UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

LUAN FRANCISCO DA SILVA

INTERPRETAÇÃO DE UM MODELO XGBOOST PARA PREVISÃO DE UTI POR VALOR DE SHAP E COMPARAÇÃO COM OS PARÂMETROS DO MODELO DE REGRESSÃO LOGÍSTICA

CAMPINA GRANDE - PB
2023

LUAN FRANCISCO DA SILVA

INTERPRETAÇÃO DE UM MODELO XGBOOST PARA PREVISÃO DE UTI POR VALOR DE SHAP E COMPARAÇÃO COM OS PARÂMETROS DO MODELO DE REGRESSÃO LOGÍSTICA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira.

CAMPINA GRANDE - PB
2023

S586i Silva, Luan Francisco da.

Interpretação de um modelo XGBOOST para previsão de UTI por valor de SHAP e comparação com os parâmetros do modelo de regressão logística [manuscrito] / Luan Francisco da Silva. - 2023.

44 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Coordenação do Curso de Estatística - CCT. "

1. Valor de SHAP. 2. MLG e SHAP. 3. Interpretação de aprendizado de máquina. I. Título

21. ed. CDD 519

Elaborada por Geovani S. de Oliveira - CRB - 15/1009

Biblioteca Central BC/UEPB

LUAN FRANCISCO DA SILVA

INTERPRETAÇÃO DE UM MODELO XGBOOST PARA PREVISÃO DE UTI POR VALOR DE SHAP E COMPARAÇÃO COM OS PARÂMETROS DO MODELO DE REGRESSÃO LOGÍSTICA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística.

Aprovado em: 07/12/2023.

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira
(Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Silvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba (UEPB)

Profa. Débora de Sousa Cordeiro
Universidade Estadual da Paraíba (UEPB)

AGRADECIMENTOS

Agradeço a minha família pelo suporte nesses longos anos de graduação, em especial minha mãe, Dona Rosa e minha irmã, Laís. Agradeço também aos professores que tanto se dedicaram a me ajudar.

“Ao verme que primeiro roeu as frias carnes do meu cadáver dedico como saudosa lembrança estas memórias póstumas”

- Machado de Assis

RESUMO

Os MLG são soluções para modelos lineares quando se tem impasses quanto as pressuposições necessárias para tal, como linearidade dos parâmetros e normalidade nos resíduos, ou ainda, problemas caracterizados no conjunto de dados, como valores de contagem, respostas binárias ou excesso de zeros, que corroboram para baixa qualidade no ajuste e, consequentemente, na interpretabilidade dos coeficientes. Esta teoria apresenta um leque de soluções sem perda da capacidade inferencial de um modelo linear. O uso da teoria de SHAP, nos permite obter informações sobre a importância de cada recurso (variável) no modelo de aprendizado de máquina, sob aspectos de comparação entre elas e de ranqueamento das mesmas. O principal objetivo deste trabalho foi fazer comparações entre esses dois tipos de interpretação, respeitando o fato de que são diferentes tipos de abordagem. Foi mostrado que a contribuição para o preditor linear e a contribuição dos valores de SHAP à probabilidade logarítmica, podem ser parecidas se considerarmos um contexto de efeito geral numa análise descritiva e de interpretação, visto que, sob certos aspectos, há variáveis ou recursos que possuem a mesma importância em ambos os modelos, mas que diferem em termos de direção, pois podem impactar tanto positivamente, quanto negativamente. As variáveis FADIGA, CARDIOPATI e HOSPITAL, são exemplos deste caso, já que no efeito geral, são parecidas em ambos os modelos. Apesar disso, a desproporcionalidade no impacto da variável HOSPITAL influencia muito na precisão do modelo de aprendizado de maquina, tornando-o um modelo com baixa capacidade de predizer a internação na UTI quando isto seria o correto.

Palavras-chaves: valor de SHAP; MLG e SHAP; interpretação de aprendizado de máquina.

ABSTRACT

MLG are solutions for linear models when there are impasses regarding the necessary assumptions for this, such as linearity of parameters and normality in residuals, or problems characterized in the data set, such as count values, binary responses or excess zeros, which corroborate the low quality of the adjustment and, consequently, the interpretability of the coefficients. This theory presents a range of solutions without losing the inferential capacity of a linear model. The use of SHAP theory allows us to obtain information about the importance of each resource (variable) in the machine learning model, in terms of comparison between them and their ranking. The main objective of this work was to make comparisons between these two types of interpretation, respecting the fact that they are different types of approach. It was shown that the contribution to the linear predictor and the contribution of SHAP values to the logarithmic probability can be similar if we consider a context of general effect in a descriptive and interpretative analysis, since, under certain aspects, there are variables or resources that have the same importance in both models, but which differ in terms of direction, as they can impact both positively and negatively. The variables FADIGA, CARDIOPATI and HOSPITAL are examples of this case, since in general effect, they are similar in both models. Despite this, the disproportionality in the impact of the HOSPITAL variable greatly influences the accuracy of the machine learning model, making it a model with a low ability to predict ICU admission when this would be correct.

Keywords: SHAP value; MLG and SHAP; machine learning interpretation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

De acordo com Paula (2004), a escolha do modelo para um determinado problema é um ponto fundamental quando se fala sobre análise de dados partindo do ponto inferencial da estatística. Neste momento, já se pressupõem que uma análise descritiva básica e simples tenha sido feita pelo pesquisador ou analista. O passo adiante se trata, então, do aprofundamento técnico no intuito de obter um melhor apuramento no que se refere a análise e, portanto, é necessário um aparato matemático e estatístico mais avançado. Partindo da ideia de correlação, é fácil compreender que essa relação pode ser explicada através de um modelo linear, mas é necessário que o modelo atenda todas as pressuposições possíveis para que o problema prático possa ser englobado em um mesmo sentido dos fundamentos teóricos.

A normalidade, a constância de variância e também a linearidade nem sempre é encontrada quando partimos para problemas reais.

As contribuições de Nelder e Wedderburn (1972) trazem a tona varias técnicas estatísticas, que eram comumente estudadas separadamente, sendo formuladas de forma unificada, como uma classe de modelos de regressão. Esta teoria, sendo então, uma extensão dos modelos clássicos de regressão, recebeu o nome de modelos lineares generalizados (MLG). Esses modelos são feitos pela base de uma variável resposta, variáveis explanatórias (regressoras) e uma amostra de n observações independentes, em que tem-se:

i) a variável resposta é considerado o componente aleatório do modelo e deve ter uma distribuição pertencente à família exponencial de distribuições;

ii) as variáveis explanatórias compõem a estrutura linear do modelo, ou seja, o componente sistemático;

iii) a função de ligação, a ser escolhida adequadamente, tem o papel de fazer a ligação entre os componentes aleatório e sistemático.

Este trabalho trás um resumo geral para o caso em que utiliza-se a distribuição binomial e a importância de usar uma ligação canônica, mas que nem sempre garante o melhor ajuste do modelo. A distribuição binomial com a ligação logito, que liga o preditor linear à média, culmina no modelo de regressão logística, que apresenta boas interpretações em relação aos fatores presentes nas variáveis explanatórias.

Sobre aprendizado de máquina, pode-se dizer que nos dias atuais uma série de técnicas complexas veem sendo utilizadas no intuito de obter ganhos de previsão, principalmente no campo de aprendizado supervisionado, no qual será desenvolvido o presente trabalho. O principal propósito, quando se pretende utilizar algum modelo deste tipo, é a capacidade de obter boas previsões, sem se preocupar com as interpretações sugeridas pelo modelo, ao contrario do que se propõe na Estatística inferencial. No contexto dos principais modelos atualmente, as estruturas feitas com base em árvores de decisão, *boosting* e *ensemble* dão aos modelos maior precisão nas previsões combinado com ganho operacional (Chen e Guestrin, 2016).

Nos últimos anos, a literatura sobre interpretabilidade nos modelos de aprendizado de

máquina veem ganhando relevância nas principais áreas da ciência da Computação e Estatística, tendo desenvolvido formas de mensurar a importância de cada variável no modelo, principalmente com o uso do SHAP, uma metodologia baseada na teoria dos jogos que é utilizada para estimar a contribuição dos valores de cada recurso (Molnar, Casalicchio e Bischl, 2020). Dessa forma, o potencial de usabilidade e a facilidade de interpretação na pratica, transformam a contribuição do SHAP imprescindível.

Finalmente, a comparação entre SHAP e MLG apresenta uma boa perspectiva para futuros trabalhos, visto que num contexto de interpretação, podem ser parecidos. A combinação dessas duas ferramentas contribuem para uma visão de interpretação que utiliza a estatística inferencial e a previsão.

Desse modo, o principal objetivo deste trabalho é fazer comparações entre esses dois tipos de interpretação, respeitando o fato de que são diferentes tipos de abordagem. Procura-se também, responder perguntas oriundas destas comparações, tais como: se é possível que esses modelos possam trabalhar juntos para prescrição e predição; se há variáveis que apresentam a mesma importância em ambos os modelos, quando olhamos para magnitude e direção de efeito; se é razoável pensar que um modelo de aprendizado de máquina possa vir acompanhado de um modelo inferencial que explore o problema de forma mais consistente.

2 MATERIAL E MÉTODOS

2.1 Material

O banco de dados fornecida pelo Datasus, são notificações por síndrome respiratória aguda grave (SRAG) ocorridas entre janeiro de 2021 e setembro de 2023, incluindo dados de COVID-19. Foram selecionados apenas casos registrados na região metropolitana da cidade de São Paulo. Realizou-se alguns tratamentos no conjunto de dados, como a exclusão de variáveis irrelevantes ou com alta quantidade de NAs e modificações em alguns rótulos de valores visando a melhor compreensão dos dados.

A base de dados a ser analisada está disposta na Tabela 12, no dicionário de variáveis, que se encontra no Apêndice A.

Neste sentido, o modelo de aprendizado de máquina utilizado na análise deste estudo foi o *XGBoost* para classificação binária, com alguns parâmetros básicos e hiperparâmetros, *learning_rate = 0.1*, *max_depth = 5*, *n_estimators = 200* e *random_state = 42*, que se referem a taxa de redução do tamanho do passo (também conhecida como η), profundidade máxima de árvore, número de árvores e valor de semente aleatória, respectivamente. Os demais parâmetros e hiperparâmetros relacionados a *TreeBoost* não precisaram ser ajustados, pois tais parâmetros, assim como os termos de regularização L1 e L2 (λ e α) não resultaram em melhora de desempenho do modelo. O *software* utilizado para a criação do modelo foi o Python (versão 3.9), através da biblioteca *xgboost* (Chen e Guestrin, 2016).

A aplicação do valor de SHAP é feita através dos dados de treino (80% dos dados) utilizados no ajuste do modelo, portanto, os valores foram calculados para as 190200 observações selecionadas. O *software* Python, através da biblioteca *shap* (Lundberg e Lee, 2017a), é o mais utilizado para fazer este tipo de aplicação, visto que apresenta uma gama de opções gráficas de fácil uso e compreensão. Conforme a Figura 1, *model* é o modelo a ser explicado, a função *shap.Explainer* é usada para chamar o método explicador no modelo e este, é aplicado ao conjunto de dados de treino.

