Módulo 3. Regresión Logística y Clasificación

1. Modelado de eventos binarios y fundamentos de la regresión logística

Introducción

En el análisis profesional de datos, existen problemas donde la pregunta no es «¿cuánto?» sino «¿ocurre o no ocurre?». ¿Un cliente incumplirá un préstamo? ¿Un paciente desarrollará una enfermedad? ¿Un estudiante abandonará sus estudios? Estas situaciones comparten una característica estructural: la variable de interés adopta solo dos posibles valores. Este tipo de fenómeno, denominado evento binario, exige herramientas distintas a las utilizadas en la regresión lineal.

La regresión logística surge precisamente para modelar la probabilidad de ocurrencia de este tipo de eventos. A diferencia de los modelos lineales tradicionales, su objetivo no es estimar un valor continuo, sino cuantificar la probabilidad de que un evento suceda en función de determinados factores explicativos. Según De la Fuente Fernández (s.f.), la regresión logística constituye una de las técnicas estadístico-inferenciales más empleadas cuando la variable dependiente es dicotómica,

especialmente en contextos médicos, financieros y sociales. Este dato resulta ilustrativo: gran parte de la investigación clínica contemporánea utiliza modelos logísticos para estimar riesgos relativos y odds ratios.

Un aspecto curioso es que muchos sistemas de decisión automatizada —desde la aprobación de créditos hasta los sistemas de detección de fraude— operan internamente con modelos de clasificación binaria. Sin embargo, una clasificación no es simplemente una etiqueta; es el resultado de estimar una probabilidad y luego aplicar un umbral de decisión. Aquí surge una pregunta relevante: ¿qué ocurre si se modifica ese umbral? ¿Se altera el equilibrio entre verdaderos positivos y falsos positivos? La respuesta a estas preguntas conecta directamente con el análisis de curvas ROC, ampliamente utilizado en evaluación de clasificadores (Fawcett, 2006).

En esta unidad abordaremos los fundamentos conceptuales de la regresión logística, la modelización de probabilidades asociadas a eventos binarios y la interpretación de sus parámetros. Se analizará cómo se construye el modelo, cómo se estiman sus coeficientes y cómo se interpretan en términos de riesgo relativo. Este recorrido permitirá comprender el pasaje desde la asociación entre variables hacia la clasificación probabilística,

ampliando el marco del modelado predictivo trabajado previamente en la materia.

Probabilidad de ocurrencia de eventos binarios

En numerosos problemas profesionales, el fenómeno de interés adopta únicamente dos posibles estados: presencia o ausencia de enfermedad, aprobación o rechazo de crédito, abandono o permanencia en un programa educativo. Este tipo de variable, denominada dicotómica o binaria, exige un tratamiento estadístico específico. La regresión logística se desarrolla precisamente para modelar la probabilidad asociada a la ocurrencia de estos eventos.

Desde una perspectiva conceptual, el objetivo central consiste en estimar la probabilidad de que un evento ocurra dadas determinadas características observables. De la Fuente Fernández (s.f.) explica que el modelo logístico pertenece al conjunto de métodos diseñados para analizar variables de respuesta cualitativas, y que su utilidad resulta particularmente evidente cuando solo existen dos categorías posibles. Esta estructura binaria permite traducir la variable dependiente en valores convencionales 0 y 1, donde 1 representa la ocurrencia del evento y 0 su no ocurrencia.

Ahora bien, estimar una probabilidad implica trabajar en un rango acotado entre 0 y 1. En este punto aparece una diferencia sustantiva respecto de la regresión lineal: una función lineal puede producir valores inferiores a cero o superiores a uno, lo que resulta incompatible con la interpretación probabilística. Por ello, la regresión logística introduce una función de enlace que transforma una combinación lineal de predictores en una probabilidad válida. Según De la Fuente Fernández (s.f.), esta transformación se apoya en la denominada función logística, que garantiza que los valores estimados permanezcan dentro del intervalo probabilístico.

Un concepto central en este contexto es el de *odds*, entendido como el cociente entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra. En ámbitos médicos y financieros, este indicador permite interpretar cambios relativos en términos de riesgo. Cuando se aplica el logaritmo natural al *odds*, se obtiene el denominado *logit*, que establece una relación lineal entre los predictores y la transformación del riesgo. Esta linealización facilita la estimación de parámetros mediante métodos de máxima verosimilitud, tal como describe De la Fuente Fernández (s.f.).

Desde el punto de vista operativo, el modelo permite responder preguntas del tipo:

- ¿Cómo varía la probabilidad de incumplimiento crediticio cuando aumenta el nivel de endeudamiento.

- ¿Qué sucede con el riesgo de una enfermedad cuando se incrementa la edad manteniendo constantes otros factores?

