UNIVERSIDADE FEDERAL DE MINAS GERAIS
Departamento de Estatística - ICEx
Programa de Especialização em Estatística

Fernanda França e Souza

ANÁLISE DE REGRESSÃO LOGÍSTICA PARA PREDIZER A EVASÃO DE COLABORADORES DE UMA EMPRESA

Belo Horizonte
2023

Fernanda França e Souza

Regressão Logística

Versão 1

Monografia de especialização apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Orientadora: Ilka Afonso Reis

Belo Horizonte
2023

Souza, Fernanda França e

S729a Análise de regressão logística para predizer a evasão de colaboradores de uma empresa [manuscrito] / Fernanda França e Souza— 2023.
69.f. il.

Orientadora: Ilka Afonso Reis.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 68-69.

1. Estatística. 2. Regressão logística. 3. Colaboradores - Empresa – Desligamento. I. Reis, Ilka Afonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 269º. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE FERNANDA FRANÇA E SOUZA.

Aos vinte dias do mês de dezembro de 2022, às 14:30 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Fernanda França e Souza**, intitulado: "Análise de Regressão Logística para Predizer a Evasão de Colaboradores de uma Empresa", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Ilka Afonso Reis – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada por unanimidade condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje. O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 20 de dezembro de 2022.

Prof.ª Ilka Afonso Reis (Orientadora)
Departamento de Estatística / ICEX / UFMG

Prof.ª Edna Afonso Reis
Departamento de Estatística / ICEX / UFMG

DEDICATÓRIA

Dedico este trabalho ao meu irmão caçula, Lucas, que tão cedo partiu deixando toda nossa família sem chão. Irmão que em pouco tempo nos mostrou como ser gentil, bondoso e carinhoso. Aquele que sempre me incentivou a seguir em frente e me elogiava por eu ser quem sou.
Lucas, a cada momento de dificuldade ou que pensei em desistir, saiba que me lembrava dos seus elogios e como preciso ser forte e te deixar orgulhoso, de onde quer que você esteja.

Te amo!

AGRADECIMENTOS

Agradeço à Deus por ter me dado forças de finalizar esse TCC em um ano tão doloroso. Agradeço aos meus pais Luiz e Flaviani por serem exemplo de força e determinação. Ao meu irmão Renato com seu otimismo e força de vontade. Ao meu namorado, Vinícius, por me incentivar e me acalmar nos momentos de ansiedade. Amo vocês!
Agradeço também ao meu amigo Diego que foi um presente da Especialização em Estatística, com quem dividi vários momentos durante o curso e agora faz parte da minha vida.
Agradeço à professora Ilka que foi minha inspiração para a escolha do tema de Regressão para o TCC, agradeço por sua paciência principalmente.
Por último, aos familiares e amigos que estiveram comigo durante este período.

RESUMO

O objetivo do presente trabalho foi encontrar um modelo de regressão logística que pudesse estimar qual a probabilidade de um colaborador deixar uma empresa em função das variáveis explicativas- nível de satisfação, nota da última avaliação, número de projetos, número médio de horas trabalhadas por mês, tempo de trabalho na empresa, ocorrência de acidente de trabalho, promoção nos últimos 5 anos, departamento, faixa salarial (baixo, médio, alto) e tipo de trabalho (campo, escritório e laboratório). Para isso, foram estudados 4 modelos diferentes de regressão, o Modelo 1, o modelo 2, o modelo 3 e o modelo 4, os quais foram criados com modificações em algumas variáveis como departamento e fator interação. Todos os modelos foram divididos em base de treino e de teste para avaliar qual modelo possuía o melhor ajuste e previsibilidade. Os modelos foram avaliados em relação à área sobre a curva ROC (AUC). O modelo com maior AUC foi o Modelo 4 (87,09). Foi possível observar que as variáveis Número médio de horas trabalhadas e Faixa salarial contribuem positivamente para saída do colaborador; já as variáveis Nível de satisfação, Número de projetos e Tempo trabalhado diminuem a chance de saída da empresa. Nota-se também que colaboradores com ocorrência de acidente de trabalho (78,57% menor) e promoção nos últimos 5 anos (76,04% menor) são menos propensos a sair da empresa. Aqueles colaboradores que trabalham em laboratório (34,87% menor) ou escritório (6,83% menor) têm menos chance de saída da empresa se comparados aos que trabalham em campo. Quando tempo trabalhado na empresa é baixo, a impacto de notas altas consiste na diminuição da chance de saída do colaborador. Quando o tempo de casa é maior, notas altas aumentam a chance de saída do colaborador. Quando a nota da última avaliação é baixa, o aumento do tempo trabalhado consiste na diminuição da chance de saída do colaborador. No entanto, quando a nota da última avaliação é máxima, o aumento do tempo trabalhado consiste no aumento da chance de saída. Por fim, mesmo com uma boa capacidade de ajuste ROC e previsibilidade, existem variáveis que não são mensuradas e podem ter influência na saída ou não do colaborador, como por exemplo, o fator psicológico.

Palavras-chave: regressão logística; evasão empresarial

ABSTRACT

The objective of this study was to find a logistic regression model that could estimate the probability of an employee leaving a company based on the explanatory variables - level of satisfaction, last evaluation score, number of projects, average number of hours worked per month , working time at the company, occurrence of work accident, promotion in the last 5 years, department, salary range (low, medium, high) and type of work (field, office and laboratory). For this, 4 different regression models were studied, Model 1, Model 2, Model 3 and Model 4, which were created with modifications in some variables such as department and interaction factor. All models were split on a training and test basis to assess which model had the best fit and predictability. The models were evaluated in relation to the area under the ROC curve (AUC). The model with the highest AUC was Model 4 (87.09). It was possible to observe that the variables Average number of hours worked and Salary range contribute positively to the employee's departure; the variables Level of satisfaction, Number of projects and Time worked decrease the chance of leaving the company. It is also noted that employees with an accident at work (78.57% lower) and promotion in the last 5 years (76.04% lower) are less likely to leave the company. Those employees who work in the laboratory (34.87% smaller) or office (6.83% smaller) are less likely to leave the company compared to those who work in the field. When time worked in the company is low, the impact of high grades is a decrease in the employee's chance of leaving. When seniority is longer, high grades increase the employee's chance of leaving. When the last evaluation score is low, the increase in the time worked consists of a decrease in the employee's chance of leaving. However, when the last evaluation score is maximum, the increase in the time worked means an increase in the chance of leaving. Finally, even with a good ability to adjust ROC and predictability, there are variables that are not measured and can influence whether or not the employee leaves, such as the psychological factor.

Keywords: logistic regression; business evasion

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC	Area Under Curve
ROC	Receiver Operating Characteristics

SUMÁRIO

1 INTRODUÇÃO

Uns dos principais ativos de uma empresa são os colaboradores, uma vez que esses detêm conhecimento técnico, formam a cultura e compartilham da visão e valores de uma empresa. O desenvolvimento de uma organização está diretamente atrelado aos seus colaboradores. Sendo assim, desligamentos (voluntários ou não) provocam impactos na rotina, funcionamento e desenvolvimento de uma companhia.

O desligamento de um colaborador gera custos à empresa, que podem ser tangíveis - como recrutamento e seleção, encargos e treinamentos - e intangíveis, como conhecimento, entregas de projetos e processos (OLIVEIRA et al., 2018).

Como forma de mensuração de movimento de empregados, as empresas utilizam-se de um indicador denominado *turnover*, termo costumeiramente usado em ambientes empresariais. Este indicador mede o número de colaboradores desligados em determinado período em comparação com o quadro de colaboradores ativos (MARRAS, 2002).

Para alguns setores, como o de tecnologia e pesquisa, existe claramente um desafio para a retenção de talentos, uma vez que há um grande investimento de tempo e dinheiro que é destinado a qualificação da equipe. Em sua pesquisa, Patias et al. (2015) defendeu que, para setores da economia que não possuem um processo de treinamento estruturado e não goza de treinamento com alta qualificação, o gasto com treinamento corresponde a 24% dos custos de desligamento e contratação de um novo empregado, considerando apenas gastos tangíveis. Empresas com alto grau de especialização podem ter esse percentual ainda maior.

