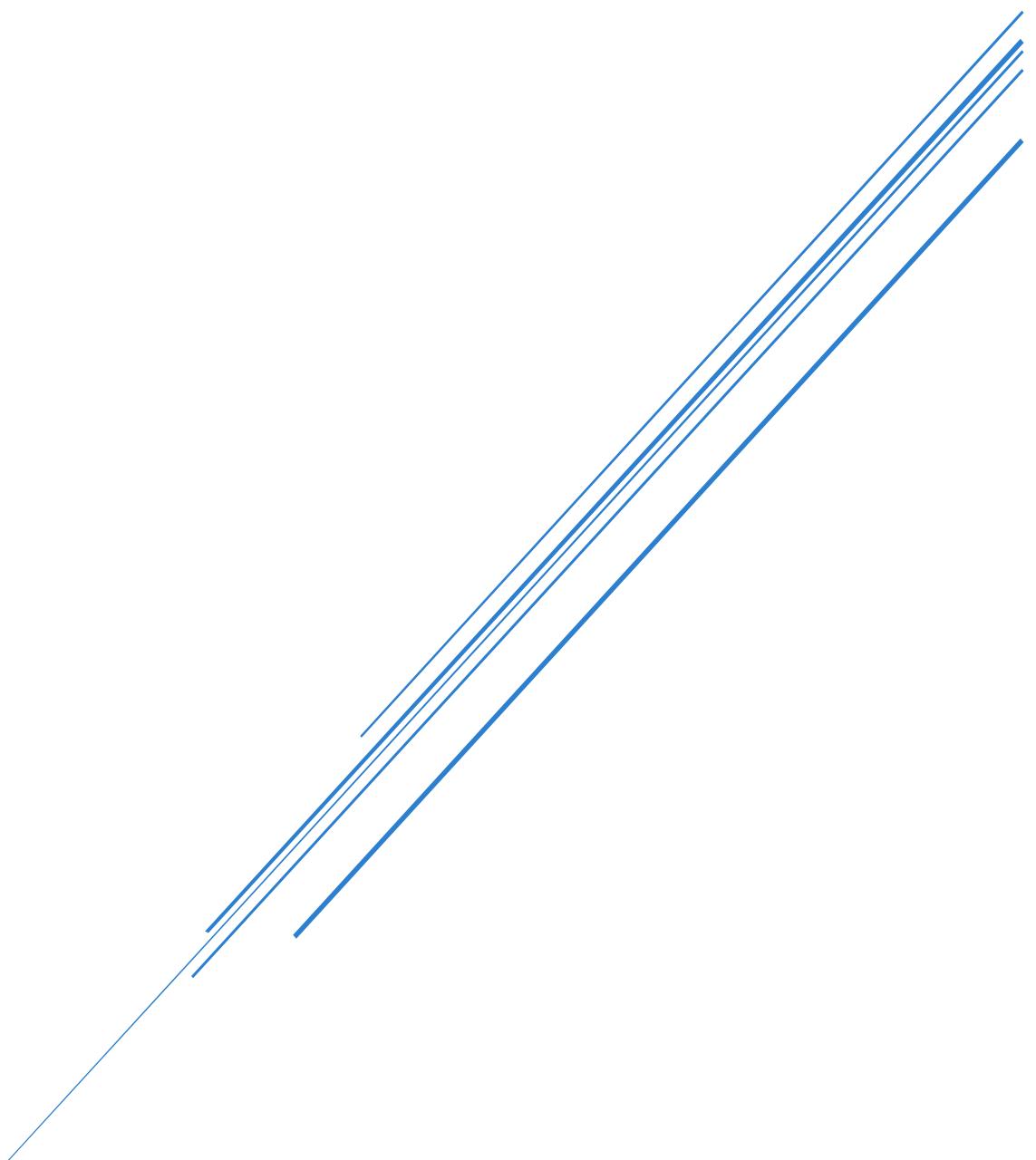


DIAGNÓSTICO PREDITIVO DE TURNOVER

Mudando o cenário de reação para a antecipação
estratégica



Departamento de Gestão de Pessoas
João Pedro – People Analytics

Diagnóstico Preditivo de Turnover: Mudando o cenário de reação para a antecipação estratégica