Estas preguntas reflejan el carácter condicional del modelo: cada probabilidad estimada depende de un conjunto específico de valores de las variables explicativas.

La regresión logística no predice valores continuos, sino probabilidades condicionadas que luego pueden transformarse en decisiones de clasificación.

En el marco de la evaluación de clasificadores, Fawcett (2006) señala que la predicción probabilística permite establecer distintos umbrales de decisión, generando diferentes equilibrios entre verdaderos positivos y falsos positivos. Esta característica

resulta especialmente relevante en contextos donde los costos asociados a errores de clasificación no son simétricos.

Elementos estructurales del modelo logístico binario

Tabla 1. Componentes conceptuales del modelo de regresión logística binaria

Elemento	Descripción técnica	Aplicación profesional
Variable dependiente	Evento binario codificado como 0 y 1	Morosidad: Sí / No
Variables independientes	Factores explicativos cuantitativos o cualitativos	Ingreso, edad, historial crediticio
Probabilidad estimada	Valor entre 0 y 1 asociado a la	Probabilidad de incumplimiento

	ocurrencia del evento	
Odds	Cociente entre probabilidad de ocurrencia y no ocurrencia	Relación de riesgo
Logit	Logaritmo natural del *odds*	Transformación lineal del riesgo

Fuente: elaboración propia con base en De la Fuente Fernández (s.f.)

Un aspecto metodológico relevante consiste en la estimación de los parámetros. A diferencia de la regresión lineal, donde se emplea el método de mínimos cuadrados, la regresión logística utiliza el principio de máxima verosimilitud. Este procedimiento busca identificar el conjunto de coeficientes que maximiza la probabilidad de observar los datos muestrales. De la Fuente Fernández (s.f.) explica que el ajuste del modelo puede evaluarse mediante contrastes de bondad global y pruebas basadas en estadísticos de desviación.

En la práctica profesional, la interpretación de los coeficientes se realiza a través del *odds ratio*, que expresa el cambio multiplicativo en el riesgo asociado a una variación unitaria del

predictor. Si el coeficiente es cercano a cero, el *odds ratio* se aproxima a uno, lo que indica ausencia de efecto relevante sobre la probabilidad del evento.

Desde el punto de vista analítico, el modelo logístico permite:

- Estimar probabilidades individuales condicionadas a múltiples factores.
- Analizar la contribución parcial de cada variable manteniendo constantes las demás.

Tabla 2. Interpretación del *odds ratio* en regresión logística

Valor del *odds ratio*	Interpretación estadística	Lectura aplicada
Igual a 1	No hay cambio en el riesgo	Variable sin efecto apreciable
Mayor que 1	Incremento en el riesgo	Aumenta la probabilidad del evento

Menor que 1	Disminución en el riesgo	Reduce la probabilidad del evento

Fuente: elaboración propia con base en De la Fuente Fernández (s.f.)

La modelización de eventos binarios constituye, por tanto, un paso fundamental en el análisis predictivo aplicado. Su fortaleza radica en integrar una formulación probabilística con una estructura interpretativa que permite traducir resultados estadísticos en términos de riesgo relativo y decisiones operativas. Además, la posibilidad de evaluar clasificadores mediante curvas *ROC* amplía el marco de análisis hacia la comparación de modelos en función de su desempeño discriminativo, aspecto que será desarrollado en los próximos apartados.

Función logística e interpretación de coeficientes

En la regresión logística, la estimación de probabilidades se apoya en una transformación específica que permite vincular una combinación lineal de variables explicativas con una probabilidad comprendida entre 0 y 1. Esta transformación es la función logística. Su incorporación al modelo resuelve el problema que surgiría si se utilizara una función lineal directa para estimar

probabilidades, ya que una combinación lineal puede tomar cualquier valor real.

De la Fuente Fernández (s.f.) explica que el modelo logístico establece una relación entre los predictores y el logaritmo del odds del evento. Esta relación lineal sobre el logit implica que cada coeficiente representa la variación en el logaritmo del cociente de probabilidades ante un incremento unitario en la variable correspondiente, manteniendo constantes las demás. Esta condición de análisis parcial resulta central para la interpretación en contextos multivariados.

Desde una perspectiva aplicada, la función logística puede entenderse como un mecanismo que suaviza la transición entre probabilidades bajas y altas. Cuando la combinación lineal de predictores toma valores muy negativos, la probabilidad estimada se aproxima a cero; cuando toma valores muy positivos, la probabilidad se aproxima a uno. En el rango intermedio, pequeñas variaciones en los predictores pueden generar cambios apreciables en la probabilidad.

La función logística permite traducir una relación lineal entre variables en una probabilidad interpretable y acotada.

