Churn de Colaboradores

Cliente: HR Insights
Analista de Dados: Cristiane Thiel
Links: Apresentação - GitHub - Vídeo de Apresentação
Ferramentas e Tecnologias: Python e VS Code

Objetivo da Análise

O objetivo central da análise é prever a probabilidade de um funcionário deixar a empresa, usando um modelo de aprendizado de máquina supervisionado. Com isso, o time de RH poderá:

- Identificar funcionários em risco de saída
- Agir preventivamente com estratégias de retenção
- Reduzir custos com demissões e novas contratações
- Melhorar o clima organizacional e o engajamento dos colaboradores

Contexto do Negócio

O contexto de negócio é a **retenção de talentos** em um mercado competitivo, onde a alta rotatividade de funcionários gera prejuízos financeiros, perda de conhecimento e impacto negativo na produtividade. A empresa deseja usar dados históricos de RH para prever quais funcionários têm maior risco de sair, permitindo ao time de Recursos Humanos tomar decisões baseadas em dados para reter esses talentos e manter a estabilidade operacional.

Possíveis Stakeholders

- Equipe de Recursos Humanos – principal usuária do modelo para ações de retenção.
- Gestores de Equipe – interessados em manter talentos nos times.
- Diretoria/Executivos – preocupados com o impacto financeiro e estratégico da rotatividade.
- Equipe de People Analytics – responsável por análises e interpretação dos dados.

Perguntas de Negócios

1. Quais funcionários estão em risco de deixar a empresa?
2. Quais fatores influenciam a saída de um funcionário?
3. É possível prever a rotatividade antes que aconteça?

Metodologia do Projeto

- ➔ Tratamento e Qualidade dos Dados
  - ◆ Limpeza de inconsistências e exclusão de registros nulos irreparáveis
  - ◆ Correção de relações lógicas entre variáveis (ex: experiência vs. empresas anteriores)
- ➔ Análise Exploratória (EDA)
  - ◆ Identificação de padrões críticos:
    - Perfis de alto risco (viagens frequentes, cargos estratégicos)
    - Relação entre salário, tempo de empresa e churn
- ➔ Engenharia de Features e Modelagem
  - ◆ Seleção de variáveis preditoras estratégicas
  - ◆ Divisão robusta: treino/teste + validação cruzada (5 folds)
  - ◆ Treinamento de dois modelos:
    - XGBoost (priorizando recall)
    - Random Forest (priorizando precisão)
- ➔ Otimização e Validação Rigorosa
  - ◆ Ajuste de hiperparâmetros via busca aleatória
  - ◆ Avaliação com métricas múltiplas:
    - Curvas ROC, Precision-Recall e Learning Curve
    - Análise de fairness por faixa etária e renda
- ➔ Benchmark e Decisão Estratégica
  - ◆ Comparação detalhada de desempenho entre modelos
  - ◆ Seleção final baseada em trade-off entre recall e precisão

Processamento e Análise

Esses são os tipos corretos para cada variável da base.

Variável	Tipo	Subtipo	Escala
Attrition	Qualitativa Dicotômica	Categórica binária	Nominal
EmployeeID	Identificador		
Age, MonthlyIncome	Quantitativa Contínua	Numérica	Razão
NumCompaniesWorked, TotalWorkingYears, YearsAtCompany, DistanceFromHome, PercentSalaryHike, StockOptionLevel, TrainingTimesLastYear, YearsSinceLastPromotion, YearsWithCurrManag	Quantitativa Discreta	Contagem	Razão
JobLevel, Education	Qualitativa ordinal	Categórica com ordem	Ordinal

Gender, MaritalStatus, Department, BusinessTravel, JobRole, EducationField	Qualitativa nominal	Categórica sem ordem	Nominal

A base original contém **4.410 registros**. Durante a inspeção inicial, foram identificados os seguintes ajustes e tratamentos necessários:

Tratamentos Necessários

1. A variável ATTRITION está armazenada como texto (Yes/No) e foi convertida para booleana (True/False), facilitando a análise e a modelagem supervisionada.
2. As variáveis EmployeeCount, Over18 e StandardHours apresentaram valores constantes em todos os registros:

- EmployeeCount: sempre 1
- Over18: sempre "Y"
- StandardHours: sempre 8

Por não contribuírem com variabilidade ou informação útil para o modelo, foram removidas da base.

Valores ausentes identificados

- NumCompaniesWorked: 19 registros nulos
- TotalWorkingYears: 9 registros nulos

Análise de Inconsistências

Contradição Lógica entre Experiência Total e Número de Empresas

A relação lógica esperada entre as variáveis é:

Se TotalWorkingYears > YearsAtCompany, então NumCompaniesWorked ≥ 1

No entanto, foram identificados 584 registros onde:

- NumCompaniesWorked = 0
- TotalWorkingYears > YearsAtCompany

Diante disso, podemos interpretar que:

- A variável NumCompaniesWorked não contabiliza a empresa atual
- A diferença entre os anos pode refletir experiências anteriores informais, autônomas ou não registradas formalmente
- Pode existir uma regra de negócio implícita que considera apenas vínculos ou similares

Tratamento de Dados Nulos

Com base na lógica observada, foi aplicada a seguinte regra de imputação:

- Se o número de empresas for nulo e o total de anos for igual aos anos na empresa atual, então a pessoa nunca trabalhou em outra empresa, NumCompaniesWorked = 1.
- Se o total de anos está nulo e a pessoa nunca trabalhou em outra empresa (zero empresas), então YearsAtCompany = TotalWorkingYears.

Valores ausentes restantes

- NumCompaniesWorked: 14 registros nulos
- TotalWorkingYears: 5 registros nulos

Representando aproximadamente 0,43% da base, optei por excluir esses registros, pois não havia base lógica ou estrutural confiável para a imputação.

Tipo de Variáveis

- NumCompaniesWorked de float para inteiro
- TotalWorkingYears de float para inteiro

Análise Exploratória dos Dados

Após o tratamento temos 4.391 registros.

- 1 variável (Attrition) booleana (int64) = 0 ou 1
- 14 variáveis numéricas (int64)
- 6 variáveis categóricas (object)

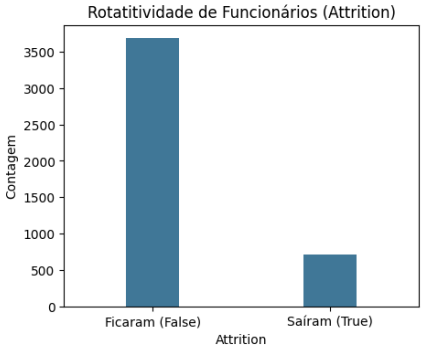
Resumo das Estatísticas Descritivas

- **Idade (Age)**: Média de 36.92 anos, com desvio padrão de 9.14. A faixa etária varia de 18 a 60 anos.
- **Distância de casa (DistanceFromHome)**: A média é 9.19 km, com valores que vão de 1 a 29 km.
- **Nível de Educação (Education)**: A maioria dos empregados possui entre 2 e 4 anos de educação, com a média em 2.91.
- **Nível de Cargo (JobLevel)**: Média de 2.06, com a maioria dos empregados ocupando cargos de nível 1 ou 2.
- **Renda Mensal (MonthlyIncome)**: A média é de 65.025, com um desvio significativo de 47.112, sugerindo possíveis outliers.

- **Número de Empresas Trabalhadas (NumCompaniesWorked)**: Em média, os empregados trabalharam em 2.69 empresas, variando de 0 a 9.
- **Aumento Salarial Percentual (PercentSalaryHike)**: A média é de 15.21%, com um aumento mais comum em torno de 14%.
- **Anos Totais de Trabalho (TotalWorkingYears)**: Média de 11.28 anos, com a maioria tendo entre 6 a 15 anos de experiência.
- **Anos Trabalhados na Empresa (YearsAtCompany)**: A média é de 7.01 anos, com a maioria dos empregados trabalhando entre 3 a 9 anos.
- **Anos Desde a Última Promoção (YearsSinceLastPromotion)**: Média de 2.19 anos, sugerindo que a maioria das promoções ocorre em até 3 anos.
- **Anos com o Atual Gestor (YearsWithCurrManager)**: A média é de 4.12 anos, com variação de 0 a 17 anos.

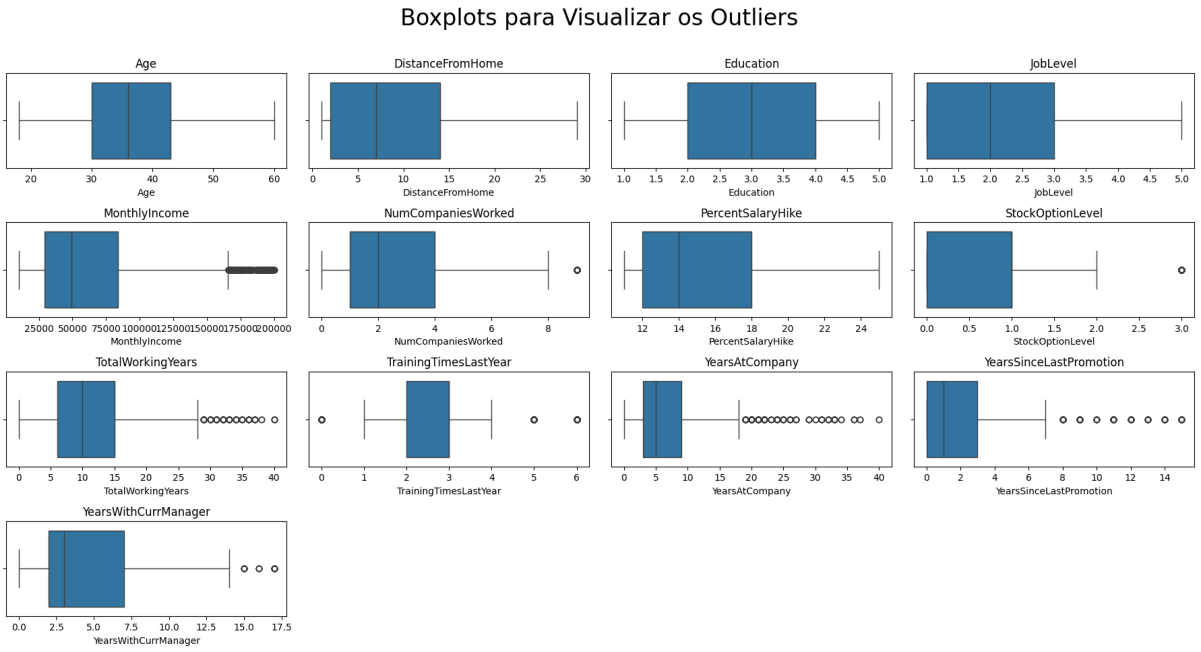
Rotatividade de Funcionários (Attrition)

Percentual de pessoas que saíram da empresa: 16.10%



Análise de Distribuição e Outliers com Boxplots

O boxplot é excelente para rapidamente identificar a dispersão, a mediana (Q2) e os outliers. Ele resume os dados em quartis, mas esconde a forma exata da distribuição. Por isso, depois vamos analisar a distribuição com histogramas.



