Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

ESTUDO DE ATRASO DE VOOS NO AEROPORTO DE UBERLÂNDIA VIA REGRESSÃO LOGÍSTICA

Aline Gonçalves Silva

Uberlândia-MG
2020

Aline Gonçalves Silva

ESTUDO DE ATRASO DE VOOS NO AEROPORTO DE UBERLÂNDIA VIA REGRESSÃO LOGÍSTICA

Trabalho de conclusão de curso de graduação apresentado à Faculdade de Matemática da Universidade Federal de Uberlândia (UFU) como requisito parcial para a obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Janser Moura Pereira

Uberlândia-MG
2020

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, 18 de Novembro de 2020

BANCA EXAMINADORA

Prof. Dr. Janser Moura Pereira

Prof. Dr. Lúcio Borges de Araújo

Prof. Dr. José Waldemar da Silva

Uberlândia-MG
2020

" Dedico esse trabalho aos meus pais que não mediram esforços para que eu pudesse completar mais essa etapa da minha vida. Aos meus irmãos e a meu esposo."

AGRADECIMENTOS

Agraço a Deus por essa conquista, pela força e energia para concluir o curso.

Agradeço a minha mãe Denise e ao meu pai Evando que me deram força e todo apoio necessário no decorrer do curso e não me deixaram desistir nunca.

Agradeço aos meus irmãos, Jalles e Joel por sempre me incentivar.

Agradeço ao meu esposo Edirley que compartilhou comigo esse momento, me ensinou no decorrer do curso e me incentivou para o desenvolvimento desse trabalho.

Agradeço aos meus professores, em especial ao meu orientador Janser Moura Pereira, pelo apoio, incentivo, ensinamento e paciência.

Agradeço a Universidade Federal de Uberlândia (UFU) pela oportunidade e possibilitar a aprendizagem e conclusão deste trabalho.

E a todos, que direta ou indiretamente fizeram parte da minha formação.

"Bem aventurado o homem que acha sabedoria, e o homem que adquire conhecimento."

Pv. 3.13

RESUMO

O presente trabalho tem o objetivo avaliar quais os fatores que influenciam o tempo de atraso dos voos no aeroporto de Uberlândia de uma determinada companhia aérea por meio de regressão logística. Os dados foram coletados no setor administrativo de chegada e decolagem de aeronaves de uma companhia aérea que pousa no aeroporto. Foram analisados mais de 3.121 voos no período de maio de 2017 a abril de 2019. Analisou-se 8 variáveis, que são possíveis motivos do tempo das aeronaves em solo que podem causar o atraso nos voos. Observou-se que entre as variáveis estudadas que 4 foram realmente significativas em relação a variável resposta. Utilizou-se para selecionar o modelo mais parcimonioso o critério de Stepwise com a estatística de Akaike e para aferir a qualidade do ajuste do modelo aplicou-se a matriz de classificação, a Curva ROC para organizar, visualizar e classificar o modelo com base em sua performance preditiva e Pseudo R^2 de Nagelkerke.

Palavras-chave: *Logit*, Inferência, Stepwise.

ABSTRACT

The current work aims to evaluate, using logistic regression, which factors have influence on the flights delay at Uberlândia airport, concerning a certain airline. The datas were collected on the aircraft arrival and departure administrative sector of an airline that lands at the airport. More than 3,121 flights were analyzed from May 2017 to April 2019. 8 variables were analyzed, which are possible reasons for aircraft time on the ground that can cause flight delays. It was observed, among the studied variables, there were 4 really significant, concerning the response variable. The Stepwise criterion with the Akaike statistic was used to select the most parsimonious model and to measure the quality of the model fit. The classification matrix, the ROC Curve, was applied to organize, visualize and classify the model based on its predictive performance and Pseudo R^2 of Nagelkerke.

Keywords: *Logit*, Inference, Stepwise.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Atrasos, cancelamentos e interrupções de voos são muitas vezes inevitáveis pelas companhias aéreas, podem ser decorrentes de fatores climáticos, problemas não programados nas aeronaves, tráfego aéreo ou até mesmo falha no sistema na prestação do serviço (ANAC, 2018).

Na aviação, um pequeno atraso pode influenciar nos demais voos e virar uma cadeia de atrasos devido a utilização intensa das aeronaves. Em 2001 essa tendência ficou evidente com a entrada de empresas de baixo custo no mercado brasileiro, e dessa forma as empresas buscaram uma forma de aumentar o uso da frota, assim qualquer imprevisto pode provocar atraso, gerando altos custos para as empresas e também insatisfação dos clientes (MELO, 2016).

Voos atrasados têm se mostrado bastante relevante no cenário mundial, independente da grandeza da malha aérea. Em 2013, 36% dos voos atrasaram mais de cinco minutos na Europa, 31.1% dos voos atrasaram mais de 15 minutos nos Estados Unidos e 16.3% dos voos foram cancelados ou sofreram atrasos maiores que 30 minutos no Brasil (EUROCONTROL, 2015).

Em função dos atrasos, as companhias aéreas voltam a atenção para os motivos que os provocam, pois as causas são visíveis e mensuráveis. Dada a importância do tema, serão utilizados dados de uma empresa que atua no aeroporto de Uberlândia, cujo propósito será identificar as causas mais decorrentes e os maiores impactos para a empresa aérea e passageiros que utilizam o aeroporto. O terminal tem capacidade de receber 2,4 milhões de passageiros por ano, sendo que, em 2017, transitaram 1,1 milhão de passageiros pelo local, média diária de 2 mil passageiros. Portanto pretende-se nesse trabalho avaliar os fatores ou causas que influenciam o tempo de atraso dos voos (INFRAERO, 2017).

A seguir destacam-se os objetivos geral e específicos. O objetivo geral da pesquisa consiste em avaliar os motivos que mais afetam os atrasos no terminal aeroportuário em Uberlândia. Como consequência tem-se os seguintes objetivos específicos: verificar se o tempo de atraso tem maior impacto em função: do setor responsável da operação em solo; da quantidade de pessoas embarcadas; do turno de trabalho; do tipo de aeronave; do

período de maior e menor demanda durante o ano. Além disso, avaliar a probabilidade de ocorrer atrasos em função dos fatores mencionados.

Dessa forma iremos identificar quais fatores mais influenciam a ocorrência dos atrasos. Nesse contexto, os modelos de regressão logística, são uma solução viável que permite identificar as variáveis que contribuem para esses acontecimentos, o que torna esses modelos uma ferramenta que permite auxiliar os gestores da área a tomar decisões e melhorar o desempenho da companhia a fim de diminuir a incidência de atraso.

2 MATERIAIS E MÉTODOS

Os dados foram coletados no setor administrativo de pouso e decolagem de aeronaves de uma companhia aérea que utiliza o aeroporto de Uberlândia. Foram analisados 3.121 voos no período de maio de 2017 a abril de 2019. A base de dados consiste em informações de todos os voos no período de maio de 2017 a abril de 2019. Consideramos Y como a variável dependente do modelo. Nesta base de dados foram coletadas as seguintes variáveis:

Y: Atraso, recebe valor 0 para voos sem atraso e valor 1 para voos com atraso. A companhia aérea considera atraso a partir de 1 minuto que excede o horário marcado para a saída do voo;

X_1: Passageiros, quantidade de passageiros e tripulantes extras embarcados na aeronave;

X_2: Área Responsável (três níveis). Denotada por: "AR1", recebe valor 1 para atrasos motivados pela área ar (no aeroporto a área ar é onde fica os aviões, um atraso do tipo manutenção ou meteorologia são ocorridas na parte ar do aeroporto) e valor 0, caso contrário; "AR2", recebe valor 1 para atrasos ocorridos na parte terra (a parte terra do aeroporto são as área comuns, comercial, como check-in, embarque, onde pode ocorrer problemas com o sistema ou até mesmo passageiro indisciplinado provocando atraso) e valor 0, caso contrário;

X_3: Equipamento, denotada por "Equipamento1", caracteriza o tamanho e modelo de aeronave que vai operar aquele voo, recebe valor 0 para aeronaves A319 com capacidade de 144 passageiros e recebe valor 1 as aeronaves A320 com capacidade de 174 passageiros a bordo sem contar a tripulação, nesta variável é incluída a quantidade de tripulantes a bordo;

X_4: Turno em que ocorreu o voo (manhã, tarde ou noite). Denotada por: "TurnoNoite" recebe valor 1 se o voo ocorreu no turno da noite e recebe valor 0, caso contrário; "TurnoTarde" recebe o valor 1 se o voo ocorreu no turno da tarde e recebe valor 0, caso contrário. Considera-se como turno da manhã, o período de 6h as 12h, turno noite o período entre 18h e 24h, turno da tarde o período entre 12h e 18h;

X_5: Ano em que foi coletado os dados 2017, 2018 ou 2019. Denotada por "Ano2018" e "Ano2019". Em que: "Ano2018" recebe valor 1 se o voo ocorreu no ano de 2018 e recebe valor 0, caso contrário; "Ano2019" recebe valor 1 se o voo ocorreu no ano de 2019 e recebe valor 0, caso contrário.