João Santos

Resumo

A saída voluntária de colaboradores (*Turnover*) representa um dos maiores problemas financeiros e intelectuais para as organizações atuais. Este artigo descreve o desenvolvimento e a implementação de um modelo preditivo de *Machine Learning* focado em antecipar o risco de saída dos funcionários em uma operação corporativa real. Utilizando dados históricos extraídos do sistema transacional de folha de pagamento (FOPAG), aplicamos técnicas de Engenharia de Variáveis (*Feature Engineering*) para traduzir fatores demográficos, temporais e de geolocalização em matrizes preditivas. Como obtivemos uma base naturalmente desbalanceada dos dados de evasão, utilizamos uma técnica de SMOTE (*Synthetic Minority Over-sampling Technique*) combinada ao algoritmo *Random Forest*. O modelo final atingiu uma capacidade de identificação (Recall) de 60% com base na classe minoritária, permitindo a transição da gestão de Recursos Humanos (RH) de um modelo totalmente reativo para uma atuação preventiva-cirúrgica baseada em probabilidades de risco de fuga (Flight Risk Score).

Palavras – Chaves: People Analytics, Turnover, Machine Learning, Random Forest, SMOTE, Retenção de Talentos.

1. Introdução

No atual cenário corporativo, a retenção de talentos ultrapassa a área de apenas clima organizacional, se consolidando como uma métrica crítica de sustentabilidade financeira. De acordo com a *Society for Human Resource Management* (SHRM, 2022), o custo de substituição de um colaborador pode variar entre 50% e 200% do seu salário anual, englobando as despesas de recrutamento, *onboarding* e a curva de aprendizado inerente à nova contratação. No Brasil, cálculos da BGC Brasil apontam que a reposição de uma posição que custa R\$ 120 mil ao ano pode gerar um custo de reposição de R\$ 81,5 mil um custo de reposição de 47%, sem considerar os efeitos em produtividade, adaptação cultural e engajamento de equipe.

Apesar da alta relevância do tema, grande parte dos departamentos de Recursos Humanos (RH) ainda opera sob um paradigma analítico puramente descritivo e operacional, focando em relatórios *post-mortem*, como entrevistas de desligamentos e métricas de turnover histórico (BERSIN, 2021). Embora muito importantes para a compreensão do passado, tais métodos falham em fornecer *insights* açãoáveis em tempo hábil para evitar a saída de talentos-chaves.

Com a criação do People Analytics 3.0, a literatura tem apontado para a necessidade de adoção e criação de modelos preditivos que utilizem dados preexistentes dos colaboradores para calcular a probabilidade futura de atrito (*Attrition Prediction*) (DAVENPORT et al., 2010).

Esse artigo tem como objetivo apresentar a arquitetura, o desenvolvimento e os resultados de um modelo preditivo de *Turnover* aplicado a uma base de dados corporativa real. O estudo demonstra como aplicação de algoritmos não lineares, especificamente o Random Forest, aliados a tratamentos matemáticos de balanceamento de classes (SMOTE), pode identificar padrões ocultos de evasão, como o impacto do tempo de deslocamento geográfico e os ciclos de maturidade do colaborador (“Mortalidade Infantil” organizacional).

2. M&Todologia e Engenharia de Dados

A construção dos modelos preditivos exigiu um rigoroso processo de *Data Preparation* (Preparação de Dados), essencial para garantir que a inteligência artificial modelasse os padrões de negócio, e não ruídos sistêmicos. A extração inicial (*Data Extraction*) foi realizada via consultas SQL diretamente no banco de dados transacional da empresa (FOPAG), garantindo a integridade referencial dos registros (identificados pela chave principal `colaborador_sk`).

2.1. Saneamento de Dados (Data Cleansing)

A base original contava com 91 registros históricos e 22 variáveis iniciais. Foram identificados e tratados valores nulos (*Missing Values*) em colunas vitais. Registros com ausência de informação crítica irrecuperável (ex.: `salario_contratual`) foram removidos (*Listwise Deletion*), resultando em um dataset útil de 89 observações. Variáveis temporais em formato de texto (object) foram convertidas para o padrão *datetime*, permitindo operações

aritméticas. Para evitar o vazamento de dados (*Data Leakage*) e o sobreajuste (Overfitting), a chave primária de identificação dos funcionários foi retirada do conjunto de treinamento.