Diante do exposto, faz-se extremamente importante estudar o fenômeno do *turnover*, com o objetivo de identificar as possíveis causas que levam os funcionários a serem desligados, traçar planos de ação assertivos a fim de minimizar a saída de colaboradores e atrair novos funcionários aderentes a cultura empresarial (BERTOTTI, 2013).

Com o avanço de conhecimento notório em tecnologias de tratamento de dados e a maior utilização dessas técnicas por empresas, a partir de coleta de dados é possível fazer uma anamnese de uma organização e definir o perfil desejado, realizar uma modelagem que possa identificar os principais fatores que causam desligamentos a fim de reduzi-lo e criar

mecanismos de previsibilidade a serem usados no momento da contratação. Dessa forma, empresas podem se tornar mais atrativas, obter maior retenção de talentos e realizar orçamentos assertivos à realidade.

1.1 Objetivo

O objetivo do presente trabalho é, a partir de um banco de dados real, aplicar a metodologia de Regressão Logística para identificar os principais fatores que influenciam a saída de colaboradores da empresa.

2 DESENVOLVIMENTO

2.1 A Regressão Logística

A Regressão Logística é um dos métodos estatísticos mais populares na análise de diversas variáveis independentes na previsão ou explicação de uma variável dependente dicotômica. De acordo com, Gonzalez (2018), a regressão logística é um método de pesquisa valioso, uma vez que pode ser aplicado em diferentes contextos:

- Estimação da probabilidade de um evento ocorrer baseado em variáveis independentes;
- Estimação da probabilidade de um evento ocorrer em relação a probabilidade de ele não ocorrer a partir de uma observação (chance);
- Análise do impacto de uma variável em relação à variável resposta – se uma variável contribui mais do que outra para a ocorrência do evento;
- Classificação das observações como mais suscetíveis ou não à ocorrência do evento estudado de acordo com valores das variáveis independentes.

2.2 A Regressão Logística Binária

Assim como a Regressão Linear, a Regressão Logística Binária pode avaliar múltiplas variáveis independentes. No entanto, como dito anteriormente, na regressão logística, a variável resposta é uma variável categórica dicotômica.

Uma vez que a variável dependente (Y) é dicotômica, ela segue a distribuição de Bernoulli, no qual Y = 1, em caso de sucesso, e Y = 0, em caso de fracasso. Na regressão logística, a estimação da probabilidade de sucesso (p) é dada a partir da combinação linear de variáveis independentes por meio da função logística, f(z), que é definida por (SOUZA, 2006):

$$f(z) = \frac{1}{1+ e^{-z}} \qquad (1)$$

Intervalo: $0 \leq f(z) \leq 1$

Onde: $z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

Assim,

$$f(z) = \frac{1}{1+ e^{-(\alpha+ \sum \beta_i X_i)}} \qquad (2)$$

A função logística *f(z)* varia no intervalo de 0 a 1. Assim a interpretação do modelo logístico se torna mais atraente quando deseja-se avaliar a probabilidade de um evento ocorrer ou não (sucesso/fracasso) em função de diversas variáveis independentes (KLEINBAUM & KLEIN., 2010).

Os coeficientes α e β´s são parâmetros desconhecidos.

Logo, a probabilidade de um evento ocorrer dadas as variáveis independentes X´s é dado por P(X):

$$P(X) = \frac{1}{1+ e^{-(\alpha+ \sum \beta_i X_i)}} \qquad (3)$$

Dessa forma, a regressão logística identifica a combinação linear das variáveis independentes com maior probabilidade da variável resposta (Y) assumir o valor 1, ou seja, de o evento ser o sucesso. (STOLZFUS, 2011). Diferentemente da regressão linear, a regressão logística não estima os parâmetros baseado no método dos mínimos quadrados e sim a utilizando o método da máxima verossimilhança (KLEINBAUM & KLEIN, 2010).

2.3 Razão de Chances (*Odds Ratio*)

A chance ou *odds* de um evento é a probabilidade de um evento ocorrer dividido pela probabilidade de esse evento não ocorrer (SAIANI , 2011).

Assim, a equação (4) expressa a chance de o evento ocorrer dado que o vetor de variáveis explicativas $X=X_0$.

$$\frac{P(X_0)}{1-P(X_0)} = \frac{1}{1+ e^{-(\alpha+ \sum \beta_i X_i)}} \cdot \frac{1+ e^{-(\alpha+ \sum \beta_i X_i)}}{1+ e^{-(\alpha+ \sum \beta_i X_i)-1}} = \frac{1}{e^{-(\alpha+ \sum \beta_i X_i)}} = e^{(\alpha+ \sum \beta_i X_i)} \qquad (4)$$

Definindo como $\frac{P(X_1)}{1-P(X_1)}$ a chance de ocorrência do evento dado que o vetor de explicativas vale X=X_1, então a Razão de Chances (RC) comparando as duas situações (X=X_0 e X=X_1) é dada por

$$RC = \frac{P(X_1)}{1-P(X_1)} \Big/ \frac{P(X_0)}{1-P(X_0)} \qquad (5)$$

Supondo que o modelo regressão logística seja simples, ou seja, tenha somente uma variável explicativa X, e que seu coeficiente no modelo seja β, substituindo a equação (4) na equação (5), temos

$$RC = e^{(\alpha+\beta X_1)} \cdot e^{-(\alpha+\beta X_0)} = e^{\alpha+\beta X_1-\alpha-\beta X_0} = e^{\beta X_1-\beta X_0} = e^{\beta(X_1-X_0)} \quad (6)$$

Ou seja, a cada aumento de (X_1-X_0) unidades na variável X, a chance de sucesso será multiplicada por $e^{\beta(X_1-X_0)}$. No caso de X ser uma variável dicotômica definindo dois grupos e valores 0 (grupo 0) ou 1 (grupo 1), então a chance de sucesso do grupo 1 (X=1) será e^{β} vezes a chance de sucesso do grupo 0 (X=0). Sendo assim, e^{β} será a razão de chances comparando o grupo 1 com o grupo 0.

2.4 Curva ROC (Receiver Operating Characteristics)

Uma das maneiras de avaliar a qualidade de predição de um critério de classificação é usar a chamada curva ROC.

O primeiro passo para o entendimento da curva ROC é definir os conceitos de especificidade e sensibilidade. Ambas são medidas de acerto de um critério de classificação.

A Sensibilidade corresponde à proporção de observações classificadas como positivas dentre aquelas que são realmente positivas e a Especificidade corresponde à proporção de observações negativas dentre aquelas que são realmente negativas (BEWICK, 2004).

No caso da regressão logística, a probabilidade de sucesso é predita por meio da aplicação do modelo estimado aos valores das variáveis independentes. Para predizer a variável resposta (sucesso ou fracasso), usa-se um ponto de corte, p_c, para avaliar a probabilidade de sucesso predita pelo modelo. Se essa probabilidade predita for maior ou igual a p_c, então a predição para a variável resposta é o evento sucesso. Caso contrário, a predição é o evento fracasso. Assim, para cada ponto de corte, podemos calcular o valor da sensibilidade e da especificidade do

critério de classificação usando aquele valor. Variando-se o ponto de corte entre 0 e 1, teremos vários pares de valores de sensibilidade e especificidade.

O gráfico dos valores da sensibilidade (taxa de verdadeiros positivos) versus os valores de (1 – especificidade), também chamada de taxa de falsos positivos, é chamado de *Receiver Operating Characteristics Curve* ou Curva ROC (BEWICK, 2004). A performance de um modelo pode ser mensurada a partir a área abaixo da curva ROC (AUC). Um critério de classificação perfeito possuiria AUC= 1.

Assim, a curva ROC é uma ferramenta que pode ser utilizada para avaliar a performance de um modelo de classificação como a regressão logística e pode ser um critério de comparação entre modelos: quanto maior o valor da área abaixo da curva ROC, melhor é a qualidade de predição do modelo.

