Aplicação de modelos GAMs na modelagem da pontuação geral em ciências

Universidade Federal da Paraíba - UFPB

Abril 2024

Sumário

Introdução

Na jornada de uma criança, a educação desempenha um papel crucial, moldando a mente dos jovens e preparando indivíduos para enfrentar os desafios do mundo. Dentro deste vasto campo educacional, as ciências assumem um papel central e fundamental. Através da educação científica, as crianças não apenas adquirem conhecimento sobre os princípios básicos do universo, mas também desenvolvem habilidades críticas, como pensamento análitico, resolução de problemas. Além disso, ao integrar as ciências na educação das crianças, estamos preparando o terreno para um futuro sustentável e inovador.

Neste contexto, o presente trabalho tem como objetivo utilizar técnicas de Modelo Aditivo Generalizado, fazendo uso de um banco de dados sobre educação. Utilizando esse banco de dados, foi modelada a nota de disciplinas de ciências de crianças de 15 anos. As outras variáveis que estão no banco de dados é a **explicação**, a pontuação ao se explicar o fenômeno científico, **interesse** em ciência, **apoio** a pesquisa científica, **renda**, índice de **educação** e o **IDH**.

Metodologia

A base de dados

O banco de dados de dados utilizado contém informações de inadimplência de clientes de cartão de crédito. Na base de dados temos 4 variáveis, das quais 2 são nominais e 2 são numéricas. Dessa forma, temos as seguintes variáveis:

- **Inadimplência** : Esta variável representa se o cliente entrou em inadimplência no cartão de crédito ou não. É uma das variáveis nominais com dois níveis: "Não", indicando que o cliente não entrou em inadimplência, e "Sim", indicando que o cliente entrou em inadimplência. Esta será a variável dependente para a modelagem com regressão logística, em que o objetivo será classificar se um cliente entrará em inadimplência em sua dívida no cartão de crédito com base em outras características no conjunto de dados.
- **Estudante**: A variável *Estudante* é a segunda variável nominal e ela indica se o cliente é ou não estudante. Ela possui dois níveis: "Não", indicando que o cliente não é estudante, e "Sim", indicando que é estudante. Esta variável independente pode ser importante para discriminar o comportamento de um cliente que é estudante para um que não é estudante, pois isso poderia significar diferentes perfis de risco.
- **Saldo**: O Saldo é uma das variáveis contínua numérica representa o saldo médio que o cliente tem remanescente em seu cartão de crédito após efetuar o pagamento mensal. O saldo reflete o valor da dívida no cartão de crédito que o cliente carrega em média. Saldos mais altos podem indicar níveis mais elevados de dívida e potencialmente maior risco de inadimplência.
- **Renda**: Por último, temos a renda do cliente, que é a última variável numérica do banco de dados. A renda é um fator importante na avaliação de solvência, já que indivíduos com rendas mais altas podem ser mais propensos a pagar dívidas pontualmente. Maiores rendas podem indicar um menor risco de inadimplência, embora isso possa variar dependendo de outros fatores.

Além disso, este banco de dados contém 10.000 observações. 20% foram reservados para teste, enquanto os outros 80% foram usados para ajustar um modelo de regressão logística.

Recursos computacionais

Para realizar a modelagem, a exploração dos dados e todas as outras análises que estão presente nesse trabalho, foi utilizada a linguagem de programação R. Como produto da linguagem R, foram utilizados pacotes para modelagem estatística e criação de gráficos. Para a modelagem estatística foi utilizado o framework *tidymodels* e para a visualização de gráficos, foi utilizado o *ggplot2*. Além disso, foi também utilizado o Quarto, que serve para fazer apresentações e documentos de escrita, o que é o caso desse documento.