2.2. Engenharia de Variáveis (*Feature Engineering*)

Para ser possível o aprendizado de máquina (*Machine Learning*), as variáveis brutas sofreram transformações físico-lógicas fundamentais:

- **Maturidade Operacional:** A variável contínua `meses_de_casa` calculada pela diferença entre a `data_admissao` e a `data_demissao` (ou data atual, para colaboradores ativos), tornando-a o principal eixo temporal da análise.
- **Geocodificação por Zonas:** A literatura de geografia urbana aponta o tempo de deslocamento pendular como um dos motivos principais do estresse dos colaboradores. Códigos de Endereçamento Postal (CEP) brutos não possuem valor numérico contínuo para algoritmos baseados em árvores. Como estratégia de aproximação espacial (*Spatial Proxy*), os CEPs foram divididos em seus dois primeiros dígitos, agrupando os colaboradores em macrozonas geográficas mapeadas (ex.: Zonas 70, 71, 72, 73 correspondentes ao Distrito Federal e entorno). Essa transformação categórica foi posteriormente binarizada através da técnica de *One-Hot Encoding*.
- **Aglomeração de Cauda Longa:** Variáveis categóricas com alta cardinalidade e baixa frequência, como “Departamentos” e “Perfis Comportamentais” com menos de 4 representantes, foram aglutinadas em uma categoria unificada (“OUTROS”). Esta técnica reduz o risco da Maldição da Dimensionalidade (*Curse of Dimensionality*) ao restringir a criação excessiva de variáveis *dummies* que não possuem uma significância estatística.

3. Modelagem Preditiva e Estratégia de Validação

A seleção e o treinamento do algoritmo preditivo foram desenhados para mitigar os principais riscos associados a bases de dados empresariais em fases embrionárias: a alta dimensionalidade e o severo desbalanceamento da variável alvo (*Turnover*).

3.1. Seleção do Algoritmo e Redução de Dimensionalidade

Devido à presença de interações não lineares entre as variáveis de RH (ex.: a relação entre idade, departamento e tempo de casa), optou-se pela utilização do algoritmo Random Forest (*Floresta Aleatória*). Por ser um método de Ensemble Learning, ele constrói múltiplas árvores de decisão independentes e agrupa seus resultados, conferindo ao modelo uma resistência superior ao sobreajuste (*Overfitting*).

Inicialmente, a base contava com mais de 30 variáveis após o processo de *One-Hot Encoding*. No entanto, respeitando a lógica estatística de Events Per Variable (EPV) para amostras limitadas, seguimos com uma agressiva Engenharia de Seleção de Variáveis (*Feature Selection*). O modelo foi forçado a focar estritamente nas variáveis de maior relevância técnica e de negócios (Tempo de Casa, Salário, Idade, Dependentes, Departamento de Relacionamento e Ausência de Perfil Comportamental). Essa redução de dimensionalidade foi vital para evitar que o algoritmo dispersasse seu poder de generalização em ruídos estatísticos, como zonas de CEP com baixa ou nenhuma representatividade.

3.2. Desenho de Validação e Limitações Amostrais

Diante da grande limitação do tamanho da amostra ($n = 89$ observações úteis), o rigor na estratégia de validação torna-se o pilar central para garantir a capacidade de generalização do modelo.

Para a avaliação preditiva, adotou-se a estratégia de *Holdout Split* estratificado, particionando o conjunto de dados em 80% para treinamento e 20% para testes isolados (Blind Test). A estratificação (`stratify = y`) garantiu que a proporção original da classe minoritária (demissões) fosse preservada de forma idêntica em ambos os conjuntos. Adicionalmente, fixou-se uma semente pseudoaleatória (`random_state = 42`) para assegurar a reproduzibilidade exata do experimento.

É prudente classificar este desenho estatístico inicial como uma **Prova de Conceito Empírica (PoC)**. Embora a validação *Holdout* tenha sido suficiente para comprovar o aprendizado de máquina nesta etapa, a recomendação para interações futuras do modelo (*Future Works*), à medida que o volume de dados operacionais ultrapasse a marca de centenas de registros, é a adoção da Validação Cruzada (*K – Fold Cross Validation*), visando atestar a estabilidade da variância das métricas em múltiplas partições aleatórias do espaço amostral.