3 MATERIAIS E MÉTODO

3.1 Banco de Dados

O banco de dados utilizado neste trabalho é um banco de dados extraído do site Kaggle (https://www.kaggle.com). O banco de dados contém informações de 14.599 colaboradores de uma empresa não identificada e sem período identificado.

A unidade de análise e o colaborador(a) da empresa.

As variáveis quantitativas coletadas foram: nível de satisfação, nota da última avaliação, número de projetos, média de horas trabalhadas por mês, tempo de empresa em anos. No banco também há variáveis categóricas, tais como envolvimento em acidente de trabalho (sim ou não), promoção nos últimos 5 anos (sim ou não), departamento e faixa salarial.

A variável resposta do problema é dicotômica, informando se o (a) colaborador(a) saiu da empresa ou não.

O Quadro 1 descreve as variáveis explicativas (x) e a variável resposta (y) associada a cada observação.

Quadro 1. Variáveis do banco de dados

Id	**Nome no banco**	**Nome**	**Tipo de variável**	
x1	satisfaction_level	Nível de satisfação (min-máx)	Numérica	Contínua
x2	last_evaluation	Nota da última avaliação (min-máx)	Numérica	Contínua
x3	number_project	Número de projetos	Numérica	Discreta
x4	average_montly_hours	Número Médio de horas trabalhadas por mês	Numérica	Contínua
x5	time_spend_company	Tempo trabalhado, em anos	Numérica	Discreta
x6	Work_accident	Ocorrência de acidente de trabalho	Categórica	Dicotômica
x7	promotion_last_5years	Promoção nos últimos 5 anos	Categórica	Dicotômica
x8	Department	Departamento (RH, IT, Mng, Mkt, Pro, RandD, Vendas, Spt, Tec, Cont)	Categórica	Nominal
x9	salary	Faixa salarial (baixo, médio e alto)	Categórica	Ordinal
x10	Tipo de trabalho	Campo, escritório e laboratório	Categórica	Nominal
y	**left**	**Saída da empresa**	**Categórica**	**Dicotômica**

Foi criada uma décima variável explicativa (x10) chamada Tipo de Trabalho. Os departamentos foram divididos em três grupos: escritório, laboratório e campo. Os colaboradores alocados como escritório são as pessoas dos departamentos: *accounting*, HR, IT, management, product_mng e marketing. As pessoas alocadas no tipo laboratório são as pessoas do departamento: RandD (*Research and Development*). As pessoas alocadas no tipo *field* são as pessoas do departamento: *sales, support, tecnical.*

3.2 Método

Para avaliar quais os fatores podem influenciar a saída de um colaborador da empresa estudada, foram construídos quatro modelos diferentes de Regressão Logística com a finalidade de identificar aquele com melhor performance. Para todas as análises, o software estatístico utilizado foi o R (versão 3.6.3) com o apoio da interface gráfica do R Studio (versão 3.6.3). A função utilizada para a construção dos modelos de Regressão Logística foi a função "glm" do R e, para a confecção das curvas ROC, foi utilizado o pacote "dplyr" e "pROC".

Diferentes modelos estatísticos foram criados com o objetivo de determinar aquele com melhor ajuste e previsibilidade. Além disso, foi avaliado o efeito das interações das variáveis explicativas e a comparação dos diferentes modelos desenvolvidos.

Para o ajuste dos modelos, a base de dados foi dividida aleatoriamente em base de teste e base de treino. A base de teste possui 66% das observações e a base de treino, 34%.

O primeiro modelo ajustado continha todas as variáveis independentes disponíveis. Uma vez construído um modelo a partir da base de treino, a significância dos seus coeficientes foi testada usando um nível de significância de 5%. As variáveis consideradas significativas foram retidas no modelo:

i. No Modelo 1, foram utilizadas todas as 9 variáveis explicativas.
ii. No Modelo 2, a variável Departamento (x8) foi removida e foi adicionada a variável tipo de trabalho (x10).
iii. No Modelo 3, foi inserido um fator de interação entre as variáveis tipo de Trabalho (x10) e acidente de trabalho (x6).
iv. No Modelo 4, foi inserido um fator de interação entre nota da última avaliação (x2) e tempo trabalhado (x5).

A área sob a curva ROC (AUC) foi calculada para cada modelo desenvolvido utilizando o conjunto de teste com o objetivo de avaliar a qualidade das predições do modelo e compará-los.

Com o modelo final, foram calculadas as razões de chances para as variáveis retidas no modelo, assim como seus respectivos intervalos de 95% de confiança.

4 RESULTADOS

4.1 Análise Descritiva

As Tabelas 1 e 2 apresentam a análise descritiva das variáveis explicativas em relação à saída ou não da empresa.

Na Tabela 1 vemos que, para a variável última avaliação, verifica-se que não existe diferença significativa entre as médias da nota da última avaliação entre os colaboradores que saíram e não saíram da empresa.

Os colaboradores que não saíram da empresa tiveram uma média superior somente para a variável nível de satisfação.

Para as variáveis médias de horas trabalhadas por mês e tempo trabalhado, os colaboradores que saíram da empresa possuem uma média superior.

Para variável número de projetos, verifica-se que as médias amostrais são numericamente iguais (com uma casa decimal), porém as médias populacionais foram consideradas estatisticamente diferentes a 5% de significância.

Tabela 1. Estatísticas descritivas para os dados das variáveis quantitativas: média e desvio-padrão (entre parênteses)

Característica	***Total de colaboradores***		***Saída da empresa***				
			Não		***Sim***		***P- valor***
	n = 14999	*100*	*11428*	*76.2 %*	*3571*	*23.8 %*	
Nível de satisfação	*0.61 (0.25)*		*0.67 (0.22)*		*0.44 (0.26)*		< 0.0001
Última Avaliação	*0.72 (0.17)*		*0.72 (0.16)*		*0.72 (0.20)*		*0.*4683
Média de horas trabalhadas por mês	*201.05 (49.94)*		*199.06 (45.68)*		*207.42 (61.20)*		< 0.0001
Tempo trabalhado em anos	*3.498 (1.460)*		*3.380 (1.562)*		*3.876 (0.978)*		< 0.0001
Número de projetos	*3.8 (1.2)*		*3.8 (1.0)*		*3.8 (1.8)*		0.03034

Figura 1. Histograma da variável nível de satisfação por grupo (saída da empresa (1) ou não (0))

