**Regressão Logística**

Regressão logística é uma técnica estatística usada para modelar a probabilidade de uma variável dependente binária (0 ou 1, sim ou não, sucesso ou fracasso) baseada em variáveis independentes.

**Fórmula Matemática**

A regressão logística modela a probabilidade como uma função logística (sigmoide) baseada em uma combinação linear das variáveis independentes.

Ícone

O conteúdo gerado por IA pode estar incorreto.

Nome da empresa

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

Onde:

* p: Probabilidade estimada de uma observação pertencer à classe positiva.
* β0​: Intercepto. (É o valor da função quando todas as variáveis independentes são zero - Log-odds)
* β1, β2, ..., βn ​: Coeficientes (impacto) das variáveis independentes - Parâmetros.
* x1, x2, ..., ​: Variáveis independentes.

**OBS:** Se β>0, a variável aumenta a chance da classe 1.

Se β<0, a variável diminui a chance.

Essa equação utiliza a função sigmoide, que transforma qualquer valor real em um número entre 0 e 1, permitindo interpretar a saída como uma probabilidade.

A decisão final é feita com base em um limiar (threshold), geralmente 0.5   
Classe= {1 se p≥0.5, 0 se p<0.5​}

**Assunções do Modelo**

* **Dependência linear**: A relação entre variáveis independentes e o logit (log das odds) é linear.
* **Independência das observações**: As amostras devem ser independentes umas das outras.
* **Ausência de multicolinearidade**: As variáveis independentes não devem ser altamente correlacionadas.
* **Homogeneidade da variância (não essencial)**.

**Métricas de Avaliação**

* **Acurácia**: Porcentagem de classificações corretas.
* **Precisão, Recall e F1-score**: Medem a performance em cada classe.
* **ROC e AUC**: Avaliam o desempenho do modelo em diferentes limiares.
* **Matriz de Confusão**: Analisa verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

**Divisão da Amostra**

Para validação do modelo, usamos o recurso da divisão de amostra, aonde particionamos a amostra em duas partes

* Modelagem (treino) de 50% a 75% do tamanho da base
* Teste de 50% a 25% do tamanho da base

Então usamos a parte de treino para treinarmos/criarmos o modelo e então a parte de teste para validarmos a “qualidade” do modelo.

OBS: Se a amostra for muito pequena, validar a função no mesmo grupo que foi utilizado para desenvolver a função

**Medidas de Ajuste do Modelo**

Quando criamos um modelo de regressão logística, precisamos avaliar o quão bem o modelo se ajustou aos dados. E fazemos isso com:

**Log-Verossimilhança (-2LL) ou -2 Log Verossimilhança**

**A verossimilhança mede:**

Quão provável é observar os dados Yi que temos dado o modelo e os parâmetros β. O -2LL é simplesmente o dobro do valor negativo da log-verossimilhança:

**Interpretação:**

Quanto menor o -2LL, melhor o ajuste do modelo.

Um -2LL menor significa que o modelo está atribuindo probabilidades maiores para as observações corretas (0s e 1s reais).

**OBS:** Quando a verossimilhança for 1, indica ajuste perfeito e o valor do -2LL é 0 (Praticamente impossível de acontecer)

**Log-Likelihood e Pseudo R²**

**Log-Likelihood:** quão provável os dados observados são, dados os parâmetros do modelo.

Quanto maior (menos negativo) o log-likelihood, melhor o ajuste.

**Pseudo R²**

**Cox & Snell R² (Semelhante ao R² da regressão linear múltipla)**

Uma imagem contendo nome da empresa

O conteúdo gerado por IA pode estar incorreto.

Obs: Nunca atinge 1 (ideal)

**Nagelkerke R²**

É uma correção do Cox & Snell para forçar o máximo a 1:

Ícone

O conteúdo gerado por IA pode estar incorreto.

**Curva ROC**

Quanto mais distante a curva estiver da diagonal melhor será o poder discriminatório do Modelo, ou seja, mais próximo do canto superior esquerdo, melhor.

**Onde o AUC:**

* Abaixo de 0.5: Não há discriminação
* Entre 0,7 e 0,8: Discriminação aceitável
* Maior que 0,8: Discriminação excelente