

## Métricas de Regressão e Desempenho em IA

Adaptado por: Dr. Arnaldo de Carvalho Junior – Jul 31, 2024.

### 1. INTRODUÇÃO

Como se verifica o desempenho de um modelo?

Nenhum modelo está correto. Então, melhor o modelo, melhor ele se ajusta aos dados. É tão simples quanto isso. Há muitas maneiras, como serão apresentados a seguir.

#### 1.1. Regressão Linear/Múltipla

Primeiro, segue uma introdução da Regressão Linear/Múltipla Linear

$$y = B_0 + B_1x_1 + B_2x_2 + E \tag{1}$$

Na equação (1) de regressão acima, 'E' significa o erro que é introduzido quando o modelo realmente não se encaixa nos dados. O objetivo de um bom modelo é minimizar esse erro. A seguir é apresentado como se comparam os diferentes modelos:

1. R-Quadrado
2. R-Quadrado Ajustado
3. MSE
4. RMSE

#### 1.2. R-Quadrado

A proporção de variância na variável dependente que é prevista a partir das variáveis independentes. O valor varia de 0 a 1.

1 mostra que a regressão explica perfeitamente a relação. 0 o oposto.

Se  $R^2$  (Leia como *R-Squared*) = 0,43 para a equação (1) de regressão acima, então significa que 43% da variabilidade em  $y$  é explicado pelas variáveis  $x_1$  e  $x_2$ .

Mas há uma falha. À medida que o número de termos aumenta, o  $R^2$  pode permanecer constante ou aumentar (Existem provas estatísticas deste — Não discutido aqui). Isso acontece mesmo que não haja uma boa relação entre as variáveis e a variável dependente. Então agora o quê? Ajuste-o!

### 1.3. R-Quadrado Ajustado

Um valor ajustado que considerará a relação entre as variáveis. Diminuirá o valor para variáveis que não melhoram o modelo existente.

### 1.4. RMSE

O erro quadrático médio (*mean square error* – MSE) nada mais é do que a média dos quadrados da diferença entre os valores observados e previstos.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \widehat{y_i})^2 \quad (2)$$

n -> número de termos. y(i) significa o valor observado para o i-ésimo termo e y(^) fala sobre o valor previsto para esse termo específico. A diferença é o termo Erro. Você soma os quadrados de todos os termos de erro e divide pelo grau de liberdade (número de variáveis independentes).

Por que quadrado? Para que o erro seja acumulado em vez de cancelar um ao outro.

Mas o que isso tem a ver com RMSE? É a raiz do MSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \widehat{y_i})^2} \quad (3)$$

Depois de elevá-lo ao quadrado, é necessário reduzir a escala para as unidades originais. De tudo isso, deve ter tido agora alguma ideia sobre como os modelos podem ser comparados em uma regressão.

### 1.5. E quanto à Regressão Logística?

Para testar o desempenho de um modelo de classificação, uma matriz de confusão pode ser usada. Em palavras simples, é uma matriz que compreende instâncias de eventos previstos e reais. Representação do mesmo na Figura 1 a seguir

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figura 1 – Matriz de instâncias de eventos previstos e reais.

Os quatro termos a seguir são os principais:

**a) Verdadeiros Positivos** — Qualquer evento positivo que tenha sido corretamente previsto. Por exemplo: Passar em um exame (*positive event*) e foi previsto corretamente.

**b) Verdadeiros Negativos** — Qualquer evento negativo que tenha sido corretamente previsto. Por exemplo: Não passar em um exame (*negative event*) e foi previsto corretamente como '*failed*'.

**c) Falsos Positivos** — Qualquer evento negativo que tenha sido previsto positivo. Por exemplo: O paciente não está tendo nenhuma doença. Mas os resultados do teste disseram (previsto) a pessoa está infectada. Isto pode ter os efeitos nocivos de “ser submetido desnecessariamente a tratamento médico indesejado”.

**d) Falso Negativo** — Qualquer evento positivo que tenha sido previsto como negativo. Por exemplo: Há uma ameaça a um local público e a equipe de inteligência não consegue identificá-lo, ou seja, denunciá-lo como não uma ameaça. Isso pode ser um problema sério.

Os termos falam sobre as previsões explícita e implicitamente sobre o evento real. (Verdadeiramente/Falsamente) previu o evento como (Positivo/Negativo). Se for verdade, os eventos reais são iguais à previsão e se for falsa os eventos reais são o oposto.

Existem outros termos comumente usados (frequentemente solicitados nas entrevistas):

- **Erro Tipo I** — Falsos Positivos (FP).
- **Erro Tipo II** — Falsos Negativos (FN).

