Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

People Analytics: Previsão de desligamentos por meio das técnicas de regressão logística e análise de sobrevivência

Fernanda Buffon Bianchi

19 de abril de 2023

Fernanda Buffon Bianchi

People Analytics: Previsão de desligamentos por meio das técnicas de regressão logística e análise de sobrevivência

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Profa. Dr. Lisiane Priscila Roldão Selau

Porto Alegre
Abril de 2023

Fernanda Buffon Bianchi

People Analytics: Previsão de desligamentos por meio das técnicas de regressão logística e análise de sobrevivência

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):_____
Profa. Dr. Lisiane Priscila Roldão Selau, UFRGS
Doutor(a) pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Silvana Schneider, UFRGS
Doutor pela Universidade Federal de Minas Gerais – Belo Horizonte, MG

Porto Alegre
Abril de 2023

"It takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!". (Alice Through the Looking-Glass)

Agradecimentos

Gostaria de agradecer, antes de tudo, à minha família, por estar sempre ao meu lado, mesmo quilômetros de distância. Não existem palavras bonitas o suficiente que expressem o amor que eu sinto por todos vocês. Obrigada.

À minha mãe, por todas às vezes em que me senti frustrada e era a voz dela dizendo o quanto acreditava em mim que me acalmava; por ser minha referência em força e generosidade, e se sacrificar tanto para me ver feliz. Ao meu pai, que me fazia sorrir todos os dias, sem falta, com seu "Bom dia, filha" pontual às 08h da manhã, me lembrando do quão sortuda sou pelo pai que tenho. À minha irmã, por me entender, me aguentar, me apoiar, reclamar e rir comigo como mais ninguém faz. Obrigada por tudo, e por ser a minha certeza no mundo de que eu nunca vou estar sozinha. Pra sempre, nós duas.

Às minhas melhores amigas, obrigada pelo apoio nas horas boas e ruins. Jamais imaginei que teria a chance de ter pessoas tão incríveis caminhando junto comigo pela vida por tanto tempo. Obrigada pelas risadas, pela escuta, pelas histórias e por darem significado a palavra amizade.

Às minhas amigas Andressa e Raquel, que desde o início foram muito mais que colegas de faculdade. Nunca me senti só em todos esses anos de curso, porque sempre tive vocês do meu lado, tornando tudo mais leve. Obrigada por estarem comigo desde a primeira semana da graduação, mas especialmente por seguirem ao fim dela.

À minha equipe e colegas de trabalho, que me ensinaram tanto ao longo do tempo e foram fomento para o desenvolvimento deste trabalho. Obrigada por todos aprendizados, compreensão e apoio.

À minha orientadora, Lisiane, por ter aceitado meu convite para orientação e pela ajuda durante esse processo. Obrigada por me apoiar, vibrar com minhas conquistas e, pelo mais importante e que fazia sem notar, sempre me passar a sensação de que tudo iria dar certo no final. Minha admiração por ti é enorme, mas minha gratidão é ainda maior.

A todos os professores do Instituto de Matemática e Estatística da UFRGS, por todo o aprendizado oferecido durante a graduação.

Por fim, agradeço a UFRGS, que foi por tanto tempo meu maior sonho até se tornar minha maior fonte de aprendizado e crescimento, tanto profissional quanto pessoal.

Resumo

O mercado de trabalho se tornou mais competitivo e globalizado nos últimos anos, acarretando altos indíces de rotatividade para as empresas. Com isso, surge uma necessidade cada vez maior de alocar ações de retenção de forma mais assertiva, identificando os colaboradores com maior risco de pedir desligamento e atuando em relação a eles, evitando assim impactos financeiros e de performance, e investindo tempo e dinheiro de forma mais estratégica. A técnica mais amplamente utilizada para predição de desligamentos voluntários é a Regressão Logística. Um modelo pouco conhecido na área, mas que é cada vez mais utilizado em problemas similares é a Análise de Sobrevivência. Sendo assim, este trabalho propõe tanto a utilização do modelo mais tradicional, quanto a introdução do modelo de sobrevivência. Visto que cada modelo prediz um aspecto diferente em relação ao desligamento do colaborador, probabilidade e tempo, respectivamente, o objetivo é cruzar as variáveis resposta dos dois modelos a fim de se desenvolver uma ferramenta de visualização de dados que forneça aos tomadores de decisão a possibilidade de alocar seus investimentos nos indivíduos com maior risco de desligamento, e que são, portanto, de alta prioridade em ações de retenção. Os modelos foram desenvolvidos com uma base de dados do período de Janeiro de 2021 e formada por 534 colaboradores. Os modelos foram analisados com base em três medidas de desempenho: percentual de acerto, teste KS e área abaixo da curva ROC, sendo a primeira medida apenas para o modelo logístico. Neste estudo, ambas as técnicas obtiveram desempenhos satisfatórios e demonstraram boa capacidade de predição. A ferramenta de visualização de dados desenvolvida propôs a divisão dos colaboradores em 3 níveis diferentes de priorização, cruzando os riscos de perda sob os aspectos probabilidade e tempo.

Palavras-Chave: People Analytics, Regressão Logística, Análise de Sobrevivência, Visualização de dados.

Abstract

The job market has become more competitive and globalized in recent years, leading to high turnover rates for companies. As a result, there is an increasing need to allocate retention actions more assertively, identifying employees with a higher risk of leaving and acting on them, thus avoiding financial and performance impacts and investing time and money more strategically. The most widely used technique for predicting voluntary turnover is Logistic Regression. A less well-known model in the field but increasingly used in similar problems is Survival Analysis. Therefore, this work proposes both the use of the traditional model and the introduction of the survival model. Since each model predicts a different aspect of employee turnover, probability and time, respectively, the goal is to cross the response variables of both models to develop a data visualization tool that provides decision-makers with the ability to allocate their investments in individuals with the highest risk of turnover, and who are therefore a high priority for retention actions. The models were developed using a database from January 2021 and consisting of 534 employees. The models were evaluated based on three performance measures: accuracy rate, KS test, and area under the ROC curve, with the first measure only for the logistic model. In this study, both techniques achieved satisfactory performance and demonstrated good predictive ability. The data visualization tool developed proposed dividing employees into three different prioritization levels, crossing the risks of loss under the probability and time aspects.

Keywords: People Analytics, Logistic Regression, Survival Analysis, Data visualization.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Contexto, tema e delimitação

A retenção de colaboradores é atualmente um dos maiores problemas enfrentados pelas empresas, em um mercado de trabalho cada vez mais competitivo e globalizado, especialmente em empresas da área de tecnologia, cujos profissionais são muito disputados e consequentemente as taxas de desligamento são bastante altas. O setor de Recursos Humanos (RH) é o responsável pelos desligamentos, bem como quaisquer outros processos que o colaborador venha passar em seu tempo na empresa. Segundo (2), a rotatividade de funcionários custa caro para as empresas, que sofrem financeiramente ao perder um colaborador pelos investimentos em tempo, dinheiro e capacitação daquele indivíduo. As saídas voluntárias podem ser definidas pela decisão de desligamento não ter partido da empresa, e, portanto, muitas vezes são saídas das quais a empresa tem interesse em evitar. Devido ao setor de RH envolver-se diretamente com as pessoas, muitas decisões são tomadas baseadas na experiência do decisor. Dado os altos custos financeiros envolvidos nos processos desta área, a necessidade de decisões baseadas em pensamento analítico e suportadas com dados fez desenvolver-se a área de *People Analytics*, que integra a cultura de dados ao ambiente de RH. Essa integração ocorre por meio de análises descritivas, visuais e estatísticas de informações relacionadas ao capital humano da organização e dos processos de RH, proporcionando decisões orientadas a dados (3). Além disso, o impacto do *People Analytics* na organização é maior quando voltado para o futuro (4), isto é, é mais útil quando é preditivo e fornece uma visão para o futuro em relação a prováveis resultados de negócio.

