Modelo de previsão de desligamento de funcionários por meio da regressão logística

Gleyce Mariana Costa dos Santos[1]*; Daniele Aparecida Cicillini Pimenta[2]

[1] Rei do Pitaco. FP&A Senior - Data. Alameda Ministro Rocha Azevedo, 912 – Jardim Paulista; 01410-002 São Paulo, São Paulo, Brasil
[2] PECEGE, USP. Mestre em Engenharia. Rua Padre João Manoel da Silva, 413 – Nova América; 13417- 770 Piracicaba, São Paulo, Brasil
*autor correspondente: gleycemariana1@gmail.com

Modelo de previsão de desligamento de funcionários por meio da regressão logística

Resumo

Com o desenvolvimento digital e a crescente utilização do “Machine Learning”, é observada a incorporação da ciência de dados como ferramenta essencial para a tomada de decisões corporativas de maneira mais assertiva. Essa tendência vem se expandindo para diversos ramos dentro das empresas. Um desses ramos é a área de Recursos Humanos, a qual dentre os seus maiores desafios é a saída de funcionário, ou seja, o “turnover” (rotatividade). Diante disso este trabalho consiste na previsão do desligamento dos funcionários, através do modelo de regressão logística utilizando o software R, apresentando uma alternativa para compreender o fenômeno. O modelo construído mostrou-se eficiente, obtendo uma acurácia de 79% e um excelente poder discriminante.
Palavras-chave: Regressão Logística; Machine Learning; Recursos Humanos; Turnover; Desligamento.

Employee termination prediction model through logistic regression

Abstract

With digital development and the growing use of Machine Learning, the incorporation of data science is observed as an essential tool for making corporate decisions in a more assertive way. This trend has been expanded to several branches within companies. One of these branches is the Human Resources area, which among its biggest challenges is the departure of employees, known as turnover. Therefore, this work consists of predicting the termination of employees, through the logistic regression model using the R software, presenting an alternative to understand the phenomenon. The built model proved to be efficient, obtaining an accuracy of 79% and an excellent discriminating power.
Keywords: Logistic Regression; Machine Learning; Human Resources; Turnover; Employee Termination.

Introdução

O mercado de trabalho é o setor com maior volatilidade no mundo atual. Em uma era de constantes mudanças é imprescindível que as empresas estejam atualizadas para que elas continuem no mercado competitivo (Orsso, 2017). Segundo Chiavenato (1999), para competir no mercado não é somente necessário o material físico, também é preciso deter o talento humano.

A oportunidade crescente de acesso à universidade e à qualificação profissional faz com que os trabalhadores obtenham novas perspectivas de mercado e uma colocação em ambientes agradáveis, ou seja, que não representem riscos a sua saúde e que ofereçam a infraestrutura necessária para desempenhar suas funções com qualidade de vida, como áreas administrativas sucedidas (Zanella et al., 2015).

Em face dessa situação as empresas privadas sofrem frequentemente com o alto índice de rotatividade dos funcionários, uma vez que eles encontram novas chances muito atrativas em um mercado voraz por profissionais. Segundo, Chiavenato (2002) a palavra

“Rotatividade” é expressa por uma relação matemática sendo o percentual entre as admissões e os desligamentos com relação ao número médio de participantes da organização, no decorrer de certo período de tempo. Quase sempre, o “Turnover” ou Rotatividade é expresso em índices mensais ou anuais para permitir comparações, desenvolver diagnósticos, promover providências, ou ainda com caráter preditivo.

O setor de recursos humanos de uma empresa é responsável pela contratação e desligamento dos funcionários, em vista desse cenário esse departamento necessita de diversas técnicas que busquem os melhores empregados e os mantenham satisfeitos em seu local de trabalho, evitando as evasões e desligamentos, uma delas é a regressão logística.

A regressão logística é uma técnica de mineração de dados. De acordo com Gonzalez (2018) tal ferramenta é usada na tomada de decisões baseada em dados computacionais, dos quais pretende-se extrair informações relevantes de uma grande base de dados, de forma a buscar vantagens competitivas ou elaborações estratégicas.

O trabalho tem como objetivo investigar a saída dos melhores e mais experientes funcionários de uma determinada empresa, identificando características que possam prever esta classificação com um determinado nível de confiança utilizando a regressão logística.