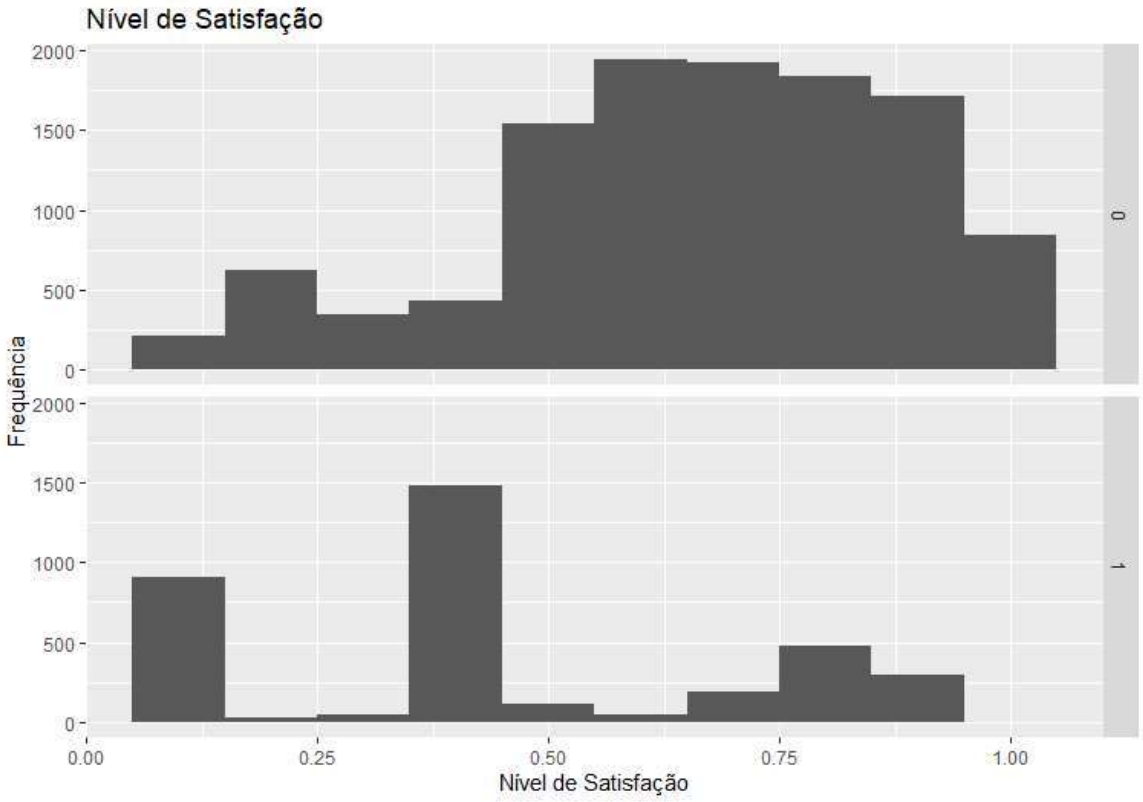


Figura 2. Histograma da variável última avaliação por grupo (saída da empresa (1) ou não (0))

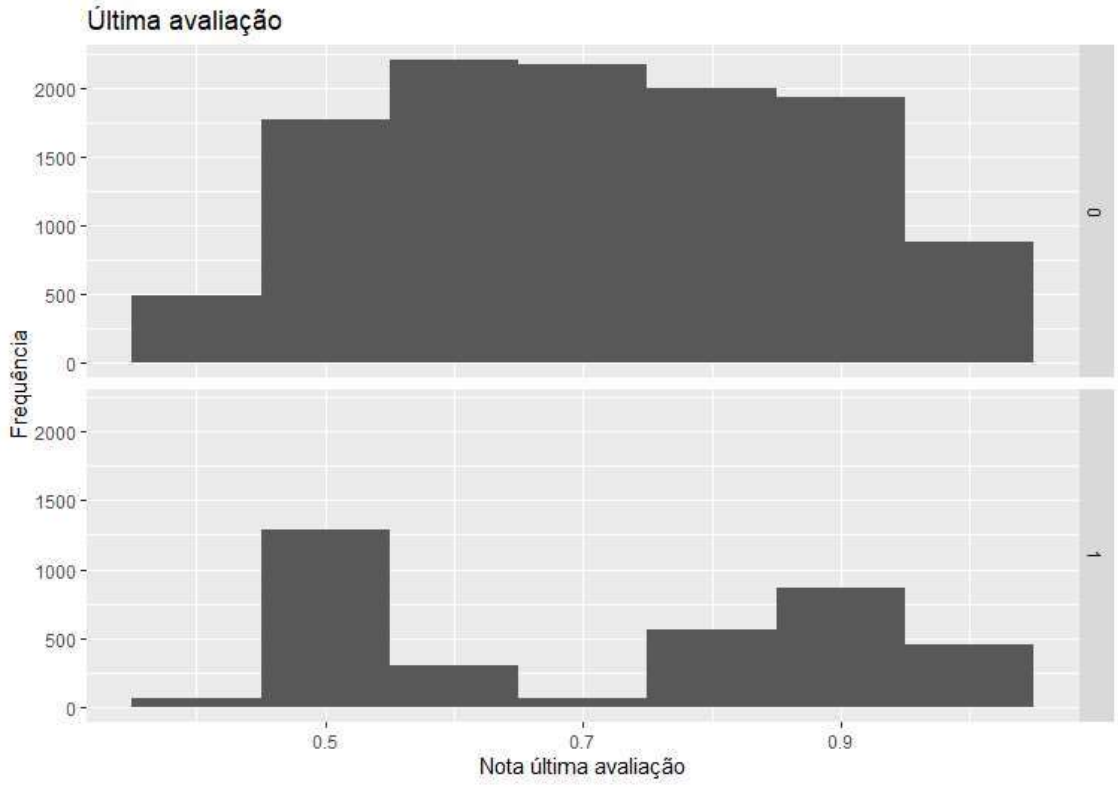


Figura 3. Histograma da variável média de horas trabalhadas por grupo (saída da empresa (1) ou não (0))

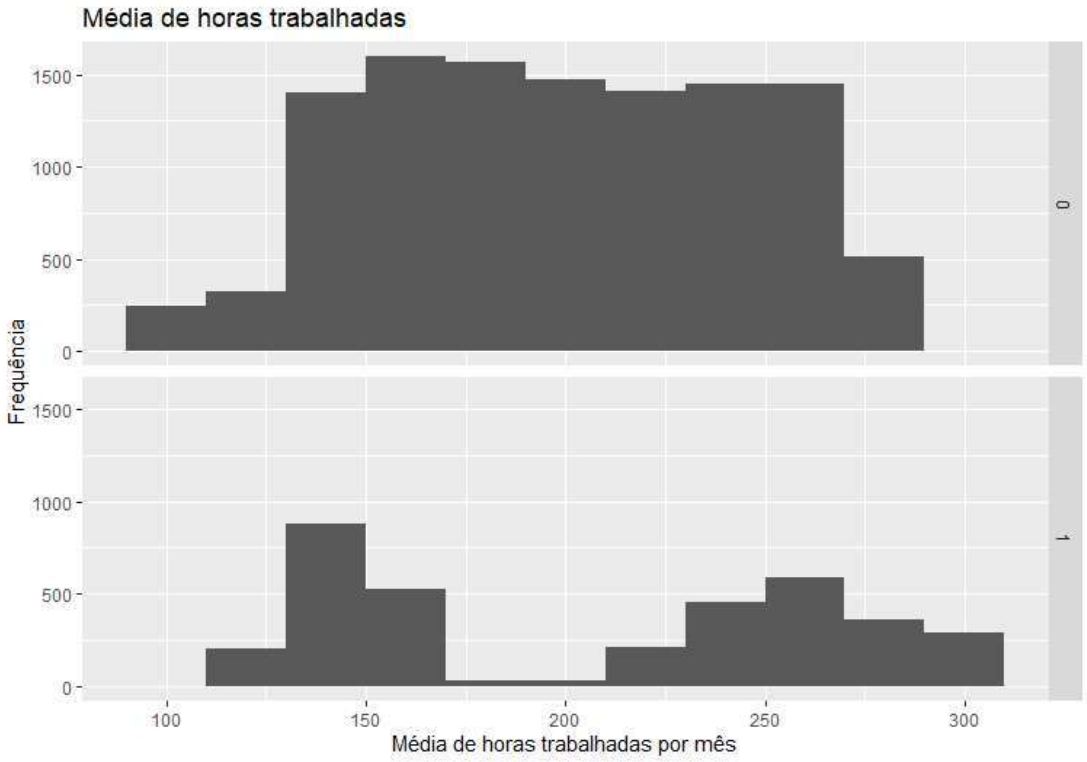


Figura 4. Histograma da variável tempo trabalhado em anos por grupo (saída da empresa (1) ou não (0))