1.2 Problematização

Identificar quando um colaborador solicitará seu desligamento é uma ferramenta importante para a tomada de decisões estratégicas, investindo financeiramente de maneira mais objetiva e até mesmo evitando a saída de funcionários com performances diferenciadas dos demais. Dado que o setor de RH desconhece o real motivo que possivelmente levou ao desligamento do colaborador e que o conhecimento dessa motivação pode levar a empresa a evitar desligamentos indesejáveis através da previsão de desligamentos, se constrói uma problemática em entender se determinadas variáveis têm influência sobre o desligamento e construir um modelo de previsão de saídas. De maneira mais específica, entender se o perfil do indivíduo, ou seja,

variáveis demográficas (como idade, sexo) e variáveis laborais (como salário, cargo) associadas ao colaborador tem alguma influência sobre seu pedido de desligamento, e, se houver, identificar através delas os funcionários mais propensos a pedir demissão. Dessa forma, a área de RH teria um embasamento, por meio de análise de dados, como fomento nas tomadas de decisão no que tange desligamentos e modelos estatísticos para atuar de forma preventiva sobre as possíveis saídas.

A utilização de modelos de previsão classificatórios são os mais comumente usados em bancos de dados similares, em que há interesse na ocorrência de um determinado desfecho binário. Alguns exemplos de técnicas classificatórias são análise discriminante, redes neurais, regressão logística, entre outras (5). A regressão logística é uma técnica muito utilizada em previsões de desligamento ou *turnover* (movimentações de entrada e saída em empresas) modelando a probabilidade de saída voluntária de um funcionário

Outra abordagem relevante seria o uso de análise de sobrevivência, em que a variável de interesse seria o tempo até o evento ocorrer (6); neste caso, o tempo até o pedido de desligamento do colaborador. Há poucos registros desta técnica sendo aplicada em problemas desta área, mas há artigos que tratam de problemas similares ao usar da análise de sobrevivência para prever a inadimplência de clientes na área financeira (7). Esta técnica é vantajosa pois, ao ter como variável resposta o tempo até o evento ocorrer, possibilita que gestores priorizem desligamentos mais próximos e sejam mais assertivos ao agirem.

Porém, adiciona-se ao problema o fato de que não é viável agir em relação a todos possíveis desligamentos, visto os custos financeiros e o tempo que isso demandaria; ademais, nem todos os desligamentos impactam o suficiente a ponto que seja de interesse da empresa evitar esta saída. Avaliar dentre todas possíveis combinações de tempo e probabilidade de saída consome um tempo talvez fundamental a fim de evitar o desligamento de um colaborador. Assim, torna-se pertinente não só o desenvolvimento de métodos estatísticos para prever os pedidos de desligamento, mas uma ferramenta que auxilie na visualização destas previsões para grandes volumes de dados. Uma das possíveis ferramentas, e a de estudo neste projeto, seria um ranking de prioridade em relação à probabilidade de saída (regressão logística) e tempo até desligamento (análise de sobrevivência), cruzando as variáveis resposta e ordenando as possíveis saídas em uma matriz de prioridade, considerando um senso de urgência em relação ao desfecho de interesse, que é o pedido de desligamento do colaborador.

1.3 Questões de pesquisa

As questões de pesquisa são:

- Quais covariáveis relacionadas ao perfil do indivíduo mais impactam para a ocorrência do desfecho de interesse?
- As técnicas estatísticas de regressão logística e análise de sobrevivência podem auxiliar na previsão de pedidos de desligamento?
- É possível cruzar as previsões de ambas as técnicas estatísticas a fim de criar uma ferramenta de priorização nas tomadas de decisão em relação ao desfecho de interesse?

1.4 Objetivo Principal

Este trabalho visa utilizar as técnicas estatísticas de regressão logística e análise de sobrevivência, através da modelagem de determinadas covariáveis, para prever a probabilidade e o tempo até o desligamento de um determinado colaborador. De forma complementar, fornecer aos gestores um recurso para priorização de seus focos de ação dentre os possíveis pedidos de desligamento ao cruzar as variáveis respostas dos modelos estimados e desenvolver uma ferramenta de visualização, categorizando os possíveis desligamentos em uma matriz de priorização de ações.

1.5 Objetivos Específicos

Os objetivos específicos são:

- Aplicação dos modelos a um banco de dados reais;
- Implementação das técnicas estatísticas no *software* R;
- Categorização das observações de acordo com sua prioridade para implementação da ferramenta de visualização.

1.6 Fonte de dados

Este estudo utilizará a base de dados de colaboradores de uma empresa da área de educação e tecnologia. O banco de dados é constituído de variáveis demográficas e laborais. Os dados referem-se a colaboradores que estavam ativos na empresa no período de janeiro de 2021, sendo este o tempo inicial do experimento, e a extração dos dados foi feita em novembro de 2022, o tempo final do estudo. Entende-se que colaboradores que não pediram desligamento até aquele momento, isto é, o desfecho de interesse não foi observado, serão tratados como censura na modelagem de análise de sobrevivência.

2 Referencial teórico

Nesta seção serão apresentadas revisões e discussões feitas por outros autores acerca dos temas que serão abordados no trabalho.

2.1 People analytics

O setor de recursos humanos é o responsável por todos os processos relacionados ao capital humano de uma empresa: seleção e admissão, remuneração e benefícios, pesquisas de clima, rescisões, entre tantos outros. De acordo com (8), o capital humano destaca-se como sendo a dimensão mais valiosa para as organizações, e sob essa perspectiva, entende-se que é um dos ativos mais importantes para tornar uma empresa competitiva. Assim, houve crescente necessidade do RH mudar seu comportamento de área administrativa para parceiro estratégico da organização. Nesse contexto, surge o *People Analytics*, que consiste num conjunto de processos, facilitados por tecnologia, que tira partido de métodos descritivos, visuais e estatísticos para interpretar dados de pessoas e processos de RH (3). Esta cultura orientada à dados pode ser implementada nos mais diferentes processos do setor, como, por exemplo: no desenvolvimento dos perfis que melhor atendem a uma vaga na seleção de talentos, modelagens para descrever as práticas de remuneração da empresa, visualizações de métricas e indicadores para observar o desempenho da empresa em relação ao mercado, bem como modelagens preditivas para evitar a rescisão de funcionários.

Para (1), há 4 níveis de maturidade do *People Analytics* em uma empresa, conforme a Figura 2.1, e a grande maioria das empresas se encontra nos níveis 1 e 2. Porém, há uma grande dificuldade de avanço para os estágios 3 e 4, ligados a um nível maior de contribuição estratégica por parte do RH, através de modelos e análises preditivas. Angrave et al. (9) acreditam que uma das barreiras para o avanço de um RH mais analítico é que muitos profissionais de RH não entendem de análise de dados, enquanto os times de análise não entendem de RH. Logo, fica evidente a importância de profissionais especializados na área de *People Analytics* no contexto atual do setor do RH, na busca por transformar-se para além de uma área operacional, mas um colaborador na tomada de decisões estratégicas da empresa e dessa forma agregar ainda mais valor em sua contribuição na organização. Ademais, a proposta desse trabalho é auxiliar o avanço da organização para os níveis 3 e 4 de maturidade, com modelagens preditivas para que o foco de atuação do RH deixe de ser reativo.

Assim, um dos assuntos de maior interesse em prever no setor é a movimentação

de um funcionário. Para (2), a definição de rotatividade, também conhecida como *turnover*, é "o fluxo de entrada e saída de pessoas em uma organização, ou seja, as entradas para compensar as saídas das pessoas nas organizações". Essas saídas podem ser por iniciativa da empresa, as quais são chamadas de desligamentos involuntários, ou por parte do colaborador, os então desligamentos voluntários. Os desligamentos involuntários tendem a ser justificados por causas como baixa performance do colaborador ou ainda necessidade de redução de pessoal. Porém, os desligamentos voluntários acarretam na perda de colaboradores que poderiam ser de grande interesse da empresa em reter, seja por sua alta performance ou simplesmente pelo processo de subtituição em si, visto que ao contratar um novo colaborador, há oscilações no nível de produtividade, além dos altos custos nos processos de demissões e admissões (10). Análises de estudos de caso demonstram que a mediana do custo da rotatividade é de 21% do salário anual do empregado (11), portanto, a menos que a substituição do colaborador seja estratégica a longo prazo, o custo envolvido em sua saída pode trazer sérios impactos financeiros à empresa.

Figura 2.1: Níveis de maturidade do *People Analytics*, adaptado de (1)

2.2 Análise Estatística

2.2.1 Regressão Logística Múltipla

A Regressão Logística é adequada em situações em que a variável resposta é medida em uma escala binária, como presença ou ausência de uma característica, em que termos genéricos usados para as duas categorias são "sucessso"e "falha"(12). No contexto deste estudo, pode-se pensar em sucesso como o pedido de desligamento e falha a não ocorrência deste evento, ou seja, a permanência do colaborador.