Material e Métodos

A Pesquisa

Trata-se de uma pesquisa descritiva, pois serão expostas características da população em estudo. Explicativa, devido visar esclarecer quais fatores contribuem de alguma forma para o desligamento dos funcionários.

Dados

Os dados utilizados neste trabalho é um conjunto de dados simulado da comunidade online “Kaggle”. A base “Human Resource” possui 14.999 indivíduos e 10 variáveis. A Tabela 1 apresenta as variáveis utilizadas neste trabalho, bem como as suas descrições e valores.

Tabela 1. Variáveis da base de dados utilizada

Variável	Descrição	Valores
nivel_satisfacao	Nível de satisfação	Média: 0,6
ultima_avaliacao	Última avaliação	Média: 0,7
numero_projeto	Número de projetos	Média: 3,8
horas_mensais_medias	Média de horas mensais	Média: 201,1
tempo_empresa	Tempo de empresa	Média: 3,5
acidente_trabalho	Se tiveram acidente de trabalho	1. 0 2. 1
desligado	Se o funcionário saiu	1. 0 2. 1
promocao_ultimos_5_anos	Se teve uma promoção nos últimos 5 anos	1. 0 2. 1
area	Departamentos	1. contabilidade 2. gestão 3. marketing 4. produto_mng 5. R e D 6. rh 7. suporte 8. técnico 9. TI 10. vendas
salario	Salário	1. baixo 2. medio 3. alto

Fonte: Dados originais da pesquisa

Análise dos dados

Análise Exploratória de Dados

A Análise Exploratória de Dados conhecida como análise descritiva ou estatística descritiva, compreende ferramentas utilizadas com a finalidade de organizar, resumir e descrever características de conjuntos de dados para obter conclusões sobre as varáveis que se deseja estudar, utiliza-se gráficos e tabelas (Magalhães et al., 2004).

A Estatística Descritiva é considerada como a etapa inicial de uma pesquisa, que tem como objetivo observar e descrever fenômenos da mesma natureza, coletando, organizando e classificando dados numéricos, apresentando gráficos e tabelas dos dados observáveis e realizando cálculos de coeficientes (Bussab et al., 2010).

Modelo de Regressão Logística

De acordo com Kutner et al. (2004), a análise de regressão é um conjunto de métodos estatísticos com o objetivo de interpretar a relação funcional entre variáveis que apresente

relação de causa e efeito, de maneira que seja possível a estimação ou previsão de uma variável resposta por meio de uma ou mais variáveis preditoras.

Existem muitos modelos de regressão, dentre eles o mais conhecido e utilizado é o modelo de regressão linear simples que apresenta apenas uma variável preditora. Este tipo de modelagem é utilizado para estimar ou prever valores de um conjunto de variável ($\boldsymbol{Y_i}$) a partir de outro conjunto de variáveis ($\boldsymbol{X_i}$), ambas quantitativas, que apresente boa aproximação, de modo que entre estes conjuntos de variáveis possa estabelecer-se um relacionamento funcional entre eles, que pode ser descrito como:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ I = 1, 2, \ldots, n} \qquad (1)$$

em que Y_i é a variável resposta e X_i é a variável preditora para a i-ésima observação, $\boldsymbol{\varepsilon_i}$ é o erro associado ao modelo ajustado, que tem como suposição o comportamento que se adere a distribuição normal com $E(\varepsilon) = 0$ e $Var(\varepsilon) = \sigma^2$.

Contudo em muitas situações práticas de um estudo estatístico, pode-se obter como resposta ao evento em estudo, uma variável categórica que assuma dois ou mais níveis de resposta, onde estes podem apresentar-se com comportamento binário (que assume valores 0 ou 1), ordinal, nominal entre outros.

Segundo Kutner et al. (2004), a regressão logística binária é indicada para estudos em que a aplicação de um modelo de regressão seja permitido, porém, a variável resposta de interesse apresente apenas duas possíveis respostas não métricas (qualitativa), por exemplo, sucesso ou fracasso de um evento, onde este pode ser representado de forma binária assumindo valores 0 (zero) ausência ou 1 (um) presença da característica em estudo.

