Ficha Técnica – Projeto 06

People Analytics: Previsão de rotatividade com Machine Learning

Ferramentas e tecnologias:

- Google Colab
- Python
- Notion
- Validação de códigos: Chat GPT e Gemini
- Loom
- Streamlit

Equipe:

Vanessa Santana do Amaral

Objetivo do projeto

Este projeto tem como objetivo analisar dados de recursos humanos para desenvolver um **modelo de machine learning supervisionado**, capaz de prever a probabilidade de desligamento de funcionários.

Benefícios esperados:

- Antecipar e mitigar a perda de talentos críticos;
- Apoiar decisões estratégicas de retenção de colaboradores;
- Otimizar processos de gestão de pessoas com base em dados concretos.

Base de dados

Variável	Tipo	Descrição
Age	int	Idade do funcionário
Attrition	object	Indica se o funcionário deixou a empresa (Yes/No)
BusinessTravel	object	Frequência de viagens a trabalho
Department	object	Departamento de atuação
DistanceFromHome	int	Distância da residência até a empresa (km)
Education	int	Nível de escolaridade
EducationField	object	Área de formação
Gender	object	Sexo do funcionário
JobLevel	int	Nível hierárquico
JobRole	object	Cargo desempenhado
MaritalStatus	object	Estado civil
MonthlyIncome	int	Salário mensal
NumCompaniesWorked	float	Número de empresas em que trabalhou
PercentSalaryHike	int	Percentual de aumento salarial
StockOptionLevel	int	Nível de opção de ações
TotalWorkingYears	float	Total de anos de experiência
TrainingTimesLastYear	int	Número de treinamentos no último ano
YearsAtCompany	int	Tempo na empresa
YearsSinceLastPromotion	int	Tempo desde a última promoção
YearsWithCurrManager	int	Tempo com o gestor atual
EmployeeCount	int	Número fixo, apenas 1
EmployeeID	int	Identificador único
Over18	object	Indicação de maioridade
StandardHours	int	Horas padrão de trabalho

Processamento inicial

- Ambiente: Google Colab (`rotaML-RHipynb`)
- Upload da base de dados;
- Importação das bibliotecas principais;

- Verificação do tamanho da base e tipos de dados.

Limpeza e preparação da base

1. **Tipos de dados:**
   - `NumCompaniesWorked` e `TotalWorkingYears` convertidos para int após tratamento de nulos.
2. **Duplicados:** Nenhum identificado.
3. **Valores constantes ou irrelevantes:**
   - Removidas: `EmployeeCount`, `Over18`, `StandardHours`, `EmployeeID`.
   - Removida: `Gender` (potencial discriminatório).
4. **Valores nulos:**
   - `NumCompaniesWorked`: 19 nulos → preenchidos com mediana.
   - `TotalWorkingYears`: 9 nulos → preenchidos com mediana.
5. **Alterações adicionais:**
   - Padronização de variáveis categóricas;
   - Criação de faixas etárias e faixas salariais para análise exploratória.

Análise exploratória (EDA)

1. **Verificação das estatísticas descritivas (média, mediana, mínimo, máximo, contagem, quartil).**
2. **Bloxplot para verificação de outliers:** Os outliers identificados foram mantidos na base de dados, pois refletem padrões comuns em um contexto de RH.
   A maioria dos funcionários está concentrada em cargos, salários e níveis de entrada, apresentando perfis mais juniores. Já os valores extremos observados no MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion e YearsWithCurrManager correspondem a profissionais de maior senioridade - um padrão esperado que se reflete em salário, promoções e tempo de empresa.

Para TrainingTimesLastYear, a concentração está em 2 a 3 treinamentos anuais, mas não é incomum haver funcionários que realizem mais ou menos treinamentos ao longo do ano.

3. **Criação da matriz de correlação para as variáveis numéricas:**

- YearsAtCompany ↔ YearsWithCurrManager (0.77): funcionários tendem a permanecer com o mesmo gestor à medida que aumentam seu tempo de empresa.
- TotalWorkingYears ↔ Age (0.68): esperado, já que a idade avança junto com a experiência profissional.
- TotalWorkingYears ↔ YearsAtCompany (0.63): maior tempo de carreira está associado a mais anos na mesma empresa.
- YearsAtCompany ↔ YearsSinceLastPromotion (0.62): quanto mais tempo na empresa, maior a chance de ter vivenciado uma promoção.
- YearsSinceLastPromotion ↔ YearsWithCurrManager (0.51): permanência com o mesmo gestor tende a aumentar a visibilidade, refletindo em promoções.