**Sensibilidade = Recordar (*Recall*) = Taxa positiva verdadeira (*true positive rate* – TPR)**

= Com que frequência o modelo previu o evento positivo corretamente. A proporção entre eventos positivos corretamente previstos para o total de eventos positivos.

$$\frac{TP}{TP + FN} \quad (4)$$

O numerador, TP, representa o número de verdadeiros positivos (*true positive*), enquanto o denominador (TP + FN) (*true positive + false negative*) representa o número total de casos

positivos reais no conjunto de dados. Ao dividir o número de verdadeiros positivos pela soma de verdadeiros positivos e falsos negativos, obtemos o valor da sensibilidade.

Na Ciência, sensibilidade é a capacidade de identificar corretamente casos positivos. A sensibilidade, também chamada de taxa positiva ou recordação positiva, é uma medição estatística que determina a proporção de casos positivos reais corretamente identificados por um modelo de aprendizado de máquina. Em outras palavras, a sensibilidade quantifica a capacidade do modelo de detectar corretamente as instâncias positivas dentro de um conjunto de dados.

A sensibilidade é uma métrica crítica no aprendizado de máquina, pois reflete a capacidade do modelo de detectar corretamente instâncias positivas dentro de um conjunto de dados. É especialmente importante em aplicações em que a identificação de casos positivos verdadeiros é de extrema importância, como diagnóstico médico, detecção de fraude ou detecção de anomalia.

**Especificidade = Taxa negativa verdadeira (*true negative rate* – TNR) =** Com que frequência o modelo previu os eventos negativos corretamente. Ela avalia a capacidade do modelo de identificar corretamente a ausência da variável de destino em um conjunto de dados. Pode ser calculada como a relação de eventos negativos corretamente previstos (*true negative* - TN) corretamente e o total de eventos negativos.

$$\frac{TN}{TN + FP} \quad (5)$$

Para entender a especificidade, precisamos entender o conceito de verdadeiros negativos (TN) e falsos positivos (*false positive* - FP). Os verdadeiros negativos representam os casos em que o modelo classifica corretamente as instâncias como negativas, enquanto falsos positivos ocorrem quando o modelo identifica incorretamente os casos negativos como positivos.

O numerador, TN, representa o número de negativos verdadeiros, enquanto o denominador (TN + FP) representa o número total de casos negativos reais no conjunto de dados. Ao dividir o número de negativos verdadeiros pela soma de verdadeiros negativos e falsos positivos, obtêm-se o valor da especificidade.

A especificidade é particularmente valiosa em cenários em que a identificação correta de casos negativos é crítica. Por exemplo, em exames médicos, a especificidade mede a

capacidade do modelo de identificar corretamente indivíduos sem uma doença ou condição específica. Um alto valor de especificidade implica que o modelo pode efetivamente filtrar indivíduos que são verdadeiramente negativos, reduzindo as chances de falsos positivos e minimizando intervenções ou tratamentos desnecessários.

**Taxa de Falso Positivo (*false positive rate* - FPR) = 1- Especificidade** = Com que frequência o modelo classificou os eventos negativos como positivos. A proporção de eventos positivos denominados incorretamente em relação ao total de eventos negativos.

$$\frac{FP}{TN + FP} \quad (6)$$

**Precisão** = Quantas vezes o modelo previu que o evento seria positivo e acabou sendo verdade. Seria a proporção entre positivos verdadeiros (TP) e casos previstos como positivos.

$$\frac{TP}{TP + FP} \quad (7)$$

**Acurácia** = Quantas vezes o modelo previu corretamente. A proporção dos casos verdadeiros para todos os casos.

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Onde  $TP$  = Verdadeiros positivos,  $TN$  = Verdadeiros negativos,  $FP$  = Falsos positivos e  $FN$  = Falsos negativos.

**Exemplo:** Um modelo classificou 100 tumores como [malignante](#) (a classe positiva) ou [benign](#) (a classe negativa):

a) Verdadeiro positivo (TP):

- Realidade: maligna
- Modelo de ML previsto: maligno

- **Número de resultados de TP: 1**

b) Falso positivo (FP):

- Realidade: benigno
- Modelo de ML previsto: maligno
- **Número de resultados de FP: 1**

c) Falso negativo (FN):

- Realidade: maligna
- Modelo de ML previsto: Benign
- **Número de resultados de FN: 8**

d) Verdadeiro negativo (TN):

- Realidade: benigno
- Modelo de ML previsto: Benign
- **Número de resultados de TN: 90**

$$Accuracy = A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1+90}{1+90+1+8} = 0.91 \quad (9)$$

A acurácia é de 0,91, ou 91% (91 previsões corretas de 100 exemplos totais). Embora a precisão de 91% possa parecer boa à primeira vista, outro modelo de classificador de tumor que sempre prevê o benigno alcançaria exatamente a mesma precisão (91/100 previsões corretas) do exemplo. Em outras palavras, o modelo não é melhor do que um que não tem capacidade preditiva de distinguir os tumores malignos dos tumores benignos.