Em problemas de regressão, o interesse chave é no valor esperado da variável resposta Y, dado os valores das variáveis independentes X, em que esse valor é

chamado de média condicional (13). Segundo estes mesmos autores, quando trata-se de variáveis dicotômicas, a média condicional precisa estar no intervalo $[0, 1]$, assim, informando a probabilidade de ocorrência do evento de interesse. Para isso, assume-se uma relação entre as variáveis independentes e dependentes que segue a distribuição sigmóide, relação esta que é não-linear e, portanto, impossibilita o uso de modelos lineares de regressão.

Dado um conjunto de variáveis independentes $\mathbf{X}$, a forma do modelo de regressão logístico múltiplo, segundo (13), é expressa pela equação

$$E(Y|x) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_j x_j}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_i x_i}} \text{ para } i = 1, ..., n \text{ e } j = 1, ..., p, \tag{2.1}$$

cuja função de ligação é

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij}, \tag{2.2}$$

A equação $log(\frac{\pi}{1-\pi})$ é chamada de função logito e é interpretada como o logaritmo de chances (12).

A comparação da ocorrência do evento com a não ocorrência do evento é, segundo (14),

$$\frac{\text{Prob (evento ocorrer)}}{\text{Prob (evento não ocorrer)}} = e^{B_0+B_1 x_1+...+B_j x_j}, \tag{2.3}$$

e também é conhecida como razão de chances, indicando quantas vezes o sucesso é mais provável de acontecer em relação ao fracasso. A mudança percentual na razão de desigualdades de um determinado coeficiente é dado pela sua exponenciação, de forma que coeficientes exponenciados menores que 1 refletem relações negativas entre o coeficiente e a ocorrência do evento, enquanto valores acima de 1 denotam relações positivas (14).

Para ajustar a regressão logística, é necessário estimar os valores de β, parâmetros desconhecidos. Para a regressão logística, o método de estimação utilizado é a máxima verossimilhança ao invés do método tradicional de mínimos quadrados, devido a natureza não-linear da transformação logística (14). Sua equação é dada por

$$l(\beta) = \sum_{i=1}^{n} \Big[y_i x_i^T \beta - \ln\Big(1 + \exp\Big(x_i^T \beta\Big)\Big)\Big]. \tag{2.4}$$

A etapa subsequente a estimação dos coeficientes refere-se ao teste de significância das variáveis através da estatística de Wald (13).

2.2.2 Análise de Sobrevivência

Em análise de sobrevivência a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse, denominado falha. Segundo (15), a principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Sem a presença de censuras, as técnicas estatísticas clássicas, como análise de regressão, poderiam ser usadas na análise deste tipo de dados. No entanto, quando há censura, tais técnicas não podem ser utilizadas pois elas necessitam de todos os tempos de falha. No contexto do atual trabalho, as censuras são os

colaboradores que permaneceram ativos no momento em que os dados foram extraídos, portanto não se sabe qual é seu tempo de falha (pedido de desligamento). Nesse sentido, trabalha-se com dados censurados à direita, isto é, o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Os dados de sobrevivência para o indivíduo i, $i = 1, ..., n$ são representados pelo par (t_i, δ_i), sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é um tempo de falha} \\ 0 & \text{se } t_i \text{ é um tempo censurado.} \end{cases}$$

Na presença de covariáveis medidas no i-ésimo indivíduo, os dados ficam representados por (t_i, δ_i, x_i).

Uma das principais funções no contexto de análise de sobrevivência é a função de sobrevivência, $S(t)$, definida como a probabilidade de um indivíduo sobreviver além de um tempo t, isto é, da falha não ocorrer até t, que é dada pela seguinte expressão

$$S(t) = P(T \geq t). \tag{2.5}$$

A função taxa de falha $\lambda(t)$, ou função risco, é bastante útil para descrever a distribuição do tempo de vida dos indivíduos, descrevendo a forma em que a taxa instantânea muda com o tempo (15) e é definida como a probabilidade de, dado que o indivíduo sobreviveu até o tempo t, ele falhe no próximo menor intervalo de tempo, divido pela amplitude deste intervalo.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.6}$$

Para os cálculos das funções acima, é necessário utilizar de um método de estimação, que neste caso será o Modelo de Regressão de Cox, que permite a análise de dados de sobrevivência ajustando por covariáveis (15). Considerando p covariáveis, de modo que x é um vetor com os componentes $x = (x_1, ..., x_p)$, a forma geral do modelo é dada por

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_j\right) = \lambda_0(t) \exp(\mathbf{x}\beta'). \tag{2.7}$$

Este modelo é composto pelo produto de dois componentes, um não-paramétrico e outro paramétrico. O componente não paramétrico, $\lambda_0(t)$, não é especificado e é usualmente chamado de função de base, pois $\lambda(t) = \lambda_0(t)$ quando $x = 0$ (15). O componente paramétrico é frequentemente usado na forma multiplicativa

$$g(x) = \exp\left(\sum_{j=1}^{p} \beta_j x_j\right) = \exp(\mathbf{x}\beta'), \tag{2.8}$$

em que β é o vetor de parâmetros associado às covariáveis.

A suposição básica do modelo de Cox é que a razão das taxas de falha de dois indivíduos são proporcionais, não importa o tempo que eles sobrevivam. Isto é, a razão das funções de taxa de falha para dois indivíduos diferentes i e j é

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp\{\mathbf{x}_i'\beta\}}{\lambda_0(t) \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i'\beta - \mathbf{x}_j'\beta\}. \tag{2.9}$$

A parte paramétrica do modelo é usualmente estimada pelo método da máxima verossimilhança parcial. No contexto deste trabalho, desligamentos são registrados na unidade discreta de tempo dias, e posteriormente transformado na unidade anos, de forma que podem ser observados mais de um desligamento em um ponto do tempo. Estes tempos de falha são chamados de "empates"e é necessário modificar a função de verossimilhança para a estimação do modelo com esta característica. Breslow veio a propor o método que seria o mais amplamente utilizado, levando em consideração o número de empates em um determinado tempo t e também a soma dos coeficientes de cada variável dos indivíduos que passaram pelo empate (16). Para encontrar os estimadores dos coeficientes β é necessária a utilização de métodos numéricos.

A parte não paramétrica, $\lambda(t)$ será estimada através do método de máxima verossimilhança. Considerando d_j os empates em um tempo t_j, $D(t_j)$ os indivíduos que tiveram empates no j-ésimo tempo e $R(t_j)$ o conjunto de indivíduos sob risco de ocorrência da falha no j-ésimo tempo, a equação é dada por

$$\hat{h}_0(t_j) = 1 - \sum_{l \in D(t_j)} \frac{\exp\left(\widehat{\boldsymbol{\beta}} x_l\right)}{1 - \hat{\xi} \exp\left(\widehat{\beta} x_l\right)} = 1 - \sum_{l \in R(t_j)} \exp\left(\widehat{\boldsymbol{\beta}} x_l\right) \tag{2.10}$$

Dessa forma, têm-se a estimativas dos componentes paramétricos e não paramétricos do modelo, sendo é possível estimar a taxa de falha pelo modelo de regressão de Cox (17).

2.3 Visualização de dados

Uma visualização de dados eficaz pode significar a diferença entre o sucesso e o fracasso na hora de comunicar constatações (18). Quando se trata de modelagem estatística, a variável resposta de um modelo pode ser de difícil compreensão para àqueles que não têm conhecimento estatístico. Porém, ao traduzir o número em dados visuais, como, por exemplo, gráficos, o resultado torna-se acessível a este público, por haver uma maior familiaridade com esta forma de comunicação de informações, especialmente no contexto do setor de RH (1). Um exemplo de visualização de dados no contexto de recursos humanos é a "Matriz 9-*box*", apresentada na Figura 2.2, que classifica o potencial de posições de liderança em organizações, com o eixo x referindo-se à performance e o eixo y ao potencial do gestor (19). Adaptando a Matriz 9-*box*, é possível cruzar as variáveis resposta dos modelos nos eixos e criar classificações de quadrantes, de forma que se indique nos quadrantes vermelhos, tal qual na Figura 2.2, quais os colaboradores com maior propensão a pedir desligamento e que a empresa deve priorizar ações de desenvolvimento e retenção. Conclui-se que, quando o volume de dados é muito grande, um resumo visual facilita a visualização bem como permite que pontos de maior atenção se destaquem, facilitando a comunicação da informação.