4. Comparação da variável Atrittion perante as demais, para entender distribuição através de visualizações e histogramas.
5. Panorama geral:

- 83,9% ativos e 16,1% desligados
- **Faixa etária:**
  Jovens: maior proporção de desligamentos
  Adultos: maior concentração de ativos
  Média de idade: 37 anos (mín: 18, máx: 60)
- **Estado civil:**
  Casados: maior proporção de ativos
  Solteiros: maior proporção de desligamentos
- **Tempo de empresa:**
  Maior saída: 0–2 anos
  +10 anos: minoria, baixa taxa de saída
- **Educação:**
  Maioria dos ativos: Bacharelado (Education 3)
  Menor representatividade: Doutorado

- **Setor:**
  Research & Development: maior número de ativos
  RH: alta rotatividade
  Sales: alta proporção de desligamentos
- **Nível Salarial:**
  Equilíbrio entre entradas e saídas em todas as faixas

Feature engineering

Feature engineering consiste em criar ou transformar variáveis para melhorar o desempenho do modelo.

1. Exclusão de variáveis não relevantes:
   - `EducationField`, `EmployeeID`, `MaritalStatus`, `Age_Category`, `TempoEmpresa_cat`, `Faixa_Salarial`
2. Definição de variável alvo: `Attrition`
3. Conversão de categóricas em dummies (one-hot encoding)
4. Divisão treino/teste: 80% treino / 20% teste, com `random_state=42` e `stratify=y` para manter proporção da variável alvo

---

Modelos testados

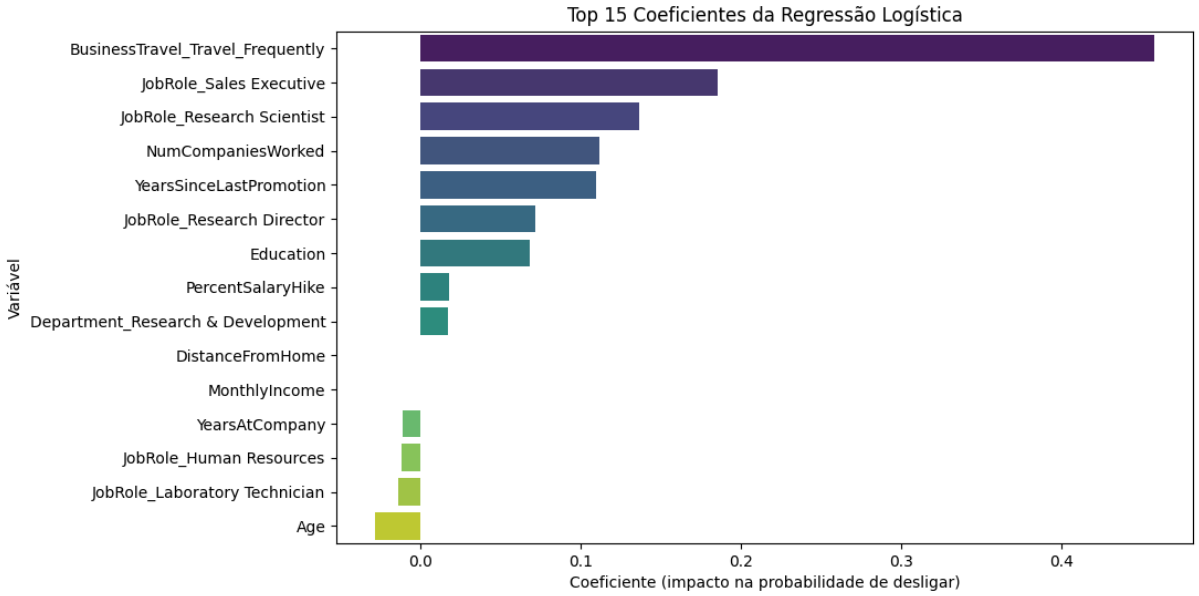
1. Regressão Logística

- Modelo clássico de classificação binária, que estima a probabilidade de saída de funcionários com base nas variáveis preditoras.

Resultados:

- Acurácia: 0.837
- Recall (classe positiva): 0.007
- F1-score (classe positiva): 0.014
- AUC: 0.701

> O modelo captura bem os funcionários ativos, mas falha em prever desligamentos (alta taxa de falso negativo).



2. XGBoost

- Modelo de boosting, combina múltiplas árvores de decisão para melhorar a previsão e reduzir overfitting.