Análise exploratória

Um dos primeiros passos para qualquer estudo estatístico, é fazer a análise exploratória dos dados. Dessa forma, esse foi o primeiro passo tomado nesse projeto, utilizando medidas de tendência central e de dispersão, como média, mediana, desvio-padrão e coeficiente de correlação, além da utilização de gráficos e tabelas com o objetivo de caracterizar e compreender melhor os eventos em questão. Além disso, foram elaborados gráficos para observar o comportamento da variável resposta e das variáveis preditoras.

Construção do modelo e métricas para sua validação

Uma das etapas mais importantes para a validação de um modelo, é a escolha de métricas para analisar a sua performance e formas de verificar se as variáveis que estão sendo utilizadas para a modelagem são estatisticamente significantes. Nesse caso, para a análise da regressão logística, foram escolhidas métricas que são geradas a partir de uma matriz de confusão e a escolha das variáveis foi feita pelo teste de significância dos coeficientes da regressão. Por fim, foi analisado como o modelo de regressão estava performando durante a classificação.

Sensitividade

A sensitividade, também chamado de frequência de verdadeiros positivos, é calculado como a quantidade verdadeiros positivos dividido pela quantidade total de positivos. Dessa forma, temos que:

$$Sensitividade = \frac{TP}{TP + FN}$$

em que TP é a quantidade de verdadeiros positivos e FN é a quantidade de falsos negativos. Ainda, a melhor sensitividade seria 1, enquanto a pior seria 0.

Especificidade

Frequência de verdadeiros negativos ou especificidade, é calculado como a quantidade de verdadeiros negativos dividido pelo total de negativos. Portanto, temos que:

$$Especificidade = \frac{TN}{TN + FP}$$

o TN é a quantidade de verdadeiros negativos e FP é a quantidade de falsos positivos. Temos também que a melhor sensitividade para o modelo de regressão logística se apróxima de 1, enquanto a pior se aproxima de 0.

Acurácia

A acurácia, que também foi utilizada para analisar a performance do modelo, é definida como a quantidade de todas as corretas classificações dividido pela total do banco de dados. Assim, temos que:

$$Acurcia = \frac{TP + TN}{TP + TN + FN + FP}$$

É importante saber também que quanto mais próximo de 1 está a acurácia, mais o modelo está classificando bem.

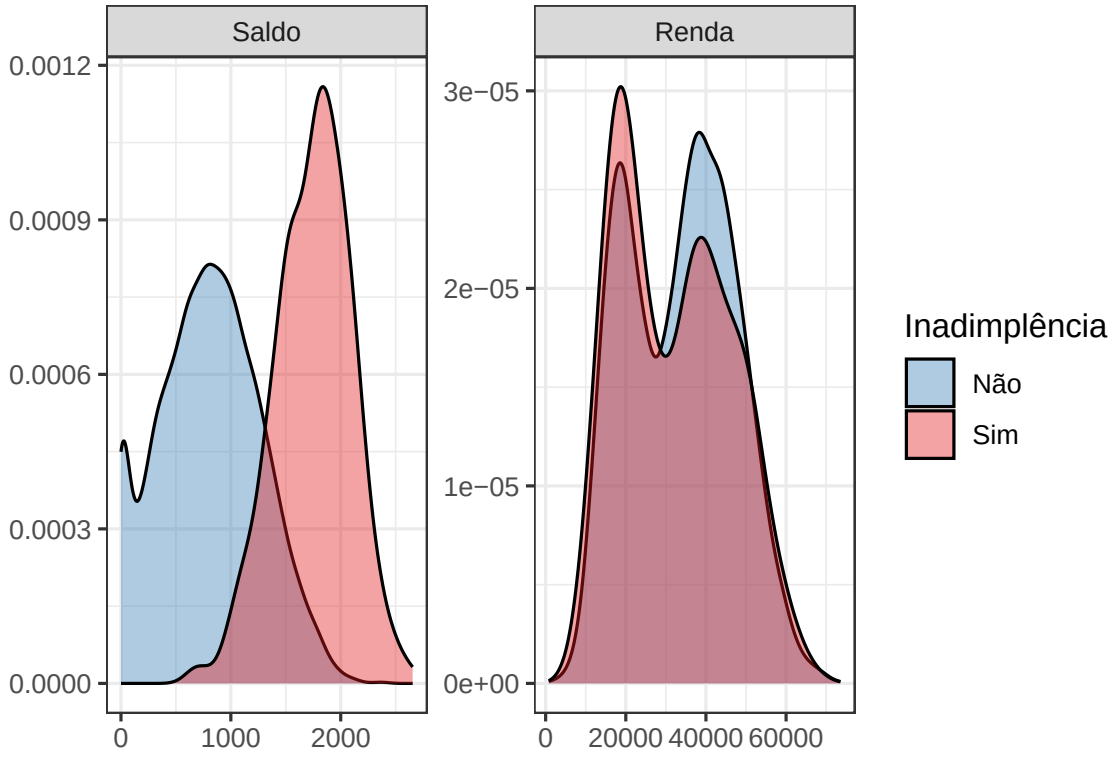
Área Sob a Curva ROC (AUC)