3.3. Tratamento de Classes Desbalanceadas (SMOTE)

O fenômeno de *Turnover (Saída Voluntária)* é, por natureza, um evento incomum em relação à permanência (Retenção). No conjunto de treinamento, algoritmos expostos a essa assimetria tendem a desenvolver um viés majoritário, classificando todos os colaboradores como “ativos” para inflar artificialmente a sua Acurácia Global.

Para reduzir esse viés sem recorrer à espera por novos dados, aplicou-se a técnica de SMOTE (Synthetic Minority Over-sampling Technique). De forma isolada e exclusiva no conjunto de treinamento (respeitando as boas práticas para evitar Data Leakage¹), o algoritmo gerou observações sintéticas baseadas na interpolação geométrica dos colaboradores que pediram demissão. O resultado foi um espaço de treinamento perfeitamente balanceado, forçando o *Random Forest* a aprender com clareza as fronteiras de decisão (*Decision Boundaries*) que separam um talento engajado e sem risco crítico de um talento não engajado com um risco de fuga.

4. Resultados e Discussão de Negócios

A avaliação de um modelo de People Analytics com foco em evasão não deve ser mensurada exclusivamente métricas de laboratório, devem ser avaliados pela sua capacidade de superar o senso comum e gerar Retornos sobre Investimentos (ROI) em cima de uma matriz de custos assimétrica. A transição de um modelo de “caixa preta” (*Black-Box*) para uma arquitetura interpretável (*White-Box*) é fundamental para a extração de inteligência de negócios.

4.1. Baseline Ingênuo vs. Matriz de Custos Assimétrica.

Para estabelecer o ganho real do algoritmo, foi necessário compará-lo a um *Baseline Ingênuo (Naive Classifier)*. Em uma operação com retenção histórica majoritária, um gestor (ou um modelo estatístico enviesado) que simplesmente previsse que “nenhum funcionário

¹ A aplicação de técnicas de *oversampling* (geração de dados sintéticos) previamente à divisão do conjunto de dados (*Train-Test Split*) constitui uma falha metodológica severa que resulta em Data Leakage (Vazamento de Dados). Se o SMOTE for aplicado a toda a base antes da separação, instâncias sintéticas geradas a partir de dados de testes serão inseridas no ambiente de treinamento. Isso permite que o modelo “memorize” as variações dos registros que deveriam ser inéditos na fase de validação, resultando em métricas de performance ilusórias e artificialmente infladas (*Overoptimistic Performance*), incapazes de refletir a real capacidade de generalização do modelo em um ambiente de produção não supervisionado.

Figura 1: Matriz de Confusão - Modelo de Risco de Fuga

		Previsão do Algoritmo (Target List)	
		Retido (0)	Evasão (1)
Realidade Histórica	Retido (0)	7	6
	Evasão (1)	2	3

Figura 1: Matriz de confusão do Modelo Final evidenciando o Recall de 60% na classe minoritária

pedirá demissão” alcançaria uma Acurácia global superior a 70%, porém com **0% de Recall**. Neste cenário, a companhia continuaria sendo financeiramente surpreendida com por 100% dos desligamentos.

O modelo desenvolvido por este estudo (Random Forest + SMOTE + *Feature Selection* com 6 variáveis vitais) abdicou da Acurácia ilusória para maximizar a Sensibilidade. No conjunto de validação cega (Holdout Test Set), o algoritmo conseguiu alcançar **60% de Recall** na classe minoritária, rastreando antecipadamente 3 de 5 eventos reais de fuga, e atingiu um **Score AUC de 0.646**, provando sua capacidade de discriminação probabilística (*Flight Risk Score*).

Sob a ótica da Economia Comportamental Empresarial, esta arquitetura obedece a uma matriz de custos radicalmente assimétrica:

- **O Custo de um Falso Negativo (O Ponto Cego):** Não prever uma demissão custa à empresa o acerto rescisório irrecuperável, a ociosidade da cadeira, o novo ciclo de *Recruiting* e a curva de aprendizado inerente à nova contratação (estimado historicamente entre 50% e 200% do salário anual do talento evadido).
- **O Custo de um Falso Positivo (O Falso Alarme):** Prever uma demissão que não ocorreria custa apenas o tempo alocado por um gestor ou um *Business Partner* para uma entrevista de retenção e alinhamento (*Stay Interview*). Trata-se de um

custo marginal quase nulo, que gera externalidades positivas para o clima organizacional.

