Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Modelo de Regressão Logística aplicado a Indicadores de Risco do Fracionamento de Cargas no Transporte Aéreo

Lucas Queiroz Gongora

Orientador: Professor Dr. Gladston Luiz da Silva

Brasília

2018

Lucas Queiroz Gongora

Modelo de Regressão Logística aplicado a Indicadores de Risco do Fracionamento de Cargas no Transporte Aéreo

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Professor Dr. Gladston Luiz da Silva

Brasília

2018

Lucas Queiroz Gongora
Modelo de Regressão Logística aplicado a Indicadores de Risco do Fracionamento de Cargas no Transporte Aéreo / Lucas Queiroz Gongora. – Brasília, 2018-
45 p.
Orientador: Professor Dr. Gladston Luiz da Silva

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação, 2018.

Lucas Queiroz Gongora

Modelo de Regressão Logística aplicado a Indicadores de Risco do Fracionamento de Cargas no Transporte Aéreo

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. Brasília, 15 de junho de 2018:

Professor Dr. Gladston Luiz da Silva
Orientador

Claudete Ruas
Membro da Banca

Helton Saulo Bezerra dos Santos
Membro da Banca

Brasília
2018

Agradecimentos

Primeiramente dedicado este trabalho aos meus pais e irmão que deram todo o suporte possível para a realização do curso de Estatística e consequentemente o projeto final.

Também gostaria de dedicar aos professores com quem tive aula ao longo do curso e em especial ao meu orientador, professor Dr. Gladston Luiz da Silva. Agradeço, ainda, ao Sr. João Marcelo Silva, funcionário da Empresa Brasileira de Correios e Telégrafos, responsável pela disponibilização dos dados que subsidiaram as análises realizadas neste trabalho.

E por fim, aos meus colegas de graduação que me incentivaram a ser o Estatístico que sou hoje e que me acompanharam desde 2014 nessa jornada.

Resumo

Este texto apresenta a aplicação de Regressão Logística para a Empresa Brasileira de Correios e Telégrafos, no qual foi estudado a predição do fracionamento de cargas no transporte aéreo, com base em indicadores de risco com o objetivo de ajustar um modelo que contribua para a redução de gastos e otimização do uso de aviões como meio de transporte. A variável resposta da Regressão Logística foi dividida entre a ocorrência ou não do fracionamento da carga, e como explicativa foram considerados o ano, o período de férias, a capacidade máxima de carga, a quantidade de paletes, a companhia aérea, origem e destino (linha), e um novo indicador com a proporção de carga desembarcada por carga transportada. O modelo final foi ajustado em um banco de dados restrito a uma Linha e uma Transportadora específicas, no caso a Linha 6 e a Transportadora 11. Este modelo indicou bons resultados de qualidade de ajustamento baseado na análise de resíduos, além de todas as variáveis terem sido adicionadas como significativas.

Palavras-chave: Regressão Logística, Logito, Transporte de cargas, Capacidade de Carga, Análise de Resíduos.

Abstract

This paper introduces the aplication of Logistic Regression to the Correios Brazilian entrprise, in which, it was studied the predection of air parcel fraction, basaead on the behavior of risk indicators with the objective of fit a model that contributes to cut costs and optimization of the airliners. The response variable of Logistic Regression was broken in happened or not happened the parcel fraction, and as predictor variable was used year, vacations, max parcel capacity, number of pallets, airline, origin and destiny (line) and a new indicator that is the proportion of landed parcel by transported parcel. The final model was adjusted in a data base restricted by Line 6 and onde Airline 11. This model showed goodness of fit by the analysis of residuals, besides, all other variables was added as significant.

Keywords: Logistic Regression, Logit, parcel transportation, Parcel capacity, Analysis of residuals.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A Gestão de Transporte da Empresa Brasileira de Correios e Telégrafos (Correios) teve sua origem em 1963 e tem por objetivo integração e inclusão social, além de oferecer soluções para atender às necessidades de comunicação de empresas e instituições.

Em 1982, como complemento à sua missão de conectar pessoas, instituições e negócios por meio de soluções postais e logística acessíveis, confiáveis e competitivas, foi criado o Sedex. Atualmente um dos principais produtos da organização e um dos líderes do setor de encomenda no Brasil.

Ao longo dos anos, embora o crescimento da área de transporte de encomendas tenha gerado incremento de receitas para os Correios, apresentou também aspectos negativos, como o aumento de riscos em produtos extraviados e, consequentemente, a elevação dos gastos em indenização, além de outros riscos como o uso desnecessário de meios de transportes.

Essa situação foi agravada já que a Empresa não possuía claras políticas de Conformidade e de Gerenciamento de Riscos, adequadas ao seu porte, complexidade e risco das operações realizadas. Para sanar esse problema, a organização busca a implementação da metodologia de Auditoria Baseada em Riscos, que visa a prevenção de perdas por erros ou fraudes.

Os resultados encontrados nesse trabalho foram utilizados como subsídios na dissertação Silva (2017), cujo objeto é propor metodologia de Gestão Baseada em Riscos em processos de gestão do transporte de cargas dos Correios.

A metodologia de Gestão Baseada em Risco será elaborada com base no artigo de Saaty (1987), no qual aborda o Processo Analítico Hierárquico (em inglês *AHP*). O AHP é uma teoria geral de mensuração que é usada para comparações entre medidas baseadas em fatos com uma escala criada a partir de ponderações relativas e do instinto de especialistas, tanto de informações discretas quanto contínuas

Nessa busca pelo aprimoramento da metodologia, a Empresa cedeu informações coletadas sobre o transporte aéreo de cargas para que pudesse ser feito o ajuste de modelos preditivos para indicadores de risco relacionados ao fracionamento de cargas no transporte aéreo. No qual, visa fornecer informação que contribuam para a redução de gastos e o uso otimizado de aviões como meio de transporte.

Os resultados encontrados foram entregues aos Correios, no qual os indicadores estimados e seus efeitos serão comparados na AHP. Dessa forma será possível realizar um estudo comparativo entre a opinião de especialistas com os resultados estatísticos

encontrados, para assim, conseguir aprimorar o método de decisão dos Correios

1.1 Objetivos

O objetivo geral deste trabalho foi ajustar modelos de Regressão Logística Binária baseados em indicadores de risco relacionados ao fracionamento de cargas no transporte aéreo, que permitam o cálculo da probabilidade da utilização da capacidade máxima de carga, sem a necessidade de demandar uma segunda locomoção.

O fracionamento de cargas ocorre quando uma aeronave precisa transportar mais encomenda do que tem de capacidade, fazendo com que a carga excedida seja fracionada e transportada em uma segunda aeronave.

1.1.1 Objetivos Específicos

Para alcançar o objetivo geral foram realizados os seguintes objetivos específicos:

- Realizar Análise Exploratória em dados provenientes dos Correios relativos ao transporte aéreo de cargas fracionadas e não fracionadas, e verificar possíveis associações entre as variáveis consideradas no modelo;
- Ajustar modelos de Regressão Logística Binária com base em variáveis relacionadas ao transporte aéreo de cargas realizado pelos Correios, assim como avaliar a qualidade dos ajustes realizados.

1.2 Metodologia

Este trabalho utilizou dados armazenados pelos Correios no período compreendido entre novembro de 2011 e agosto de 2017, relativos ao sistema de transporte aéreo de cargas realizadas pela Empresa. cujas principais variáveis de interesse são o peso da carga, a capacidade de carga, origem e destino dos trajetos.

De início, buscou-se entender a Organização em questão e o possíveis fatores de risco do transporte de cargas dos Correios. Em seguida foi realizada revisão bibliográfica de Regressão Logística Binária.