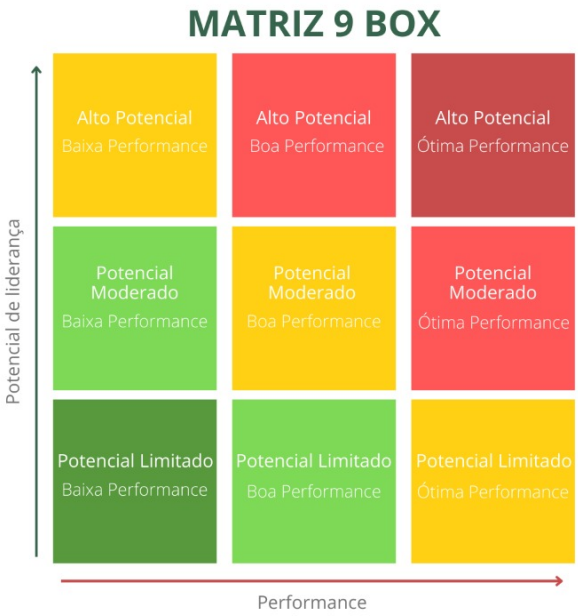

Figura 2.2: Representação da matriz 9-*box*

3 Metodologia de Pesquisa

O método para a construção dos modelos de predição de pedido de desligamento que será proposto está dividido em 6 etapas e foi baseado em (5). Dentro de cada etapa, estão no total 17 subitens, conforme descrito na Figura 3.1.

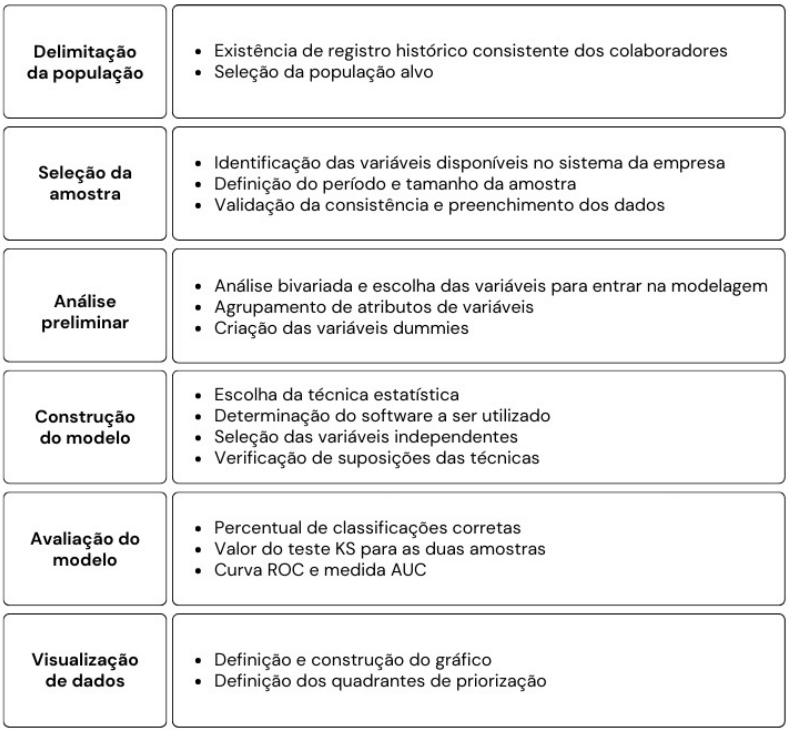

Figura 3.1: Etapas do método

3.1 Delimitação da população

Primeiro, é necessário um histórico consistente de registro de dados dos colaboradores da empresa. Os dados da amostra têm que constituir toda a informação disponível sobre o colaborador em seu cadastro admissional, suas informações enquanto capital da empresa (cargo que ocupa, setor do qual faz parte) e seu status

subsequente como ativo ou como desligado voluntário.

Também é preciso decidir para qual segmento da população o modelo vai ser construído, visto que nem toda população tem a mesma suscetibilidade a pedidos de rescisão, devido ao tipo de contrato laboral.

3.2 Seleção da amostra

É necessário o mapeamento das variáveis disponíveis através da avaliação das diferentes bases acessíveis no sistema. São elas as bases cadastrais do colaborador, contendo variáveis de caráter demográfico, e as bases de registro do funcionário enquanto membro da empresa, que contêm suas movimentações na estrutura da instituição entre outros registros provenientes de processos internos.

Para definição do período de extração da amostra, foi levado em consideração o tempo que se gostaria de predizer o desfecho, que seria de aproximadamente 2 anos a partir do momento observado.

De posse do banco de dados, é feita a etapa de análise exploratória do banco a fim de identificar observações inconsistentes ou faltantes, que podem vir a prejudicar o modelo. Se tais observações forem encontradas, é necessário fazer algum tipo de tratamento para os dados. A análise exploratória é feita através de uma análise descritiva completa do banco, com gráficos e tabelas para avaliar o comportamento das variáveis.

3.3 Análise preliminar

O primeiro passo é, aliado à análise exploratória feita anteriormente, o agrupamento de variáveis com baixa densidade de observações bem como transformação das variáveis contínuas do banco em variáveis categóricas, agrupadas em determinados intervalos.

Inicia-se a análise preliminar identificando, dentre as variáveis disponíveis, quais entrarão na análise. Isso será feito através da análise bivariada das variáveis pelo uso de tabelas de contingência, calculando o risco relativo (RR) associado aos diferentes atributos das variáveis. O cálculo do RR é dado pela equação abaixo

$$\text{Risco Relativo} = \frac{\text{Ocorrências no grupo/ Total de ocorrências}}{\text{Não-ocorrências no grupo/Total não-ocorrências}} \tag{3.1}$$

A interpretação é a mesma para ambos modelos - quanto mais o percentual de pedidos de desligamento/ocorrência do evento diferir do percentual de ativos/censuras para um atributo da variável, maior será a capacidade de discriminação entre os dois grupos deste atributo. Esse cálculo é feito para cada categoria das variáveis presentes no banco.

Neste trabalho, será utilizado as classificações de risco relativo em 7 classes, conforme proposto por (20) e apresentado na Figura 3.2. Além de identificar variáveis com boa capacidade preditora, o método auxilia no agrupamento e filtragem de variáveis. Agrupam-se atributos com mesma classificação do RR (quando fizer sentido, respeitando a natureza dos dados) por possuírem comportamento semelhante, e os

atributos classificados como neutros não são utilizados na análise, por não demonstrarem capacidade discriminatória significante entre os grupos.

Figura 3.2: Classes de risco relativo

Por fim, dada a seleção de quais variáveis e respectivos atributos que entrarão na análise, é feita a criação de variáveis *dummies* para cada um dos níveis analisados destas variáveis. A variável *dummy* dicotomiza a observação, assumindo apenas dois valores: 1 (presença do atributo) ou 0 (ausência do atributo).

3.4 Construção do modelo

Após a seleção das variáveis, agrupamento das mesmas e criação das variáveis *dummies*, inicia-se o processo de construção do modelo com o banco final que se têm em mãos. É feita a escolha das técnicas estatísticas utilizadas para modelagem, que neste trabalho serão a regressão logística múltipla e o modelo de regressão de Cox, e determinado o *software* em que será realizada a a construção do modelo.

Tendo escolhida a técnica estatística e o *software*, a etapa seguinte é a de aplicação de métodos de seleção automáticos para a escolha das variáveis preditoras que entrarão no modelo. Segundo (21), o método mais adequado para seleção é o *stepwise*, que é vantajoso em comparação a métodos como *forward* e *backward* por não avaliar apenas uma vez a entrada da variável no modelo e por desconsiderar variáveis com indícios de multicolinearidade.

Depois da seleção de variáveis resultar no modelo mais parcimonioso, é feita a verificação das suposições das técnicas. O pressuposto para aplicação da técnica de regressão logística é a ausência de multicolinearidade, isto é, que as variáveis do modelo não sejam altamente correlacionadas entre si. Para avaliar a multicolinearidade, utiliza-se da medida VIF, *Variance Inflaction Factor*, que mede a proporção da variância da variável analisada que pode ser explicada pelas demais variáveis do modelo(21), e valores acima de 10 indicam alta multicolinearidade. Para o modelo de regressão de Cox, a suposição básica é a de riscos proporcionais. Esta suposição pode ser verficada tanto através de testes estatísticos quanto de diagnósticos gráficos pela análise dos resíduos de Schoenfeld (22), em que o que se procura observar é um padrão aleatório em relação ao tempo. Outro pressuposto que deve ser verificado para o modelo de regressão de Cox é a linearidade das variáveis contínuas.

