PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Supervised Machine Learning: Modelos Logísticos Binários e Multinomiais I
Data: 24/08/2021

Maria Conceição De Andrade
Prof, apenas por curiosidade, sobre seu comentário à respeito de grandes investimentos de empresas na compra de BD. Como são definidas quais Variáveis que irão/poderão aumentar a qualidade do modelo?
Maria, o procedimento stepwise é o procedimento utilizado para seleção de variáveis que aumentem o poder preditivo dos modelos logísticos.

Daniele de Barros Crespo
O coeficiente de Gini tem relação com o Índice de Gini (instrumento para medir o grau de concentração de renda em determinado grupo)?
Daniele, aqui estamos falando de uma estatística calculada com base no coeficiente de ROC que analisa a capacidade preditiva/classificatória de um modelo de classificação.

Tiago Borges Alves
com o cutoff eu defino true ou false em uma observação, mas com o resultado da ROC, como eu defino o mesmo em cada observação?
Tiago, a curva ROC é construída com base na sensitividade e especificidade em função de diversos valores de *cutoff.* Ou seja, é um gráfico que pode auxiliar em sua análise visual dos diferentes ajustes para diferentes pontos de corte. O *cutoff* pode ser definido utilizando a literatura, os modelos anteriores, a experiência de mercado, e com base no problema de pesquisa, considerando o que seria mais importante para cada caso.

Ariston Farias
Boa noite, eu consigo escolher um cutoff com base na máxima LL, ou na máxima ROC?
Ariston, a curva ROC é construída com base na sensitividade e especificidade em função de diversos valores de *cutoff.* Ou seja, é um gráfico que pode auxiliar em sua análise visual dos diferentes ajustes para diferentes pontos de corte. O *cutoff* pode ser definido utilizando a literatura, os modelos anteriores, a experiência de mercado, e com base no problema de pesquisa, considerando o que seria mais importante para cada caso.

Priscila Schall
prof, perdi de onde veio o resultado de área de 0,799. pode repetir, por favor?
Priscila, está no plot da curva ROC referente ao "EXEMPLO 01 - CONSTRUÇÃO DA CURVA ROC", disponível no script visto em aula.

Laila Monte Neto Donni
Como podemos saber qual variável tem maior importância para o modelo?
Laila, é necessário verificar a significância estatística de cada parâmetro. No caso da Regressão Logística é utilizada a estatística z.

Luiz Rodriguez Fontini 00537461981

Rafael Viegas De Carvalho Carlos Gomes
onde esta o valor da área embaixo da curva?
Rafael, está no plot da curva ROC referente ao "EXEMPLO 01 - CONSTRUÇÃO DA CURVA ROC", disponível no script visto em aula.

Carlos Rodrigo Costa
0,799 saiu onde, perdi essa parte
Carlos, essa é a área abaixo da curva ROC. Está no plot da curva ROC referente ao "EXEMPLO 01 - CONSTRUÇÃO DA CURVA ROC", disponível no script visto em aula.

William Henrique Stenico
Curva ROC só faz sentido para variáveis dependentes Qualitativas?
William a curva ROC é construída com base na sensitividade e especificidade em função de diversos valores de *cutoff*. Ou seja, é um gráfico que pode auxiliar em sua análise visual dos diferentes ajustes para diferentes pontos de corte e faz sentido somente para verificação de variáveis dependentes dicotômicas.

Vicente Souza Neto
Professor pode-se usar uma média para determinar um cut off ? como isso se relaciona ao ajuste dos dados em uma curva de frequência e a ROC
Vicente, o *cutoff* é definido com base em critérios relacionados ao seu problema de pesquisa. Valores maiores ou menores de *cutoff* geram maiores e menores valores de acurácia, especificidade e sensitividade, para cada problema de pesquisa podemos querer maximizar um destes indicadores em particular.

Paulo Henrique Real Leite
Por favor, poderia dar a definição de sensitividade e especificidade novamente? Obrigado.
Paulo, seguem as definições:
Sensitividade: diz respeito ao percentual de acerto, para um determinado *cutoff* quando consideramos apenas as observações que de fato são evento.
Especificidade: diz respeito ao percentual de acerto, para um determinado *cutoff* quando consideramos apenas as observações que de fato não são evento.

Henry Abellan Bovolon
pode colocar uma probabilidade para sim e outra para não?
Henry, é possível. A Regressão Logística é apropriada para variáveis dependentes do tipo dicotômicas. Sim/Não se enquadra nesse caso.

Ana Carolina Dos Santos Custódio
como que o professor calculou o PHAT desse modelo mesmo?
Ana Carolina, o PHAT foi calculado a partir da equação estimada. De posse dos parâmetros da equação é possível estimar a probabilidade associada a cada uma das observações.