Também, foi realizada Análise Exploratória de Dados para entender as relações entre as variáveis de interesse, mediante uso de gráficos e medidas resumo.

Em seguida, foram verificados os pressupostos necessários para os métodos estatístico que foram empregados. Posteriormente, ajustados os modelos de Regressão Logística, cujo objetivo foi indicar a probabilidade de se alterar o fracionamento de determinada carga, no transporte aéreo.

Por fim, foi feita a validação dos modelos regressivos visando verificar a qualidade dos mesmos como ferramente de suporte para as Auditorias Baseada em Risco.

Para construir os modelos preditivos foi utilizada a técnica de Regressão Logística, que consiste em fazer o melhor ajuste e mais parcimonioso, de tal maneira que seja possível explicar a variável dependente por meio de um conjunto de variáveis independentes. O uso da Regressão Logística neste trabalho decorre do fato de que a variável dependente do modelo a ser ajustado para o transporte aéreo de cargas ser binária.

2 Referencial Teórico

De acordo com Agresti (2002), a análise de regressão é um método estatístico que utiliza da relação entre duas ou mais variáveis para que a variável resposta possa ser predita, a partir das demais. Mais especificamente para a Regressão Logística Binária, essa variável resposta é qualitativa, ou seja, a resposta será dividida em escalas de mensuração.

As demais variáveis são chamadas de variáveis independentes, ou de explicativas, são as variáveis que explicam a sua resposta. Isto é, o quanto que a probabilidade da sua variável resposta ser um sucesso altera para cada valor dessa variável explicativa.

Assim como há diferença entre o tipo de variável resposta, há diferença para a explicativa. Ao utilizar uma variável independente pode-se classificá-la como quantitativa ou qualitativa. No primeiro caso, a interpretação é feita de acordo com o próprio valor da observação. Já para a variável qualitativa, são criadas variáveis *dummy* para o auxílio da interpretação desses fatores.

2.1 Regressão Logística Binária

Baseado no contexto dos Correios, a Regressão Logística foi utilizada para construir um modelo estatístico que calculou a probabilidade do transporte sofrer corte de carga. Essa variável resposta, também chamada de variável dependente, foi dividida em "sim"ou "não"e portanto, enquadra-se nos critérios de regressão logística.

No estudo de caso dos Correios, a Regressão Logística aplicou-se ao transporte aéreo, na qual a variável resposta foi 1 caso tenha ocorrido o corte de carga, ou 0 caso não tenha ocorrido o corte de carga.

Para calcular a probabilidade de ocorrer o corte na carga ($\pi(y = 1)$), ajustou-se um modelo estatístico com uma matriz de variáveis explicativas influentes ($\underset{\sim}{x}$) como mostra a seguir:

$$\pi(y = 1) = \alpha + \underset{\sim}{\beta}\underset{\sim}{x} + \epsilon \tag{2.1}$$

em que:

- α é o intercepto do modelo;
- ϵ é o erro aleatório;
- β representa, em uma matriz, o coeficiente de cada uma das variáveis .

De maneira geral, o β indica o peso que determinada variável tem na probabilidade. Em caso de β negativo, a medida que a variável explicativa aumenta a probabilidade do sucesso ocorrer diminui, dado as demais variáveis constantes. E quando é positivo, a probabilidade de sucesso aumenta sob as mesmas circunstâncias. Quando $\beta_i = 0$ significa que a resposta não depende da variável i, ou seja, que determinado parâmetro não influencia o fator de interesse.

Na expressão 2.1, a distribuição de y_i pode ser descrita como uma distribuição de Bernoulli, uma vez que, dentro da Regressão Logística Binária, os valores atribuídos a y podem assumir apenas 0 ou 1.

$$y_i \mid X_i = x_i \sim Ber(\pi(x_i)) \tag{2.2}$$

No qual Y_i é a i-ésima variável resposta e X_i é o vetor de covariáveis da i-ésima observação. Além disso, para a implementação do modelo de Regressão Logístico Y sofre uma transformação logito no qual pode ser escrito por: $exp(\alpha + \underset{\sim}{\beta}\underset{\sim}{x})$, e dessa forma, o modelo passa a ser uma função linear dos parâmetros de β. Assim, o efeito de um indicador x sobre a variável resposta y pode ser encontrada por:

$$\pi(y \mid X = x) = \frac{exp(\alpha + \underset{\sim}{\beta}\underset{\sim}{x})}{1 + exp(\alpha + \underset{\sim}{\beta}\underset{\sim}{x})} \tag{2.3}$$

As estimativas β são obtidas a partir do cálculo do estimador de máxima verossimilhança. Dessa forma, é obtido $\hat{\pi}$ em função dos parâmetros estimados:

$$\hat{\pi}(y \mid X = x) = \frac{exp(\hat{\alpha} + \underset{\sim}{\hat{\beta}}\underset{\sim}{x})}{1 + exp(\hat{\alpha} + \underset{\sim}{\hat{\beta}}\underset{\sim}{x})} \tag{2.4}$$

O resultado estimado da probabilidade de x permite encontrar a influência desse indicador na variável resposta, de forma que para cada acréscimo de uma unidade da variável x, mantendo os demais indicadores constantes, a chance é acrescida em e^{β_i}, como indicado na equação 2.5

Sabe-se que a *odds*, também chamada de chance, é dada por $\pi(1-\pi)^{-1}$, então ao fazer a razão de chances (OR) de x_1 e de $x_1 + 1$, ou seja, o quanto que o primeiro indicador influencia a variável resposta, encontra-se o seguinte resultado:

$$OR(x_1 + 1; x_1) = \frac{\pi(x_1 + 1)(1 - \pi(x_1 + 1))^{-1}}{\pi(x_1)(1 - \pi(x_1))^{-1}} = \frac{e^{\alpha + \beta_1 x_1 + 1}}{e^{\alpha + \beta_1 x_1}} = e^{\beta_1} \tag{2.5}$$

Então, como já dito, para β negativo, implica que $e^{\beta} < 1$, então dessa forma, a chance de ocorrência de $Y = 1$ é reduzida. De maneira análoga, para β positivo, a chance aumenta.

A estimativa da razão de chances é dada por: $\hat{OR}(x_1+1, x_1) = e^{\hat{\beta}_1}$. De forma geral, a estimativa da razão de chances para o acréscimo em u unidades de x_1 é dado por:

$$\hat{OR}(x_1 + u, x_1) = e^{u\beta_1} \tag{2.6}$$

O β estimado no modelo regressivo 2.4 pode ser dado de acordo com seu intervalo de confiança, supondo que $Z_{\frac{\alpha}{2}}$ segue uma distribuição Normal Padrão (média igual a zero e variância igual a um):

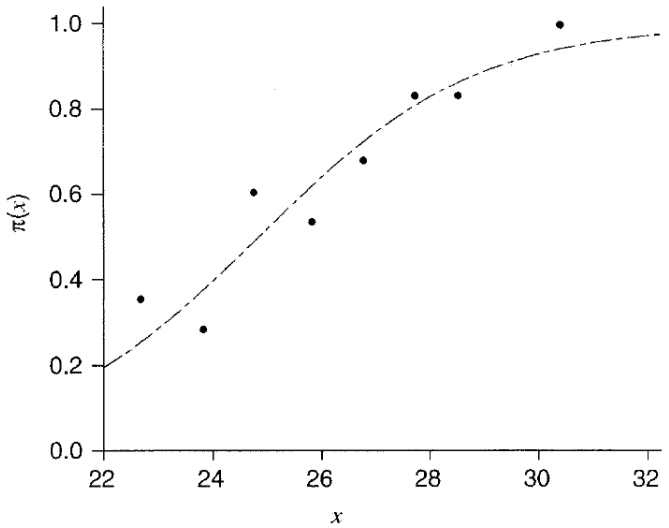
$$(\alpha + \beta_i x_i) \in \left\{\hat{\alpha} + \hat{\beta}_i x_i \pm Z_{\frac{\alpha}{2}} ASE\right\} \tag{2.7}$$

no qual o ASE (*Asymptotic Standard Error*), erro padrão assintótico, pode ser escrito como:

$$ASE = \sqrt{var(\hat{\alpha}) + x_i^2 var(\hat{\beta}_i) + 2xcov(\hat{\alpha}, \hat{\beta}_i)} \tag{2.8}$$