AUC significa Área Sob a Curva ROC. É uma métrica bastante conhecida para avaliar o desempenho de um modelo de classificação binária. A curva ROC plota a frequência de verdadeiros positivos **(sensibilidade)** contra a taxa de falsos positivos **(1 - especificidade)** para diferentes valores de limiar. A métrica de AUC quantifica o poder discriminativo geral do modelo em todos os valores de limiar possíveis. A interpretação do AUC é feita de forma bastante simples. Um AUC igual a 1 indica um classificador perfeito que separa perfeitamente as instâncias positivas e negativas. A curva ROC alcança o canto superior esquerdo, significando alta sensibilidade e especificidade.

Resultados

Exploração dos dados

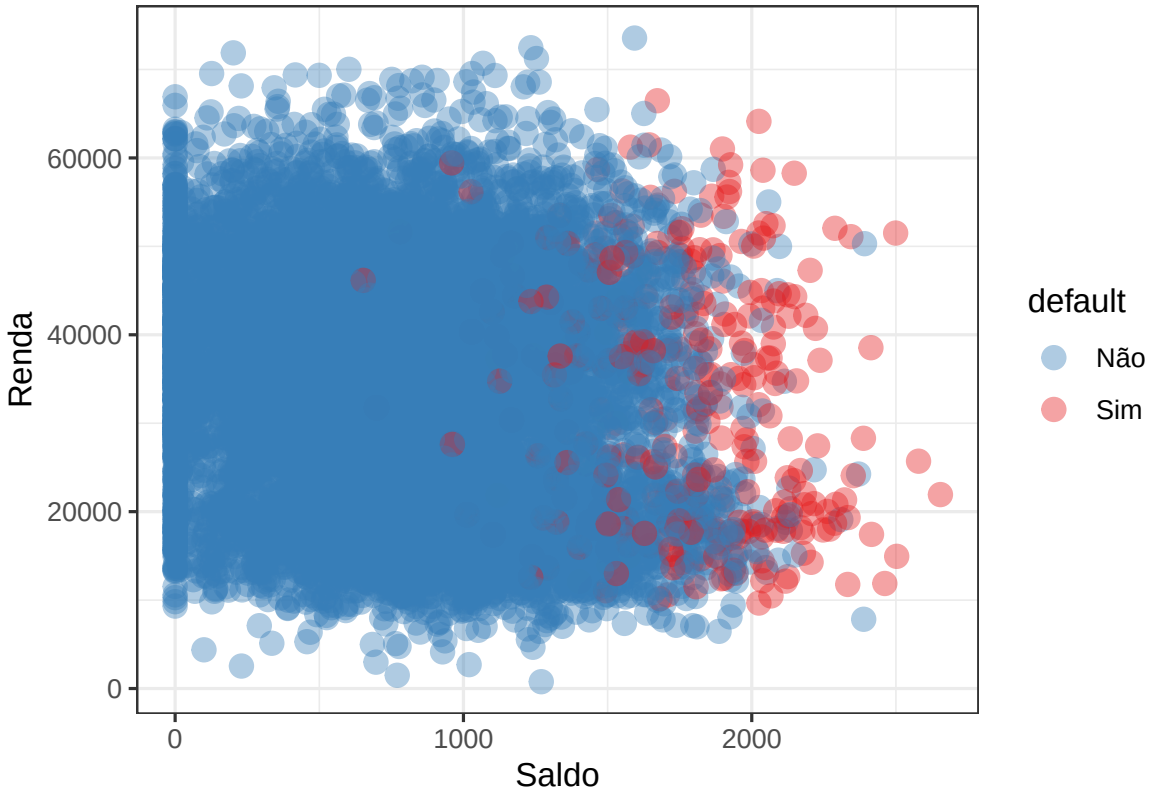
Figura 1: Distribuição da Renda e Saldo do cliente