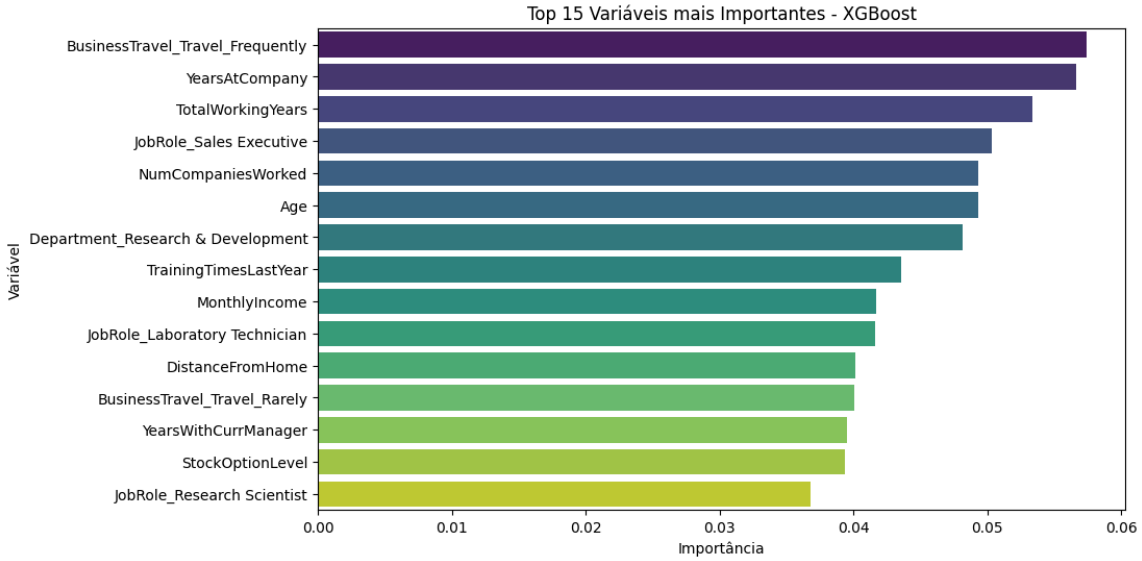
Resultados:

- Acurácia: 0.980
- Recall (classe positiva): 0.873
- F1-score (classe positiva): 0.932
- AUC: 0.996

> O modelo equilibra acurácia e capacidade de identificar desligamentos, sendo mais robusto que regressão logística.

- **Cross-validation:**
  - Média da acurácia: 0.992 ± 0.004

- Média da AUC-ROC: 0.994 ± 0.006



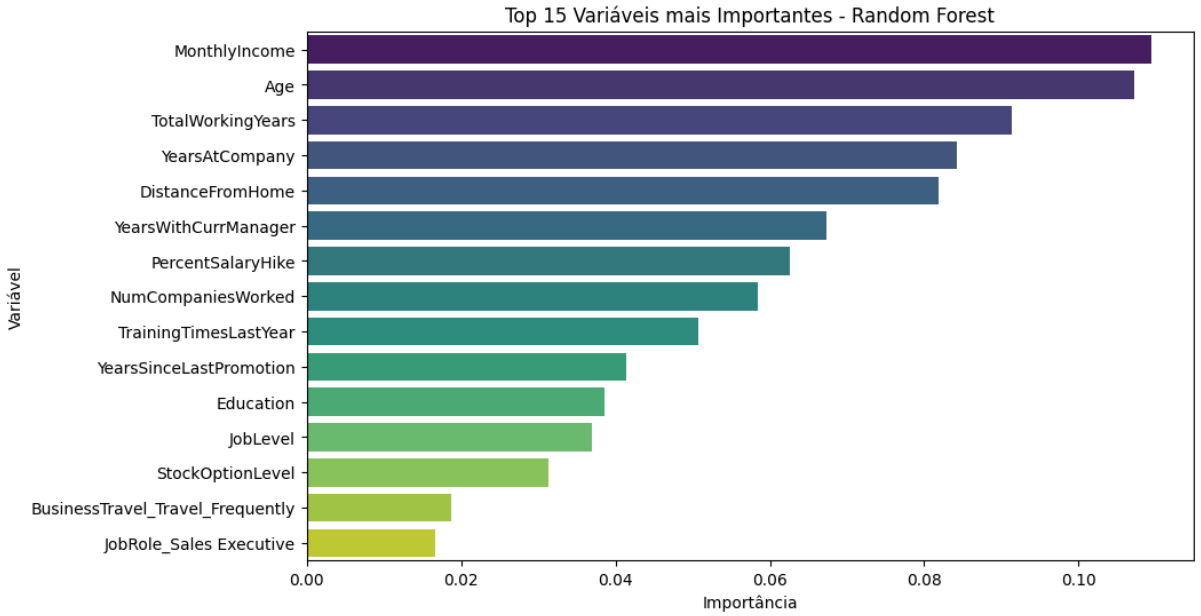
3. Random Forest

- Modelo baseado em múltiplas árvores de decisão, usando amostras aleatórias e agregando resultados (bagging).