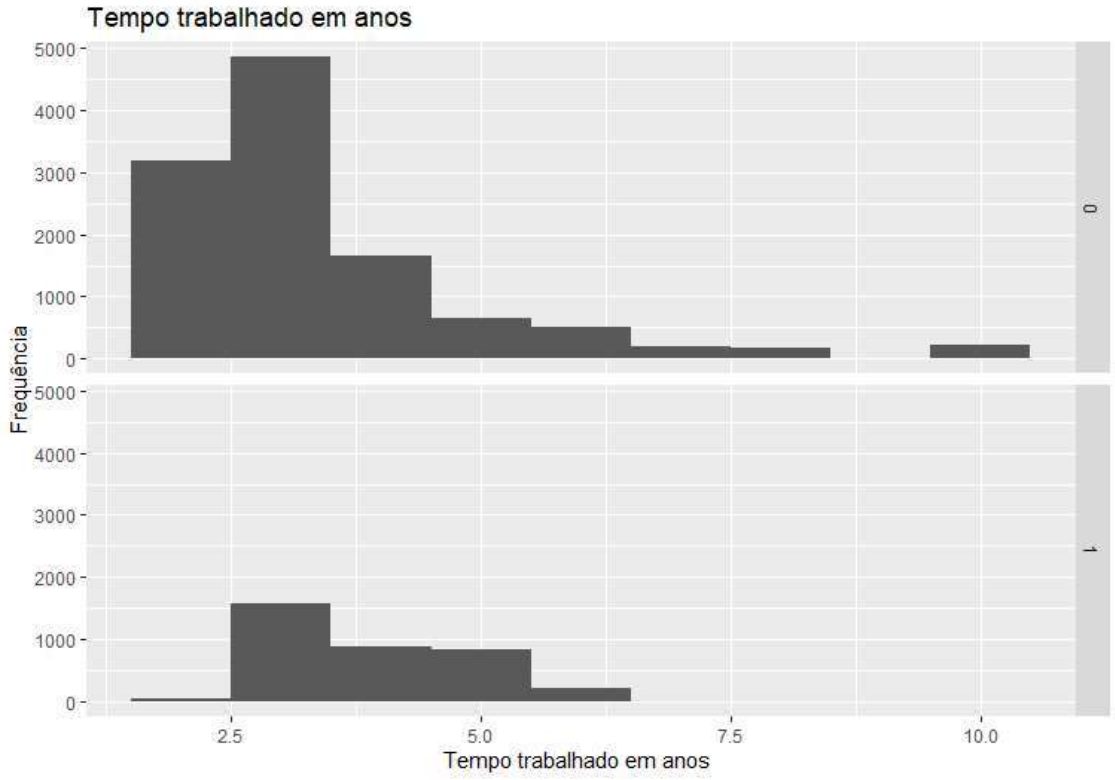
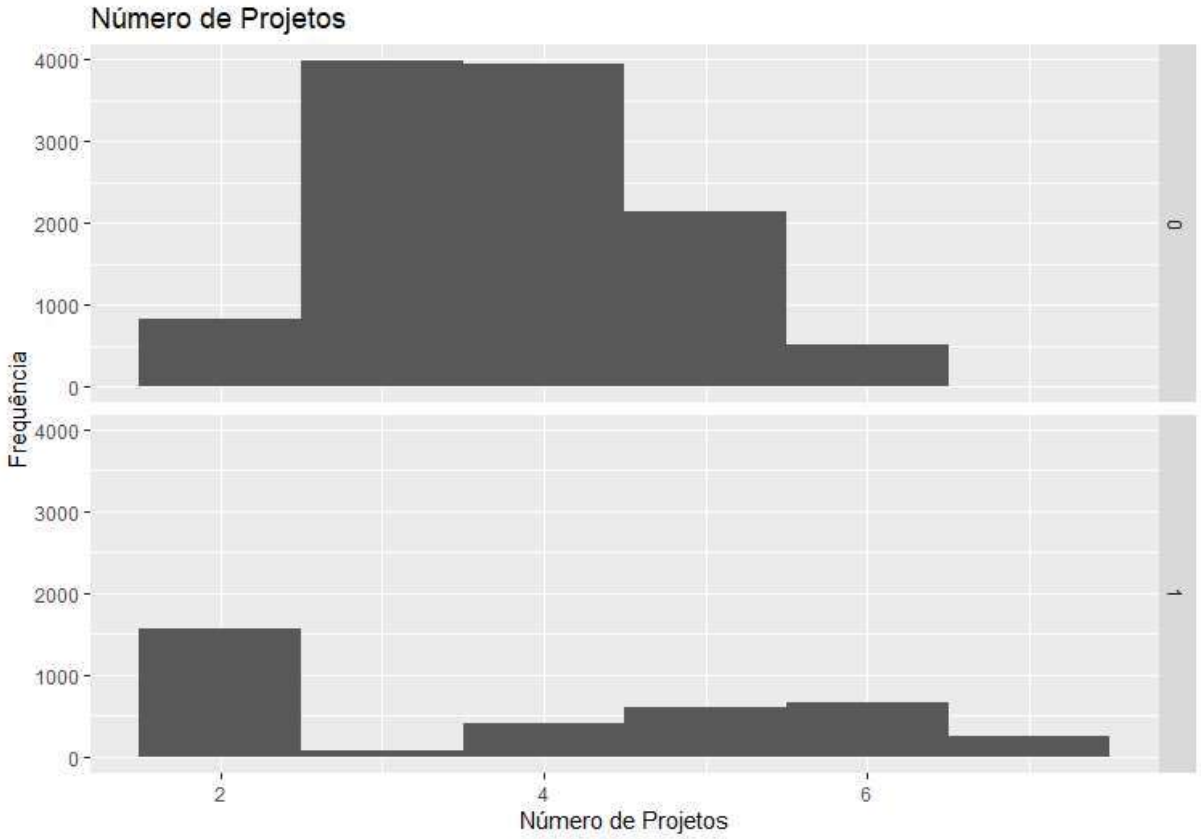


Figura 5. Histograma da variável número de projeto por grupo (saída da empresa (1) ou não (0))



Na Tabela 2, observa-se que os colaboradores que não saíram da empresa são maioria em todos os departamentos. Os departamentos de Gestão e RandD possuem o menor percentual de evasão, comparado aos outros departamentos.

Analisando a variável Faixa salarial, nota-se que o os colaboradores que não saíram da empresa, são maioria em todas as faixas salariais (baixa média e alta). No entanto, a faixa salarial alta possui um percentual de saída menor do que as outras 2 faixas.

Para a variável Acidente de trabalho, os colaboradores que não saíram da empresa são maioria tanto para os colaboradores que tiveram quanto para os que não tiveram acidente de trabalho. Colaboradores que sofreram acidente de trabalho possuem um percentual menor de saída da empresa quando comparado aos que não sofreram.

No caso de Promoção nos últimos 5 anos, a maior parte dos colaboradores da empresa não recebeu promoção. Quando se analisa os colaboradores que saíram da empresa, aqueles sem promoção são maioria em relação aos colaboradores que tiveram promoção.

Tabela 2. Descrição das variáveis categóricas

Característica	*Total de colaboradores*	*Saída da empresa*		
		Não	*Sim*	*P- valor*
Departamento				*< 0.001*
Gestão	*630 (4.20%)*	*539 (85.56%)*	91 (14.44%)	
Recursos Humanos	*739 (4.93%)*	*524 (70.91%)*	215 (29.09%)	
Contabilidade	*767 (5.11%)*	*563 (73.40%)*	204 (26.60%)	
RandD	*787 (5.25%)*	*666 (84.63%)*	121 (15.37%)	
Marketing	*858 (5.72%)*	*655 (76.34%)*	203 (23.66%)	
Produção	*902 (6.01%)*	*704 (78.05%)*	198 (21.95%)	
TI	*1227 (8.12%)*	*954 (77.75%)*	273 (22.25%)	
Suporte	*2229 (14.86%)*	*1674 (75.10%)*	555 (24.90%)	
Técnico	*2720 (18.13%)*	*2023 (74.38%)*	697 (25.63%)	
Vendas	*4140 (27.60%)*	*3126 (75.51%)*	1014 (24.49%)	
Faixa salarial				*< 0.001*
Baixa	*7316 (48.78%)*	*5144 (70.31%)*	2172 (29.69%)	
Média	*6446 (42.97%)*	*5129 (79.57%)*	1317 (20.43%)	
Alta	*1237 (8.25%)*	*1155 (93.37%)*	82 (6.63%)	
Acidente de trabalho				*< 0.001*
Não	*12830 (85.54%)*	*9428 (73.48%)*	3402 (26.52%)	
Sim	*2169 (14.46%)*	*2000 (92.21%)*	169 (7.79%)	
Promoção nos últimos 5 anos				*< 0.001*
Não	*14680 (97.87%)*	*11128 (75.80%)*	3552 (24.20%)	
Sim	*319 (2.13%)*	*300 (94.04%)*	19 (5.96%)	

A Tabela 3 apresenta a distribuição de frequências da variável de saída da empresa nas bases de Treino e de Teste. Nota-se que em ambas as bases a maior parte dos dados refere-se aos colaboradores que não saíram da empresa. Nota-se também que a proporção entre colaboradores que saíram e não saíram da empresa são numericamente similares nas duas bases.

