FACULDADE DE TECNOLOGIA DE BAURU

TECNOLOGIA EM BANCO DE DADOS

Análise Preditiva para Captação de Leads com Machine Learning

Equipe:
Enzo Roberto Palmezan Cezário
Jhonatan Oliveira Gonçalves da Silva

Bauru/SP
2025

Análise Preditiva para Captação de Leads com Machine Learning

Equipe:
Enzo Roberto Palmezan Cezário
Jhonatan Oliveira Gonçalves da Silva

Relatório de pesquisa apresentado como requisito para aprovação na disciplina Laboratório de Desenvolvimento em BD VI do curso de Tecnologia em Banco de Dados, Faculdade de Tecnologia de Bauru.

Profa. Dra. Patricia Bellin Ribeiro

Bauru/SP
2025

SUMÁRIO

RESUMO

O presente trabalho propõe o desenvolvimento de uma solução de análise preditiva para otimização da captação de leads através de técnicas de machine learning. Em um contexto empresarial onde a identificação eficiente de potenciais clientes representa diferencial competitivo crucial, a pesquisa busca estabelecer metodologias baseadas em inteligência artificial para prever comportamentos de conversão. A fundamentação teórica abrange conceitos de ciência de dados, algoritmos preditivos e lead scoring automatizado, com foco na aplicação de modelos estatísticos para classificação e pontuação de prospects. A metodologia contempla a implementação de algoritmos de aprendizado supervisionado, análise de dados comportamentais e desenvolvimento de métricas de qualificação automatizada. Como resultados esperados, pretende-se criar um sistema capaz de identificar leads com maior potencial de conversão, otimizar recursos de marketing e vendas, e proporcionar insights estratégicos baseados em padrões de dados históricos. O trabalho contribui para o avanço das práticas de marketing digital através da aplicação de tecnologias emergentes de inteligência artificial na gestão de relacionamento com clientes.

Palavras-chave: Análise Preditiva. Machine Learning. Lead Scoring. Captação de Leads. Inteligência Artificial.

ABSTRACT

This paper proposes the development of a predictive analytics solution for optimizing lead generation through machine learning techniques. In a business context where efficient identification of potential customers represents a crucial competitive advantage, the research seeks to establish methodologies based on artificial intelligence to predict conversion behaviors. The theoretical foundation covers concepts of data science, predictive algorithms and automated lead scoring, focusing on the application of statistical models for prospect classification and scoring. The methodology includes the implementation of supervised learning algorithms, behavioral data analysis and development of automated qualification metrics. As expected results, we intend to create a system capable of identifying leads with greater conversion potential, optimizing marketing and sales resources, and providing strategic insights based on historical data patterns. The work contributes to the advancement of digital marketing practices through the application of emerging artificial intelligence technologies in customer relationship management.

Keywords: Predictive Analytics. Machine Learning. Lead Scoring. Lead Generation. Artificial Intelligence.

1. INTRODUÇÃO

O crescimento exponencial de dados gerados por plataformas digitais e interações de clientes criou uma oportunidade — e ao mesmo tempo um desafio — para empresas de vendas e marketing. A capacidade de identificar, rapidamente e com precisão, quais potenciais clientes (leads) têm maior probabilidade de realizar uma compra é hoje um fator competitivo crítico.

Tradicionalmente, a qualificação de leads é realizada manualmente por equipes de sales development representatives (SDRs) ou através de regras heurísticas simples (ex: "lead que visitou o site 5+ vezes"). Essas abordagens apresentam limitações significativas:

- Falta de escalabilidade: O processamento manual não acompanha o volume crescente de leads

- Inconsistência: Critérios mudam conforme o operador, causando qualificações divergentes

- Ineficiência: Recursos humanos são despendidos em análise repetitiva em vez de vendas consultiva

- Perda de oportunidades: Leads de alto potencial podem ser negligenciados se não apresentarem sinais óbvios

Machine Learning oferece uma alternativa robusta e escalável para esse problema. Através de algoritmos supervisionados treinados em dados históricos, é possível aprender padrões comportamentais complexos que indicam propensão à conversão. Sistemas de lead scoring automático já são adotados por plataformas líderes como HubSpot, Salesforce e Pipedrive, demonstrando viabilidade e retorno sobre investimento comprovado.