Segundo Kutner et al. (2004) o modelo de regressão logística binária simples admite somente dois valores de forma codificada, ou seja, para os níveis de resposta da variável do tipo, sim ou não, certo ou errado, presença ou ausência, sucesso ou fracasso, entre outros. Sua distribuição de probabilidade é representada por um modelo Bernoulli, com parâmetro $\boldsymbol{\pi_i}$. E considerando apenas uma variável preditora $\boldsymbol{X_i}$, tem-se o modelo de regressão logística binária simples, sendo descrito por

$$\boldsymbol{E\,(Y_i \mid X_i) = \pi_i(X_i) = \frac{\exp\,(\beta_0+\beta_1 X_i)}{1+\exp\,(\beta_0+\beta_1 X_i)}} \qquad (2)$$

onde $\boldsymbol{\beta_0}$ é o ponto onde a curva corta o eixo dos Y ($\boldsymbol{X_i}$ = 0); $\boldsymbol{\beta_1}$ é o coficiente angular e $\boldsymbol{X_i}$ é a variável independente, com i = 1,2,...,n.

Para a construção do modelo foi utilizado algoritmos de “machine learning”

(Aprendizado de máquina) com o uso de uma série de pacotes que fazem parte do pacote "tidymodels" do "Software" *R*. O algoritmo aplicado foi o Regressão logística penalizada que é do pacote "glmnet*"* que se ajusta a modelos lineares generalizados e similares por meio da máxima verossimilhança penalizada. A Tabela 2 apresenta algumas informações sobre o algoritmo Regressão logística penalizada.

Tabela 2. Características do algoritmo de Regressão logística penalizada

Algoritmo (Pacote)	**Hiperparâmetros otimizados**	**Principais características**	**Ranking de importância dos preditores**
Regressão logística penalizada (*glmnet*)	α (alpha): porcentagem correspondente ao tipo de regularização a ser aplicada: 0: ridge; 1: lasso; > 0 e < 1: elastic net. λ (lambda): parâmetro de regularização. Pode assumir valores no intervalo [0, ∞]. Quanto maior o valor, maior a penalização das estimativas dos parâmetros da regressão.	Tem como objetivo obter estimadores viesados, porém com variância reduzida, para os parâmetros de um modelo de regressão usual, o que resulta em um modelo preditivo menos complexo e, portanto, menos sujeito ao sobreajuste em novas observações.	Valor absoluto dos coeficientes correspondentes ao modelo que foi selecionado na etapa de otimização dos hiperparâmetros.

Fonte: Adaptado de Santos et al. (2019, p.5)

Entre os vários algoritmos disponíveis, alguns são pouco flexíveis (menos complexo), porém são interpretáveis, como é o caso da regressão logística (James et al., 2014). O ajuste de modelos preditivos, pode ser descrito pelas seguintes etapas: divisão do conjunto de dados, pré-processamento, técnicas de reamostragem e avaliação de performance do modelo.

- **Divisão do conjunto de dados**

Foi realizado a divisão da amostra em dados de treino e de teste para verificar se o modelo apresenta boa performance nos dados de treino e também nos dados de teste. Na divisão da amostra as proporções mais utilizadas são 60:40, 70:30 ou 80:20, quanto maior o número de observações, maior será a proporção do conjunto inicial utilizado para treinamento (Raschka, 2015).

Neste estudo foram utilizados 75% dos dados para treinamento do algoritmo (n = 11.249) e 25% para teste da performance preditiva dos modelos ajustados (n = 3.750). O pacote "rsample" foi usado para a divisão da base em treino e teste.

- **Pré-processamento**

O pré-processamento dos dados é conduzido pelos algoritmos que serão utilizados para o ajuste do modelo preditivo. A sua aplicação está relacionada a algumas atividades como transformação de variáveis quantitativas, redução de dimensionalidade do conjunto de dados, exclusão de variáveis com dados faltantes ou utilização de técnicas de imputação e organização de variáveis qualitativas (Raschka, 2015).

Neste trabalho foi aplicado o pacote “recipes” para o pré-processamento dos dados, criação de dummies e remoção de variáveis com variância zero.

- **Técnicas de reamostragem**

Entre as técnicas de reamostragem a mais utilizada em problemas de “machine learnig” é a validação cruzada “k-fold”, esta técnica corresponde na divisão aleatória do banco de treinamento em *k* partes de tamanhos iguais, em que *k-1* irá compor os dados de treinamento para o ajuste de modelos e a outra parte ficará designada para a estimativa de sua performance. O processo segue até que todas as partes tenham participado do treinamento e da validação do modelo, resultando em *k* estimativas de performance. As repeticões desse processo podem ser aplicadas para aumentar a precisão dessas estimativas (Kuhn et al., 2013).

