Faculdade XPe

Capítulo 07

ESTATÍSTICA PARA CIÊNCIA DE DADOS

Arnaldo Satoru Gunzi

2024

Regressão Logística

No mundo real, frequentemente precisamos prever uma categoria de saída em vez de um número. Por exemplo, em uma financeira, precisamos determinar a probabilidade de um novo cliente ser Adimplente (1) ou Inadimplente (0). Se a probabilidade de ser Adimplente for maior, estaremos mais seguros em conceder o empréstimo.
A Regressão Linear é usada quando a variável resposta é numérica, mas, para prever uma classe, utilizamos a Regressão Logística, que é uma variação dos modelos lineares generalizados. A Regressão Logística modela a probabilidade de uma observação pertencer a uma categoria binária de saída, como 1 (ocorrência do evento) e 0 (não ocorrência).
A fórmula ajustada da Regressão Logística é:

$$\hat{y} = Pr\,(y = 1 \mid x) = \frac{e^{x^T\beta+\beta_0}}{1 + e^{x^T\beta+\beta_0}}$$

Onde X é matriz de preditores, β são os coeficientes da equação e 'e' é a função exponencial.
A estimação dos coeficientes β na Regressão Logística usa o método dos mínimos quadrados iterativos ou o método da máxima verossimilhança, frequentemente com o método de Newton-Raphson. Vamos utilizar pacotes computacionais para encontrar os valores da regressão, no link ao final do capítulo.

Interpretando o modelo ajustado

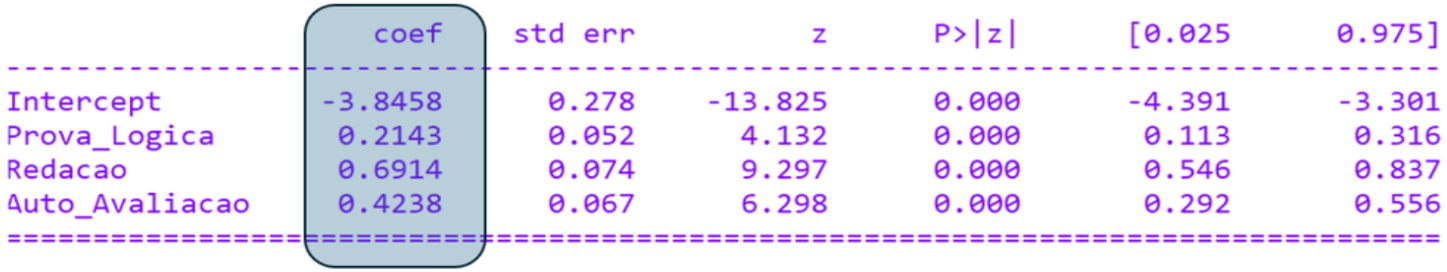
Assim como na Regressão Linear, a Regressão Logística serve tanto para explicação como para predição.
Suponha que você trabalha em uma grande empresa do segmento de RH e deseja agilizar suas contratações para uma determinada empresa que é um grande cliente. Esse cliente demanda sempre uma grande quantidade de contratações simultâneas, pois sempre que fecha um novo contrato, uma nova operação é inicializada e há uma grande demanda por novos funcionários.
A empresa de RH reuniu em um conjunto de dados as notas obtidas nas avaliações em processos seletivos anteriores e vinculou a performance dos candidatos, contendo n = 699 observações e rotulando como 'Boa' os funcionários que tiveram bom desempenho e 'Ruim' os funcionários que não obtiveram bom desempenho.
Utilizaremos a Regressão Logística tanto para mensurar o impacto de cada preditor na variável resposta, como também a utilizaremos para prever se um novo candidato será da classe 'Boa' ou da classe 'Ruim'.