X_6: Mês, refere-se ao mês do ano em que ocorreu aqueles voos. Denotada por: "Mes1" recebe valor 0 para os meses de março a outubro, recebe valor 1 para os meses de novembro a fevereiro, onde são os meses mais chuvosos do ano, o que pode causar atrasos meteorológicos, e são meses de férias para os brasileiros o que acarreta aeronaves com a capacidade máxima ou próxima disso;

X_7: Agentes, quantidade de agentes fazendo o embarque, recebe valor 0 para dois ou mais agentes no portão de embarque, 1 para um agente no portão de embarque agilizando todo o processo.

Analisou-se também a variável tempo de atraso, representada em minutos, o tempo que a aeronave ficou em solo até que se resolvesse todas as adversidades para que o voo decolasse com segurança. Está variável não foi considerada no ajuste do modelo de regressão logística.

Usou-se como metodologia de pesquisa a regressão logística para analisar as variáveis para ver se tem relação com a variável resposta.

2.1 CONTEXTUALIZAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA

2.1.1 Regressão logística simples

A regressão logística é uma técnica estatística utilizada para descrever o comportamento entre uma variável binária e variáveis independentes métricas ou não métricas. Ou seja, destina-se a investigar o efeito das variáveis pelas quais os indivíduos, objetos ou sujeitos estão expostos sobre a probabilidade de ocorrência de determinado evento de interesse (FÁVERO, 2009).

Considere o modelo de regressão linear simples conforme equação (1) (FIGUEIRA, 2006):

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (1)$$

em que $Y_i = 0$ ("fracasso") ou $Y_i = 1$ ("sucesso").

Para Figueira (2006), em problemas de regressão modela-se a média condicional, que é o valor médio da variável resposta Y dado os valores da variável independente, x_i, designada por $E(Y_i \mid X = x_i)$, cujo valor será:

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_i, \qquad (2)$$

Já que Y_i assume dois resultados possíveis pode-se afirmar que Y_i é uma variável aleatória com distribuição Bernoulli (FIGUEIRA, 2006):

$$Y_i = 1 \rightarrow P(Y_i = 1|x_i) = \pi_i \qquad (3)$$

$$Y_i = 0 \rightarrow P(Y_i = 0|x_i) = 1 - \pi_i \qquad (4)$$

Pela definição de esperança matemática tem-se que (FIGUEIRA, 2006),

$$E(Y_i|x_i) = \pi_i \qquad (5)$$

Igualando (2) e (5) tem-se,

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_i = \pi_i = \pi(x_i) \qquad (6)$$

Então, pode-se afirmar que $Y_i = \pi_i + \varepsilon_i$, em que a quantidade ε_i admite o valor $1 - \pi_i$ para $Y_i = 1$, ou $-\pi_i$ para 0, assim ε_i segue uma distribuição Bernoulli com média zero e variância igual a $\pi_i(1-\pi_i)$, que podem ser verificadas. Sendo $P(\varepsilon_i = -\pi_i) = 1 - \pi_i$ e $P(\varepsilon_i = 1 - \pi_i) = \pi_i$, então o valor esperado de ε_i é (FIGUEIRA, 2006):

$$E(\varepsilon_i) = -\pi_i(1 - \pi_i) + (1 - \pi_i)\pi_i = -\pi_i + {\pi_i}^2 + \pi_i - {\pi_i}^2 = 0.$$

E a variância (PAGANO & GAUVREAU, 2008):

$$\begin{aligned} Var(\varepsilon_i) &= E(\varepsilon_i^2) - \left[E(\varepsilon_i)\right]^2 = E(\varepsilon_i^2) = (-\pi_i)^2(1-\pi_i) + (1-\pi_i)^2\pi_i \\ &= (-\pi_i)^2(1-\pi_i) + (1-\pi_i)^2\pi_i = \pi_i^2 - \pi_i^3 + (1 - 2\pi_i + \pi_i^2)\pi_i \\ &= \pi_i^2 - \pi_i^3 + \pi_i - 2\pi_i^2 + \pi_i^3 = \pi_i - \pi_i^2 \\ &= \pi_i(1-\pi_i). \end{aligned}$$

Por (PAGANO & GAUVREAU, 2008) quando a variável resposta é binária, assume o valor 1 (um) para representar o "sucesso" ou o "evento de interesse" ou assume o valor 0 (zero) para representar o "fracasso", a resposta média representará a probabilidade de Y_i ser igual a 1, ao nível da variável preditora x_i. Sendo assim, a princípio poderia considerar um modelo da seguinte forma,

$$\pi_i = \pi(x_i) = \beta_0 + \beta_1 x_i. \qquad (7)$$

Porém, como π_i é uma probabilidade, seus valores variam entre 0 e 1. Logo, $0 \leq \beta_0 + \beta_1 x_i \leq 1$. No entanto, é uma restrição inapropriada para um modelo linear, que assume valores reais, isto é, no intervalo $(-\infty, \infty)$. Uma alternativa seria (PAGANO & GAUVREAU, 2008):

$$\pi_i = \pi(x_i) = e^{\beta_o + \beta_1 x_i} \qquad (8)$$

A expressão (8) assegura que π_i seja positiva, mas o termo $e^{\beta_o + \beta_1 x_i}$ pode gerar um valor maior que 1. Assim, para satisfazer essa restrição, ajusta-se um modelo do tipo (PAGANO & GAUVREAU, 2008):

$$\pi_i = \pi(x_i) = \frac{e^{\beta_o + \beta_1 x_i}}{1 + (e^{\beta_o + \beta_1 x_i})} \qquad (9)$$

O termo acima, chamado de função logística, não gera valores negativos ou maiores que 1, delimitando assim, o valor estimado de π_i de acordo com o intervalo requerido (PAGANO & GAUVREAU, 2008).

Para (PAGANO & GAUVREAU, 2008) uma propriedade interessante é que a função logística pode ser linearizada por meio da transformação logarítmica. Mas, primeiramente deve-se conhecer o conceito de chance. Os autores definem como chance, a probabilidade de ocorrência de um evento dividida pela probabilidade da não ocorrência do mesmo evento. A transformação é interpretada como o logaritmo da chance entre π_i e $1 - \pi_i$ (chance do sucesso). Os autores afirmam que esta interpretação é bastante empregada em estudos toxicológicos, epidemiológicos e em outras áreas, sendo definida como:

$$g(x_i) = ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = ln\left[\frac{\frac{e^{\beta_o + \beta_1 x_i}}{1 + (e^{\beta_o + \beta_1 x_i})}}{\frac{1}{1 + (e^{\beta_o + \beta_1 x_i})}}\right] = ln[e^{\beta_o + \beta_1 x_i}] = \beta_o + \beta_1 x_i. \qquad (10)$$

Com isso, em vez de assumir que a relação entre π_i e x_i é linear, assume-se que a relação entre $\ln\left[\pi(x_i)/(1-\pi(x_i))\right]$ e x_i é linear. Essa técnica é conhecida como Regressão Logística (CORRAR *et al.*, 2009).

A distribuição acumulada de probabilidade tem sido utilizada para a obtenção de modelos para a média condicional de *Y* dado *x* quando as variáveis resposta são dicotômicas. Embora muitas funções de distribuição de probabilidade têm sido propostas para darem suporte a modelos estatísticos de respostas binárias, a distribuição logística de probabilidade continua imperando nas escolhas para modelagem devido a sua extrema flexibilidade e fácil utilização, isso do ponto de vista matemático, assim como sua

capacidade de proporcionar interpretações ricas em significados práticos (HOSMER & LEMESHOW, 2000).

A probabilidade de ocorrência de um evento de interesse na regressão logística é expressa da seguinte forma (HOSMER & LEMESHOW, 2000):

$$\pi_i = \pi(x_i) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (11)$$

e a probabilidade de fracasso (HOSMER & LEMESHOW, 2000):

$$1-\pi_i = 1-\pi(x_i) = P(Y_i = 0|X = x_i) = \frac{1}{1+(e^{\beta_o+\beta_1 x_i})} \qquad (12)$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ é o vetor de parâmetros.

2.1.2 Estimação dos Parâmetros

Para estimar os parâmetros β_0 e β_1 do modelo de regressão logística simples utiliza-se o método da máxima verossimilhança. Os estimadores de máxima verossimilhança de β_0 e β_1 são os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximizam o logaritmo da função de verossimilhança. Considere uma amostra independente (x_i, y_i) de tamanho *n,* em que x_i é o valor da variável independente da *i-ésima* observação, com $i = 1, \dots, n$ e y_i é o valor da variável resposta dicotômica. Sendo que $Y_i \sim Ber(\pi_i)$, a distribuição de probabilidade de Y_i é dada por (ABREU, 2004):

$$f(y_i, \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad (13)$$

Como as observações y_i são independentes, a função de máxima verossimilhança é dada por:

$$\prod_{i=1}^{n} f(y_i, \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, y_i \in [0,1]$$

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \qquad (14)$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ é o vetor de parâmetros.