La interpretación de los coeficientes requiere distinguir entre el parámetro estimado y su transformación exponencial. El coeficiente representa el cambio en el *logit*; al aplicar la función exponencial al coeficiente se obtiene el *odds* ratio, que indica el cambio multiplicativo en el riesgo asociado a una variación unitaria del predictor. Esta lectura resulta especialmente utilizada en estudios clínicos y epidemiológicos.

Según De la Fuente Fernández (s.f.), cuando el *odds* ratio es mayor que uno, el predictor incrementa la probabilidad del evento; cuando es menor que uno, la reduce. Si el valor es cercano a uno, el efecto resulta poco apreciable. Esta interpretación se mantiene siempre bajo la condición de que las demás variables permanezcan constantes.

Desde el punto de vista metodológico, la correcta especificación del modelo implica considerar posibles interacciones entre variables y respetar la estructura jerárquica cuando se introducen términos combinados. El autor enfatiza que la incorporación de interacciones exige mantener también los términos de orden inferior, ya que su omisión puede distorsionar la estimación de efectos parciales (De la Fuente Fernández, s.f.).

En el ámbito de la evaluación de clasificadores, Fawcett (2006) señala que el modelo logístico genera puntuaciones continuas que pueden interpretarse como estimaciones de probabilidad. Estas puntuaciones permiten establecer distintos puntos de corte para clasificar observaciones, lo que influye en la tasa de verdaderos positivos y falsos positivos.

Entre los aspectos clave en la interpretación de coeficientes se destacan:

- El coeficiente estimado refleja cambios en el logit, no directamente en la probabilidad.
- El odds ratio permite expresar el efecto en términos multiplicativos sobre el riesgo.

Tabla 3. Interpretación de coeficientes en regresión logística

Elemento interpretado	Significado técnico	Aplicación práctica
Coeficiente estimado	Cambio en el *logit* ante variación	Impacto parcial manteniendo

	unitaria del predictor	constantes otras variables
Exponencial del coeficiente	*Odds ratio*	Cambio multiplicativo en el riesgo
Signo positivo	Incremento en el *logit*	Mayor probabilidad del evento
Signo negativo	Disminución en el *logit*	Menor probabilidad del evento

Fuente: elaboración propia con base en De la Fuente Fernández (s.f.)

La relación entre coeficientes y probabilidad no es lineal. Esto significa que un mismo incremento en un predictor puede producir cambios distintos en la probabilidad según el punto de partida. En probabilidades intermedias, el efecto marginal suele ser más pronunciado; en probabilidades cercanas a cero o uno, el efecto se atenúa. Esta característica responde a la forma sigmoidal de la función logística.

Asimismo, la interpretación debe considerar la presencia de variables categóricas codificadas mediante variables indicadoras. De la Fuente Fernández (s.f.) describe que, cuando una variable

nominal posee varias categorías, se incorporan variables *dummy* que permiten comparar cada categoría con un nivel de referencia. El coeficiente asociado a cada variable indicadora expresa el cambio relativo en el riesgo respecto de esa categoría base.

Tabla 4. Interpretación de variables categóricas en regresión logística

Tipo de variable	Tratamiento en el modelo	Interpretación del coeficiente
Variable cuantitativa	Se incorpora directamente	Cambio en el *logit* por unidad adicional
Variable categórica (k categorías)	Se crean k−1 variables *dummy*	Comparación con categoría de referencia
Interacción entre variables	Se incluyen términos combinados	Efecto conjunto condicionado

Fuente: elaboración propia con base en De la Fuente Fernández (s.f.)

En síntesis, la función logística establece el puente entre una estructura lineal de predictores y una interpretación probabilística coherente. La lectura adecuada de los coeficientes exige comprender la diferencia entre efecto sobre el *logit* y efecto sobre la probabilidad, así como la relevancia del *odds ratio* como medida de riesgo relativo. Esta comprensión resulta imprescindible para aplicar el modelo en contextos donde la clasificación binaria tiene consecuencias prácticas significativas.

Estimación del modelo y supuestos en regresión logística

La regresión logística se estima mediante el principio de máxima verosimilitud. A diferencia de la regresión lineal, donde se minimiza la suma de los errores al cuadrado, aquí se busca el conjunto de parámetros que maximiza la probabilidad de observar los datos disponibles. Este enfoque resulta coherente con la naturaleza probabilística del modelo, ya que cada observación contribuye a la función de verosimilitud en función de su probabilidad estimada.

De la Fuente Fernández (s.f.) señala que el proceso de estimación requiere procedimientos iterativos debido a que no existe una solución analítica cerrada. El algoritmo ajusta progresivamente los coeficientes hasta alcanzar convergencia, es decir, hasta que

las variaciones entre iteraciones resultan suficientemente pequeñas. Este carácter iterativo explica por qué la implementación computacional resulta indispensable en la práctica profesional.

