



# Data Science

Hands on - M4L3 | Aprendizaje supervisado  
II: Regresión logística



→ [soyhenry.com](http://soyhenry.com)



# Objetivos

- Distinguir la regresión logística de la regresión lineal, evaluando su propósito y aplicación en modelos de clasificación.
- Analizar la función sigmoide y su relación con la probabilidad, interpretando la transformación de valores en términos de confianza en una clasificación binaria.
- Examinar el odds ratio y los coeficientes de la regresión logística, identificando su impacto en la toma de decisiones basada en modelos predictivos.





# Agend a

COMENCEMOS →

#TEMAS

- .01 Definición y diferencia con regresión lineal
- .02 Función sigmoide y probabilidad
- .03 Interpretabilidad de los resultados
- .04 Avance de PI





<->

# ¿Qué vimos en la lecture?





<01>

# Definición y diferencia con regresión lineal





# Introducción a la regresión logística

- La regresión logística se usa cuando queremos **clasificar** entre dos categorías (por ejemplo, alta o baja demanda).
- A diferencia de la regresión lineal, **no predice valores continuos**, sino **probabilidades** entre 0 y 1.
- En CityScoot, una regresión lineal podría dar valores imposibles como -0.3 o 1.4; la logística los traduce en probabilidades válidas: 12% o 95%.
- Cambia también la **función de costo**: usa **log-loss (entropía cruzada)**, que penaliza más fuerte las predicciones erróneas con alta confianza.
- La regresión logística **mantiene la estructura lineal**, pero transforma la salida con la **función sigmoide**, garantizando que el resultado sea interpretable como probabilidad.





# Supuestos básicos de la regresión logística

- **Independencia de observaciones:** cada caso debe aportar información nueva; si hay dependencia temporal, conviene usar modelos de series de tiempo.
- **Linealidad en el logit:** la relación lineal se da entre las variables predictoras y el logaritmo de los odds, no con la probabilidad directa.
- **Ausencia de multicolinealidad:** si dos variables aportan la misma información, los coeficientes se vuelven inestables; se pueden combinar o regularizar (Ridge/Lasso).
- **No separabilidad perfecta:** si las clases se separan de forma exacta, los coeficientes tienden a infinito. El modelo funciona mejor con solapamientos y probabilidades intermedias.





# Casos de uso típicos de la regresión logística

- ★ **Marketing:** predecir si un cliente hará clic en una campaña.
- ★ **Medicina:** estimar si un paciente padece una enfermedad.
- ★ **Finanzas:** anticipar la probabilidad de default.
- ★ **CityScoot:** decidir si un día es de alta demanda o baja, según clima, marketing y eventos





<02>

# Función sigmoide y probabilidad





# La función sigmoide: corazón de la regresión logística

La **sigmoide** convierte cualquier valor real en una probabilidad entre 0 y 1:

$$\sigma(z) = 1 / (1 + e^{-z})$$

Cuando  $z \rightarrow +\infty \rightarrow p \rightarrow 1$

Cuando  $z \rightarrow -\infty \rightarrow p \rightarrow 0$

Cuando  $z = 0 \rightarrow p = 0.5$  (incertidumbre)

Esta función permite **interpretar el resultado como probabilidad**, manteniendo el modelo matemático estable.





# Del log-odds a la probabilidad

El modelo aprende en la escala de los **log-odds**:

$$\text{logit}(p) = \ln(p / (1 - p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$



Luego, la sigmoide convierte esos log-odds en una probabilidad interpretativa.

Si  $\text{log-odds} = 1.1$ , los odds  $\approx 3$ , y  $p \approx 0.75 \rightarrow 75\%$  de probabilidad de alta demanda.

Esto permite entender la regresión logística como un **modelo lineal dentro de una escala logarítmica**, transformada después en probabilidad.





# Evaluar el rendimiento del modelo

La sigmoide también se usa en:

- **Redes neuronales:** como función de activación en salidas binarias.
- **Modelos probabilísticos:** vincula variables continuas y binarias.
- **Epidemiología o física:** describe curvas de crecimiento o saturación.



En CityScoot, puede aplicarse para estimar la **probabilidad de reserva de un usuario**:

- $z = -2.0 \rightarrow p = 0.12$
- $z = 3.1 \rightarrow p = 0.96$

**Su fuerza está en transformar valores infinitos en probabilidades interpretables, conectando modelos matemáticos con fenómenos reales.**





<03>

# Interpretabilidad de los resultados





# Introducción a la interpretabilidad

La **interpretabilidad** es entender **cómo y por qué** el modelo toma decisiones.

En CityScoot, no basta con saber que hay 90% de probabilidad de alta demanda; interesa saber **qué variables influyen**:

- Temperatura (+)
- Lluvia (-)
- Marketing (+)



Un modelo interpretable permite **accionar**: ajustar flota, planificar campañas, anticipar caídas.





# Lectura de coeficientes en regresión logística

- Cada coeficiente  $\beta$  indica cuánto cambian los **log-odds** ante una variación de la variable, manteniendo las demás constantes.
- $\beta > 0 \rightarrow$  aumenta la probabilidad del evento.
- $\beta < 0 \rightarrow$  la reduce.
- Ejemplo CityScoot:
  - $\text{temp\_c} = +0.15 \rightarrow$  cada grado aumenta los log-odds.
  - $\text{rain\_mm} = -0.07 \rightarrow$  cada mm de lluvia los reduce.





