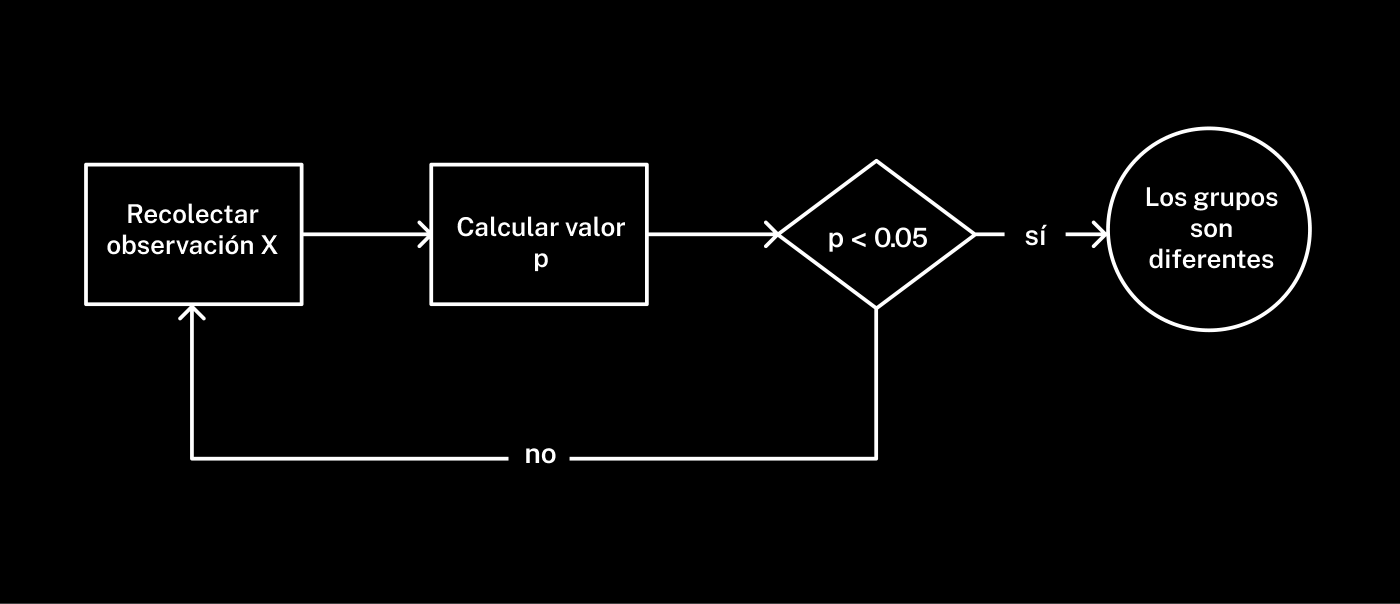
Sin embargo, debido a las fluctuaciones, la significación estadística se alcanza durante los primeros días de la prueba. Si se tratara de una prueba real, y no de una prueba simulada, la decisión tomada al alcanzar la significación estadística sería incorrecta.

Para solucionar el peeking problem, el tamaño de la muestra debe establecerse antes del inicio de la prueba.

Este es el procedimiento correcto de las pruebas *A/B*:



Y así es como no debemos realizar la prueba. Si añadimos nuevos datos al principio del experimento, aparecerá el peeking problem:



La forma más fácil de calcular el tamaño de la muestra es utilizar una calculadora en línea como esta: <https://vwo.com/tools/ab-test-duration-calculator/> (materiales en inglés)

Supongamos que estás probando un formulario de pedido para elegir entre 4 o 5 campos. Estás dispuesto a cambiar una parte de los datos del usuario recogidos por un aumento de la conversión. Ahora utiliza la calculadora:

* *La tasa de conversión existente estimada* es del 1 %.
* *La mejora mínima de la tasa de conversión que se quiere detectar* es el 5 % del cambio relativo en la conversión: 1.05 % frente al 1 % de la conversión.
* *Número de variaciones/combinaciones* equivale a 2 en nuestro caso. También hay pruebas A/B/n en las que se comparan varios grupos a la vez.
* *El número promedio de visitantes diarios* es de 500 000.
* *Porcentaje de visitantes incluidos en la prueba*: establece el valor en el 20% de tu número promedio de visitantes diarios, que será 500,000 \* 0.2= 100,000.

Necesitamos al menos 13 días para realizar la prueba.

Pregunta

En una nueva pestaña, abre la calculadora de duración de la prueba: <https://vwo.com/tools/ab-test-duration-calculator/>

Encuentra la duración mínima de la prueba para el cambio relativo de 2 % entre los grupos, siempre que:

1. La tasa de conversión existente estimada sea del 10 %
2. El número de grupos (variaciones) sea de 2
3. El número promedio de visitantes diarios sea de 50 000

Uno de cada dos visitantes participa en la prueba.

14 días

6 días

29 días

La prueba durará todo el mes. Tenemos que rastrear un pequeño cambio en la conversión para un número reducido de visitantes.

3 días

¡Perfecto!

**Comparación de las medias**

**Hemos realizado la prueba *A/B*. Ahora, ¿cómo podemos determinar si la nueva funcionalidad es útil? ¿Qué hipótesis debemos aceptar y cuál rechazar?**

Analicemos los resultados de la prueba *A/B*. ¿Qué describe el comportamiento de todos los usuarios? ¡El valor medio de la métrica! Podría ser el tiempo medio de navegación en el sitio web, el importe medio de las compras o el número medio de usuarios únicos.

Los resultados de las mediciones y los valores medios contienen un elemento aleatorio. Por lo tanto, tienen un componente de error aleatorio. No podemos predecir el valor exacto de cada observación con exactitud absoluta, pero podemos estimarlo utilizando métodos estadísticos.

Supongamos que nuestra hipótesis nula H₀ dice: la nueva funcionalidad no mejora las métricas. Entonces nuestra hipótesis correspondiente H₁ será: la nueva funcionalidad mejora las métricas.

En la fase de comprobación de la hipótesis, son posibles dos tipos de errores:

1. Error de tipo I: se produce cuando la hipótesis nula es correcta, pero se rechaza (resultado *falso positivo*. En este caso, la nueva funcionalidad se aprueba y, por lo tanto, es *positiva*)
2. Error de tipo II: se produce cuando la hipótesis nula es incorrecta, pero se acepta (resultado *falso negativo*)

Hipoˊtesis ciertaResultado de aplicarun criterio𝐻0𝐻1se aceptaConclusioˊn Conclusioˊn incorrecta 𝐻0correcta(Error de tipo II) 𝐻0𝐻0Conclusioˊn incorrectaConclusioˊn rechazada(Error de tipo I)correctaun criterioResultado de aplicar​​Hipoˊtesis ciertase acepta *H*0​ *H*0​rechazada​*H*0​Conclusioˊn correcta*H*0​Conclusioˊn incorrecta(Error de tipo I)​*H*1​Conclusioˊn incorrecta(Error de tipo II)Conclusioˊn correcta​​​

Para aceptar o rechazar la hipótesis nula, calculemos el nivel de significación, también conocido como valor p (valor de probabilidad). Muestra la probabilidad del error de tipo I, pero no revela nada sobre el error de tipo II.

Ten en cuenta que si el valor p es mayor que el valor de umbral, la hipótesis nula no debería rechazarse. Si es menor que el umbral, puede que no valga la pena aceptar la hipótesis nula. Los umbrales generalmente aceptados son del 5 % y del 1 %. Pero solo el data scientist toma la decisión final sobre qué umbral podría considerarse suficiente.

Los valores medios se comparan utilizando los métodos de prueba de hipótesis unilateral. La hipótesis unilateral se acepta si el valor que se está comprobando es mucho mayor o mucho menor que el de la hipótesis nula. A nosotros nos interesa la desviación en una sola dirección, que es "mayor que".

Si la distribución de los datos se aproxima a la normalidad (no hay valores atípicos significativos en los datos), se utiliza la prueba estándar para comparar las medias. Este método supone una distribución normal de las medias de todas las muestras y determina si la diferencia entre los valores comparados es lo suficientemente grande como para rechazar la hipótesis nula.

La tienda online de calzado KicksYouCanPayRentWith agregó a su sitio web una mascota animada de zapatillas deportivas. Va saltando por la pantalla mientras el usuario está navegando por los productos.

Observa dos muestras de importe promedio de compra antes y después de la implementación de la mascota. Imprime los valores medios anteriores y posteriores.

Comprueba la hipótesis de que el importe promedio de compra ha aumentado. Establece el nivel de significación en 5 %. Imprime en la pantalla el *valor p* y el resultado de la comprobación de la hipótesis.

**Pistas**

Realiza la prueba*t*: llama a la función st.ttest\_ind de la librería scipy.stats.

from scipy import stats as st

results = st.ttest\_ind(*# pasa las muestras anteriores y posteriores como argumentos)*

import pandas as pd

from scipy import stats as st

sample\_before = pd.Series([

436, 397, 433, 412, 367, 353, 440, 375, 414,

410, 434, 356, 377, 403, 434, 377, 437, 383,

388, 412, 350, 392, 354, 362, 392, 441, 371,

350, 364, 449, 413, 401, 382, 445, 366, 435,

442, 413, 386, 390, 350, 364, 418, 369, 369,

368, 429, 388, 397, 393, 373, 438, 385, 365,

447, 408, 379, 411, 358, 368, 442, 366, 431,

400, 449, 422, 423, 427, 361, 354])

sample\_after = pd.Series([

439, 518, 452, 505, 493, 470, 498, 442, 497,

423, 524, 442, 459, 452, 463, 488, 497, 500,

476, 501, 456, 425, 438, 435, 516, 453, 505,

441, 477, 469, 497, 502, 442, 449, 465, 429,

442, 472, 466, 431, 490, 475, 447, 435, 482,

434, 525, 510, 494, 493, 495, 499, 455, 464,

509, 432, 476, 438, 512, 423, 428, 499, 492,

493, 467, 493, 468, 420, 513, 427])

# Calcular las medias de ambas muestras

print("La media de antes:",sample\_before.mean())

print("La media de después:", sample\_after.mean())

# nivel crítico de significación

# la hipótesis se rechaza si el valor p es menor que ese

alpha = 0.05# < escribe tu código aquí >)

results = st.ttest\_ind(sample\_before, sample\_after, alternative='less')

# prueba unilateral (de una cola): el valor p será la mitad

pvalue = results.pvalue # < escribe tu código aquí >)

print('p-value: ', pvalue)

if pvalue < alpha:

print(

"La hipótesis nula se rechaza, a saber, es probable que el importe promedio de las compras aumente"

)

else:

print(

"La hipótesis nula no se rechaza, a saber, es poco probable que el importe medio de las compras aumente"

)

Resultado

La media de antes: 396.9714285714286

La media de después: 470.5285714285714

p-value: 3.090979110409132e-29

La hipótesis nula se rechaza, a saber, es probable que el importe promedio de las compras aumente

¡Es correcto!

Parece que a los amantes de las zapatillas deportivas les gusta la compañía cuando quieren comprar otro par (o dos). El importe promedio de compra ha aumentado.