La evaluación del ajuste puede realizarse mediante distintos contrastes. El autor describe, por ejemplo, la utilización del estadístico de desviación y el contraste global de Hosmer-Lemeshow para analizar la adecuación del modelo a los datos observados. Estos procedimientos permiten contrastar si las probabilidades estimadas difieren significativamente de las frecuencias observadas en grupos de riesgo similares.

Desde el punto de vista aplicado, la estimación del modelo implica atender a varios aspectos:

- Verificar la significación estadística de los coeficientes individuales.
- Evaluar la bondad global del ajuste mediante contrastes apropiados.
- Analizar posibles factores de confusión que alteren la estabilidad de los coeficientes.

- Revisar la presencia de interacciones relevantes entre variables explicativas.

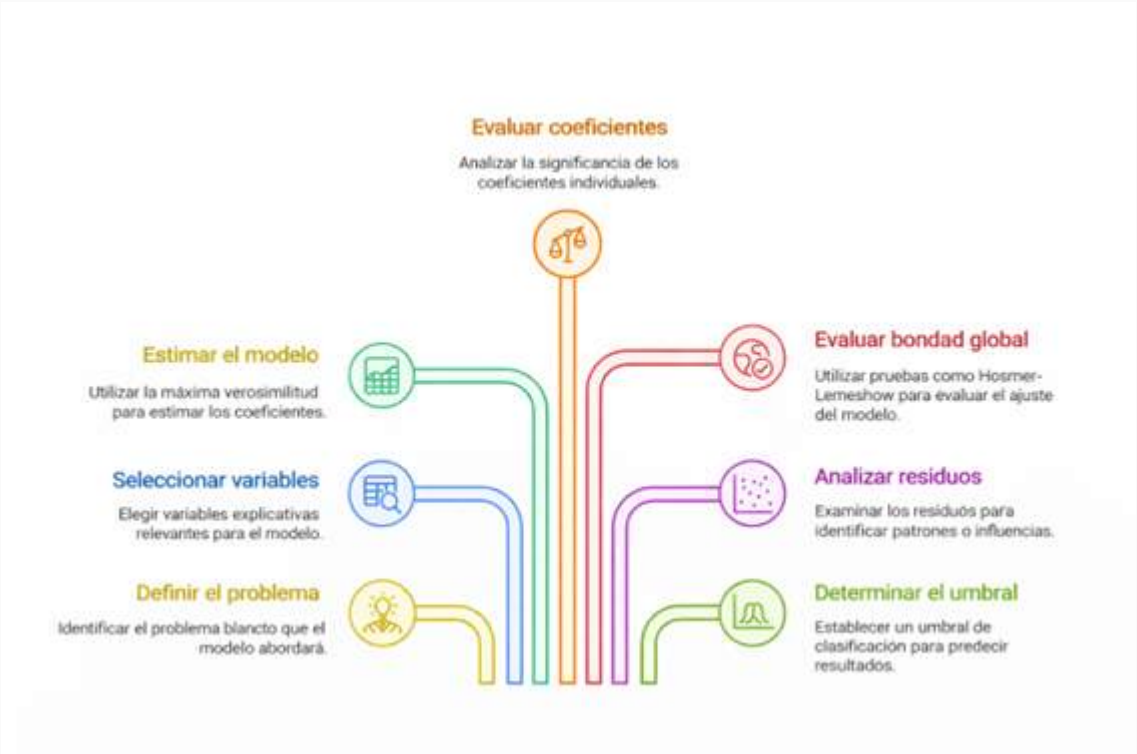
La estimación del modelo logístico no concluye con la obtención de coeficientes; requiere un análisis integral de ajuste, estabilidad e interpretación.

Un aspecto metodológico relevante es el tratamiento de variables categóricas. Como explica De la Fuente Fernández (s.f.), las variables nominales deben transformarse en variables *dummy* para poder incorporarse al modelo. Esta codificación permite comparar cada categoría respecto de un nivel de referencia, facilitando la interpretación en términos de riesgo relativo.

Asimismo, el análisis de residuos y medidas de influencia constituye una herramienta útil para identificar observaciones con alto impacto en la estimación. El autor describe diferentes indicadores que permiten evaluar el efecto de casos individuales sobre los parámetros estimados, lo que resulta particularmente importante en muestras de tamaño reducido o con valores extremos.

En el marco de la clasificación, Fawcett (2006) destaca que los modelos probabilísticos permiten generar puntuaciones continuas que pueden transformarse en decisiones binarias mediante la elección de un umbral. Esta posibilidad conecta la estimación del modelo con su desempeño como clasificador, ya que diferentes puntos de corte producen distintos equilibrios entre sensibilidad y especificidad.