3.5 Avaliação do modelo

Nesta etapa são avaliadas as medidas de desempenho dos modelos para analisar seu poder de predição em relação as variáveis resposta de interesse.

A primeira medida de desempenho, utilizada para o modelo logístico, é o percentual de classificações corretas feitas pelo modelo. Essa medida é analisada em

conjunto com sua matriz de confusão, em que compara-se os valores preditos com os valores observados.

Outra medida de desempenho analisada, tanto para o modelo de regressão logística quanto para o modelo de regressão de Cox, é o valor do teste de Kolmogorov-Smirnov (KS) para as duas amostras. Este valor é obtido pelas funções de distribuição acumulada de dois diferentes grupos, e a medida é dada por um percentual que representa o quão bem é possível diferenciar estes dois grupos (23).

Por fim, é analisada a curva ROC, ferramenta utilizada na avaliação de modelos (24), que é um gráfico da sensibilidade (taxa de verdadeiros positivos) versus a especificidade (taxa de falsos positivos). Espera-se que, se o modelo tem poder de discriminação alto, a curva esteja no canto superior esquerdo da figura, enquanto modelos com poder mais baixo estejam próximos da linha diagonal do gráfico. A área sob a curva ROC (AUC) é uma medida da qualidade geral do modelo, em que valores maiores indicam um modelo com melhor desempenho.

3.6 Visualização de dados

Construído os modelos, é necessário definir-se a ferramenta descritiva de interesse, que pode ser uma medida, tabela ou gráfico, e o que se deseja observar por meio dela, e a consequente construção da ferramenta de visualização. Por fim, define-se categorias de priorização em relação as variáveis resposta dos modelos, ou seja, os pontos de corte destas variáveis resposta que indicam que o indivíduo é mais ou menos suscetível ao desfecho de interesse.

4 Resultados

Nesta seção são apresentados os resultados dos dois modelos desenvolvidos conforme cada abordagem utilizada.

4.1 Delimitação da população

Conforme descrito anteriormente, para desenvolvimento deste projeto, foi utilizado o conjunto de dados de uma empresa de médio porte brasileira, que atua no setor da educação e tecnologia. As informações disponibilizadas pela empresa para a criação do modelo são oriundas do cadastro do colaborador no sistema da instituição, feito no momento de admissão e atualizadas e validadas mensalmente, enquanto há vínculo entre colaborador e empresa.

A população-alvo são os colaboradores, com exceção dos cargos de aprendiz e estagiário, por serem níveis com término de contrato pré-definido.

Vale salientar que colaboradores que foram desligados por iniciativa da empresa, representando um 3º desfecho, não constam no banco, pois entende-se que se diferenciam dos demais perfis (ativos e desligados voluntários).

4.2 Seleção da amostra

A amostra é constituída pelos colaboradores ativos na empresa no mês de janeiro de 2021. O critério para o período escolhido foi a data mais antiga da qual se tinha um registro histórico consistente. Dessa forma, o desfecho foi observado no tempo de quase 2 anos decorridos, em novembro de 2022.

O banco é formado por 18 variáveis, sendo elas variáveis demográficas (sexo, idade, raça), variáveis laborais (tempo de empresa, salário, nível de atuação), bem como variáveis decorrentes de processos internos da empresa (reconhecimentos, avaliação de performance). Destas 18 variáveis, três delas são contínuas e as demais categóricas.

Após análise exploratória do conjunto de dados acerca da qualidade de preenchimento, dados faltantes e inconsistentes, o banco ficou formado por 534 observações. Não foi necessária a exclusão de nenhuma variável ou observação, visto que, como foi citado anteriormente, as informações são atualizadas mensalmente e portanto não havia inconsistências no histórico. A distribuição das observações pelo seu desfecho é indicada pela Tabela 4.1.

Tabela 4.1: Distribuição das observações pelo desfecho

Desfecho	Quantidade	Percentual
Ativo na empresa	365	68.4%
Pedido de desligamento	169	31.6%

A divisão da amostra entre treino e teste é usualmente recomendada como uma das formas de validação do modelo preditivo. A divisão neste banco não será realizada tanto por se tratar de uma amostra pequena, quanto pelas técnicas de regressão logística e análise de sobrevivência não serem tão suscetíveis ao sobreajuste como algoritmos de *machine learning*, em que há um risco muito maior de *overfitting* (25).

4.3 Análise preliminar

Foi feita a categorização das três variáveis contínuas do banco (tempo de empresa, idade e salário) em faixas de valores, bem como o agrupamento de atributos de variáveis categóricas como, por exemplo, mestrado e doutorado na variável escolaridade, entre outros. Esses agrupamentos foram feitos devido a baixa densidade de observações nas categorias, que prejudicava o posterior cálculo de RR. Os agrupamentos categóricos foram feitos respeitando as hierarquias dos dados ordinais.

Após o agrupamento inicial de atributos, foi feito o cálculo do risco relativo, dividindo o percentual de desligamentos voluntários pelo percentual de colaboradores ativos. Esta etapa permitiu o agrupamento de atributos com mesmo RR, e também a retirada de atributos considerados neutros e que portanto não acrescentavam informação à variável. A Tabela 4.2 demonstra a classificação de RR nos atributos da variável faixa etária do colaborador, em que as categorias Faixa4 e Faixa5 foram agrupadas. A partir da utilização do RR, retirou-se do estudo a variável indicadora de deficiência, por não contribuir para discriminação de pedidos de desligamento, além de vários atributos com RR neutro. Também foram desconsiderados atributos que não tinham o mínimo de 10 observações.

Tabela 4.2: Risco Relativo - Faixa etária do colaborador

Atributo	RR	Classificação
Faixa1	1.64	Muito bom
Faixa2	1.34	Bom
Faixa3	0.77	Mau
Faixa4	0.61	Muito mau
Faixa5	0.51	Muito mau

Em seguida, para cada uma das variáveis foi feita uma transformação dos seus atributos que permaneceram na análise em variáveis *dummies*, resultando em 52 variáveis *dummies* de um total de 17 variáveis categóricas do banco.

4.4 Regressão Logística

4.4.1 Construção do modelo

A técnica estatística escolhida foi a regressão logística, cuja variável resposta é o desfecho binário em que 1 é o pedido de desligamento, e 0 a permanência na empresa. O resultado do modelo é a probabilidade de pedido de desligamento do colaborador. Para a realização das análises e modelagens estatísticas, foi utilizada a linguagem de programação R na versão 4.2.2 (26), com apoio da interface Rstudio (27). A sintaxe utilizada está disponível no Apêndice .1.

Para a seleção das variáveis que entrariam no modelo logístico, foi utilizado o método de seleção de variáveis *stepwise*. Dentre as 52 variáveis *dummies* analisadas, o método *stepwise* selecionou inicialmente 19 variáveis para composição do modelo. Após feito o modelo, é necessário verificar o atendimento das suposições para utilização da técnica em questão. No caso da regressão logística, o pressuposto exigido é a ausência de multicolinearidade. A utilização do método *stepwise* para seleção das variáveis preditoras auxilia a minimizar a presença de multicolinearidade, mas ainda é preciso avaliar sua ausência. Assim, das 19 variáveis iniciais selecionadas pelo modelo, 2 delas foram retiradas por ter VIF alto, maior que 10. A especificação das variáveis finais do modelo é apresentada na Tabela 4.3.

Tabela 4.3: Relação de variáveis dummy selecionadas pelo método *stepwise* - Modelo Logístico

Variável	Descrição
DEMP1	Empresa 1
DIDAD1	Faixa etária 1
DIDAD2	Faixa etária 2
DIDAD3	Faixa etária 3
DSEX1	Sexo 1
DN6	Nível de atuação 6
DT1	Tempo de empresa 1
DSAL3	Faixa salarial 3
DSAL7	Faixa salarial 7
DDIR3	Diretoria 3
DDIR6	Diretoria 6
DDIR8	Diretoria 8
DDIR29	Diretorias 2 e 9
DRECS	Recebeu reconhecimento
DPERF1	Avaliação de performance 1
DPERF23	Avaliações de performance 2 e 3
DPERF4	Avaliação de performance 4

Das variáveis finais do modelo, a nível de significância de 5%, 11 tiveram poder discriminatório significativo. Na Tabela 4.4 tem-se os coeficientes do modelo e suas medidas associadas.