**Intervalo de confianza**

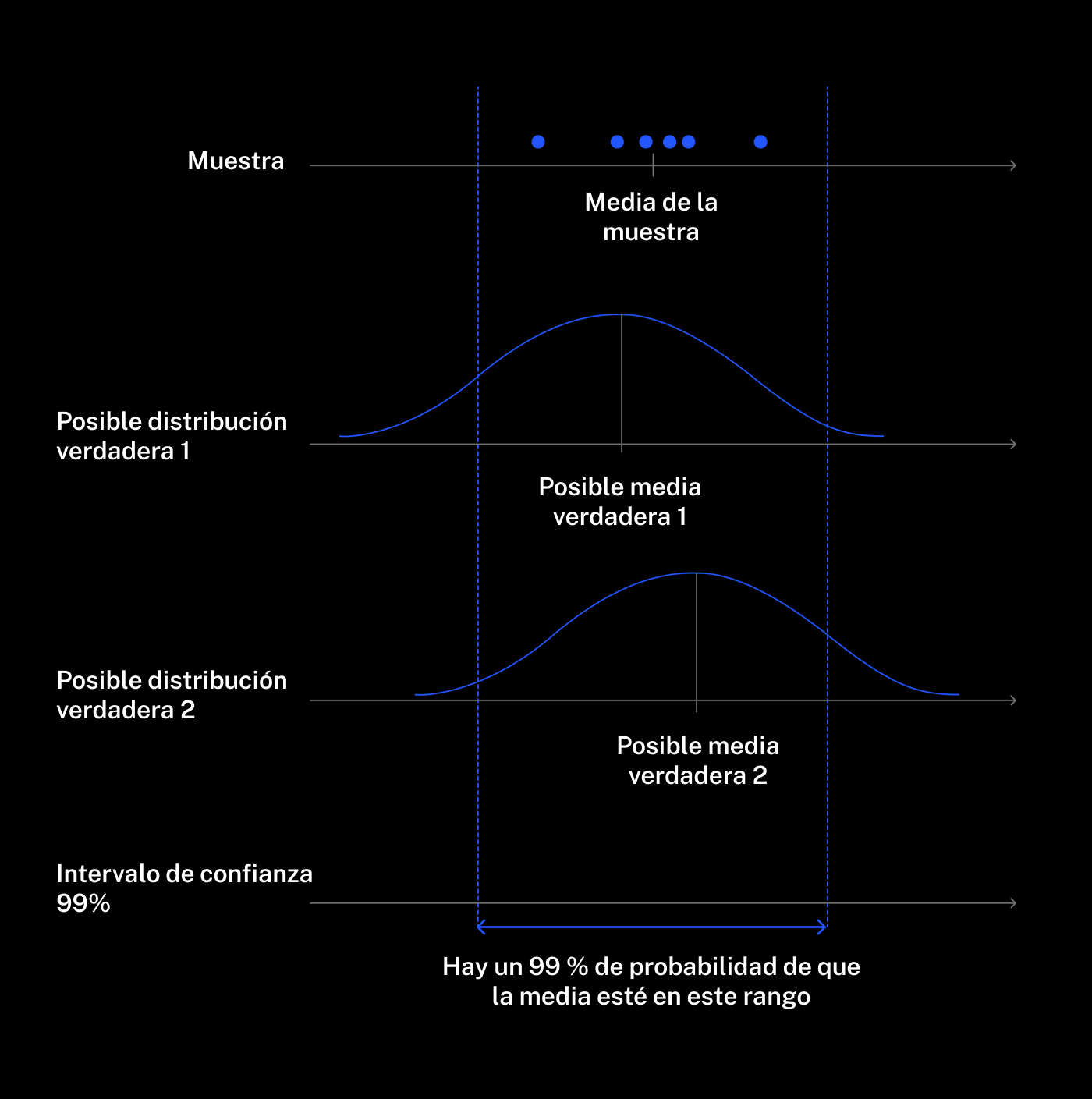
**En los negocios, es importante entender los riesgos y evaluar los posibles valores que puede tomar el parámetro de interés.**

Supongamos que tomamos una muestra y obtenemos una media de 472. Se trata de nuestra estimación de la media de toda la población, pero ¿qué podemos hacer para averiguar su exactitud? ¿Qué pasaría si accidentalmente tomáramos una muestra con una media muy alejada de la media real? Es poco probable, pero posible.

Un intervalo de confianza representa un segmento del eje numérico dentro del que cae el parámetro poblacional de interés, con una probabilidad predeterminada. El parámetro es desconocido, pero podemos estimarlo a partir de la muestra. Si el valor cae dentro del rango de 300 a 500 con una probabilidad del 99 %, entonces el intervalo de confianza del 99 % para este valor es (300, 500).

Al calcular el intervalo de confianza, normalmente descartamos la misma cantidad de valores de cada uno de sus extremos. Por ejemplo, para construir un intervalo de confianza del 99 %, tenemos que descartar el 1 % de los valores extremos, lo cual supone el 0.5 % de los valores más grandes y el 0.5 % de los más pequeños. El punto donde termina el 0.5 % de los valores más pequeños se denomina percentil de 0.5 %. El punto donde empieza el 0.5 % de los valores más grandes es el percentil de 99.5 %.

El intervalo de confianza no es solo un rango de valores aleatorios. El valor que evaluamos no es aleatorio debido a su diseño. Por ejemplo, el importe promedio de compra es un número único y constituye un parámetro del comportamiento del usuario sin dispersión ni varianza. La causa de la variabilidad radica en el hecho de que el número es desconocido y que se calcula a partir de la muestra. El carácter aleatorio de la muestra introduce aleatoriedad en la estimación. El intervalo de confianza mide la confianza en dicha estimación.



Pregunta

Elige la afirmación correcta:

El intervalo de confianza del importe promedio de compra es igual a la dispersión de los valores del importe de compra.

El intervalo de confianza del importe promedio de compra muestra qué valores puede tomar el verdadero importe promedio de compra como parámetro aleatorio.

El intervalo de confianza del importe promedio de compra representa un rango construido alrededor de la media verdadera del importe promedio de compra y que contiene la media de la muestra.

El intervalo de confianza del importe promedio de compra es un rango construido alrededor de la media de la muestra y contiene la media real del importe promedio de compra.

¡Correcto!

¡Buen trabajo!

Pregunta

¿Qué percentiles se deberían elegir para construir un intervalo de confianza del 95 %? Elige la respuesta correcta:

El percentil de 5 % superior y el percentil de 5 % inferior.

El percentil de 5 % y el percentil de 95 %

El percentil de 2.5 % superior y el percentil de 2.5 % inferior

El percentil de 2.5 % y el percentil de 97.5 %

¡Correcto! 97.5 % - 2.5 % = 95 %.

¡Perfecto!

**Cálculo del intervalo de confianza**

**Vamos a ver cómo construir un intervalo de confianza para la media, basado en la muestra, mediante el teorema central del límite.**

Supongamos que tomamos nuestra muestra a partir de una distribución con los siguientes parámetros:

𝜇=media poblacional*μ*=media poblacional

𝜎2=varianza poblacional*σ*2=varianza poblacional

Denota la media de la muestra:

𝑋ˉ=media poblacional*X*ˉ=media poblacional

El teorema del límite central dice que todas las medias de todas las muestras posibles con un tamaño *n* se distribuyen normalmente alrededor de la verdadera media poblacional. "Alrededor" significa que la media de esta distribución de todas las medias muestrales será igual a la verdadera media poblacional. La varianza será igual a la varianza poblacional dividida entre *n* (el tamaño de la muestra).

𝑋ˉ∼𝑁(𝜇,𝜎2𝑛)*X*ˉ∼**N**(*μ*,*nσ*2​)

La desviación estándar de esta distribución se denomina error estándar (*error estándar de la media*, o *SEM*, abreviado del "standard error of mean"):

SEM(𝑋ˉ)=𝜎𝑛SEM(*X*ˉ)=*n*​*σ*​

Cuanto mayor sea el tamaño de la muestra, menor será el error estándar, es decir, todas las medias muestrales estarán más cerca de la media real. Cuanto mayor sea la muestra, más precisa será la estimación.

Vamos a estandarizar esta distribución normal. Para obtener la distribución normal estándar con media = 0 y error estándar = 1, hay que restar la media y luego dividir el resultado entre el error estándar:

𝑋ˉ−𝜇SEM(𝑋ˉ)∼𝑁(0,12)SEM(*X*ˉ)*X*ˉ−*μ*​∼**N**(0,12)

A partir de la distribución normal estándar, toma el percentil de 5 % *F*(0.05) y el percentil de 95 % *F*(0.95) para obtener el intervalo de confianza del 90 %:

𝑃(𝐹(0.05)<𝑋ˉ−𝜇SEM(𝑋ˉ)<𝐹(0.95))=90%*P*(*F*(0.05)<SEM(*X*ˉ)*X*ˉ−*μ*​<*F*(0.95))=90%

Volvamos a escribirlo todo:

𝑃(𝑋ˉ+𝐹(0.05)⋅SEM(𝑋ˉ)<𝜇<𝑋ˉ+𝐹(0.95)⋅SEM(𝑋ˉ))=90%*P*(*X*ˉ+*F*(0.05)⋅SEM(*X*ˉ)<*μ*<*X*ˉ+*F*(0.95)⋅SEM(*X*ˉ))=90%

¡Aquí lo tenemos! El intervalo de confianza del 90 % para la media real.

Solo nos queda un problema. Para calcular el error estándar, utilizamos la varianza poblacional, pero la desconocemos al igual que la media poblacional. La estimamos a partir de la muestra.

Este hecho también afecta a la distribución de las medias muestrales de modo que, si la varianza es desconocida, no podemos utilizar la distribución normal y tenemos que describirla con la distribución de Student. Al poner en la fórmula el percentil de 5 % *t*(0.05) y el percentil de 95 % *t*(0.95), obtenemos:

𝑃(𝑋ˉ+𝑡(0.05)⋅SEM(𝑋ˉ)<𝜇<𝑋ˉ+𝑡(0.95)⋅SEM(𝑋ˉ))=90%*P*(*X*ˉ+*t*(0.05)⋅SEM(*X*ˉ)<*μ*<*X*ˉ+*t*(0.95)⋅SEM(*X*ˉ))=90%

Es posible simplificar el cálculo utilizando la distribución de Student *scipy.stats.t*. Tiene una función para el intervalo de confianza, *interval()*, que toma:

* *alpha*: nivel de significación
* *df*: número de grados de libertad (igual a n - 1)
* loc (de *localización*): la distribución media igual a la estimación media. Se calcula de la siguiente manera: sample.mean().
* *scale*: el error estándar de la distribución igual a la estimación del error estándar. Se calcula de la siguiente manera: sample.sem().

Ejercicio   
  
Construye un intervalo de confianza del 95 % para el importe promedio de compra en KicksYouCanPayRentWith después de la implementación de la mascota de las zapatillas.  
  
  
import pandas as pd

from scipy import stats as st

sample = pd.Series([

439, 518, 452, 505, 493, 470, 498, 442, 497,

423, 524, 442, 459, 452, 463, 488, 497, 500,

476, 501, 456, 425, 438, 435, 516, 453, 505,

441, 477, 469, 497, 502, 442, 449, 465, 429,

442, 472, 466, 431, 490, 475, 447, 435, 482,

434, 525, 510, 494, 493, 495, 499, 455, 464,

509, 432, 476, 438, 512, 423, 428, 499, 492,

493, 467, 493, 468, 420, 513, 427])

print('Media:', sample.mean())

confidence\_interval = st.t.interval(0.95, len(sample)-1, loc=sample.mean(), scale=st.sem(sample)) # < escribe tu código aquí >

print('Intervalo de confianza del 95 %:', confidence\_interval)  
  
Resultado

Media: 470.5285714285714

Intervalo de confianza del 95 %: (463.357753651609, 477.6993892055338)

¡Es correcto!

La mascota de las zapatillas de deporte nos ha proporcionado la dispersión de ingresos de solo 7 dólares.