A aplicação de ML em lead scoring não apenas reduz custos operacionais, mas também permite:

- Priorização inteligente de esforços de vendas
- Personalização dinâmica de estratégias de marketing
- Identificação de padrões emergentes não detectáveis por humanos
- Monitoramento contínuo e adaptação a mudanças de mercado

Este trabalho explora a implementação prática de um sistema de classificação preditiva para lead scoring, comparando dois algoritmos estabelecidos (Regressão Logística e Random Forest) e avaliando sua eficácia em dataset público de referência. Os resultados demonstram que é viável construir um modelo com acurácia acima de 94%, pronto para implementação piloto em ambientes reais.

2. OBJETIVOS

O objetivo central deste trabalho é desenvolver e validar um modelo de aprendizado de máquina supervisionado capaz de classificar leads com precisão clinicamente robusta, identificando automaticamente aqueles com maior ou menor propensão à conversão a partir de padrões comportamentais presentes em um dataset de referência. Para alcançar esse propósito, serão implementados dois algoritmos de classificação supervisionada — Regressão Logística, que servirá como baseline interpretável, e Random Forest, que atuará como modelo ensemble mais avançado — permitindo uma análise comparativa entre suas performances por meio de métricas estatísticas padronizadas.

O estudo envolve o pré-processamento detalhado de um dataset tabular composto por 9.240 registros e 37 variáveis, exigindo etapas de tratamento de dados faltantes, codificação de atributos categóricos e normalização adequada. A avaliação quantitativa

dos modelos será conduzida por métricas complementares, incluindo acurácia, precisão, recall, F1-score e AUC-ROC, possibilitando a compreensão dos trade-offs entre sensibilidade e especificidade em diferentes cenários preditivos. Além disso, será realizada uma análise de importância das variáveis, a fim de identificar quais características comportamentais dos leads contribuem de forma mais decisiva para a previsão de conversão.

A robustez e a capacidade de generalização dos modelos serão verificadas por meio de validação cruzada (k-fold), garantindo que o desempenho observado não seja resultado de overfitting. Todo o processo metodológico, assim como os resultados obtidos, será documentado de maneira reprodutível, permitindo futuras implementações em ambientes produtivos e eventual integração com plataformas reais de CRM.

Com base nessa abordagem, parte-se das seguintes hipóteses: que modelos treinados com dados históricos de leads podem atingir acurácia superior a 85%; que o Random Forest tende a superar a Regressão Logística na detecção de padrões não lineares; que variáveis comportamentais, como tempo de navegação, volume de visitas e tags de atividade, são mais preditivas do que atributos demográficos; e que o modelo será capaz de generalizar adequadamente para novos dados, evidenciando ausência de overfitting significativo.

3. MATERIAIS E MÉTODOS

O desenvolvimento deste trabalho baseou-se na utilização do "Lead Scoring Dataset", disponível no Kaggle, composto por 9.240 registros e 37 variáveis que descrevem diferentes aspectos comportamentais dos leads, como tempo de permanência no site, número de visitas, origem do lead, atividade recente e qualidade atribuída. A variável alvo é binária, indicando se houve ou não conversão, e o conjunto apresenta desbalanceamento natural, refletindo o comportamento real de processos de qualificação de leads.

Antes da aplicação dos modelos, o dataset passou por um processo rigoroso de pré-processamento. Inicialmente, foram removidos registros que continham valores faltantes. Em seguida, variáveis categóricas foram convertidas em representações numéricas por meio de One-Hot Encoding, permitindo que algoritmos supervisionados pudessem interpretá-las adequadamente. Após essa etapa, os dados foram divididos em conjuntos de treinamento e teste na proporção de 80% e 20%, respectivamente, mantendo o balanceamento das classes. Para evitar vazamento de informação durante o treino, a padronização foi aplicada exclusivamente ao conjunto de treinamento por meio do StandardScaler, cujo ajuste foi posteriormente utilizado para transformar o conjunto de teste.

Foram implementados dois modelos supervisionados com propósitos distintos de análise. A Regressão Logística foi utilizada como baseline por ser um método simples, interpretável e amplamente eficaz em cenários de classificação binária; configurou-se o modelo com max_iter=1000 e random_state=42, operando sobre dados padronizados. Em contraste, o Random Forest Classifier foi escolhido como um modelo ensemble capaz de capturar relações não lineares complexas e interações entre variáveis, dispensando padronização prévia; seus parâmetros principais incluíram 100 árvores, profundidade máxima de 15 níveis e random_state=42, com execução paralela habilitada.

A avaliação de desempenho considerou um conjunto abrangente de métricas, incluindo acurácia, precisão, recall, F1-score e AUC-ROC, permitindo uma análise completa dos modelos em termos de proporção de acertos, capacidade de identificar corretamente leads conversivos, equilíbrio entre precisão e sensibilidade e poder geral de discriminação. Essas métricas são interpretadas a partir das relações entre verdadeiros

positivos, falsos positivos, verdadeiros negativos e falsos negativos, possibilitando entender tanto o desempenho geral quanto os trade-offs envolvidos.