Tabela 3. Distribuição de frequências da variável de saída da empresa nas bases de Treino e de Teste.

Saiu da empresa?	*Frequência Treino (%)*	*Frequência Teste (%)*
Não	7527 (76.04)	3901 (76.49)
Sim	2372 (23.96)	1199 (23.51)
Total	**9899 (100)**	**5100 (100)**

A Tabela 4 apresenta os resultados da função VIF para avaliação de multicolineriedade das variáveis explicativas. Os resultados apresentados para todas as variáveis explicativas são maiores do que 1 e menores do que 5, indicando algum tipo de correlação, mas não o suficiente para serem retiradas do modelo.

Tabela 4. Coeficientes estimados para o Modelo 1 (predição do evento Sair da Empresa).

Descrição	*GVIF*	*DF*	*GVIF^(1/(2*Df))*
Nível de satisfação	1,159	1	1,077
Nota da última avaliação	1,442	1	1,201
Número de projetos	1,786	1	1,336
Número Médio de horas trabalhadas por mês	1,524	1	1,234
Tempo trabalhado	1,120	1	1,058
Ocorrência de acidente de trabalho	1,011	1	1,006
Promoção nos últimos 5 anos	1,017	1	1,008
Departamento	1,052	9	1,003
Salário	1,048	2	1,012

4.2 Desenvolvimento do Modelo de Regressão Logística

A Tabela 5 apresenta os coeficientes estimados para as variáveis do Modelo 1 (todas as explicativas). Verifica-se que a variável departamento não é significativa a 5%.

Tabela 5. Coeficientes estimados para o Modelo 1 (predição do evento Sair da Empresa).

ID	*Descrição*	*Coeficientes Estimados*	*Desvio Padrão*	*Valor z*	*Valor-p)*
	(Intercept)	-1,531	0,239	-6,395	<0,0001
x1	Nível de satisfação	-4,180	0,121	-34,442	<0,0001
x2	Nota da última avaliação	0,690	0,184	3,745	<0,0001
x3	Número de projetos	-0,309	0,026	-11,761	<0,0001
x4	Número Médio de horas trabalhadas por mês	0,004	0,001	6,489	<0,0001
x5	Tempo trabalhado	0,271	0,019	14,084	<0,0001
x6	Ocorrência de acidente de trabalho	-1,581	0,110	-14,331	<0,0001
x7	Promoção nos últimos 5 anos	-1,591	0,338	-4,705	<0,0001
x8	DepartmentoRH	0,249	0,161	1,543	0,1227
	DepartmentoIT	-0,114	0,151	-0,757	0,4491
	DepartmentoMng	-0,298	0,195	-1,524	0,1274
	DepartmentoMkt	0,122	0,161	0,759	0,4478
	DepartmentoPro	-0,135	0,161	-0,84	0,4007
	DepartmentRandD	-0,485	0,175	-2,762	0,0058
	DepartmentoVendas	0,060	0,126	0,481	0,6308
	DepartmentoSpt	0,141	0,134	1,049	0,2944
	DepartmentoTec	0,104	0,131	0,79	0,4298
	DepartamentoCont	----	----	----	-----
x9	Faixa Salarial Low	2,038	0,160	12,723	<0,0001
	Faixa Salarial Med	1,481	0,161	9,184	<0,0001

A Tabela 6 apresenta os coeficientes estimados para as variáveis do Modelo 2 (sem a variável Departamento e com a inserção da variável Tipo de trabalho). Verifica-se que todas as variáveis são significativas a 5%, exceto Tipo de Trabalho_Escritório em relação à Tipo de trabalho_Campo.

Tabela 6. Coeficientes estimados do Modelo 2 (predição do evento Sair da Empresa).

ID	*Descrição*	*Coeficientes Estimados*	*Desvio Padrão*	*Valor z*	*Pr(>|z|)*
	(Intercept)	-1,301	0,208	-6,268	<0,0001
x1	Nível de satisfação	-4,227	0,122	-34,548	<0,0001
x2	Nota da última avaliação	0,833	0,183	4,546	<0,0001
x3	Número de projetos	-0,325	0,026	-12,418	<0,0001
x4	Média de horas trabalhadas por mês	0,004	0,001	6,138	<0,0001
x5	Tempo trabalhado	0,266	0,019	13,832	<0,0001
x6	Ocorrência de acidente de trabalho	-1,497	0,108	-13,827	<0,0001
x7	Promoção nos últimos 5 anos	-1,580	0,321	-4,921	<0,0001
x9	Faixa Salarial Low	1,900	0,149	12,775	<0,0001
X9	Faixa Salarial Med	1,394	0,150	9,311	<0,0001
x10	Tipo de trabalho_Lab	-0,782	0,146	-5,354	<0,0001
x10	Tipo de trabalho_Escritório	-0,086	0,058	-1,479	0,139

A Tabela 7 apresenta os coeficientes estimados para as variáveis do Modelo 3 (mesmas variáveis do Modelo 2 com adição de interação entre as variáveis Tipo de trabalho e Acidente de trabalho). Nota-se que a interação entre Tipo de trabalho e Acidente de trabalho não é significativa a 5% de significância. Logo tal coeficiente é estatisticamente igual a zero e o termo de interação foi retirado do modelo.

Tabela 7. Coeficientes Estimados do Modelo 3 (predição do evento Sair da Empresa).

ID	*Descrição*	*Coeficientes Estimados*	*Desvio Padrão*	*Valor z*	*Pr(>|z|)*
	(Intercept)	-1,724	0,213	-8,077	<0,0001
x1	Nível de satisfação	-4,027	0,121	-33,321	<0,0001
x2	Nota da última avaliação	0,799	0,184	4,348	<0,0001
x3	Número de projetos	-0,291	0,026	-11,111	<0,0001
x4	Média de horas trabalhadas por mês	0,004	0,001	6,892	<0,0001
x5	Tempo trabalhado	0,266	0,019	13,873	<0,0001
x6	Ocorrência de acidente de trabalho	-1,586	0,145	-10,959	<0,0001
x7	Promoção nos últimos 5 anos	-1,481	0,320	-4,635	<0,0001
x9	Faixa Salarial Low	2,007	0,157	12,782	<0,0001
x10	Faixa Salarial Med	1,477	0,158	9,350	<0,0001
	Tipo de trabalho_Lab	-0,605	0,149	-4,073	<0,0001
	Tipo de trabalho_Escritório	-0,038	0,060	-0,636	0,525
x6*x10	Work_accident1:Type_workLab	-0,002	0,581	-0,004	0,997
	Work_accident1:Type_workOffice	-0,396	0,259	-1,527	0,127

A Tabela 8 apresenta os coeficientes estimados para as variáveis do Modelo 4 (mesmas variáveis do Modelo 2 com adição de interação entre as variáveis Nota da última avaliação e Tempo trabalhado). Nota-se que a interação entre Nota da última avaliação e Tempo trabalhado é significativa e, portanto, será mantida no modelo.

Tabela 8. Coeficientes Estimados do Modelo 4 (predição do evento Sair da Empresa).

ID	*Descrição*	*Coeficientes Estimados*	*Desvio Padrão*	*Valor z*	*Pr(>|z|)*
	(Intercept)	8,617	0,479	17,992	<0,0001
x1	Nível de satisfação	-4,660	0,131	-35,484	<0,0001
x2	Nota da última avaliação	-13,160	0,608	-21,660	<0,0001
x3	Número de projetos	-0,332	0,028	-11,738	<0,0001
x4	Média de horas trabalhadas por mês	0,005	0,001	7,742	<0,0001
x5	Tempo trabalhado	-2,642	0,128	-20,673	<0,0001
x6	Ocorrência de acidente de trabalho	-1,540	0,118	-13,040	<0,0001
x7	Promoção nos últimos 5 anos	-1,429	0,337	-4,235	<0,0001
x9	Faixa Salarial Low	2,094	0,171	12,259	<0,0001
	Faixa Salarial Med	1,601	0,172	9,309	<0,0001
x10	Tipo de trabalho_Lab	-0,429	0,143	-3,004	0,003
	Tipo de trabalho_Escritório	-0,071	0,061	-1,156	0,248
*x2*x5*	Nota da última avaliação*Tempo trabalhado	3,899	0,165	23,600	<0,0001

4.2.1 Escolha do Modelo Final

A escolha do modelo final baseou-se na comparação entre as áreas sob a Curva ROC de cada um dos modelos construídas usando-se a base de dados de teste.