Luis Rodriguez Fantini 005.374.619-81

Laila Monte Neto Donni
Como a gente sabe qual variável é mais ou menos significativa para explicar o y
Laila, é necessário verificar a significância estatística de cada parâmetro. No caso da Regressão Logística é utilizada a estatística z de Wald.

Raphael Fidelis Valadares
Professor, se bem entendi, a probabilidade de ocorrência do evento é inferida a partir dos dados (daí, suponho, o nome do algoritmo: logístico binário por MÁXIMA VEROSSIMILHANÇA)... exemplo, 60% mas a frequência de ocorrência no banco de dados não representa esta frequência real, mas, suponhamos, 64%, como o algoritmo faz o ajuste se, em nenhum momento, foi-lhe informada… a probabilidade correta - P(evento) - da Função densidade de probabilidade da distribuição de Bernoulli?
Raphael, a equação procura os parâmetros de chance que melhor explicam o Logito com a máxima verossimilhança. Para tanto, o algoritmo usa métodos de programação linear avançados.

Matheus Garcia
Professor, ao limpar dados, caso seja pedido o comportamento de uma variável quali binária, não transformada em 0 e 1, é correto fazer o procedimento (de transformação em 0 e 1)?
Matheus, nesse caso, tem que ver se essa transformação faz sentido. Toda transformação deve ser estudada com cuidado, pois depende muito do fenômeno estudado.

Rafael Viegas De Carvalho Carlos Gomes
No resultado do export_summs, o que são os valores em parenteses?
Rafael, são os erros padrões dos parâmetros. Quanto menor melhor.

Juliana Garcez de Oliveira
Para comparar modelos usando os Pseudos R^2 , eu escolho o que tiver maior ou menor pseudo R^2?
Juliana, o R quadrado representa uma espécie de ajuste ou qualidade do modelo proposto. Nesse caso é interessante que o modelo tenha um R quadrado maior. Lembre-se que trata-se de um indicador que apresenta algumas inconsistências, conforme explicado pelo professor Fávero em aula.

Jailson de Oliveira Arieira
Prof. Fávero, boa noite, já que os indicadores AIC, BIC, Pseudo R2 são mecanismos para comparar modelos, qual eu escolho usar quando os resultados foram contraditórios, e se isto pode ocorrer?
Pode ocorrer, Jaílson. Nesse caso, há a questão da experiência do analista em escolher um modelo ou tentar voltar a campo e pesquisar parâmetros melhores para montar um novo modelo.

Luiz Rodrigues Fantini 001.374.619-07

Matheus Garcia
Professor, há comparações de modelos em que o BIC é maior em um e o AIC é maior em outro? Se sim, como verificar o melhor modelo?
Pode ocorrer, Matheus. Nesse caso, entra a questão da experiência do analista em escolher um modelo ou tentar voltar a campo e pesquisar parâmetros melhores para montar um novo modelo.

Rodrigo Alves Pereira Gitirana
pode explicar mais sobre as ultimas informações do summary do modelo ? "Null deviance: 135.37 on 99 degrees of freedom Residual deviance: 100.93 on 97 degrees of freedom"
Rodrigo, sobre o tema encontrei na internet um debate explicando a questão que pode ser útil, a seguir: https://stats.stackexchange.com/questions/108995/interpreting-residual-and-null-deviance-in-glm-r

Maria Clara Barreiros Rodrigues
nao entendi de onde veio o -2 da formula do qui quadrado e aic
Maria Clara, é somente o modo como é construída a equação.

Rodrigo Alves Pereira Gitirana
pode explicar mais sobre o final do sumury do modelo ?
Rodrigo, o AIC é uma medida de comparação de modelos. Quanto menor esse parâmetro, melhor.

Gabriel Rodrigues Coutinho Pereira
Esse valor de -2, no cálculo do AIC e do teste chi², diz respeito ao valor de graus de liberdade? Como esses gl são calculados?
Gabriel, é somente o modo como é construída a equação.

Rodrigo Alves Pereira Gitirana
o que significa 3 graus de liberdade para o Log-Likelihood ?
Rodrigo, são os três parâmetros no modelo.

Marleide Ferreira Alves
Professor! Na fórmula do LL tem a exponencial de z, mas no excel o senhor não usou a exponencial, não entendi.
Marleide, atente-se que na função do excel é utilizado o Logaritmo Natural.

Rafael Viegas De Carvalho Carlos Gomes
Por que df=3? Não sao varias amostras?
Rafel, são os três parâmetros no modelo.

Luiz Rodriguez Fantini 005.374.619-01

Carlos Henrique de Oliveira

Professor, por que não posso utilizar na linha 68 glm(formula = atrasado ~ .? Onde seria o atrasado contra todos.