A partir da função que calcula a probabilidade da ocorrência do evento (Equação 2.4) é possível observar e ajustar as proporções da variável resposta para uma única variável independente quantitativa, como mostra a Figura 1

Figura 1 – Modelo Logito sob uma única variável x



Fonte: imagem adaptada do livro (AGRESTI, 2002), pag. 169

Como mostrado na Figura 1, a curva de regressão logística ajustada possui no eixo x o efeito de uma variável explicativa (no caso, que varia de 22 à 32) e no eixo y a probabilidade de ocorrer o sucesso do modelo.

2.2 Seleção de Modelos

Após definido o modelo estudado, encontra-se quais variáveis explicativas que compõe o modelo. Para isso utilizou-se os métodos de seleção automática de variáveis, *Forward, Backward* e *Stepwise*.

O método *Forward* consiste em supor que não existem variáveis no modelo, apenas o intercepto. E a partir desse pressuposto adiciona-se uma variável por vez ao modelo utilizando como critério o Teste da Razão de Verossimilhança (Equação) e, dessa forma, é adicionada ao modelo a variável com menor $p - valor$ obtido. Utiliza-se como critério de parada quando nenhuma variável obtiver $p - valor$ significativo.

Ou seja, é realizado a hipótese nula (H_0) de $\beta_j = 0$ contra a hipótese alternativa (H_1) de $\beta_j \neq 0$, em seguida analisa-se o menor $p - valor$ encontrado dentre os j teste realizados e verificou-se se esse valor é menor que o α definido em 0,05. Caso seja, adiciona a $j - ésima$ variável ao modelo. Realiza-se o procedimento até que o $p - valor$ seja maior que o α.

O método *Backward* funciona de forma semelhante ao $Forward$, mas ao invés de supor que não há variáveis no modelo, supõe-se que todas as variáveis já estão inseridos. Então, pelo critério do Teste da Razão de Verossimilhança, retira do modelo a variável que obtém o maior $p - valor$, e realiza o mesmo procedimento até que todas as variáveis sejam consideradas significativas.

E dessa forma, têm-se as mesmas hipóteses do método anterior, mas busca-se o $p - valor$ que é maior que o α, e o procedimento repete-se até que todos sejam menores.

Por fim, o *Stepwise* é a junção dos dois métodos anteriores. O modelo começa sem nenhuma variável e da mesma forma que acontece no *Forward* acrescenta-se a variável com menor $p - valor$. Então, com apenas uma variável no modelo realiza-se o mesmo procedimento do *Backward* para verificar se essa variável deve sair. O processo é repetido até que nenhuma variável entre mais no modelo e nenhuma saia.

O Teste de Razão Verossimilhança, compara o modelo completo com o modelo reduzido e verifica se todos os β da hipótese nula (H_0) são iguais a zero, ou seja, que a variável reposta não depende da variável i. A estatística para o teste é dada por G^2:

$$G^2 = -2\log_e\left[\frac{L(R)}{L(F)}\right] \tag{2.9}$$

No qual,

- L(R) representa o modelo reduzido a ser testado;
- L(F) o modelo completo;

Além disso, para tamanho de amostra considerada grande, G^2 assume uma distribuição aproximada qui-quadrado ($\chi^2_{(p-q)}$). Portanto, não rejeita a H_0 para $G^2 \leq \chi^2_{(1-\alpha;p-q)}$.

Embora os três métodos de seleção automática desempenharem o mesmo papel, é importante a realização de todos eles para a construção do modelo, uma vez que é possível que esses métodos gerem resultados diferentes, e, então, é feita uma comparação para ver qual é mais adequado.

2.3 Critérios de Decisão

Através dos métodos de seleção automática, foram obtidos três modelos. Para verificar qual desses modelos foi o melhor, tanto de ajuste quanto de erro, foram analisados os resultados dos Critérios de Informação, valores influentes, valores discrepantes, teste de ajustamento, coeficiente de determinação, entre outros critérios.

Vale ressaltar que para os critérios descritos acima, o aumento no número de variáveis, resultou em uma redução em seus resultados, no entanto, verificou-se a significância da redução, já que modelos com elevado número de variáveis tendem a ser mais complexos para a utilização.

Um desses critérios utilizado, além dos testes de significância, foi o Critério de Informação de Akaike Corrigido (AIC_C). O critério baseia-se no quanto que os valores ajustados do modelo estão próximos dos valores reais, essa técnica penaliza o modelo quando há muitos parâmetros:

$$AIC_C = 2k - 2ln(\hat{L}) + \frac{2k(k+1)}{n-k-1} \tag{2.10}$$

no qual,

- k é o número de parâmetros do modelo;
- n é o tamanho da amostra.

Já o Critério de Informação Bayesiano (BIC) penaliza de forma mais severa o modelo que possui muitos parâmetros. E pode ser descrito como:

$$BIC = k \times ln(n) - 2ln(\hat{L}) \tag{2.11}$$

2.4 Análise de Resíduos

Além dos critérios de informação, realizou-se o teste de Hosmer & Lemeshow (Teste de Qualidade de Ajustamento), que foi utilizado para verificar se o modelo ajustado foi

bem adequado ou não. O teste utiliza-se de variáveis categóricas para criar repetições nas categorias e assim verificar o ajuste. No caso de variáveis quantitativas, para realização do teste, foram categorizados as observações conforme a separação por decis e, dessa forma, foram criadas as repetições nas categorias e calculado a estatística do teste. Suas hipóteses foram:

$$H_0 : \text{o modelo está bem ajustado}$$
$$H_1 : \text{o modelo não está bem ajustado}$$

E assim, como os demais testes, caso o $p - valor$ encontrado seja menor do que o α definido, considera-se que há evidências para a rejeição de H_0, e, nesse caso, que o modelo não foi bem ajustado.

Outra técnica que foi utilizada foi a do resíduo de Pearson Studentizado (r_{sp_i}) que funciona de forma similar ao resíduo da Regressão Normal Linear, e objetivou avaliar o quão bem as observações foram previstas pelo modelo, de forma que funciona como um complemento do teste de Hosmer & Lemeshow. De modo, que as observações que não foram bem ajustadas pelo modelo apresentam resíduos elevados, que é dado por:

$$t_{s_i} = \frac{1}{\sqrt{1 - \hat{h}_{ii}}} \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}} \tag{2.12}$$

no qual,

- h_{ii} é a posição ii da matriz dada por $X(X^TX)^{-1}X^T$.

Como está sendo trabalhado com resíduos, quanto maior o resíduo, maior a sua discrepância para os dados reais. Além de verificar a qualidade do ajuste por observação, também é possível medir a influência que cada uma dessas observações tem nas estimativas dos coeficientes (β), para isso, utilizou-se a Distância de Cook:

$$LD_i = \frac{\hat{h}_{ii}}{(1 - \hat{h}_{ii})^2} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)} \tag{2.13}$$

O cálculo da distância de Cook utiliza-se das mesmas informações do cálculo do Pearson Studentizado. De acordo com Paula (2013), essas são as análises recomendadas e suficientes para o estudo dos resíduos, de t_{s_i} e de LD_i contra as probabildiade de $\hat{\pi}_i$.