Entendendo o Perfil dos Funcionários

A análise mostra que a maioria dos colaboradores está em níveis iniciais ou plenos, com menor tempo de casa e remuneração mais baixa. Variáveis como *MonthlyIncome*, *YearsAtCompany* e *TotalWorkingYears* indicam concentração em faixas baixas, mas também a presença de um grupo menor de profissionais sêniores e executivos com valores elevados.

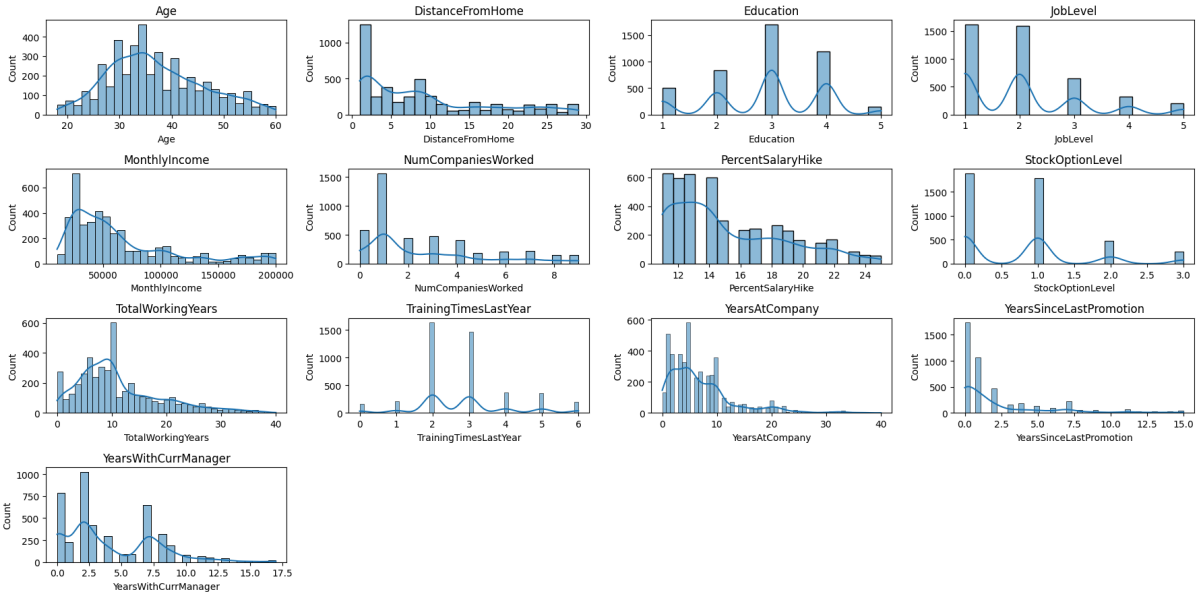
A maior parte dos funcionários reside relativamente próxima ao trabalho (*DistanceFromHome*) e os aumentos salariais (*PercentSalaryHike*) têm baixa variação.

Decisão: Os outliers encontrados (ex: altos salários ou longa permanência) foram mantidos por representarem perfis estratégicos. Removê-los poderia comprometer a capacidade do modelo em aprender padrões de retenção desses grupos.

Análise de Frequências com Histogramas

O histograma é perfeito para ver a forma da distribuição. Podemos ver picos (modas), vales, se os dados são simétricos, assimétricos, unimodais, bimodais (como em Education e JobLevel). Ele nos dá uma visão muito mais granular de onde os dados se concentram.

Histogramas para Visualizar a Distribuição das Variáveis



Assimetria à direita (Right-Skewed): Variáveis como MonthlyIncome, TotalWorkingYears e YearsAtCompany apresentam cauda longa à direita. Isso indica que a maioria dos funcionários possui **salários menores**, **menos tempo de casa e de carreira**, com poucos casos extremos em valores altos. Esse padrão é comum em dados de Recursos Humanos.

Distribuições multimodais: Variáveis como Education e JobLevel apresentam **múltiplos picos**, reforçando que são variáveis ordinais, e não contínuas. A maioria dos funcionários está concentrada nos **níveis 3 (educação)** e nos **níveis 1 e 2 (nível de cargo)**.

Concentração em valores baixos

- DistanceFromHome: A maior parte dos funcionários **mora a até 5 km (ou milhas)** da empresa.
- YearsSinceLastPromotion: A maioria foi **promovida há 0 ou 1 ano**, indicando ou promoções recentes ou longos períodos sem promoção. Isso pode ser um fator importante na previsão de Attrition.
- Distribuição aproximadamente normal: A variável Age é a que mais se aproxima de uma distribuição normal, **com concentração na faixa dos 30 a 35 anos**.

A análise combinada de boxplots e histogramas permitiu compreender bem o comportamento das variáveis. Identifique padrões esperados para dados de RH, validei os outliers como valores legítimos e confirmei a importância dessas variáveis para futuras etapas, como engenharia de atributos e modelagem preditiva.

Análise de Correlação entre Variáveis Numéricas

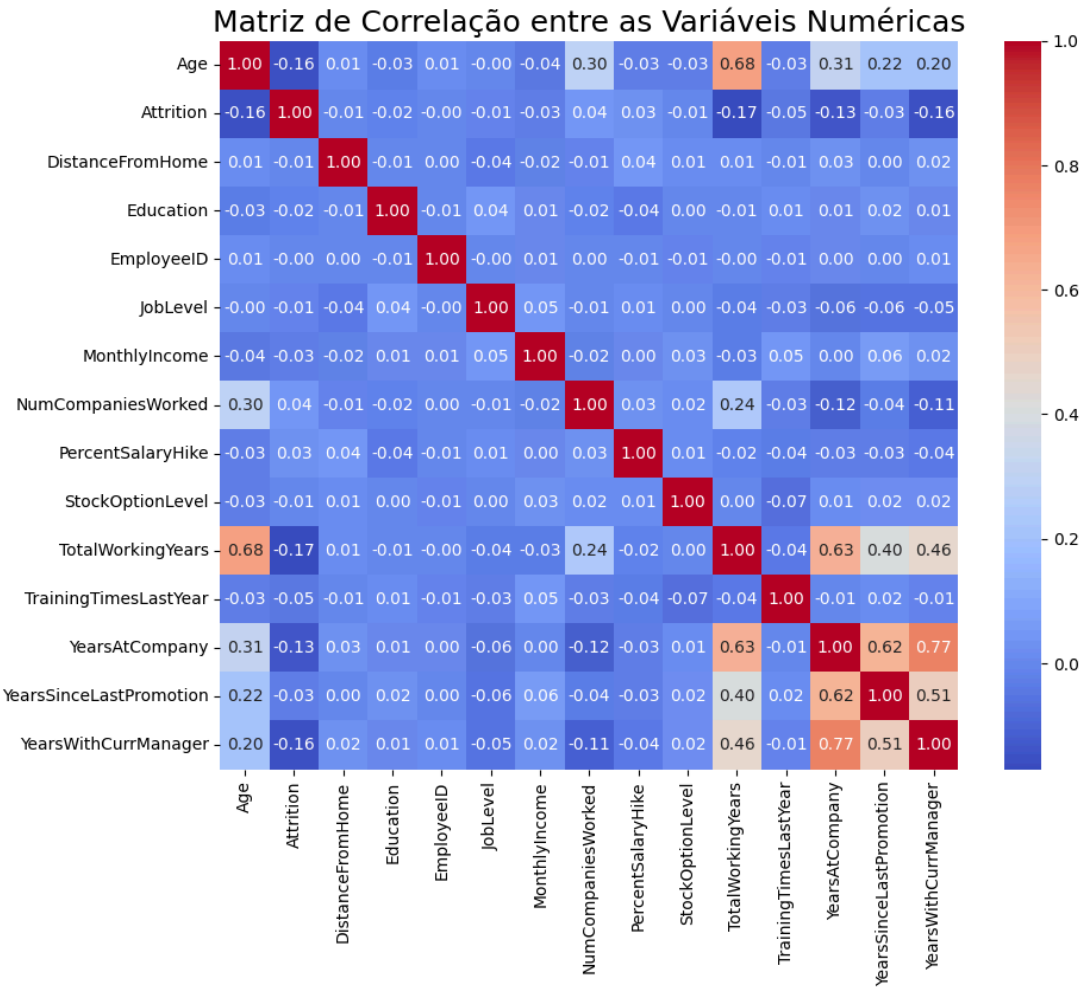
A matriz de correlação (heatmap) foi utilizada para quantificar a **relação linear entre as variáveis numéricas**. As principais observações são:

Correlações Positivas Fortes (Esperadas)

- Como esperado, há uma forte correlação positiva entre **YearsAtCompany (anos na empresa) e YearsWithCurrManager (anos com o gestor atual)**, com um coeficiente de 0.77. Isso indica que, para muitos funcionários, o tempo na empresa é muito semelhante ao tempo com seu líder direto.
- A **Age (idade) e TotalWorkingYears (total de anos de experiência)** também apresentam uma correlação forte (0.68).
- Variáveis relacionadas ao tempo de serviço, como **TotalWorkingYears, YearsAtCompany e YearsSinceLastPromotion**, estão moderada a fortemente correlacionadas entre si (coeficientes entre 0.40 e 0.77).

Correlações Negativas e Fracas

Não foram identificadas correlações negativas fortes. A variável alvo, Attrition, apresenta correlações negativas fracas com variáveis como TotalWorkingYears (-0.17), Age (-0.16) e YearsWithCurrManager (-0.16), sugerindo que funcionários com mais experiência, mais idade e mais tempo com o mesmo gestor têm uma **tendência ligeiramente menor de deixar a empresa**.



Implicações para a Modelagem: Risco de Multicolinearidade

A **forte correlação** entre as variáveis de tempo (YearsAtCompany, TotalWorkingYears, etc.) indica um potencial **problema de multicolinearidade**. Isso significa que essas variáveis carregam **informações redundantes**.

Na etapa de construção do modelo, essa multicolinearidade deve ser levada em conta. Embora algoritmos baseados em árvore (como os utilizados neste projeto) sejam robustos a este fenômeno em termos de performance preditiva, a redundância de informações pode impactar a interpretação da importância de cada feature. O modelo pode "dividir" a importância entre variáveis correlacionadas, o que exige uma análise crítica dos resultados para entender os verdadeiros drivers do churn.

Análise Bivariada: Onde o Churn Acontece?

Analisando como a rotatividade se distribui entre diferentes departamentos, cargos e rotinas de trabalho para identificar os fatores de churn na organização.