A precisão por si só não conta a história completa quando se está trabalhando com um **conjunto de dados desequilibrado**, como este, em que há uma diferença significativa entre o número de rótulos positivos e negativos.

$$Precisão = P = \frac{TP}{TP + FP} = \frac{1}{1 + 1} = 0.5 \quad (10)$$

O modelo tem uma precisão de 0,5. Em outras palavras, quando prevê que um tumor é maligno, é correto em 50% do tempo.

$$Recall = R = \frac{TP}{TP + FN} = \frac{1}{1 + 8} = 0.11 \quad (11)$$

O modelo teve um *Recall* de 0,11. Em outras palavras, identifica corretamente 11% de todos os tumores malignos.

Em resumo:

- **Sensibilidade** refere-se a quanto bom **todos os eventos positivos atuais** estão em **previstos eventos positivos**.
- **Precisão** refere-se quanto **todos os eventos positivamente previstos** foram **corretamente previstos**.
- **Acurácia** é o quão corretamente todos os eventos foram previstos.
- **Especificidade** é o quão bom está a previsão de eventos negativos.

Como saber qual deles procurar? Bem, se o pesquisador está mais interessado em casos de Falso Positivo, deve optar pela precisão. Se os casos Falsos Negativos são o que o pesquisador procura, então Recordar (*Recall*) é uma boa medida.

A **Acurácia** é melhor se houver um cenário tendencioso. Se for imparcial — Precisão (*Precision*) e Recordar (*Recall*).

Pontuação F1 (**F1 Score**) ajuda se o pesquisador está incomodado com ambos os valores. É uma média harmônica, ou seja, uma média ponderada de precisão e recordação.

$$2 * \frac{Precision * Recall}{Precision + Recall} \quad (12)$$

Assim, do exemplo anterior:

$$F1\ Score = 2 * \frac{P * R}{P + R} = 2 * \frac{0.5 * 0.11}{0.5 + 0.11} = 0.18 \quad (13)$$

Lembrar dos resultados da regressão logística da probabilidade. Tem que haver algum fator decisivo ou valor que decida. “Ei amigo! Você tem menos de 0.5, você vai lá”. Este fator

decisivo, 0.5 (*default*) é o valor limite. Este limiar ajuda a regressão logística a classificar. Após a classificação, são calculados os quatro termos principais. E a **matriz de confusão** é construída.

## 1.6. ROC e AUC

A Curva Característica do Operador Receptor (*receiver operator characteristic curve* - ROC) ajuda a decidir o melhor valor limite.

A curva ROC é basicamente um gráfico traçado entre a **taxa de verdadeiros positivos** e a **taxa de falsos positivos**. Portanto, é um gráfico entre a frequência com que um modelo prevê eventos positivos como positivos e a frequência com que um modelo prevê eventos negativos como positivos.

A Área sob a curva (area under curve – AUC) é uma das métricas amplamente usadas e basicamente usada para classificação binária. A AUC de um classificador é definida como a probabilidade de um classificador classificará (*rank*) um exemplo positivo escolhido aleatoriamente maior que um exemplo negativo.

A Figura 2 a seguir apresenta a curva ROC para o exemplo da sessão anterior. O modelo ao longo da linha tracejada seria o pior classificador. Não se pode discriminar entre as classes. A Área Sob a Curva (AUC) seria 0.5 neste caso. O modelo ao longo da linha verde paralelo ao eixo x na parte superior é o melhor modelo. A AUC seria 1. Classifica perfeitamente os eventos positivos e negativos.



Figura 2 – Curva ROC do exemplo.  
Fonte: Adaptado de [1].



Qualquer modelo ou curva (*curve*) entre estes dois terá uma área maior que 0.5 e menor que 1. Isto resulta na sobreposição de classes e, portanto, são introduzidos erros do Tipo 1 e do Tipo 2.

Para diferentes limiares, é calculada a Sensibilidade e o FPR (1- Especificidade). FP baixo significa negativos verdadeiros mais altos. Uma curva é traçada. Dependendo de quantos Falsos Positivos aceitar, o limite é selecionado.

Para comparar modelos, aquele com AUC maior oferece o melhor. Na figura acima, o vermelho é melhor que o azul.