Desta forma, a captura de 60% dos eventos de Turnover representa uma contenção direta e imediata da hemorragia financeira e perca de capital intelectual.

4.2. Interpretabilidade do Modelo e Direcionadores de Risco (Features Importance)

A fim de entregar um plano de ação tático para a área de Recursos Humanos, o peso de cada variável na decisão do algoritmo foi extraído através do Índice de Impureza de Gini (*Gini Importance / Mean Decrease in Impurity*).

O modelo revelou que a evasão não é um fenômeno randômico, mas sim governado por três maiores direcionadores ocultos na companhia:

1. **A Linha do Tempo de Evasão:** A variável quantitativa `meses_de_casa` dominou a árvore decisória. Esta descoberta técnica referendou a tese central deste estudo empírico, detalhada na seção a seguir.
2. **O Epicentro Operacional:** A alocação no “Departamento de Relacionamento” atua como um severo multiplicador do risco de fuga. O algoritmo isolou o problema sistêmico para uma área de negócio específica, descartando a hipótese de uma crise global de clima organizacional.
3. **O Fator Econômico (Competitividade e Atratividade Financeira):** Ao contrário das premissas iniciais que apontavam a falha de recrutamento como o principal ofensor, a variável `salario_contratual` despontou como o grande gatilho decisivo para o Turnover. O modelo provou matematicamente que o pacote de remuneração atua como o principal fiel da balança: quando o colaborador atinge a janela crítica de tempo de casa, o seu patamar salarial dita se ele mantém engajado ou se busca recolocação no mercado.

A Desmistificação da Falha de Processo:

É imperativo notar que a variável “Perfil Comportamental Não Mapeado” (ausência de dados no *Onboarding*) figurou na base da pirâmide de importância do algoritmo. A inteligência Artificial demonstrou que as falhas burocráticas no registro de entrada possuem impacto marginal no desligamento. A real alavanca para a evasão é estrutural: **Tempo, Remuneração e Lotação.**

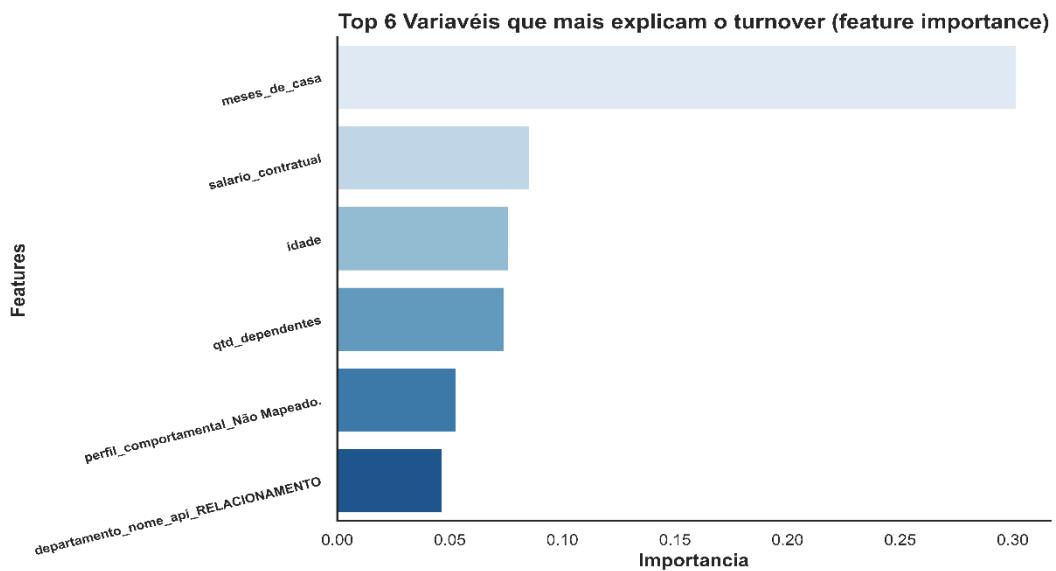


Figura 2: Importância Relativa das variáveis [Gini Importance] com dominância do Tempo de Casa e do Salário Contratual

4.3. A “Mortalidade Infantil Organizacional”

O cruzamento da Análise Exploratória de Dados (EDA) prévia com o alto peso algorítmico da variável **meses_de_casa** permitiu a cunhagem e a constatação empírica de um padrão comportamental sistêmico na operação: a **Mortalidade Infantil Organizacional**.