Sprint 10

Capítulo 3/8

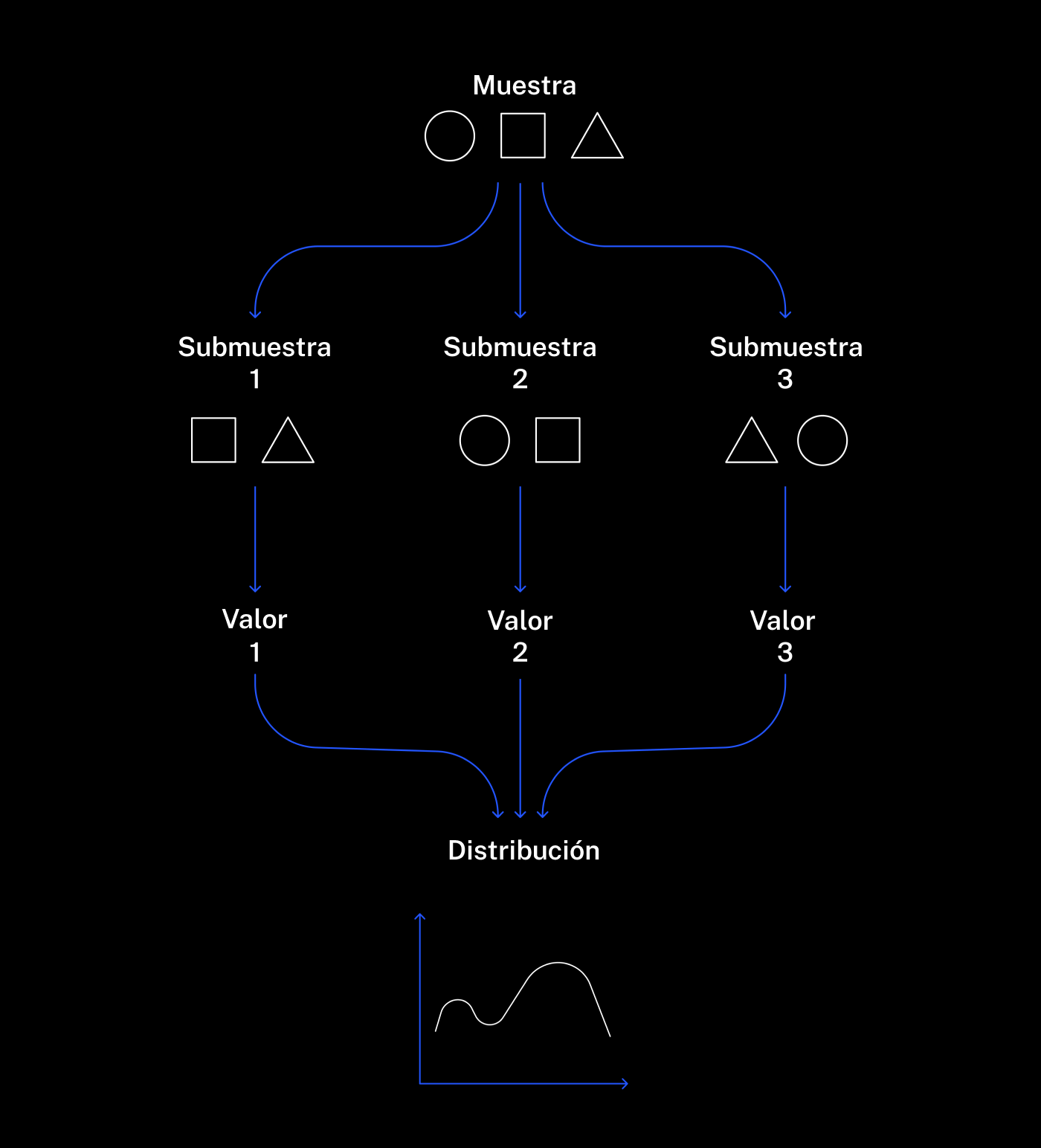
Implementar una nueva funcionalidad

**Bootstrapping**

**Los valores complejos se pueden calcular con la ayuda de la técnica del**bootstrapping**. Es la forma más fácil de hacerlo. Vamos a ver cómo funciona.**

Para conseguir un valor deseado, por ejemplo, la media, podemos obtener las submuestras (pseudomuestras) del conjunto fuente de datos. A continuación, calcularemos la media de cada una de ellas. En teoría, podemos formar submuestras y calcular el valor deseado a partir de ellas muchas veces. De este modo, podemos obtener varios valores para el parámetro de interés y estimar la distribución.

El nombre de la técnica proviene de la expresión en inglés *to pull oneself up by one’s bootstraps*, o *elevarse uno mismo tirando de los cordones de sus botas*. Esto significa que algo que parece imposible de conseguir sin ayuda externa, en realidad sí se puede hacer. Así es como pasó en los cuentos de Raspe cuando el barón Munchausen se sacó a sí mismo y a su caballo del pantano tirando de su propio pelo.



El bootstrapping es aplicable a cualquier muestra. Es útil cuando:

* Las observaciones no pueden ser descritas mediante una distribución normal;
* No hay pruebas estadísticas para el valor objetivo.

De hecho, no siempre se puede confiar en la distribución normal. Veamos la técnica del bootstrapping en el contexto de la evaluación del intervalo de confianza y el análisis de los resultados de una prueba *A/B*.

Pregunta

Elige la afirmación correcta:

El bootstrapping forma una muestra ampliada con base en varias muestras fuente.

El bootstrapping forma nuevas muestras con base en una muestra fuente.

¡Correcto!

El bootstrapping transforma la muestra fuente cambiando la distribución de los datos.

El bootstrapping comprueba si la distribución de los valores medios coincide con la muestra.

¡Bien hecho!

Pregunta

Elige la afirmación correcta:

El bootstrapping puede utilizarse a condición de que los datos se repartan según la distribución normal.

El bootstrapping no puede utilizarse si los datos se reparten de acuerdo con la distribución normal.

El bootstrapping funciona con cualquier distribución de datos.

¡Correcto!

¡Perfecto!

**Bootstrapping para el intervalo de confianza**

**Vamos a utilizar el bootstrapping para evaluar el intervalo de confianza.**

El Banco de Salvación Monetaria está probando un nuevo sistema de atención al cliente cuyo objetivo es determinar cuánto tiempo pasan los clientes en las colas de espera. Después de las pruebas, disponemos de una muestra de valores de tiempo de espera.

data = pd.Series([

10.7, 9.58, 7.74, 8.3 , 11.82, 9.74, 10.18, 8.43, 8.71,

6.84, 9.26, 11.61, 11.08, 8.94, 8.44, 10.41, 9.36, 10.85,

10.41, 8.37, 8.99, 10.17, 7.78, 10.79, 10.61, 10.87, 7.43,

8.44, 9.44, 8.26, 7.98, 11.27, 11.61, 9.84, 12.47, 7.8,

10.54, 8.99, 7.33, 8.55, 8.06, 10.62, 10.41, 9.29, 9.98,

9.46, 9.99, 8.62, 11.34, 11.21, 15.19, 20.85, 19.15, 19.01,

15.24, 16.66, 17.62, 18.22, 17.2, 15.76, 16.89, 15.22, 18.7,

14.84, 14.88, 19.41, 18.54, 17.85, 18.31, 13.68, 18.46, 13.99,

16.38, 16.88, 17.82, 15.17, 15.16, 18.15, 15.08, 15.91, 16.82,

16.85, 18.04, 17.51, 18.44, 15.33, 16.07, 17.22, 15.9, 18.03,

17.26, 17.6, 16.77, 17.45, 13.73, 14.95, 15.57, 19.19, 14.39,

15.76])

La métrica objetivo del experimento es el percentil de 99 % de esta distribución. Es decir, descartamos el 1 % de los valores más grandes y encontramos el máximo para el resto de los valores.

A la dirección del banco no le entusiasma la implementación de características poco fiables, así que hay que determinar el intervalo de confianza del 90 % para el percentil de 99 %.

Vamos a averiguar cómo formar submuestras para el bootstrapping. Ya conoces la función *sample()*. Para esta tarea necesitamos llamarla en un bucle. Pero aquí nos encontramos con un problema:

for i in range(5):

*# extrae un elemento aleatorio de la muestra 1*

*# especifica random\_state para su reproducción*

print(data.sample(1, random\_state=12345))

69 13.68

dtype: float64

69 13.68

dtype: float64

69 13.68

dtype: float64

69 13.68

dtype: float64

69 13.68

dtype: float64

Como especificamos el random\_state, el elemento aleatorio es siempre el mismo. Para solucionarlo, crea una instancia RandomState() del módulo numpy.random:

from numpy.random import RandomState

state = RandomState(12345)

Esta instancia se puede pasar al argumento random\_state de cualquier función. Es importante que, con cada nueva llamada, su estado cambie a aleatorio. Así obtendremos diferentes submuestras:

for i in range(5):

*# extrae un elemento aleatorio de la muestra 1*

print(data.sample(1, random\_state=state))

69 13.68

dtype: float64

27 8.44

dtype: float64

36 10.54

dtype: float64

84 18.44

dtype: float64

59 15.76

dtype: float64

Otro detalle importante a la hora de crear submuestras consiste en que deben proporcionar una selección de elementos con reemplazo. Es decir, el mismo elemento puede caer en una submuestra varias veces. Para ello, especifica replace=True para la función sample(). Compara:

example\_data = pd.Series([1, 2, 3, 4, 5])

print('Sin reemplazo')

print(example\_data.sample(frac=1, replace=False, random\_state=state))

print('Con reemplazo')

print(example\_data.sample(frac=1, replace=True, random\_state=state))

Sin reemplazo

1 2

4 5

3 4

0 1

2 3

dtype: int64

Con reemplazo

0 1

2 3

1 2

0 1

0 1

dtype: int64

1.

Utilizando la técnica del bootstrapping, crea 10 submuestras y un cuantil del 99 % para cada una de ellas. Imprímelas en la pantalla utilizando el string break.

Comprueba la función *quantile()* para las instancias *pandas.Series*.

import pandas as pd

import numpy as np

from numpy.random import RandomState

data = pd.Series([

10.7 , 9.58, 7.74, 8.3 , 11.82, 9.74, 10.18, 8.43, 8.71,

6.84, 9.26, 11.61, 11.08, 8.94, 8.44, 10.41, 9.36, 10.85,

10.41, 8.37, 8.99, 10.17, 7.78, 10.79, 10.61, 10.87, 7.43,

8.44, 9.44, 8.26, 7.98, 11.27, 11.61, 9.84, 12.47, 7.8 ,

10.54, 8.99, 7.33, 8.55, 8.06, 10.62, 10.41, 9.29, 9.98,

9.46, 9.99, 8.62, 11.34, 11.21, 15.19, 20.85, 19.15, 19.01,

15.24, 16.66, 17.62, 18.22, 17.2 , 15.76, 16.89, 15.22, 18.7 ,

14.84, 14.88, 19.41, 18.54, 17.85, 18.31, 13.68, 18.46, 13.99,

16.38, 16.88, 17.82, 15.17, 15.16, 18.15, 15.08, 15.91, 16.82,

16.85, 18.04, 17.51, 18.44, 15.33, 16.07, 17.22, 15.9 , 18.03,

17.26, 17.6 , 16.77, 17.45, 13.73, 14.95, 15.57, 19.19, 14.39,

15.76])

state = np.random.RandomState(12345)

for i in range(10):

subsample = data.sample(frac=1, replace=True, random\_state=state)# < escribe tu código aquí >

print(subsample.quantile(0.99))

Resultado

19.192200000000003

20.85

19.20660000000001

19.028400000000012

19.42440000000001

19.42440000000001

20.85

19.42440000000001

19.42440000000001

19.19

Es correcto!

De manera intuitiva, podemos estimar que encontrar el cuantil del 99 % nos llevó unos 19-20 minutos. Ahora vamos a encontrar el intervalo de confianza.

2.

Utilizando la técnica del bootstrapping, encuentra el intervalo de confianza del 90 % para el cuantil del 99 %. Guarda su valor más bajo en la variable lower, y el valor más alto en la variable upper. Imprímelos (en el precódigo).