Pelo gráfico acima podemos perceber que os clientes que estão ou não em situação de inadimplência tem uma distribuição de renda bastante parecida, tendo bastante interseção entre os dois níveis. No entanto, os clientes que não estão em situação de inadimplência, tem uma frequência de renda maior, o que faz sentido, já que esses clientes tem mais condições de arcar com as suas dívidas. Já na distribuição do saldo dos clientes, podemos ver que os clientes com saldos menores tendem a não não estar inadimplentes, enquanto os que estão em

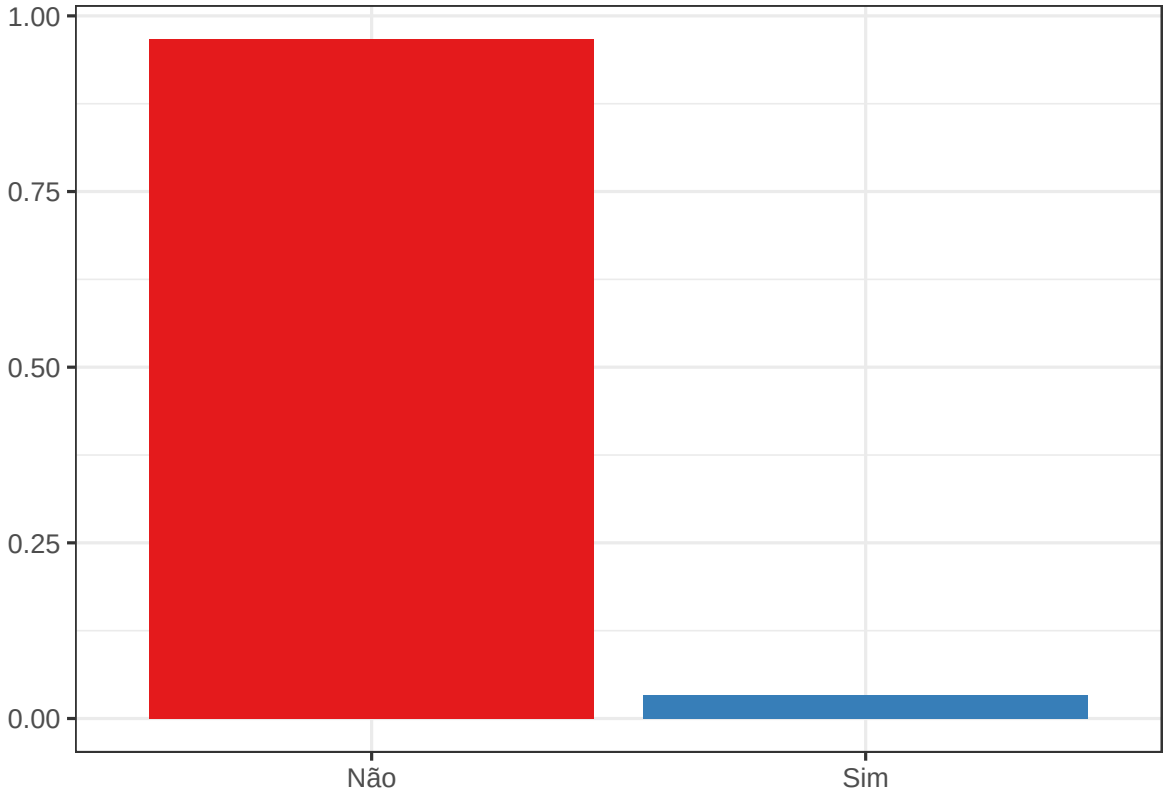
situação de inadimplência tem saldos maiores. Isto faz sentido, pois saldos mais altos podem indicar níveis mais elevados de dívida.