Figura 1 – Algoritmo em Python para aplicação do valor de SHAP.

```
# get shap values
explainer = shap.Explainer(model)
shap_values = explainer(X_treino)
```

Fonte: Elaborado pelo autor, 2023.

As demais bibliotecas python utilizadas foram:

- *pandas* (McKinney, 2010), para a manipulação de objetos do tipo data.frame;
- *numpy* (Harris et al., 2020), para a manipulação de objetos do tipo array;

- *sklearn* (Pedregosa et al., 2011), para as métricas de precisão do modelo de aprendizado de maquina.

O modelo linear generalizado foi construido através do *software* R (versão 4.2.2) utilizando o pacote *stats* (R Core Team, 2013). Também foram utilizados os pacotes *hnp* (Moral, Hinde e Demétrio, 2017) e *readxl* (Wickham e Bryan, 2023).

2.2 Métodologia

2.2.1 Modelo Linear Generalizado

Segundo Paula (2004), considerando $Y_1, \ldots, Y_n$ variáveis aleatórias independentes, com função densidade ou função de probabilidades dada pela forma abaixo.

$$f(y_i;\theta_i,\phi) = exp[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)]. \tag{2.1}$$

É possível mostrar sob as condições usuais de regularidade, que

$$\begin{aligned} E\left\{\frac{\partial log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right\} &= 0 \ e \\ E\left\{\frac{\partial^2 log f(Y_i;\theta_i,\phi)}{\partial\theta_i^2}\right\} &= -\left[E\left\{\frac{\partial log f(Y_i;\theta_i,\phi)}{\partial\theta_i}\right\}^2\right] \end{aligned}$$

$\forall i$. Além disso, $E(Y_i) = \mu_i = b'(\theta_i)$ e $Var(Y_i) = \phi^{-1}V(\mu_i)$, em que $V_i = V(\mu_i) = d\mu_i/d\theta_i$ é a função de variância e $\phi^{-1} > 0$ ($\phi > 0$) é conhecido como parâmetro de dispersão, enquanto ϕ é o parâmetro de precisão. Um importante papel é desempenhado pela função de variância neste contexto, pois dada a função de variância, tem-se uma classe de distribuições correspondentes, e vice-versa.

Além da forma estrutural (2.1), os modelos lineares generalizados são definidos pela parte sistemática,

$$g(\mu_i) = \eta_i,$$

no qual $g(\mu_i) = \mathbf{x}_i^T\beta$ é chamado de preditor linear, $\beta = (\beta_1, \ldots, \beta_p)^T$, $p < n$, é um vetor de parâmetros que serão estimados, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ são os valores das variáveis explicativas e $g(\cdot)$ é uma função monótona e diferenciável, denominada função de ligação (Cordeiro e Demétrio, 2008). Dentre as distribuições pertencentes à família exponencial, uma das mais conhecidas e também muito utilizada é a distribuição binomial, que é dada da seguinte forma.

Sendo Y a proporção de sucessos em n ensaios independentes, com probabilidade de ocorrência μ. Assumindo que $nY \sim B(n,\mu)$, sua função de probabilidade é dada por

$$p(y;\mu) = exp\left\{\log\binom{n}{ny} + ny\log\left(\frac{\mu}{1-\mu}\right) + n\log(1-\mu)\right\},$$

em que $0 < \mu, y < 1$. Com isso, $\phi = n$, $\theta = \log\{\frac{\mu}{1-\mu}\}$, $b(\theta) = \log(1+e^\theta)$ e $c(y;\phi) = log\binom{\phi}{\phi y}$. A função de variância é dada por $V(\mu) = \mu(1-\mu)$.

A distribuição binomial possui grande usabilidade para casos em que se deseja estimar proporções ou uma variável resposta do tipo dicotômica, em que os resultados possíveis são 0 ou 1. Neste sentido, a finalidade da utilização da distribuição binomial neste trabalho se dar por conta da necessidade de se estimar os resultados possíveis para a variável UTI, em que 0 significa a não ocorrência da internação e 1 significa a ocorrência da internação na UTI.

2.2.2 *Variável binária*

Segundo Dobson e Barnett (2018), modelos lineares generalizados podem ser aplicados a variáveis em que os resultados são medidas numa escala binária. Por exemplo, as respostas podem ser vivas ou mortas, presentes ou ausentes. Sucesso e fracasso são usados como termos genéricos das duas categorias. Assim, definindo a variável aleatória binária, tem-se que,

$$z = \begin{cases} 1, & \text{se o resultado for sucesso} \\ 0, & \text{se o resultado for fracasso} \end{cases} \tag{2.2}$$

Com probabilidades $Pr(Z=1)=p$ e $Pr(Z=0)=1-p$, que segue à distribuição de bernoulli $B(p)$. Se existirem n variáveis aleatórias $Z_1,...,Z_n$, independentes com $Pr(Z_j=1)=\pi_j$, então sua probabilidade conjunta é

$$\prod_{j=1}^{n} \pi_j^{z_j}(1-\pi_j)^{1-z_j} = exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right], \tag{2.3}$$

sendo assim, pertencente à família exponencial. A seguir, para o caso em que os π_j's são todos iguais, pode-se definir

$$Y = \sum_{j=1}^{n} Z_j,$$

de modo que Y é o número de sucessos em n tentativas e, neste sentido, a variável aleatória Y tem distribuição $Bin(n,p)$, de modo que

$$Pr(Y=y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \;\; y=0,1,...,n. \tag{2.4}$$

Finalmente, considerando o caso geral de N variáveis aleatórias independentes $Y_1,Y_2,...,Y_N$, correspondente ao número de sucessos em N diferentes subgrupos ou estratos. Se $Y_i \sim Bin(n_i,\pi_i)$, a função log-verossimilhança é

$$l(\pi_1,...,\pi_N;y_1,...,y_n) = \sum_{i=1}^{N}\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i) + \log\binom{n_i}{y_i}\right]. \tag{2.5}$$

2.2.3 *Modelo linear para proporções*

Pretende-se descrever a proporção de sucessos, $P_i = Y_i/n_i$, em cada subgrupo em termos de níveis de fatores e outras variáveis explicativas que caracterizam o subgrupo. Como $E(Y_i) = n_i\pi_i$ e então $E(P_i)=\pi_i$, modela-se as probabilidades π_i como

$$g(\pi_i) = x_i^T\boldsymbol{\beta},$$

em que x_i é um vetor de variáveis explicativas, $\boldsymbol{\beta}$ é um vetor de parâmetros e $g(.)$ é uma função de ligação.

O caso mais simples é o modelo linear

$$\pi = x^T\boldsymbol{\beta}. \tag{2.6}$$

Entretanto, embora π seja uma probabilidade, os valores ajustados $x^T\boldsymbol{\beta}$ podem ser menores que zero ou maiores que um. Para garantir que π esteja restrito ao intervalo [0,1], muitas vezes é modelado usando uma distribuição de probabilidade cumulativa

$$\pi = \int_{-\infty}^{t} f(s)ds,$$

no qual $f(s) \geq 0$ e $\int_{-\infty}^{\infty} f(s)ds = 1$. A função de densidade de probabilidade $f(s)$ é chamada de distribuição de tolerância (Dobson e Barnett, 2018).

Um modelo que fornece bons resultados é o modelo logístico ou logit. Assim, dada a distribuição de tolerância

$$f(s) = \frac{\beta_2 exp(\beta_1 + \beta_2 s)}{[1 + exp(\beta_1 + \beta_2 s)]^2},$$

então

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{exp(\beta_1 + \beta_2 x)}{1 + exp(\beta_1 + \beta_2 x)},$$

que resulta na função de ligação

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x.$$

O termo $\log\left(\frac{\pi}{1-\pi}\right)$ é algumas vezes chamado de função logit e tem uma interpretação natural como o logaritmo das probabilidades.

2.2.4 Modelo geral de regressão logística

O modelo simples de regressão logística usado em (2.6) é uma caso especial do modelo geral de regressão logística (Dobson e Barnett, 2018)

$$logit\,\pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i^T}\boldsymbol{\beta},$$

em que $\mathbf{x_i}$ é um vetor de variáveis que podem ser contínuas ou variáveis dummy correspondentes a níveis de fator e $\boldsymbol{\beta}$ é o vetor de parâmetros. Este modelo é muito utilizado para análise de dados envolvendo respostas binárias ou binomiais e diversas variáveis explicativas.

Estimativas de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ e, consequentemente, das probabilidades $\pi_i = g^{-1}(\boldsymbol{x_i^T}\boldsymbol{\beta})$, são obtidas maximizando a função de log-verossimilhança,

$$l(\pi;\mathbf{y}) = \sum_{i=1}^{N}\left[y_i \log \pi_i + (n_i - y_i)\log(1-\pi_i) + \log\binom{n_i}{y_i}\right].$$

A função deviance é a estatística de log-verossimilhança, resultante da diferença da função de verossimilhança de um modelo saturado (com o número máximo de parâmetros) e de um modelo de interesse a ser testado (com a quantidade de parâmetros reduzida), utilizada para verificar a qualidade do modelo de interesse. Pode ser escrita como

$$D = 2\sum o \log\frac{o}{e},$$

em que o denota os "sucessos" y_i e "fracassos" $(n_i - y_i)$ observados e e denota as frequências esperadas estimadas correspondentes ou valores ajustados $\hat{y}_i = n_i\hat{p}_i$ e $(ni - \hat{yi}) = (n_i - n_i\hat{p}_i)$ (Dobson e Barnett, 2018).

Observe que D não envolve nenhum parâmetro desconhecido, portanto, a qualidade do ajuste pode ser avaliada e as hipóteses podem ser testadas diretamente usando a aproximação

$$D \sim \chi^2(N-p),$$

com p sendo o número de parâmetros estimados e N o número de padrões de covariáveis.

2.2.5 Risco Relativo

Segundo Paula (2004), supondo que indivíduos de uma determinada população são classificados por um fator de dois níveis, A e B, e a presença ou ausência de uma doença denotados, respectivamente, por D_1 e D_0. As proporções são descritas conforme a Tabela 9.

Tabela 1 – Tabela de exemplificação para o Risco Relativo

	Fator	
Doença	A	B
D_1	P_1	P_3
D_0	P_2	P_4

Fonte: Paula (2004).

Essas proporções podem ser definidas como:

$P_1/(P_1+P_2)$: proporção de indivíduos classificados como doentes no grupo A;

$P_3/(P_3+P_4)$: proporção de indivíduos classificados como doentes no grupo B.