- **O Desgaste das Viagens:** A frequência de viagens é um dos fatores mais críticos. Funcionários que viajam frequentemente têm a maior taxa de saída, enquanto aqueles que não viajam quase nunca deixam a empresa. Isso aponta para um problema de desgaste e qualidade de vida.

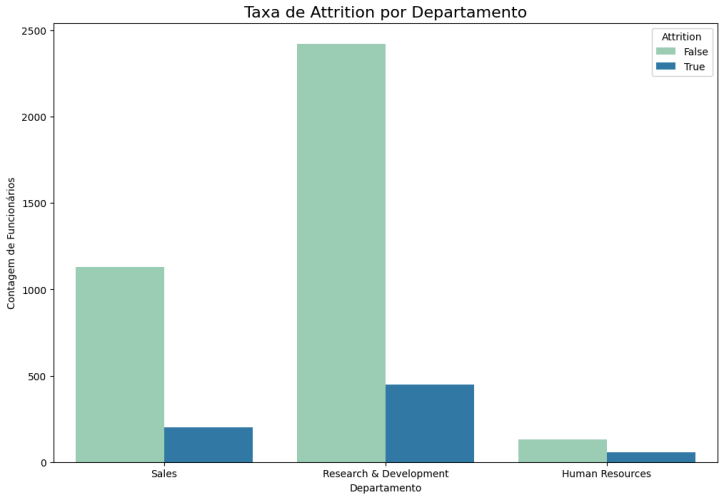
- **A Pressão dos Cargos de Entrada e Operacionais:** O cargo é um dos maiores determinantes da rotatividade. Cargos como Técnico de Laboratório, Representante de Vendas e profissionais de Recursos Humanos apresentam as maiores taxas de churn. Em contraste, cargos de liderança (Gerente, Diretor) mostram uma retenção muito alta.

- **Departamentos Críticos:** Embora o departamento de P&D tenha mais saídas em números absolutos, os departamentos de Vendas e RH são proporcionalmente os que mais perdem talentos, indicando que podem ser as áreas com os maiores desafios de retenção.

O risco de churn não está distribuído de forma homogênea. Ele está concentrado em perfis específicos: **funcionários operacionais e de vendas, que viajam com frequência.** As estratégias de retenção devem ser direcionadas a esses grupos, com ações focadas em melhorar as **condições de trabalho**, oferecer **planos de carreira** claros para cargos de entrada e **gerenciar a carga de viagens**.

Variáveis Categóricas vs. Rotatividade (Attrition)

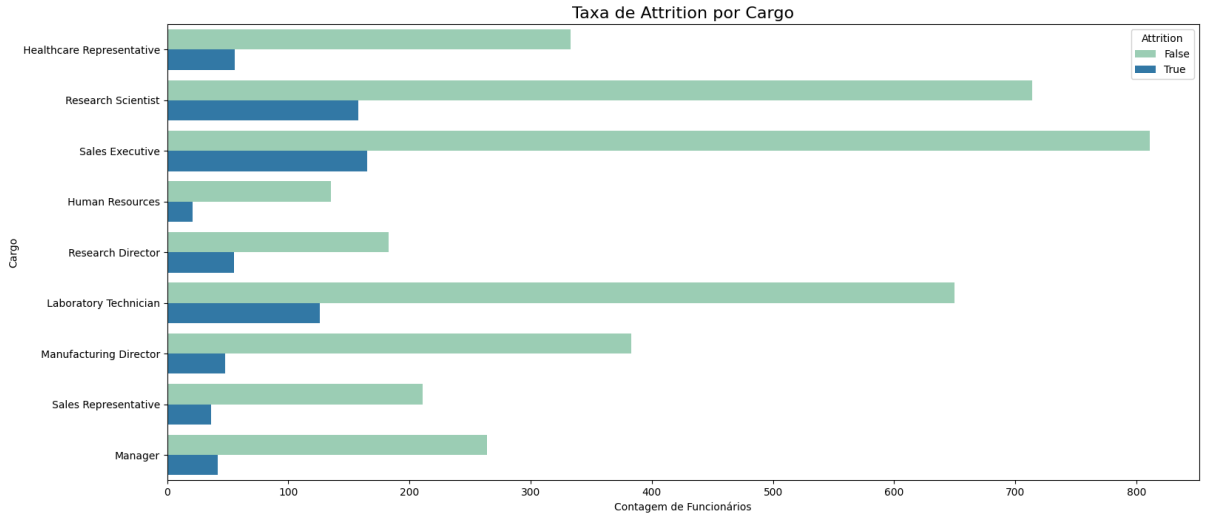
Departamento: Apesar de P&D ter mais saídas absolutas, Sales e RH apresentam maior proporção de rotatividade. **Principalmente RH.**



Cargo: Cargos operacionais ou de entrada têm maior propensão a sair, enquanto cargos de liderança mostram maior retenção.

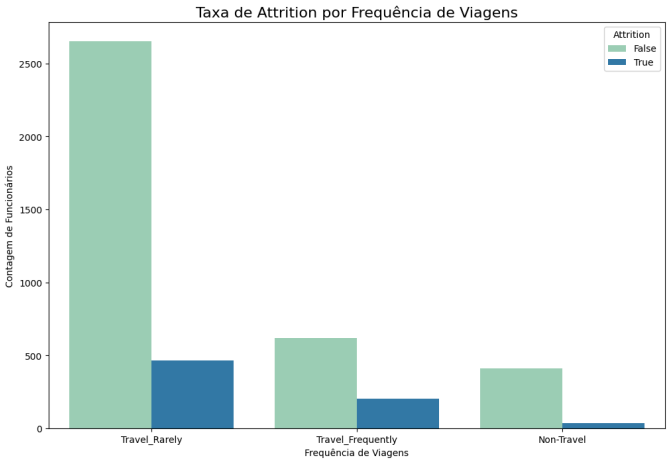
Alta rotatividade: Técnico de Laboratório, Representante de Vendas, Pesquisador Cientista e RH.

Baixa rotatividade: Gerente, Diretor de Manufatura e Representante de Saúde.



Frequência de Viagens: A frequência de viagens impacta negativamente a permanência.

Funcionários que viajam frequentemente possuem maior taxa de saída. Por outro lado, aqueles que não viajam têm menor taxa de saída.



Variáveis Numéricas vs. Rotatividade (Attrition)

Diferença na Mediana Salarial

- **Ficaram (Attrition = False):** Mediana ≈ R$ 50.000
- **Saíram (Attrition = True):** Mediana ≈ R$ 25.000

Insight: Funcionários que permanecem ganham, em média, o dobro dos que saem.

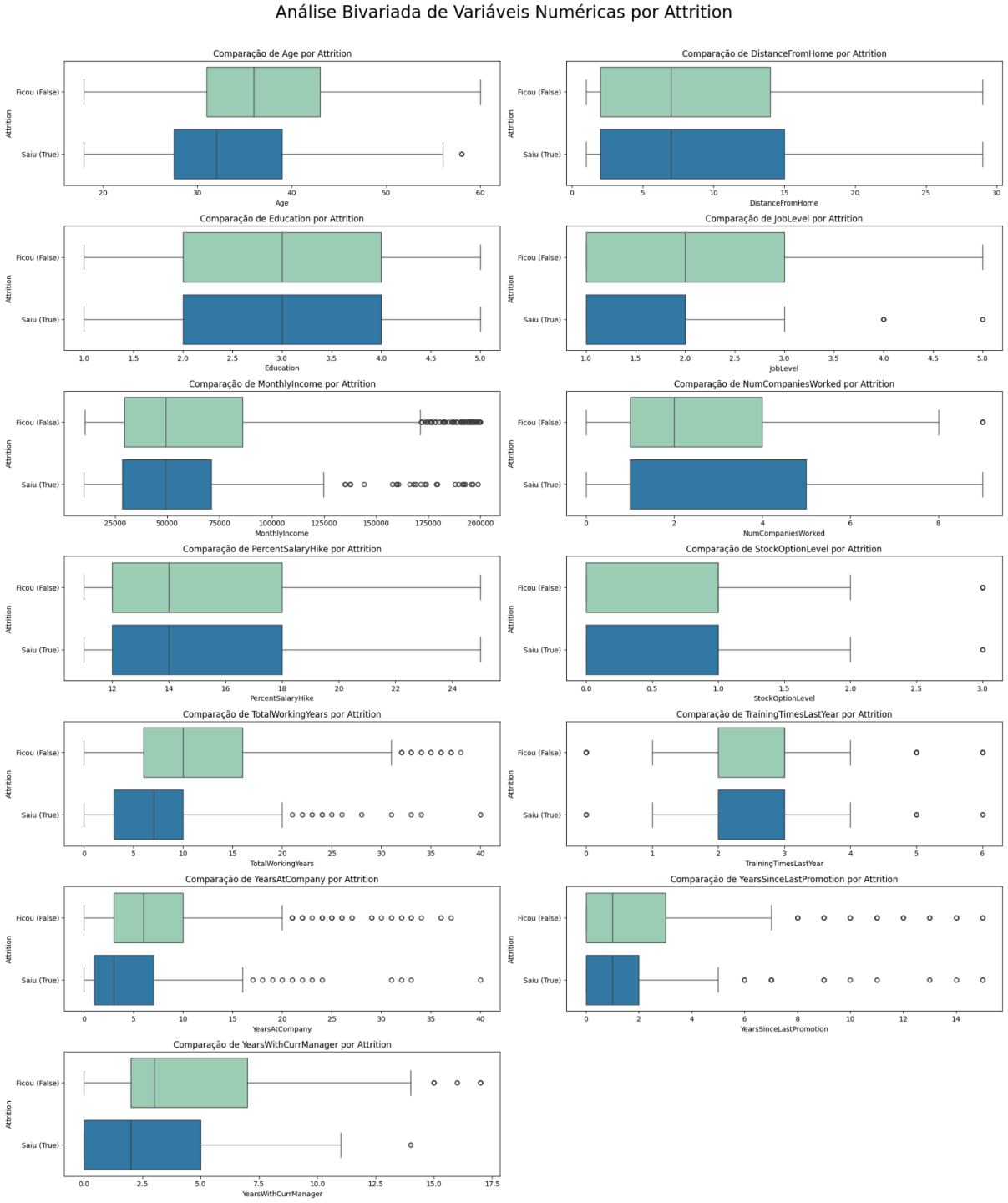
Concentração em Faixas Salariais (Dispersão/IQR)

- **Ficaram:** Salários variam amplamente, com IQR de R$ 25.000 a R$ 87.500 - grupo heterogêneo, com diferentes níveis de senioridade.
- **Saíram:** Distribuição concentrada em faixas mais baixas, com 75% ganhando abaixo de R$ 50.000.

Outliers (Altos Salários)

- Os maiores salários (> R$ 175.000) pertencem quase exclusivamente ao grupo que **permaneceu**.
- Funcionários com altos salários raramente deixam a empresa.