# Casos de uso típicos de la regresión logística

- ★ **Odds:**  $p / (1 - p)$  → mide la razón entre ocurrencia y no ocurrencia.
- ★ **Log-odds:**  $\ln(\text{odds})$  → escala continua donde aprende el modelo.
- ★ **Odds-ratio:**  $e^{\beta}$  → cuánto se multiplican los odds si la variable aumenta una unidad.





# Escalado y variables categóricas

- **Variables escaladas:**  $\beta$  refleja el efecto de una **desviación estándar**, no una unidad bruta.
- **Categóricas (dummies):** cada  $\beta$  se interpreta **respecto a la categoría base**.
- **Ejemplo CityScoot:**
  - marketing\_spend (escalada):  $\beta=0.6 \rightarrow$  un aumento de 1 $\sigma$  en marketing  $\uparrow$  odds 1.8 $\times$ .
  - is\_holiday=1  $\rightarrow$  OR=3.3  $\rightarrow$  feriados triplican odds de alta demanda.





# Interpretabilidad como valor de negocio

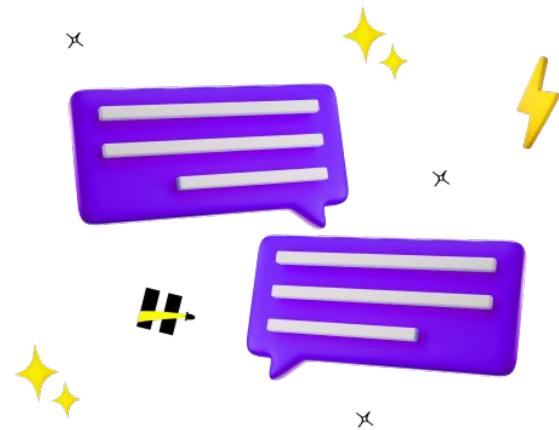
La regresión logística traduce datos en **decisiones accionables**.

Permite entender **qué variables son palancas clave** y en qué dirección influyen.

Genera **confianza**: los líderes pueden explicar y justificar decisiones basadas en el modelo.

CityScoot puede usarlo para:

- Prever picos de demanda.
- Reforzar flota en fines de semana.
- Planificar marketing cuando baja la probabilidad de viajes.





<DATA SCIENCE/>

# Vayamos a la práctica



Avance de PI



→ soyhenry.com



# Consigna

FinanceGuard es un banco digital que enfrenta un incremento del **20% anual en la tasa de abandono de clientes (churn)**. Como **Científico de Datos Junior**, tu objetivo en este primer avance es construir un **modelo baseline de predicción de churn** mediante **Regresión Logística**, comprendiendo cómo las variables demográficas y financieras influyen en la probabilidad de que un cliente abandone el banco. El análisis comienza con la carga y exploración del dataset de 50.000 clientes, la detección de valores faltantes, el tratamiento de variables categóricas y numéricas, y la división del conjunto de datos para entrenamiento y prueba.

El modelo se implementará con **scikit-learn**, interpretando la **función sigmoide**, los **coeficientes** y los **odds ratios**, para luego evaluar su desempeño mediante **matriz de confusión, precision, recall, F1-score y curva ROC-AUC**. Este trabajo se entregará en el notebook **1\_EDA\_RegresionLogistica.ipynb**, que servirá como base de referencia para comparar y mejorar modelos más complejos en los próximos avances.



# Tareas a realizar

- 1. Comprepción del problema y análisis exploratorio básico:**
  - Investigar qué es el churn bancario
  - Carga y exploración inicial del dataset (50,000 clientes)
  - Variables demográficas: edad, género, ubicación, antigüedad
  - Variables financieras: saldo promedio, productos contratados, transacciones
  - Variable objetivo: churn (1 = abandonó, 0 = activo)
  - Análisis de desbalanceo de clases (típicamente 80-20)
- 2. Preparación de datos para regresión logística:**
  - Tratamiento de valores faltantes básico
  - Encoding de variables categóricas (One-Hot, Label Encoding)
  - Escalamiento de variables numéricas (StandardScaler)
  - Split básico: train (80%), test (20%)
  - Identificación de multicolinealidad





# Tareas a realizar

## 3. Implementación de Regresión Logística:

- **Regresión Logística Simple:**
  - Implementación desde cero (opcional)
  - Uso de scikit-learn
  - Interpretación de la función sigmoide
- **Análisis de coeficientes:**
  - Interpretación de pesos/coeficientes
  - Odds ratios y su significado
  - Intervalos de confianza
- **Evaluación específica:**
  - Matriz de confusión
  - Curva ROC y AUC
  - Precision, Recall, F1-Score





## Notas extra

La regresión logística será tu modelo baseline. Enfócate en entender cómo funciona el algoritmo, el procesamiento de las variables, y la interpretación de los resultados. Este modelo debe ser tu referencia para comparar modelos más complejos.



**HENRY**



#OpenQuestion



¿Preguntas?



→ soyhenry.com

~~HENRY~~

**¡Muchas gracias!**



→ soyhenry.com