Figura 1. Proceso de estimación y validación del modelo logístico



Fuente: elaboración propia

Este esquema refleja que la estimación constituye un proceso estructurado que integra modelización, evaluación y decisión.

Tabla 5. Etapas del proceso de estimación en regresión logística

Etapa	Objetivo técnico	Impacto en la práctica profesional
Especificación del modelo	Definir variables dependiente e independientes	Asegura coherencia conceptual
Estimación iterativa	Obtener coeficientes por máxima verosimilitud	Permite cuantificar efectos
Evaluación global	Analizar bondad de ajuste	Verifica adecuación del

		modelo
Diagnóstico de observaciones	Identificar influencia o casos atípicos	Mejora estabilidad del modelo
Definición de umbral	Transformar probabilidad en decisión binaria	Optimiza desempeño clasificatorio

Fuente: elaboración propia con base en De la Fuente Fernández (s.f.) y Fawcett (2006)

En síntesis, la estimación del modelo logístico combina fundamentos probabilísticos con procedimientos de validación estadística. Su aplicación profesional exige integrar la interpretación de coeficientes, la evaluación global del ajuste y el análisis del desempeño como clasificador. Este enfoque integral permite garantizar que el modelo no solo sea estadísticamente consistente, sino también operativamente útil.

CONTINUAR

2. Evaluación, validación y aplicaciones de modelos de clasificación

Introducción

En el análisis profesional de modelos de clasificación, estimar probabilidades constituye solo una etapa del proceso. La pregunta decisiva aparece inmediatamente después: ¿qué tan confiables son las predicciones generadas? ¿Cómo se mide el desempeño de un modelo cuando su objetivo es clasificar correctamente eventos binarios? Estas preguntas conducen al estudio de métricas específicas y técnicas de validación que permiten evaluar la capacidad discriminativa del modelo.

En el ámbito del aprendizaje automático y la estadística aplicada, la simple tasa de aciertos puede resultar insuficiente, especialmente cuando las clases están desbalanceadas. Fawcett (2006) advierte que la precisión global puede ocultar comportamientos deficientes si no se analiza la relación entre verdaderos positivos y falsos positivos. De allí la importancia de herramientas como la curva ROC, que permite visualizar el desempeño del clasificador bajo distintos umbrales de decisión.

Asimismo, la validación cruzada se ha consolidado como un procedimiento central para estimar la capacidad predictiva fuera de muestra. Andrle (2018) explica que esta técnica divide los datos en subconjuntos sucesivos de entrenamiento y prueba, permitiendo obtener estimaciones más estables del desempeño. Este enfoque contribuye a reducir el riesgo de sobreajuste y mejora la generalización del modelo.

En esta unidad abordaremos las principales métricas de evaluación, los procedimientos de validación y diversas aplicaciones prácticas en salud, marketing y educación. El objetivo consiste en integrar la estimación probabilística trabajada previamente con criterios rigurosos de evaluación y contextos reales de aplicación profesional.

Validación cruzada y métricas de precisión

En el análisis de modelos de clasificación, evaluar el desempeño implica estimar qué tan bien generaliza el modelo a datos no observados previamente. Un modelo puede ajustarse adecuadamente al conjunto de entrenamiento y, sin embargo, presentar un rendimiento inferior cuando se aplica a nuevos

datos. Este fenómeno, conocido como sobreajuste, motiva la necesidad de procedimientos de validación sistemáticos.

La validación cruzada constituye una técnica ampliamente utilizada para estimar la capacidad predictiva fuera de muestra. Andrle (2018) explica que el procedimiento consiste en dividir el conjunto de datos en subconjuntos o particiones. En cada iteración, uno de estos subconjuntos se utiliza como conjunto de prueba, mientras que los restantes se emplean para entrenar el modelo. Este proceso se repite tantas veces como particiones existan, y los resultados se promedian para obtener una estimación más estable del desempeño.

El método más difundido es la validación cruzada de tipo *k-fold*, donde los datos se dividen en *k* bloques aproximadamente del mismo tamaño. Cada bloque actúa una vez como conjunto de validación. Esta estrategia reduce la variabilidad asociada a una única partición y permite aprovechar de manera más eficiente la información disponible.

Desde el punto de vista profesional, la validación cruzada permite:

- Estimar de manera más robusta la precisión del modelo, reduciendo la dependencia de una única división

entrenamiento-prueba.

- Detectar posibles problemas de sobreajuste antes de implementar el modelo en un entorno operativo.