Llama a la función quantile() dos veces: primero, para obtener el cuantil del 99 % de cada submuestra, y luego, para obtener el intervalo de confianza.

import pandas as pd

import numpy as np

from numpy.random import RandomState

data = pd.Series([

10.7 , 9.58, 7.74, 8.3 , 11.82, 9.74, 10.18, 8.43, 8.71,

6.84, 9.26, 11.61, 11.08, 8.94, 8.44, 10.41, 9.36, 10.85,

10.41, 8.37, 8.99, 10.17, 7.78, 10.79, 10.61, 10.87, 7.43,

8.44, 9.44, 8.26, 7.98, 11.27, 11.61, 9.84, 12.47, 7.8 ,

10.54, 8.99, 7.33, 8.55, 8.06, 10.62, 10.41, 9.29, 9.98,

9.46, 9.99, 8.62, 11.34, 11.21, 15.19, 20.85, 19.15, 19.01,

15.24, 16.66, 17.62, 18.22, 17.2 , 15.76, 16.89, 15.22, 18.7 ,

14.84, 14.88, 19.41, 18.54, 17.85, 18.31, 13.68, 18.46, 13.99,

16.38, 16.88, 17.82, 15.17, 15.16, 18.15, 15.08, 15.91, 16.82,

16.85, 18.04, 17.51, 18.44, 15.33, 16.07, 17.22, 15.9 , 18.03,

17.26, 17.6 , 16.77, 17.45, 13.73, 14.95, 15.57, 19.19, 14.39,

15.76])

state = np.random.RandomState(12345)

# Guarda los valores del cuantil del 99 % en la variable de valores

values = []

for i in range(1000):

subsample = data.sample(frac=1, replace=True, random\_state=state)

values.append(subsample.quantile(0.99))# cambia el tipo por comodidad

values = pd.Series(values)

lower = values.quantile(0.05) # < escribe tu código aquí >

upper = values.quantile(0.95) # < escribe tu código aquí >

print(lower)

print(upper)

Resultado

19.011400000000002

20.85

¡Es correcto!

Con una alta probabilidad, podemos suponer que el tiempo máximo de espera para el 99 % de los clientes es de entre 19 y 21 minutos. La estabilidad es buena. Por supuesto. ¿Pero qué pasa si contratamos a otro empleado?

**Bootstrapping para el análisis de pruebas A/B**

**El bootstrapping también se utiliza para analizar los resultados de las pruebas *A/B*.**

Digamos que hemos modificado ligeramente la interfaz de una tienda online. Necesitamos comprobar si el importe promedio de compra ha cambiado. Podemos utilizar esos datos para responder a la pregunta.

Mientras se realiza la prueba, recopilamos datos sobre el importe de cada compra en el grupo de control y en el grupo experimental.

Probablemente reconozcas que la comparación de dos grupos por sus medias en alguna variable se parece mucho a la prueba de hipótesis estadísticas. El bootstrapping no elimina el concepto de la hipótesis estadística, pero ofrece otra forma de probarlas.

Vamos a formular la hipótesis asociada a la pregunta, pero en términos de comparación de la diferencia entre los promedios, que es

𝐷=importe promedio de compra en 𝐵 − importe promedio de compra en 𝐴*D*=importe promedio de compra en *B* − importe promedio de compra en *A*

Entonces, la hipótesis estadística puede formularse como

𝐻0:𝐷=0𝐻𝐴:𝐷>0*H*0​:*HA*​:​*D*=0*D*>0​

La hipótesis nula consiste en que no hay diferencia entre los importes promedio de compra en los dos grupos. La hipótesis alternativa dice que el importe promedio de compra es mayor en el grupo experimental. Ahora bien, si calculamos simplemente la diferencia de forma aritmética, es probable que esta varíe ligeramente a causa de la aleatoriedad. Vamos a investigar cuál es la probabilidad de que dicha diferencia se haya obtenido por casualidad (este será nuestro valor p). Si el valor p está por debajo del umbral de significación estadística, podemos rechazar 𝐻0*H*0​ a favor de 𝐻𝐴*HA*​.

La prueba mediante el bootstrapping se realiza de la siguiente manera:

* Se calcula la diferencia entre los importes promedio de compra en los dos grupos.
* Dada la hipótesis 𝐻0*H*0​ suponemos que los grupos son iguales. Por consiguiente, las observaciones de ambos grupos pueden interpretarse como una sola muestra (extraída de la misma distribución), con lo cual concatenamos las observaciones en una muestra combinada.
* Simulamos numerosos experimentos seleccionando dos muestras (una para representar al grupo A y otra para representar al B) de la muestra combinada. Para cada experimento simulado, se calcula la diferencia del importe promedio de compra para esa iteración. Podemos llamarlo 𝐷𝑖*Di*​.
* El valor p puede estimarse entonces como la razón entre el número de veces cuando 𝐷𝑖*Di*​ no fue inferior a 𝐷*D* y el número de experimentos simulados.

Si la disparidad entre los promedios de las muestras originales (D) es simplemente un resultado de la casualidad (aleatoriedad), el proceso aleatorio (combinando las muestras, dibujando nuevas muestras aleatorias y calculando la diferencia entre sus promedios) es probable que no se encuentre con ninguna dificultad para reunirse o superar D. Habrá un número sustancial de casos en los que esto ocurra. Sin embargo, si la diferencia es genuina y no solo un resultado de la aleatoriedad (por ejemplo, refleja un verdadero efecto de tratamiento), el bootstrapping puede tener dificultades para producir datos remuestreados con diferencias similares entre las medias, ya que no puede crear nueva información más allá de lo que está presente en los datos originales.

Calculamos la *diferencia real de los importes promedio de compra* entre los grupos de acuerdo con la fórmula anterior.

𝐷=importe promedio de compra en 𝐵 − importe promedio de compra en 𝐴*D*=importe promedio de compra en *B* − importe promedio de compra en *A*

A continuación, combinamos las observaciones originales de los grupos A y B.

Determinamos el número de experimentos simulados. Normalmente, este oscila entre cientos y miles. Vamos a utilizar el índice $i$ para referirnos al número de cualquier experimento simulado.

Para llevar a cabo un experimento simulado, se extraen dos muestras aleatorias (con reemplazo) del conjunto combinado de observaciones.

* 𝐴𝑖*Ai*​: la muestra del grupo A en el *i-*º experimento, su tamaño debe ser igual al tamaño de la muestra original del grupo A (según lo observado),
* 𝐵𝑖*Bi*​: la muestra del grupo B para el *i-*º experimento, su tamaño debe ser igual al tamaño de la muestra original del grupo B (según lo observado)

Encuentra la diferencia del importe promedio de compra entre estos dos grupos:

𝐷𝑖=importe promedio de compra en 𝐵𝑖 − importe promedio de compra en 𝐴𝑖*Di*​=importe promedio de compra en *Bi*​ − importe promedio de compra en *Ai*​

Encuentra la estimación del valor p:

p-value=𝑃{𝐷𝑖≥𝐷}=nuˊmero de experimentos simulados cuando 𝐷𝑖≥𝐷nuˊmero de experimentos simuladosp-value=*P*{*Di*​≥*D*}=nuˊmero de experimentos simuladosnuˊmero de experimentos simulados cuando *Di*​≥*D*​

Si el valor p es inferior a un determinado umbral (normalmente, 0,05), podemos rechazar la hipótesis nula y decir que existe una diferencia significativa entre los dos grupos.

Analiza las dos muestras y comprueba la hipótesis que dice que el importe promedio de compra ha aumentado. Guarda la diferencia entre los importes promedio de compra en la variable AB\_difference e imprímela en la pantalla. Asigna un nivel de significación del 5 % (.05) a la variable alpha. Guarda el valor p en la variable pvalue e imprímelo. Imprime el resultado de la prueba de hipótesis.

Concatena las muestras de la siguiente manera:

united\_samples = pd.concat([samples\_A, samples\_B])

Encuentra la submuestra:

subsample = united\_samples.sample(frac=1, replace=True, random\_state=state)

Divídela por la mitad:

subsample\_A = subsample[:len(samples\_A)]

subsample\_B = subsample[len(samples\_A):]

import pandas as pd

import numpy as np

# datos del grupo de control A

samples\_A = pd.Series([

98.24, 97.77, 95.56, 99.49, 101.4 , 105.35, 95.83, 93.02,

101.37, 95.66, 98.34, 100.75, 104.93, 97. , 95.46, 100.03,

102.34, 98.23, 97.05, 97.76, 98.63, 98.82, 99.51, 99.31,

98.58, 96.84, 93.71, 101.38, 100.6 , 103.68, 104.78, 101.51,

100.89, 102.27, 99.87, 94.83, 95.95, 105.2 , 97. , 95.54,

98.38, 99.81, 103.34, 101.14, 102.19, 94.77, 94.74, 99.56,

102. , 100.95, 102.19, 103.75, 103.65, 95.07, 103.53, 100.42,

98.09, 94.86, 101.47, 103.07, 100.15, 100.32, 100.89, 101.23,

95.95, 103.69, 100.09, 96.28, 96.11, 97.63, 99.45, 100.81,

102.18, 94.92, 98.89, 101.48, 101.29, 94.43, 101.55, 95.85,

100.16, 97.49, 105.17, 104.83, 101.9 , 100.56, 104.91, 94.17,

103.48, 100.55, 102.66, 100.62, 96.93, 102.67, 101.27, 98.56,

102.41, 100.69, 99.67, 100.99])

# datos del grupo experimental B

samples\_B = pd.Series([

101.67, 102.27, 97.01, 103.46, 100.76, 101.19, 99.11, 97.59,

101.01, 101.45, 94.8 , 101.55, 96.38, 99.03, 102.83, 97.32,

98.25, 97.17, 101.1 , 102.57, 104.59, 105.63, 98.93, 103.87,

98.48, 101.14, 102.24, 98.55, 105.61, 100.06, 99. , 102.53,

101.56, 102.68, 103.26, 96.62, 99.48, 107.6 , 99.87, 103.58,

105.05, 105.69, 94.52, 99.51, 99.81, 99.44, 97.35, 102.97,

99.77, 99.59, 102.12, 104.29, 98.31, 98.83, 96.83, 99.2 ,

97.88, 102.34, 102.04, 99.88, 99.69, 103.43, 100.71, 92.71,

99.99, 99.39, 99.19, 99.29, 100.34, 101.08, 100.29, 93.83,

103.63, 98.88, 105.36, 101.82, 100.86, 100.75, 99.4 , 95.37,

107.96, 97.69, 102.17, 99.41, 98.97, 97.96, 98.31, 97.09,

103.92, 100.98, 102.76, 98.24, 97. , 98.99, 103.54, 99.72,

101.62, 100.62, 102.79, 104.19])

# diferencia real entre las medias de los grupos

AB\_difference = samples\_B.mean() - samples\_A.mean()# < escribe tu código aquí >

print("Diferencia entre los importes promedios de compra:", AB\_difference)

alpha = 0.05# <escribe aquí el nivel de significación>

state = np.random.RandomState(12345)

bootstrap\_samples = 1000

count = 0

for i in range(bootstrap\_samples):

# concatena las muestras

united\_samples = pd.concat([samples\_A, samples\_B])# < escribe tu código aquí >

# crea una submuestra

subsample = united\_samples.sample(frac=1, replace=True, random\_state=state)# < escribe tu código aquí >

# divide la submuestra por la mitad

subsample\_A =subsample[:len(samples\_A)]

subsample\_B = subsample[len(samples\_A):]

# encuentra la diferencia entre las medias

bootstrap\_difference = subsample\_B.mean() - subsample\_A.mean()# < escribe tu código aquí >

# si la diferencia no es menor que la diferencia real, añade "1" al contador

if bootstrap\_difference >= AB\_difference:

count += 1

# el valor p es igual al porcentaje de valores excedentes

pvalue = 1. \* count / bootstrap\_samples

print('p-value =', pvalue)

if pvalue < alpha:

print("La hipótesis nula se rechaza, a saber, es probable que el importe promedio de las compras aumente")

else:

print("La hipótesis nula no se rechaza, a saber, es poco probable que el importe medio de las compras aumente")

Resultado

Diferencia entre los importes promedios de compra: 0.7682000000000357

p-value = 0.034

La hipótesis nula se rechaza, a saber, es probable que el importe promedio de las compras aumente

¡Es correcto!