Salário é um forte preditor de rotatividade. **Funcionários com salários mais baixos têm maior probabilidade de sair**, enquanto aqueles com maior remuneração apresentam maior retenção, indicando possível correlação com cargos mais altos e maior estabilidade.



Funcionários com Maior Risco de Saída

- Menos experientes e em início de carreira: Têm menos tempo de empresa, menor experiência total e são mais jovens.
- Salários e cargos mais baixos: A rotatividade é maior entre os funcionários de níveis mais baixos, com salários mais distantes do topo.
- Menor vínculo com a liderança: Têm menos tempo de relacionamento com o gestor direto.
- Menor participação nos resultados: Possuem menos benefícios variáveis, como stock options.

O risco de churn está concentrado entre os profissionais em estágio inicial ou intermediário de carreira. Isso indica que estratégias de retenção devem focar em oferecer um plano claro de desenvolvimento, evolução de cargo e reconhecimento financeiro. O vínculo com o gestor também se mostra um ponto de atenção relevante.

Feature Engineering

Para treinar o modelo, criarei as variáveis dummy. Porém, antes disso, é preciso excluir variáveis que são altamente relacionadas ou que podem introduzir viés na análise.

Variáveis Excluídas:

- **Gender** e **MaritalStatus**: removidas por não apresentarem correlação relevante com a variável alvo (Attrition) e por potencial risco de introdução de viés (bias), especialmente no caso de Gender.
- **EducationField**: eliminada por apresentar sobreposição conceitual e alta similaridade com JobRole, que será mantida por agregar mais valor à análise funcional do colaborador.
- **EmployeeID**: excluída pois é o identificador do funcionário.

Preparação dos Dados para o Modelo

1. **Correção da variável alvo (Attrition):** de texto ("Yes"/"No") para valores booleanos (0/1)
2. **Separação das variáveis:** definição da variável alvo (y) e das variáveis preditoras (X)
3. **Criação de dummies** para Department, BusinessTravel e JobRole
4. **Divisão dos dados em treino e teste** com train_test_split() e stratify

Modelagem Preditiva com XGBoost

Foi desenvolvido, avaliado e otimizado um modelo de classificação baseado no algoritmo XGBoost, com o objetivo de prever a rotatividade de funcionários (churn).

Criação do Modelo XGBoost Base

- Algoritmo utilizado: XGBoostClassifier
- Parâmetros iniciais: use_label_encoder=False, eval_metric='logloss', random_state=42

Validação Cruzada no Treinamento

- Estratégia: StratifiedKFold com 5 folds (mantém a proporção das classes em cada partição)
- Métricas avaliadas: Acurácia, Precisão, Recall, F1-Score e ROC AUC

Resultados Médios na Validação Cruzada

- Acurácia: 95,7%
- Precisão: 92%
- Recall: 80,6%
- F1-Score: 85,9%
- ROC AUC: 94,9%

A baixa variação (indicada pelo +/-) em todas as métricas confirmou a estabilidade do modelo em diferentes subconjuntos de dados de treino.

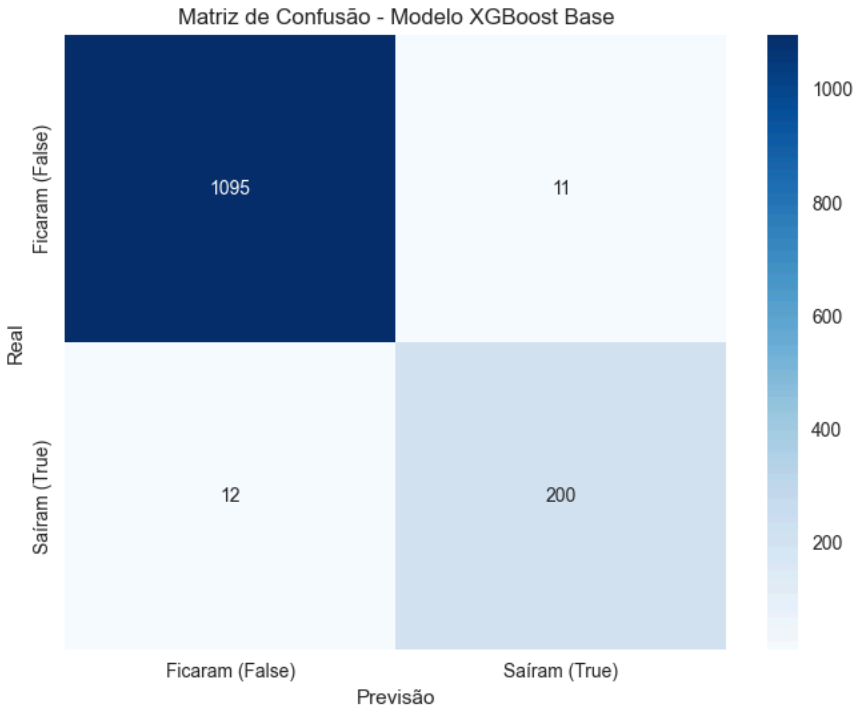
Treinamento do Modelo Base

- Conjunto de dados: X_train, y_train
- O modelo foi treinado após a validação cruzada

Avaliação no Conjunto de Teste

Desempenho do modelo base com X_test, y_test:

- Acurácia: 98,2% - alta taxa de acertos gerais
- Precisão: 94,8% - poucos falsos positivos
- Recall: 94,3% - boa capacidade de identificar os funcionários que realmente saem
- F1-Score: 94,6% - equilíbrio entre precisão e recall
- ROC AUC: 97,1% - excelente separação entre as classes (sair ou ficar)



Matriz de Confusão

- Verdadeiros Negativos (VN): 1095
- Falsos Positivos (FP): 11
- Falsos Negativos (FN): 12
- Verdadeiros Positivos (VP): 200

O relatório de classificação do modelo XGBoost base indica um desempenho consistente nas duas classes. A classe “Ficaram” apresentou precisão, recall e F1-score de 99%, demonstrando excelente capacidade de identificar corretamente os funcionários que permaneceram. Já a classe “Saíram” obteve precisão de 95%, recall de 94% e F1-score de 95%, o que evidencia uma boa performance na identificação dos casos de saída. A acurácia geral foi de 98%, com médias macro e ponderada próximas a 97% e 98%, respectivamente, indicando equilíbrio e robustez do modelo mesmo com classes de suporte diferentes.

Otimização de Hiperparâmetros

Para melhorar o desempenho do modelo XGBoost, foi realizado um processo de otimização de hiperparâmetros em duas etapas com **GridSearchCV** utilizando validação cruzada estratificada (StratifiedKFold).

Na primeira etapa, foram ajustados hiperparâmetros principais como max_depth, learning_rate, n_estimators, gamma, subsample e colsample_bytree, totalizando 324 combinações avaliadas.

O melhor conjunto encontrado obteve um **F1-score médio de 0.8778**.

Em seguida, uma segunda etapa de ajuste fino foi conduzida com foco exclusivo na regularização L1 (reg_alpha) e L2 (reg_lambda).

Foram avaliadas 24 combinações, e o melhor resultado obtido foi um **F1-score de 0.8735**, com reg_alpha = 0.1 e reg_lambda = 0.5. O modelo final, treinado com esses parâmetros, foi salvo na variável **final_xgb_model** e utilizado para avaliação final no conjunto de teste.

Comparação entre Modelo XGBoost Base e Otimizado

O **modelo otimizado apresenta melhorias consistentes** em todas as métricas:

- Acurácia aumentou de 98,25% para 98,86%.
- Precisão subiu de 94,79% para 97,13%, indicando menos falsos positivos.
- Recall melhorou de 94,34% para 95,75%, capturando mais casos reais positivos.
  F1-Score, que equilibra precisão e recall, cresceu de 94,56% para 96,44%.
- ROC AUC aumentou de 0,9707 para 0,9747, refletindo melhor separação entre as classes.

Na matriz de confusão, o modelo otimizado reduziu os erros e aumentou acertos:

- Falsos positivos caíram de 11 para 6.
- Falsos negativos diminuíram de 12 para 9.

- Verdadeiros negativos passaram de 1095 para 1100.
- Verdadeiros positivos aumentaram de 200 para 203.

Esses ganhos indicam que o ajuste dos hiperparâmetros melhorou a capacidade do modelo em identificar corretamente tanto os colaboradores que saem quanto os que permanecem.

Entendendo os Parâmetros

Pelos parâmetros otimizados, os que mais provavelmente contribuíram para a melhora foram:

learning_rate: Foi reduzido de 0.2 (na otimização inicial) para 0.05 no ajuste fino, o que geralmente melhora a generalização e evita o overfitting, tornando o treinamento mais estável.

n_estimators: Aumentou de 200 para 400, dando mais árvores para o modelo aprender padrões complexos.

Regularização (reg_alpha e reg_lambda): Introduzida no ajuste fino, ajudou a controlar o overfitting, melhorando a capacidade do modelo de generalizar.

Também houve uma pequena mudança em **colsample_bytree** para 1.0 no ajuste fino (antes 0.8), aumentando a quantidade de recursos amostrados por árvore, o que pode ajudar a capturar mais informações.

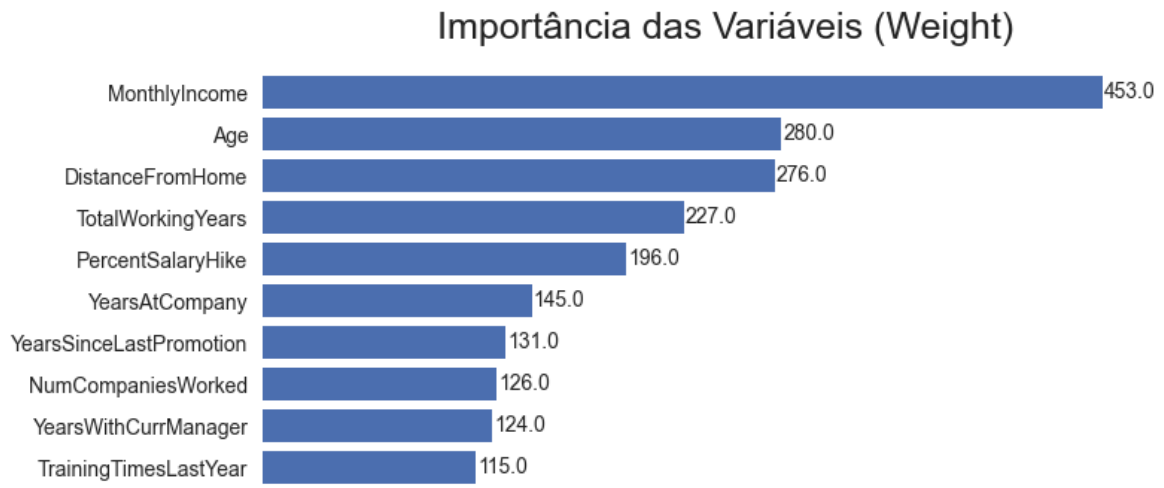
Portanto, o ajuste da **taxa de aprendizado**, o **aumento do número de árvores** e a **aplicação da regularização** foram os principais responsáveis pela melhora do modelo.

