FIOCRUZ

Concurso Público Fiocruz 2023

Pesquisador em Saúde Pública

Prova Discursiva

PE 62

Cientista de Dados

Espelho de Resposta

Pontuação de cada Questão Discursiva conforme Anexo II do Edital nº 3, de acordo com a Unidade detentora da vaga.

Espera-se que o candidato, no desenvolvimento do tema, tenha feito considerações técnicas adequadas sobre os seguintes pontos:

Questão 01

1) A divisão da base em treino e teste permite medir a qualidade do modelo ajustado, isto é, a capacidade de previsão, tanto para a base de treino quanto para a base de teste. A base de treino é aquela que contém as observações utilizadas no ajuste do modelo, já a base de teste é composta por novas observações, não utilizadas no ajuste. Esta prática permite que o pesquisador identifique a existência de sobreajuste e identifique variáveis independentes relevantes para o problema.

   A identificação do sobreajuste é feita a partir da comparação da medida de qualidade do ajuste na base de treino e teste. Caso o valor desta medida seja parecido nas duas bases, conclui-se que o modelo captou a informação nos dados e foi capaz de realizar boas previsões em novos dados. Caso a medida de qualidade do ajuste na base de treino seja significativamente melhor do que na base de teste, isto indica a existência de sobreajuste.

   Já a identificação de variáveis relevantes pode ser feita comparando a medida de qualidade do ajuste, tanto na base de treino quanto na base de teste, do modelo completo (com todas as variáveis) com o modelo completo a menos da variável em questão. Se as medidas de qualidade do ajuste pouco alterarem com a exclusão da variável em questão, isto indica que tal variável não agrega muita informação ao modelo. Por outro lado, se as medidas de qualidade do ajuste pioraram significativamente com a exclusão de uma variável, isto indica que esta variável agrega muita informação ao modelo.

2) Ao comparar os Modelos 1 e 2 observamos que a retirada da variável “ganho de peso gestacional” trouxe uma diminuição significativa no valor do R2, que indica piora no ajuste. Isso nos faz concluir que esta variável, “ganho de peso gestacional” contribui significativamente na qualidade da previsão e ela deve ser mantida no modelo.

Ao comparar os Modelos 1 e 3 observamos que a retirada da variável “sexo do bebê” trouxe pouca variação nos valores do R2. Isso nos faz concluir que esta variável, “sexo do bebê” contribui pouco na qualidade da previsão e pode ser desconsiderada.

As mesmas conclusões ocorrem quando comparamos os modelos 4 e 5, a variável “ganho de peso gestacional” deve ser mantida, e quando comparamos os modelos 4 e 6, a variável “sexo do bebê” pode ser desconsiderada.

3) Ao comparar o par de modelos que se diferem apenas pelo número de árvores percebemos, em todos os pares apresentados, que o modelo com 500 árvores apresenta maiores (melhores) valores de R2 na base de treino quando comparados com o modelo equivalente com 100 árvores. Porém, os valores do R2 na base de teste não crescem como os da base de treino quando mudamos do modelo com 100 para o com 500 árvores. Isso indica que aumentar o número de árvores não melhora as previsões, ou seja, as 100 árvores já captam toda a informação possível e aumentar o número de árvores só gera sobreajuste.

4) A partir dos argumentos apresentados nos itens anteriores foi observado que a variável “sexo do bebê” não agrega valores significativos ao ajuste e que o modelo com 100 árvores já capta a mesma informação captada pelo modelo de 500 árvores. Sendo assim, eu escolheria o Modelo 3, XGBoost com 100 árvores e todas as variáveis menos “sexo do bebê”.

Questão 02

1) O modelo de regressão logística é apropriado para analisar variáveis dependentes binárias, como é o caso do câncer de mama (variável resposta dicotômica: sim ou não). Em contrapartida, o modelo de regressão normal é adequado para variáveis dependentes quantitativas. Dessa forma, ao investigar fatores associados ao câncer de mama, optamos pelo modelo de regressão logística. Esse modelo estima o efeito das variáveis independentes na probabilidade de sucesso, neste contexto, na probabilidade de desenvolver câncer de mama. Se o efeito de uma variável for positivo, ela é considerada um fator de risco para o câncer de mama. Por outro lado, se o efeito for negativo, a variável é considerada um fator protetor. Uma das grandes vantagens desse modelo é a capacidade de obter a razão de chances (*odds ratio*), uma medida de fácil interpretação. A razão de chances indica a chance de o desfecho ocorrer no grupo exposto ao fator em estudo. Isso significa que, se a razão de chances for maior que 1, há um aumento nas chances do desfecho, enquanto um valor menor que 1 sugere uma diminuição nas chances. Essa interpretação simplificada facilita a compreensão dos resultados, tornando o modelo de regressão logística uma ferramenta valiosa na investigação de associações em estudos na área da saúde.

2) Inicialmente vamos obter a razão de chances para a categoria 40 a 69 anos da variável idade:

$$RC = \exp(1{,}975) = 7{,}2.$$

Sendo assim, uma mulher com idade entre 40 e 69 anos possui 7,2 vezes mais chances de ter câncer de mama do que uma mulher com 40 anos ou menos, considerando que as mesmas possuem as demais variáveis independentes iguais.

A razão de chances para insulina no sangue é dada por:

$$RC = \exp(0{,}084) \cong 1{,}09.$$

Sendo assim, o aumento de uma unidade de insulina no sangue de uma mulher aumenta em 9% a chance de câncer de mama, considerando as demais variáveis independentes iguais.

3) Com base em um nível de significância de 5%, as variáveis idade, insulina e resistina no sangue são fatores importantes para a predição de câncer de mama, pois idade possui uma categoria com p-valor inferior a 0,05 e as variáveis quantitativas, insulina e resistina no sangue, possuem p-valor inferior ao nível de significância adotado.

4) Segundo os critérios apresentados, o melhor modelo é aquele que possui o menor Critério de Akaike e a maior área sob a curva ROC. Sendo assim, o Modelo 2 é o mais indicado, pois $AIC(M_1) = 119,26 > AIC(M_2) = 106,15$ e $AUC(M_1) = 0,78 < AUC(M_2) = 0,85$.