Figura 22 - Conjunto de dados reunido pelo RH

	Prova_Logica	Redacao	Psicotecnico	Dinamica_Grupo	Fit_Cultural	Ingles	Avaliacao_RH	Auto_Avaliacao	Classe
1	2	1	1	1	2	1	2	1	Ruim
2	2	1	1	1	2	1	3	1	Ruim
3	5	1	1	1	2	1	2	1	Ruim
4	5	4	6	8	4	1	8	1	Boa
5	5	3	3	1	2	1	2	1	Ruim
6	2	3	1	1	3	1	1	1	Ruim
7	3	5	7	8	8	9	7	1	Boa
8	2	5	6	1	6	1	7	7	Boa
9	1	9	8	7	6	4	7	1	Boa
...									
697	6	1	1	1	8	1	7	1	Boa
698	5	7	1	1	5	1	1	1	Boa
699	1	1	1	1	2	1	2	1	Ruim

Vamos examinar o output que a ferramenta nos fornece para regressão logística. Para simplificar por enquanto, vamos utilizar as pontuações do candidato em *Prova_Logica, Redacao* e *Auto_Avaliacao* para explicar/prever a Classe.

Figura 23 - Output da Regressão Logística



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.8458	0.278	-13.825	0.000	-4.391	-3.301
Prova_Logica	0.2143	0.052	4.132	0.000	0.113	0.316
Redacao	0.6914	0.074	9.297	0.000	0.546	0.837
Auto_Avaliacao	0.4238	0.067	6.298	0.000	0.292	0.556

Marcados de azul, temos os coeficientes para montar a equação, ou seja, são os coeficientes β. Quando o coeficiente é positivo, nos diz que na medida que o preditor aumenta, a probabilidade do evento de interesse (classe positiva) ocorrer também aumenta. Quando o coeficiente é negativo, nos diz que na medida que o preditor aumenta, a probabilidade do evento oposto (classe negativa) ocorrer diminui. Lembrando que neste caso o nosso evento de interesse é prever a classe 'Boa'. Ao contrário da Regressão Linear, a interpretação não é direta, para saber o quanto o preditor impacta nas chances de o evento de interesse ocorrer, devemos aplicar a função exponencial ao coeficiente β.

Vamos interpretar os coeficientes:

β^1 = Prova_Logica = exp(0,2143) = 2,7182^0,2143 = **1,23** – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de lógica, o candidato aumenta em média 1,23 vezes as chances de pertencer a classe 'Boa'.

β^2 = Redacao = exp(0,6914) = 2,7182^0,6914= **1,99** – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de redação, o candidato aumenta em média, 1,99 vezes as chances de pertencer a classe 'Boa'.

β^3 = Auto_Avaliacao = exp(0,4238) = 2,7182^0,4238= **1,5278** – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na auto avaliação, o candidato aumenta em 1,52 vezes as chances de pertencer a classe 'Boa'.

A equação utilizando os coeficientes fornecidos fica:

$$\hat{y} = \frac{e^{-3,8458+(0,2143*Prova\_Logica)+(0,6914*Redacao)+(0,4238*Auto\_Avaliacao)}}{1+\ e^{-3,8458+(0,2143*Prova\_Logica)+(0,6914*Redacao)+(0,4238*Auto\_Avaliacao)}}$$

Supondo que o candidato tire 3 em Prova_Logica, 5 em Redacao e 1 em Auto_Avaliacao. A probabilidade dele(a) pertencer a classe de interesse 'Boa' fica:

$$\hat{y} = \frac{e^{-3,8458+(0,2143*3)+(0,6914*5)+(0,4238*1)}}{1+\ e^{-3,8458+(0,2143*3)+(0,6914*5)+(0,4238*1)}}$$

$$\hat{y} = \frac{1,9697}{2,9697}$$

$$\hat{y} = \ 0,6632\ (ou\ 66,32\%)$$

Portanto, um candidato com essa pontuação nas três variáveis teria 66,32% de probabilidade de pertencer à classe 'Boa'.