Resultados:

- Acurácia: 0.995
- Recall (classe positiva): 0.972
- F1-score (classe positiva): 0.986
- AUC: 0.999

> Melhor desempenho geral, identificando quase todos os casos de desligamento sem sacrificar a precisão para funcionários ativos.



Comparação dos modelos

Modelo	Acurácia	Recall	F1-score	AUC
Regressão Logística	0.837	0.007	0.014	0.701
XGBoost	0.980	0.873	0.932	0.996
Random Forest	0.995	0.972	0.986	0.999

> Conclusão: Random Forest apresenta melhor equilíbrio entre precisão e recall, sendo ideal para prever desligamentos e reduzir perdas estratégicas de RH.

Conclusões do Projeto

- É **viável prever com alto grau de confiabilidade** o churn de colaboradores a partir de dados de RH.
- Modelos tradicionais, como **Regressão Logística**, têm limitações para capturar padrões complexos. Já modelos como **Random Forest** e **XGBoost**, entregam performance superior, com maior capacidade de generalização.

- As variáveis mais relevantes para previsão estão ligadas a:

Deslocamento e viagens: colaboradores com trajetos longos ou que realizam viagens frequentes apresentam maior propensão ao desligamento.

Liderança e tempo de empresa: profissionais com pouco tempo de casa, histórico de mudanças frequentes ou pouca convivência com o gestor atual tendem a ser mais suscetíveis à saída.

Idade e renda: colaboradores mais jovens e com remuneração inferior demonstram maior tendência a deixar a empresa de forma precoce.

Recomendações para RH

1. **Monitoramento proativo**: utilizar o modelo para identificar colaboradores em alto risco de desligamento e priorizar ações preventivas.
2. **Programas de retenção focados**: direcionar esforços para profissionais jovens, recém-contratados (0–2 anos de empresa) e em cargos estratégicos.
3. **Planejamento de carreira estruturado**: alinhar expectativas por meio de planos de carreira claros, promoções transparentes e oportunidades de crescimento interno.
4. **Análise contínua**: atualizar periodicamente o modelo com novos dados para manter sua precisão e relevância.

Estratégias sugeridas

- **Mentoria e desenvolvimento**: programas de capacitação e acompanhamento de carreira voltados para colaboradores em início de jornada.
- **Revisão de salário e benefícios**: garantir remuneração competitiva, bônus, incentivos e políticas de reconhecimento.
- **Programas de engajamento**: promover eventos, iniciativas de pertencimento e reconhecimento contínuo.
- **Integração com gestores**: incentivar check-ins regulares e feedback estruturado em forma de PDI, especialmente para quem tem pouco tempo

sob a mesma liderança.

- **Treinamento de liderança**: capacitar gestores em práticas de retenção, feedback e gestão de equipes.
- **Flexibilidade de jornada**: oferecer home office ou horários flexíveis para colaboradores com longos deslocamentos.
- **Gestão de viagens**: reduzir viagens excessivas e fornecer suporte logístico/financeiro adequado.
- **Benefícios de mobilidade**: subsídios de transporte ou apoio para minimizar impactos do trajeto.

Aplicativo: People Analytics - Previsão de Churn

Com o objetivo de operacionalizar o modelo de Machine Learning e disponibilizar suas análises de forma acessível para o time de RH, foi desenvolvido um aplicativo interativo no Streamlit, utilizando o algoritmo de melhor desempenho identificado: Random Forest.

O modelo foi treinado no Google Colab, a partir das variáveis preditoras selecionadas (como idade, tempo de empresa, promoções, histórico de cargos, renda, entre outras). Após o treinamento, o modelo e o mapeamento das colunas foram exportados e versionados no GitHub, garantindo reprodutibilidade e integração direta com o aplicativo.

No aplicativo, o usuário pode preencher as informações de um colaborador em um formulário simples. A partir disso, o modelo calcula a probabilidade de desligamento e apresenta, de forma visual e intuitiva (gráficos interativos), o nível de risco de saída.

Além do resultado numérico, o sistema fornece ações recomendadas de retenção de acordo com o perfil identificado (revisão de benefícios, plano de carreira, capacitação ou políticas de flexibilidade).

Assim, o aplicativo funciona como uma ferramenta prática de People Analytics, permitindo que gestores e equipes de RH tomem decisões baseadas em dados para reduzir a rotatividade e aumentar o engajamento dos colaboradores.

💡 Para usar o aplicativo, é necessário a versão 3.13 do python.