Tabela 4.4: Variáveis do Modelo via Regressão Logística, seus coeficientes e p-valores associados

Variável	Coeficiente	Exp(Coef)	P-valor
Intercepto	1.9496	7.03	0.00015
DEMP1	-4.2090	0.01	2.74e-09
DIDAD1	1.6202	5.05	6.73e-05
DIDAD2	0.9149	2.50	0.01091
DIDAD3	0.7340	2.08	0.05207
DSEX1	-0.6561	0.52	0.01638
DN6	-2.3208	0.10	4.21e-08
DT1	0.4898	1.63	0.05698
DSAL3	-0.5393	0.58	0.11060
DSAL7	-1.2561	0.28	0.03793
DDIR3	1.0972	3.00	0.05864
DDIR6	0.5763	1.78	0.07396
DDIR8	-2.1837	0.11	0.00272
DDIR29	-1.0432	0.35	0.00380
DRECS	-2.4278	0.09	2.80e-16
DPERF1	-2.0598	0.13	0.02320
DPERF23	-1.7691	0.14	4.26e-06
DPERF4	-1.9814	0.17	0.10707

As variáveis em que os coeficientes possuem sinais positivos revelam associações com o desfecho binário pedido de desligamento, indicando que indivíduos com a presença daquela variável são mais suscetíveis a pedir o desligamento. Ou seja, um colaborador que está na faixa etária 1 (DIDAD1), por exemplo, tem maior probabilidade de pedir desligamento (Coeficiente = 1.6202). Da mesma forma, coeficientes com sinal negativo revelam associações com a não ocorrência do desfecho binário pedido de desligamento, indicando que indivíduos com a presença daquela variável são menos propensos a sairem da empresa. Por exemplo, um colaborador que recebeu algum tipo de reconhecimento no último ano (DRECS) tem menor probabilidade de querer sair da empresa (Coeficiente = -1.7691).

4.4.2 Avaliação do modelo

As medidas de desempenho do modelo avaliadas foram a taxa de acerto, o valor do teste KS e a curva sob a área ROC. Na Tabela 4.5, são apresentados os valores observados e os valores preditos pelo modelo.

Tabela 4.5: Matriz de confusão - Modelo logístico

	Valor predito	
Valor observado	Ativo	Pedido desligamento
Ativo	336	29
Pedido desligamento	64	105

A taxa de acerto foi de 82,58%, demonstrando desempenho satisfatório do modelo. O valor do teste KS foi de 59,62% e da medida AUC foi de 87%, portanto considera-se que o modelo foi eficiente na predição dos dois grupos. Essas medidas demonstram que o modelo possui bom poder de separação entre os desligamentos voluntários e os colaboradores ativos. Na Figura 4.1 é possível ver a representação da curva ROC.

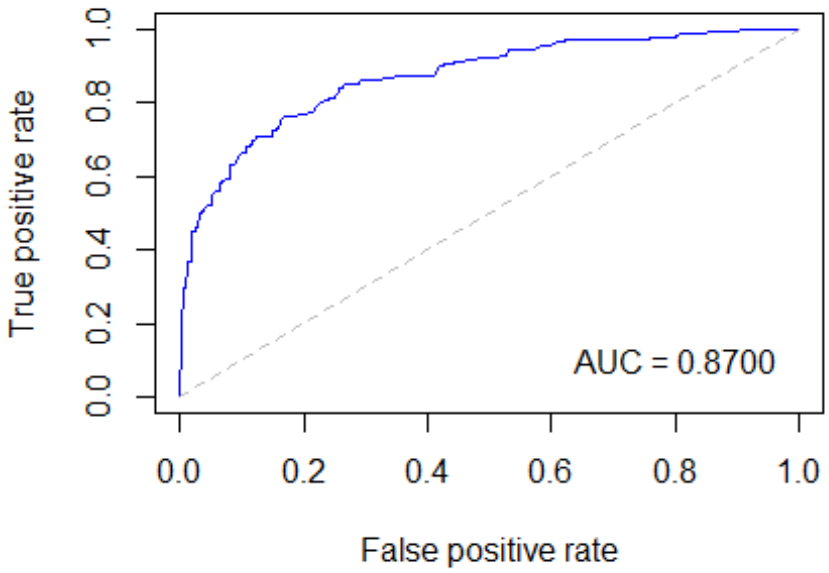

Figura 4.1: Área sob a Curva ROC - Modelo Logistico

4.5 Análise de Sobrevivência

4.5.1 Construção do modelo

A técnica estatística escolhida foi análise de sobrevivência, cuja variável resposta de interesse é o tempo até a ocorrência do evento. Para a realização das análises e modelagens estatísticas, foi utilizada a linguagem de programação R na versão 4.2.2 (26), com apoio da interface Rstudio (27). A sintaxe utilizada está disponível no Apêndice .2.

Para a seleção das variáveis que entrariam no modelo de Cox, foi utilizado o método de seleção de variáveis *stepwise*. Dentre as 52 variáveis *dummies* analisadas, o método *stepwise* selecionou 16 variáveis para composição do modelo. A especificação das variáveis escolhidas é apresentada na Tabela 4.6.

Tabela 4.6: Relação de variáveis dummy selecionadas pelo método *stepwise* - Modelo de Cox

Variável	Descrição
DEMP1	Empresa 1
DIDAD1	Faixa etária 1
DIDAD2	Faixa etária 2
DIDAD3	Faixa etária 3
DSEX1	Sexo 1
DN6	Nível de atuação 6
DT1	Tempo de empresa 1
DSAL456	Faixas salariais 4,5 e 6
DDIR3	Diretoria 3
DDIR6	Diretoria 6
DDIR8	Diretoria 8
DDIR29	Diretorias 2 e 9
DRECS	Recebeu reconhecimento
DPERF1	Avaliação de performance 1
DPERF23	Avaliações de performance 2 e 3
DPERF4	Avaliação de performance 4

Das variáveis selecionadas pelo *stepwise*, 12 variáveis tiveram poder discriminatório significativo. Na Tabela 4.7 tem-se os coeficientes do modelo e suas medidas associadas.

Tabela 4.7: Variáveis do Modelo via Modelo proporcional de Cox, seus coeficientes, exponencial dos coeficientes e p-valores associados

Variável	Coeficiente	Exp(Coef)	P-valor
DEMP1	-2.99263	0.05016	2.01e-08
DIDAD1	1.05363	2.86804	2.92e-05
DIDAD2	0.58395	1.79311	0.012809
DIDAD3	0.54377	1.72248	0.040405
DSEX1	-0.37907	0.68450	0.024511
DN6	-1.24166	0.28890	3.87e-06
DT1	0.32914	1.38977	0.055064
DSAL456	0.39422	1.48322	0.057267
DDIR3	0.60867	1.83798	0.081745
DDIR6	0.39193	1.47983	0.040649
DDIR8	-1.67910	0.18654	0.005575
DDIR29	-0.82843	0.43674	0.000813
DRECS	-1.70332	0.18208	$<$ 2e-16
DPERF1	-1.60634	0.20062	0.028736
DPERF23	-1.44335	0.23614	6.44e-12
DPERF4	-1.92949	0.14522	0.061862

As variáveis cujo exponencial do coeficiente estimado pelo modelo foram maiores que 1 representam fatores de risco, isto é, tornam o indivíduo mais suscetível a ocorrência do evento. Por exemplo, indivíduo da faixa etária 1 (DIDAD1), cujo exponencial do coeficiente foi maior que 1 (Exponencial coeficiente = 2.86804), indicando que esta variável contribui para que o colaborador peça o desligamento. Variáveis com exponencial do coeficiente menores que 1, como recebimento de reconhecimento no último ano (DRECS; Exponencial coeficiente = 0.18208), representam fatores de proteção, isto é, tornam o indivíduo menos suscetível a ocorrência do evento.