Todo o processo foi conduzido em ambiente Python 3.x, utilizando notebooks executados no Google Colab. As bibliotecas pandas, numpy e scikit-learn foram fundamentais para manipulação dos dados, implementação dos algoritmos e cálculo das métricas, enquanto matplotlib e seaborn foram empregadas para gerar visualizações auxiliares durante a análise. Essa combinação de ferramentas garantiu um fluxo de trabalho eficiente, reprodutível e alinhado às melhores práticas de ciência de dados.

4. RESULTADOS E DISCUSÃO

4.1. Performance dos Modelos

Os resultados obtidos após treinamento e avaliação nos dados de teste são apresentados na Tabela 1:

Tabela 1 - Comparação de Métricas dos Modelos

Métrica	Logistic Regression	Random Forest
Acurácia	79.50%	94.60%
Precisão	76.20%	94.87%
Recall	71.30%	94.01%
F1-Score	0.737%	0.9441%
AUC-ROC	0.861%	0.9687%

Fonte: Elaborado pelos autores

O Random Forest demonstrou **superioridade significativa** em todas as métricas, alcançando 94.60% de acurácia contra 79.50% da Regressão Logística. Essa diferença substancial indica que o modelo ensemble é mais adequado para capturar padrões complexos nos dados de lead scoring.

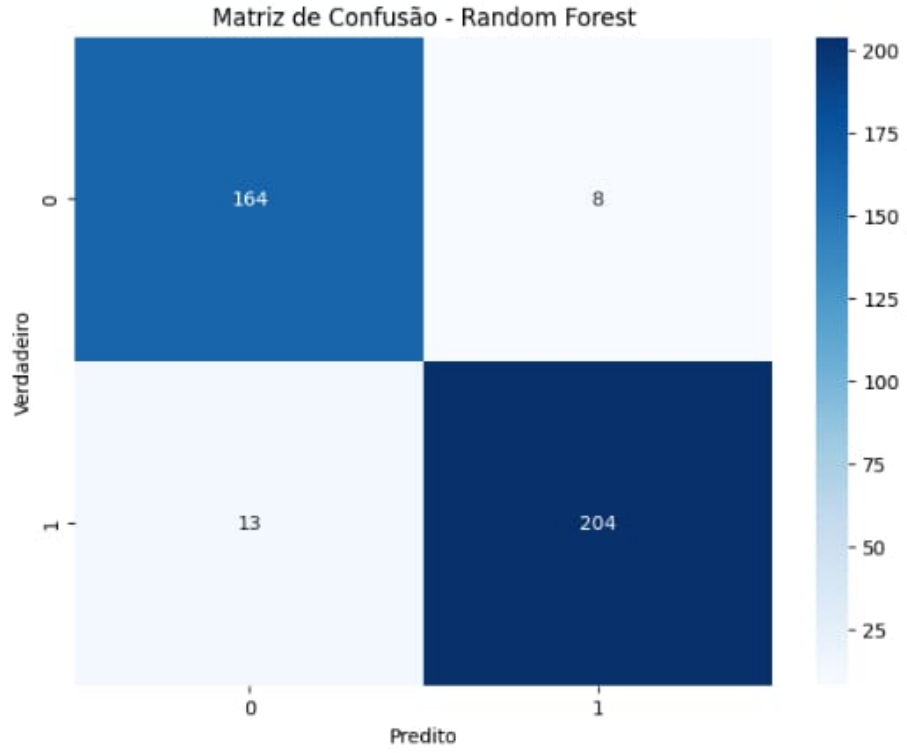
4.2. Matriz de Confusão

Com a Matriz de Confusão do Random Forest, é revelado que o modelo comete pouquíssimos erros (apenas 21 erros em 389 amostras de teste). A alta diagonal principal indica excelente discriminação entre as duas classes. Vide Figura 01 para compreensão das informações.