El valor p es menor o igual que el nivel de significación. ¡La funcionalidad está lista para la producción!

**Bootstrapping para modelos**

**Ha llegado el momento de utilizar el bootstrapping para evaluar los intervalos de confianza en los modelos de ML.**

Channel Tunnel es una escuela de francés que ofrece clases exprés de francés a un precio razonable. Está en el proceso de desarrollo de un modelo que determine la probabilidad de que un alumno asista o no a una clase. Diariamente se reciben muchas solicitudes. La prioridad se asigna por orden de solicitud, o sea, la primera en llegar, será la primera en ser atendida. Aproximadamente la mitad de quienes solicitan una clase, no se presentan y no la pagan. La dirección de la escuela ha decidido impartir clases solo a estudiantes con mayor probabilidad de asistir a clase. Debido a los posibles riesgos para su reputación, la empresa introducirá el nuevo sistema únicamente a condición de que se demuestre el aumento de los ingresos. Para tomar la decisión correcta, hay que evaluar la distribución de probabilidad de los ingresos.

Aquí tienes las condiciones importantes para la tarea:

* El modelo de predicción de la probabilidad de asistencia a clase ya está entrenado. Las predicciones se encuentran en el archivo eng\_probabilities.csv, y las respuestas correctas en eng\_target.csv.
* El coste de una lección es de 10 dólares. Se pueden programar hasta 10 lecciones por día. Los ingresos diarios actuales ascienden a 50 dólares (la mitad de los estudiantes cancelan la lección).
* La media diaria de solicitudes recibidas es de 25.
* El ingreso objetivo para la implementación del nuevo sistema se establece en 75 dólares, y la probabilidad de alcanzar este objetivo debe ser de al menos el 99%.

1.

Escribe la función *revenue()* que calcula y devuelve el valor de los ingresos. Esta función utiliza:

* la lista de respuestas (*target*): si el estudiante asistirá a la clase
* la lista de probabilidades (*probabilities*): el modelo evalúa si el estudiante asistirá o no
* el número de estudiantes que asisten a las clases por día (*count*).

La función debe seleccionar a los estudiantes con mayor probabilidad de asistir a clase y, basándose en las respuestas, calcular los posibles ingresos. Ten en cuenta que la función toma series de datos así que no es necesario utilizar ningún dataset para esta tarea.

En el precódigo tenemos un ejemplo de ejecución de la función en el que las listas de respuestas y probabilidades son cortas y el número de estudiantes es solo de 3.

import pandas as pd

def revenue(target, probabilities, count):

#Por comodidad, primero se ordenan las probabilidades:

probs\_sorted = probabilities.sort\_values(ascending=False)

#Luego, para seleccionar las respuestas, toma sus índices

selected = target[probs\_sorted.index][:count]

return 10 \* selected.sum()# < escribe tu código aquí >

target = pd.Series([1, 1, 0, 0, 1, 0])

probab = pd.Series([0.2, 0.9, 0.8, 0.3, 0.5, 0.1])

res = revenue(target, probab, 3)

print(res)  
  
Resultado

20

¡Es correcto!

Ahora podemos calcular los ingresos. Así que estamos a mitad de camino. Vamos a averiguar cómo calcular la distribución de los ingresos.

2.

Para encontrar el cuantil de ingresos del 1 %, realiza el proceso de bootstrapping con 1000 repeticiones.

Guarda la lista de estimaciones del bootstrapping en la variable *values* y el cuantil del 1 % en la variable *lower*. Imprime los ingresos promedio y el cuantil del 1 % (en precódigo)

import pandas as pd

import numpy as np

# Abre los archivos

# toma el índice “0” para convertir los datos a pd-Series

target = pd.read\_csv('/datasets/eng\_target.csv')['0']

probabilities = pd.read\_csv('/datasets/eng\_probabilites.csv')['0']

def revenue(target, probabilities, count):

probs\_sorted = probabilities.sort\_values(ascending=False)

selected = target[probs\_sorted.index][:count]

return 10 \* selected.sum()

state = np.random.RandomState(12345)

values = []

for i in range(1000):

# Muestreo con reemplazo

target\_subsample = target.sample(n=25, replace=True, random\_state=state)

probs\_subsample = probabilities[target\_subsample.index]

# Calcular los ingresos para el conjunto de bootstrap

values.append(revenue(target\_subsample, probs\_subsample, 10))

values = pd.Series(values)

lower = values.quantile(0.01) # Cuantil del 1%

mean = values.mean()

print("Ingresos promedio:", mean)

print("Cuantil del 1 %:", lower)

Resultado

Ingresos promedio: 91.67

Cuantil del 1 %: 60.0

¡Es correcto!

Casi 100 dólares. El ingreso promedio se acerca al máximo posible. Y con una probabilidad del 99 %, podemos obtener ingresos de al menos 60 dólares.

Parece que todavía no es un buen momento para cambiar las cosas.

Sprint 10

Capítulo 3/8 · Última lección

Implementar una nueva funcionalidad

**Conclusión**

En este capítulo aprendiste a:

* Analizar los resultados de las pruebas *A/B*;
* Calcular los intervalos de confianza;
* Calcular valores complejos utilizando el bootstrapping.

En el próximo capítulo, aprenderás a recopilar los datos y a utilizarlos para probar modelos con mayor fiabilidad.

**Llévate esto contigo**

Descárgate el resumen del capítulo y la hoja informativa para poder consultarlos cuando los necesites.

* [Resumen del capítulo: Implementación de nuevas funciones](https://practicum-content.s3.us-west-1.amazonaws.com/new-markets/DS_sprint_9/moved_Resumen_del_captulo_Implementacin_de_nuevas_funciones.pdf)
* [Hoja informativa: Implementación de nuevas funciones](https://practicum-content.s3.us-west-1.amazonaws.com/new-markets/DS_sprint_9/moved_Hoja_informativa_Implementacin_de_una_nueva_funcionalidad.pdf)

Sprint 10

Capítulo 4/8

Recopilación de datos

**Introducción**

**Para entrenar modelos, primero debemos recopilar datos.**

**Aprenderás a:**

* controlar la calidad del etiquetado por mayoría de votos;
* evitar la fuga de información;
* utilizar la validación cruzada para mejorar la fiabilidad de la evaluación del modelo.

**¿Cuánto tiempo tomará?**

9 lecciones de 5-10 minutos cada una.

**Descripción del ejercicio**

Utilizaremos diferentes casos para investigar diferentes métodos de recopilación de datos.

Capítulo 4/8

Recopilación de datos

**Fuentes de datos**

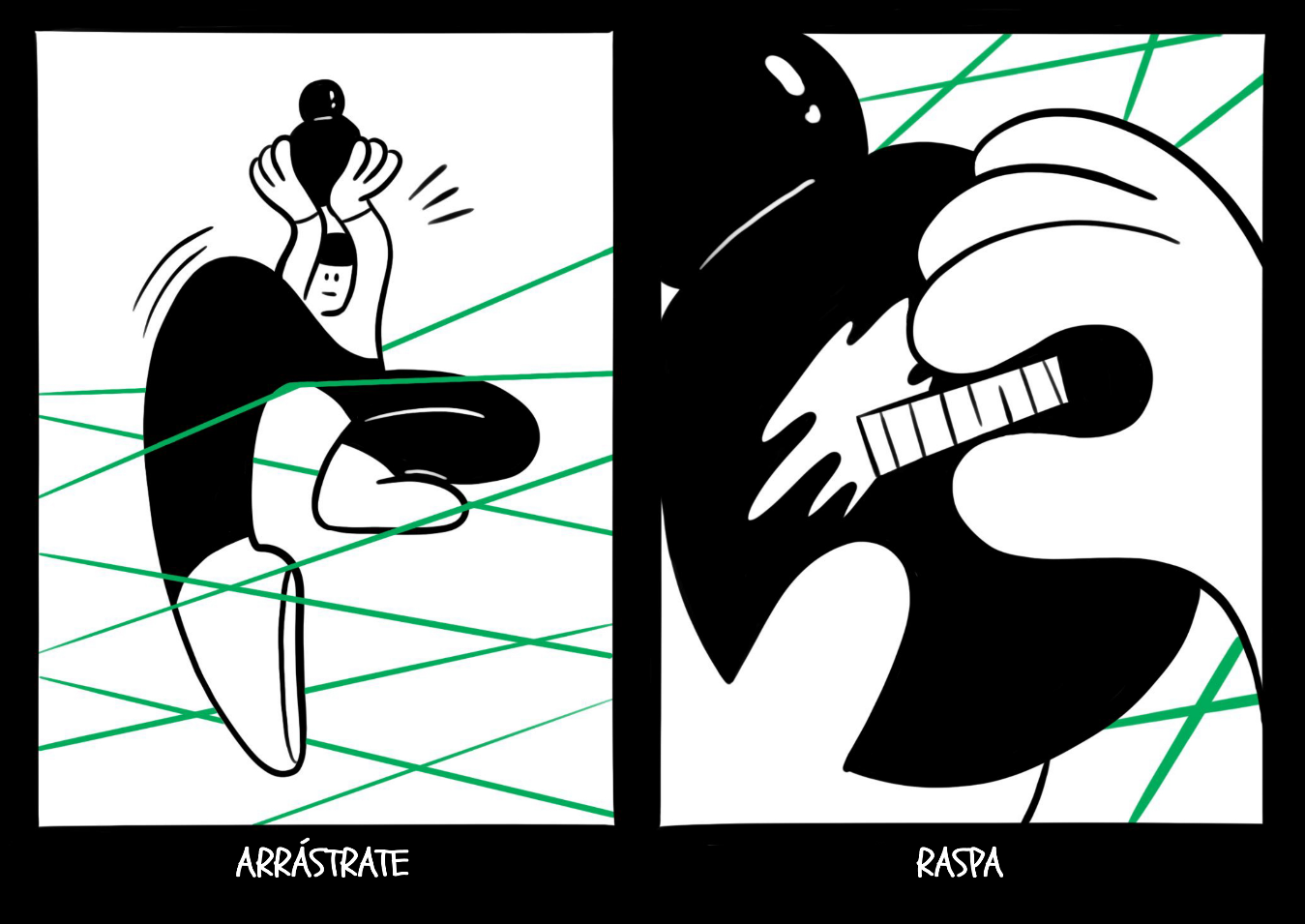
**Sin datos no hay machine learning. ¿De dónde los podemos obtener?**

¡Desde el almacén de datos de la empresa! Por ejemplo, un banco quiere construir un modelo para predecir los pagos de préstamos. Ya tiene un historial de atención a diferentes clientes: sus perfiles e información sobre pagos. Todo lo que tenemos que hacer es extraer ese historial y prepararnos para el análisis. Pero ese es un escenario bastante optimista.

La mayoría de las veces, la empresa no puede proporcionar los datos. Si la tarea es común, los conjuntos de datos se pueden encontrar en fuentes abiertas (materiales en inglés):

* [Plataforma de competencia Kaggle Data Science;](https://www.kaggle.com/)
* [Repositorio de machine learning UC Irvine;](https://archive.ics.uci.edu/ml/index.php)
* [Base de datos abierta del gobierno de los EE. UU.](https://www.data.gov/about)
* [FiveThirtyEight: datos abiertos sobre análisis de encuestas de opinión, política, economía y más.](https://data.fivethirtyeight.com/)

Para algunas tareas, los datos se pueden recopilar en Internet. Descarga todas las páginas del portal requerido y usa el software crawler (rastreador) o scraper (extractor) para extraer los datos. Por ejemplo, para construir un modelo que funcione con vocabulario coloquial, puedes usar publicaciones de Twitter.