Não há uma regra para a escolha de *k*, porém é mais comum a divisão dos dados em 5 e 10 partes. À medida que *k* aumenta, a diferença de tamanho do conjunto de treinamento original e dos subconjuntos reamostrados se torna menor, e à medida que essa diferença se torna menor, o viés da validação cruzada se torna maior. Contudo, o tempo necessário para obter o resultado final da validação cruzada se torna maior (Kuhn et al., 2013).

Para o estudo utilizou-se validação cruzada com *k = 10*, repetida 10 vezes e o pacote “rsample” foi usado para realizar reamostragens dos dados e a criação de validação cruzada.

- **Avaliação de Performance**

Depois de definido o valor de k, é necessário definir uma medida para estimar a performance dos modelos ajustados. Estas medidas são importantes tanto na etapa de seleção quanto na avaliação dos modelos preditivos. Seu cálculo tem como objetivo mensurar o quanto o valor predito para uma observação se aproxima do seu valor observado. (Meurer et al., 2017).

É possível obter uma matriz de confusão, que é a tabulação cruzada de classes observadas e preditas para os dados de teste. A sensibilidade é a proporção de verdadeiros positivos (VP) e a especificidade é a proporção de verdadeiros negativos (VN). (Meurer et al., 2017).

Uma ferramenta adequada para avaliar a sensibilidade e a especificidade é a curva ROC (“Receiver Operating Characteristic”), um desempenho geral de um classificador é

avaliado pela área abaixo da curva ROC, quanto mais próxima de 1, melhor é a performance do modelo.

Neste trabalho foi aplicado o pacote “yardstick” para a criação de medidas de performance e desempenho dos modelos, os pacotes “tune” e “dials” para encontrar hiperparâmetros ideais dos modelos e o pacote “workflow” que une todas estas funções.

Resultados e Discussão

Análise Exploratória de Dados

A Figura 1 apresenta o cruzamento das variáveis “ultima_avaliacao”, “nivel_satisfacao”, “salario” que está classificado como baixo, médio e alto, “numero_projeto” que vai de 2 a 7 e “status”. A variável “status” foi criada com base na variável “desligado”. Observa-se que os funcionários desligados estão mais aglomerados nos salários baixo e médio. A menor e a maior quantidade de projetos apresenta um baixo nível de satisfação entre os desligados.

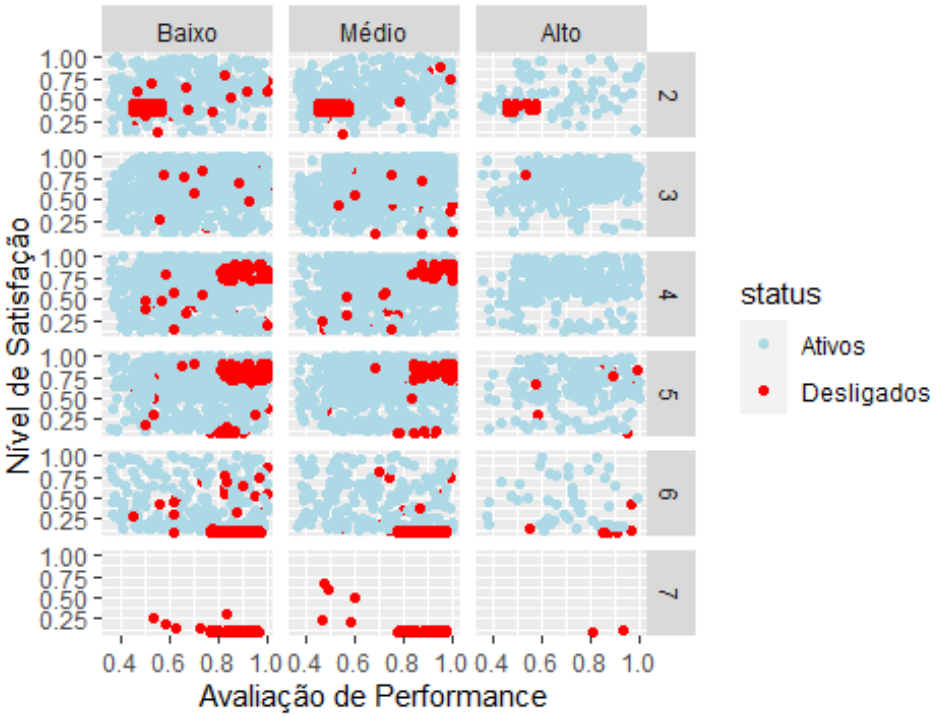

Figura 1. Avaliação de performance por nível de satisfação, salário, projetos e status
Fonte: Resultados originais da pesquisa

A Figura 2 apresenta o cruzamento das variáveis “salario”, “grupos” e “status”. A variável “grupos” foi criada com base nas varíaveis “nivel_satisfacao” e “ultima_avaliacao”.