Cornfield (1951) denominou a razão entre essas proporções como sendo o risco relativo de doença entre os níveis A e B, em que

$$RR = \frac{P_1/(P_1+P_2)}{P_3/(P_3+P_4)} = \frac{P_1/(P_3+P_4)}{P_3/(P_1+P_2)}, \tag{2.7}$$

que assume a forma simplificada

$$\psi = \frac{P_1P_4}{P_2P_3}, \tag{2.8}$$

denominada por *Odds ratio*, conhecida também como razão de chances.

No contexto do modelo de regressão logística, a razão de chances é utilizada na interpretação dos parâmetros.

Desse modo, supondo um modelo de regressão logística simples, em que x é uma variável dicotômica com as categorias A e B, sendo B a variável indicadora, ou seja, B ocorre quando $x = 1$, paralelamente, A ocorre quando $x = 0$, tem-se que

$$\psi_B = \frac{odds\{B\}}{odds\{A\}} = \frac{\pi_B/(1-\pi_B)}{\pi_A/(1-\pi_A)} = \frac{e^{\beta_0+\beta_1 x1}}{e^{\beta_0+\beta_1 x0}} = e^{\beta_1}$$

no qual π_A é a probabilidade de ocorrência de um individuo na categoria A e π_B a probabilidade de ocorrência de um individuo na categoria B. Dessa forma, e^{β_1} é a razão de chances de resposta para a categoria B em relação a A.

Para o caso, em que x é uma variável numérica, e^{β_1} corresponde ao acréscimo na chance de resposta para um aumento unitário em x.

2.2.6 Teste de Wald

O teste de Wald é um dos testes estatísticos utilizados na regressão logística para verificar a significância dos coeficientes (βs) estimados do modelo. Especificamente, consiste em testar o efeito dos preditores nas log-odds (ou logaritmo das razões de chances) da variável resposta. Dado que a razão de chances mede a relação do efeito de uma categoria B contra uma categoria A, quando $\beta = 1$ o efeito de ambas as categorias é o mesmo, ou seja, a presença ou não de uma doença não tem efeito sobre a variável resposta. Neste sentido, o teste tem o objetivo de verificar se $\beta_j \neq 1$ ou ainda, se o intervalo de confiança para o parâmetro β_j não contém o valor 1.

Inicialmente, segundo Paula (2004), supõem-se as hipóteses simples

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^0 \;\; vs \;\; H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}^0$$

em que $\boldsymbol{\beta}^0$ é um vetor p-dimensional conhecido e ϕ é conhecido. Desse modo, o teste de Wald é definido por

$$\xi_W = [\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0]^T \hat{V}ar^{-1}(\widehat{\boldsymbol{\beta}})[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0],$$

em que $\hat{V}ar^{-1}(\widehat{\boldsymbol{\beta}})$ é a matriz de variância-covariância assintótica de $\widehat{\boldsymbol{\beta}}$. A estatística de Wald pode ser reescrita como

$$\xi_W = \phi[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0]^T (\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0],$$

com W sendo uma matriz de pesos que muda a cada passo do processo iterativo. Para o caso $p = 1$, quando há apenas uma variável a ser testada no modelo, o teste de Wald se equivale ao teste t^2.

A região de confiança para $\boldsymbol{\beta}$, baseada no teste de Wald e com confiança $(1-\alpha)$, pode ser dada por

$$[\widehat{\boldsymbol{\beta}}; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \phi^{-1} \chi^2_p (1-\alpha)],$$

em que $\chi^2_p(1-\alpha)$ denota o percentil $(1-\alpha)$ de uma distribuição qui-quadrado com p graus de liberdade.

2.2.7 Avaliação da qualidade do ajuste

2.2.7.1 AIC e BIC

Uma das mais importantes medidas usada para avaliar a qualidade do ajuste é o critério de informação de Akaike (**Akaike Information Criterion - AIC**), proposto por Akaike (1974). Para

um modelo qualquer, sendo $\widehat{l}$ a log-verossimilhança maximizada e p o número de parâmetros estimados, segue que

$$AIC = -2\widehat{l}_p + 2p.$$

Observa-se que esta medida é penalizada pela quantidade de parâmetros presentes no modelo (complexidade do modelo). E neste sentido, busca-se um modelo com o menor valor de AIC.

Outra opção, é o critério de informação de Bayes (BIC), proposto por Schwarz (1978), representado por

$$BIC = -2\widehat{l}_p + p\log(n).$$

Neste caso, a penalidade de modelos mais complexos se dá também por amostras maiores.

2.2.7.2 Gráficos semi-normais com envelopes simulados

Esta técnica, relativamente fácil, consiste em traçar os valores absolutos ordenados de um modelo de diagnóstico *versus* as estatísticas de ordem esperada de uma distribuição semi-normal (Moral, Hinde e Demétrio, 2017), que pode ser aproximada como

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right),$$

em que i é a estatística de i-ésima ordem, $1 \leq i \leq n$ e n é o tamanho da amostra, como em McCullagh e Nelder (1989), seguindo os resultados de Blom (1958) e Royston (1982).

A obtenção do envelope simulado para um gráfico semi-normal é simples e consiste em:

1. Ajustar um modelo;
2. Extrair diagnósticos de modelos e calcular valores absolutos ordenados;
3. Simular variáveis de resposta usando a mesma matriz de modelo, distribuição de erros e estimativas de parâmetros;
4. Ajustar o mesmo modelo para cada variável de resposta simulada, extrair o mesmo diagnóstico do modelo e ordenar novamente os valores absolutos;
5. Calcular os percentis desejados (por exemplo, 2,5 e 97,5) dos valores de diagnóstico simulados em cada valor da estatística de ordem esperada e usá-los para formar o envelope.

2.3 Modelo XGBoost

Extreme Gradient Boosting (XGBoost) é um algoritmo de aprendizado de maquina que vem ganhando muita notoriedade nos últimos anos, visto que tem bons resultados tanto para problemas de regressão como classificação, sendo eficiente e escalável para diversos tipos de predição. Sua estrutura é baseada no aprendizado supervisionado, pertencendo à família de modelos de *boosting* (Friedman, 2001), baseado em árvores de decisão. De forma geral, consiste em usar um conjunto de árvores de decisão, em que cada árvore é construída de forma sequencial, a fim de corrigir os erros residuais das árvores anteriores, utilizando árvores de decisão rasas, combinadas para formar um modelo mais robusto.

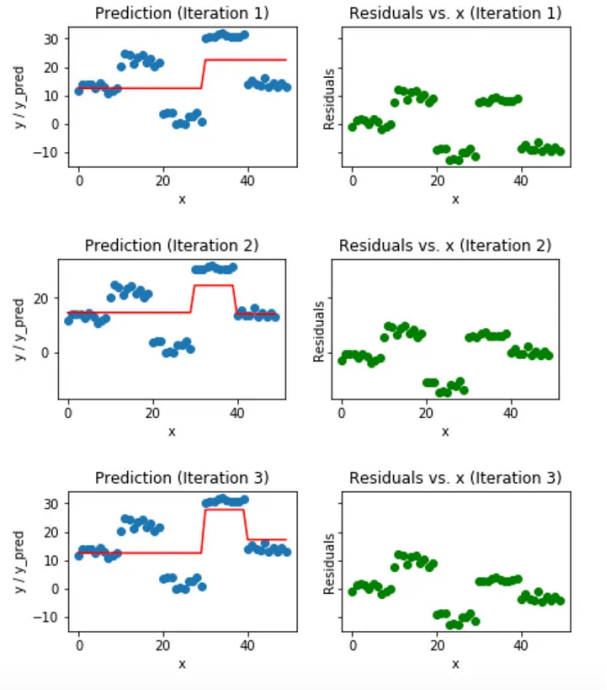
2.3.1 Gradiente Boost

Pode-se dizer que o XGBoost é uma otimização dos modelos gradiente boost (Chen e Guestrin, 2016). Este, tem como ideia central, ajustar árvores sequencialmente, visando corrigir os erros residuais (diferença entre as previsões atuais e os valores reais). Este processo é feito através de um processo iterativo de ajuste.

A estrutura da árvore é formada por pontos de divisão, conhecidos como nós, que segmentam os dados de acordo com as suas características, dividindo-os em ramos. Essas árvores podem ser ajustadas por hiperparâmetros para evitar *overfitting*. As interações podem ser visualizadas nas Figuras 2 e 3. Em que, na primeira interação, a partir de um pequeno obstáculo, na etapa seguinte, é usado como um reforço para aumentar o desempenho e construir um reforço mais forte, reduzindo assim, a função de perda. Sendo assim, este reforço é adicionado iterativamente a cada modelo e calcula-se a perda, e, então, com este valor de perda as previsões são atualizadas a fim de minimizar os resíduos.

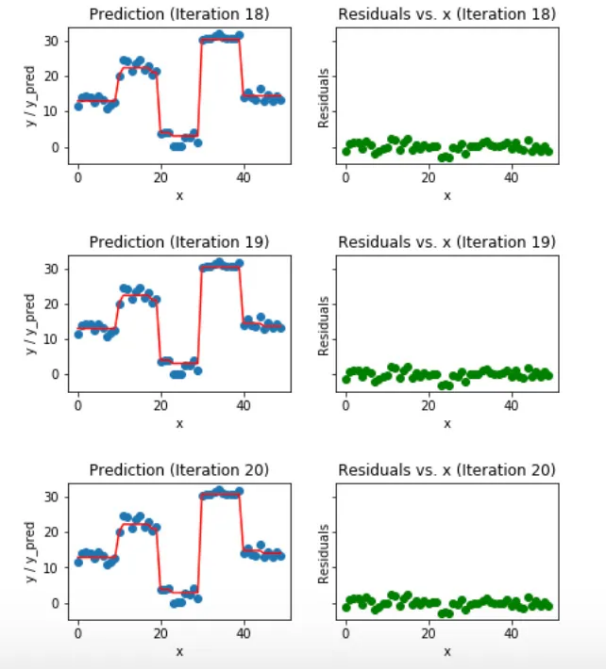
Dessa forma, a partir das Figuras observa-se que os gráficos de pontos azuis (esquerda) são entrada (x) *versus* saída (y). Neste sentido, a linha vermelha (esquerda) mostra os valores previstos pela árvore de decisão. Por outro lado, os pontos verdes (direita) mostram os resíduos *versus* a entrada (x) para a i-ésima iteração. A iteração representa a sequencial ordem de ajuste da árvore de aumento de gradiente (Grover, 2017).

Figura 2 – Visualização de previsões de aumento de gradiente (3 primeiras iterações).



Fonte: Grover (2017)

Figura 3 – Visualização de previsões de aumento de gradiente (18ª a 20ª iterações).



Fonte: Grover (2017)

2.3.2 Função Objetivo

Durante o treinamento, o XGBoost utiliza uma função de perda (*loss function*) para medir o quão bem o modelo está se saindo. No caso de problemas de classificação binária, a função de perda frequentemente utilizada é a função logística (*log loss*).