Por meio da máxima verossimilhança o que se deseja é determinar os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximiza $L(\boldsymbol{\beta})$. Logo, aplicando o logaritmo na expressão (13) tem-se (ABREU, 2004):

$$l(\boldsymbol{\beta}) = ln(\, L(\boldsymbol{\beta})) = ln\left(\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} [y_i \, ln(\, \pi_i) + (1-y_i) \, ln(\, 1-\pi_i)]$$

$$= \sum_{i=1}^{n} \left[y_i \, ln\left(\frac{\pi_i}{1-\pi_i}\right) + ln(\, 1-\pi_i)\right] \quad (15)$$

Substituindo as equações (11) e (12) em (15) tem-se:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\beta_o + \beta_1 x_i) + ln\left(\frac{1}{1+exp(\,\beta_o + \beta_1 x_i)}\right)\right]$$

$$= \sum_{i=1}^{n} [y_i(\beta_o + \beta_1 x_i) - ln(1 + exp(\,\beta_o + \beta_1 x_i))] \quad (16)$$

Derivando a expressão (16) em relação a cada parâmetro tem-se:

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \frac{1}{1+exp(\beta_o+\beta_1 x_i)}(\beta_o + \beta_1 x_i)\right] \quad (17)$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} \left[y_i x_i - \frac{1}{1+exp(\beta_o+\beta_1 x_i)}(\beta_o + \beta_1 x_i) x_i\right] \quad (18)$$

No entanto, não existe solução analítica para os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros β_0 e β_1 (PAGANO & GAUVREAU, 2008). Esse procedimento é feito por meio de recursos computacionais, e o método iterativo utilizado pelo *software* R é o de Newton-Raphson.

O valor ajustado para o *i-ésimo* valor é dado por:

$$\pi_i = \frac{e^{\hat{\beta}_o + \hat{\beta}_1 x_i}}{1+\left(e^{\hat{\beta}_o + \hat{\beta}_1 x_i}\right)} \quad (19)$$

A função resposta ajustada é dada por:

$$\pi = \frac{e^{\hat{\beta}_o + \hat{\beta}_1 x}}{1+\left(e^{\hat{\beta}_o + \hat{\beta}_1 x}\right)} \quad (20)$$

Vale ressaltar que o método de estimação de máxima verossimilhança, é uma das alternativas para estimação dos parâmetros.

2.1.3 Regressão logística múltipla

Analogamente, obtêm-se os estimadores dos parâmetros para um modelo de regressão logística múltipla. Até o momento foi apresentado a contextualização da

regressão logística de forma mais simples e conceitual, isto é, considerou-se apenas uma variável explicativa. Contudo é importante a visualização de um modelo mais completo quando se utiliza mais variáveis independentes na modelagem (RODESKI, 2010).

Em termos conceituais a diferença reside apenas no acréscimo de p variáveis ao modelo, o que conduz a algumas sutis adaptações na formulação matemática do modelo de regressão simples, e que podem ser vistas como segue (RODESKI, 2010):

$$g(X) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad (21)$$

em que $\boldsymbol{X} = (x_1, x_2, \ldots, x_p)$,

Do mesmo modo, as demais expressões antes apresentadas para o caso do modelo de regressão logística simples seguem valendo para o modelo de regressão logística múltipla, onde a notação $\boldsymbol{\beta' x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ passa a ser inserida. Além do mais, a probabilidade de *sucesso* no caso modelo múltiplo poderá ser obtida através da expressão (RODESKI, 2010):

$$\pi(\boldsymbol{x}) = \frac{e^{\beta' x}}{1+e^{\beta' x}} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \ . \qquad (22)$$

A notação $\boldsymbol{\beta'}$ expressa o vetor transposto de $\boldsymbol{\beta}$.

Em se tratando dos modelos lineares, os parâmetros são estimados pelo Método dos Mínimos Quadrados ou Método de Máxima Verossimilhança (PAGANO, 2008). Quando o Método dos Mínimos Quadrados é utilizado em modelos com desfecho dicotômico, os estimadores não apresentam as propriedades estatísticas desejáveis. Por esta razão, utiliza se o Métodos da Máxima Verossimilhança, que produz valores para os parâmetros desconhecidos que minimizam a probabilidade de obtenção dos conjuntos de dados observados (HOSMER & LEMESHOW,2000). Além do Método de Máxima Verossimilhança, pode-se utilizar também o método de estimação via Inferência Bayesiana (Método Bayesiano).

O logaritmo da função de verossimilhança, para estimar os parâmetros do modelo de regressão logística múltipla, passa a ser escrito como (HOSMER & LEMESHOW,2000):

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \boldsymbol{x_i^T \beta} - ln(\,1 + exp(\,\boldsymbol{x_i^T \beta}))] \qquad (23)$$

2.1.3.1 Razão de Chances (*Odds Ratio*)

A Odds Ratio (OR) é uma razão entre chances, e a chance é a probabilidade de que um evento ocorra dividido pela probabilidade de que ele não ocorra (BEZERRA, 2012). É obtida pela comparação de indivíduos que diferem apenas na variável de interesse e que apresentam outras características constantes (HORMER & LEMESHOW, 2000).

No caso em que a variável independente também é dicotômica, ou seja, $x = 0$ e $x = 1$, então, a chance de resposta quando $x = 1$ é dada por $\frac{\pi(1)}{(1-\pi(1))}$ e quando $x = 0$ é dada por $\frac{\pi(0)}{(1-\pi(0))}$. O logaritmo da chance é dado por (CORRAR *et al.*, 2009):

$$g(1) = ln\left[\frac{\pi(1)}{1-\pi(1)}\right] \qquad (24)$$

$$g(0) = ln\left[\frac{\pi(0)}{1-\pi(0)}\right] \qquad (25)$$

Na Tabela 1 é apresentado o resumo das probabilidades de sucesso e fracasso referente à situação quando a variável independente é dicotômica.

Tabela 1 - Resumo das probabilidades de sucesso e fracasso referente à situação quando a variável independente é dicotômica.

Variável resposta (Y)	**Variável independente (X)**	
	$x = 1$	$x = 0$
$Y = 1$	$\pi(1) = \frac{e^{\beta_o+\beta_1}}{1+(e^{\beta_o+\beta_1})}$	$\pi(0) = \frac{e^{\beta_o}}{1+(e^{\beta_o})}$
$Y = 0$	$1-\pi(1) = \frac{1}{1+(e^{\beta_o+\beta_1})}$	$1-\pi(0) = \frac{1}{1+(e^{\beta_o})}$
Total	1,0	1,0

Dessa forma, pode-se definir razão das chances ("odds ratio"), cuja finalidade é identificar a probabilidade associada à ocorrência de determinado evento, como (CORRAR *et al.*, 2009):

$$\psi = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} = \frac{\frac{\frac{e^{\beta_o+\beta_1}}{1+(e^{\beta_o+\beta_1})}}{\frac{1}{1+(e^{\beta_o+\beta_1})}}}{\frac{\frac{e^{\beta_o}}{1+(e^{\beta_o})}}{\frac{1}{1+(e^{\beta_o})}}} = \frac{e^{\beta_o+\beta_1}}{e^{\beta_o}} = e^{\beta_1} \qquad (26)$$

Então, o logaritmo da expressão (26) é dado por:

$$ln(\psi) = ln\left(e^{\beta_1}\right) = \beta_1 \qquad (27)$$

A *odds ratio* pode ser interpretada como o aumento/decréscimo esperado na probabilidade de sucesso devido à uma mudança de uma unidade no valor da variável preditora. A interpretação dos coeficientes de regressão para o caso de um modelo de regressão logística múltipla é similar ao caso em que o modelo tem apenas um único regressor. Nestes casos, a quantidade de $\exp(\hat{\beta}_j)$ é a *odds ratio* por regressor x_j, assumindo que todas as outras várias preditivas são constantes (MONTGOMERY; PECK; VINING, 2001).

2.1.4 Avaliador da qualidade do ajuste

Após obter as estimativas dos coeficientes da regressão logística, é necessário avaliar a qualidade do modelo ajustado. Para o ajuste do modelo, utilizou-se o critério de Stepwise para a seleção de variáveis, tendo como referência o critério de informação de Akaike (*AIC*) para inclusão ou exclusão da variável. Além disso, a significância dos parâmetros dos modelos foi avaliada pelo teste de Wald.