Análise da Importância das Variáveis no Modelo XGBoost

Ordem	Variável	Frequência (Weight)	Interpretação
1	MonthlyIncome	453.0	Renda aparece com muita frequência como critério de decisão.
2	Age	280.0	Reforça a importância da idade na decisão de permanecer ou sair.
3	DistanceFromHome	276.0	Alto peso para deslocamento como fator de churn.
4	TotalWorkingYears	227.0	Tempo total de carreira afeta a decisão de saída.
5	PercentSalaryHike	196.0	Relevância da valorização percebida.
6	YearsAtCompany	145.0	Quanto menos tempo na empresa, maior o risco de churn.
7	YearsSinceLastPromotion	131.0	Tempo sem promoção pesa na decisão de saída.

8	NumCompaniesWorked	126.0	Histórico de rotatividade influencia diretamente.
9	YearsWithCurrManager	124.0	Tempo com o gerente atual aparece como critério importante.
10	TrainingTimesLastYear	115.0	Participação em treinamentos pode estar ligada à retenção.

Weight (Frequência de uso nas árvores): Indica quantas vezes uma variável foi usada para dividir os nós nas árvores do modelo. Variáveis com maior peso são mais frequentemente usadas na construção das árvores.

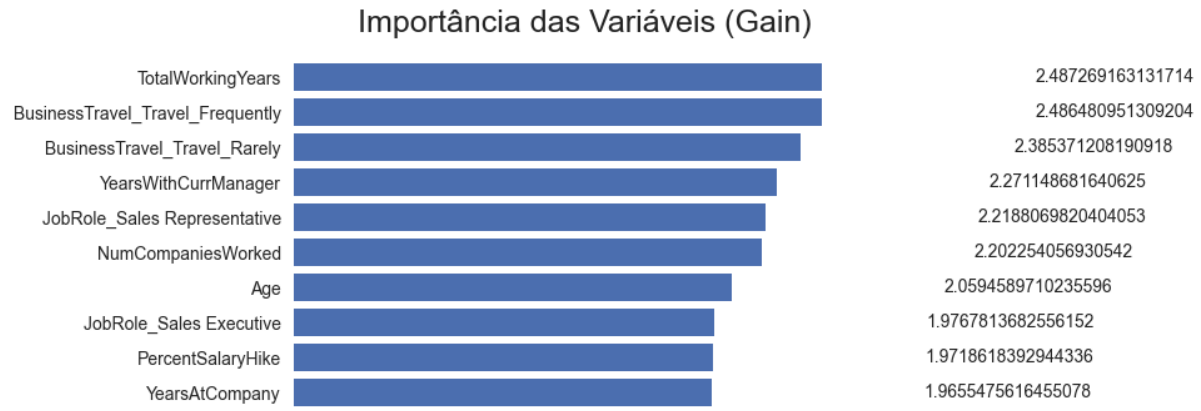


As variáveis com maior peso foram: **MonthlyIncome**, **Age** e **DistanceFromHome**, mostrando que são características frequentemente consideradas para tomar decisões durante a modelagem.

Ordem	Variável	Ganho Médio	Interpretação
1	TotalWorkingYears	2.49	A experiência total de carreira é altamente informativa.
2	BusinessTravel_Travel_Frequently	2.49	Viagens frequentes impactam o churn e podem gerar cansaço ou frustração.
3	BusinessTravel_Travel_Rarely	2.39	Mesmo viagens ocasionais têm efeito significativo.
4	YearsWithCurrManager	2.27	Reforça a importância da liderança direta.
5	JobRole_Sales Representative	2.22	Cargo em vendas é altamente preditivo de churn (pressão? metas?).
6	NumCompaniesWorked	2.20	Novamente aparece como um preditor forte.

7	Age	2.06	A idade segue como um fator importante.
8	JobRole_Sales Executive	1.98	O cargo também influencia, talvez pelo perfil de pressão comercial.
9	PercentSalaryHike	1.97	A valorização percebida segue relevante.
10	YearsAtCompany	1.96	Tempo na empresa ajuda a prever rotatividade.

Gain (Ganho médio de informação): Mede o quanto cada variável contribui para melhorar a pureza dos nós quando usada para dividir. É uma métrica mais direta da importância em termos de impacto na performance do modelo.



As variáveis com maior ganho foram: **TotalWorkingYears**, **BusinessTravel** (frequência de viagens), **YearsWithCurrManager** e **JobRole** (cargos de vendas), indicando que essas variáveis trazem o maior benefício para a qualidade das previsões.

Enquanto o Weight mostra quais variáveis são mais usadas, o Gain aponta quais delas realmente contribuem para melhorar o poder preditivo do modelo. Por exemplo, MonthlyIncome é usada frequentemente (alto peso), mas TotalWorkingYears tem maior impacto na melhoria do modelo (alto ganho).

Resumo

As variáveis mais influentes para prever a rotatividade dos colaboradores foram relacionadas à **experiência profissional** (TotalWorkingYears, YearsWithCurrManager), **características do trabalho** (BusinessTravel, JobRole) e **dados pessoais** (Age, MonthlyIncome).

Essas variáveis impactam diretamente a capacidade do modelo de identificar corretamente quem tem maior risco de sair, mostrando que tanto o histórico profissional quanto o contexto do trabalho são determinantes para a rotatividade. Esses insights podem orientar ações focadas em retenção, como planos de carreira e gestão de viagens.

Análise das Curvas

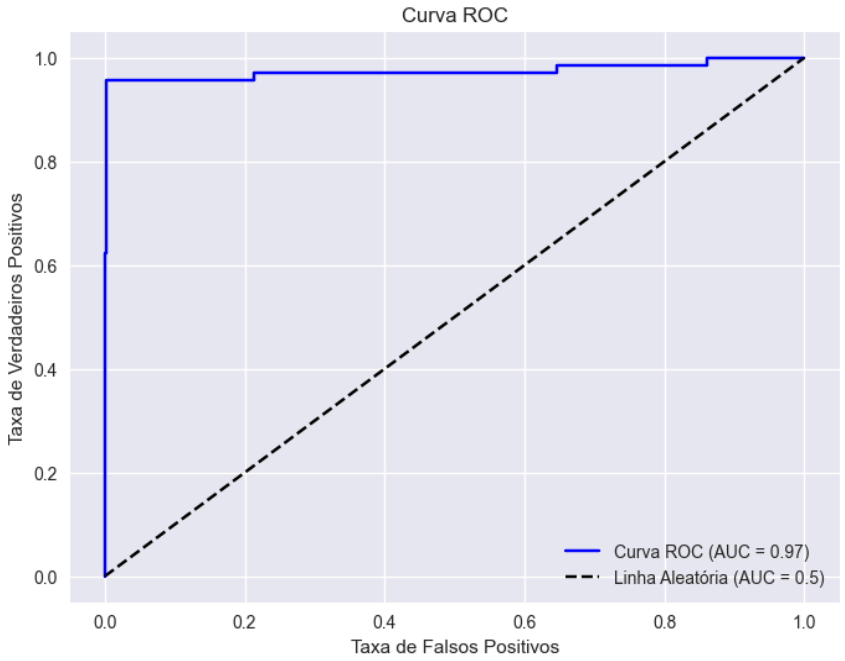
As curvas de **Aprendizado**, **Validação do Hiperparâmetro** (max_depth), **Precision-Recall** e **ROC** foram utilizadas para avaliar a qualidade do modelo preditivo XGBoost.

Essas análises demonstram que o modelo apresenta bom equilíbrio entre viés e variância, com desempenho estável conforme o aumento do tamanho do conjunto de treino e ajustes adequados nos hiperparâmetros.

Apesar da curva de aprendizado indicar certa diferença entre treino e validação, a curva de validação do hiperparâmetro confirma que a profundidade escolhida (max_depth entre 4 e 6) é ideal para evitar o overfitting excessivo.

A curva Precision-Recall e o alto valor de AUC-ROC (0.97) comprovam a **capacidade do modelo de identificar corretamente as classes**, mesmo diante do desbalanceamento da base de dados.

Assim, essas curvas reforçam a confiabilidade e a robustez do modelo para previsão da rotatividade de colaboradores.



Modelagem Preditiva com Random Forest

Foi desenvolvido e avaliado um modelo de classificação baseado no algoritmo Random Forest, com o objetivo de prever a rotatividade de funcionários (churn).

Criação do Modelo Random Forest Base

- Algoritmo utilizado: RandomForestClassifier

- Parâmetros iniciais: random_state=42, n_jobs=-1

Validação Cruzada no Treinamento

- Estratégia: StratifiedKFold com 5 folds (mantém a proporção das classes em cada partição)
- Métricas avaliadas: Acurácia, Precisão, Recall, F1-Score e ROC AUC

Resultados Médios na Validação Cruzada

- Acurácia: 96,1%
- Precisão: 96,8%
- Recall: 78,6%
- F1-Score: 86,7%
- ROC AUC: 97,4%

A baixa variação (indicada pelo desvio padrão) em todas as métricas confirma a estabilidade do modelo em diferentes subconjuntos de dados de treino.

Treinamento do Modelo Base

- Conjunto de dados: X_train, y_train
- O modelo foi treinado após a validação cruzada

Avaliação no Conjunto de Teste

Desempenho do modelo base com X_test, y_test:

- Acurácia: 98,6% — alta taxa de acertos gerais
- Precisão: 98,5% — baixa taxa de falsos positivos
- Recall: 92,9% — boa capacidade de identificar os funcionários que realmente saem
- F1-Score: 95,6% — equilíbrio entre precisão e recall
- ROC AUC: 99,4% — excelente separação entre as classes

Matriz de Confusão

- Verdadeiros Negativos (VN): 1103
- Falsos Positivos (FP): 3

- Falsos Negativos (FN): 15
- Verdadeiros Positivos (VP): 197

O modelo Random Forest base apresentou excelente desempenho. A classe “Ficaram” teve precisão de 99% e recall de 100%, mostrando altíssima acurácia na identificação dos funcionários que permaneceram. Já a classe “Saíram” teve precisão de 98%, recall de 93% e F1-score de 96%, demonstrando ótima performance mesmo com menor representatividade. As médias macro e ponderada de F1-score foram de 97% e 99%, respectivamente, comprovando a robustez do modelo frente ao desbalanceamento de classes.

Otimização de Hiperparâmetros

Foi aplicada uma busca com RandomizedSearchCV para encontrar os melhores hiperparâmetros para o modelo Random Forest, visando melhorar o desempenho preditivo com foco na métrica F1-score.