Nota-se que o percentual de salário baixo, insatisfeito e produtivo é maior nos funcionários desligados. Há uma predominância de funcionários ativos com salário médio e que estão satisfeito e improdutivo.

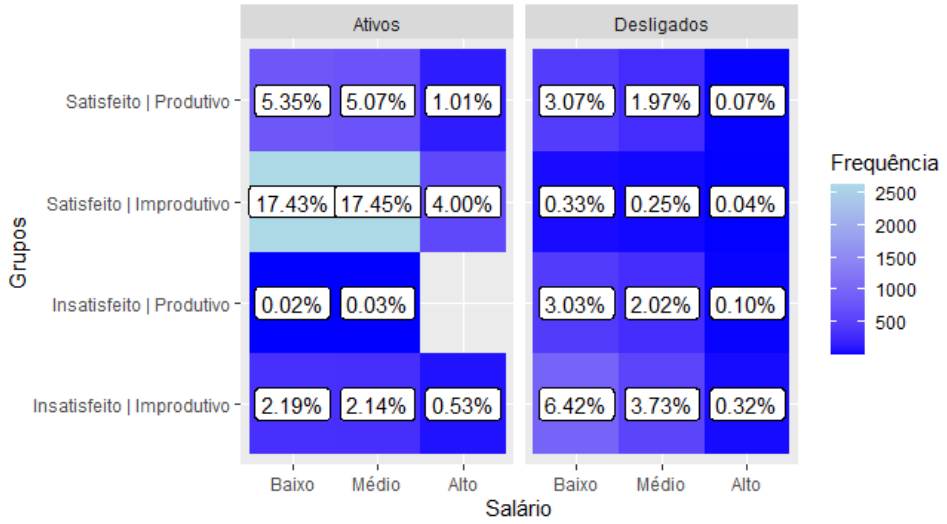

Figura 2. Distribuição do salário por grupos e status
Fonte: Resultados originais da pesquisa

Modelo de Regressão Logística

Segundo Freund et al. (1998), o efeito da multicolinearidade em um modelo não afeta seu poder preditivo nem sua adequação aos dados; porém, diminui a efetividade de serem previstos os efeitos das variáveis independentes sobre a variável dependente. Para verificação da existência ou não de multicolinearidade entre as variáveis analisadas foi calculado a matriz de correlação de Pearson. Na Tabela 3 estão presentes os resultados dessa correlação, nota-se que todas as correlações são fracas.

Tabela 3. Correlações de Pearson entre as variáveis do estudo.

Variáveis	ultima_ avaliacao	numero_ projeto	horas_mensais_ medias	tempo_ empresa
horas_mensais_medias				0,13
numero_projeto			0,42	0,2
ultima_avaliacao		0,35	0,34	0,13
nivel_satisfacao	0,11	-0,14	-0,02	-0,01

Fonte: Resultados originais da pesquisa

A Figura 3 mostra as variáveis que entraram no modelo bem como a importância de cada variável. É possível verificar que os valores das variáveis “ultima_avaliacao” e “tempo_empresa” são positivos. Logo, influenciam positivamente no modelo, ou seja, quanto maior o valor destes índices, maior será a probabilidade do indivíduo ser desligado. Em contraposição, as variáveis “nível_satisfacao” apresenta valor negativo, o que indica que quanto menor o valor deste índice, maior é a probabilidade de um funcionário ser desligado.