Figura 2: Renda em função do saldo do cliente



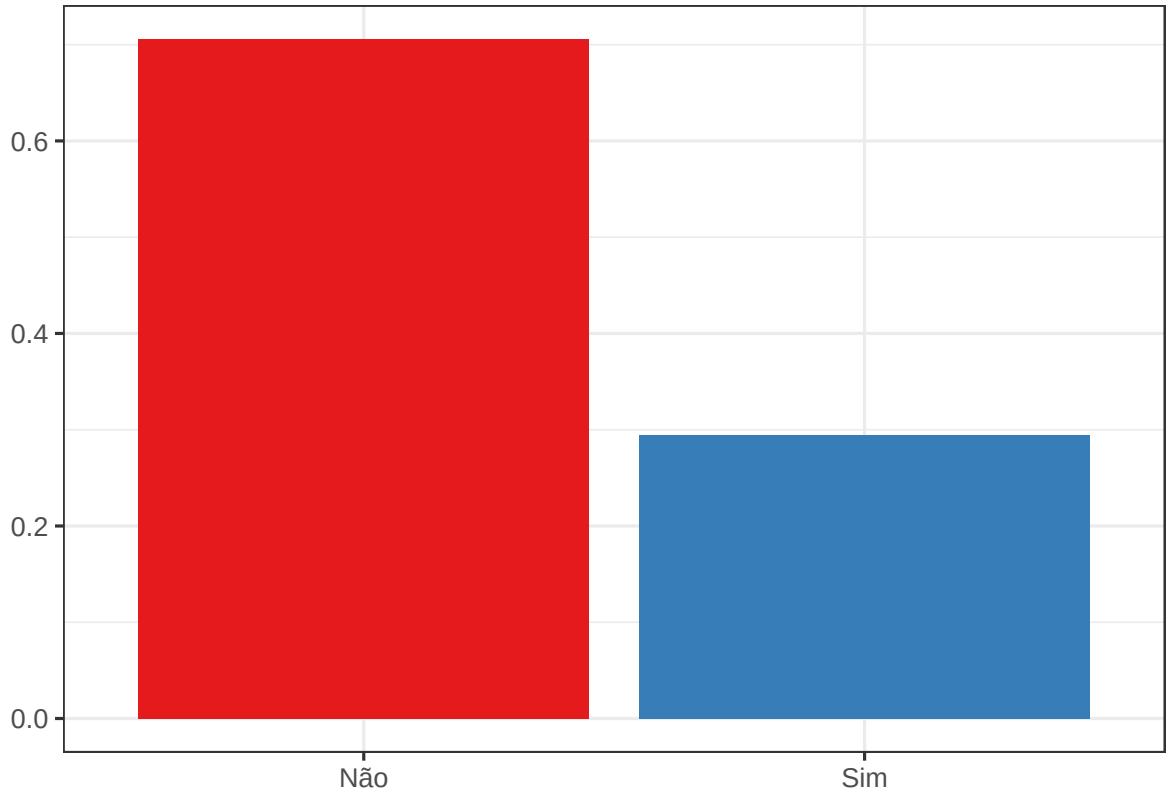
Assim como foi observado anteriormente no gráfico de densidade, a renda dos clientes não parecem indicar diferenças em situação de inadimplência. No entando, podemos observar novamente analizando apenas o eixo das abscissas, que saldos maiores parecem caracterizar aqueles clientes que costumam estar mais endividados.