Após feito o modelo, é necessário verificar o atendimento das suposições para utilização da técnica em questão. No caso do modelo de Cox, o pressuposto necessário é a proporcionalidade dos riscos. As Figuras **??** a **??** demonstram a proporcionalidade das taxas de falha ao longo do tempo para as variáveis selecionadas, pela plotagem dos resíduos de Schoenfeld. Apesar do valor teste demonstrar violação do pressuposto, o teste é sensível à distribuição dos resíduos e tamanho da amostra, portanto a análise gráfica é mais recomendada para avaliação do pressuposto. Após a realização da análise gráfica, não identifica-se padrões ao longo do tempo nos resíduos. Como todas as variáveis são categóricas, não é preciso se preocupar com o pressuposto da linearidade.

4.5.2 Avaliação do modelo

As medidas de desempenho do modelo avaliadas foram o valor do teste KS e a curva sob a área ROC.

O valor do teste KS foi de 57% e da medida AUC foi de 86,7%, portanto considera-se que o modelo foi eficiente na predição dos dois grupos. Essas medidas demonstram que o modelo possui bom poder de separação entre os desligamentos

voluntários e os colaboradores ativos. Na Figura 4.2 pode-se ver a representação da curva ROC.

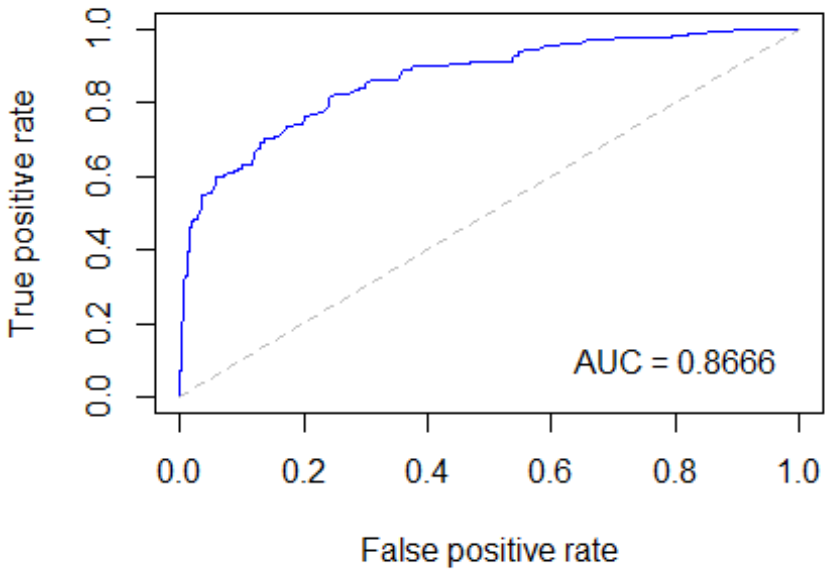

Figura 4.2: Área sob a Curva ROC - Modelo de Cox

4.6 Visualização de dados

A ferramenta de visualização utilizada é uma matriz que cruza as variáveis resposta de cada modelo, isto é, probabilidade de ocorrência e tempo até evento, para que haja mais assertividade em relação a previsão pedido de desligamento. Assim, constrói-se o gráfico com o eixo y sendo a probabilidade de ocorrência e o eixo x o tempo estimado até o evento, na unidade de anos. Em seguida, é decidido os pontos de corte para divisão de cada uma das variáveis resposta em categorias. Sendo assim, as categorias para a variável resposta probabilidade de ocorrência são descritas na Tabela 4.8 e para a variável resposta tempo até o evento na Tabela 4.9.

Tabela 4.8: Categorias da probabilidade de ocorrência do pedido de desligamento

Probabilidade	Intervalo de probabilidade
Baixa	[0 - 0.25)
Média	[0.25 - 0.75)
Alta	[0.75 - 1]

Tabela 4.9: Categorias do tempo estimado até o pedido de desligamento

Ocorrência do evento	Intervalo de tempo
Próximo	[0 - 0.25 anos)
Médio	[0.25 - 0.75 anos)
Distante	A partir de 0.75 anos

Dessa forma, o cruzamento entre as categorias de probabilidade e tempo até ocorrência do evento geram quadrantes de priorização, classificados em Baixa, Média e Alta prioridade. A ferramenta de visualização construída e proposta neste trabalho é apresentada na Figura 4.3, com suas respectivas proporções, e as proporções em cada categoria de priorização são exibidas na Tabela 4.10.

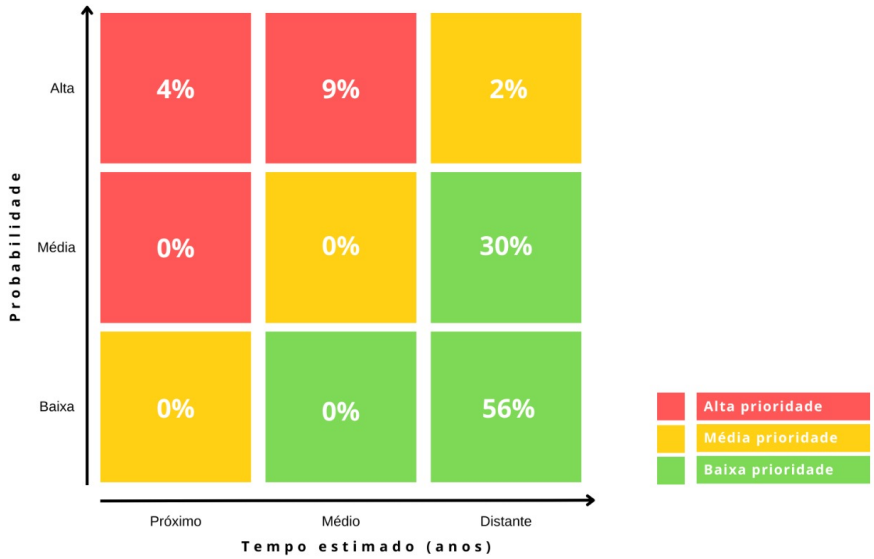

Figura 4.3: Matriz de prioridades de ação em relação à desligamentos voluntários

Tabela 4.10: Proporção por categoria da matriz de prioridade

Categoria de prioridade	Proporção
Alta prioridade	13%
Média prioridade	2%
Baixa prioridade	86%

A distribuição da matriz indica que 13% da amostra são indivíduos com alta prioridade de ação em relação à sua retenção na empresa, visto que possuem alta probabilidade de saída em um espaço curto de tempo. Dessa forma, a matriz possibilita ao gestor que tome sua decisão entre manter ou não o colaborador. Caso não

seja de interesse reter aquele indivíduo, a ferramenta de visualização auxilia para que a empresa tenha um tempo hábil maior para o processo de substituição do colaborador. Caso a decisão seja agir para evitar o desligamento do indivíduo, recorre-se aos modelos desenvolvidos a fim de avaliar variáveis que atuam como fatores de proteção e agir sobre elas como forma de reter o colaborador.

5 Considerações finais

Este trabalho fez o uso de um banco de dados real utilizado no setor de RH, para o desenvolvimento de modelos de previsão de pedidos de desligamento por meio de duas metodologias distintas: Regressão Logística e Análise de Sobrevivência. A previsão de desligamentos voluntários é pertinente pois observa-se custos financeiros e de produtividade ao perder um funcionário, substituí-lo e capacitar o novo colaborador.

O objetivo principal deste trabalho foi a predição do desligamento voluntário do colaborador sob dois aspectos distintos: a probabilidade de ocorrência e o tempo até o evento. A metodologia de Regressão Logística é bastante utilizada em problemas da área, enquanto a técnica de Análise de Sobrevivência ainda é pouco conhecida. O modelo de Regressão Logística foi utilizado tanto pelo interesse em sua variável resposta, quanto pelo seu histórico como modelo usualmente utilizado neste tipo de problema. O modelo de Análise de Sobrevivência prediz o tempo até ocorrência do evento levando em consideração as observações em que o evento ainda não ocorreu, sendo mais vantajoso que outros modelos de predição que ignorariam estes dados por serem faltantes, visto que na problemática do trabalho a grande maioria das observações são censuras e isso reduziria expressivamente a amostra.

As variáveis mais importantes para o ajuste do modelo de Regressão Logística foram as de faixa etária do colaborador, recebimento de reconhecimentos, performance do indivíduo e setor do qual faz parte. Enquanto na análise de sobrevivência, o nível de atuação impactou sobre a variável resposta de forma mais significativa que no logístico. As variáveis selecionadas em ambos modelos são as mesmas, com exceção da seleção de diferentes *dummies* relacionadas a faixa salarial nos modelos. As variáveis presentes em ambos modelos tem comportamentos consistentes em relação a variável resposta, seja aumentando o risco de pedido de desligamento ou diminuindo. Em linhas gerais, observa-se que as variáveis demográficas tem uma influência bem menos expressiva nos pedidos de desligamento em comparação as variáveis laborais.