Sprint 10

Capítulo 4/8

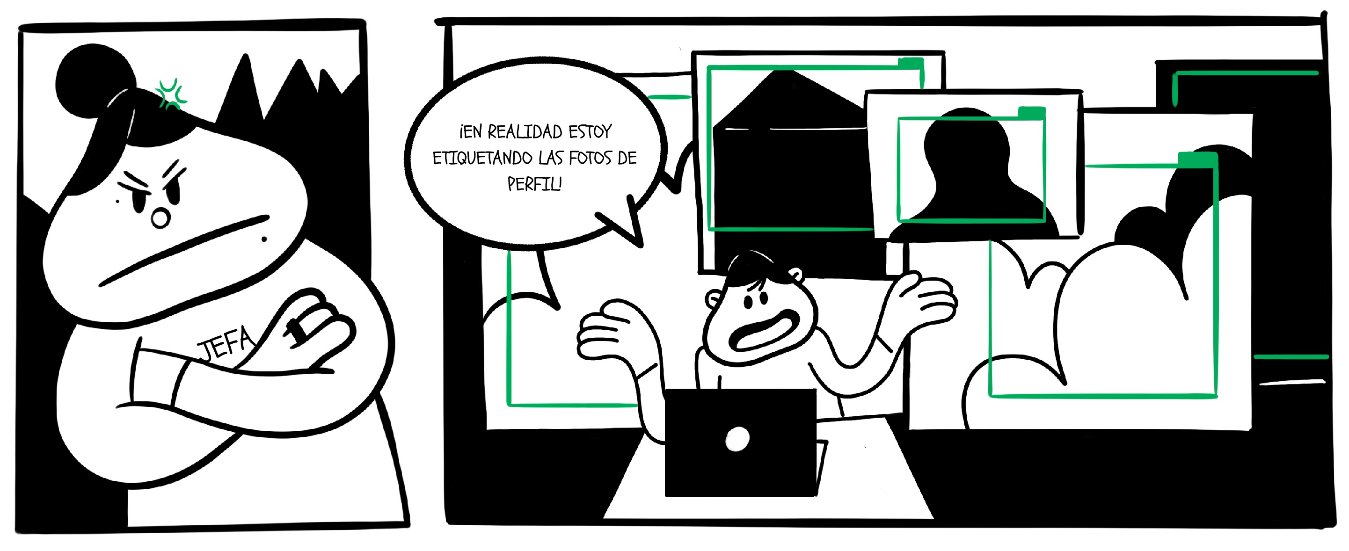
Recopilación de datos

**Etiquetado de datos**

**Si no puedes encontrar los datos, créalos.**

Pregúntale a tus colegas. Quizás la empresa tiene los datos, pero le faltan los valores objetivo. Dichos datos no están etiquetados, pero sí puedes usarlos. Supongamos que una red social planea entrenar a un modelo que reconocería si hay una persona en la foto de perfil o no. Todas las fotos en el almacén son datos sin etiquetar con los que podemos trabajar.

Para obtener un conjunto de entrenamiento, debemos realizar etiquetado de datos o anotación de datos. Es decir, establecer la respuesta correcta para cada foto (una persona o algo más). Esto etiquetará efectivamente los datos. En ocasiones, el etiquetado se puede realizar sin conocimientos especiales (como es el caso de las fotos de perfil), pero cuando los datos se relacionan con la salud y el bienestar de las personas, es posible que necesites asesoría profesional.



Pregunta

Elige los enunciados correctos:

Los datos sin etiqueta son datos con valores ausentes.

Los datos etiquetados son los datos con pesos de clase ajustados.

El etiquetado introduce el objetivo en los datos.

¡Correcto!

El etiquetado aumenta el número de observaciones en datos no etiquetados.

¡Tu comprensión del material es impresionante!

Hay servicios en línea dedicados para el etiquetado. Los datos sin etiquetar se cargan en el recurso y se especifica el precio del etiquetado por observación. Cualquiera puede ir al servicio, etiquetar los datos y cobrar por ello.

Estos son algunos servicios populares (materiales en inglés):

* [Amazon Mechanical Turk](https://www.mturk.com/)
* [Y.Toloka](https://toloka.yandex.com/)

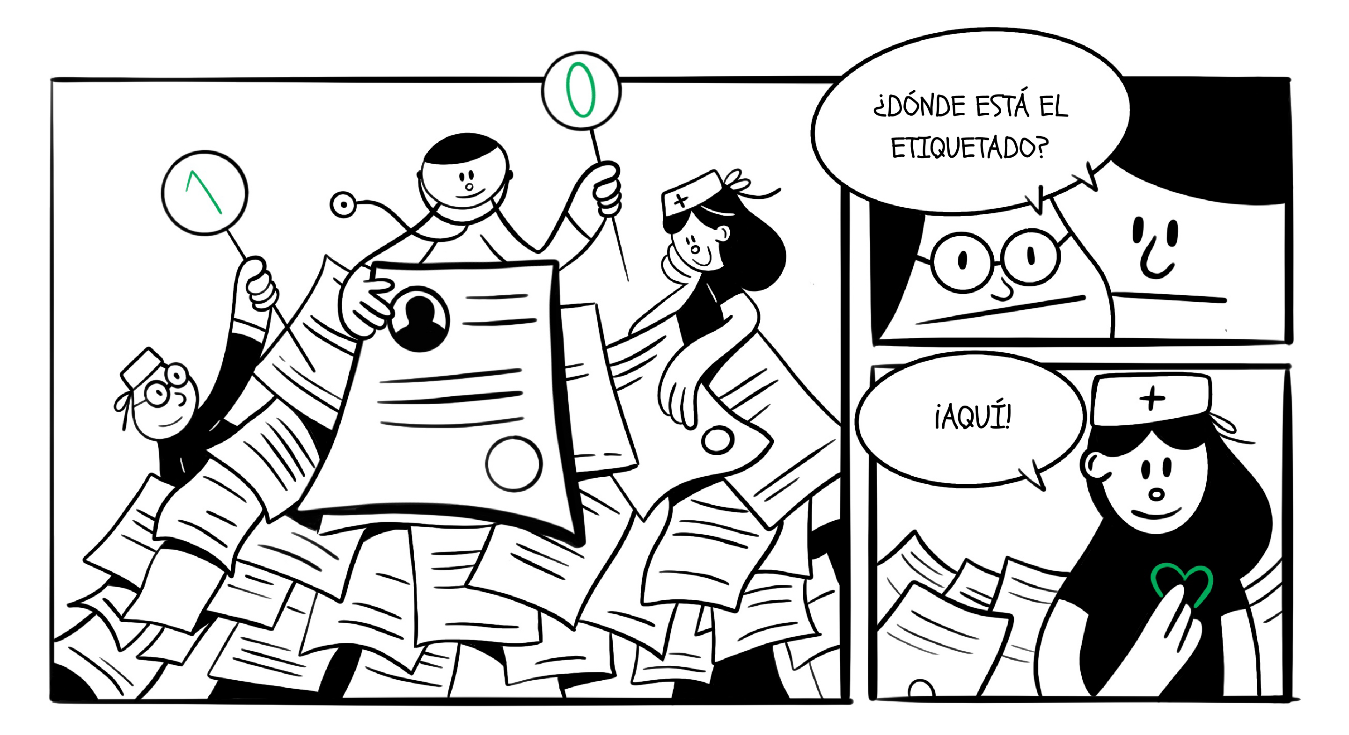
**Control de calidad de etiquetado**

**¿Qué pasa si el etiquetador comete un error? Afectará las predicciones del modelo. Averigüemos cómo evitar esto.**

La calidad de los datos después del etiquetado se puede mejorar utilizando los métodos para el control de calidad del etiquetado. ¿Cómo funcionan? Todas las observaciones, o una parte de ellas, se etiquetan varias veces y luego se forma la respuesta final.

Veamos uno de esos métodos, el voto mayoritario. ¿Quién está "votando" y cómo? Por ejemplo, cada observación está etiquetada por tres evaluadores. La respuesta final es la elegida por la mayoría.

Aquí verás cómo este método trata con un conjunto de datos. Una empresa médica lanza un sistema para el diagnóstico automatizado de enfermedades del corazón. Los datos de 303 pacientes están etiquetados por tres profesionales de la salud.



Las siguientes características están disponibles para nosotros:

* *age* — edad del paciente
* *sex* — sexo biológico del paciente
* *cp* — tipo de dolor en el pecho
* *trestbps* — presión arterial en reposo
* *chol* — colesterol sérico
* *fbs* — azúcar en sangre en ayunas (si > 120 mg/dl)
* *restecg* — resultados electrocardiográficos en reposo
* *thalach* — frecuencia cardíaca máxima alcanzada
* *exang* — angina inducida por el ejercicio
* *oldpeak* — depresión del ST inducida por el ejercicio en relación con el reposo
* *slope* — la pendiente del segmento ST del ejercicio máximo
* *ca* — número de vasos principales (0-3) coloreados por fluoroscopia
* *thal* — Resultado de la prueba de esfuerzo con talio

Etiquetado:

* *label\_1* — respuesta del Profesional 1
* *label\_2* — respuesta del Profesional 2
* *label\_3* — respuesta del Profesional 3

Ejercicio

Realiza el voto mayoritario para el conjunto de datos. Almacena el objetivo en la variable objetivo. Muestra en pantalla las cinco primeras filas de la tabla resultante (en precódigo).

Hay muchas soluciones al problema, elige la que más te convenga.

import pandas as pd

data = pd.read\_csv('/datasets/heart\_labeled.csv')

target = []

for i in range(data.shape[0]):

labels = data.loc[i, ['label\_1', 'label\_2', 'label\_3']]

true\_label = int(labels.mode()[0]) # < escribe tu código aquí >

target.append(true\_label)

data['target'] = target# < escribe tu código aquí >

print(data.head())

Resultado

age sex cp trestbps chol ... thal label\_1 label\_2 label\_3 target

0 50 1 0 150 243 ... 3 0 0 1 0

1 55 0 1 135 250 ... 2 1 1 1 1

2 54 1 0 140 239 ... 2 1 0 1 1

3 59 1 3 178 270 ... 3 1 0 1 1

4 70 1 2 160 269 ... 3 1 0 0 0

[5 rows x 17 columns]

Es correcto!

Un experto es bueno, pero tres son mejores. ¡Y 33 lo son mucho más!

Capítulo 4/8

Recopilación de datos

**Fuga de información**

**Ya recopilaste los datos. Ahora podemos verificar si hay**fugas de objetivos**.**

La fuga de información ocurre cuando la información sobre el objetivo se filtra accidentalmente en las funciones.

Por ejemplo: el juego online "Age of Grinding" está desarrollando un modelo que predecirá qué jugadores van a hacer compras en el juego para subir de nivel a sus personajes, logrando que jueguen durante más horas. Si el modelo predice que el jugador no va a realizar compras, se le ofrece un código de descuento después de dos horas de juego. Digamos que hemos recopilado el historial de comportamiento del usuario y es una tabla con las siguientes características:

* Edad
* Sexo
* Sobrenombre
* Número de compras en otros juegos de la plataforma
* Nivel del personaje
* Si el jugador completó la misión de entrenamiento (primera hora del juego)
* Horas totales dedicadas al juego
* Número de publicaciones en el foro del juego

Pregunta

¿Qué características no se conocen antes de las predicciones del modelo?

Elige tantas como quieras

Edad

Sexo

Sobrenombre

Número de compras en otros juegos de la plataforma

Nivel del personaje

Sí, el nivel del personaje depende de si el jugador ha gastado previamente dinero real en el juego o no.

Si el jugador completó la misión de entrenamiento

Horas totales dedicadas al juego

Esta característica también depende indirectamente del pago: los jugadores que gastan dinero real tienden a pasar más tiempo en el juego.

¡Bien hecho!