A distribuição dos dados evidenciou que o risco de fuga voluntária não é de forma linear ao longo da jornada do colaborador. Existe uma “zona de arrebatamento” altamente concentrada, com um pico de eventos de *Turnover* ocorrendo de forma precoce, especificamente com uma mediana situada entre **9º e 11º mês de contratação**.

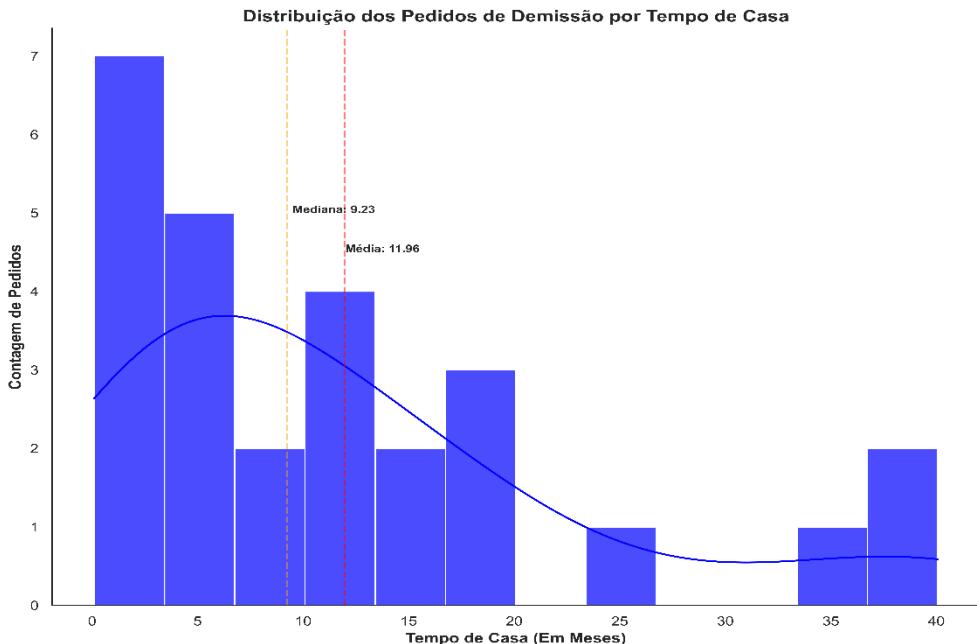


Figura 3: Distribuição de densidade apontando o pico de evasão - "Mortalidade Infantil" - entre 9º e 11º meses

O conceito de “Mortalidade Infantil Organizacional” descreve a janela crítica onde o “Contrato Psicológico” inicial estabelecido entre o talento e a empresa entra em colapso. Representa o exato momento em que o período de “lua de mel” e a adaptação se encerra, as expectativas reais de encarreiramento se chocam com a rotina da operação (especialmente no setor de Relacionamento) e o colaborador toma a decisão íntima de saída.

Se não houver uma intervenção estruturada de Gestão de Pessoas (ex.: avaliação de desempenho com feedbacks claros ou bônus de retenção) por volta do 8º mês, o colaborador entra em propensão de fuga irreversível antes mesmo de completar o seu primeiro ciclo anual na companhia. A eficácia preditiva do modelo reside, em grande parte, em sua capacidade de rastrear colaboradores ativos que estão adentrando essa janela de vulnerabilidade temporal.

4.4. Justificativa de Complexidade: Modelagem Não Linear vs. Parcimônia Linear

Sob a ótica da parcimônia estatística (Navalha de Ockham), modelos de alta complexidade (*Ensembles*) só devem ser adotados se provarem superioridade clara sobre abordagens lineares mais simples e transparentes.