2.5 Poder Preditivo

Outra medida para verificar a qualidade do ajuste foi por meio da análise do poder preditivo, estudo sobre a Curva ROC. Elaborou-se o gráfico da função (1-especificidade)

no eixo x com a sensibilidade no eixo y. No qual costuma ter formato côncavo ligando a origem ao ponto (1,1). A análise foi feita pela área abaixo dessa curva, sendo que quanto maior a área, melhor.

A sensibilidade pode ser indicada pela $P(\hat{y} = 1 \mid y = 1)$, ou seja, é a probabilidade de uma observação for diagnosticada como positiva seja de fato positivo. Já a função de especificidade é dada por $P(\hat{y} = 0 \mid y = 0)$ é a probabilidade de um valor negativo ser negativo, ou seja um verdadeiro negativo.

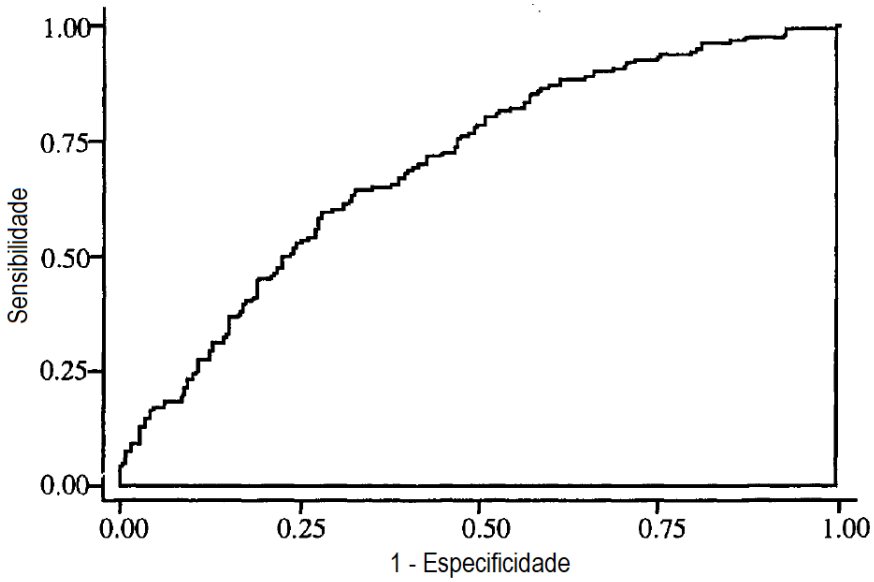
Tabela 1 – Classificação das observações

	Valor Real	
Predito	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

Que pode ser percebido através da Tabela 1, no qual VP é o Verdadeiro Positivo, ou seja, observações preditas como positiva e são positivas. Falso Positivo (FP) são observações negativas que foram estimadas como positivas, e da mesma forma para o Falso Negativo (FN), por fim, o Verdadeiro Negativo são observações negativas, classificadas como negativas.

Assim, a curva ROC pode ser dada por:

Figura 2 – Curva ROC calculada sobre o modelo



Fonte: imagem adaptada do livro Hosmer e Lemeshow (2000), pag. 163

A Figura 2 ilustra a Curva ROC de área de 0,6989, e Hosmer e Lemeshow (2000) utiliza como regra geral que a área equivalente a 0,5 sugere que não há efeito do modelo

regressivo. Para valores entre 0,7 e menores que 0,8 é considerável aceitável, entre 0,8 e menores que 0,9 é uma discriminação excelente e para áreas maiores ou iguais a 0,9 é excepcional.

Além da Curva ROC realizou-se o cálculo do Coeficiente de Determinação ajustado (R^2_{LS}). Essa medida é baseada nas informações obtidas na modelo saturado e no modelo apenas com o intercepto, que pode ser descrito por:

$$R^2_{LS} = \frac{L_0 - L_p}{L_0 - L_S} \tag{2.14}$$

no qual,

- L_S é o modelo saturado, ou seja, aquele com todos os parâmetros;
- L_0 é o modelo apenas com o intercepto;
- L_p é o modelo ajustado.

O Coeficiente de Determinação ajustado está entre 0 e 1, de modo que $L_0 < L_p < L_S$, espera-se sempre que o modelo ajustado seja próximo ao modelo saturado. Entretanto, essa proximidade não significa que o modelo seja bom, uma vez que o valor de L_p aumenta para cada parâmetro acrescido. Então espera-se encontrar a melhor combinação entre poucas variáveis e alta explicação.

Uma vez que foram satisfeitas as condições para um modelo adequado, foi possível realizar as interpretações dos coeficientes β, como indicado na equação de *Odds Ratio* (Equação 2.5), então realizar o intervalo de confiança para esses mesmos coeficientes (Equação 2.8).

3 Resultados

Os resultados apresentados neste relatório foram com base nos dados cedidos pela Empresa de Correios com informações de 11 de novembro de 2011 a 19 de setembro de 2017. Os registros indicam cada uma das viagens realizadas, no período, dentro do Brasil, contendo especificações da origem, do destino, da companhia aérea, da quantidade de carga transportada, bem como da lotação máxima e do quanto de carregamento que foi desembarcado no destino assinalado.

Para o objeto de estudo, também foi coletada a informação sobre a ocorrência de corte na carga. Dessa forma, foi possível identificar quais são as variáveis explicativas que podem justificar o fato de ter ocorrido corte na carga, e o quanto que elas realmente influenciam a chance de resposta.

Como primeira análise, foi realizado estudo base para verificar a real utilidade de determinada variável dentro do modelo, a fim de averiguar se o estudo faz, de fato, algum sentido para o texto técnico de Silva (2017) no qual será parte esta pesquisa.

3.1 Análise Exploratória

Os registros de previsão de chegada e de partida, atrasos, entre outras, foram retiradas antes de realizar o ajuste do modelo, uma vez que elas possuem resultados desatualizados ou muitos valores faltantes. Além disso, algumas dessas informações também não trariam nenhuma informação complementar para o modelo, como é o caso do *status* do voo, que apresenta todos os registros como *concluído*.

Os indicadores que não foram excluídos de imediato foram tratados de maneira mais severa na análise exploratória. Com isso, foi possível verificar indícios de multicolinearidade, eventuais interações, além de *outliers* que puderam indicar tanto erro no banco de dados, quanto algum tipo de informação relevante para os especialistas no assunto.

Desta forma, os resultados a seguir indicam as análises bivariada das variáveis que foram apontadas como candidatas a serem significantes. Os resultados obtidos foram apresentados em função da presença ou não do corte de carga. É importante informar que há seis vezes mais registros de situações nas quais não foram verificadas corte de carga.

3.1.1 Os Dados

Embora se tenha o registro diário das transportações realizadas, optou-se por realizar a análise mensal e anual das informações, em busca de possível padrão temporal para a ocorrência desses eventos.

Tabela 2 – Frequência de corte na carga por ano

Ano	Sem Corte	Com corte (%)
2011	1126	194 (14,7%)
2012	5255	1251 (19,2%)
2013	7671	1473 (16,1%)
2014	9147	1469 (13,8%)
2015	8740	1103 (11,2%)
2016	10032	619 (5,8%)
2017	5975	449 (7,0%)

Tabela 3 – Frequência de corte na carga por mês

Mês	Sem Corte	Com corte (%)
Janeiro	4441	522 (10,5%)
Fevereiro	3756	500 (11,7%)
Março	3863	564 (12,7%)
Abril	3584	591 (14,2%)
Maio	3812	666 (14,8%)
Junho	3994	510 (11,3%)
Julho	4400	588 (11,8%)
Agosto	4767	456 (8,7%)
Setembro	4037	397 (9,0%)
Outubro	3769	448 (10,6%)
Novembro	3606	516 (12,5%)
Dezembro	3917	800 (17,0%)

Os anos de 2011 e 2017 não estão completos, 2011 possui registros relativos aos meses de novembro e dezembro, enquanto 2017 tem registros até meados de Setembro. Isso significa que praticamente todos os meses aparecem o mesmo número de vezes, mas não necessariamente nos mesmos anos.