Os resultados das medidas indicadoras de desempenho utilizadas para avaliar os modelos desenvolvidos se mostraram bem próximas e pode-se considerar que ambos tiveram desempenho satisfatório em termos de capacidade de predição de desligamentos voluntários para o banco de dados analisado. O modelo de Regressão Logística teve medidas levemente melhores que o modelo de Análise de Sobrevivência nas medidas em comum de ambos: área abaixo da curva ROC e medida do teste KS. Porém, não há intenção em comparar os modelos, visto que eles predizem características diferentes a respeito dos pedidos de desligamento. É interessante, porém,

destacar o desempenho satisfatório e próximo do modelo de Análise de Sobrevivência, pouco utilizado para problemas deste tipo, em relação ao modelo de Regressão Logística, tradicionalmente empregado.

Além disso, o objetivo do trabalho era cruzar estes dois diferentes aspectos preditos de desligamentos voluntários a fim de criar uma ferramenta de visualização unificada, que permitisse aos gestores uma maior acessibilidade aos resultados, mas também um instrumento estratégico para tomada de decisões em um contexto de grande volume de dados. O resultado foi uma matriz com quadrantes de prioridade que refletem o cruzamento entre probabilidade de saída e tempo até ocorrência do evento, facilitando que se priorize colaboradores com alto risco de desligamento e fazendo com que os resultados das técnicas utilizadas sejam aplicadas de forma ainda mais assertiva por quem irá fazer uso delas, ao utilizar dos modelos desenvolvidos como uma das ferramentas para retenção do colaborador por meio dos coeficientes que atuam como fatores de proteção em relação ao desfecho.

Com relação às limitações do estudo, a dificuldade em encontrar registros históricos consistentes fez com que fosse necessário limitar, no caso da técnica de Análise de Sobrevivência, a duração do experimento pelo período de aproximadamente 2 anos. Tendo sido maior este tempo, é possível que o modelo apresentasse desempenho superior ao relatado neste trabalho. Também decorrente desta dificuldade diversas variáveis que poderiam ter influência na predição do desligamento, como banco de horas, treinamentos e engajamento interno, apenas recentemente vêm sendo mapeadas e registradas, portanto não foi possível utiliza-lás na modelagem.

Como tópicos de pesquisas futuras, é possível utilizar-se da rescisão por iniciativa da empresa, desconsiderada no atual trabalho, para modelagens na área de atração e seleção de colaboradores. Também enquanto sugestão para trabalhos futuros, propõe-se uma etapa seguinte a construção da matriz de prioridades que seria sua disponibilização através da implementação em *softwares* de visualização de dados, como o *Power BI*. De forma complementar, a construção de um *dashboard* que disponibilize o perfil dos indivíduos no que diz respeito as variáveis significativas dos modelos, para visualizar, de forma mais direta, quais covariáveis que o colocam em determinada situação de risco de desligamento.

Referências Bibliográficas

[1] E. Vulpen, *The basic principles of People Analytics.* AIHR, 2019.

[2] I. Chiavenato, *Gestão de Pessoas - O Novo Papel dos Recursos Humanos nas Organizações.* São Paulo: Elsevier, 4th ed., 2014.

[3] J. Marler and J. Boudreau, "An evidence-based review of hr analytics," *The International Journal of Human Resource Management*, vol. 28, pp. 3–26, 2017.

[4] J. P. Isson and J. S. Harriott, *People Analytics in the Era of Big Data: Changing the Way You Attract, Acquire, Develop.* John Wiley & Sons, 2016.

[5] L. Selau and J. Ribeiro, "Uma sistemática para a construção e escolha de modelos de previsão de risco de crédito," *Gestão Produção*, vol. 16, no. 3, pp. 398–413, 2009.

[6] E. Lee and J. Wang, *Statistical Methods for Survival Data Analysis.* John Wiley Sons, 2003.

[7] F. Louzada-Neto, "Modelagem temporal para credit scoring - uma nova alternativa à modelagem tradicional via análise de sobrevivência," *Revista tecnologia de crédito*, vol. 56, pp. 8–22, 2014.

[8] N. Bontis, "Intellectual capital: an exploratory study that develops measures and models," *Management Decision*, vol. 36, pp. 63–76, 1998.

[9] D. E. A. Angrave, "Hr and analytics: why hr is set to fail the big data challenge," *Human Resource Management Journal*, vol. 26, pp. 1–11, 2016.

[10] M. Lucena, *Planejamento de recursos humanos.* Atlas, 2007.

[11] H. Boushey and S. Glynn, "There are significant business costs to replacing employees," *Human Resource Management Journal*, vol. 22, pp. 150–166, 2012.

[12] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models.* Chapman Hall, 3rd ed., 2008.

[13] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression.* John Wiley & Sons, 3rd ed., 2013.

[14] J. e. a. Hair, *Análise Multivariada de Dados.* Bookman Companhia Editora Ltda, 2009.

[15] E. A. Colosimo and S. R. Giolo, *Análise de Sobrevivência Aplicada*. Blucher, 2006.

[16] N. Breslow, "Covariance analysis of censored survival data," *Biometrics*, vol. 30, no. 1, pp. 89–99, 1974.

[17] J. J. D. Kalbfleisch and R. Prentice, "Marginal likelihoods based on cox's regression and life model," *Biometrika*, vol. 60, pp. 267–278.

[18] C. Knaflic, *Storytelling com dados: Um guia sobre visualização de dados para profissionais de negócios*. Alta Books, 2016.

[19] B. Davies and B. Davies, "Talent management in academies," *International Journal of Educational Management*, vol. 24, pp. 418–426, 2010.

[20] E. M. Lewis, *An Introduction to Credit Scoring*. Athena Press, 1992.

[21] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models*. McGraw-Hill Irwin, 2005.

[22] D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. John Wiley Sons, 2008.

[23] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[24] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[27] RStudio Team, *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2021.

.1 Sintaxe Regressão Logística

```
library(car)
library(ROCR)
library(creditR)
library(pROC)

#Stepwise
stepwise <- step(modelo.completo,direction="both")

#Modelagem
stepwise <- glm(stepwise$formula, family=binomial,data=dados)

#Resultados modelo
summary(stepwise)

#Verificação de pressupostos
vif(stepwise)

#Predição
dados$score<-predict(stepwise,type='response',dados)
pred<-prediction(dados$score, dados$DESFECHO)

#Teste KS
Kolmogorov.Smirnov(dados,"DESFECHO","score")

#Matriz de confusão e taxa de acerto
dados$predito<-ifelse(dados$score>=0.5,1,0)
tab<-table(dados$DESFECHO,dados$predito)
taxa.acerto<-(tab[2,2]+tab[1,1])/sum(tab)

# Gráfico da área ROC e medida AUC
roc1=plot.roc(dados$DESFECHO,fitted(stepwise))
plot(roc1,
print.auc=TRUE,
auc.polygon=TRUE,
grud=c(0.1,0.2),
grid.col=c("green","red"),
max.auc.polygon=TRUE,
auc.polygon.col="lightgreen",
print.thres=TRUE)
```

.2 Sintaxe Análise de Sobrevivência

```
library(survival)
library(survminer)
library(creditR)
library(ROCR)
library(pROC)

#Stepwise
stepwise <- step(fit2,direction="both")

#Modelagem
stepwise<-coxph(stepwise$formula,data=dados,x=T,method="breslow")

#Resultados modelo
summary(stepwise)

#Verificação de pressupostos
test.ph = cox.zph(stepwise)
ggcoxzph(test.ph)

#Predição
dados$score<-predict(stepwise,type='lp',dados)
trueY <- dados$STATUS
probs <- dados$score

#Teste KS
Kolmogorov.Smirnov(dados,"STATUS","score")

# Gráfico da área ROC e medida AUC
rocplot <- function(pred, truth, ...)  {
predob = prediction(pred, truth)
perf = performance(predob, "tpr", "fpr")
plot(perf, ...)
area <- auc(truth, pred)
area <- format(round(area, 4), nsmall = 4)
text(x=0.8, y=0.1, labels = paste("AUC =", area))
segments(x0=0, y0=0, x1=1, y1=1, col="gray", lty=2) }
rocplot(probs, trueY, col="blue")
```