Para validar essa premissa, o desempenho do algoritmo *Random Forest* foi contraposto a um modelo de Regressão Logística com penalização de classes `(class_weight = 'balanced')`. No conjunto de testes, a Regressão Logística apresentou um desempenho

sensivelmente inferior, alcançando um *Recall* de apenas 40% (capturando 2 dos 5 eventos). A introdução da não-linearidade geométrica provida pela Floresta Aleatória elevou a captura para 60% (3 de 5). No contexto de Recursos Humanos, onde o custo unitário de uma rescisão surpresa atinge a casa dos milhares de reais, este ganho marginal de 20 pontos percentuais na Sensibilidade justifica financeiramente a adoção de uma arquitetura matemática de maior complexidade algorítmica.

4.5. Limitações Críticas e Instabilidade Amostral

Embora os resultados atestem a viabilidade do modelo preditivo, é imperativo dizer manter a sobriedade analítica frente ao escore AUC de 0.646. Este valor reflete uma calibração probabilística modesta, típica do fenômeno de *Data Starvation (escassez de dados)*.

Adicionalmente, dada a reduzida cardinalidade da classe minoritária no conjunto de testes (apenas 5 eventos de demissão confirmados), pequenas variações individuais impactam significativamente as métricas de performance. O deslocamento de um único indivíduo de Falso Negativo para Verdadeiro Positivo altera o *Recall* absoluto em 20%. Esta instabilidade amostral inerente não invalida o direcionamento estratégico da Prova de Conceito (PoC), mas reforça a necessidade incontornável de expansão amostra. À medida que a operação maturar, a introdução de validações cruzadas repetidas (*Repeated K-Fold Cross Validation*) será necessária para atestar a estabilidade dos intervalos de confiança do modelo e refinar a precisão contínua da probabilidade.

A Evidência Probabilística: A Curva ROC

Para materializar a capacidade de separação do modelo além da classificação binária (corte de 50%), a performance probabilística foi plotada através da Curva ROC (*Receiver Operating Characteristic*). O gráfico abaixo ilustra o distanciamento da performance do modelo *Random Forest* (com *Feature Selection*) em relação à linha de aleatoriedade (*Baseline* de 0.50).

A área sob a curva ($AUC = 0.646$) ratifica graficamente que, apesar das limitações amostrais previamente discutidas, o modelo desenvolveu uma capacidade de discriminação probabilística superior ao acaso. Esta métrica chancela a utilização do algoritmo para a ordenação de risco (*Ranking*), permitindo que a área de Gestão de Pessoas priorize abordagens aos colaboradores que se encontram no quartil superior de probabilidade da *Target List*.