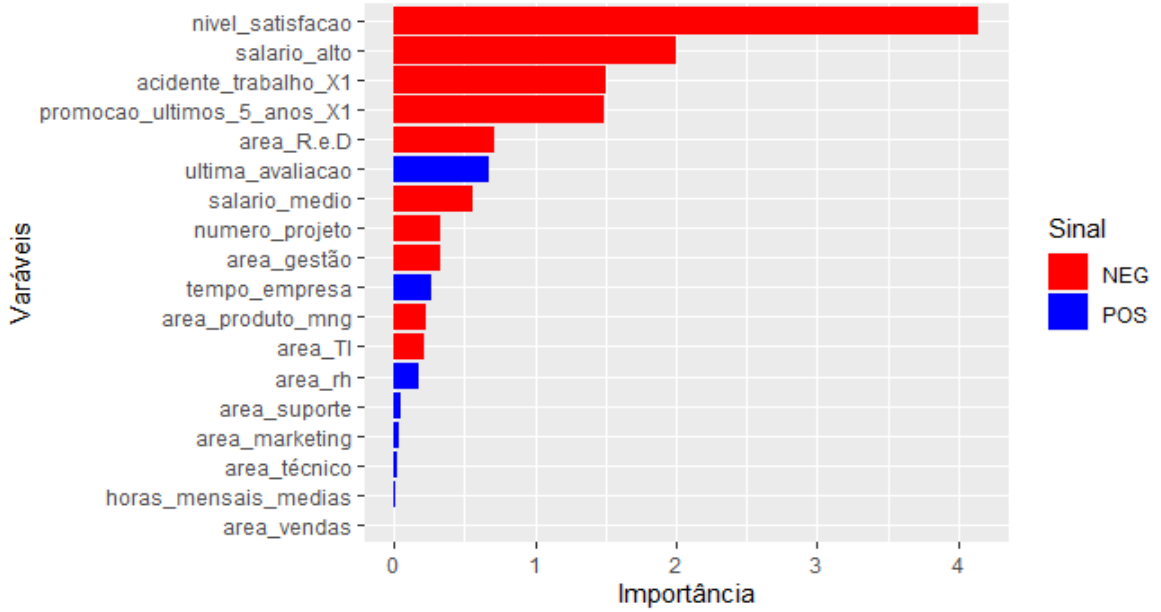

Figura 3. Importância das variáveis no modelo final
Fonte: Resultados originais da pesquisa

Verifica-se na Tabela 4 a classificação dos casos previstos pelo modelo para a variável "desligado", 79% dos funcionários são classificados de forma correta. O percentual de classificação correta para funcionários desligados é de 34% e para funcionários não desligados é de 93%.

Tabela 4. Tabela de confusão

Valores observados		Valores Previstos 0-Não	Valores Previstos 1-Sim	% Correto
Desligado	0-Não	2653	204	93
	1-Sim	592	301	34
Percentual de acerto				79

Fonte: Resultados originais da pesquisa

De cordo com Hosmer et al. (2000, p.160) a curva ROC mostra a probabilidade de detecção dos verdadeiros sinais (sensibilidade) e falso sinal (especificidade) a partir de um intervalo de possíveis pontos de corte. A partir da curva é possível avaliar a qualidade de discriminação do modelo. A Descrição dos resultados apresentados sob a curva ROC está apresentada no Tabela 5.

Tabela 5. Descriçao dos valores da curva ROC

Ponto de Corte	Descrição
ROC = 0,5	Sem poder discriminante
0,7 ≤ ROC < 0,8	Aceitável poder discriminante
0,8 ≤ ROC < 0,9	Excelente poder discriminante
ROC ≥ 0,9	Excepcional poder discriminante

Fonte: Adaptado de Hosmer et al. (2000, p.162)