O critério de informação de Akaike (*AIC*) é um método de seleção de modelos. O *AIC* foi desenvolvido por meio dos estimadores de máxima verossimilhança (EMV), para decidir qual o modelo mais adequado quando se utiliza muitos modelos com quantidades diferentes de coeficientes (Sobral & Barreto, 2011). A decisão quanto ao melhor modelo ajustado é realizada escolhendo o menor valor de *AIC*. Sobral & Barreto (2011) define *AIC* como:

$$AIC = -2l(\boldsymbol{\beta}) + 2k \qquad (28)$$

em que $l(\boldsymbol{\beta})$ é o logaritmo da função de verossimilhança do modelo e k é o número de parâmetros.

Para Corrar, Paulo e Dias Filho (2009), a finalidade da estatística Wald é testar o grau de significância de cada coeficiente da equação logística, inclusive a constante, ou seja, verificar se cada parâmetro estimado é significativamente diferente de zero. O teste

Wald assim como o teste da razão de verossimilhança depende da estimativa de máxima verossimilhança dos parâmetros β_j´s (MEZZOMO, 2009).

Hosmer & Lemeshow (2000) destacam que o teste Wald é obtido pela comparação da estimativa da máxima verossimilhança do parâmetro de inclinação β_j em relação à estimativa do seu erro padrão ($SE_{\beta j}$).

Portanto, no presente trabalho foi adotado o teste de Wald, que sob a hipótese H_0: $\beta_j = 0$, a estatística abaixo segue uma distribuição normal padrão (HOSMER & LEMESHOW, 2000):

$$Wald = \frac{\hat{\beta}_j}{S\hat{E}\left(\hat{\beta}_j\right)} \qquad (29)$$

em que: β_j é o coeficiente de uma variável independente incluída no modelo e $\hat{\beta}_j$ representa a estimativa desse coeficiente; $SE\left(\beta_j\right)$ é o erro padrão, também conhecido como desvio padrão do coeficiente de β_j e $S\hat{E}\left(\hat{\beta}_j\right)$, representa a estimativa do desvio padrão desse coeficiente. A partir da matriz de informação de Fisher, $\boldsymbol{I}(\boldsymbol{\beta})$, determina-se a matriz de variâncias e covariâncias dos coeficientes da seguinte forma: $Var(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$. A variância de β_j, denotada por $Var(\beta_j)$, corresponde ao *j-ésimo* elemento da diagonal principal da matriz variâncias e covariâncias dos coeficientes estimados. Logo, o desvio padrão do coeficiente de β_j é:

$$SE\left(\beta_j\right) = \sqrt{Var\left(\beta_j\right)} \qquad (30)$$

Mais detalhes sobre a matriz de informação de Fisher são apresentados em (HOSMER & LEMESHOW, 2000).

2.1.5 Capacidade preditiva do modelo

Um dos objetivos no ajuste de modelos de regressão logística é a predição. No entanto, é importante que o modelo tenha um ótimo poder de discriminação. Para avaliar a capacidade preditiva do modelo foi utilizado matriz de classificação, curva ROC e o Pseudo R^2.

2.1.5.1 Matriz de Classificação

Uma maneira prática de qualificar a capacidade preditiva do modelo de regressão logística é pela projeção do modelo na tabela de classificação, também conhecida como matriz de confusão. Para isto, precisa-se criar uma tabela com o resultado da classificação cruzada da variável resposta, de acordo com uma variável dicotômica em que os valores se derivam das probabilidades logísticas estimadas na regressão (HOSMER & LEMESCHOW, 2000).

De acordo com Hosmer & Lemeschow (2000), é intuitivo supor que se as probabilidades se aproximam de 1, o indivíduo estimado pode ser classificado como $\widehat{Y}_i = 1$, bem como de forma contrária, se o modelo estimar probabilidades perto de 0, classificá-la como $\widehat{Y}_i = 0$. A grande questão é determinar qual será o ponto de corte para classificar a estimação como 0 ou 1. Na literatura recomenda-se o valor de 0,5, mas dependendo do estudo proposto pode não ser limitado a este nível (HOSMER & LEMESCHOW, 2000). No presente trabalho o ponto de corte adotado foi 0,5.

Após determinado o ponto de corte, cria-se a matriz de confusão (Tabela 2) com as observações de Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN) (SMOLSKI; BATTISTI, 2019, p9). A seguir apresenta-se a matriz de confusão.

Tabela 2 – Matriz de confusão.

Valor Estimado	**Valor Observado**	
	$Y = 1$	$Y = 0$
$Y = 1$	VP	FP
$Y = 0$	FN	VN

Fonte: Adaptado de Fawcett (2006).

A partir da matriz de confusão é possível determinar os seguintes avaliadores acerca da capacidade preditiva do modelo estimado:

1. Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente, dado por (SILVA *et al.,* 2019):

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (31)$$

2. Precisão: representa a proporção das predições corretas do modelo sobre o total (FAWCETT, 2006):

$$\text{Precisão} = \frac{VP+VN}{P+N} \qquad (32)$$

em que: P representa o total de "eventos" positivos (Y = 1) e N é o total de "não eventos" (Y = 0, ou negativo).

3. Sensibilidade: representa a proporção de verdadeiros positivos, ou seja, a capacidade do modelo em avaliar o evento como $\hat{Y} = 1$ (estimado) dado que ele é evento real $Y = 1$ (FAWCETT, 2006):

$$\text{Sensibilidade} = \frac{VP}{VP+FN} \qquad (33)$$

4. Especificidade: a proporção apresentada dos verdadeiros negativos, ou seja, o poder de predição do modelo em avaliar como "não evento" $\hat{Y} = 0$ sendo que ele não é $Y = 0$ (FAWCETT, 2006):

$$\text{Especificidade} = \frac{VN}{VN+FP} \qquad (34)$$

5. Verdadeiro Preditivo Positivo (VPP): se caracteriza como proporção de verdadeiros positivos com relação ao total de predições positivas, ou seja, se o evento é real $Y = 1$ dada a classificação do modelo $\hat{Y} = 1$ (FAWCETT, 2006):

$$\text{VPP} = \frac{VP}{VN+FP} \qquad (35)$$

6. Verdadeiro Preditivo Negativo (VPN): se caracteriza pela proporção de verdadeiros negativos comparando-se com o total de predições negativas, ou seja, o indivíduo não ser evento $Y = 0$ dada classificação do modelo como "não evento" $\hat{Y} = 0$ (FAWCETT, 2006):

$$\text{VPN} = \frac{VN}{VN+FN} \qquad (36)$$

2.1.5.2 Curva ROC

A taxa de acerto do modelo de regressão logística estimado será verificada também por meio da curva ROC (*Receiver Operating Characteristic*). A curva ROC consiste em um método gráfico para avaliação, organização e seleção do sistema diagnóstico e/ou predição (BRAGA, 2000).

Baseado no modelo, a curva ROC mensura a capacidade de predição do mesmo, partir das predições da sensibilidade e especificidade. Quanto maior a parte abaixo da curva ROC, isto é, quanto maior a área sob a curva ROC (AUC), maior será a sua

capacidade de averiguação dos grupos atraso e sem atraso (FAWCETT, 2006). A seguir apresenta-se a curva ROC (Figura 1).

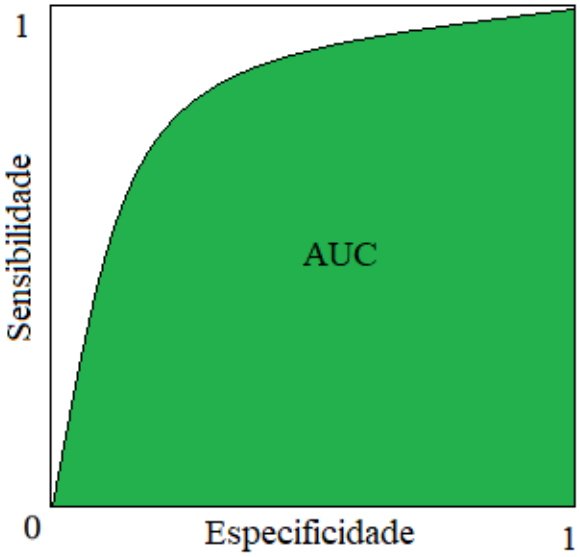


Figura 1 – Curva ROC

Assim será considerado mais preciso quando a curva se aproxima do canto superior esquerdo do gráfico. Hosmer & Lemeschow (2000) considera como aceitável, valores de AUC acima de 0,7.

Na Tabela 3 verifica-se a classificação obtida pela curva ROC (FÁVERO, 2009).

Tabela 3 - Classificações da curva ROC.