A Tabela 9 apresenta os valores da AUC e seus respectivos intervalores de confiança de 95%. Nota-se que o Modelo 4 foi o modelo com a melhor performance preditiva.

Tabela 9. Área sobre a curva ROC (AUC) dos quatro modelos ajustados e respectivos intervalos de 95% de confiança.

	AUC	***95% IC***
Modelo 1	81,76%	80,46% - 83,05%
Modelo 2	81,91%	80,65% - 83,17%
Modelo 3	82,10%	80,71% - 83,27%
Modelo 4	87,44%	86,83% - 88,05%

4.2.2 Interpretação dos resultados do Modelo 4

A equação do modelo escolhido é dada por:

$$\ln\left(\frac{p}{1-p}\right) = 8{,}617 - 4{,}660\,x_1 - 13{,}16x_2 - 0{,}332x_3 + 0{,}05x_4 - 2{,}642x_5 - 1{,}540x_6 - 1{,}429x_7 + 2{,}094x_{9(Low)} + 1{,}601x_{9(Med)} - 0{,}429x_{10(Lab)} - 0{,}071x_{10(Office)} + 3{,}899x_2 * x_5$$

A Tabela 10 mostra os resultados para o Modelo 4, utilizando toda base de dados, as estimativas pontuais e intervalares para a razão de chances.

Tabela 10. Resultados do Modelo 4

ID	*Descrição*	*Coeficientes Estimados*	*Desvio Padrão*	*Valor z*	*Pr(>|z|)*	*Razão das Chances*	*2.5 %*	*97.5 %*
	(Intercept)	8,617	0,479	17,992	<0,0001	-	-	-
x1	Nível de satisfação	-4,660	0,131	-35,484	<0,0001	0,009	0,007	0,012
x2	Nota da última avaliação	-13,160	0,608	-21,660	<0,0001	-	-	-
x3	Número de projetos	-0,332	0,028	-11,738	<0,0001	0,717	0,678	0,758
x4	Média de horas trabalhadas por mês	0,005	0,001	7,742	<0,0001	1,005	1,004	1,007
x5	Tempo trabalhado	-2,642	0,128	-20,673	<0,0001	-	-	-
x6	Ocorrência de acidente de trabalho	-1,540	0,118	-13,040	<0,0001	0,214	0,169	0,269
x7	Promoção nos últimos 5 anos	-1,429	0,337	-4,235	<0,0001	0,240	0,119	0,448
x9	Faixa Salarial Low	2,094	0,171	12,259	<0,0001	8,121	5,869	11,477
	Faixa Salarial Med	1,601	0,172	9,309	<0,0001	4,959	3,575	7,022
x10	Tipo de trabalho_Lab	-0,429	0,143	-3,004	0,003	0,651	0,490	0,858
	Tipo de trabalho_Escritório	-0,071	0,061	-1,156	0,248	0,932	0,826	1,050
*x2*x5*	Nota da última avaliação*Tempo trabalhado	3,899	0,165	23,600	<0,0001	49,377	35,845	68,508

Nota-se que as variáveis Média de horas trabalhadas por mês, Faixa Salarial possuem uma influência positiva para saída do funcionário.

Como mostrado na Tabela 10, a razão de chances para cada variável foi calculada. Abaixo, foram interpretadas aquelas razões de chances significativamente diferentes de 1 a 5% de significância.

- **Nível de satisfação (x1):** Para cada aumento de 0,1 do nível de satisfação a chance de o colaborador sair da empresa reduz em aproximadamente 37,25%[1].
- **Número de projetos (x3):** Para cada aumento unitário do número de projetos a chance do colaborador sair da empresa reduz em 75,38%.
- **Média de horas trabalhadas por mês (x4):** Para cada aumento unitário da média de horas trabalhadas por mês, a chance de o colaborador sair da empresa aumenta 100,53%.

[1] Para efeitos didáticos, os cálculos das variações nas chances são detalhados a seguir para cada variável.
Nível de satistifação (x1): 100*[exp(-4,66*0,1)-1 = 37,25%
Número de Projetos (x3): 100*[exp(0,717 - 1)] = 75,38%
Média de horas trabalhadas por mês (x4): 100*exp(1,005-1) = 100,53%
Ocorrência de acidente de trabalho (x6): (1- 0,214) = 78,57%
Promoção nos últimos 5 anos (x7): (1- 0,669) = 76,04%
Salário (x9):Faixa salarial baixa = OR = 8,12 /Faixa salarial alta= OR = 4,96
Tipo de trabalho (x10):Laboratório = (1-0,651) = 34,87% / Escritório = (1-0,932) = 6,83%

- **Ocorrência de acidente de trabalho (x6):** Os colaboradores que já tiveram algum acidente de trabalho possuem 78,57% a menos de chance de saírem da empresa, comparado àqueles que não tiveram acidente.
- **Promoção nos últimos 5 anos (x7):** Os colaboradores promovidos nos últimos 5 anos possuem 76,04% menos chance de saírem da empresa comparado aos que não foram promovidos nos últimos 5 anos.
- **Salário (x9):** a comparação ocorre em relação à faixa salarial alta.
    - Os colaboradores que possuem uma faixa salarial baixa, possuem 8 vezes mais chance de deixar a empresa, comparado aos colaboradores com faixa salarial alta.
    - Os colaboradores que possuem uma faixa salarial média, possuem aproximadamente 5 vezes mais chances de deixar a empresa, comparado aos colaboradores com faixa salarial alta.
- **Tipo de trabalho (x10):** a comparação ocorre em relação aos colaboradores que trabalham no campo (field)
    - Colaboradores que trabalham no laboratório possuem 34,87% menos chance de sair da empresa comparado aos colaboradores do campo.
    - Os colaboradores do escritório possuem 6,83% menos chances de saírem comparados ao mesmo grupo.

Por causa do termo de interação entre as variáveis nota da última avaliação (x2) e tempo trabalhado em horas (x5), a estimativa da razão de chances para uma variável deve considerar os valores da outra variável. Assim, foram utilizados os valores mínimo e máximo de uma variável para o cálculo da estimativa da razão de chances da outra variável.

Considerando a nota mínima para a última avaliação (x2= 0,36), a cada aumento de um ano no tempo trabalhado, a chance de sair da empresa diminui em 71,01%[2]. No entanto, considerando a nota máxima para a última avaliação (x2=1), a cada aumento de um ano no tempo trabalhado, a chance de sair da empresa aumenta em 251,49%.

