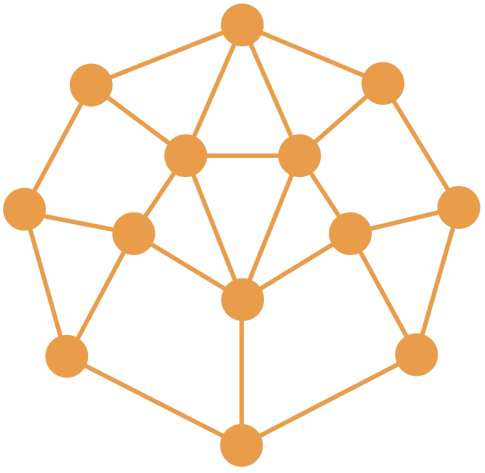
Guía Reto Diabetes Saturdays.AI

(01/02/2020)



Saturdays.AI

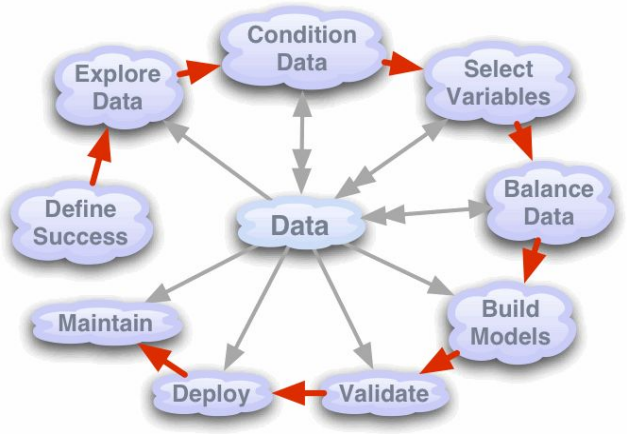
Índice

Fases de creación de un modelo

¿Qué es un modelo predictivo?

- Un modelo predictivo es simplemente una herramienta para predecir hechos/características futuras usando datos históricos
- Ejemplos del día a día:
    - Decidir la ruta en función del tráfico
    - Siguiente palabra a teclear en el teléfono móvil
- Construir un modelo sigue un proceso muy definido para tener un modelo predictivo.

Ciclo de vida de modelos predictivos



1. Definir el problema

Definir un problema es la primera parte de la construcción de modelos. Es necesario responder al "por qué". Por ejemplo: asumimos que podrías optar por la "Ciencia de los

Datos" como opción de carrera. ¿Pero por qué? Si sabes la respuesta, genial. Pero, si todavía estás pensando en una respuesta, puede que sea el momento de considerar el "por qué" estás tomando este taller.

En la construcción de modelos, la definición del problema nos ayuda a saber qué estamos tratando de resolver. Idealmente, esta etapa requiere una exhaustiva sesión de lluvia de ideas de preguntas y respuestas con los clientes de negocios para saber qué es exactamente lo que él/ella espera del equipo de análisis.

Cuando se trata de construir una definición del problema, se espera que se ponga un marco al problema y que luego se encuentren las posibles soluciones.

Por ejemplo, pongamos que la tarea en cuestión es aumentar los beneficios de las tarjetas de crédito.

Entonces, mientras calculamos la rentabilidad de las tarjetas de crédito, podemos decir que puede ser incrementada ya sea aumentando los ingresos o disminuyendo los costes.

Los ingresos pueden aumentarse cambiando varias cosas:

- Aumentar / disminuir la tasa de interés
- Cambiar los límites de crédito de los clientes
- Aumentar la tarifa de las tarjetas de crédito

De manera similar, los costes pueden reducirse realizando cualquiera de los siguientes cambios:

- Renegociar los términos del contrato para el tipo de interés que el banco paga a sus acreedores
- Reducir el costo de las operaciones reduciendo o automatizando el soporte al cliente
- ¡Cerrar las cuentas de personas que nunca pagan intereses!

Aunque no hay una base de respuesta correcta o incorrecta en la información que tenemos, probablemente habría elegido cambiar el tipo de interés para los clientes. Esto se debe a que esto podría tener el impacto más directo en la rentabilidad de los clientes que están llevando el crédito mes a mes.