AUC	*Classificação*
Menor ou igual a 0,5	Não há descriminação
Entre 0,7 e 0,8	Discriminação aceitável
Maior que 0,8	Discriminação Excelente

2.1.5.3 Pseudo R^2

O Pseudo R^2 é a redução proporcional no valor absoluto da medida verossimilhança-log e, desse modo, é uma medida de quanto a não-aderência aumenta com o resultado da inclusão de uma variável preditora. Ele pode variar de 0 a 1, isto é,

valor igual a 0 indica que os preditores são inúteis na predição da variável de saída e valor igual a 1 indica que o modelo prevê perfeitamente a variável de saída (FIELD, 2009).

R^2_{CS} de Cox e Snell (1989), é baseado na verossimilhança-log do modelo (VL(Novo)) e a verossimilhança-log do modelo original (*VL*(nulo) e o tamanho da amostra, *n* (FIELD, 2009):

$$R^2_{CS} = 1 - e^{[-2/n(VL(novo)-Vl)(nulo))]} \qquad (37)$$

Contudo, essa estatística nunca alcança o seu valor teórico máximo, 1. Portanto, Nagelkerke (1991) sugeriu a seguinte correção (de Nagelkerke) (FIELD, 2009):

$$R^2_N = \frac{R^2_{CS}}{1-e^{[2(VL(básico)/n)]}} \qquad (38)$$

Embora todas essas medidas tenham diferenças na forma de cálculo (e nos resultados), coletivamente é possível considerá-las como praticamente as mesmas. Assim, em termos de interpretação elas podem ser vistas como similares ao R^2 da regressão linear no sentido de fornecer uma medida do grau de aderência do modelo (FIELD, 2009). No presente trabalho usou-se R^2_N de Nagelkerke.

2.1.6 Multicolinearidade

Justifica-se investigar se há multicolinearidade entre as variáveis explicativas, visto que a forte correlação entre elas acarreta vários efeitos negativos no ajuste do modelo de regressão. O problema de multicolinearidade torna a estimativa dos parâmetros imprecisa, por conta de um alto valor do erro padrão, o que não é conveniente estatisticamente (KUTNER et al., 2004; TAMHANE, DUNLOP, 2000).

Uma das formas de detecção da presença de multicolinearidade é avaliar o Fator de Inflação da Variância (*VIF*). Esse fator mede o grau de associação entre as variáveis explicativas, a partir do coeficiente de determinação do modelo de regressão ajustado apenas entre as variáveis independentes. O Fator de Inflação da Variância é definido como (BERK, 1977):

$$VIF_i = \frac{1}{1-R_i^2}, \quad (39)$$

em que: R_i^2 é o coeficiente de determinação da regressão da variável explicativa X_i sobre as outras variáveis explicativas com $i = 1,2,\ldots,k$, sendo k a quantidade de variáveis explicativas no modelo.

Moreira (2008), sugere que se qualquer fator de inflação da variância exceder 10, então a multicolinearidade será um problema.

Os resultados dos testes são apresentados na seção seguinte. Cabe ressaltar que todas as análises estatísticas foram realizadas no freeware R (R CORE TEAM, 2020).

3 RESULTADOS E DISCUSSÕES

Na tabela 4 são apresentadas algumas estatísticas descritivas das variáveis número de passageiros por voo e tempo de atraso em minutos.

Tabela 4 - Estatísticas descritiva das variáveis número de passageiros por voo e tempo de atraso.

Variável	Média	Mediana	Máximo	Mínimo	Desvio Padrão
Nº de Passageiros	128,3	134	185	0	31,1082
Tempo Atraso	21,02	10	337	1	30,9143

Verifica-se que a média de passageiros por voo é de 128 passageiros, e nesta companhia, neste período de tempo estudado, teve mínima de 0 passageiros pagantes, somente a tripulação operava o voo e máxima de 185 passageiros na aeronave, a capacidade máxima de passageiros pagantes é de 174 em sua maior aeronave, A320 que pode fazer pouso em Uberlândia, porém tem assentos extras para tripulantes e comandantes que não conta como assento de passageiros pagantes, aumentando assim a capacidade geral da aeronave. Já a média de tempo de atraso do voo é de 21 minutos. Tem mínimo de 1 minuto e máxima de 5 horas e 37 minutos que pode ser causado por manutenção não programada na aeronave, troca de aeronave ou até mesmo problemas com meteorologia.

Na Tabela 5, são apresentadas estatísticas descritivas das variáveis área responsável (AR), equipamento, turno, ano, mês e agentes em relação a variável atraso.

Tabela 5 - Estatísticas descritivas das variáveis área responsável, equipamento, turno, ano, mês e agentes em relação a variável atraso.

(continua)

Variáveis Independentes	Dummy	Atraso		Nº de voos	(%)
		sem atraso	com atraso		
Área Responsável	Sem	1908 (86%)	305 (14%)	2213	71
	Ar	0 (0%)	747 (100%)	747	24
	Terra	0 (0%)	161 (100%)	161	5
Equipamento	A319	559 (62,2%)	340 (37,8%)	899	29
	A320	1349 60,7%)	873 (39,3%)	2222	71
Turno	Manhã	813 (64,5%)	448 (35,5%)	1261	40
	Noite	681 (57,8%)	497 (42,2%)	1178	38
	Tarde	414 (60,7%)	268 (39,3%)	682	21
Ano	2017	625 (68,7%)	285 (31,3%)	910	29
	2018	864 (50,6%)	842 (49,4%)	1706	55
	2019	419 (83,0%)	86 (17,0%)	505	16

(conclusão)

Variáveis Independentes	Dummy	Atraso		Nº de voos	(%)
		sem atraso	com atraso		
Mês	Nov. a Fev.	754 (67,9%)	357 (32,1%)	1111	36
	Mar. a Out.	1154 (57,4%)	856 (42,6%)	2010	64
Agentes	2	1133 (61,1%)	724 (38,9%)	1857	60
	1	775 (61,3%)	489 (38,7%)	1264	40
	Total	1908 (61,1%)	1213 (38,9%)	--	--

Na Tabela 5 observa-se que 38,9% dos voos operados pela companhia no período dos dois anos tiveram atraso. Destes a maior parte se deu na área ar do aeroporto com 24% e somente 5% foi provocado pela equipe da área terra do aeroporto. A variável equipamento nos mostra que 71% dos voos operados no aeroporto de Uberlândia por essa companhia foi realizado pela aeronave A320, com capacidade de 174 passageiros. O turno que houve mais atrasos foi o da noite com 497 voos dos 1178, isto é, 42,2% dos voos tiveram atraso nos dois anos analisados, em contra partida, o turno manhã apresentou maior quantidade dos voos sem atraso. O ano que teve mais atraso foi o de 2018 com 1706 voos atrasados, isto é, 49,4% dos voos no ano de 2018 tiveram atrasos. O ano com menos atraso foi o de 2019 com 86 voos atrasados, ou seja, cerca de 17,0% dos voos no ano de 2019 tiveram atrasos. Observa-se na Tabela 2, que o período com maior número de voos foi de março a outubro, com 2010 voos. Além disso, esse período apresentou maior proporção (42,6%) de voos em atraso. No embarque, independentemente do número de agentes, aproximadamente 39% dos voos apresentaram atrasos. No entanto, 60% dos voos foram feitos com dois colaboradores.

A seguir serão apresentados resultados referentes aos ajustes dos modelos de regressões logísticas. No primeiro momento foi proposto o ajuste do modelo de regressão logística considerando a variável atraso (variável dependente) e as variáveis independentes número de passageiros, área responsável, equipamento, turno, ano, mês e número de agentes. Na Tabela 6 são apresentadas as estatísticas referentes ao ajuste do modelo em questão.

Tabela 6 - Estatísticas referentes ao ajuste do primeiro modelo de regressão logística múltipla com todas as variáveis mencionadas anteriormente.