Mesmo com informações incompletas, a análise por ano (Tabela 2) mostra que de 2013 à 2016 o número de transportes com corte decresceu. Entretanto, para os meses não é possível verificar tal padrão na tabela 3 referente ao mês, embora dezembro tenha um valor de transportes com corte muito elevado e agosto e setembro tenham um resultado muito baixo para a mesma porcentagem, foram apenas esses resultados que podem indicar algum tipo de relação entre as variáveis.

Tabela 4 – Frequência de corte na carga na época de férias

Época	Sem Corte	Com corte (%)
Não Férias	27438	3638 (11,7%)
Férias	20508	2920 (12,5%)

Foi definido que o período de férias compreendia os meses de janeiro, fevereiro, junho, julho e dezembro que correspondem a aproximadamente 41% do total de meses. Optou-se por utilizar essa informação como possível parâmetro da regressão, uma vez que pode haver embasamento teórico sobre o potencial aumento de mercadorias consumidas via internet.

Além disso, para essa variável foi utilizada como referência o período de não férias, no qual foi possível analisar a resposta quando houve férias. Em outras palavras, como a variável é qualitativa, utilizou-se *dummys* para verificar sua real influência no modelo, e, dessa forma, fixou-se uma das classes dentro dessa variável para comparar com as demais, e, assim, averiguar se a nova classe é mais influente que a classe referência.

O critério para a definição da classe não é rigoroso, ou até mesmo não existe um critério fixo. Para este estudo foi definida como referência aquela que possuia uma das maiores diferenças absolutas e proporções entre as observações com corte e sem corte para cada uma das variáveis. Assim, buscou-se obter o referencial mais comum e a que conseguisse maior equilíbrio entre o corte de carga.

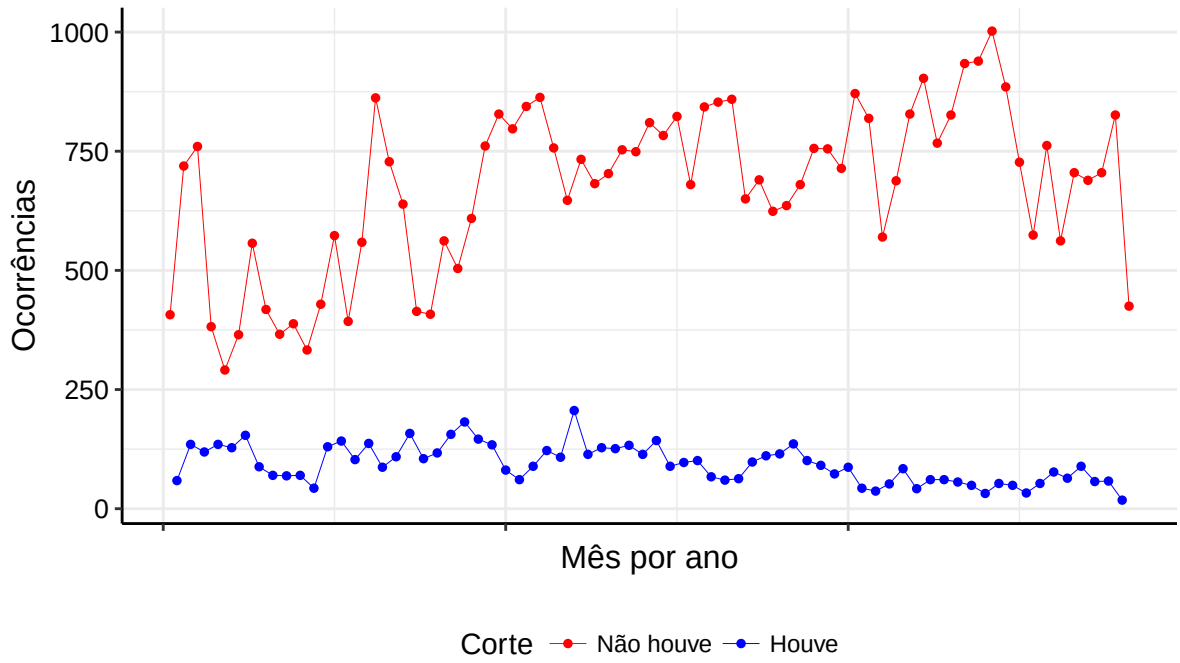


Figura 3 – Gráfico de ocorrências de corte e não corte por período

Na Figura 3, o eixo X representa o mês de cada ano separado pelo tipo do corte.

Ou seja, o primeiro ponto representa novembro de 2011, o segundo ponto, dezembro de 2011, e assim por diante até setembro de 2017. É possível verificar tanto a diferença entre as frequências observadas de transportes com corte e sem, quanto que nos últimos meses, a quantidade de transportes realizados segue em uma decrescente, para ambos os grupos.

3.1.2 Origem e Destino

As informações de origem e destino podem ser compactadas em apenas uma variável, chamada de "linha", uma vez que todas as informações de saída e aterrissagem formam uma triangulação.

Em outras palavras, um avião que saia do aeroporto A e tinha como destino o aeroporto B só poderia ir para o aeroporto C, ou seja, uma vez definido o percurso A para B, o outro destino possível deles seria apenas o C.

Entretanto, um avião que saia do aeroporto A poderia ir para o aeroporto D, e dessa forma, poderia apenas ter como um novo destino o aeroporto E. Então, os destinos e origens sempre formaram uma triangulação com outros dois aeroportos, porém, não foi obrigatório que ficasse apenas nesse triângulo, sendo possível a criação de novas origens e destinos, mas sempre mantendo essa ideia de trio.

Dessa forma, foi possível criar linhas (Tabela 5) para cada três aeroportos, reduzindo o número de variáveis, mas mantendo a informação, totalizando 25 linhas.

Tabela 5 – Frequência de corte na carga por linhas de origem e destino

Linha	Sem Corte	Com corte	Linha	Sem Corte	Com corte
L1	1253	130	L14	495	8
L2	2654	332	L15	855	303
L3	2215	122	L16	191	100
L4	1130	78	L17	269	74
L5	4461	835	L18	3611	291
L6	6061	1172	L19	1133	392
L7	4083	585	L20	467	75
L8	3	0	L21	0	1
L9	4226	193	L22	2	0
L10	4298	535	L23	0	3
L11	2283	562	L24	2	0
L12	2299	251	L25	1	0
L13	5954	506	-	-	-

Foi utilizado o mesmo critério das férias para a definição da variável de referência. E, embora não seja a maior proporção e nem a maior diferença, optou-se por realizar a comparação com a Linha 6, uma vez que a mesma, teve ambos os resultados bem elevados.

3.1.3 Companhia Aérea

De forma semelhante, a Companhia Aérea foi uma variável qualitativa que necessitou da criação de *dummys* para o ajuste ao modelo, e para o qual foi selecionada a melhor classe para a comparação na regressão. É importante dizer que as transportadoras na Tabela 6 estão todas codificadas por questão de sigilo.