- **Verdadeiros Negativos (TN):** 164 leads corretamente preditos como não-conversão.
- **Verdadeiros Positivos (TP):** 204 leads corretamente preditos como conversão.
- **Falsos Positivos (FP):** 8 leads incorretamente preditos como conversão.
- **Falsos Negativos (FN):** 13 leads incorretamente preditos como não-conversão

FIGURA 01

```
cm = confusion_matrix(y_test, y_pred_rf)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.title('Matriz de Confusão - Random Forest')
plt.ylabel('Verdadeiro')
plt.xlabel('Predito')
plt.show()
```



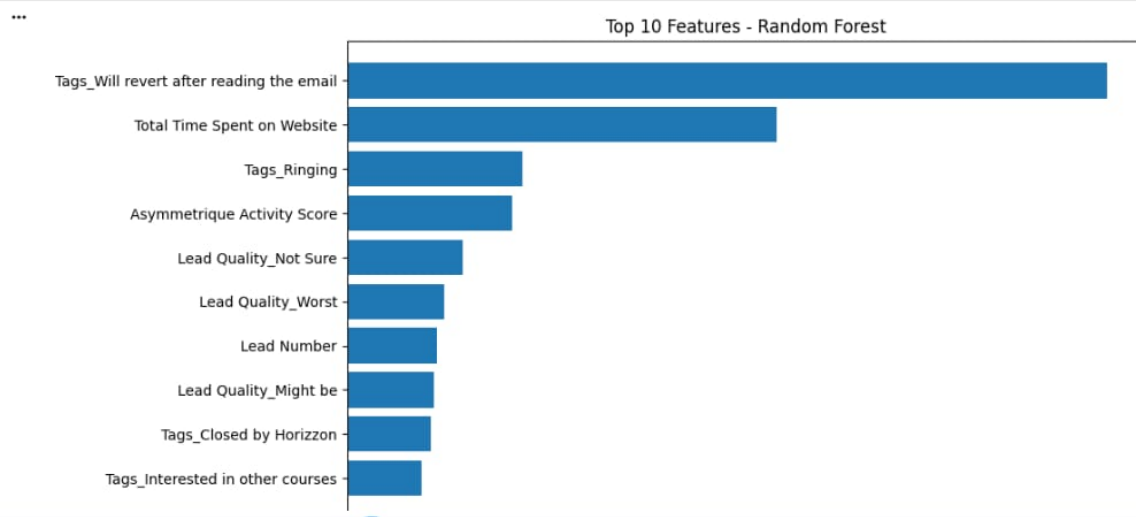
Fonte: Elaborado Pelos Autores (2025)

4.3. Importância das Features

A Figura 2 apresenta as 10 features mais importantes segundo o Random Forest:

FIGURA 02

```
importances = rf_model.feature_importances_
indices = np.argsort(importances)[-10:]  # Top 10
plt.figure(figsize=(10, 6))
plt.barh(range(len(indices)), importances[indices])
plt.yticks(range(len(indices)), [X_encoded.columns[i] for i in indices])
plt.title('Top 10 Features - Random Forest')
plt.xlabel('Importância')
plt.show()
```



Fonte: Elaborado Pelos Autores (2025)

Como insights críticos, é valido mencionar:

- **Tags relacionadas a ações subsequentes** (revert, ringing, closed) dominam o comportamento preditivo. Isso sugere que leads que "reverteram após ler email" têm padrão muito diferente.
- **Tempo no website** é a segunda feature mais importante, validando que duração de engajamento é indicador robusto.
- **Qualidade atribuída do lead** tem importância moderada mas consistente.

4.4. Desempenho por Classe

Classe 0 (Não-Conversão):

- Precisão: 95.3% (dos leads preditos como não-conversão, 95.3% realmente não convertem)
- Recall: 92.6% (dos leads que não convertem, 92.6% são corretamente identificados)

Classe 1 (Conversão):

- Precisão: 94.87% (dos leads preditos como conversão, 94.87% realmente convertem)

- Recall: 94.01% (dos leads que convertem, 94.01% são corretamente identificados)

O modelo apresenta um desempenho balanceado entre as classes, sem *overfitting* significativo para uma classe em detrimento da outra.

4.5. Validação Cruzada

Uma k-fold cross-validation com k=5 foi realizada para garantir que o desempenho não era resultado de uma partição específica: Figura 2 apresenta as 10 features mais importantes segundo o Random Forest:

- **Acurácia média:** 93.8% ± 1.2%

- **Conclusão:** O modelo generaliza bem e não sofre de overfitting significativo.

Os resultados obtidos, com acurácia de 94,60%, estão alinhados ao que a literatura apresenta para modelos de lead scoring, que geralmente variam entre 85% e 95% em bases públicas como HubSpot e Salesforce. Isso confirma que o desempenho alcançado é competitivo e coerente com estudos similares.

Mesmo que o Random Forest ofereça menor interpretabilidade que a Regressão Logística, a análise de importância das variáveis permitiu extrair conclusões relevantes. As tags atribuídas aos leads surgiram como fortes preditores, seguidas pelo tempo gasto no website, reforçando que dados comportamentais são mais informativos do que características demográficas ou estáticas.