A função *Log Loss* é definida como

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p(y_i)) + (1-y_i)\log(1-p(y_i))\right],$$

em que N é o número total de observações no conjunto de dados; y_i é a verdadeira classe da i-ésima observação (0 ou 1); $p(y_i)$ é a probabilidade prevista pelo modelo de que a i-ésima observação pertence à classe positiva (1).

O objetivo é minimizar essa função durante o treinamento, ajustando os parâmetros do modelo para que as previsões se aproximem o máximo possível dos verdadeiros rótulos, penalizando fortemente as previsões incorretas (Saha, 2018). Durante o treinamento, o modelo ajusta os pesos das previsões anteriores, com base na derivada da função de perda em relação às previsões. A interpretação da *log loss* é tal que quanto menor o valor obtido, melhor é a qualidade das previsões do modelo. Essa função penaliza mais fortemente as previsões erradas, especialmente quando estão muito longe do rótulo real.

O modelo pode ser regularizado a partir de vários parâmetros de regularização para controle de complexidade do modelo, como a profundidade máxima da árvore, o número mínimo de amostras por folha, a taxa de aprendizado (*learning rate*) e a regularização $L1/L2$ (penalização dos pesos) (Brownlee, 2020).

As previsões de todas as árvores são agregadas para formar a previsão final do modelo (*Ensemble*). Normalmente, a previsão final é obtida pela média das previsões individuais ou através de uma abordagem ponderada.

2.4 Valores Shapley

Os *Shapley Values* (ou "valores de Shapley") foram introduzidos por Shapley (1951), no contexto da teoria dos jogos. Para um jogo cooperativo qualquer, os valores de Shapley distribuem uma quantidade total de contribuição para cada jogador da equipe de forma justa.

Considera-se que uma previsão pode ser explicada assumindo que cada valor de recurso da instância é um "jogador" em um jogo em que a previsão é o pagamento. O método atribui pagamentos aos jogadores dependendo da sua contribuição para o pagamento total. Os valores de Shapley exibem, através da cooperação dos jogadores em coalizões, como distribuir de forma justa o "pagamento" entre os recursos (Molnar, 2022).

A concepção geral dos valores de Shapley pode ser entendida a partir de um exemplo, em que se tem um modelo de aprendizado de máquina para prever preços de apartamentos. Assim, em um determinado apartamento, a previsão é de €300.000. Este tem área de 50m², situa-se no 2ª andar e é próximo de um parque, mas proíbe gatos. O objetivo é explicar esta previsão, ou seja, esclarecer como cada um desses valores de recursos contribuiu para a previsão. Portanto, dado que a previsão média para os apartamentos é €310.000, é necessário descobrir a contribuição de cada recurso para a previsão real em relação a previsão média.

De forma geral, o cálculo dos valores de Shapley envolve considerar todas as combinações possíveis de recursos e calcular as contribuições marginais de cada recurso em relação à predição do modelo. Supondo um conjunto de características (recursos) $X = x_1, x_2, ..., x_n$. Para obter as contribuições marginais dos recursos, avalia-se o impacto de x_i na predição para cada combinação possível de recursos que inclui ou exclui x_i, calcula-se a predição do modelo em relação a essas combinações e a diferença resultante na predição. Daí, encontra-se a média ponderada dessas diferenças considerando todas as permutações possíveis de recursos, em que o peso é dado pelo número de recursos no conjunto.

2.4.1 SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) de Lundberg e Lee (2017b) é um método para explicar previsões individuais é baseado na teoria dos jogos. Segundo os autores, "A melhor explicação de um modelo é um modelo de explicação mais simples, que definimos como qualquer aproximação interpretável do modelo original". Neste sentido, a explicação do valor de Shapley pode ser entendida como um método aditivo de atribuição de recursos, um modelo linear. Um modelo de explicação, sendo função linear de variáveis binárias, é descrito como

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{2.9}$$

em que $z' \in \{0,1\}^M$, M é o número de recursos de entrada e $\phi_i \in \mathbb{R}$.

Estes métodos atribuem um efeito y_i a cada recurso e a soma dos efeitos de todas as atribuições de recursos aproxima a saída $f(x)$ do modelo original.

2.4.2 Estimativa clássica do valor de Shapley

As estimativas do valor de Shapley são encontradas através de um modelo de regressão (Lipovetsky e Conklin, 2001). Este método requer o retreinamento do modelo em todos os subconjuntos de recursos $S \subseteq F$, no qual F é o conjunto de todos os recursos, atribuindo um valor de importância a cada recurso, que representa o efeito na previsão do modelo com a inclusão desse recurso. Um modelo $f_{S\cup\{i\}}$ é treinado com esse recurso presente e outro modelo f_S é treinado com o recurso retido. Em seguida, as previsões dos dois modelos são comparadas na entrada atual $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, em que x_S representa os valores dos recursos de entrada no conjunto S. Como o efeito da retenção de um recurso depende de outros recursos no modelo, as diferenças anteriores são calculadas para todos os subconjuntos possíveis $S \subseteq F\backslash\{i\}$. Os valores Shapley são, então, calculados e utilizados como atribuições de recursos. Eles são uma média ponderada de todas as diferenças possíveis:

$$\phi_i = \sum_{S\subseteq F\backslash\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{2.10}$$

Para valores de regressão Shapley é feito o mapeamento 1 ou 0 para o espaço de entrada original, no qual 1 indica que a entrada está incluída no modelo e 0 indica exclusão do modelo. Se deixarmos $\phi_0 = f_\oslash(\oslash)$, então os valores da regressão de Shapley correspondem à Equação (2.11) e são, portanto, um método aditivo de atribuição de características.

2.4.3 Propriedades

Uma vantagem da classe de métodos aditivos de atribuição de características, é a presença de uma única solução nesta classe, com três propriedades desejáveis. Sendo elas:

1. **Precisão local**: Ao aproximar o modelo original f para uma entrada específica x, a precisão local exige que o modelo de explicação corresponda pelo menos à saída de f para a entrada simplificada x (que corresponde à entrada original x).

2. **Falta**: Se as entradas simplificadas representam a presença de recursos, então a falta exige que os recursos ausentes na entrada original não tenham impacto. Todos os métodos descritos na Seção 2, obedecem à propriedade de falta.

3. **Consistência**: A consistência afirma que, se um modelo mudar de modo que a contribuição de algum insumo simplificado aumente ou permaneça a mesma independentemente dos outros insumos, a atribuição desse insumo não deve diminuir.

3 RESULTADOS E DISCUSSÃO

3.1 Análise descritiva

Na Tabela 2, são apresentadas as principais medidas estatísticas de resumo para as variáveis numéricas do tipo discreta, sendo elas, idade e a diferença de semanas entre os primeiros sintomas e a notificação. A maioria dos indivíduos são adultos, isto fica evidente quando observado que a média de idade é 47 anos, bem como, 75% das pessoas tem 22 anos ou mais. Em relação a diferença entre a semana dos primeiros sintomas e a semana da notificação, pode-se observar que, em geral, as pessoas não costumam demorar muito tempo para ir ao médico, após perceberem os primeiros sintomas, visto que, em média, este intervalo de tempo dura aproximadamente 1 semana e além disso, em 75% dos casos este intervalo de tempo é de até 2 semanas.

Tabela 2 – Medidas de resumo para as variáveis idade e diferença de semanas.

Nome	**Mín**	**Máx**	**Média**	**Dp**	**1ª Q**	**Mediana**	**3ª Q**
IDADE	0	110	46,81	28,7	22	52	70
DIF_SEM_NOT_PRI	0	48	1,27	1,94	0	1	2

Fonte: Elaborado pelo autor, 2023.

As variáveis ordinais foram utilizadas através de uma escala numérica categorizada. Sobre a escolaridade dos indivíduos, desconsiderando os que ignoraram ou não se enquadram para este fator, a maioria (7,2%) possuem o ensino médio, já as minorias são os que possuem ensino superior e os analfabetos. A zona urbana é onde reside a maioria dos indivíduos, cerca de 82%. E das pessoas gestantes, dentre as que se aplicam a esta categoria, a maioria se encontra no 3ª trimestre de gestação. Essas e outras informações estão expostas na Tabela 3.

Tabela 3 – Distribuição de frequências para as variáveis escolaridade, zona geográfica e idade gestacional da paciente.

Nome	**Classificação**	**Frequência**	**Freq. %**
Escolaridade	0: Ignorado/Não se aplica	179299	75.4%
	1: Analfabeto/Sem escolaridade	8861	3.7%
	2: Fundamental 1º ciclo	14137	5.9%
	3: Fundamental 2º ciclo	9872	4.2%
	4: Ensino Médio	17164	7.2%
	5: Ensino Superior	8418	3.5%
Zona geográfica	0: Ignorado/Vazio	41844	17.6%
	1: Rural	858	0.4%
	2: Periurbano	200	0.1%
	3: Urbano	194849	82.0%
Idade gestacional da paciente	0: Ignorado/Não se aplica	236102	99.3%
	1: 1º Trimestre	200	0.1%
	2: 2º Trimestre	487	0.2%
	3: 3º Trimestre	962	0.4%

Fonte: Elaborado pelo autor, 2023.

Na Tabela 4, são apresentadas as frequências das morbidades dos indivíduos, no qual pode-se observar que os principais potenciais fatores de risco, em termos de frequência, são relacionados a doença cardiovascular crônica (24,69%), diabetes *mellitus* (15,7%) e outras morbidades não listadas (20,71%).

Tabela 4 – Distribuição de frequências para as variáveis relacionadas a morbidade.

Morbidade	**Frequência (% Freq.)**	
	Não possui: 0	Possui: 1
Puérpera	237202 (99.77%)	549 (0.23%)
Doença Cardiovascular Crônica	179049 (75.31%)	58702 (24.69%)
Doença Hematológica Crônica	235964 (99.25%)	1787 (0.75%)
Síndrome de Down	237063 (99.71%)	688 (0.29%)
Doença Hepática Crônica	236024 (99.27%)	1727 (0.73%)
Asma	227514 (95.69%)	10237 (4.31%)
Diabetes mellitus	200417 (84.30%)	37334 (15.70%)
Doença Neurológica Crônica	229372 (96.48%)	8379 (3.52%)
Pneumatopatia Crônica	229141 (96.38%)	8610 (3.62%)
Imunodeficiência ou Imunodepressão	231567 (97.40%)	6184 (2.60%)
Doença Renal Crônica	230771 (97.06%)	6980 (2.94%)
Obesidade	226308 (95.19%)	11443 (4.81%)
Outros	188505 (79.29%)	49246 (20.71%)

Fonte: Elaborado pelo autor, 2023.

Os sintomas e sinais mais presentes nos indivíduos são tosse, dispneia, saturação (O_2<95%) e desconforto respiratório, esses estão presentes em cerca de 65%, 60%, 59% e 52% dos casos notificados, respectivamente, conforme observado na Tabela 5.

Tabela 5 – Distribuição de frequências para as variáveis relacionadas a sintomas e sinais.