(continua)

Parâmetro	Estimativa	Erro Padrão	OR	Wald	valor p	AIC
Intercepto	-4,6880	0,6470	--	-7,246	0,0000	1239,24
Passageiros (x_1)	-0,0163	0,0023	0,9838	-6,834	0,0000	**Pseudo R^2**
AR1 (x_2)	25,680	534,40	1,4e+11	0,048	0,9617	0,8299
AR2 (x_3)	25,543	1179,0	1,1e+11	0,022	0,9827	

(conclusão)

Parâmetro	Estimativa	Erro Padrão	OR	Wald	valor p	AIC
Equipamento1 (x_4)	0,0348	0,1753	1,0354	0,199	0,8426	
TurnoNoite (x_5)	-0,6402	0,1798	0,5271	-3,561	0,0003	
TurnoTarde (x_6)	-0,2088	0,1994	0,8115	-1,047	0,2950	
Ano2018 (x_7)	4,1810	0,5106	65,403	8,187	0,0000	
Ano2019 (x_8)	0,9699	0,7202	2,6377	1,347	0,1780	
Mes1 (x_9)	1,9830	0,2165	7,2630	9,156	0,0000	
Agentes (x_{10})	0,2271	0,1536	1,2549	1,479	0,1392	

Com base nos resultados da Tabela 6, observa-se que, pelo teste de Wald ao nível de significância de 5%, as variáveis área responsável, equipamentos e número de agentes não foram significativas para o modelo. A seguir destaca-se descrições sobre as variáveis dummies: (i) "TurnoNoite" recebe o valor 1 para os voos operados pela noite e recebe valor 0, caso contrário; (ii) "TurnoTarde" recebe o valor 1 para os voos operados no período da tarde e recebe valor 0, caso contrário; (iii) "Ano2018" recebe o valor 1 para os voos operados em 2018 e recebe o valor 0, caso contrário; "Ano2019" recebe o valor 1 para os voos operados em 2019 e recebe valor 0, caso contrário (iv) "Mes1", recebe o valor 1, se o voo ocorreu no período de março a outubro e recebe valor 0, se o período for de novembro a fevereiro; (v) Agentes, recebe valor 0 para dois ou mais agentes no portão de embarque, recebe valor 1 para um agente no portão de embarque. Observa-se que as estimativas são todas positivas (com exceção das variáveis passageiro e turno), ou seja, existe uma relação diretamente proporcional dessas variáveis com o logaritmo da chance de atraso nos voos.

A seguir é apresentado na Tabela 7 informações sobre o Fator de Inflação da Variância (*VIF*) para avaliar a presença de multicolinearidade no primeiro modelo ajustado.

Tabela 7 - Valores do *VIF* das variáveis do primeiro modelo ajustado.

Variável	*VIF*
Passageiros (x_1)	17,1902
AR1 (x_2)	162859500,0
AR2 (x_3)	212237300,0
Equipamento1 (x_4)	19,6647
TurnoNoite (x_5)	23,7047
TurnoTarde (x_6)	21,1921
Ano2018 (x_7)	20,1670
Ano2019 (x_8)	21,9539
Mes1 (x_9)	33,5522
Agentes (x_{10})	17,7364

Observa-se na Tabela 7 que todas as variáveis têm valor do *VIF* acima de 10. Além disso percebeu-se que a variável Área responsável (AR1 e AR2) tem valor muito alto e discrepante em relação aos demais. Neste caso, procedeu-se com o ajuste de um segundo modelo de regressão logística sem a variável Área Responsável. Na tabela 8 apresentam-se as estatísticas referentes ao ajuste do modelo sem a variável Área Responsável.

Tabela 8 - Estatísticas referentes ao ajuste modelo sem a variável Área Responsável.

Parâmetro	**Estimativa**	**Erro Padrão**	**OR**	**Wald**	**valor p**	***AIC***
Intercepto	-1,1154	0,2230	0,3277	-5,000	0,0000	3940,22
Passageiros (x_1)	0,0003	0,0012	1,0003	0,285	0,7756	**Pseudo R²**
Equipamento1 (x_4)	-0,0541	0,0973	0,9472	-0,557	0,5778	0,1037
TurnoNoite (x_5)	0,3571	0,0938	1,4291	3,806	0,0001	
TurnoTarde (x_6)	0,1741	0,1073	1,1901	1,621	0,1049	
Ano2018 (x_7)	0,7477	0,0903	2,1122	8,272	0,0000	
Ano2019 (x_8)	-0,8070	0,1486	0,4461	-5,430	0,0000	
Mes1 (x_9)	0,3298	0,0822	1,3908	4,009	0,0000	
Agentes (x_{10})	-0,1533	0,0850	0,8578	-1,802	0,0715	

Conforme Tabela 8, ao nível de significância de 5%, as variáveis passageiros, equipamento e agentes não são significativas. Já as variáveis turno, ano e mês foram significativas. Percebe-se também que o valor do *AIC* aumentou e o valor do Pseudo R² diminui em relação ao primeiro modelo. Porém, o segundo modelo (sem a variável Área Responsável) não apresenta multicolinearidade, pois, conforme os resultados apresentados na Tabela 9, todos os valores de *VIF* são inferiores a 10. Portanto, percebe-se que a exclusão da variável Área Responsável é de fato necessária. Logo, o modelo com Área responsável (primeiro modelo ajustado) deve ser descartado, pois apresenta multicolinearidade.

Tabela 9 – Valores do VIF das variáveis do modelo sem a variável Área Responsável.

Variável	***VIF***
Passageiros (x_1)	5,0722
Equipamento1 (x_4)	6,0653
Turno Noite (x_5)	6,4562
Turno Tarde (x_6)	6,1464
Ano2018 (x_7)	6,3196
Ano2019 (x_8)	9,3517
Mes1 (x_9)	4,8440
Agentes (x_{10})	5,4451

Em seguida ajustou-se um terceiro modelo de regressão logística. No entanto, foi proposto a seleção de variáveis por meio do critério de Stepwise com base no critério de informação de Akaike (*AIC*). A significância dos parâmetros foi avaliada por meio do teste de Wald, ao nível de 5%. Os resultados desse ajuste estão apresentados na Tabela 10.

Tabela 10 - Estatísticas referentes ao ajuste do terceiro modelo de regressão logística, com seleção de variáveis pelo critério de Stepwise e com base no AIC.

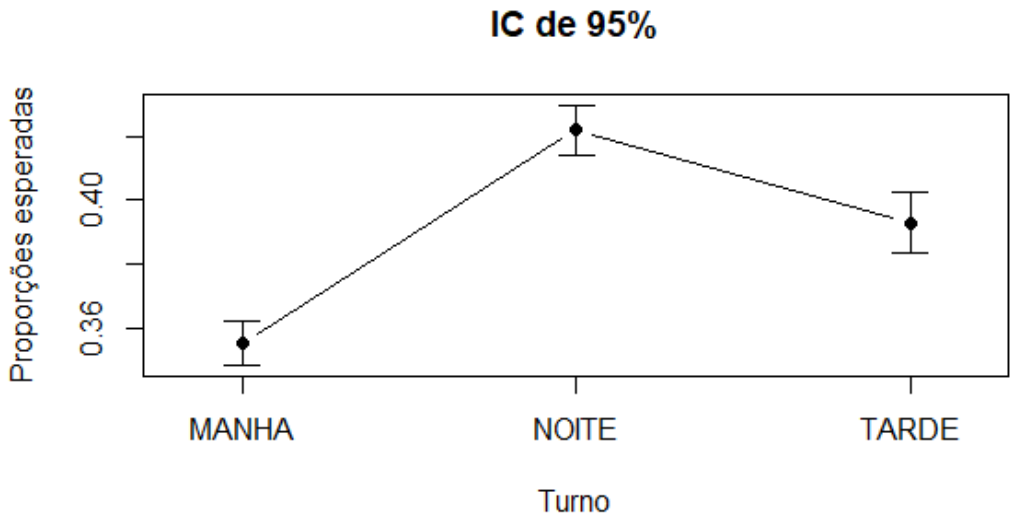
Parâmetro	**Estimativa**	**Erro Padrão**	**OR**	**Wald**	**valor p**	***AIC***
Intercepto	-1,0962	0,1181	0,3314	-9,275	0,0000	3936,57
TurnoNoite (x_5)	0,3373	0,0869	0,4011	3,882	0,0001	**Pseudo R²**
TurnoTarde (x_6)	0,1886	0,1028	0,2076	1,835	0,0664	0,1036
Ano2018 (x_7)	0,7413	0,0897	2,0988	8,263	0,0000	
Ano2019 (x_8)	-0,8124	0,1483	0,4437	-5,478	0,0000	
Mes1 (x_9)	0,3290	0,0821	1,3896	4,004	0,0000	
Agentes (x_{10})	-0,1561	0,0838	0,7257	-1,863	0,0624	

A partir do critério de Stepwise e do *AIC*, percebe-se que as variáveis, passageiros e equipamento foram retiradas do modelo. Comparando o *AIC* do segundo modelo com *AIC* do terceiro modelo, verifica-se que o terceiro modelo é melhor que o segundo, pois apresenta menor valor de *AIC*. Já em relação ao Pseudo R^2, praticamente não existe diferença entre os modelos. No entanto, o terceiro modelo possui um menor número de variáveis, consequentemente, menor número de parâmetros. Portanto, o melhor ajuste foi obtido no terceiro modelo.

Percebe-se que o turno noite é significativo a 5% de significância, em relação ao turno da manhã. Para cada voo com possibilidade de ocorrer a noite, o logaritmo da chance de atraso aumenta de 0,3373 a chance de atraso, quando comparado com o turno da manhã. Quando o logaritmo da chance aumenta, a probabilidade π também aumenta. Isso é natural, pois no turno da noite existe a possibilidade de ser afetado por atrasos ocorridos durante todo o dia, ou até mesmo o desfalque de funcionários nesse período do dia. Quanto ao turno tarde, não significativo a 5% de significância, conclui-se que as chances de atrasos entre os voos da tarde e da manhã, são as mesmas.