Avaliando a Performance Preditiva do Modelo

O output da Regressão Logística é uma probabilidade, então um ponto de corte (threshold) deve ser definido, de forma que, se a probabilidade estimada for acima do ponto de corte, a predição será considerada como pertencente ao evento positivo 1 (neste contexto, é a classe 'Boa'), caso contrário, a predição será considerada como pertencente ao evento negativo 0 (neste contexto, é a classe 'Ruim'). Se adotarmos o critério de que se a probabilidade for acima de 50%, consideraremos a classe positiva e caso seja abaixo, consideraremos a classe negativa. Teríamos a seguinte classificação para cada observação:

Figura 24 - Classe predita considerando 0,5 como ponte de corte

	Prova_Logica	Redacao	Auto_Avaliacao	Classe	Probabilidade	Classe_Predita
1	2	1	1	Ruim	0.09096250	Ruim
2	2	1	1	Ruim	0.09096250	Ruim
3	5	1	1	Ruim	0.15989589	Ruim
4	5	4	1	Boa	0.60235935	Boa
5	5	3	1	Ruim	0.43140265	Ruim
6	2	3	1	Ruim	0.28514808	Ruim
7	3	5	1	Boa	0.66331842	Boa
8	2	5	7	Boa	0.95288247	Boa
9	1	9	1	Boa	0.95325780	Boa
10	4	1	1	Ruim	0.13315862	Ruim
11	5	1	1	Ruim	0.15989589	Ruim
12	8	1	9	Boa	0.91487406	Boa
...						
697	6	1	1	Boa	0.19081999	Ruim
698	5	7	1	Boa	0.92341050	Boa
699	1	1	1	Ruim	0.07472674	Ruim

Para conferir os erros e acertos da Regressão Logística, podemos comparar a coluna Classe com a coluna Classe_Predita. A fim de fazer essa análise para muitas observações, iremos resumir em uma tabela, a **Matriz de Confusão.**

Figura 25 - Estrutura geral de uma Matriz de Confusão para um classificador binário

Classe Original		Classe Predita	
		Positiva	Negativa
	Positiva	A	B
	Negativa	C	D

A célula A trará a quantidade de observações preditas como pertencentes da classe positiva e que realmente eram da classe positiva. Ou seja, o quanto o algoritmo acertou para a classe positiva (verdadeiro positivo).
A célula B trará a quantidade de observações preditas como pertencentes da classe negativa, mas que originalmente pertencem a classe positiva. Ou seja, o quanto o algoritmo errou para a classe negativa (falso negativo).
A célula C trará a quantidade de observações preditas como pertencentes da classe positiva, mas que originalmente pertencem a classe negativa. Ou seja, o quanto o algoritmo errou para a classe positiva (falso positivo).
A célula D trará a quantidade de observações preditas como pertencentes da classe negativa e que realmente eram da classe negativa. Ou seja, o quanto o algoritmo acertou para a classe negativa (verdadeiro negativo).
Observe que os elementos da diagonal principal da matriz (células A e D) correspondem a quantidade de acertos, e os elementos fora da diagonal principal (células B e C) correspondem aos erros.
A Matriz de Confusão para o exemplo:

Figura 26 - Matriz de Confusão considerando 0,5 como ponto de corte

	Predito_Boa	Predito_Ruim
Boa	169	72
Ruim	23	435

A partir da Matriz de Confusão, iremos avaliar a taxa de acerto geral (Acurácia), a taxa de acerto para classe positiva (Sensitividade) e a taxa de acertos pra classe negativa (Especificidade).

Acurácia – Soma dos elementos da diagonal principal / Soma de todos os elementos.

$$Acurácia = \frac{169 + 435}{169 + 23 + 72 + 435} = 86{,}40\%$$

Sensitividade – A Sensitividade responde a seguinte pergunta: de todas as observações da classe 'Boa', quantas o algoritmo classificou como 'Boa'?

$$Sensitividade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$

$$Sensitividade = \frac{169}{169 + 72} = 70{,}12\%$$

Especificidade – A Especificidade responde a seguinte pergunta: de todas as observações da classe 'Ruim', quantas o algoritmo classificou como 'Ruim'?

$$Especificidade = \frac{Verdadeiros\ Negativos}{Verdadeiros\ Negativos + Falsos\ Positivos}$$