Sintomas	**Frequência (% Freq.)**	
	Não possui: 0	Possui: 1
Dor abdominal	225467 (94.83%)	12284 (5.17%)
Fadiga	189650 (79.77%)	48101 (20.23%)
Perda do Olfato	227326 (95.62%)	10425 (4.38%)
Perda do Paladar	226836 (95.41%)	10915 (4.59%)
Febre	125387 (52.74%)	112364 (47.26%)
Tosse	81939 (34.46%)	155812 (65.54%)
Dor de Garganta	207119 (87.12%)	30632 (12.88%)
Dispneia	94066 (39.56%)	143685 (60.44%)
Desconforto Respiratório	113712 (47.83%)	124039 (52.17%)
Saturação de O_2<95%	97120 (40.85%)	140631 (59.15%)
Diarreia	217074 (91.30%)	20677 (8.70%)
Vômito	217566 (91.51%)	20185 (8.49%)
Outros	167778 (70.57%)	69973 (29.43%)

Fonte: Elaborado pelo autor, 2023.

Observando os fatores sociais na Tabela 6, se tratando de raça/cor dos indivíduos, pode-se dizer que a maioria dos casos notificados são por pessoas brancas, que representam 45,5%. Outra categoria também muito frequente são os pardos, com 25,3% dos casos.

Tabela 6 – Distribuição de frequências para as variáveis relacionadas a raça/cor.

Raça/Cor	**Frequência (% Freq.)**	
	Não possui: 0	Possui: 1
Branca	129539 (54.49%)	108212 (45.51%)
Preta	227422 (95.66%)	10329 (4.34%)
Amarela	234605 (98.68%)	3146 (1.32%)
Parda	177618 (74.71%)	60133 (25.29%)
Indigena	237603 (99.94%)	148 (0.06%)

Fonte: Elaborado pelo autor, 2023.

Na Tabela 7, encontram-se informações sobre a variável resposta para este estudo, que é a internação ou não na UTI. As internações aconteceram em 31% dos casos, além de outras variáveis informativas levadas em consideração na ficha desses casos, como por exemplo, se o individuo recebeu a vacina da covid, em que 37% receberam, como também, se o paciente ficou internado no hospital, o que ocorreu em cerca de 96% dos casos.

Tabela 7 – Distribuição de frequências para a variável resposta (Internação na UTI) e variáveis relacionadas a informações adicionais.

Váriavel resposta e outras informações	**Frequência (% Freq.)**	
	Não: 0	Sim: 1
Internado em UTI?	164026 (68.99%)	73725 (31.01%)
Infecção adquirida no hospital?	232933 (97.97%)	4818 (2.03%)
Paciente trabalha ou tem contato direto com aves ou suínos?	236827 (99.61%)	924 (0.39%)
Recebeu vacina COVID-19?	149726 (62.98%)	88025 (37.02%)
Usou antiviral para gripe?	232584 (97.83%)	5167 (2.17%)
Paciente foi internado no hospital?	8437 (3.55%)	229314 (96.45%)

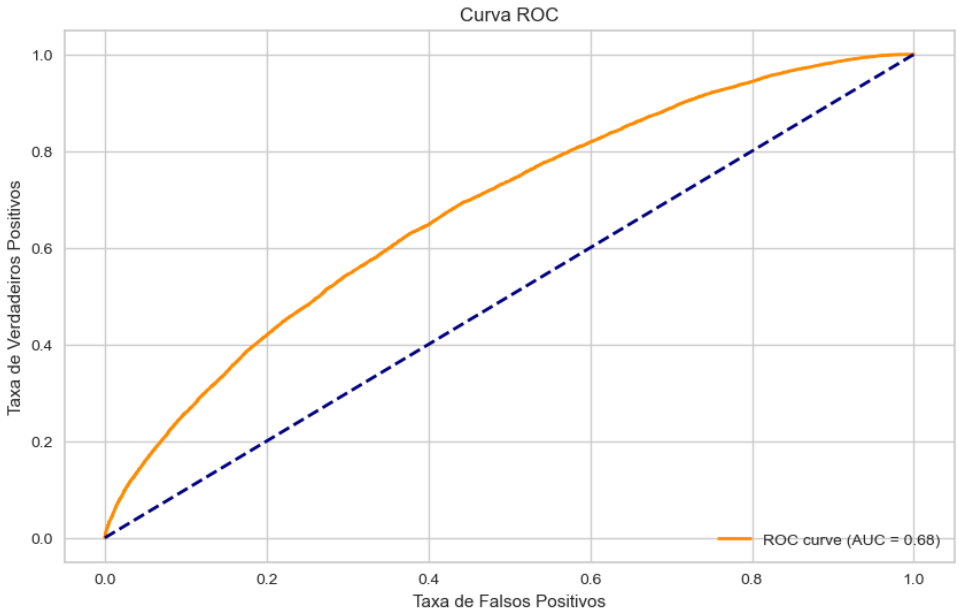
Fonte: Elaborado pelo autor, 2023.

3.2 Modelo XGBoost

Os dados foram divididos em treino e teste, em que 80% das observações foram destinadas a treino (190200). Para medir o desempenho do modelo, a validação cruzada com 5 *folders* não apresentou diferença em relação a validação utilizando os dados de teste de forma direta (47551).

O modelo a ser analisado obteve uma acurácia geral de 0,71 e área sob a curva ROC de 0,68, conforme a Figura 4. Essas são métricas importantes para a validação de um modelo de aprendizado de máquina, mas é importante observar outras métricas quando se trata de dados desbalanceados, como é o caso deste. A acurácia deste modelo é resultante de valores desequilibrados de Recall. Para valores positivos, ou seja, a previsão de casos em que era realmente necessário a internação na UTI (sensibilidade), o resultado foi de apenas 0,15. Já para valores negativos, ou seja, a previsão de casos que não precisariam da internação na UTI (especificidade), o resultado foi de 0,95.

Figura 4 – Curva ROC e área sob a curva do modelo.



Fonte: Elaborado pelo autor, 2023.

3.3 Valor de SHAP

3.3.1 Principais gráficos e interpretações

A contribuição do valor de cada recurso (variável) é estimada pelo valor de SHAP, em que cada observação têm um valor de contribuição ou impacto, respectivo, a cada variável. No gráfico *waterfall* apresenta-se, para uma observação de interesse, a contribuição dos respectivos valores de recurso para a predição. Valores de SHAP positivos aumentam a probabilidade para o individuo ser internado na UTI, enquanto que os negativos, diminuem esta probabilidade. $E[f(x)]$ é a média da probabilidade logarítmica prevista para todos os casos, pode-se dizer que este seria um "ponto de partidaṗara a predição e, $f(x)$, a probabilidade logarítmica prevista para esta observação, que resulta na predição (O'Sullivan, 2023). Para o caso de variáveis binárias, considera-se $f(x) = ln(\frac{p}{1-p})$. Assim, esta relação é dada pela forma

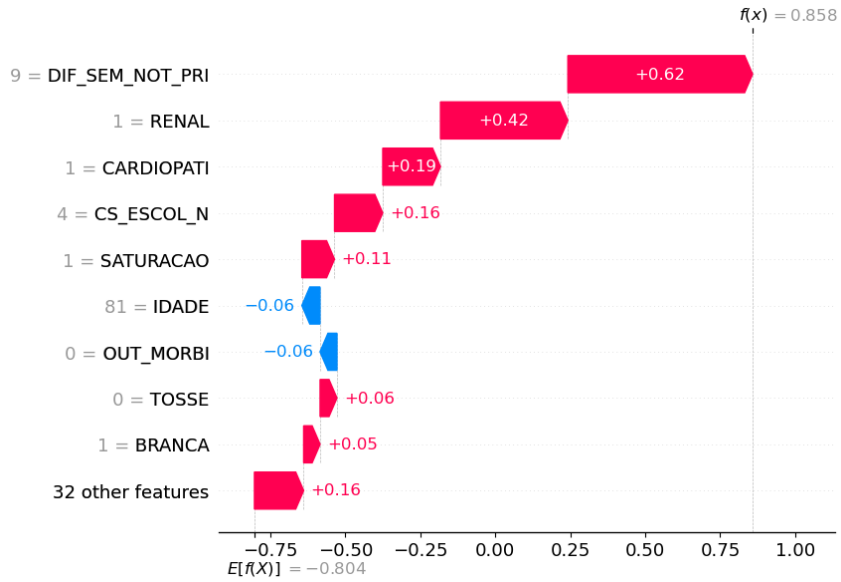
$$\ln(\frac{p}{1-p}) = E[ln(\frac{p}{1-p})] + \text{SHAP values}.$$

Na Figura 5, nota-se um caso previsto como "1", ou seja, a internação na UTI. O maior impacto foi da variável diferença de semanas entre o primeiro sintoma e a notificação, no qual essa diferença foi de 9 semanas, aumentando a probabilidade logarítmica em 0,62. Além disso,

$$p = \frac{e^{0,858}}{1+e^{0,858}} \approx 0,702.$$

Portanto, o modelo prevê a probabilidade de 0,702 para a ida do individuo à UTI.

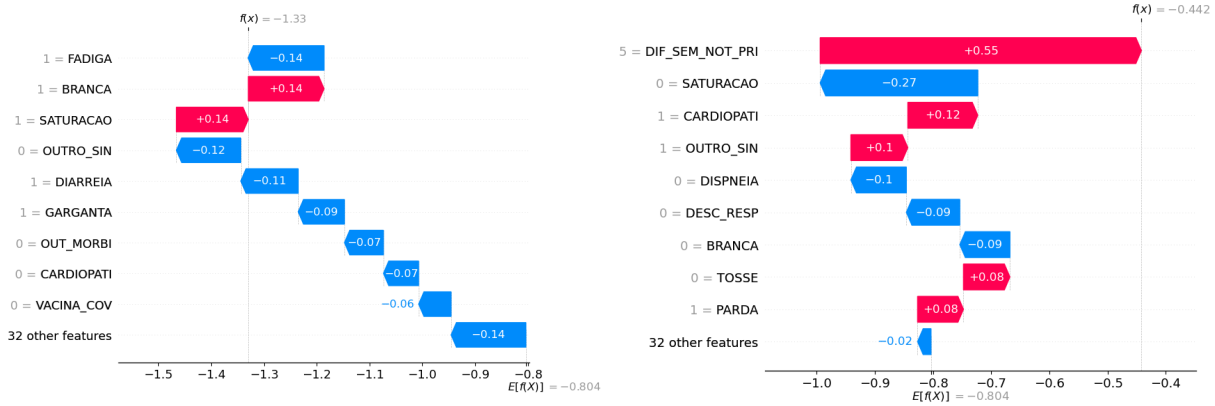
Figura 5 – Gráfico *waterplot* para uma observação predita como ida para UTI.



Fonte: Elaborado pelo autor, 2023.