Carlos, essa sintaxe também é possível, no entanto precisa lembrar de retirar a coluna relativa ao nome dos alunos neste caso.

Danilo Steckelberg

Quando há pouca probabilidade de o evento acontecer (como fraude, por exemplo), a prob. do evento não fica próxima de 1, podemos usar modelo logit? Podemos multiplicar um fator p/ ajustar, i.e. 0,1*P

Danilo, quanto mais próximo do valor 1 há maior probabilidade de ocorrência de evento, enquanto quanto mais próximo de 0 menor a possibilidade da ocorrência do evento. Não vejo a necessidade de ajustar o modelo.

Cainã Max Couto Da Silva

Ainda sobre ponderação arbitrária, existe alguma forma diferente de lidar com variáveis ordinais? Ou sempre são trabalhadas como as nominais em dummies? A "dummização" das ordinais gera perda de info?

Olá Cainã. Sempre é bom utilizar dummy que não gera perda de informação, por si.

Isabella Montanhal de Araujo

Não ficou claro o que é PLOGIT e PPROBIT e suas diferenças?

Isabella, a logística possui caudas um pouco mais achatadas, isto é, a curva probit se aproxima dos eixos mais rapidamente que a curva logit. Os modelos logit e probit, no entanto, são apenas *modelos*. Ambos os modelos permitirão *detectar* a existência de um efeito de no resultado; exceto em alguns casos muito especiais, nenhum deles será "realmente verdadeiro", e sua *interpretação* deve ser feita com cautela e levando-se em consideração as características de cada problema de pesquisa e banco de dados. Caso haja maior preocupação com a parte final da curva, em algum momento a seleção do logit ou probit será importante. Não existe uma regra exata para selecionar probit ou logit. Você pode selecionar o modelo observando a probabilidade (ou a probabilidade do log) ou AIC.

Guilherme Piva Magalhaes Da Rocha

Na verdade, podemos fazer um OLS para variavel Y binaria, porem depois preciamos tranfsormao o resultado que é o logito em sigmoide

Acho que na sua informação estão sendo confundidos conceitos. Conforme Fávero (2017) o logito Z não representa a variável dependente, denominada por Y, e o nosso objetivo neste momento é definir a expressão da probabilidade P; de ocorrência do evento de interesse para cada observação, em função do logito Z; ou seja, em função dos parâmetros estimados para cada variável explicativa. A regressão logística binária define o logito Z como o logaritmo natural da chance. Portanto, o logito não representa a variável Y da OLS.

Luiz Rodrigues Homem 005374.610.81

Heloísa Hilário
Quando vamos ter alguma aula em alguma outra linguagem? Seria legal a gente ter uma visão de código diferente do R, pra ver se o conceito ficou realmente fixado independente da linguagem.
Olá Heloísa, esse curso é todo baseado em R. Eventualmente, algum professor pode utilizar outra linguagem.

Diego Barbosa Batista
Quando veremos KS1 e KS2?
Diego, seguem as definições:
O indicador KS1 foi desenvolvido para calcular a aderência, é um indicador que tem o objetivo de mensurar a performance do modelo, mas de indicar se a população em que foi feito o modelo se alterou ou não, ou seja, se o modelo está sendo aplicado no mercado para qual ele foi desenvolvido. Dito de outro modo, o KS1 é uma estatística também para verificar possíveis distorções no perfil da população em relação à base de referência.
Já o KS2 indica o cálculo da performance do modelo, ou seja, se a separação de bons e maus está coerente com o que esperávamos na fase de estimação do modelo.

André Araújo
Professor, como realizar a transformação de Box-Cox quando há valores negativos no Y? Como fazer a transformação de Yeo-Johnson?
Olá André. Não existe problema da utilização da transformação de Box-Cox quando há valores negativos. Sobre a transformação de Yeo-Johnson segue código utilizado no Rstudio, a seguir: https://search.r-project.org/CRAN/refmans/VGAM/html/yeo.johnson.html

Paulo Renato Leite
Profe, da onde veio o 0,799 da curva ROC?
Paulo, essa é a área abaixo da curva ROC. Está no plot da curva ROC referente ao "EXEMPLO 01 - CONSTRUÇÃO DA CURVA ROC" disponível no script visto em aula.

Alexandre Gonçalves da Rocha
Coeficiente de Gini é o mesmo usado para medir desigualdade de renda?
Alexandre, aqui estamos falando de um estatística calculada com base no coeficiente de ROC que analisa a capacidade preditiva de um modelo de classificação.