La evaluación del desempeño en clasificación binaria no puede limitarse a la tasa global de aciertos. Fawcett (2006) señala que la precisión puede resultar engañosa cuando existe desbalance entre clases. En estos casos, un modelo puede clasificar correctamente la mayoría de los casos negativos y aun así presentar baja capacidad para identificar correctamente los positivos.

Por ello, se utilizan métricas derivadas de la matriz de confusión, que permite distinguir entre verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Estas métricas proporcionan una visión más detallada del comportamiento del clasificador.

Tabla 6. Principales métricas de evaluación en clasificación binaria

Métrica	Definición conceptual	Interpretación aplicada

Exactitud (*accuracy*)	Proporción total de clasificaciones correctas	Desempeño global del modelo
Sensibilidad	Proporción de positivos correctamente identificados	Capacidad para detectar el evento
Especificidad	Proporción de negativos correctamente identificados	Capacidad para descartar no eventos
Precisión	Proporción de positivos predichos que son realmente positivos	Confiabilidad de las predicciones positivas

Fuente: elaboración propia con base en Fawcett (2006)

La combinación de validación cruzada y métricas específicas permite evaluar de manera integral el desempeño del modelo. Andrle (2018) destaca que este enfoque contribuye a seleccionar modelos con mejor capacidad de generalización, evitando

decisiones basadas únicamente en el ajuste interno al conjunto de entrenamiento.

En síntesis, la validación cruzada aporta un marco metodológico para estimar la estabilidad predictiva, mientras que las métricas derivadas de la matriz de confusión permiten comprender con mayor detalle el equilibrio entre distintos tipos de errores. Esta integración constituye la base para una evaluación rigurosa de modelos de clasificación en contextos profesionales.

Matriz de confusión, sensibilidad, especificidad y curva ROC

En el análisis de modelos de clasificación binaria, la matriz de confusión constituye el punto de partida para comprender el comportamiento del clasificador. Esta herramienta organiza los resultados en cuatro categorías: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Cada una de estas categorías aporta información específica sobre el tipo de aciertos y errores cometidos por el modelo.

Fawcett (2006) explica que muchas métricas de desempeño derivan directamente de esta estructura matricial. La sensibilidad, también denominada tasa de verdaderos positivos, mide la proporción de casos positivos correctamente identificados. La especificidad, por su parte, refleja la proporción de casos negativos correctamente clasificados. Estas métricas permiten distinguir entre la capacidad del modelo para detectar el evento y su capacidad para evitar falsas alarmas.

En contextos profesionales, esta distinción adquiere relevancia concreta. En salud, una alta sensibilidad reduce la probabilidad de omitir diagnósticos positivos; en sistemas de control de fraude, una alta especificidad evita bloquear transacciones legítimas. La elección del equilibrio adecuado depende del costo asociado a cada tipo de error.

La curva *ROC* amplía esta evaluación al representar gráficamente la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para distintos umbrales de decisión. Fawcett (2006) señala que esta representación permite visualizar el desempeño relativo del clasificador en todo el rango de posibles puntos de corte, evitando depender de un único umbral fijo.

Desde el punto de vista operativo, el análisis basado en matriz de confusión y curva *ROC* permite:

- Evaluar el impacto de distintos umbrales sobre la relación entre sensibilidad y especificidad.

- Comparar modelos considerando su capacidad discriminativa a lo largo de múltiples escenarios de decisión.

Un aspecto destacado por Fawcett (2006) es que la curva *ROC* resulta insensible a cambios en la distribución de clases. Esto implica que la forma de la curva permanece estable incluso cuando la proporción de eventos positivos y negativos varía, característica particularmente útil en dominios con clases desbalanceadas.

Tabla 7. Métricas derivadas de la matriz de confusión

Métrica	Definición conceptual	Aplicación práctica
Sensibilidad	Proporción de verdaderos positivos sobre el total de positivos reales	Detección efectiva del evento

Especificidad	Proporción de verdaderos negativos sobre el total de negativos reales	Control de falsas alarmas
Tasa de falsos positivos	Proporción de negativos clasificados como positivos	Nivel de error tipo I
Precisión	Proporción de positivos predichos que son realmente positivos	Confiabilidad de alertas

Fuente: elaboración propia con base en Fawcett (2006)

La representación en el espacio *ROC* permite identificar clasificadores que dominan a otros en términos de mayor tasa de verdaderos positivos para igual o menor tasa de falsos positivos. Fawcett (2006) describe que un clasificador situado más cerca del vértice superior izquierdo del gráfico presenta mejor desempeño, ya que combina alta sensibilidad con baja tasa de falsas alarmas.

Además, el área bajo la curva *ROC*, conocida como *AUC*, resume en un único valor la capacidad discriminativa global del modelo. Un valor cercano a uno indica alta capacidad para distinguir entre clases, mientras que un valor cercano a 0,5 sugiere comportamiento equivalente al azar.