Na Figura 6 são exibidos dois casos previstos como "0", mas que recebem contribuições diferentes para as respectivas previsões. No caso **a)**, a maioria das variáveis tiveram um impacto negativo, apesar do impacto positivo na presença da saturação de O_2<95%, portanto, o modelo prevê a probabilidade $p = 0,209$ de ida para a UTI. Em **b)**, há um grande impacto positivo causado pela quantidade de semanas na diferença de semanas entre o primeiro sintoma e a notificação e também pela presença de doença cardiovascular crônica, mas que é reduzido pela não presença da saturação de O2<95% e de outras características. Neste caso, no geral, existe um impacto positivo na probabilidade logarítmica, mas que não é suficiente para a internação na UTI, resultando numa probabilidade $p = 0.391$ de ida para a UTI.

Figura 6 – Gráfico waterplot para duas observações distintas preditas como não ida para UTI.



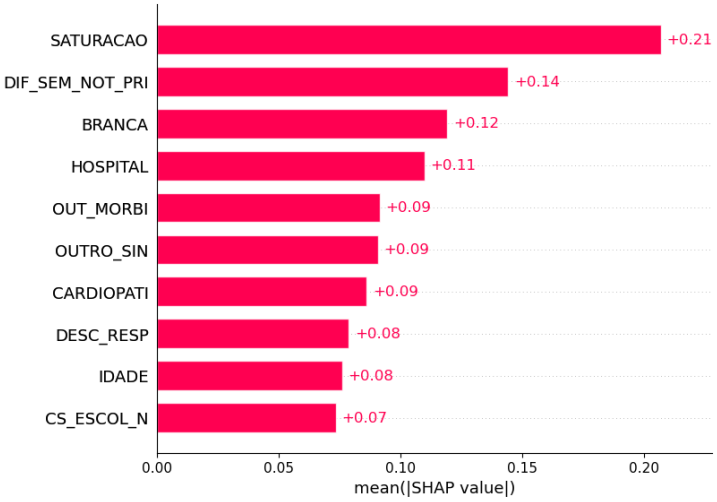
a) b)

Fonte: Elaborado pelo autor, 2023.

Uma das formas de observar a contribuição global das variáveis presentes no modelo é através do gráfico de barras, em que no eixo y tem-se o nome das variáveis e o eixo x é representado pela média dos valores de SHAP em módulo.

Na Figura 7, são apresentadas as dez variáveis que mais impactam, de forma geral, na contribuição para as predições, sem fazer distinção entre a presença ou não do evento, no caso das variáveis binárias. Do grupo de sinais e sintomas, a saturação de O_2<95 se mostra como a mais importante, seguida de outros sintomas e desconforto respiratório. As variáveis discretas também apresentam boa importância, como a idade e, principalmente, a diferença de semanas entre o primeiro sintoma e a notificação. Dentre as morbidades, destacam-se doenças cardiovasculares e outras morbidades não listadas. Outras variáveis a ser pontuadas são, internação no hospital (não confudir com internação na UTI) e a autodeclaração branca.

Figura 7 – Barplot para as dez variáveis que mais contribuem no modelo.



Fonte: Elaborado pelo autor, 2023.

Para observar a contribuição global das variáveis, não só em termos de impacto geral, mas também de sentido do impacto, faz-se necessário o gráfico *beeswarm*. Este, traz consigo informações a respeito da distribuição dos valores de SHAP, respectivos, a cada variável, além da visualização, por cores, dos valores da observação para a variável de interesse. Observe que, variáveis binárias se resumem a classificação de valores baixos/*low* (azul), quando os valores são "0"e altos/*high* (vermelho), quando os valores são "1".

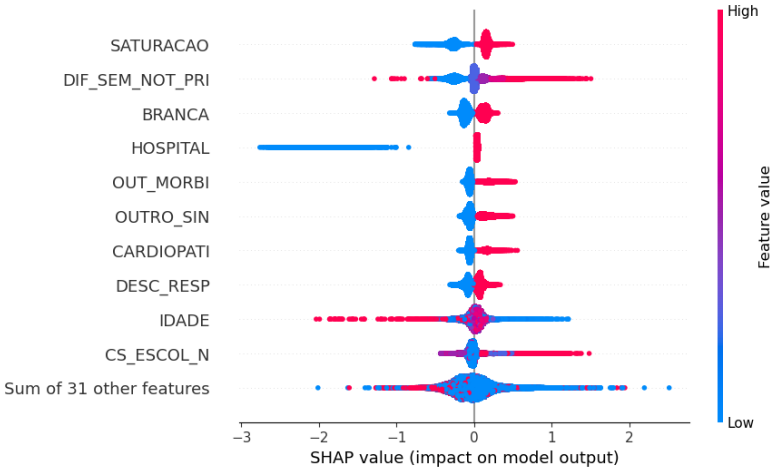
De acordo com a Figura 8 observa-se o gráfico *beeswarm* para as dez variáveis que mais contribuem para as predições. Vale ressaltar que, valores de SHAP positivo, aumentam a probabilidade logarítmica da internação na UTI, enquanto que os negativos, reduzem esta probabilidade. Para a variável de saturação, quando observada a distribuição dos valores de SHAP, apesar de terem impacto parecido para valor baixo e alto, com a diferença de que valor baixo tem impacto negativo e para valor alto ocorre o contrário, os impactos negativos dos valores baixo tendem a ser levemente maiores que os impactos dos valores altos.

Um comportamento diferente ocorre com a variável diferença de semanas entre o primeiro sintoma e a notificação, que por se tratar de uma variável numérica, é adicionada à escala a classificação de valores médios (roxo). De fato, valores médios para esta variável tem impacto em torno de zero nas previsões, entretanto, valores altos têm grande impacto positivo e maiores, quando comparados aos valores baixos, que impactam negativamente. Há ainda, alguns poucos valores altos que têm impacto negativo e razoavelmente grandes. Ainda sobre variáveis numéricas, a variável idade porta-se de outra maneira, em que idades menores impactam positivamente e idades elevadas impactam de forma negativa, além de idades medianas não impactarem tanto.

A variável hospital, que representa se o individuo ficou ou não internado no hospital, tem um papel importante no modelo, pois pode-se notar que, quando o individuo fica internado no hospital, ou seja, ocorre um valor alto, quase não tem impacto na predição feita pelo modelo, todavia, quando ocorre o contrario, ou seja, o individuo não fica internado, há um impacto

negativo muito alto na predição.

Figura 8 – Gráfico *beeswarm* para as dez variáveis que mais contribuem no modelo.



Fonte: Elaborado pelo autor, 2023.

3.3.2 Análise dos valores de SHAP para os fatores principais

Visto que, existem valores de SHAP para as observações em cada uma das variáveis, pressupõe-se pensar em uma analise descritiva desses valores, de forma que se possa verificar se a maior parte dessas contribuições se dão de forma negativa ou positiva. Uma das maneiras de realizar este estudo é, de forma separada, somando o total do valor de SHAP dos valores negativos e positivos, contando a quantidade desses valores e calculando as respectivas médias. Nas Tabelas 8 estão dispostas essas analises para as variáveis HOSPITAL, SATURACAO e CARDIOPATI, respectivamente. Este exercício de análise acrescenta bons *insights* às visualizações feitas no gráfico beeswarm anteriormente.

Observa-se então, para a variável HOSPITAL, que o impacto médio para as contribuições negativas, oriundas de casos em que o individuo não fica internado no hospital, é imensamente considerável, em relação as contribuições positivas. Além disso, este modelo dá uma ênfase muito maior para casos em que o individuo não fica internado no hospital. De acordo com o banco de dados, em 96% dos casos o individuo fica internado no hospital, e desses, 32% ficaram internados na UTI.

Tabela 8 – Análise dos valores de SHAP para as variáveis HOSPITAL, SATURACAO e CARDIOPATI.

Nome	Sentido do impacto	Soma dos valores de SHAP	Frequência	Média
HOSPITAL	Negativo	-13273.59	6799	-1.95
	Positivo	7598.151	183401	0.04
SATURACAO	Negativo	-21370.816	77804	-0.27
	Positivo	17947.912	112396	0.16
CARDIOPATI	Negativo	-8582.833	143189	-0.06
	Positivo	7770.1426	47011	0.16

Fonte: Elaborado pelo autor, 2023.

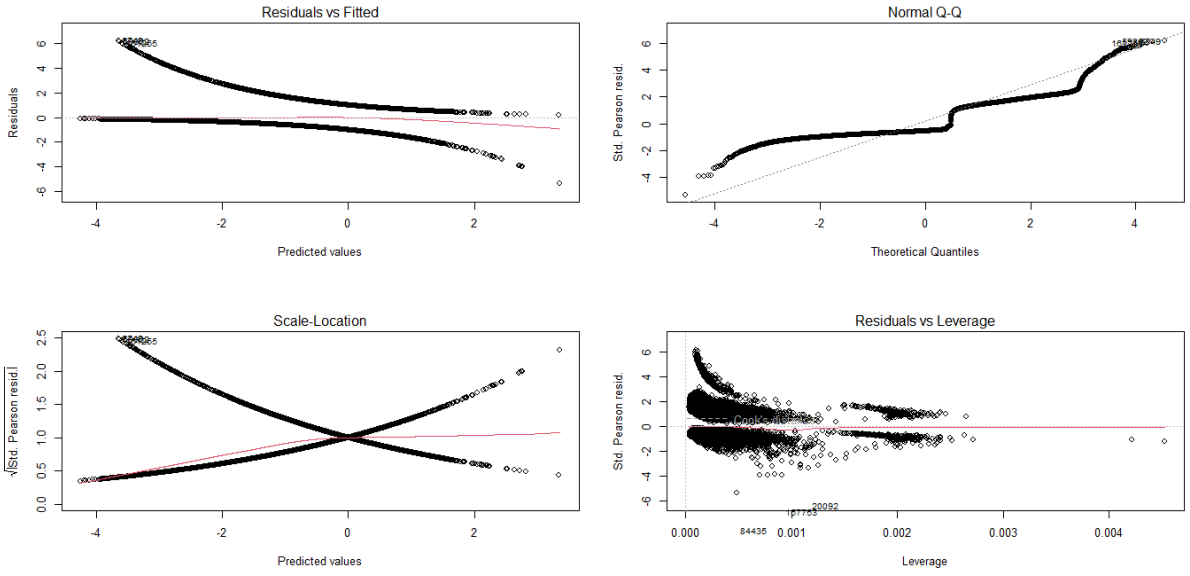
Para a variável SATURACAO, o impacto médio para as contribuições negativas, que advêm de casos em que não ocorre a saturação de O2<95%, é maior em relação as contribuições positivas. No geral, a saturação de O2<95% ocorre em 59% dos casos, e desses, 35% ficaram internados na UTI, segundo o banco de dados. Já na variável CARDIOPATI, o impacto médio para as contribuições positivas, que vem de casos em que existe a presença de doenças cardiovasculares, é maior em relação as contribuições negativas. Observando o banco de dados, essas doenças ocorrem em 25% dos casos, e desses, 38% ficaram internados na UTI. Em ambos os casos, pode-se dizer que os impactos são considerados parcimoniosos, pelo modelo, em relação a predição.