Renato Santos Luz
Professor, existe alguma área de mercado que pede uma acurácia extremamente alto?
Renato, há algumas áreas em que necessita-se de uma acurácia realmente maior. Imaginemos por exemplo o caso da área médica, em que o acerto ou o erro pode ser a vida de uma pessoa, ou ainda a área de aviação, em que a probabilidade de ocorrência de um desastre deve ser a mínima possível (tendente a zero, preferencialmente). Ninguém quer que o avião caia, nesse caso (e em vários outros) realmente necessita-se de uma acurácia extremamente elevada.

Luiz Rodrigues Fantini 005.374.119-81

Flávia Ruiz Leão

Em qual item do output aparece o cálculo da área da curva ROC?

Flávia, está no plot da curva ROC referente ao "EXEMPLO 01 - CONSTRUÇÃO DA CURVA ROC", disponível no script visto em aula.

Ronei Gomes de Almeida

Professor, o modelo foi treinado com toda a base (100 registros)? A predição não deveria ter sido realizada com novas observações?

Para qualquer modelo de predição não cabe extrapolação da predição para além da amplitude das observações.

Gustavo Murad

prof, nao eh viavel o modelo/R indicar o cutoff ideal, ja que tenho o modelo e o resultado Y real pra otimizar tal parâmetro?

Gustavo, essa definição acaba sendo um problema de pesquisa. Depende muito do que o pesquisador tem como objetivo, para cada *cutoff* estaremos definindo percentuais de acerto ou de erro em cada uma das possibilidades (evento ou não evento). E aí cabe a decisão de pesquisa: quero acertar mais os eventos ou os não eventos? Aumentar/dimunuir a especificidade ou a sensibilidade? Cada caso tem suas peculiaridades.

Alexandro Correa Gonçalves Afonso

Professor, a dúvida da interpretação dos betas ficou pela metade. Pelo exemplo do dataset, poderíamos dizer que a magnitude do beta é análogo à sensibilidade que a variação de uma unidade da

Alexandre, os betas representam a chance da ocorrência de um evento a partir da variação de uma unidade de cada parâmetro da equação.

Samya de Lara Lins de Araujo Pinheiro

No caso de bernoulli, faz algum sentido falar em odds ratio?

Com certeza, Samya. A regressão logística relaciona-se com uma distribuição de bernoulli. Conforme Fávero (2017) em modelos de regressão logística multinornial, a chance (odds ratio) também é chamada de razão de risco relativo (relative risk ratio).

Damião Flávio dos Santos

Boa noite! Meus dados advém de um censo e com base na última aula, pelo que entendi, não é necessário teste para os parâmetros. Como vou identificar as variáveis que significativas no meu modelo? OBG

Olá, Damião. Nesse caso, sugiro que rode o modelo de regressão e veja como as variáveis se comportam em relação ao fenômeno estudado, sem atentar tanto para a significância dos parâmetros.

Vanessa Hoffmann de Quadros

Professor, todos esses indicadores para comparação apontam para o mesmo modelo quando comparamos entre modelos?

Vanessa, nesse caso você deve escolher o modelo que tem melhores indicadores.

Luiz Rodrigues Fonini 005.354.619-81

Samya de Lara Lins de Araujo Pinheiro

A função step usa AIC para escolha do modelo para regressão logística? No OLS, a escolha também pelo AIC?

Samya, se entendi a pergunta, a função step busca retirar os parâmetros não significativos até conseguir ajustar o melhor modelo. O AIC acaba sendo um resultado do logaritmo.

Vanessa Hoffmann de Quadros

Professor, é possível que haja mais de um ponto de máximo na função de verossimilhança?

Olá Vanessa, os parâmetros do modelo buscam maximizar a função de verossimilhança, portanto só há um ponto máximo dentro do intervalo de dados.

Lucas Alves Dias Cardoso

(parte 1) Marcelo Sabadini e Professor, entendi a resposta, mas conceitualmente faz diferença se o que chama de Logito é um ou outro. Quando o professor chamou Logito de a+b1.x1+b2.x2... ele conservou esse significado quando manteve esta nomenclatura nas operações que fez na equação. Se o logito for o nome dado à expressão com ln(chances), não seria correto fazer isso, no meu entendimento. Complementando: digo "por coincidência" pois o logito fica, de fato, igual a "Z", mas a nomenclatura se dá, na realidade, para o outro lado da equação: a expressão com ln.

Na equação Z = alfa + beta1*x1 + beta2*x2 … betan*xn, o Z representa o logito. Conforme Fávero (2017) o logito Z não representa a variável dependente, denominada por Y, e o nosso objetivo neste momento é definir a expressão da probabilidade P; de ocorrência do evento de interesse para cada observação, em função do logito Z; ou seja, em função dos parâmetros estimados para cada variável explicativa. A regressão logística binária define o logito Z como o logaritmo natural da chance.

Luiz Rodriguez Fantini 005.374.119-81