Figura 3: Frequência dos clientes em inadimplência



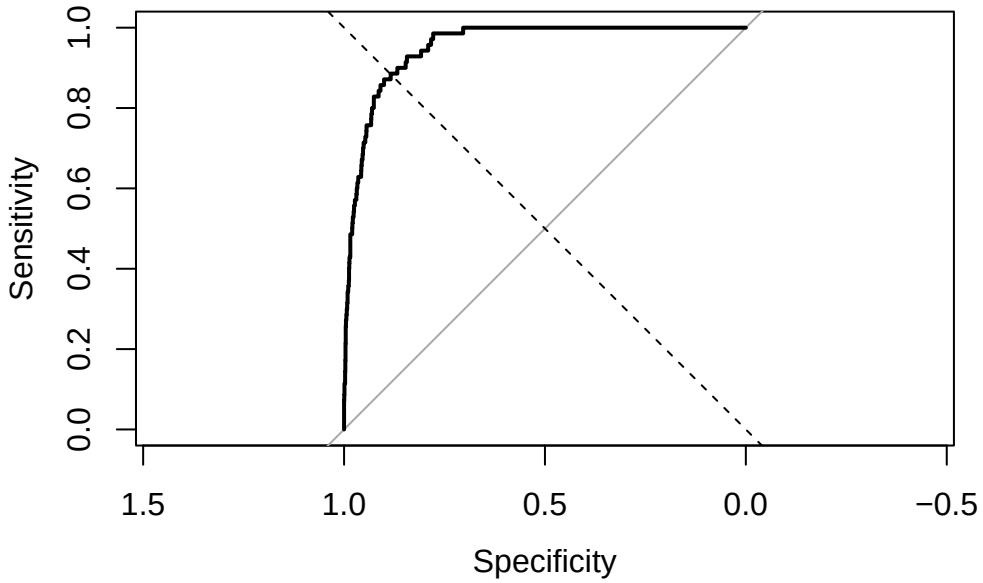
O gráfico acima nos entrega a informação da porcentagem de clientes em situação de inadimplência. Dessa forma, podemos ver que 96,67% dos clientes não estão em situação de inadimplência. Isso pode significar que o modelo de regressão linear talvez fique melhor para classificar aqueles clientes que não estão em situação de inadimplência, pois a classe que está sendo modelada tem níveis desbalanceados.

Figura 4: Frequência dos clientes que são ou não estudantes



Agora observando a porcentagem porcentagem dos clientes que são estudantes, vemos que o banco de dados é composto por maioria de clientes não estudantes, com 70,56% não sendo estudantes.

Construção do modelo



Envelope

```
                    5 %          95 %
(Intercept) -11.71182787 -10.309683897
studentYes   -0.91696052  -0.375444086
balance       0.00544956   0.006331535
```

Antes

	Coeficiente	Erro Padrão	Estatística	$Pr\,(>\|\|z\|\|)$
(Intercept)	-10.8854	0.5468227	-19.9067650	$3.555x10^{-88}$
balance	0.0056	0.0002555	22.2072284	$2.924x10^{-109}$
income	$6.6x10^{6}$	0.0000091	0.7263696	0.4676
student_Yes	- 0.6545	0.2652104	-2.4680278	0.0135

Depois

	Coeficiente	Erro Padrão	Estatística	$Pr\,(>\|\|z\|\|)$
(Intercept)	-10.6228	0.4051	-26.2187	$1.6259x10^{-151}$
balance	0.0056	0.0002	22.2254	$1.9479x10^{-109}$
student_Yes	-0.8041	0.1660	-4.8418	$1.2864x10^{-6}$

	Resultado Final
$Desvio/\phi$	1271.217
χ^2	1355.276

Conclusões

Anexos