Tabela 6 – Frequência de corte na carga por companhia aérea

Companhia Aérea	Sem Corte	Com corte	Companhia Aérea	Sem Corte	Com corte
Transportadora 1	17580	2514	Transportadora 7	2	0
Transportadora 2	2	0	Transportadora 8	3	0
Transportadora 3	7150	1430	Transportadora 9	1	0
Transportadora 4	0	1	Transportadora 10	0	3
Transportadora 5	3926	258	Transportadora 11	15658	2073
Transportadora 6	467	75	Transportadora 12	3157	204

Da mesma maneira realizada para a linha de origem e destino, ao calcular tanto a diferença entre corte e sem corte, quanto a proporção, para a companhia aérea, optou-se por definir a Transportadora 1 como padrão.

Mas não só isso, ao criar os modelos específicos para as Linhas, muitas das transportadoras ficaram de fora do modelo, já que, em sua maioria, cada rota de origem e destino continham até duas transportadoras.

3.1.4 Paletes

Paletes são suportes confeccionados com o objetivo de aumentar a capacidade de armazenagem, sendo utilizados para verticalizar o espaço, ou seja, multiplicar a capacidade de estoque. Além disso, auxiliam na racionalização do espaço e redução de danos em produtos.

Desta forma, a Tabela 7 mostra as medidas resumo do número de paletes por cada um dos tipos de corte.

Tabela 7 – Medidas resumo de corte na carga por paletes

Corte	Mínimo	Mediana	Média	Máximo
Não Houve	1,0	10,0	10,4	30,0
Houve	1,0	12,0	11,2	30,0

Essa parte da análise permitiu que fosse definida a quantidade de paletes máxima como 30, os resultados que apresentaram valores superiores foram excluídos, uma vez que valores superiores foram considerados erro de digitação. A Tabela 7 já apresenta os resultados corrigidos após a limpeza dos dados, com o objetivo de não afetar os valores da média e da mediana.

Foi possível verificar que tanto na média, quanto na mediana, a estatística obtida para a avaliação de paletes foi maior para as ocasiões em que houve corte na carga.

Entretanto, foi verificado que a variável palete não teve comportamento linear com a variável resposta do modelo, por isso, optou-se por transformar a variável em qualitativa, com quatro grupos definidos pelos quartis, a fim de conseguir a melhor aproximação para a modelagem.

3.1.5 Capacidade Máxima de Carga (Toneladas)

A tabela 8 apresenta os resultados obtidos da informação de capacidade máxima que a aeronave pode transportar de carga. Tanto os resultados de mediana quanto de média indicam em que para os registros que houve corte os valores de capacidade são maiores.

Tabela 8 – Medidas resumo de corte na carga por capacidade máxima de carga, em toneladas

Corte	Mínimo	Mediana	Média	Máximo
Não Houve	0,497	16,190	15,902	40,000
Houve	1,400	20,000	19,689	39,772

De maneira geral, assim como para os paletes, os resultados em que houve corte de carga foram superiores com exceção do máximo de carga observado, porém, foi estabelecido para essa modelagem que a capacidade máxima seria limitada em 40 toneladas.

Da mesma forma que os paletes, foi analisada a variável quanto ao seu comportamento em relação ao corte de carga. E ao separar em quartis, porém, diferente do resultado anterior, para a capacidade máxima de carga, não houve necessidade de transformar a variável em qualitativa, uma vez que foram verificadas sinais de que a relação com a variável resposta era linear.

3.1.6 Proporção Carga Transportada

Ao estudar as variáveis associadas à quantidade de carga transportada (capacidade máxima de carga, carga transportada e carga desembarcada), verificou-se, por meio da correlação de Spearman, a existência de uma forte associação entre elas, tanto para as observações com corte, quanto para as sem corte de carga.

Esse resultado poderia indicar interações e multicolinearidade no modelo regressivo, o que dificultaria a interpretação e sua aplicabilidade. Desta forma, optou-se por tomar as variáveis de Carga Transportada e de Carga Desembarcada e transformá-las em um novo indicador, calculado por meio da divisão de carga desembarcada pela transportada. No qual pode ser interpretada pela taxa de carga desembarcada em relação ao total da carga transportada.

Os resultados encontrados fora do intervalo 0 e 1 foram excluídos, uma vez que seria teoricamente impossível uma aeronave desembarcar mais carregamento do que o que foi levado.

Tabela 9 – Medidas resumo de corte na carga por proporção de carga de transporte

Corte	Mínimo	Mediana	Média	Máximo
Não Houve	0,0003	0,7374	0,6707	1,0000
Houve	0,0008	0,6012	0,6494	1,0000

E assim como todas as demais variáveis quantitativas, foi realizado estudo dos quartis para descobrir se a ocorrência de um comportamento linear do parâmetro. Entretanto, para a variável ponderação foi percebido grande presença de resultados com valor 1, representando que, a aeronave desembarcou toda carga que foi transportada.

Dessa forma, optou-se por também categorizar essa variável em quatro grupos. Com isso, ajustou-se uma regressão com todas as variáveis analisadas anteriormente, no qual buscou-se o modelo com melhor poder de previsão com a menor quantidade de indicadores.

3.2 Análise de Regressão

O modelo selecionado objetivou encontrar quais variáveis são responsáveis por alterar a chance de ocorrer ou não o corte de cargas nas aeronaves.

Dessa forma, os resultados apresentados a seguir apresentarão quais são os indicadores que proporcionam maiores resultados para a comparação com a metodologia AHP, de tal maneira, que as medidas a serem tomadas possam utilizar os resultados apresentados como base.

As precauções com correlação e multicolinearidade foram tomadas ao criar novas variáveis na seção de análise exploratória. A etapa seguinte consistiu em definir o modelo e realizar a análise e interpretação do mesmo.

3.2.1 Modelo Brasil

O primeiro modelo a ser ajustado, chamado de modelo Brasil, conteve todas as variáveis estudadas na seção de análise exploratória, foi dado esse nome uma vez que será estimado para todas as linhas do Brasil, independente de região. Após a validação dos dados, foram utilizados 54.504 observações.

Foi ajustado pelos métodos de seleção automático *Forward*, *Backward* e *Stepwise* por meio do teste da razão de verossimilhança, e os resultados foram semelhantes em todos os aspectos. Dessa forma, optou-se pelo *Stepwise*, apenas por ser o que possui o método de seleção mais completo.

Tabela 10 – Análise de efeitos do modelo Brasil ajustado

Parâmetros	Estatística	$p - valor$
Ano	760	<0,0001
Ferias	8,13	0,0043
Capacidade (mil)	1997	<0,0001
Pallete Quartil	953	<0,0001
Ponderacao Quartil	146	<0,0001
Transportadora	1688	<0,0001
Linha	1544	<0,0001

Todos os parâmetros foram significativos (Tabela 10) para um nível de significância de 5%. Ou seja, para o modelo Brasil ajustado, todos os indicadores foram considerados influentes para a explicação da variável corte de carga.

Primeiramente, verificou-se que os critérios para a convergência do modelo foram satisfeitos e como complemento da análise, também foram calculados os critérios de informação, como apresentado na tabela 11:

Tabela 11 – Critérios de decisão do modelo Brasil ajustado

Critério	Intercepto	Intercepto e Covariáveis
AICc	48 346	40 144
BIC	48 356	40 480
-2 log Verossimilhança	48 344	40 070

Além disso, esse modelo apresentou o coeficiente de determinação ajustado (R^2_{LS}) de 0,23, bem como a área da curva ROC com de 0,79, definida por Hosmer e Lemeshow (2000) como um ajuste aceitável. Porém, foi verificado que pelo teste de Hosmer & Lemeshow que o modelo não tem um bom ajuste, uma vez que a estatística encontrada foi de 8,13 com $p - valor$ de 0,004, no qual deve-se rejeitar a hipótese nula de que o modelo foi bem ajustado.