Em alguns casos, por exemplo, quando se tem muitos casos negativos, pode-se optar pela Precisão em caso de Falso Positivo. Para isso o pesquisador deve saber qual é o seu objetivo primeiro.

## 1.7. EXEMPLOS

A Figura 3 apresenta 6 imagens com a respectiva classificação.

Animal		Barn owl
		Chihuahua
		Sheep dog
Not animal		Apple
		Muffin
		Mop







Figura 3 – Animais e Coisas que se parecem animais.

A Figura 4 apresenta a matriz de confusão verdadeiro, falso, positivo e negativo.







		Predicted	
		Animal	Not animal
Actual	Animal	True Positives	False Negatives
	Not animal	False Positives	True Negatives

Figura 4 – Matriz de confusão.

A Figura 5 apresenta a mesma matriz, com predições perfeitas e imperfeitas..

		Predicted			
		Animal	Not animal		
Actual	Animal	  		True Positives	3
	Not animal		  	True Negatives	3
				False Positives	0
				False Negatives	0

(a)

		Predicted			
		Animal	Not animal		
Actual	Animal	 		True Positives	2
	Not animal		 	True Negatives	2
				False Positives	1
				False Negatives	1

(b)

Figura 5 – Matriz de confusão com predições perfeitas (a) e imperfeitas (b).

A Figura 6 apresenta a situação em que um modelo classifica todas as imagens como “animal”.







		Predicted			
		Animal	Not animal		
Actual	Animal	  		True Positives	3
	Not animal	  		True Negatives	0
				False Positives	3
				False Negatives	0
		Accuracy	50%	$\frac{3+0}{3+0+0+3}$	
		Precision	50%	$\frac{3}{3+3}$	
		Recall	100%	$\frac{3}{3+0}$	
		F1 score	67%	$2 \cdot \frac{0.5 \cdot 1}{0.5 + 1}$	

Figura 6 – Modelo com Classificação de Todas as Imagens como “animal”.

A Figura 7 apresenta a situação em que um modelo classifica todas as imagens como “Not animal”.







		Predicted			
		Animal	Not animal		
Actual	Animal		  	True Positives	0
	Not animal		  	True Negatives	3
				False Positives	0
				False Negatives	3
		Accuracy	50%	$\frac{0+3}{0+3+3+0}$	
		Precision	*100%	$\frac{0}{0+0}$	
		Recall	0%	$\frac{0}{0+3}$	
		F1 score	0%	$2 \cdot \frac{1 \cdot 0}{1+0}$	

Figura 7 – Modelo com Classificação de Todas as Imagens como “Not animal”.

Na Figura 8, é analisado o caso de mais predições (sobre predição) como “Not animal” e menos como “Animal”.

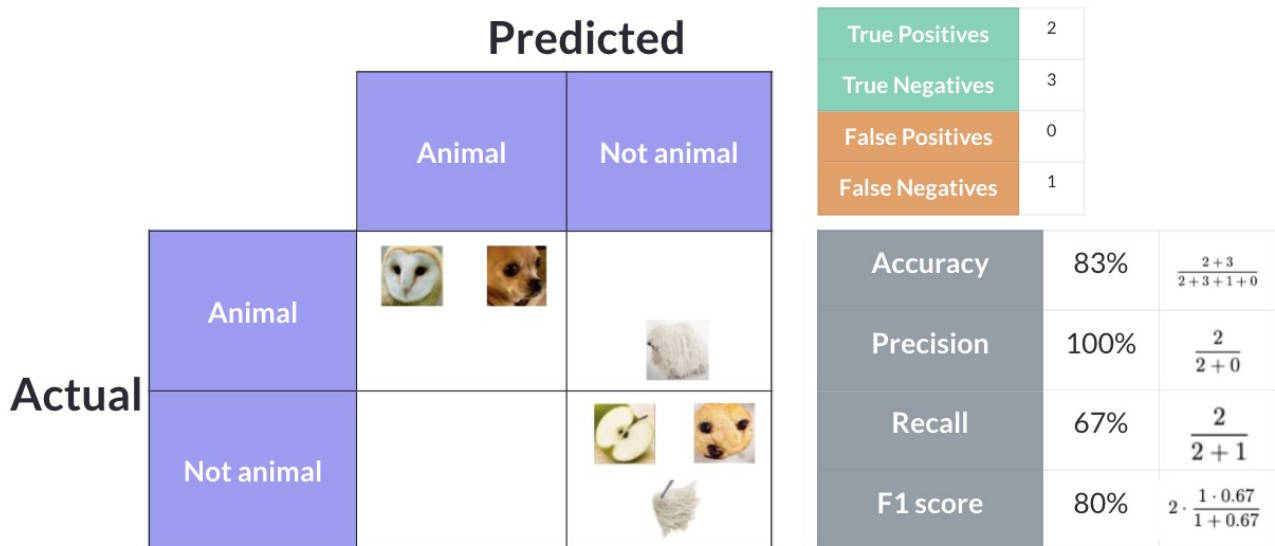


Figura 8 – Modelo com mais classificações como “Not animal” e menos de “animal”.