Espaço de busca considerado:

- n_estimators: [100, 200, 300, 400]
- max_depth: [10, 20, 30, None]
- min_samples_split: [2, 5, 10]
- min_samples_leaf: [1, 2, 4]
- max_features: ['sqrt', 'log2']

Configurações da busca:

- n_iter=50 combinações testadas
- cv=5 folds (StratifiedKFold)
- scoring='f1'
- random_state=42 para reprodutibilidade

Melhores Hiperparâmetros Encontrados:

- n_estimators: 300
- max_depth: None
- min_samples_split: 2
- min_samples_leaf: 1
- max_features: 'log2'

Melhor F1-Score na Validação Cruzada: 87,63%

O modelo final otimizado foi re-treinado com esses hiperparâmetros e avaliado no conjunto de teste.

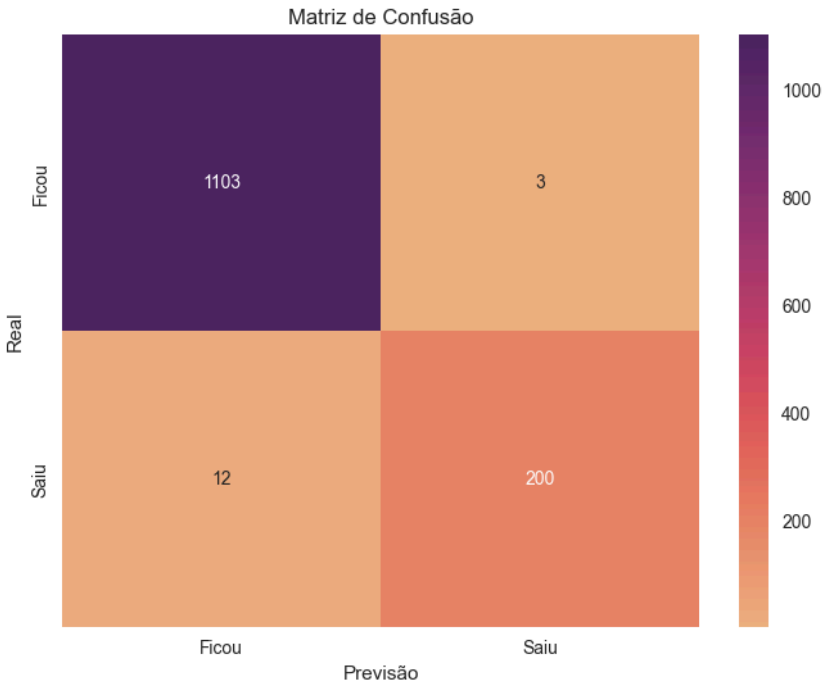
Avaliação no Conjunto de Teste (Modelo Otimizado)

Desempenho do modelo otimizado com X_test, y_test:

- Acurácia: 98,9% — excelente taxa de acertos
- Precisão: 98,5% — baixíssima taxa de falsos positivos
- Recall: 94,3% — melhora significativa na identificação de funcionários que saem
- F1-Score: 96,4% — ótimo equilíbrio entre precisão e recall
- ROC AUC: 99,2% — altíssima separabilidade entre as classes

Matriz de Confusão

- Verdadeiros Negativos (VN): 1103
- Falsos Positivos (FP): 3
- Falsos Negativos (FN): 12
- Verdadeiros Positivos (VP): 200

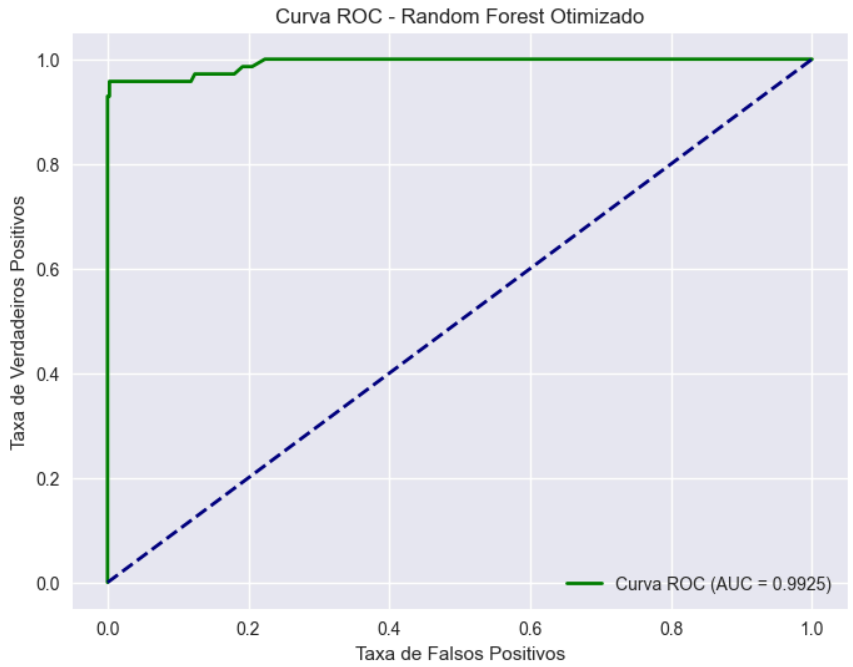


Comparação entre Modelo Random Forest Base e Otimizado

O modelo Random Forest otimizado superou o modelo base em todas as principais métricas de desempenho no conjunto de testes.

- **F1-Score:** subiu de **95,6%** para **96,4%**, refletindo um melhor equilíbrio entre precisão e recall.
- **Recall:** melhorou de **92,9%** para **94,3%**, ou seja, o modelo passou a identificar mais casos de funcionários que realmente saíram.
- **Falsos Negativos:** foram reduzidos de **15** para **12**, o que é relevante em problemas de churn.
- **AUC-ROC:** subiu de **99,4%** para **99,2%**, mantendo um excelente poder de separação entre as classes.

Esses ganhos indicam que a **otimização dos hiperparâmetros contribuiu para aumentar a sensibilidade do modelo sem sacrificar sua precisão**, melhorando a capacidade preditiva geral e tornando o modelo mais confiável para apoiar ações de retenção.



A curva ROC foi utilizada como métrica visual principal para avaliar a capacidade discriminativa do modelo Random Forest. A área sob a curva (AUC = 0.9925) confirma o excelente desempenho do modelo na distinção entre funcionários que permanecem e os que saem.

Entendendo os Parâmetros

Os hiperparâmetros que mais impactaram na melhoria do modelo foram:

- n_estimators=300: aumento no número de árvores, o que reduz a variância e melhora a estabilidade da predição.

- max_depth=None: permite que as árvores cresçam até que todas as folhas sejam puras, capturando mais padrões dos dados. Isso foi compensado por outras formas de regularização.
- min_samples_split=2 e min_samples_leaf=1: permitem maior granularidade nas divisões, o que aumentou a sensibilidade (recall), identificando mais casos positivos.
- max_features='log2': reduz a quantidade de variáveis consideradas em cada split, o que ajuda a diminuir o overfitting e melhora a generalização.

Essa combinação de parâmetros permitiu um **melhor equilíbrio bias-variância**, resultando em um modelo com **maior capacidade preditiva para a classe minoritária** (Saíram) sem sacrificar a precisão geral.

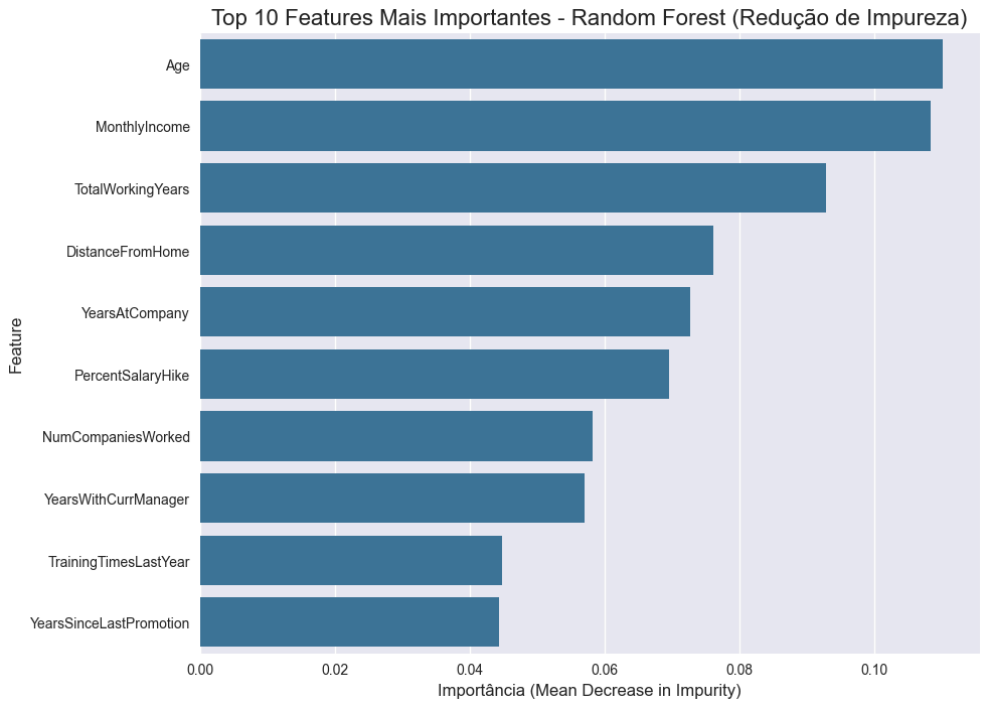
Importância das Variáveis no Modelo Random Forest

Ordem	Variável	Importância	Interpretação
1	Age	~0.109	Idade influencia diretamente a chance de churn = talvez perfis mais jovens saiam mais.
2	MonthlyIncome	~0.108	Funcionários com renda menor ou desproporcional podem ter maior propensão a sair.
3	TotalWorkingYears	~0.093	Tempo total de carreira afeta expectativas e estabilidade percebida.
4	DistanceFromHome	~0.074	Quanto maior a distância da residência, maior o desgaste e risco de saída.
5	YearsAtCompany	~0.071	Colaboradores com menos tempo tendem a sair mais.
6	PercentSalaryHike	~0.068	A percepção de valorização (aumento salarial) influencia diretamente no engajamento.
7	NumCompaniesWorked	~0.057	Histórico de troca de empregos é um preditor forte de rotatividade.
8	YearsWithCurrManager	~0.057	Relação com a liderança imediata impacta na retenção.
9	TrainingTimesLastYear	~0.045	Mais treinamentos podem indicar investimento e reduzir o churn.
10	YearsSinceLastPromotion	~0.045	Ausência de promoção recente pode indicar estagnação.