Uma vez que o modelo não foi satisfatório, optou-se por estimar outro modelo que pudesse explicar a situação especificada. Acredita-se que esse mau ajuste deve-se pelo fato do modelo Brasil conter muitas observações e por essas informações serem muito específicas para cada um dos fatores.

A alternativa encontrada foi a utilização da Tabela 5 para verificar qual linha que possui mais registros e com base nela, criar um novo banco de dados apenas com as informações da Linha 6, no qual manteve-se os demais indicadores a serem ajustados.

3.2.2 Modelo L6

Foi chamado de modelo L6, o modelo estimado com base no banco de dados com origem e destino exclusivos da Linha 6, porém, reparou-se também, que para essa linha operava apenas a Transportadora 11, então o modelo ajustado não utilizou das informações de Transportadora e Linha, já ambas não possuem nenhuma discriminação.

Para esse ajuste, foram utilizados 7.878 observações, e os resultados das variáveis significativas, incluindo o valor de cada uma das *dummys* acrescidas ao modelo, estão a seguir (Tabela 12):

Tabela 12 – Análise de efeitos do modelo L6 ajustado

Parâmetros	Estimativa	Erro Padrão	Estatística	$p - valor$
β_0 Intercepto	1126	60,40	347,8	<0,0001
β_1 Ano	-0,56	0,030	351,8	<0,0001
β_2 Ferias	0,14	0,038	14,4	0,0002
β_3 Capacidade (t)	0,33	0,025	185,5	<0,0001
β_{4_2} Palete Quartil 2	-0,30	0,116	6,6	0,0101
β_{4_3} Palete Quartil 3	0,89	0,065	192,9	<0,0001
β_{4_4} Palete Quartil 4	0,29	0,093	10,1	0,0015
β_{5_1} Ponderação Quartil 1	0,38	0,080	23,4	<0,0001
β_{5_3} Ponderação Quartil 3	-0,61	0,106	33,7	<0,0001

Pela tabela, ao analisar o $p - valor$, foi possível perceber que apenas a variável *dummy* da ponderação pelo segundo quartil não entrou no modelo. Também verificou-se que algumas variáveis apresentaram estimativas negativas, indicando que uma relação inversa do parâmetro com a chance de ocorrência do corte na carga.

Os quartis estão definidos pelos seguintes valores (Tabela 13):

Tabela 13 – Quartis por paletes

Posição	Paletes	Ponderação
Primeiro Quartil	6	0,275
Segundo Quartil	8	0,55
Terceiro Quartil	9	0,98

Os resultados da Tabela 14 apresentam os critérios de informação que foram utilizados como uma das estatísticas para a decisão desse modelo como o mais adequado.

Tabela 14 – Critérios de decisão do modelo L6 ajustado

Critério	Intercepto	Intercepto e Covariáveis
AICc	6 857	4 548
BIC	6 854	4 617
-2 log Verossimilhança	6 855	4 528

As estatísticas encontradas para o Critério de Informação de Akaike corrigido e o Bayesiano não foram as menores, porém optou-se por esse modelo em razão de apresentar valores próximos e mais parcimonioso, e da mesma forma acontece para o cálculo do logaritmo da verossimilhança.

O teste global de beta, no qual a hipótese nula de que $\beta = 0$ foi rejeitado para todos os parâmetros, indicando que não há evidência para afirmar que as variáveis do modelo não explicam a variável resposta.

Além disso, o teste de Hosmer & Lemeshow, que antes implicou na não utilização do modelo Brasil (Tabela 10), agora, teve $p - valor$ de 0,0793, no qual não se rejeita a hipótese de que o modelo foi bem ajustado.

Então, dessa forma, deu-se prosseguimento à análise de resíduos, para o qual não foi encontrado nenhuma anormalidade, seja pelo Pearson Studentizado que verifica a qualidade da previsão das observações do modelo, ou pela Distância de Cook's para a influência de cada uma das observações sob cada um dos coeficientes encontrados (Tabela 12).

Para finalizar a verificação da qualidade do modelo ajustado, calculou-se duas estatísticas, a área sob a curva ROC e o coeficiente de determinação. O primeiro, pela

definição de Hosmer e Lemeshow (2000), é dita como uma discriminação excelente, com resultado de 0,88 de área, como indica a figura 3.2.2:

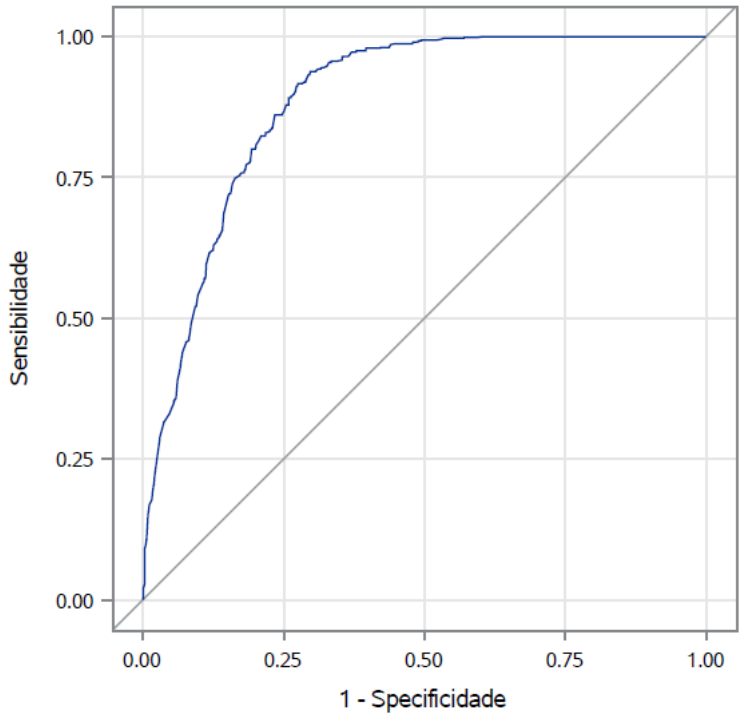


Figura 4 – curva ROC do modelo L6 ajustado

O coeficiente de determinação encontrado foi de 0,44. Embora a estatística não seja a mais adequada, o resultado baixo pode ser explicado por uma possível falta de variáveis explicativas que pudessem aumentar os critérios de discriminação do modelo.

Ademais, como se trata de estudo Brasil, não necessariamente o coeficiente seja baixo, nem alto, mas sim um valor a ser computado para futuras comparações com modelos mais completos.

3.2.3 Diagnóstico

Tabela 15 – Razão de chances estimadas para o modelo L6

Parâmetros	Odds Ratio	Intervalo de Confiança	
		Limite Inferior	Limite Superior
β_1 Ano	0,570	0,537	0,604
β_2 Ferias	1,331	1,148	1,543
β_3 Capacidade (t)	1,397	1,331	1,466
β_{4_2} Palete Quartil 2	1,817	1,241	2,660
β_{4_3} Palete Quartil 3	6,008	4,573	7,894
β_{4_4} Palete Quartil 4	3,285	2,346	4,600
β_{5_1} Ponderação Quartil 1	1,062	0,853	1,321
β_{5_3} Ponderação Quartil 3	0,389	0,290	0,522

De maneira geral a interpretação das variáveis quantitativas é mais fácil, uma vez que é feita a análise em relação ao próprio valor do parâmetro. Para cada ano que passa a chance de ocorrer o corte na carga é 57% da chance do ano anterior. Ou seja, para anos mais recentes e possíveis anos futuros, a chance de ocorrer corte na carga reduz em 57% em relação ao ano anterior.