Na Figura 9, é analisado o caso de mais predições (sobre predição) como “Animal” e menos como “Not animal”.

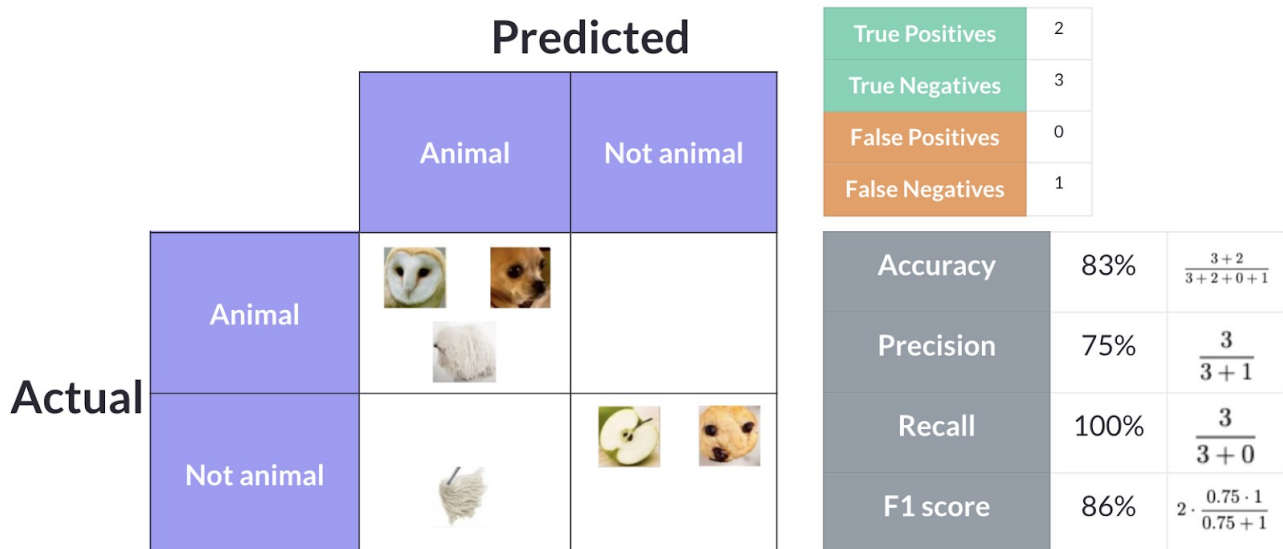


Figura 9 – Modelo com mais classificações como “Animal” e menos de “Not animal”.

## CONCLUSÃO

A **sensibilidade** é uma métrica valiosa para avaliar a capacidade de um modelo de aprendizado de máquina de identificar corretamente casos positivos. Um alto valor de sensibilidade sugere uma forte capacidade de capturar instâncias positivas, enquanto um baixo valor de sensibilidade indica uma maior probabilidade de falta de casos positivos. A interpretação da sensibilidade resulta no contexto dos requisitos do aplicativo e em conjunto

com outras métricas de desempenho ajuda a determinar a eficácia e a adequação do modelo.

A **especificidade** é uma métrica fundamental no aprendizado de máquina que mede a capacidade do modelo de identificar corretamente casos negativos. Ela desempenha um papel crucial na avaliação do desempenho do modelo e sua adequação a aplicações específicas, onde é essencial identificação precisa de instâncias negativas.

## REFERÊNCIAS

Kakanadan, U. Regression and performance metrics — Accuracy, precision, RMSE and what not!, Medium, 2024. Disponível em: <https://medium.com/analytics-vidhya/regression-and-performance-metrics-accuracy-precision-rmse-and-what-not-223348cfcafe>. Acessado em Jul 31, 2024.

STITH, L. What is Sensitivity and Specificity in Machine Learning. Robots.Net, Nov 2023. Disponível em: <https://robots.net/fintech/what-is-sensitivity-and-specificity-in-machine-learning>. Acessado em Ago 01, 2024.

TIGERSCHIOLD, T. What is Accuracy, Precision, Recall and F1 Score?, Labelf, 2022. Disponível em: <https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score>. Acessado em Nov 13, 2024.