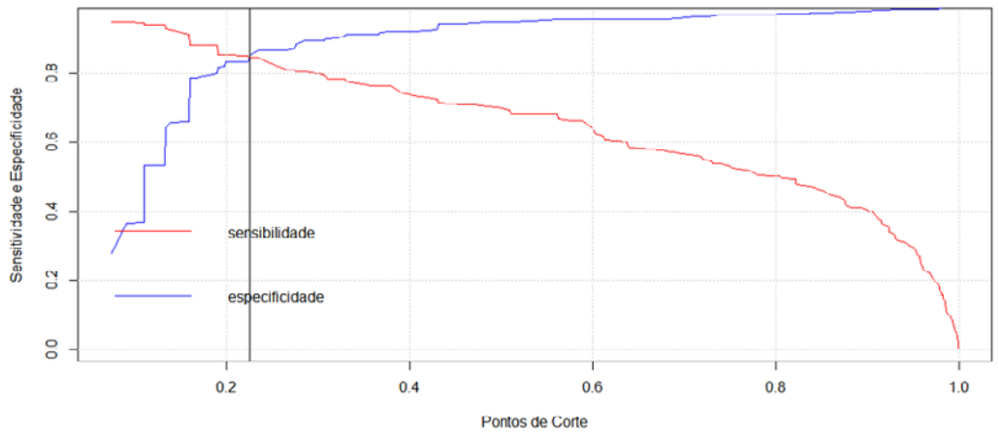
$$Especificidade = \frac{435}{435 + 23} = 94{,}97\%$$

Como podemos perceber, o algoritmo teve uma performance preditiva melhor para classe negativa. Para tentar resolver este problema, podemos tentar variar o ponto de corte. Veremos como achar o ponto de corte ideal para equilibrar a Sensitividade e a Especificidade no próximo tópico.

Análise de Sensibilidade e Especificidade

A ideia da Análise de Sensibilidade e Especificidade é simular várias matrizes de confusão, através de vários pontos de corte diferentes e identificar aquela matriz de confusão que nos dará tanto a maior Sensibilidade, quanto a maior Especificidade.

Figura 27 - Sensibilidade e Especificidade para diversos pontos de corte



Na intersecção das duas curvas, é o ponto de corte que nos dará a maior Sensitividade em equilíbrio com a maior Especificidade. Neste caso, o ponto de corte é 0,225, ou seja, caso a probabilidade seja acima de 22,5%, classificaremos o indivíduo como pertencente à classe positiva ('Boa'), caso contrário, classe negativa ('Ruim').
A matriz de confusão, considerando o ponto de corte de 22,5%, fica:

Figura 28 - Matriz de Confusão considerando 22,5% como ponto de corte

	Predito_Boa	Predito_Ruim
Boa	204	37
Ruim	75	383

A Sensibilidade e a Especificidade ficaram:

$$Sensitividade = \frac{204}{204 + 37} = 84{,}64\%$$

$$Especificidade = \frac{383}{383 + 75} = 83{,}62\%$$

Observe que comparando com o ponto de corte de 50%, que tivemos a Sensitividade de 70,12% e a Especificidade de 94,97%, o ponto de corte de 22,5% nos forneceu uma Sensitividade de 86,64% e a Especificidade de 83,62%. Foi necessário reduzir um pouco a Especificidade para ganhar na Sensitividade.

A análise de Sensitividade e Especificidade nos fornece o ponto de corte de forma matemática para equilibrar a Sensitividade e a Especificidade, mas na prática não devemos desprezar o contexto ao definir o ponto de corte, é sempre interessante analisar, o que é mais custoso: um falso positivo ou um falso negativo? E dessa forma o pesquisador pode ir variando o ponto de corte, observando que, sempre que a Sensitividade aumentar, a Especificidade cairá, e vice-versa.

Pílulas de conhecimento

Para uma visão complementar sobre regressão logística, vide https://ideiasesquecidas.com/2023/09/07/regressao-logistica-domine-esta-tecnica-poderosa

Estatística Computacional – Regressão Logística

Vide link a seguir para acompanhar os códigos para regressão logística com Python.
Link do Github: https://github.com/asgunzi/Estatistica_Analise_Dados

Faculdade XPe

xpeducacao.com.br