Já para a capacidade máxima de carga, que será estudado a cada tonelada, a influência é contrária, pois para cada uma tonelada acrescida na capacidade máxima da aeronave a chance de ocorrer corte na carga aumenta em 39,7 %.

Por outro lado, para as variáveis qualitativas, a interpretação deve ser feita em relação a um dos fatores dessa variável. Para o período de férias, utilizou-se como referência o resultado não férias, então, para o período de férias a chance de ocorrer corte na carga é 1,331 vezes a chance no período de não férias.

No caso da variável Palete foi utilizado como referência o primeiro quartil, dessa forma, ao comparar com o Palete de segundo quartil a chance de ocorrer o corte de carga no Quartil 2 é 1,817 vezes a chance do primeiro quartil, e essa mesma interpretação pode ser feita para os Quartis 3 e 4, da mesma variável, que seguem, respectivamente, 6,008 e 3,285 vezes a chance de ocorrer a variável de interesse do que qual é um Palete do primeiro quartil.

De acordo com Agresti (2002), se um parâmetro contém, em seu intervalo de confiança, o valor 1 esse parâmetro não será influente no modelo. Por mais que pelo método de seleção a variável *dummy* de Ponderação do primeiro quartil tenha entrado, a análise dela será desconsiderada, uma vez que o intervalo de confiança é de 0,853 à 1,321.

Portanto, o terceiro quartil foi o único quartil da variável Ponderação que foi selecionado como influente para o modelo, isso ao compará-los com o quarto quartil. No qual a chance do evento de interesse reduz para as observações que pertencem ao terceiro quartil, sendo 0,389 vezes a chance do quarto quartil.

De maneira geral, os indicadores coletados e seus efeitos serão utilizados na comparação com a AHP, de forma que, por exemplo, a decisão pelo uso de aeronaves com maior capacidade aumenta a chance de ocorrer corte na carga, então tem-se insumos para a análise de risco em função do uso dessas aeronaves.

4 Conclusão

Após definir que o evento de interesse seria a ocorrência de corte na carga e a validação do banco de dados a fim de retirar valores incorretos, para não afetar a estimação, ajustou-se dois modelos de regressão logística. No qual, teve-se por objetivo, estimar os principais indicadores e qual o efeito desses para o modelo, para que assim, os resultados encontrados pudessem ser comparados com a metodologia AHP realizada pelos Correios.

Os indicadores encontrados servirão como base para o aprimoramento da metodologia utilizada pela Empresa para a realização do serviço de transporte de encomendas, e dessa forma, possibilitando redução nos custos e otimizando o tempo gasto nos voos.

O primeiro modelo composto por todas as observações do banco de dados rejeitou, pelo teste de Hosmer & Lemeshow, a hipótese de que foi um modelo bem ajustado. Além disso, o R^2 ajustado não apresentou alta determinação, com coeficiente de 0,23.

Embora as demais estatísticas, tenham sido aceitáveis e o modelo tenha incluído todas as variáveis, optou-se por realizar novo ajuste, no qual analisou-se a linha que continha mais observações, tanto de corte na carga quanto de sem o corte na carga. Dessa forma, chegou em um novo banco de dados que incluiu apenas as observações que continham a Linha 6, consequentemente, apenas a Transportadora 11, já que dentro da Linha 6, apenas houve voos da Transportadora 11.

Esse novo modelo, chamado de L6, apresentou bom ajuste, tanto pelo teste de Hosmer & Lemeshow, quanto pelo teste de $\beta = 0$. Além disso, o estudo pelos critérios de informação também representaram boas estatísticas, embora não as mais adequadas, mas suficientemente boas para o estudo inicial.

Pela análise do Pearson Studentizado e da distância de Cook's, que permitem, respectivamente, analisar a qualidade da previsão de cada uma das observações do modelo ajustado e a influência dessas observações para os parâmetros de L6. Concluiu-se que não houve nenhuma observação que afete o ajuste do modelo.

Também, verificou-se que o coeficiente de determinação de 0,44 é aceitável para a situação específica, no qual tem-se poucos indicadores e é um estudo inicial e com área sobre a curva ROC de 0,88, considerada uma discriminação excelente de acordo com Hosmer e Lemeshow (2000).

Por fim, os resultados da razão de chances indicam que a chance de ocorrer corte na carga para Capacidade Máxima (a cada tonelada) aumenta a medida que aumenta esse limite. Entretanto, para a variável ano, a medida que o ano aumenta, a chance de ocorrer corte na carga reduz, mantendo as demais variáveis constantes.

Para os outros parâmetros foi feita a análise em relação a uma variável fixa. Então, para férias, palete e ponderação, fixou-se os eventos de, respectivamente, não férias, primeiro quartil e de quarto quartil.

Quando acontece o evento de férias a probabilidade de acontecer corte na carga aumenta em relação ao evento de não férias e da mesma maneira acontece para os Paletes de Quartis 2, 3 e 4 ao comparar com o primeiro quartil, alterando apenas a intensidade dessa probabilidade.

Para a variável de ponderação, optou-se por desconsiderar a influência do Quartil 1, uma vez que o intervalo de confiança contém 1, ou seja, é dito que não há influência entre as variáveis. E para as ocorrências de evento do terceiro quartil, a chance de ocorrer corte na carga é menor do que para os eventos de quarto quartil.

De maneira geral, o modelo L6 foi bem ajustado, porém, para a obtenção de resultados melhores e mais preciso, seria necessário realizar coleta mais abrangente de variáveis. E se possível obter mais informações que possam descriminar melhor o modelo.

Referências

AGRESTI, A. *Categorical Data Analysis*. 2nd. ed. [S.l.]: John Wiley & Sons, Inc., 2002. Citado 3 vezes nas páginas 21, 23 e 41.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 8th. ed. [S.l.]: Saraiva, 2013. Nenhuma citação no texto.

HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. 2nd. ed. [S.l.]: Wiley-Interscience Publication, 2000. Citado 4 vezes nas páginas 27, 37, 39 e 43.

KUTNER, M. H. et al. *Applied Linear Statistical Models*. 5th. ed. [S.l.]: McGraw-Hill/Irwin, 2005. Nenhuma citação no texto.

LOTTES, I. L.; DEMARIS, A.; ADLER, M. A. Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology, Vol 24, No 3*, July – 1996. Nenhuma citação no texto.

OLIVERIA, A. V. M.; PAMPLONA, D. A.; FILHO, D. P. Estudo de previsão de demanda do trabalho urbano coletivo público na região metropolitana de são paulo. *Revista de transportes públicos - ANTP*, 2015. Nenhuma citação no texto.

PAULA, G. A. Modelo de regressão com apoio computacional. Não publicado. 2013. Citado na página 26.

POHLMAN, J. T.; LEITNER, D. W. A comparison of ordinary least squares and logistic regression. *The Ohio Journal of Science, Vol 103, No 5*, December 2003. Nenhuma citação no texto.

RODRIGUES, A. S. *Regressão Logística com erro de medida: comparação de métodos de estimação*. Dissertação (Mestrado) — Instituto de matemática e estatística da Universidade de São Paulo, 2013. Nenhuma citação no texto.

SAATY, R. W. The analytic hierarchy process - what it is and how it is used. *Pergamon Journals Lrd*, 1987. Citado na página 17.

SILVA, J. M. da. *Auditoria baseada em riscos aplicada na gestão de transportes de carga da Empresa Brasileira de Correios e Telégrafos (Correios): um estudo de caso*. [S.l.], 2017. Citado 2 vezes nas páginas 17 e 29.

WANKE, P.; FLEURY, P. F. Transporte de cargas no brasil: Estudo exploratório das principais variáveis relacionadas aos diferentes modais e às suas estruturas de custo. *Research Gate*, 2006. Nenhuma citação no texto.