3.4 Regressão Logística

O modelo de regressão logística com função de ligação *logit* foi ajustado, utilizando a mesma amostra de dados de treino utilizada para estimar os valores de SHAP. Não houve resultados em relação a comparação de modelos, visto que os métodos de seleção de variável para o modelo não diferenciam muito em termos de resultados, pois se tratando de uma grande quantidade de dados, a exclusão e inclusão de variáveis não têm tanto efeito na diminuição do *residual deviance*, AIC ou BIC. Todavia, para maior facilidade de interpretação dos parâmetros, foram mantidas no modelo, as variáveis significativas ao nível de significância de 0,1% no teste de Wald, ou seja, as variáveis cujo o intervalo de confiança para o parâmetro β não contém o valor 1, ao nível de confiança de 99,9%, reduzindo, de 41 variáveis explicativas para 26.

Analisando os resíduos do modelo, através dos gráficos na Figura 9, inicialmente, pode-se pensar que não existe normalidade nos resíduos, possível tendência e problemas relacionados a constância na variância. Porém, boa parte desses problemas se dão devido ao tipo de variável resposta, fazendo com que esta primeira analise gráfica não tenha tanta relevância. Portanto, será feito o diagnóstico utilizando o gráfico semi-normal com envelope de simulação.

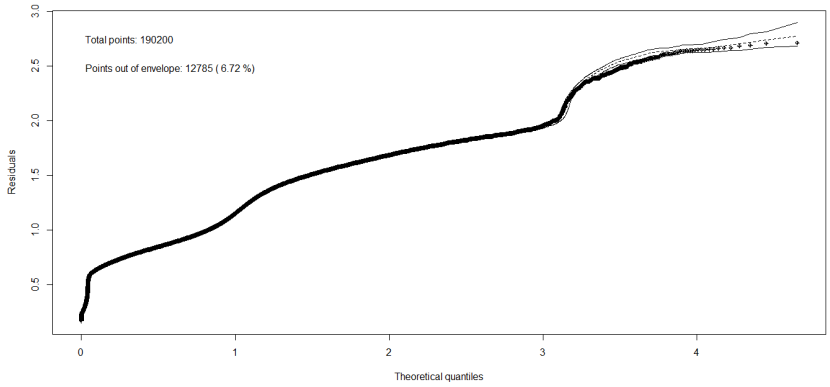
Figura 9 – Resíduos do modelo.



Fonte: Elaborado pelo autor, 2023.

Assim, de acordo com a Figura 10, a maioria resíduos se encontram no interior do envelope de simulação. Portanto, pode-se dizer que o modelo está bem ajustado.

Figura 10 – Gráfico semi-normal com envelope de simulação



Fonte: Elaborado pelo autor, 2023.

O modelo final, com $AIC = 224431$ e $BIC = 224705.7$, tem os seguintes parâmetros conforme a Tabela 9.

Tabela 9 – Modelo de Regressão Logística

β's	**Variáveis**	**Estimativa**	**Exp{Estim.}**	**Erro padrão**	**P-valor**
β_0	INTERCEPTO	-3.519	0.03	0.057	<2e-16
β_1	DIF_SEM_NOT_PRI	0.078	1.081	0.003	<2e-16
β_2	BRANCA	0.220	1.246	0.010	<2e-16
β_3	AMARELA	0.325	1.384	0.043	4.85e-14
β_4	CS_ESCOL_N	0.058	1.060	0.004	<2e-16
β_5	CS_ZONA	0.024	1.024	0.005	2.01e-7
β_6	DOR_ABD	-0.14	0.869	0.025	1.27e-8
β_7	FADIGA	-0.216	0.806	0.013	<2e-16
β_8	PERD_PALA	-0.239	0.787	0.026	<2e-16
β_9	FEBRE	0.077	1.08	0.011	3.09e-13
β_{10}	TOSSE	-0.118	0.889	0.011	<2e-16
β_{11}	GARGANTA	-0.102	0.903	0.016	2.24e-10
β_{12}	DISPNEIA	0.089	1.093	0.012	6.47e-14
β_{13}	DESC_RESP	0.146	1.157	0.012	<2e-16
β_{14}	SATURACAO	0.418	1.519	0.012	<2e-16
β_{15}	DIARREIA	-0.125	0.882	0.019	3.32e-11
β_{16}	OUTRO_SIN	0.228	1.256	0.011	<2e-16
β_{17}	CARDIOPATI	0.222	1.249	0.013	<2e-16
β_{18}	SIND_DOWN	0.414	1.513	0.088	2.43e-6
β_{19}	DIABETES	0.087	1.091	0.015	3.2e-9
β_{20}	PNEUMOPATI	0.147	1.158	0.026	1.82e-8
β_{21}	IMUNODEPRE	0.186	1.204	0.031	1.69e-9
β_{22}	RENAL	0.402	1.495	0.029	<2e-16
β_{23}	OBESIDADE	0.517	1.677	0.022	<2e-16
β_{24}	OUT_MORBI	0.265	1.303	0.012	<2e-16
β_{25}	ANTIVIRAL	0.167	1.182	0.034	6.97e-7
β_{26}	HOSPITAL	1.922	6.835	0.056	<2e-16

Fonte: Elaborado pelo autor, 2023.

3.4.1 Interpretações para β´s positivos

Aplicando a exponencial na estimativa dos β´s, é possível obter as razões de chance. Pode-se dizer que, em relação a variável HOSPITAL, a chance de uma pessoa internada no hospital ir para a UTI é 4.83 vezes maior que a chance de uma pessoa que não foi internada. Indivíduos com a saturação de O_2<95% têm 51.9% mais chance de ir para a UTI, em comparação com os indivíduos com a saturação de 0_2>95%. A presença de doença cardiovascular crônica aumenta em 24.9% a chance dos indivíduos irem para a UTI. Em relação a DIF_SEM_NOT_PRI, que representa a diferença de semanas entre o primeiro sintoma e a notificação, a cada semana de diferença a mais, a chance deste individuo ir para a UTI é um aumentada em 8,11%. Indivíduos com obesidade têm 67.7% mais chance de ir para a UTI, em relação a indivíduos sem obesidade.

3.4.2 Interpretações para β´s negativos

Alguns sintomas apresentam efeito contrário na variável resposta, ou seja, são fatores de proteção. Indivíduos com sintoma de tosse tem 11.1% menos chance de ir para a UTI, em comparação com os indivíduos sem este sintoma. A presença de fadiga diminui em 19.4% a chance do individuo ir para a UTI. Indivíduos com perda do paladar tem 21.3% menos chance de ir para a UTI, em relação a indivíduos sem este sintoma.

3.4.3 Comparação entre o Preditor Linear e o valor de SHAP

Para a variável HOSPITAL, comparando o valor estimado para β_{26}, com o valor de SHAP médio para os casos em que o individuo é internado no hospital, observa-se uma diferença muito discrepante, pois, neste caso, é adicionado 1,922 unidades ao preditor linear. Um contraponto a isto seria a probabilidade logarítmica que, neste caso, é aumentada em 0,04 apenas.

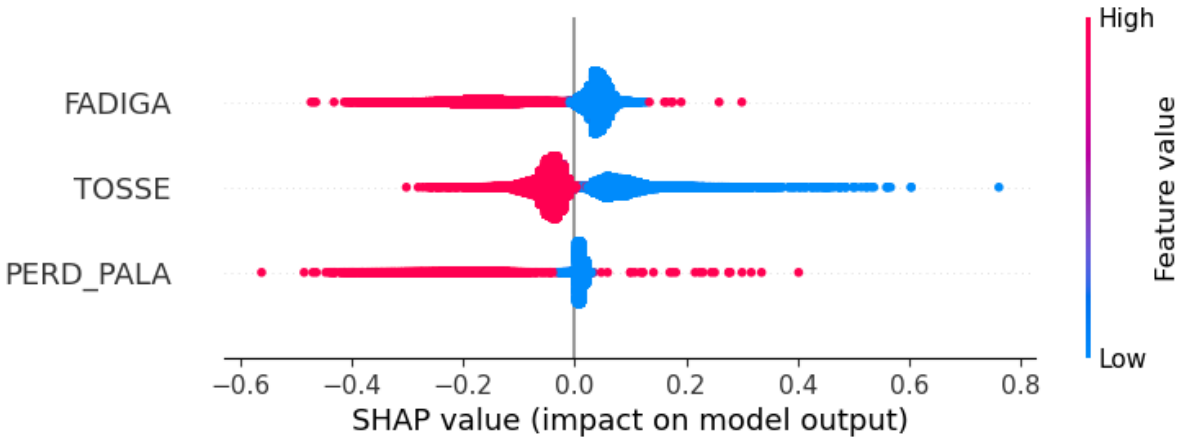
Observando os valores de SHAP da Figura 8, apesar de existirem alguns valores altos para diferença de semanas que impactam negativamente na predição e a maioria dos valores estarem distribuídos em torno de zero e baixos valores negativos, o que torna a média de contribuição de valores negativos (-0.17) ter mais impacto do que os valores positivos (0.12). Em geral, valores altos da diferença de semanas tendem a contribuir gradativamente de forma positiva na predição para indivíduos irem para a UTI, corroborando com o valor estimado para $\beta_1 = 0,078$.

O valor incrementado no preditor linear quando ocorre a saturação de O2<95% no individuo é de 0,418. Em contrapartida, o aumento na probabilidade logarítmica é, em média, de 0,159.

A presença de doença cardiovascular crônica no individuo reflete num aumento de 0,222 unidades no preditor linear. No valor de SHAP, o aumento na probabilidade logarítmica é, em média, de 0,165.

Algumas variáveis apresentaram efeito inverso no modelo de regressão logistica e no modelo de aprendizado de maquina, conforme observado na Figura 11.

Figura 11 – Gráfico beeswarm para variáveis que contribuem de forma inversa no modelo.



Fonte: Elaborado pelo autor, 2023.

Neste sentido, a presença de tosse no individuo tem uma diminuição de 0,118 unidades no preditor linear. Em contrapartida, a diminuição na probabilidade logaritmica é, em média, de 0.046. A presença de fadiga no individuo tem uma diminuição de 0,216 unidades no preditor linear. Em contrapartida, a diminuição na probabilidade logaritmica é, em média, de 0.153. A presença de perda do paladar no individuo tem uma diminuição de 0,239 unidades no preditor linear. Em contrapartida, a diminuição na probabilidade logaritmica é, em média, de 0.188.

Outra forma de entender a relação da contribuição das variáveis pelos valores de SHAP e as estimativas do preditor linear, para casos em que as variáveis são binárias, é compreendendo a contribuição como um efeito geral, particionado em impacto positivo e negativo. Somando o valor absoluto das médias dos impactos dos valores de SHAP, obtém-se valor próximo as estimativas dos parâmetros do preditor linear, conforme apresentado nas Tabelas 10 e 11.