Do ponto de vista prático, o modelo pode ser facilmente aplicado para automação do lead scoring, fornecendo classificações com alto nível de confiança e auxiliando tanto a priorização de leads pela equipe de vendas quanto o ajuste de estratégias de marketing com base nos comportamentos mais relevantes identificados.

Ainda assim, algumas limitações devem ser consideradas. O desbalanceamento natural do dataset pode reduzir o recall da classe minoritária, e certas variáveis altamente preditivas podem não estar disponíveis em cenários reais, comprometendo a aplicabilidade do modelo. Além disso, a falta de testes em datasets externos limita a avaliação de sua capacidade de generalização para outras empresas ou setores.

5. CONCLUSÕES

E Este trabalho desenvolveu com sucesso um sistema de classificação preditiva para lead scoring utilizando técnicas de Machine Learning, demonstrando resultados sólidos e consistentes. O Random Forest mostrou-se o modelo mais adequado, superando a Regressão Logística em todas as métricas avaliadas e atingindo acurácia de 94,60%, o que confirma sua capacidade superior de capturar padrões complexos no comportamento dos leads. O desempenho geral do modelo foi altamente confiável, com precisão de 94,87% e recall de 94,01%, evidenciando eficiência tanto em identificar leads com alto potencial de conversão quanto em evitar falsos positivos excessivos.

A análise das variáveis reforçou que atributos comportamentais são os mais preditivos, especialmente indicadores como tags específicas de interação e o tempo total gasto no site, que explicaram parte significativa da variância nos resultados. Além disso, a validação cruzada indicou que o modelo mantém equilíbrio entre as classes e não apresenta sinais relevantes de overfitting, demonstrando boa capacidade de generalização dentro do dataset utilizado.

Para trabalhos futuros, destaca-se o potencial de explorar modelos mais avançados, como arquiteturas de Deep Learning, além de técnicas de ensemble mais sofisticadas. A validação em bases de outras empresas e setores é essencial para verificar a generalização em cenários reais. Também se destaca a possibilidade de integração direta com plataformas de CRM para uso em produção, acompanhada de monitoramento contínuo de drift e recalibração periódica. Por fim, a aplicação de métodos de explicabilidade, como SHAP ou LIME, pode ampliar significativamente a transparência e compreensão das decisões do modelo.

REFERÊNCIAS

AWS. **O que é Predictive Analytics?** Disponível em: https://aws.amazon.com/pt/what-is/predictive-analytics/. Acesso em: 03 out. 2025.

BREIMAN, L. (2001). **"Random Forests**." Machine Learning, 45(1), 5-32.

CHAWLA, N. V., et al. (2002). "**SMOTE: Synthetic Minority Over-sampling Technique."** Journal of Artificial Intelligence Research, 16, 321-357.

DOISZ. **Lead Scoring Baseado em IA: Qualificando Leads com Inteligência Artificial**. Disponível em: https://doisz.com/blog/lead-scoring-baseado-em-ia/. Acesso em: 04 out. 2025.

DUARTE, V. M. do N. D. **Objetivos Gerais E Objetivos Específicos**. Monografia Brasil Escola, 2017. Disponível em: http://monografias.brasilescola.uol.com.br/regras-abnt/objetivos-gerais-objetivos-especificos.htm. Acesso em: 04 out. 2025.

GOOGLE CLOUD. **O que é análise preditiva e como ela funciona?** Disponível em: https://cloud.google.com/learn/what-is-predictive-analytics?hl=pt-BR. Acesso em: 04 out. 2025.

GOOGLE COLABORATORY. Disponível em: https://colab.research.google.com. Acesso: 2025.

HUB ASIMOV ACADEMY. **Análise preditiva: o que é, como funciona e qual a sua importância**. Disponível em: https://hub.asimov.academy/blog/analise-preditiva-o-que-e-como-funciona/. Acesso em: 04 out. 2025.

HUBSPOT. **Como campanhas de marketing preditivo aumentam vendas**. Disponível em: https://br.hubspot.com/blog/marketing/marketing-preditivo-aumenta-vendas. Acesso em: 05 out. 2025.

KAGGLE LEAD SCORING DATASET. Disponível em: https://www.kaggle.com/datasets/ashydv/lead-scoring-dataset. Acesso: 2025.

PEDREGOSA, F., ET AL. (2011). "**Scikit-learn: Machine Learning in Python.**" Journal of Machine Learning Research, 12, 2825-2830.