En la presentación del reto deberéis explicar cómo habéis definido el problema.

2. Hipótesis

Lo más importante es hacer una lista/brainstorming de las variables que consideráis que son más importantes.

Recuerda que la calidad de tu modelo va a depender directamente de la calidad de tu hipótesis.

Por ejemplo, ¿cuáles son las variables que más pueden afectar cuando intentamos predecir si una tarjeta de crédito entra dentro del grupo de morosos?

- Salario: un salario alto es un predictor de estabilidad financiera
- Tipo de trabajo: una persona con un trabajo más estable tiene mayor estabilidad financiera
- Historial
- Nivel socioeducativo
- Y muchas más…

¿Debemos pensar en nuestra hipótesis antes o después de explorar lo datos? ¡Siempre antes! Esto puede permitirte pensar sin un sesgo previo y explorar fuentes de datos adicionales.

En la presentación deberéis explicar cuales son vuestras hipótesis

3. Extracción y análisis de datos

No es el caso de este reto. Sin embargo, es muy útil recoger datos de diferentes fuentes y combinarlos antes de la exploración de datos:

- Identificación de variables. Cuales son nuestro target y cuáles no. Tipo de datos y categoría.
- Análisis univariable
- Análsisi multivariable
- Datos perdidos.
- Eliminar outliers
- Transformación de variables

Tan importante como entender el problema es entender los datos que tenemos disponibles. Es común hacer un análisis exploratorio de datos para familiarizarnos con ellos.

En el análisis exploratorio se suelen hacer gráficos, correlaciones y estadísticas descriptivas para comprender mejor qué historia nos están contando los datos. Además ayuda a estimar si los datos que tenemos son suficientes, y relevantes, para construir un modelo.

4. Desarrollo del modelo predictivo

- Cómo explorar el dataset
- Preparar el dataset
- Elegir un modelo
- Evaluación del modelo
- Interpretar el modelo e insights.

👉 LINK

Criterios de evaluación del reto

El jurado evaluará las siguientes características* de la solución presentada por el grupo:

- Visualización de los datos.
- Imputación de datos, transformación de variables y exploración de datos.
- Desarrollo y evaluación del modelo.
- Hipótesis y soluciones más allá de las variables del dataset.
- Score del modelo.

Tanto en la presentación como en el código, cada parte de la solución debe estar debidamente justificada y comentada. El jurado ***no*** *realizará ninguna pregunta al equipo.*

¿En qué consiste el reto?

El objetivo del reto de datos es predecir de forma diagnóstica si un paciente tiene o no diabetes, basándose en ciertas medidas diagnósticas incluidas en el conjunto de datos.

El conjuntos de datos consiste en varias variables médicas de predicción y una variable objetivo. Las variables predictoras incluyen el número de embarazos que los pacientes paciente (todas mujeres) han tenido, su IMC, nivel de insulina, edad, etc.

¿Qué es la diabetes?

Según el NIH, "La diabetes es una enfermedad que ocurre cuando la glucosa en la sangre, también llamada azúcar en la sangre, es demasiado alta. La insulina, una hormona producida por el páncreas, ayuda a que la glucosa de los alimentos entre en las células para ser utilizada como energía. A veces el cuerpo no produce suficiente o nada de insulina o no utiliza bien la insulina. La glucosa permanece en la sangre y no llega a las células.

Con el tiempo, tener demasiada glucosa en la sangre puede causar problemas de salud.

¿Cuáles son los diferentes tipos de diabetes? Los tipos más comunes de diabetes son el tipo 1, el tipo 2 y la diabetes gestacional.

Diabetes de tipo 1: Si tienes diabetes de tipo 1, tu cuerpo no produce insulina. El sistema inmunológico ataca y destruye las células del páncreas que producen insulina. La diabetes de tipo 1 se suele diagnosticar en niños y adultos jóvenes, aunque puede aparecer a cualquier edad. Las personas con diabetes de tipo 1 necesitan tomar insulina diariamente.