O fato do efeito em um sentido ter, em média, um peso maior que o efeito do sentido contrário, pode ser interpretado como a importância dada pelo modelo para a ocorrência do evento associado a variável, levando em consideração a proporção da ocorrência desses fatores com a variável resposta (UTI). A maioria das variáveis listadas tem um maior peso no efeito associado a ocorrência do evento, ou seja, quando $x = 1$. Isto ocorre, inclusive, nas variáveis que funcionam como um fator de proteção.

Tabela 10 – Possíveis incrementos no preditor linear para os β's positivos e contribuição média para as probabilidades logarítmicas.

Variável	**Incremento Preditor linear**		**Valor de SHAP absoluto médio**		
	0 - Não	**1 - Sim**	**Negativo (0)**	**Positivo (1)**	**Soma**
HOSPITAL	0	1.922	**1.952**	0.041	1.993
SATURACAO	0	0.418	**0.275**	0.160	0.435
CARDIOPATI	0	0.222	0.060	**0.165**	0.225
OBESIDADE	0	0.517	0.025	**0.418**	0.443
BRANCA	0	0.220	0.114	**0.125**	0.239
OUT_MORBI	0	0.265	0.063	**0.197**	0.260
OUTRO_SIN	0	0.228	0.064	**0.152**	0.216
DESC_RESP	0	0.146	**0.085**	0.073	0.158

Fonte: Elaborado pelo autor, 2023.

Tabela 11 – Possíveis incremento no preditor linear para os β's negativos e contribuição média para as probabilidade logarítmica.

Variável	**Incremento Preditor linear**		**Valor de SHAP absoluto médio**		
	0 - Não	**1 - Sim**	**Negativo (1)**	**Positivo (0)**	**Soma**
TOSSE	0	-0.118	0.046	**0.082**	0.128
FADIGA	0	-0.216	**0.153**	0.042	0.195
PERD_PALA	0	-0.239	**0.188**	0.010	0.198
DOR_ABD	0	-0.140	**0.104**	0.005	0.109

Fonte: Elaborado pelo autor, 2023.

Para a variável HOSPITAL, apesar de, na maioria dos casos, os indivíduos não ficarem na UTI, é razoável pensar que este impacto na predição seja desproporcional a realidade. Este pode ser um dos motivos para que o modelo tenha problemas de sensibilidade ou, em outras palavras, baixa capacidade de predizer a internação na UTI quando isto seria o correto.

Para a variável TOSSE, apesar de ser um fator de proteção, o efeito positivo quando o individuo não apresenta este sintoma, em média, é maior que o efeito negativo quando ocorre o sintoma.

4 CONCLUSÃO

A contribuição para o preditor linear e a contribuição dos valores de SHAP à probabilidade logarítmica, podem ser parecidas se considerarmos um contexto de efeito geral numa análise descritiva e de interpretação. Ligado a isto, muitas possibilidades de análises veem a tona. O modelo linear generalizado pode funcionar como um medidor de ajuste para os modelos de aprendizado de maquina ou vice-versa. Ou ainda, pode auxiliar na melhoria do desempenho desses modelos, pois pode comparar diferentes tipos, bem como pode auxiliar na inclusão ou exclusão de variáveis. Portanto, os dois métodos podem ser utilizados juntos em um panorama de tomada de decisão.

Pode-se dizer que, há variáveis ou recursos que possuem a mesma importância em ambos os modelos, mas que diferem em termos de direção, pois podem impactar tanto positivamente, quanto negativamente. As variáveis FADIGA, CARDIOPATI e HOSPITAL, são exemplos deste caso já que, no efeito geral, são parecidas em ambos os modelos. Apesar disso, a desproporcionalidade no impacto da variável HOSPITAL, influencia muito na precisão do modelo de aprendizado de maquina, tornando-o com baixa precisão.

Considerando que existam dois modelos bem ajustados e com boa precisão, em ambos os métodos, tem-se um cenário em que o ponto de vista inferencial do MLG destacando o modelo de aprendizado de máquina pode trazer muitos *insights*, por exemplo, relacionado a saúde pública, *marketing* ou num contexto de *business intelligence* ou *business analytics*.

As internações nas UTI por síndrome respiratória grave podem ser melhor compreendidas com este tipo de análise, não só apenas do ponto de vista da previsão de novos casos, relacionando com a disponibilidade dos leitos hospitalares de UTI, mas também para campanhas extensivas contra a SRAG, visto que é possivel elencar os fatores que mais impactam na saúde dos indivíduos, bem como o entendimento de que esses temas estão relacionados, assim como é possivel relacionar as predições com as prescrições.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 18.

BLOM, G. *Statistical Estimates and Transformed Beta Variables*. [S.l.]: Wiley, 1958. Citado na página 19.

BROWNLEE, J. *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. 2020. Disponível em: <https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>. Acesso em: 25 de novembro 2023. Citado na página 22.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>. Citado 3 vezes nas páginas 10, 12 e 20.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP*, p. 31, 2008. Citado na página 14.

CORNFIELD, J. A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, Oxford University Press, v. 11, n. 6, p. 1269–1275, 1951. Citado na página 17.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: CRC press, 2018. Citado 3 vezes nas páginas 15, 16 e 17.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado na página 20.

GROVER, P. *Gradient Boosting from scratch*. 2017. Disponível em: <https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d>. Acesso em: 25 de novembro 2023. Citado 2 vezes nas páginas 20 e 21.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 12.

LIPOVETSKY, S.; CONKLIN, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, Wiley Online Library, v. 17, n. 4, p. 319–330, 2001. Citado na página 24.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>. Citado na página 12.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 23.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models (Second edition)*. [S.l.]: London: Chapman & Hall, 1989. 500 p. Citado na página 19.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61. Citado na página 12.

MOLNAR, C. *Interpretable Machine Learning*: A guide for making black box models explainable. 2. ed. [s.n.], 2022. Disponível em: <https://christophm.github.io/interpretable-ml-book>. Citado na página 23.

MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: ____. *Communications in Computer and Information Science*. Springer International Publishing, 2020. p. 417–431. ISBN 9783030659653. Disponível em: <http://dx.doi.org/10.1007/978-3-030-65965-3_28>. Citado na página 11.

MORAL, R. A.; HINDE, J.; DEMéTRIO, C. G. B. Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 2017. Citado 2 vezes nas páginas 13 e 19.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Disponível em: <http://www.jstor.org/stable/2344614>. Citado na página 10.

O'SULLIVAN, C. *SHAP for Binary and Multiclass Target Variables*. 2023. Disponível em: <https://towardsdatascience.com/shap-for-binary-and-multiclass-target-variables-ff2f43de0cf4>. Acesso em: 25 de novembro 2023. Citado na página 30.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP, 2004. Citado 4 vezes nas páginas 10, 14, 17 e 18.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 13.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>. Citado na página 13.

RATKOWSKY, D. Nonlinear regression modelling. New York, Marcel Dekker, 1983. Nenhuma citação no texto.

ROYSTON, J. Algorithm as 177: Expected normal order statistics (exact and approximate). *Journal of the royal statistical society. Series C (Applied statistics)*, JSTOR, v. 31, n. 2, p. 161–165, 1982. Citado na página 19.

SAHA, S. *Understanding the log loss function of XGBOOST*. 2018. Disponível em: <https://medium.datadriveninvestor.com/understanding-the-log-loss-function-of-xgboost-8842e99d975d>. Acesso em: 25 de novembro 2023. Citado na página 22.

SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, JSTOR, p. 461–464, 1978. Citado na página 19.

SHAPLEY, L. S. Notes on the n-person game—ii: The value of an n-person game. Rand Corporation, 1951. Citado na página 23.

WICKHAM, H.; BRYAN, J. *readxl: Read Excel Files*. [S.l.], 2023. Https://readxl.tidyverse.org, https://github.com/tidyverse/readxl. Citado na página 13.

APÊNDICE A – DICIONÁRIO DE VARIÁVEIS

Tabela 12 – Tabela com o dicionário das variáveis presentes no banco de dados

Nome	Tipo	Descrição	
IDADE	Numérica	Idade (anos)	
DIF_SEM_NOT_PRI	Numérica	Diferença de semanas entre primeiros sintomas e a semana de notificação	
CS_ESCOL_N	Numérica Ordinal	Nível de escolaridade do paciente	0: Ignorado/Não se aplica 1: Analfabeto/Sem escolaridade 2: Fundamental 1º ciclo 3: Fundamental 2º ciclo 4: Ensino Médio 5: Ensino Superior
CS_ZONA	Numérica Ordinal	Zona geográfica do endereço de residência do paciente	0: Ignorado/Vazio 1: Rural 2: Periurbano 3: Urbano
CS_GESTANT	Numérica Ordinal	Idade gestacional da paciente	0: Ignorado/Não se aplica 1: 1º Trimestre 2: 2º Trimestre 3: 3º Trimestre
UTI	Dicotômica	O paciente foi internado em UTI?	0: Não 1: Sim
NOSOCOMIAL		Caso de SRAG com infecção adquirida após internação	
AVE_SUINO		Caso com contato direto com aves ou suínos	
VACINA_COV		Recebeu vacina COVID-19?	
ANTIVIRAL		Fez uso de antiviral para tratamento da doença?	
HOSPITAL		Paciente foi internado no hospital?	
PUERPERA	Dicotômica	Paciente é puérpera ou parturiente?	0: Não 1: Sim
CARDIOPATI		Paciente possui Doença Cardiovascular Crônica?	
HEMATOLOGI		Paciente possui Doença Hematológica Crônica?	
SIND_DOWN		Paciente possui Síndrome de Down?	
HEPATICA		Paciente possui Doença Hepática Crônica?	
ASMA		Paciente possui Asma?	
DIABETES		Paciente possui Diabetes *mellitus*?	
NEUROLOGIC		Paciente possui Doença Neurológica?	
PNEUMOPATI		Paciente possui outra pneumopatia crônica?	
IMUNODEPRE		Paciente possui Imunodeficiência ou Imunodepressão?	
RENAL		Paciente possui Doença Renal Crônica?	
OBESIDADE		Paciente possui obesidade?	
OUT_MORBI		Paciente possui outro(s) fator(es) de risco?	
DOR_ABD		Paciente apresentou dor abdominal?	
FADIGA		Paciente apresentou fadiga?	
PERD_OLFT		Paciente apresentou perda do olfato?	

PERD_PALA		Paciente apresentou perda do paladar?	
FEBRE		Paciente apresentou febre?	
TOSSE		Paciente apresentou tosse?	
GARGANTA		Paciente apresentou dor de garganta?	
DISPNEIA		Paciente apresentou dispneia?	
DESC_RESP		Paciente apresentou desconforto respiratório?	
SATURACAO		Paciente apresentou saturação O_2<95%?	
DIARREIA		Paciente apresentou diarreia?	
VOMITO		Paciente apresentou vômito?	
OUTRO_SIN		Paciente apresentou outro(s) sintoma(s)?	

Fonte: Elaborado pelo autor, 2023.