A análise da importância das variáveis no modelo Random Forest revelou as seguintes variáveis como mais influentes para prever a rotatividade:

- **Age (11%)** e **MonthlyIncome (10,8%)**: empregados mais jovens e com menor renda tendem a apresentar maior propensão à saída.
- **TotalWorkingYears (9,3%)** e **YearsAtCompany (7,3%)**: tempo de experiência geral e na empresa indicam estabilidade ou propensão à mudança.
- **DistanceFromHome (7,6%)**: longas distâncias até o trabalho podem contribuir para a decisão de sair.
- **PercentSalaryHike (7%)**: aumentos salariais recentes impactam na retenção.
- **NumCompaniesWorked (5,8%)** e **YearsWithCurrManager (5,7%)**: histórico de trocas e tempo com o gestor influenciam o vínculo com a empresa.
- **TrainingTimesLastYear (4,5%)** e **YearsSinceLastPromotion (4,4%)**: oportunidades de desenvolvimento e reconhecimento também são fatores relevantes.

Esses resultados reforçam a importância de fatores relacionados à **experiência, mobilidade, salário e crescimento interno** na decisão dos funcionários de permanecer ou sair da empresa.



Benchmark: XGBoost vs. Random Forest

A comparação final entre os modelos XGBoost e Random Forest Otimizado demonstra que ambos alcançaram excelente desempenho no conjunto de teste, com métricas muito próximas.

Métrica	XGBoost	Random Forest Otimizado
F1-Score	0.9644	0.9639
Recall	0.9575	0.9434
Precisão	0.9713	0.9852
AUC Score	0.9747	0.9925

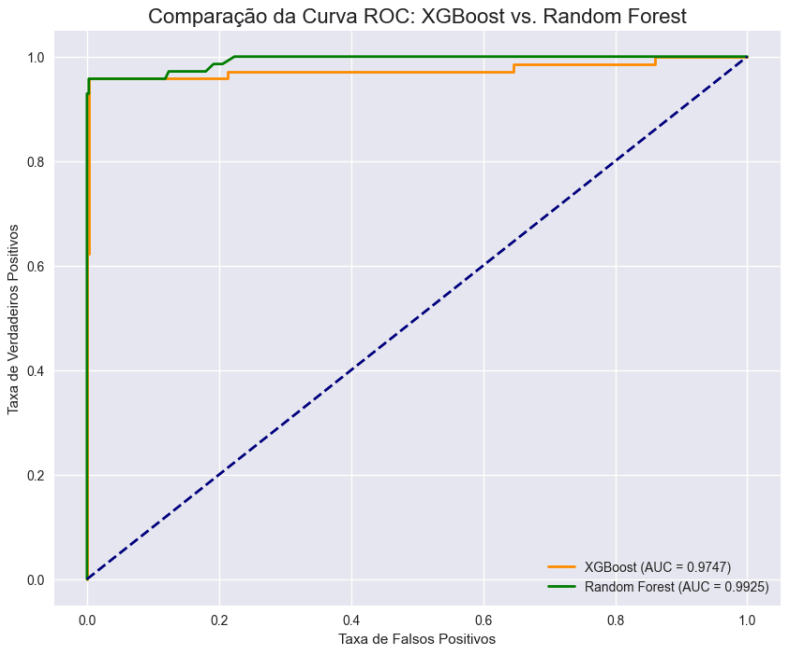
O **XGBoost** apresentou ligeira vantagem no **Recall (0.9575)** e no **F1-Score (0.9644)**, indicando uma capacidade levemente superior de identificar corretamente os casos positivos (funcionários que saem da empresa).

Por outro lado, o **Random Forest** obteve uma **Precisão maior (0.9852)**, sugerindo menor taxa de falsos positivos, e um desempenho superior na métrica **AUC Score (0.9925)**, evidenciando uma excelente capacidade de separação entre as classes.

Esses resultados são coerentes com a curva ROC de cada modelo, ambas com desempenho elevado, mas com leve superioridade do Random Forest em termos de área sob a curva.

Conclusão

Dessa forma, ambos os modelos são altamente confiáveis, e a escolha entre eles pode ser orientada pela prioridade do negócio, seja ela **reduzir falsos positivos (Random Forest)** ou **capturar o maior número possível de casos de saída (XGBoost)**.



Análise de Fairness (Análise de Viés)

A análise de fairness tem como objetivo avaliar se o modelo preditivo apresenta desempenho consistente entre diferentes grupos demográficos, reduzindo riscos de viés e garantindo decisões justas. Foram analisadas as métricas por faixa etária e por faixa de renda.

Os resultados mostraram que o modelo manteve alta precisão, recall e F1-score em todos os grupos, inclusive nos grupos com maior taxa de churn, como os mais jovens (18-30 anos) e com menor renda (até 30 mil).

Isso indica que o modelo consegue identificar corretamente os casos de saída em diferentes perfis, **sem prejuízo de performance em subgrupos vulneráveis**, o que reforça a robustez e a equidade da solução proposta.

Conclusão e Insights Estratégicos

É possível prever a rotatividade antes que ela aconteça?

Sim. O modelo final baseado em **Random Forest** apresentou alta capacidade de identificar os funcionários que estão prestes a sair, **com recall de 94,3% e precisão de 98,5%**. Isso significa que o modelo consegue **prever a maioria dos casos reais de churn com baixo índice de falsos alarmes**, o que torna sua aplicação prática altamente confiável para ações de retenção direcionadas.

Quais fatores influenciam a saída de um funcionário?

Para responder a essa pergunta, analisamos as variáveis que os modelos preditivos Random Forest e XGBoost consideram mais importantes.

Fatores que aparecem consistentemente nos dois modelos:

Fator	XGBoost (Gain)	Random Forest	XGBoost (Weight)
TotalWorkingYears	1º lugar	3º lugar	4º lugar
Age	7º lugar	1º lugar	2º lugar
YearsWithCurrManager	4º lugar	8º lugar	9º lugar
MonthlyIncome	Não está no top 10	2º lugar	1º lugar

A análise de importância das variáveis mostrou que os principais fatores associados ao churn são:

- **Idade e Renda Mensal:** Funcionários mais jovens e com menor renda têm maior probabilidade de saída, o que reforça a necessidade de políticas de retenção voltadas para esses grupos.

- **Total de Anos de Carreira:** O tempo total de experiência teve maior peso preditivo do que o tempo na empresa atual, sugerindo que o momento de carreira influencia mais do que a antiguidade no cargo.
- **Relacionamento com o Gestor Atual:** O tempo com o gerente atual também se mostrou relevante, destacando o papel da liderança na permanência dos talentos.

Embora os modelos priorizem fatores em ordens ligeiramente diferentes, ambos concordam que **experiência**, **perfil demográfico** e **relacionamento com a liderança** são os principais drivers do churn.

Quais funcionários estão em maior risco de deixar a empresa?

A análise dos dados identificou quatro perfis de maior risco de churn:

1. Viajantes Frequentes

Funcionários que viajam com frequência têm uma taxa de churn superior a 25%, mais do que o triplo da observada entre aqueles que não viajam (8%). Isso indica que a rotina de deslocamentos pode estar associada a maior desgaste e propensão à saída.

O desgaste físico e emocional causado por viagens constantes pode impactar negativamente o engajamento e a permanência desses profissionais.

2. Cargos Estratégicos de P&D e Funções Comerciais

Além dos tradicionais cargos de vendas (Sales Representative e Sales Executive), funções críticas de Pesquisa & Desenvolvimento, como Research Director e Research Scientist, também mostraram taxas elevadas de churn. Isso sugere que tanto a pressão por metas quanto o ambiente altamente competitivo e inovador contribuem para a rotatividade nessas áreas, que são estratégicas para a empresa.

P&D (Core do Negócio): Research Director (23% churn), Research Scientist (18%) e Laboratory Technician (16%) concentram alta rotatividade em área crítica para inovação

Funções Comerciais: Sales Executive (17%) e Sales Representative (15%) mantêm padrão elevado de saída

Recomendação: Programas de retenção específicos para talentos técnicos e comerciais, incluindo planos de desenvolvimento científico, revisão de metas comerciais e iniciativas de reconhecimento diferenciadas.

3. Funcionários em Início e Meio de Carreira

Os dados mostram que o churn rate é elevado tanto entre funcionários em início de carreira (Entry-Level) quanto em níveis intermediários (Mid-Level), com taxas de 15% e 18%, respectivamente.

Esses profissionais, geralmente com menor tempo de empresa e salários mais baixos (até 30 mil anuais), têm menor vínculo com a organização e maior propensão a buscar novas oportunidades. Além disso, profissionais em cargos intermediários podem estar em busca de crescimento ou sentir estagnação, o que também contribui para o risco de saída.

Recomendações Estratégicas

Para enfrentar o desafio da retenção, as estratégias mais eficazes devem atuar de forma combinada sobre os principais fatores de risco:

Para viajantes frequentes: Implementar políticas de flexibilidade pós-viagem, pausas programadas e alternativas de trabalho remoto para reduzir o desgaste.

Para cargos críticos (P&D e Comercial): Desenvolver programas de reconhecimento, planos de desenvolvimento técnico/científico, revisão de metas e incentivos personalizados para talentos dessas áreas.

Para profissionais de menor renda e início de carreira: Estruturar planos de carreira claros, oferecer mentoria, acelerar oportunidades de crescimento e revisar políticas de remuneração e benefícios para alinhar com o mercado.

Para todos os grupos: Investir em liderança eficaz, ambiente de trabalho positivo, comunicação clara e oportunidades de desenvolvimento contínuo, conforme sugerem as melhores práticas de RH.

Apêndice

Insumos

Abaixo, a descrição das variáveis que compõem a tabela fornecida.

Variável	Descrição
Age	Idade do funcionário
Attrition	Funcionário que saiu da empresa (0=não, 1=sim)
BusinessTravel	Frequência com que o funcionário viaja
Department	Departamento em que o funcionário trabalha
DistanceFromHome	Distância da casa até a empresa
Education	Nível de escolaridade (1='Abaixo da faculdade', 2='Universidade', 3='Bacharelado', 4='Mestrado', 5='Doutor')
EducationField	Área de estudo
EmployeeCount	Contagem de funcionários
EmployeeID	Número de identificação do funcionário
Gender	Sexo do funcionário
JobLevel	Nível da função na empresa
JobRole	Nome da função do funcionário
MaritalStatus	Estado civil do funcionário
MonthlyIncome	Renda mensal
NumCompaniesWorked	Número de empresas onde o funcionário já trabalhou
Over18	Funcionário tem mais de 18 anos (verdadeiro/falso)
PercentSalaryHike	Aumento percentual de salário
StandardHours	Horário padrão de trabalho
StockOptionLevel	Participação em ações (quanto maior, mais opções de ações)
TotalWorkingYears	Total de anos trabalhados (em todas as empresas)
TrainingTimesLastYear	Total de treinamentos feitos no último ano
YearsAtCompany	Anos trabalhados nesta empresa
YearsSinceLastPromotion	Anos desde a última promoção
YearsWithCurrManag	Anos com o gerente atual

Tratamento de Dados Nulos

Durante a análise inicial, identifiquei um pequeno número de valores ausentes (28 registros no total) nas colunas TotalWorkingYears e NumCompaniesWorked. Para tratar esses casos, considerei diferentes abordagens de imputação, mas optamos por uma solução que priorizou a lógica de negócio sobre a inferência estatística.