Diabetes de tipo 2: Eu cuerpo no produce ni utiliza bien la insulina. Puedes desarrollar diabetes de tipo 2 a cualquier edad, incluso durante la infancia. Sin embargo, este tipo de diabetes se presenta con mayor frecuencia en personas de mediana edad y mayores. El tipo 2 es el tipo más común de diabetes.

Diabetes gestacional: La diabetes gestacional se desarrolla en algunas mujeres cuando están embarazadas. La mayoría de las veces, este tipo de diabetes desaparece después del nacimiento del bebé. Sin embargo, si ha tenido diabetes gestacional, se tiene una mayor probabilidad de desarrollar diabetes de tipo 2 más adelante. A veces, la diabetes diagnosticada durante el embarazo es en realidad diabetes de tipo 2.

Fases del reto

Parte 1 - Descripción de datos:

El conjunto de datos que obtengamos y queramos introducir en el modelo será nuestro dataset.

Dado el dataset, se espera que cada equipo lea el nombre de la variable y entienda la información que codifica. De esta información debe extraer los rangos en los que debe moverse cada variable y por lo tanto buscar si existen nulos.

Además de realizar esta tarea, se espera que obtengan una descripción de la distribución de datos de cada variable, una información básica de media, moda, correlaciones, etc.

Parte 2 - Alteraciones del conjunto de datos

Una vez entendido el conjunto de datos, el siguiente paso es editarlo y limpiarlo.

Lo haremos en un *notebook/cuaderno* (ese archivo .ypinb) de Jupyter. Un notebook es un documento de programación en la nube.

Ya que solo se dispone de una fuente de datos, estas variables (en caso de que se consideren relevantes) se obtendrán como combinación de las variables ya presentes en el dataset. De cara a la obtención de estas variables, se recomienda la representación de las mismas en distintos diagramas (barplot o scatterplot) para estudiar su distribución de manera visual y ver la posible existencia de patrones.

Parte 3 - Entrenamiento

Una vez editado el conjunto de datos a gusto y consideración del grupo, se procederá a buscar y entrenar un modelo que resuelva el ejercicio. Ya que cada modelo depende de uno o más hiperparámetros que han de ser configurados por cada grupo, se espera que se haga un estudio de los distintos valores que puede tomar y dar un motivo por el cual se está eligiendo un determinado valor. Una vez más, las representaciones gráficas resultan de gran ayuda en este apartado.

Finalmente, se deben analizar los resultados del modelo, en clasificación por lo general se utiliza una matriz de confusión, para analizar qué casos son más conflictivos para el modelo y analizar si es un escenario conveniente o si por el contrario se debe buscar otra métrica para la selección del modelo.

¿Cómo tratar los datos?

Los datos en el mundo real rara vez son limpios y homogéneos. Se debe a las siguientes razones:

- Tienden a ser incompletos.
- Encontramos ruido.
- Datos corruptos.
- Fallos en la carga de información.
- Extracción incompleta de los datos.

Por lo tanto, es una tarea importante de un científico de datos el tratar los datos llenando los valores faltantes para tener un modelo robusto. Es importante manejar la falta de datos

ya que podrían llevar a una predicción o clasificación errónea para cualquier modelo que se utilice. Los datos del mundo real a menudo tienen valores perdidos.

- Si quieres leer más (paso a paso) cómo afrontar un reto de datos. [LINK]
- Tutorial de cómo lanzar el notebook en Sagemaker y entrenar un modelo [LINK]
- Otro step by step de XGBoost [LINK]

¿Cómo presento mi solución?

Ceñiros a los criterios de evaluación. Explicad de forma clara cómo habéis definido el problema y la hipótesis, y después centraros en detallar el tratamiento de los datos y el modelo. Tendréis muy pocos minutos. Parte del jurado estará revisando vuestra solución de código.

Wi-Fi

Nombre: IMMUNERS
Pass: Castellana89Imm

Ahora… ¡A trabajar! 💪🟧

Si quieres saber más:

www.saturdays.ai

madrid@saturdays.ai