Na Figura 2 são apresentadas as proporções esperadas de atrasos por turno, determinadas a partir das predições do modelo final.

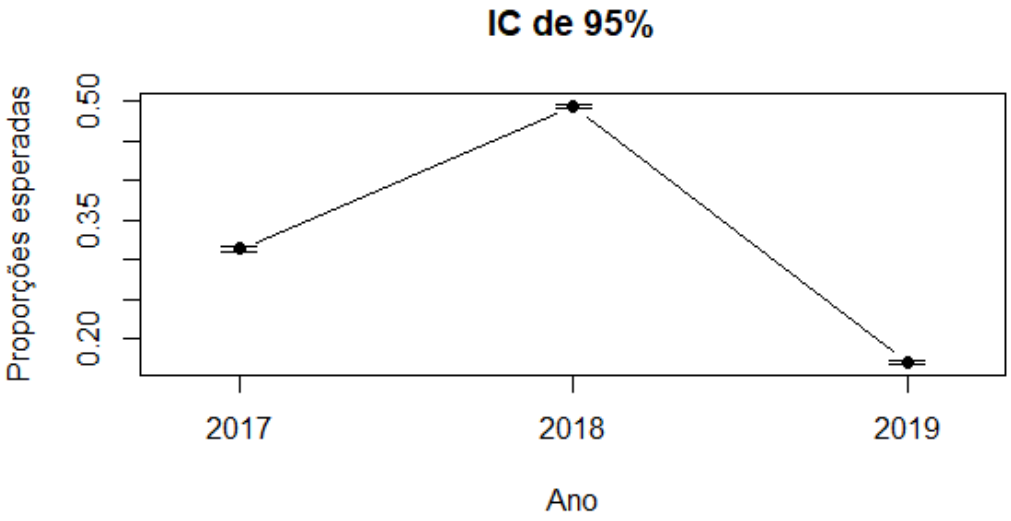
Figura 2 - Proporção esperadas de atraso em relação ao fator Turno.



A partir da Figura 2, percebe-se que as proporções indicam que o período da noite tem maior probabilidade de atraso em relação aos demais turnos.

Com base na Tabela 10, tem-se que o ano de 2018 é significativo a 5% de significância, conclui-se imediatamente que se o voo for do ano de 2017, as chances de atrasos de voos caem de forma considerável quando comparado com o ano de 2018. A ocorrência de atrasos no ano de 2018 foi dada pela transição de sistema que ocorreu nesse ano e alinhamento de procedimentos, já que a empresa passava por união com outra empresa do mesmo ramo. Além disso, conclui-se que os voos operados no ano de 2018, as chances de ter ocorrido atraso é 2,09 vezes mais provável quando comparado com o ano de 2017. Quanto ao ano de 2019, significativo a 5% de significância, quando comparado com o ano de 2017. Logo, os voos operados no ano de 2019 apresentaram menores chances de atraso, quando comparado com o ano de 2017. Na Figura 3 são apresentadas as proporções esperadas de atrasos por ano, determinadas a partir do modelo final.

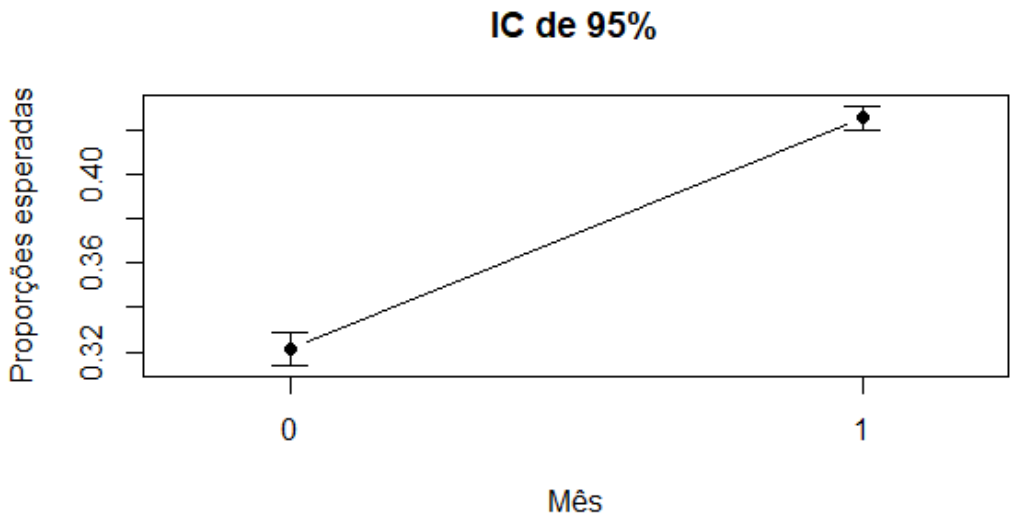
Figura 3 - Proporção esperadas de atraso em relação ao fator Turno.



É possível observar na Figura 3 que a proporção de voos com atraso foi maior no ano de 2018, seguido pelos anos 2017 e 2019.

A variável “Mes1” representa os voos ocorridos nos meses de março a outubro. Percebe-se que no período de março a outubro tem-se 1,38 vezes mais chance de ocorrer atraso quando comparado aos meses de novembro a fevereiro. Uma justificativa plausível é que de novembro a fevereiro consiste no período de férias escolares e meses chuvosos. A companhia se antecipa e contrata mais pessoas, propõe treinamentos que antecedem a data para que os funcionários estejam alinhados para trabalhar diante das adversidades tanto meteorológicas como ajustes não programados. Na Figura 4 são apresentadas as proporções esperadas de atrasos em relação ao fator Mês, determinadas a partir do modelo final.

Figura 4 - Proporções esperadas de atraso em relação ao fator Mês.



Com base nos resultados apresentados na Tabela 10, tem-se que o modelo estimado apresenta-se da seguinte forma:

$$ln\left[\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right] = -1{,}10 + 0{,}34x_5 - 0{,}19x_6 + 0{,}74x_7 - 0{,}81x_8 + 0{,}33x_9 - 0{,}16x_{10}$$

Para aferir a qualidade do ajuste do terceiro modelo (modelo escolhido), utilizou-se: a matriz de classificação (matriz de confusão), a acurácia, a precisão, a especificidade, a sensibilidade, a curva ROC e o Pseudo R^2.

A matriz de confusão consiste numa classificação cruzada da variável resposta (atraso e sem atraso), a partir de um ponto de corte (regra de predição), para saber se houve ou não acerto da probabilidade estimada com base nos valores reais, pois as probabilidades variam de 0 a 1 enquanto os valores reais binários possuem valores fixo de 0 "ou" 1 (Hosmer & Lemeschow 2000). No presente trabalho, o ponto de corte utilizado foi 0,5. Portanto, valores de probabilidades acima desse ponto indica, ocorrência de atraso e para valores abaixo de 0,5, indica a ausência de atraso. Na Tabela 11 pode-se comparar a classificação prevista versus a classificação observada, a partir do ponto de corte estabelecido.

Tabela 11 - Matriz de confusão para o modelo.

Observadas	**Predito**		**Total**
	Sem atraso	*Atraso*	
Sem Atraso	1634	274	1908
Atraso	816	397	1213
Total	2450	671	3121

A partir da Matriz de Confusão (Tabela 11), pode-se medir a acurácia, precisão, especificidade e a sensibilidade do modelo. A seguir, na Tabela 6, são apresentados resultados acerca do modelo final.

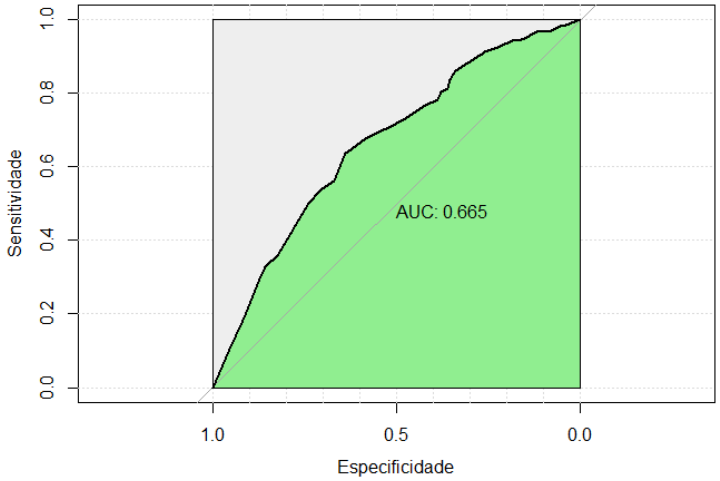
Tabela 12 - Medidas determinadas a partir da Matriz de Confusão.