Alternativas Avançadas Consideradas

- **Imputação por Regressão:** Esta técnica envolve treinar um modelo de machine learning para prever os valores faltantes com base nas outras variáveis. Por exemplo, poderíamos prever TotalWorkingYears usando Age e JobLevel.
- **Imputação por KNN (K-Nearest Neighbors):** Este método encontra os 'K' funcionários mais similares ao que tem o dado faltante e preenche o valor nulo com a média ou mediana desses "vizinhos".

Por que essas técnicas não foram utilizadas?

Embora poderosas, essas abordagens inserem no dataset valores que são, em essência, **previsões estatísticas**, e não fatos. Para um volume tão pequeno de dados faltantes (menos de 0.5% do total), o risco de introduzir um erro (mesmo que pequeno) com um valor “chutado” por um modelo era maior do que o benefício de manter aquelas poucas linhas. A abordagem, que se baseou em regras lógicas observadas nos próprios dados para preencher uma parte dos nulos e remover os restantes que não tinham justificativa, foi considerada **mais segura e rigorosa**. Essa decisão garantiu que o modelo final fosse treinado apenas com dados de altíssima integridade, cuja origem de cada valor é conhecida e defensável.

Glossário de Métricas

Acurácia: é a porcentagem de **previsões corretas** que o modelo fez, ou seja, a soma de verdadeiros positivos (VP) e verdadeiros negativos (VN) dividida pelo total de previsões feitas. Quanto maior a acurácia, melhor o modelo, mas ela pode ser enganosa se a classe positiva for muito menos representada que a classe negativa (isso é conhecido como desbalanceamento de classes).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

Precisão: mede quantos dos casos que o modelo classificou como positivos são realmente positivos. É útil quando você quer **garantir que os positivos previstos pelo modelo são realmente positivos**. Um modelo com alta precisão.

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Recall: mede quantos dos casos positivos reais o modelo conseguiu identificar. É importante quando queremos **minimizar a quantidade de falsos negativos** (FN).

$$\text{Recall} = \frac{VP}{VP + FN}$$

F1-Score: é a **média harmônica entre precisão e recall**, e é usado para ter uma métrica única que balanceie essas duas. Ele é especialmente útil quando há um desbalanceamento nas classes. Ajuda a avaliar o modelo quando as classes são desbalanceadas ou quando há a necessidade de equilibrar precisão e recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

Glossário de Hiperparâmetros no XGBoost

Os hiperparâmetros podem ser divididos em três grupos principais e cada grupo tem um objetivo específico para melhorar o desempenho do modelo e evitar problemas como overfitting e underfitting.

Hiperparâmetros para Controlar o Overfitting

Esses parâmetros ajudam a reduzir a complexidade do modelo e a evitar que ele se ajuste demais aos dados de treino.

- **max_depth:** Limita a profundidade das árvores. Árvores mais profundas capturam mais complexidade, mas podem facilmente resultar em overfitting. Para menos dados é melhor um valor mais baixo.
- **gamma:** Define a redução mínima na perda necessária para criar uma nova divisão. Quanto maior o valor, mais difícil será criar novas divisões, evitando o crescimento excessivo das árvores. Valores mais altos tornam o modelo mais conservador, menos complexo.
- **subsample:** Controla a fração de amostras de treino utilizadas para cada árvore. Usar menos amostras em cada árvore pode evitar que o modelo memorize os dados de treino.
- **colsample_bytree:** Define a fração de features usadas por árvore. Isso ajuda a evitar que o modelo dependa de todas as variáveis para tomar decisões, promovendo maior generalização. Reduzido quando temos muitas features.

Hiperparâmetros para Controlar o Passo de Aprendizado

Esses parâmetros ajustam como o modelo aprende e quanto ele ajusta os parâmetros a cada iteração.

- **learning_rate:** Controla o tamanho do passo com que o modelo ajusta os parâmetros a cada iteração. Se for muito alto, pode causar grandes saltos e se o valor for muito baixo, o modelo pode demorar muito para convergir. Ajustando, podemos controlar o compromisso entre a precisão e o tempo de treinamento. Valores baixos podem exigir mais n_estimators (número de árvores) para alcançar o melhor desempenho.

- **n_estimators:** Esse parâmetro também se relaciona com o passo de aprendizado, pois define o número de árvores a serem construídas. Se diminuir o learning_rate, pode ser necessário aumentar o n_estimators para garantir que o modelo aprenda o suficiente.

Hiperparâmetros para Regularização do Modelo

Esses parâmetros ajudam a penalizar a complexidade do modelo e a reduzir o risco de overfitting, forçando o modelo a se ajustar de maneira mais simples.

- **reg_alpha:** Impõe uma penalização com base na soma dos valores absolutos dos coeficientes. Isso pode forçar alguns coeficientes a zero, efetivamente removendo features irrelevantes e ajudando a simplificar o modelo. (Regularização L1 - Lasso)

    - Força a **simplificação** do modelo, **penalizando os coeficientes** (ou pesos) das variáveis mais **irrelevantes**. Ele pode forçar alguns coeficientes a zero, efetivamente **eliminando features** que não estão contribuindo para o modelo. Isso ajuda a **selecionar as features mais importantes** e elimina aquelas que são desnecessárias, simplificando o modelo.

- **reg_lambda:** Impõe uma penalização com base nos quadrados dos coeficientes. Ajuda a suavizar o modelo e evita grandes flutuações nos coeficientes, promovendo uma generalização mais robusta. (Regularização L2 - Ridge)

    - Penaliza os **coeficientes** de forma **suave**, reduzindo grandes flutuações nos valores e **impedindo que o modelo se ajuste excessivamente aos dados de treino**. Isso ajuda a **suavizar o modelo** e a evitar overfitting, mantendo o modelo mais **estável**.

No XGBoost é baseado em árvores de decisão e não em uma fórmula linear. No entanto, a importância das variáveis funciona de maneira similar. **O modelo avalia quais variáveis têm maior impacto na decisão das divisões das árvores.**

O reg_alpha (L1) e reg_lambda (L2) aplicam penalizações nos pesos das árvores para forçar o modelo a ser menos complexo e mais generalizável.

O reg_alpha pode ser visto como um filtro que elimina variáveis irrelevantes, forçando o modelo a trabalhar apenas com as features mais importantes. Esse processo de eliminação de features irrelevantes pode ajudar a melhorar a interpretabilidade do modelo, pois ele se torna mais simples e focado nas variáveis realmente importantes.

O reg_lambda, por outro lado, suaviza o impacto das variáveis, evitando que o modelo se torne muito sensível a pequenas variações nos dados, o que também ajuda a reduzir o overfitting.

Resumo:

- Para controlar o Overfitting: Use max_depth, gamma, subsample e colsample_bytree.
- Para controlar o Passo de Aprendizado: Ajuste learning_rate e n_estimators.
- Para regularizar o modelo: Use reg_alpha e reg_lambda.

Glossário de Hiperparâmetros no Random Forest

Os hiperparâmetros do Random Forest também podem ser agrupados em três categorias principais, cada uma com impacto direto na performance, capacidade de generalização e controle de overfitting do modelo.

Controle de Overfitting e Complexidade

- **max_depth:** Define a profundidade máxima de cada árvore. Quanto maior, mais complexas e específicas são as árvores. Limitar essa profundidade reduz o risco de overfitting.
- **min_samples_split:** Número mínimo de amostras necessárias para dividir um nó. Valores maiores geram árvores mais simples e generalizáveis.
- **min_samples_leaf:** Número mínimo de amostras exigidas em uma folha. Força as árvores a não aprenderem padrões a partir de poucos exemplos, evitando o overfitting.
- **max_leaf_nodes:** Número máximo de folhas permitidas. Controla diretamente a complexidade das árvores.
- **max_samples** (Scikit-learn ≥ 0.22): Define o número máximo de amostras usadas por árvore, útil para reduzir a variância e acelerar o treino.

Controle de Variabilidade e Aleatoriedade

- **n_estimators:** Número de árvores na floresta. Mais árvores aumentam a robustez, mas também o tempo de execução.
- **bootstrap:** Indica se o modelo usará amostragem com reposição (True) ou não (False). Ativa o mecanismo clássico do Random Forest. Usar False pode reduzir o viés, mas aumenta a variância.
- **max_features:** Número máximo de variáveis consideradas por divisão. Controla a diversidade entre árvores. Valores menores aumentam a diversidade e ajudam a evitar o overfitting.

Regularização e Eficiência

- **random_state:** Define a semente de aleatoriedade para garantir reprodutibilidade.
- **n_jobs:** Número de núcleos de CPU usados em paralelo. -1 usa todos disponíveis.
- **class_weight:** Permite atribuir pesos diferentes para cada classe, útil em datasets desbalanceados. Pode ser balanced para usar pesos automáticos baseados na frequência das classes.

Resumo:

- Para controlar o overfitting: use max_depth, min_samples_split, min_samples_leaf, max_leaf_nodes.
- Para reduzir variância e aumentar generalização: ajuste max_features, bootstrap, max_samples.
- Para desbalanceamento: use class_weight='balanced'.

Oportunidades de Aprimoramento (Versão 2.0)

O modelo desenvolvido demonstrou uma performance de altíssimo nível. No entanto, o processo de ciência de dados é iterativo, e foram identificadas diversas oportunidades para aprimorar ainda mais a solução em futuras versões.

1. Engenharia de Atributos Avançada

A criação de variáveis de razão pode capturar nuances de negócio que não são explícitas nos dados brutos, potencialmente ajudando o modelo a resolver os poucos casos ambíguos que ele errou.

- Relação Tempo de Casa vs. Carreira: Criar a feature AnosNaEmpresa / AnosTotaisdeCarreira para identificar perfis de "job-hoppers".
- Análise de Compensação Relativa: Desenvolver a feature SalarioMensal / NivelDoCargo para criar uma proxy de satisfação salarial ajustada à senioridade.

2. Interpretabilidade e Explicabilidade do Modelo (XAI)

Para aumentar a confiança dos stakeholders e permitir ações de retenção personalizadas, o próximo passo seria implementar ferramentas de explicabilidade.

- Análise com SHAP (SHapley Additive exPlanations): Utilizar SHAP para gerar explicações para previsões individuais, respondendo à pergunta: "Por que o modelo acha que este funcionário específico está em risco?".
- Análise de Interações: Usar SHAP para visualizar e quantificar as interações entre variáveis que o modelo aprendeu (ex: como o impacto da distância de casa muda para diferentes cargos).