En síntesis, la matriz de confusión ofrece una evaluación puntual en un umbral específico, mientras que la curva *ROC* proporciona una visión integral del desempeño a lo largo de distintos escenarios de decisión. Esta combinación fortalece el análisis profesional, permitiendo seleccionar modelos con base en criterios consistentes y comparables.

En conclusión, la evaluación de modelos de clasificación requiere integrar múltiples perspectivas analíticas. La matriz de confusión permite comprender con precisión la distribución de aciertos y errores en un punto de corte específico, mientras que la curva *ROC* amplía el análisis hacia un rango continuo de umbrales posibles. Esta doble mirada facilita decisiones fundamentadas cuando los costos asociados a falsos positivos y falsos negativos difieren significativamente.

Desde una perspectiva profesional, el uso combinado de sensibilidad, especificidad y área bajo la curva fortalece la comparación entre modelos y contribuye a seleccionar alternativas con mayor capacidad discriminativa. Tal como

plantea Fawcett (2006), el análisis basado en *ROC* permite trascender la simple tasa de aciertos y adoptar un enfoque más robusto y contextualizado del desempeño clasificatorio. Este marco evaluativo prepara el terreno para analizar aplicaciones concretas en distintos campos profesionales, donde la clasificación binaria adquiere implicancias estratégicas.

Matriz de confusión, sensibilidad, especificidad y curva ROC

La regresión logística y los modelos de clasificación encuentran aplicaciones directas en múltiples campos profesionales donde la decisión se estructura en términos binarios. En estos contextos, el objetivo consiste en estimar la probabilidad de ocurrencia de un evento y, posteriormente, transformar esa probabilidad en una acción concreta. La integración entre modelización probabilística y criterios de evaluación rigurosos permite que estas herramientas se incorporen en procesos estratégicos de gestión.

En el ámbito de la salud, la regresión logística se utiliza para estimar el riesgo de desarrollar una enfermedad en función de variables clínicas, demográficas y conductuales. De la Fuente Fernández (s.f.) describe que el modelo permite interpretar los coeficientes en términos de riesgo relativo mediante el odds

ratio, facilitando la comunicación de resultados en estudios epidemiológicos. En este escenario, la sensibilidad adquiere especial relevancia, ya que omitir un caso positivo puede implicar consecuencias significativas.

En marketing, los modelos de clasificación se aplican para predecir la probabilidad de que un cliente responda a una campaña, abandone un servicio o incurra en incumplimiento de pago. La posibilidad de ajustar el umbral de decisión permite equilibrar la captación de clientes potenciales con el costo de contactar perfiles con baja probabilidad de respuesta. En este tipo de aplicaciones, el análisis de la curva ROC contribuye a seleccionar puntos de corte acordes con los objetivos comerciales.

En educación, la regresión logística se emplea para identificar factores asociados al abandono escolar o al bajo rendimiento académico. Andrle (2018) destaca la importancia de validar adecuadamente los modelos mediante técnicas de partición y validación cruzada, especialmente cuando los datos disponibles son limitados. Esta precaución metodológica mejora la estabilidad de las conclusiones y fortalece su utilidad en la formulación de políticas educativas.

Desde una perspectiva aplicada, estas herramientas permiten:

- Estimar probabilidades individuales de ocurrencia de eventos relevantes para la toma de decisiones.

La clasificación binaria convierte estimaciones probabilísticas en decisiones operativas con impacto concreto en distintos sectores profesionales.