Acurácia	**Precisão**	**Especificidade**	**Sensibilidade**
0,6507	0,5916	0,8563	0,6669

A ponto de corte de 0,5, a matriz de confusão tem uma boa acurácia de 65,07%, a proporção de positivos que foram corretamente identificados é de 59,16% e a proporção de verdadeiros positivos foi de 66,69%. Sendo que a precisão, que representa a proporção das predições corretas e a proporção dos verdadeiros negativos, representada pela especificidade foi de 85,63%. Conclui-se que o modelo tem um bom poder de descriminação.

A área sob a curva ROC é que representa um importante indicador de precisão do modelo. Essa curva permite evidenciar os valores para os quais existe maior otimização da sensibilidade em função da especificidade. O valor da área abaixo da diagonal (0,5 ou 50%) não tem validade, pois os acertos ou erros entram na mesma proporção; enquanto que, um valor igual a 1,0 ou 100% não pode ser alcançado pois sempre a superposição das proporções dos grupos (CAMARA, 2009). A seguir apresenta-se a Curva ROC (Figura 5).

Figura 5 - Curva ROC do modelo de regressão logística.



Como área sob a curva ROC (AUC - Area Under the ROC Curve) resultou em aproximadamente 0,7, pode-se dizer que o modelo apresenta uma boa discriminação.

De acordo com Louviere et al., (2000), o Pseudo R^2 na regressão logística pode ser interpretado como o coeficiente de determinação (R^2) na regressão linear. No entanto, os autores ressaltam que o valor do Pseudo R^2 não será tão grande, isto é, valores entre 0,2 e 0,4 são considerados indicativos de excelentes ajustes. Nas simulações de Domenich & McFadden (1975), os valores de Pseudo R^2 entre 0,2 e 0,4 equivalem ao intervalo de 0,7 a 0,9 para uma função linear. No presente trabalho, o modelo final (terceiro modelo) apresentou um Pseudo R^2 no valor de 0,1036, ou seja, o modelo apresentou um ajuste razoável.

4 CONCLUSÃO

No presente trabalho foi possível estudar a relação do atraso dos voos e os fatores que influenciam o atraso por meio do ajuste de modelos de regressão logística. A partir da avaliação da multicolinearidade por meio do Fator da Inflação da Variância (*VIF*), excluiu-se a variável área responsável do modelo. Em seguida, a partir do critério de Stepwise e com base no critério de informação de Akaike (*AIC*), as variáveis equipamento e passageiros foram excluídas. As demais variáveis: Turno, Ano, Mês e Agentes permaneceram no modelo com base no critério de informação de Akaike (*AIC*).

Evidenciou-se que o turno da noite tem maior probabilidade de atraso contrastando aos demais, destaca-se a organização dos aviões para que atrasos durante o dia não afete os outros turnos e quantidade de funcionários seja a necessária para a operação diária.

Em relação ao ano, deve-se salientar que o ano de 2018 houve mais atraso quando comparado aos demais anos. Já os meses do ano com mais chuvas e época de férias é o período com menor probabilidade de atraso pois, a companhia promove treinamentos durante todo o ano podendo evitar eventuais imprevistos.

Para estudos futuros pretende-se considerar outros aeroportos de mesmo porte, incluindo aeroportos do mesmo estado. Acredita-se que assim será possível identificar falhas e pontos a serem melhorados.

5 REFERÊNCIAS

ABREU, H. J. **Aplicação da análise de sobrevivência em um problema de Credit Scoring e comparação com a regressão logística**. 2004 118p. Dissertação (Mestrado) – Universidade Federal de São Carlos, São Carlos, 2004.

Aeroporto de Uberlândia completa 83 anos de operações. **Infraero Aeroportos**, Brasília, 2017. Disponível em: <https://www4.infraero.gov.br/imprensa/noticias/aeroporto-de-uberlandia-completa-83-anos-de-operacoes/>. Acesso em: 14 de Out. de 2019.

Atrasos, cancelamentos, interrupção do serviço e preterição. **ANAC**, Brasília, 2018. Disponível em: <https://www.anac.gov.br/assuntos/legislacao/legislacao-1/rbha-e-rbac/passageirodigital/atraso-cancelamento-e-pretericao/atrasos-cancelamentos-pretericao>. Acesso em: 17 de Out. de 2019.

Annual Network Operations Report 2014. **Technical report**, 2014, Disponível em: <https://www.eurocontrol.int/sites/default/files/publication/ performance/2014_annual/final/annual_network_operations_report_2014_main_report_final_edition.pdf.>. Acesso em: 15 de Out. De 2019

BERK, KENNETH N. "**Tolerance and Condition in Regression Computations**,". Proceedings of the Ninth Interface Symposium on Computer Science and Statistics. Edited by: Hoaglin, David C. and Welsch, Roy E. Boston: Prindle, Weber, and Schmidt.

BEZERRA, G.K.A. **Logistic Regression Model for prediction of death in the Intensive Care Unit**. João Pessoa, 2012. 90p. Dissertação (Mestrado) – Departamento de Estatística, Universidade Federal da Paraíba.

CAMARA, F. P. **Psiquiatria e estatística V**: validação de procedimentos diagnósticos pela curva ROC. Phychiatry on line Brasil. V. 14, n.4,2009.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada: para os cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, 2009.

DOMENICH, T.; MCFADDEN, D.L. **Urban Travel Demand: A Behavioral Analysis**. North-Holland Publishing Co., 1975.

FÁVERO: L. P. FÁVERO, P. B., SILVA, F. BETTY L. C.; Elsevier. **Análise de dados**. CAMPUS, 2009.

FAWCETT, T. 2006. «**An introduction to ROC analysis**». Pattern Recognition Letters 27: 861– 74. https://doi.org/10.1016/j.irbm.2014.09.001.

FIELD, A. **Descobrindo a estatística usando SPSS**. 2. ed. Porto Alegre: Artmed, 2009.

FIGUEIRA, C. V. *Modelos de regressão logística*. 2006 149 p. Dissertação (Mestrado em Matemática) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006

HAIR Jr., J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005a.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. New York. 2ª ed.: John Wiley, 2000.

KUTNER, M. H. et al. **Applied linear models**. 5th ed. New York: McGraw-Hill Irwin, 2005.

LOUVIERE, J.J.; HENSHER, A.D.; SWAIT, D.J. **Stated choice methods**. New York: Cambridge University Press, 2000.

MELO, L, **Análise da Propagação de Atrasos na Malha Aérea Brasileira Usando Padrões Frequentes**, Rio de Janeiro, 2016

MEZZOMO, M. Estudo da mortalidade Infantil – um estudo de regressão logística múltipla. **Monografia de Especialização em Estatística e Modelagem Quantitativa**.

Centro de Ciências Naturais e Exatas – Universidade Federal de Santa Maria. Santa Maria, RS, Brasil, 2009).

MOREIRA, L.F. **Multicolinearidade em Análise de Regressão**. XII ERMAC, 2008.

MONTGOMERY, D., E. PECK, e G. VINING (2001). **Introduction to linear regression analysis**. John Wiley and Sons.

NAGELKERKE, N.. "**A note on a general definition of the coefficient of determination**." Biometrika 78 (1991): 691-692.

PAGANO, M. **Princípios de Bioestatística**. Tradução de Luiz Sergio de Castro Paiva. São Paulo: Pioneira Thomson Learning, 2008)

R Core Team (2020). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

RODESKI W. 2010. Geomarketing: **O uso de regressão logística múltipla no mapeamento de regiões geográficas de alto potencial mercadológico** (Bacharel em Estatística) - Universidade Federal do Rio Grande do Sul, [S. l.], 2010. Disponível em: https://www.lume.ufrgs.br/handle/10183/29109. Acesso em: 26 set. 2020.

SILVA, L. C. C.; SILVA, A. W. S.; FERNANDES FILHO, A. N.; BARRADAS FILHO, A. O. **Estudo Comparativo de Métodos de Aprendizagem de Máquina Aplicados em Sistemas de Detecção de Intrusão**. *In*: ESCOLA REGIONAL DE COMPUTAÇÃO CEARÁ, MARANHÃO, PIAUÍ (ERCEMAPI), 7., 2019, São Luís. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 135-142.

SMOLSKI, F.; BATTISTI, I. **Software R: curso avançado**. [S. l.: s. n.], 2019. 9 p. Disponível em: https://smolski.github.io/livroavancado/index.html. Acesso em: 26 set. 2020.

SOBRAL, T. E. L.; BARRETO, G. **Análise dos critérios de informação para a seleção de ordem em modelos auto regressivos**. *Conferência brasileira de Dinâmica, Controle e Aplicações.* Águas de Lindóia-SP v.1, n. único, 2011.

TAMHANE, A. C. & Dunlop D. D. **Statistics and Data Analysis**: from elementary to intermediate**.** Upper Saddle River: Prentice-Hall, 2000.