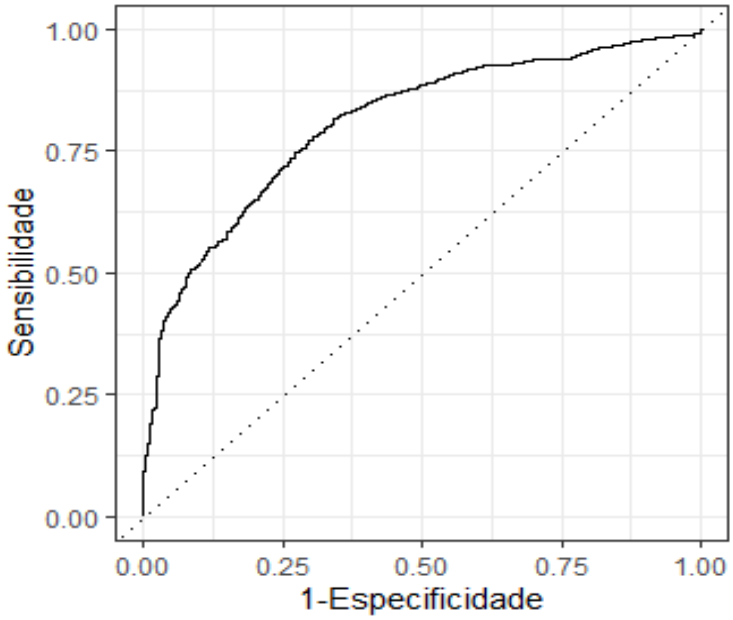

Figura 4. Curava ROC para modelo de regressão logística binário
Fonte: Resultados originais da pesquisa

Observa-se na Figura 4 que a área abaixo da curva corresponde ao valor de 81%, demonstrando um excelente poder de discriminação de acordo com a classificação dada por Hosmer et al. (2000).

Considerações Finais

Neste trabalho com o uso da regressão logística foi possível evidenciar as principais variáveis que impactam no desligamento de funcionários da empresa, oriunda do conjunto de dados simulado da comunidade online “Kaggle”. As variáveis “ultima_avaliacao”, “tempo_empresa”, “área_rh”, “area_suporte”, “area_marketing”, “area_técnico”, “horas_mensais_medias” e “área_vendas” possuem impacto positivo para o “turnover”, enquanto “nivel_satisfação”, “salario_alto”, “acidente_trabalho_X1”, “promocao_ultimos_5-_anos_X1”, “área_R.e.D”, “salario_medio”, “numero_projeto”, “area_gestao”, “area_produto_mng” e “área_TI” afetam esse indicador de forma negativa.

Os resultados obtidos com o modelo de regressão logística permitem sugerir à empresa, visando melhorar o “turnover”, que direcione seus esforços para o reconhecimento salarial daqueles que cumprem suas metas, pois existem funcionários produtivos ganhando menos que funcionários considerados improdutivos. Melhorar a qualidade na contratação a fim de minimizar a insatisfação e improdutividade dos funcionários. Avaliar as áreas de pessoas e suporte que apresentam grande chance de desligamento. Realizar o controle na

quantidade de projetos que os funcionários atuam, pois bons profissionais estão sendo sobrecarregado.

Referências

Bussab, W. O.; Morettin, P. A. 2010. Estatística básica, v. 4.

Chiavenato, I. 2002. Administração de Recursos Humanos. Fundamentos Básicos. São Paulo: Editora: Atlas S.A.

Freund, R. J.; Wilson, W. J. 1998. Regression Analysis – Statistical Modeling of a Response Variable. San Diego: Academic Press.

Hosmer Jr, D. W.; Lemeshow, S. 2000. Applied logistic regression. New York: John Wiley & Sons 2: 260-280.

Gonzalez, L. D. A. 2018. Regressão logística e suas aplicações.

James G.; Witten D.; Hastie T.; Tibshirani R. 2014. An introduction to statistical learning: with application in R. New York: Springer.

Kaggle. Disponível em: <https://www.kaggle.com/>. Acesso em: 17 de janeiro de 2022.

Kuhn, M.; Johnson, K. 2013. Applied predictive modeling. New York: Springer 26: 13.

Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. Applied linear statistical models. Boston, Mass.: McGraw-Hill 5: 1398.

Magalhães, M. N.; De Lima, A. C. P. 2004. Noções de probabilidade e estatística. São Paulo: Edusp, 6.

Meurer, W. J.; Tolles, J. 2017. Logistic regression diagnostics: understanding how well a model predicts outcomes. Jama, 317: 1068-1069.

Orsso, G. C. 2017. Rotatividade de funcionários em escritórios contábeis.

Raschka, S. 2015. Python machine learning. Packt publishing ltd.

Santos, H. G. D.; Nascimento, C. F. D.; Izbicki, R.; Duarte, Y. A. D. O.; Porto Chiavegatto, A. D. 2019. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. Cadernos de Saúde Pública, 35.

Zanella, T.; Araldi, J.; de Almeida Silva, T. 2015. Influência da Rotatividade de Funcionários da Construção Civil e outras Variáveis no Custo Final de uma Obra. In: XV Mostra de Iniciação Científica, Pós-graduação, Pesquisa e Extensão.