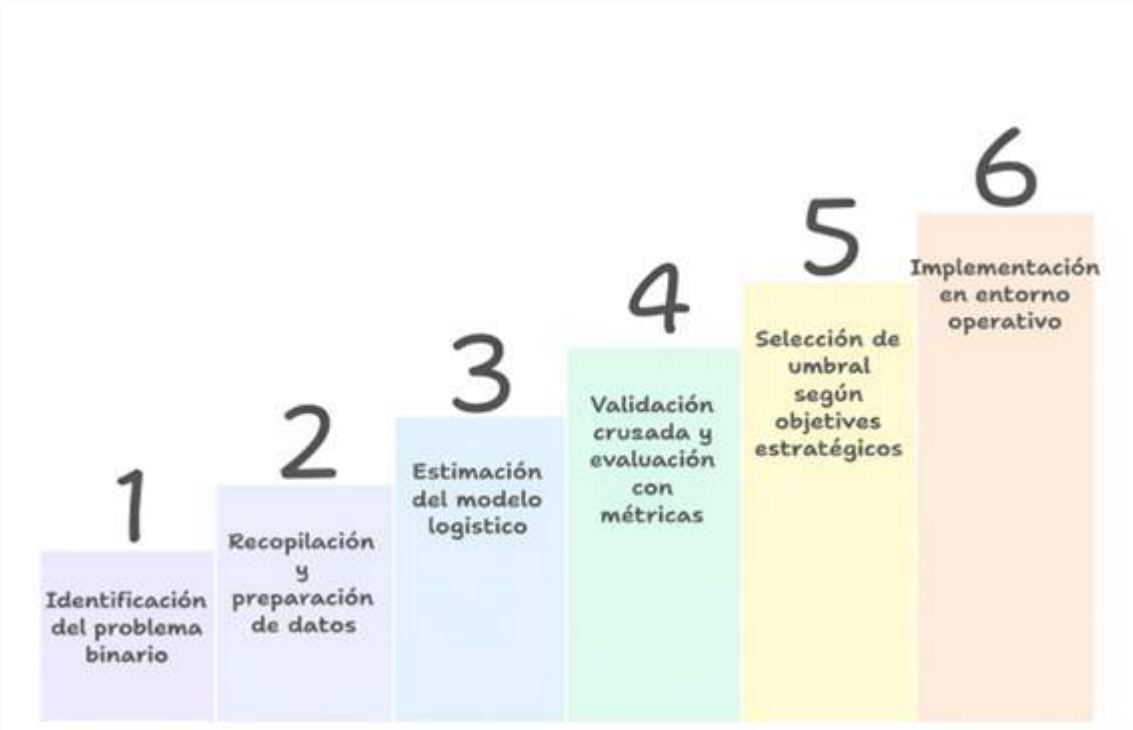
Tabla 8. Aplicaciones profesionales de la regresión logística y clasificación

Campo de aplicación	Evento modelado	Impacto de la clasificación
Salud	Presencia o ausencia de enfermedad	Diagnóstico y prevención temprana
Marketing	Respuesta a campaña o	Optimización de estrategias comerciales

	abandono de cliente	
Educación	Abandono escolar o aprobación académica	Diseño de intervenciones focalizadas

Fuente: elaboración propia con base en De la Fuente Fernández (s.f.) y Andrle (2018)

Figura 2. Aplicación profesional del modelo de clasificación



Fuente: elaboración propia

Este esquema sintetiza el recorrido desde la definición del problema hasta la implementación del modelo en un contexto real. La aplicación efectiva de la regresión logística requiere integrar fundamentos estadísticos, criterios de evaluación rigurosos y comprensión del entorno donde se tomará la decisión.

A modo de cierre, resulta pertinente subrayar que la aplicación de modelos de clasificación no se agota en la estimación de probabilidades ni en la evaluación técnica del desempeño. La implementación en contextos reales exige comprender las implicancias prácticas de cada decisión derivada del modelo. En salud, una predicción puede orientar un tratamiento; en marketing, puede definir la asignación de recursos; en educación, puede activar intervenciones preventivas.

Fawcett (2006) sostiene que la evaluación rigurosa mediante herramientas como la curva ROC contribuye a evitar

interpretaciones simplificadas del rendimiento de un clasificador. Esta advertencia adquiere especial relevancia cuando los modelos se utilizan en entornos donde los errores tienen consecuencias diferenciadas. La selección de un umbral de decisión implica adoptar una postura estratégica frente al equilibrio entre sensibilidad y especificidad.

Asimismo, Andrle (2018) enfatiza que la validación sistemática mejora la capacidad de generalización del modelo y reduce el riesgo de decisiones basadas en patrones espurios. Esta perspectiva metodológica fortalece la confianza en los resultados y consolida el vínculo entre modelado estadístico y práctica profesional.

En síntesis, la regresión logística y los métodos de clasificación constituyen herramientas analíticas que integran modelización probabilística, evaluación rigurosa y aplicación estratégica. Su uso profesional requiere considerar:

- La coherencia conceptual entre el problema planteado y la estructura del modelo.
- La calidad y representatividad de los datos utilizados en la estimación.

- La elección fundamentada del umbral de decisión según los costos asociados a cada tipo de error.
- La validación cruzada como mecanismo de control de estabilidad predictiva.
- La interpretación contextualizada de los resultados en función del sector de aplicación.

Este enfoque integral permite que el análisis de eventos binarios trascienda el ámbito técnico y se convierta en un instrumento sólido para la toma de decisiones basada en evidencia cuantitativa.

CONTINUAR

Referencias

Andrle, M. (2018). *Machine learning for economists: Part 1 – Cross-validation*. International Monetary Fund.

De la Fuente Fernández, S. (s.f.). *Regresión logística*. Facultad de Ciencias Económicas y Empresariales, Departamento de Economía Aplicada.

Fawcett, T. (2006). *An introduction to ROC analysis*. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

CONTINUAR