Entonces, las características "Nivel del personaje" y "Total de horas dedicadas al juego" se conocen después de que el modelo hace predicciones. Para evitar la fuga de información, no utilices dichas funciones.

Capítulo 4/8

Recopilación de datos

**Validación cruzada**

**Verificar la calidad del modelo con el dataset de prueba a veces no es suficiente. ¿Cómo podemos verificar los resultados para asegurarnos de que nuestro modelo funciona bien?**

Ya aprendiste a dividir un conjunto de datos en subconjuntos de entrenamiento, prueba y validación. Cada subconjunto representa la distribución del conjunto de datos original, pero con algunas desviaciones. Estas desviaciones son causadas por la aleatoriedad que ocurre al formar un subconjunto a través de una técnica de muestreo aleatorio utilizada en la función train\_test\_split.

La pregunta es: ¿cómo podemos asegurarnos de que la distribución de cada subconjunto sea relativamente cercana al conjunto de datos original? Es de vital importancia entrenar y evaluar el modelo en los subconjuntos que representan los datos originales lo más fielmente posible. ¡La solución es tomar varias muestras al azar!

La validación cruzada es una técnica que ayudará a entrenar y probar un modelo utilizando varias muestras formadas aleatoriamente. Cuando hablamos de validación cruzada, por lo general, nos referimos específicamente a validación cruzada k-fold.

¿Como funciona? Averigüémoslo en un ejemplo.

Supongamos que tenemos un conjunto de datos con 10 mil muestras. Para realizar una validación cruzada k-fold, primero tenemos que decidir un valor k. Este valor indica el número de subconjuntos iguales en los que dividiremos nuestro conjunto de datos. Por ahora, establezcamos el valor de k en 5.

Una vez que se establece k, dividimos el conjunto de datos (las diez mil muestras) en k partes iguales. Terminamos con cinco subconjuntos de 2000 muestras cada uno.

A continuación, iteramos sobre estos 5 subconjuntos y les asignamos diferentes propósitos. Por ejemplo, durante la primera iteración, el primer subconjunto se usará para la validación y los cuatro restantes (del 2 al 5) son para entrenamiento.

En la segunda iteración, el segundo subconjunto ahora se usa para la validación, mientras que el primero, el tercero, el cuarto y el quinto ahora se asignan con fines de entrenamiento.

Podemos seguir con la misma lógica. El número total de iteraciones es igual a k. Podemos visualizar este proceso con el siguiente diagrama:

Conjunto de entrenamientoConjunto de validacioˊn𝑘-veces1Subconjunto 1Subconjunto 2Subconjunto 3Subconjunto 4Subconjunto 52Subconjunto 1Subconjunto 2Subconjunto 3Subconjunto 4Subconjunto 53Subconjunto 1Subconjunto 2Subconjunto 3Subconjunto 4Subconjunto 54Subconjunto 1Subconjunto 2Subconjunto 3Subconjunto 4Subconjunto 55Subconjunto 1Subconjunto 2Subconjunto 3Subconjunto 4Subconjunto 5Conjunto de entrenamiento​Conjunto de validacioˊn​*k*-veces12345​Subconjunto 1​Subconjunto 1​Subconjunto 1​Subconjunto 1​Subconjunto 1​​Subconjunto 2​Subconjunto 2​Subconjunto 2​Subconjunto 2​Subconjunto 2​​Subconjunto 3​Subconjunto 3​Subconjunto 3​Subconjunto 3​Subconjunto 3​​Subconjunto 4​Subconjunto 4​Subconjunto 4​Subconjunto 4​Subconjunto 4​​Subconjunto 5​Subconjunto 5​Subconjunto 5​Subconjunto 5​Subconjunto 5​​

En el diagrama puedes ver que estamos "cruzando" los datos, tomando cada vez un nuevo subconjunto para la validación. En cada paso de la iteración, el modelo se entrena con los datos de entrenamiento y se evalúa con los datos de validación.

Cuando se completan todas las iteraciones de k, se realiza la evaluación final del modelo tomando la media de todas las puntuaciones de evaluación de k.

El método de validación cruzada se asemeja a un arranque en el que se forman varias muestras, pero la diferencia es que la validación cruzada usa subconjuntos con contenido fijo que no cambia en cada etapa de entrenamiento y validación. Cada observación pasa por el conjunto de entrenamiento y el conjunto de validación.

Pregunta

Elige los enunciados correctos:

Elige tantas como quieras

Antes del entrenamiento, los datos se dividen en muestras una vez. Durante la validación cruzada, los datos se dividen varias veces.

La validación cruzada implica k número de ciclos de entrenamiento y validación. Para cada ciclo se intercambian las muestras de entrenamiento y validación.

Después de la validación cruzada, el modelo se evalúa en función de la peor puntuación de evaluación entre todas las etapas.

Durante la validación cruzada, los datos se dividen en muestras una vez.

En cada etapa de la validación cruzada, un nuevo bloque se convierte en la muestra de validación.

¡Correcto!

¡Buen trabajo!

La validación cruzada es útil cuando necesitamos comparar modelos, seleccionar hiperparámetros o evaluar la utilidad de las funciones. Minimiza la aleatoriedad de la división de datos y proporciona un resultado más preciso. Compara: con una sola división, obtenemos solo una muestra de validación (25% de todos los datos) y una muestra de entrenamiento (75%). Para la validación cruzada, el conjunto de validación es del 20%, ¡pero tenemos cinco de ellos! Entonces, con cada división, observamos un nuevo 20% y validamos el modelo para el 100% de los datos en total.

El único inconveniente de la validación cruzada es el tiempo de cálculo, especialmente con muchas observaciones o un valor alto de k. Es mucho tiempo.

Pregunta

Selecciona los enunciados correctos:

Elige tantas como quieras

La validación cruzada se utiliza para interpretar los parámetros del modelo.

Ayuda a elegir el mejor modelo y evaluarlo.

Las pruebas en diferentes muestras y el promedio de la puntuación de evaluación brindan resultados confiables.

Cuanto mayor sea el conjunto de datos, menos tiempo llevará la validación cruzada.

La validación cruzada se utiliza para la selección de hiperparámetros.

Ayuda a comparar modelos con diferentes hiperparámetros y elegir el mejor.

¡Lo has entendido bien!

Validación cruzada en Python

Completa el código del bucle para obtener una validación cruzada en tres bloques del mismo tamaño. En cada etapa del ciclo, tienes el número del primer elemento de la muestra de validación y el tamaño del bloque (sample\_size).

1. Crea las matrices: *valid\_indexes* y *train\_indexes*. Estas contienen números de observaciones para muestras de validación y entrenamiento. Cambia los números en cada etapa del ciclo.
2. Divide las variables *features* y *target* en las muestras *features\_train, target\_train, features\_valid* y *target\_valid* para que contengan solo observaciones con los números necesarios.
3. Evalúa la calidad del modelo entrenado en cada muestra.

Calcula la calidad promedio del modelo y guárdala en la variable *final\_score*. Imprime el valor de la pantalla (en precódigo).

import pandas as pd

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score

data = pd.read\_csv('/datasets/heart.csv')

features = data.drop(['target'], axis=1)

target = data['target']

scores = []

# establece el tamaño del bloque si solo hay tres de ellos

sample\_size = int(len(data)/3)

for i in range(0, len(data), sample\_size):

# Crear índices para el bloque de validación

valid\_indexes = list(range(i, min(i + sample\_size, len(data))))

# Crear índices para el conjunto de entrenamiento

train\_indexes = list(range(0, i)) + list(range(min(i + sample\_size, len(data)), len(data)))

# Dividir las características y las etiquetas en muestras de entrenamiento y validación

features\_train = features.iloc[train\_indexes]

target\_train = target.iloc[train\_indexes]

features\_valid = features.iloc[valid\_indexes]

target\_valid = target.iloc[valid\_indexes]

# Entrenar el modelo

model = DecisionTreeClassifier(random\_state=0)

model.fit(features\_train, target\_train)

# Evaluar el modelo

predictions = model.predict(features\_valid)

score = accuracy\_score(target\_valid, predictions)

# Almacenar la puntuación

scores.append(score)

final\_score = sum(scores) / len(scores)

# < calcula la calidad media del modelo >

print('Valor de calidad promedio del modelo:', final\_score)

Valor de calidad promedio del modelo: 0.7689768976897691

Hemos validado de forma cruzada todos los datos, pero ¿se puede hacer en menos tiempo?

**Validación cruzada en Sklearn**

**La validación cruzada puede llevar menos tiempo si usamos las herramientas de *sklearn*.**

Para evaluar el modelo por validación cruzada usaremos la función cross\_val\_score del módulo *sklearn.model\_selection*.

Así se llama a la función:

from sklearn.model\_selection import cross\_val\_score

cross\_val\_score(model, features, target, cv=3)

La función toma varios argumentos, como:

* *model*: modelo para validación cruzada. Está entrenado en el proceso de validación cruzada, por lo que tenemos que pasarlo sin entrenar. Supongamos que necesitamos este modelo para un árbol de decisión:

from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier()

* *features*
* *target*
* *cv* — número de bloques para validación cruzada (son 3, por defecto)

La función no requiere dividir los datos en bloques o muestras para la validación y el entrenamiento. Todos estos pasos se realizan de forma automática. La función devuelve una lista de valores de evaluación del modelo de cada validación. Cada valor es igual a *model.score()* para la muestra de validación. Por ejemplo, para una tarea de clasificación, esto es *exactitud*.

Ejercicio

Calcula el puntaje de evaluación promedio usando el método de validación cruzada y guárdalo en la variable *final\_score*. Llama a la función cross\_val\_score.

Muestra en pantalla los valores *final\_score* (en precódigo).

import pandas as pd

from sklearn.tree import DecisionTreeClassifier

from sklearn.model\_selection import cross\_val\_score

data = pd.read\_csv('/datasets/heart.csv')

features = data.drop(['target'], axis=1)

target = data['target']

model = DecisionTreeClassifier(random\_state=0)

cross\_value\_score = cross\_val\_score(model, features, target, cv=5)# < calcula las puntuaciones llamando a la función con cinco bloques >

final\_score = cross\_value\_score.mean()

print('Puntuación media de la evaluación del modelo:', final\_score)

Resultado

Puntuación media de la evaluación del modelo: 0.7856284153005464

¡Es correcto!

Eso es lo que sucede cuando agregas algo de velocidad a la cruza (validación)

Sprint 10

Capítulo 4/8 · Última lección

Recopilación de datos

# Conclusión

En este capítulo aprendiste a:

* controlar la calidad del etiquetado utilizando el voto de mayoría;
* identificar y eliminar la fuga de información;
* realizar una validación cruzada utilizando las herramientas *sklearn*.

A continuación, tenemos lecciones sobre habilidades blandas y el proyecto del curso.

### Llévate esto contigo

Descárgate el resumen del capítulo y la hoja informativa para poder consultarlos cuando los necesites.

* [Resumen del capítulo: Recopilación de datos](https://practicum-content.s3.us-west-1.amazonaws.com/new-markets/DS_sprint_9/ES/DS_ES_9_Sprint_Resumen_del_captulo_Recopilacin_de_datos.pdf?etag=30f2253e5aa92bfbace60f599b276f68)
* [Hoja informativa: Recopilación de datos](https://practicum-content.s3.us-west-1.amazonaws.com/new-markets/DS_sprint_9/ES/DS_ES_9_Sprint_Hoja_informativa_Recopilacin_de_datos.pdf?etag=aa0867794095a79051e118197a8d2363)

Sprint 10

Capítulo 5/8

Habilidades socioemocionales

# Lo que necesitas aprender sobre la empresa durante tu primera semana

Ya te diste cuenta de lo importante que es el contexto de los datos para un data scientist. El contexto global de los datos es la actividad empresarial y sus objetivos. Necesitas comprender qué lugar tendrán tus objetivos personales en la empresa. Luego puedes prepararte para una entrevista y trabajar de manera efectiva en el nuevo equipo.