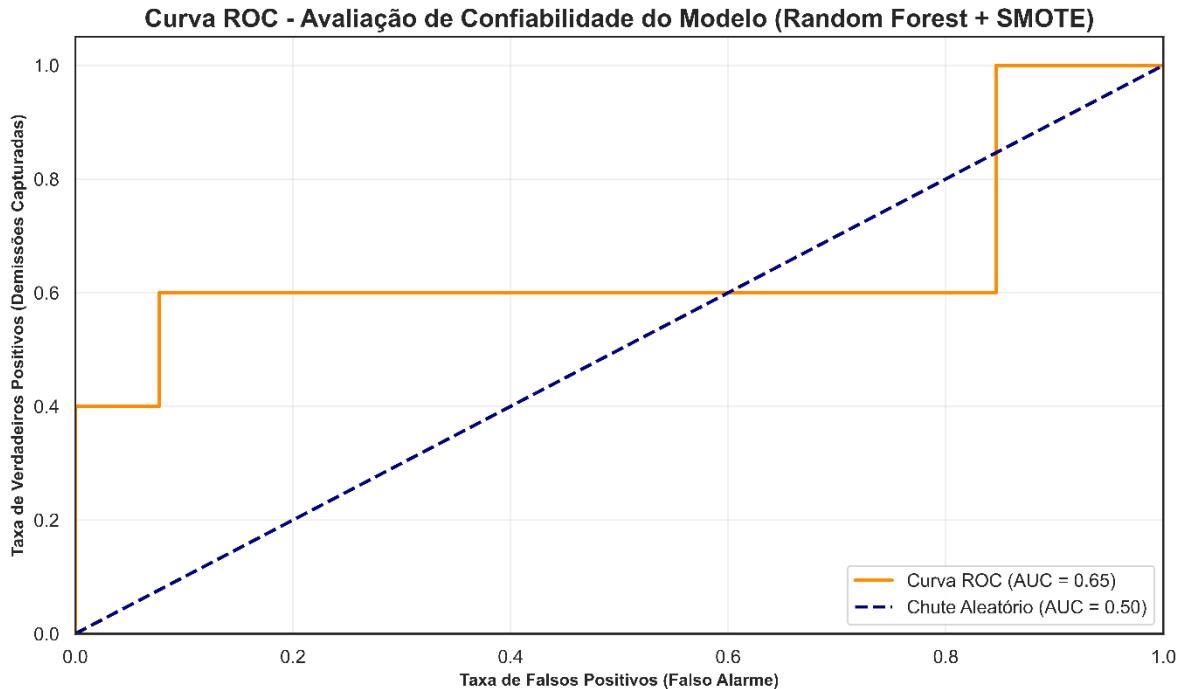


Figura 4: Curva ROC ilustrando o Score AUC de 0.646 e a superioridade do modelo sobre a classificação aleatória

5. Conclusão e Plano de Ação Estratégico

Este estudo comprovou a viabilidade técnica, metodológica e financeira de se transicionar o departamento de Recursos Humanos de uma postura historicamente reativa para um paradigma preditivo (*People Analytics 3.0*).

Apesar das limitações inerentes a uma amostra de dados operacionais embrionária ($n = 89$), a Prova de Conceito (PoC) utilizando o algoritmo *Random Forest*, otimizado pela técnica de balanceamento sintético (SMOTE) e por uma severa redução de dimensionalidade (*Feature Selection*), demonstrou capacidade de antecipar 60% das evasões voluntárias da companhia.

O distanciamento de um modelo “caixa preta” permitiu a refutação empírica de dogmas corporativos: a análise provou que o pacote de remuneração e a lotação no “Departamento de Relacionamento” suplantam massivamente eventuais falhas comportamentais de recrutamento no momento da decisão de fuga.

Mais importante, a modelagem descortinou o fenômeno da “**Mortalidade Infantil Organizacional**”, alertando a Diretoria para a janela crítica de evasão estabelecida entre o 9º e o 11º mês de contrato.

5.1. Plano de Ação e Próximos Passos

Para materializarmos o ganho analítico em Retorno sobre Investimento (ROI), recomendam-se três frentes de ação imediatas para a Gestão de Pessoas:

- 1. Operacionalização da Target List Mensal:** A saída probabilística do modelo deve ser conectada a um painel de *Business Intelligence* (ex.: Power BI). Os *Business Partner* (BPs) de RH e gestores de área devem passar a utilizar a lista de colaboradores que cruzarem a fronteira de 75% de Risco de Fuga (*Flight Risk Score*) para conduzir conversas cirúrgicas de retenção, renegociação ou alinhamento de plano de carreira, com foco especial naqueles que se aproximam do 8º mês de casa.
- 2. Revisão de Política de Remuneração e Benefícios:** Uma vez que a variável `salario_contratual` provou ser um gatilho decisório superior às falhas de perfil Sólides, recomenda-se uma auditoria mercadológica na tabela salarial vigente, a fim de mitigar a perda de profissionais já treinados para a concorrência direta.
- 3. Treinamento de Liderança:** Como o comprovado na EDA e sendo uma das variáveis que elevam o índice de turnover, é altamente necessária uma revisão nas rotinas e no clima do “Departamento de Relacionamento”, tendo em vista que o setor possui uma taxa de turnover 2x maior que o segundo setor com o maior turnover.
- 4. Maturidade e Governança de Dados (*Continuous Learning*):** O Score AUC atual de 0.646 valida o uso gerencial imediato, mas exige refinamento. Estabelece-se a necessidade imperativa de re-treinamento do modelo a cada ciclo semestral. O incremento orgânico do N amostral, atrelado à adoção futura de *Repeated K-Fold Cross Validation*, permitirá elevar a calibração probabilística do algoritmo a patamares de excelência, espera-se incremento progressivo da capacidade discriminatória com expansão amostral.