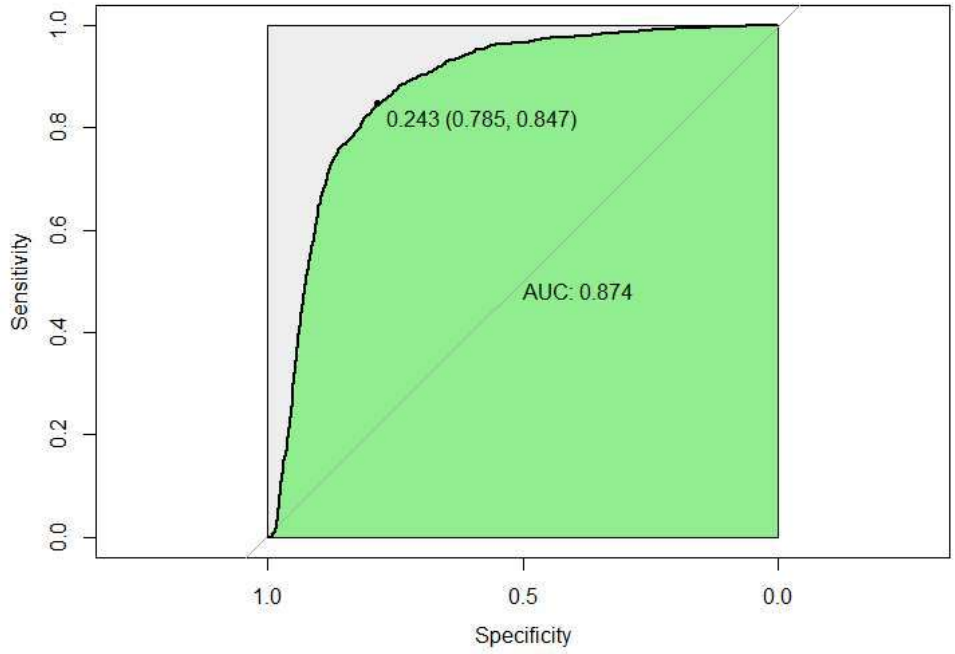
[2] Para efeitos didáticos, os cálculos das variações nas chances de cada uma das variáveis que interagem são detalhados a seguir.
Tempo de trabalho (x5) quando a nota da última avaliação (x2) vale 0,36: exp(-2,642+3,899*1*0,36)-1= -71,01%.
Tempo de trabalho (x5) quando a nota da última avaliação (x2) vale 1: exp(-2,642+3,899*1*1)-1= 251,49%.
Nota da última avaliação (x2) quando tempo de trabalho (x5) vale 2 anos: exp(-13,160*0,1+3,899*0,1*2)-1= -41,5%.
Nota da última avaliação (x2) quando tempo de trabalho (x5) vale 10 anos: exp(-13,160*0,1+3,899*0,1*10)-1= 12,24.

Considerando o tempo mínimo trabalhado na empresa (x5=2 anos), a cada aumento de 0,10 ponto na nota da última avaliação, a chance de sair da empresa diminui em 41,5%. No entanto, considerando o tempo máximo trabalhado na empresa (x5=10 anos), a cada aumento de 0,10 ponto na nota da última avaliação, a chance de sair da empresa aumenta quase 12 vezes (OR=12,24).

4.2.3 Curva ROC e AUC do Modelo 4

A Figura 6 mostra a curva ROC para as predições do Modelo final (modelo 4) quando aplicado ao conjunto de teste. Como visto na Tabela 08, o valor de AUC para esse modelo é de 87,44%, indicando uma boa qualidade preditiva.

Figura 6. Curva ROC para as predições do Modelo 4 quando aplicado ao conjunto de teste.



O ponto de corte para o qual existe a maior sensibilidade em relação à sensibilidade dado pelo ponto de 78,5% de sensibilidade e 84,7% de 1-especificidade.

5 CONCLUSÃO

Este trabalho investigou a influência das variáveis: nível de satisfação (x1), nota da última avaliação (x2), número de projetos (x3), média de horas trabalhadas por mês (x4), tempo trabalhado (x5), ocorrência de acidente de trabalho (x6), promoção nos últimos 5 anos (x7), departamento (x8) e faixa salarial (x9) na ocorrência da saída do colaborador da empresa. Utilizando um banco de dados com 14999 registros, foi utilizada a técnica de regressão logística para modelar a probabilidade de saída do colaborador da empresa em função das variáveis citadas.

Foi possível observar que as variáveis Médias de horas trabalhadas e Faixa salarial contribuem positivamente para saída do colaborador. Foi observado também que os colaboradores de faixa salarial baixa e média são mais propensos a sair da empresa quando comparados aos colaboradores de faixa salarial alta, assim como aqueles que possuem uma jornada maior de horas trabalhadas por mês.

Já as variáveis Níveis de satisfação, Número de projetos e Tempo trabalhado diminuem a chance de saída da empresa. Nota-se também que colaboradores com ocorrência de acidente de trabalho e promoção nos últimos 5 anos são menos propensos a sair da empresa. Aqueles colaboradores que trabalham em laboratório ou escritório têm menos chance de saída da empresa se comparados aos que trabalham em campo.

Quando tempo trabalhado na empresa é baixo, a impacto de notas altas consiste na diminuição da chance de saída do colaborador. Quando o tempo de casa é maior, notas altas aumentam a chance de saída do colaborador.

Quando a nota da última avaliação é baixa, o aumento do tempo trabalhado consiste na diminuição da chance de saída do colaborador. No entanto, quando a nota da última avaliação é máxima, o aumento do tempo trabalhado consiste no aumento da chance de saída.

Assim, o departamento de Recursos Humanos pode atentar-se aos colaboradores com muito tempo de empresa e com altas notas na pesquisa de satisfação, visto que, pelo modelo escolhido, entre trabalhadores com muito tempo de empresa, o aumento da nota da última avaliação pode aumentar em até 13 vezes a chance de o colaborador sair da empresa.

Neste contexto, algumas ações podem ser tomadas, tal como o diagnóstico mais qualitativo para entender, a partir da opinião dos colaboradores, o que os fazem deixar a empresa mesmo

satisfeita ou com altas notas na última avaliação. Assim, é possível rever o programa de pagamento de bônus, participação em lucros etc.

Em linhas gerais, o modelo mostrou que as variáveis utilizadas podem prever de maneira satisfatória a saída de um colaborador da empresa. No entanto, mais informações podem ser coletadas com objetivo de deixar o modelo cada vez mais robusto (melhor AUC).

É de suma importância o entendimento dos fatores quantitativos e qualitativos para que o modelo ajude a prever e as ações mantenham os índices de turnover cada vez menores.

Mesmo com um modelo de boa capacidade preditiva, existem fatores não mapeados que podem ser de suma importância na decisão de um colaborador de deixar ou uma empresa e limitam o modelo. Fatores psicológicos podem ter uma influência que o modelo não consegue prever.

REFERÊNCIAS

BERTOTTI, Andressa. **Identificação das causas da rotatividade de pessoal: um estudo de caso**. Caxias do Sul: Universidade de Caxias do Sul, 2013.

BEWICK, Viv; CHEEK, Liz; BALL, Jonathan. **Statistics review 13: Receiver operating characteristic curves.** BioMed Central Ltd, 2004. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065080/pdf/cc3000.pdf. Acesso em: 02 de novembro de 2022.

GONZALEZ, Leandro de A., **Regressão Logística e suas Aplicações**. São Luís, MA, 2018.

KLEINBAUM, David G.; KLEIN, Mitchel. **Logistic Regression: A Self-Learning Text**. 3 Ed. Atlanta, GA. Springer, 2010.

MARRAS, J. P. **Administração de Recursos Humanos: Do operacional ao estratégico**. São Paulo, SP, 2000.

OLIVEIRA, Áurea de F.; et al. **Análise dos Fatores Organizacionais Determinantes da Intenção de Rotatividade**. 2018 Disponível em: https://www.scielo.br/j/tpsy/a/4PrYpVxh7TbNnVJ5wv8rZBw/?lang=pt. Acesso em: 02 de novembro de 2022.

PACOTE "dplyr". **A Grammar of Data Manipulation.** Versão 1.0.10. Disponível em: https://dplyr.tidyverse.org

PACOTE "pROC". **Display and Analyze ROC Curves.** Versão 1.18.0. Disponível em: https://cran.r-project.org/web/packages/pROC/index.html

PATIAS, Tiago Z.; et al. **Custos de Rotatividade de Pessoal: Evidências no setor de supermercados**. UNOESC, 2015. Disponível em: http://editora.unoesc.edu.br/index.php/race

SAIANI, Kristin L., **Understanding Odds Ratio**. American Academy of Physical Medicine and Rehabilitation, Vol. 3, 263-267, 2011

SOUZA, Édila Cristina de. **Análise de influência local no modelo de regressão Logística**. Piracicaba, 2006. 101p.

STOLTZFUS, Jill C. (2011). **Logistic Regression: A Brief Primer**. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/j.1553-2712.2011.01185.x. Acesso em: 12 de outubro de 2022.