Pregunta

Jack está siendo entrevistado para un puesto de data scientist. La persona a cargo del equipo hace la pregunta: "¿Qué métricas importan cuando necesitas medir la participación del usuario en la aplicación?"

Jack responde de inmediato: "Echaré un vistazo al tiempo promedio de sesión. Cuanto más tiempo pasen los usuarios en la aplicación, mejor”.

¿Estás de acuerdo con Jack?

Sí

No

No hay suficiente información para responder

¡Tienes razón! No tenemos suficientes datos. Veamos.

¡Yo soy Jack! ¿Has intervenido mi teléfono?

¡Perfecto!

Jack fue entrevistado por una empresa que desarrolla una aplicación de seguimiento llamada GoThin. La mayoría de las veces, la aplicación se ejecuta en segundo plano y determina la frecuencia cardíaca, cuenta los pasos y monitorea otros parámetros. La mayoría de los usuarios abren la aplicación por la mañana y por la noche para comprobar cómo les va. Las sesiones para cada usuario son realmente cortas y, a menudo, las notificaciones automáticas son suficientes.

Jack no se tomó su tiempo y no se fijó en las condiciones de la tarea. A la persona encargada del equipo probablemente le pareció incompleta su respuesta.

Trata de dar la mejor impresión. Este formulario te ayudará a recopilar información útil sobre la empresa.

1. Qué debes investigar antes de la entrevista:
   * Campo de actividad
   * La actividad principal de la empresa que aporta dinero.
   * ¿La empresa trabaja con particulares u otras empresas?
2. Conoce los detalles de la cultura empresarial: qué valoran los empleados, cómo está organizado el equipo y cuál es el horario de trabajo.
3. Esto es lo que debes averiguar durante las primeras semanas:
   * Los objetivos clave de la empresa y su departamento para el año en curso y los periodos de informe
   * El lugar que ocupa tu departamento en la estructura de la empresa
   * ¿Cuáles son los KPI para tu departamento y para ti personalmente?
   * ¿A qué proceso del negocio perteneces?
   * ¿Con quién estás trabajando? ¿Quiénes son tus colegas, clientes internos y contratistas?

Los nuevos empleados suelen obtener esta información de forma caótica. Llena este formulario durante las primeras dos semanas y podrás resolver problemas de forma sistemática.

### Ejercicio

Inténtalo: elige la empresa para la que quieres trabajar. Realiza consultas basadas en el primer elemento de este formulario y comparte los resultados en el chat #*teamwork*.

Sprint 10

Capítulo 6/8

Proyecto del curso

¡Felicidades! Has completado otro curso de la plataforma de entrenamiento. Ahora es el momento perfecto para probar tus habilidades y resolver un nuevo problema de machine learning. Realizarás este proyecto individualmente.

Cuando termines, envía tu trabajo al revisor del proyecto. Recibirás feedback dentro de las siguientes 48 horas. Después de eso, harás los cambios necesarios en tu trabajo y lo enviarás para una segunda revisión.

Por lo general, este proceso se repetirá varias veces hasta que recibas el visto bueno de la revisión y se aprueben todas las correcciones. Todo eso es parte del trabajo.

Tu proyecto se considerará completado una vez que el revisor del proyecto lo apruebe.

### Descripción del proyecto

Trabajas en la compañía de extracción de petróleo OilyGiant. Tu tarea es encontrar los mejores lugares donde abrir 200 pozos nuevos de petróleo.

Para completar esta tarea, tendrás que realizar los siguientes pasos:

* Leer los archivos con los parámetros recogidos de pozos petrolíferos en la región seleccionada: calidad de crudo y volumen de reservas.
* Crear un modelo para predecir el volumen de reservas en pozos nuevos.
* Elegir los pozos petrolíferos que tienen los valores estimados más altos.
* Elegir la región con el beneficio total más alto para los pozos petrolíferos seleccionados.

Tienes datos sobre muestras de crudo de tres regiones. Ya se conocen los parámetros de cada pozo petrolero de la región. Crea un modelo que ayude a elegir la región con el mayor margen de beneficio. Analiza los beneficios y riesgos potenciales utilizando la técnica *bootstrapping*.

### Condiciones:

* Solo se debe usar la regresión lineal para el entrenamiento del modelo.
* Al explorar la región, se lleva a cabo un estudio de 500 puntos con la selección de los mejores 200 puntos para el cálculo del beneficio.
* El presupuesto para el desarrollo de 200 pozos petroleros es de 100 millones de dólares.
* Un barril de materias primas genera 4.5 USD de ingresos. El ingreso de una unidad de producto es de 4500 dólares (el volumen de reservas está expresado en miles de barriles).
* Después de la evaluación de riesgo, mantén solo las regiones con riesgo de pérdidas inferior al 2.5%. De las que se ajustan a los criterios, se debe seleccionar la región con el beneficio promedio más alto.

Los datos son sintéticos: los detalles del contrato y las características del pozo no se publican.

### Descripción de datos

Los datos de exploración geológica de las tres regiones se almacenan en archivos:

* geo\_data\_0.csv. [Descarga el conjunto de datos](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/geo_data_0.csv)
* geo\_data\_1.csv. [Descarga el conjunto de datos](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/geo_data_1.csv)
* geo\_data\_2.csv. [Descarga el conjunto de datos](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/geo_data_2.csv)
* *id* — identificador único de pozo de petróleo
* *f0, f1, f2* — tres características de los puntos (su significado específico no es importante, pero las características en sí son significativas)
* *product* — volumen de reservas en el pozo de petróleo (miles de barriles).

## Instrucciones del proyecto

1. Descarga y prepara los datos. Explica el procedimiento.
2. Entrena y prueba el modelo para cada región en geo\_data\_0.csv:
   1. Divide los datos en un conjunto de entrenamiento y un conjunto de validación en una proporción de 75:25
   2. Entrena el modelo y haz predicciones para el conjunto de validación.
   3. Guarda las predicciones y las respuestas correctas para el conjunto de validación.
   4. Muestra el volumen medio de reservas predicho y *RMSE* del modelo.
   5. Analiza los resultados.
   6. Coloca todos los pasos previos en funciones, realiza y ejecuta los pasos 2.1-2.5 para los archivos 'geo\_data\_1.csv' y 'geo\_data\_2.csv'.
3. Prepárate para el cálculo de ganancias:
   1. Almacena todos los valores necesarios para los cálculos en variables separadas.
   2. Dada la inversión de 100 millones por 200 pozos petrolíferos, de media un pozo petrolífero debe producir al menos un valor de 500,000 dólares en unidades para evitar pérdidas (esto es equivalente a 111.1 unidades). Compara esta cantidad con la cantidad media de reservas en cada región.
   3. Presenta conclusiones sobre cómo preparar el paso para calcular el beneficio.
4. Escribe una función para calcular la ganancia de un conjunto de pozos de petróleo seleccionados y modela las predicciones:
   1. Elige los 200 pozos con los valores de predicción más altos de cada una de las 3 regiones (es decir, archivos 'csv').
   2. Resume el volumen objetivo de reservas según dichas predicciones. Almacena las predicciones para los 200 pozos para cada una de las 3 regiones.
   3. Calcula la ganancia potencial de los 200 pozos principales por región. Presenta tus conclusiones: propón una región para el desarrollo de pozos petrolíferos y justifica tu elección.
5. Calcula riesgos y ganancias para cada región:
   1. Utilizando las predicciones que almacenaste en el paso 4.2, emplea la técnica del bootstrapping con 1000 muestras para hallar la distribución de los beneficios.
   2. Encuentra el beneficio promedio, el intervalo de confianza del 95% y el riesgo de pérdidas. La pérdida es una ganancia negativa, calcúlala como una probabilidad y luego exprésala como un porcentaje.
   3. Presenta tus conclusiones: propón una región para el desarrollo de pozos petrolíferos y justifica tu elección. ¿Coincide tu elección con la elección anterior en el punto 4.3?

# Evaluación del proyecto

Hemos definido los criterios de evaluación para el proyecto. Lee esto con atención antes de pasar al ejercicio.

Esto es lo que los revisores buscarán cuando evalúen tu proyecto:

* ¿Cómo preparaste los datos para el entrenamiento?
* ¿Seguiste todos los pasos de las instrucciones?
* ¿Consideraste todas las condiciones del negocio?
* ¿Cuáles son tus hallazgos sobre el estudio de tareas?
* ¿Aplicaste correctamente la técnica *bootstrapping*?
* ¿Sugeriste la mejor región para el desarrollo de pozos? ¿Justificaste tu elección?
* ¿Evitaste la duplicación de código?
* ¿Mantuviste la estructura del proyecto y el código limpio?

Ya tienes tus hojas informativas y los resúmenes de los capítulos, por lo que todo está listo para continuar con el proyecto.

¡Buena suerte!

Capítulo 7/8 · Faltan 2 lecciones

Conclusión

**Conclusión**

¡Felicidades! Ha completado el curso "Machine learning en los negocios".

Has aprendido a:

* Distinguir entre métricas fuera de línea y métricas en línea;
* Usar bootstrap para calcular intervalos de confianza y métricas comerciales;
* Realizar validación cruzada para que las conclusiones sobre la evaluación del modelo sean más fiables.

A continuación, se presenta el proyecto integrado donde puedes poner a prueba los conocimientos que adquiriste. ¡Y luego las vacaciones!

¿Qué te pareció la lección?

Capítulo 8/8

Mensaje CPC 60%

**Mensaje CPC 60%**

🌟 ¡Excelente trabajo alcanzando el 60% en tu trayecto académico! 🌟

En el Centro de Carreras, estamos emocionados por tu dedicación y logros, ¡y estamos aquí para impulsarte hacia nuevas alturas de éxito profesional! 🚀

Y ahora, te presentamos nuestro Curso de Preparación para la Carrera, ¡una experiencia que transformará tu trayectoria y abrirá puertas a emocionantes oportunidades! 💫

Durante las próximas 2 semanas, te sumergirás en un viaje de desarrollo personal y profesional, donde moldearás tu propio *Kit de Empleabilidad* con detalles que te harán destacar :

📝 CV y Carta de Presentación: Aprende a mostrar tus logros y habilidades de manera impactante, dejando una impresión inolvidable en los reclutadores.

💻 LinkedIn: Descubre los secretos para crear un perfil atractivo y profesional que te haga destacar en el mercado laboral.

🐈‍⬛ GitHub (Portafolio): Construye un portafolio que no solo destaque tu experiencia, sino que también refleje tu pasión y dedicación.

¡Pero eso no es todo! ¡Nuestro equipo de expertos del Centro de Carreras estará contigo en cada paso del camino! Recibirás retroalimentación personalizada para perfeccionar tus documentos y perfiles, guiado por la sabiduría de su experiencia. 🧙🏼

Recuerda, ¡cada esfuerzo que haces hoy es una inversión en el futuro que sueñas! ¡Estás al borde de desbloquear tu potencial y llevar tu carrera a nuevos horizontes emocionantes! 🚀 ¡Adelante, valiente guerrero, el éxito te espera con los brazos abiertos! 🌟

Si no es de tu interés, agrega aquí la razón y cómo podemos apoyarte

Ingres [aquí](https://docs.google.com/forms/d/e/1FAIpQLSeuCmbLMY_mYa9-9QOiMCs0-qrPwrmy0VapbQFzRP2BO4CqUQ/viewform) al formato de registro

⌛ Recuerda que una vez que te gradúas tienes solo una ventana de tiempo de no más de 2 semanas para hacer el Curso de Preparación de Carrera.

Si tienes preguntas o necesitas ayuda, no dudes en contactarnos. Estamos aquí para apoyarte en cada paso del camino 🙂

Gabriela Rojas

Career Success Manager LatAm

gabriela-r@tripleten.com