Investigando Fatores de Sucesso em Marketing Bancário para Depósitos a Prazo usando Aprendizado de Máquina Supervisionado

Felipe Oliveira do Espirito Santo - 11925242 [**Universidade de São Paulo** | ***felipe.santo@usp.br***]
Samuel Origa da Silva - 17049253 [**Universidade de São Paulo** | ***samuel.origa@usp.br***]

Resumo Este trabalho investiga os fatores que influenciam a adesão a depósitos a prazo por clientes durante campanhas de telemarketing bancário e avalia a viabilidade de prever tal resultado usando aprendizado de máquina supervisionado. O estudo é feito com base no dataset ”Bank Marketing” da UCI, que contém dados detalhados de campanhas reais. O método proposto envolve uma análise exploratória de dados para identificar padrões e relações relevantes, seguida pelo desenvolvimento e avaliação de modelos de classificação. Os procedimentos planejados para pré-processamento e tratamento de dados, bem como as estratégias de avaliação para a tarefa de classificação, também são incluídos no método. Espera-se que este estudo forneça conhecimentos aplicáveis sobre os direcionadores de sucesso em marketing bancário e demonstre o potencial preditivo de modelos de ML neste contexto, contribuindo com uma análise metodologicamente rigorosa de um problema de negócio relevante.

Palavras-chave: Marketing Bancário, Aprendizado de Máquina Supervisionado, Classificação, Depósito a Prazo.

1 Introdução

A otimização de campanhas de marketing direto é um desafio constante para instituições financeiras que buscam maximizar o retorno sobre seus investimentos e aprimorar o relacionamento com o cliente. Campanhas de telemarketing, especificamente para produtos como depósitos a prazo, representam um investimento significativo, cuja eficácia é modulada por uma miríade de fatores [Moro *et al*., 2014b]. Compreender quais características do cliente, detalhes da interação de marketing e condições socioeconômicas influenciam a decisão de adesão é fundamental para aprimorar a segmentação, personalizar abordagens e aumentar a taxa de conversão.

O advento do Aprendizado de Máquina (ML) proporcionou novas ferramentas para analisar grandes volumes de dados de campanhas e extrair padrões complexos que podem não ser aparentes através de análises tradicionais. Modelos de classificação supervisionada, em particular, podem ser treinados com dados históricos para prever a probabilidade de um cliente aceitar uma oferta, permitindo que os bancos concentrem seus esforços nos clientes mais promissores.

Diversos estudos têm explorado a aplicação de ML em marketing bancário. Moro *et al*. [2014b], por exemplo, apresentaram uma abordagem baseada em dados para prever o sucesso do telemarketing bancário, utilizando o mesmo dataset que fundamenta a investigação deste trabalho. No entanto, a identificação precisa dos fatores mais influentes e a construção de modelos preditivos robustos e realistas (que não dependam de informações conhecidas apenas após o evento, como a duração da chamada) continuam sendo itens de interesse.

Neste contexto, este estudo tem o intuito de responder a seguinte questão: *‘Quais são os fatores que influenciam a adesão a depósitos a prazo em uma campanha de marketing bancário?’*

Para responder a essas questão, é proposto um método detalhado que combina Análise Exploratória de Dados (EDA) com a aplicação e avaliação de diversos algoritmos de classificação supervisionada. A contribuição deste trabalho reside na aplicação sistemática de técnicas de experimentação em ML para extrair conhecimentos aplicáveis de um dataset público relevante e avaliar o potencial preditivo de diferentes abordagens, com atenção especial às boas práticas de modelagem, como evitar vazamento de dados.

Este artigo está estruturado da seguinte forma: a Seção 2 detalha o método proposto, incluindo a descrição dos dados e os planos de pré-processamento, EDA e modelagem. A Seção 3 discute os resultados esperados e potenciais implicações. Finalmente, a Seção 4 apresenta as considerações finais e os próximos passos.

2 Método

Esta seção descreve a abordagem metódica planejada para conduzir a investigação, seguindo as diretrizes de experimentação em aprendizado de máquina [Roman, 2025a,b].

2.1 População e fonte de dados

A população de interesse compreende todo cliente bancário exposto a uma campanha de marketing com objetivo de venda de algum produto. Após buscas na web, foi encontrado o dataset ‘Bank Marketing’[1], que contém dados referentes a campanhas de telemarketing de um banco português entre os anos de 2008 e 2010. A escolha desse dataset se baseia nos fatos de que, até onde se sabe, é o único dataset público com dados reais de campanhas de marketing bancário e pela diversidade de variáveis dependentes, que podem proporcionar diferentes ângulos de análise. É importante notar que os dados referem-se a uma única insti-

[1] Dataset disponível em `https://archive.ics.uci.edu/dataset/222/bank+marketing`

tuição bancária portuguesa e a um período específico. Portanto, a generalização direta dos resultados para outros bancos, países, períodos de tempo e até outros canais de marketing (para além do telefônico) deve ser feita com as devidas ressalvas. Além disso, a própria natureza de ser um dataset de campanhas significa que a população estudada é a que foi propriamente contatada, não necessariamente todos os clientes elegíveis a algum contato.

O dataset está no formato Comma Separated Values (CSV) e possui 45.212 registros, em que cada linha representa um contato telefônico realizado a um cliente da instituição bancária, sendo que mais de um contato pode ser feito para um mesmo cliente. Para fins de planejamento e EDA inicial, coletou-se uma amostra aleatória de 20% dos dados contidos no dataset, a partir de uma distribuição uniforme, totalizando 9042 registros. Mais detalhes sobre o dataset serão apresentados nas seções seguintes.

2.2 Caracterização e pré-processamento dos dados

Os dados contêm 16 variáveis dependentes com informações do cliente (idade, profissão, etc.), do contato (mês, dia, tipo), de campanhas anteriores e socioeconômicas, além da variável independente 'y' (adesão ao depósito a prazo, 'yes'/'no'). A caracterização detalhada das variáveis foi realizada com base na documentação apresentada em [Moro *et al*., 2014a]. As seguintes variáveis dependentes estão presentes no dataset:

- `age` (inteiro, quantitativo de razão): Idade do cliente.
- `job` (texto, qualitativo nominal): Tipo de profissão.
- `marital` (texto, qualitativo nominal): Estado civil.
- `education` (texto, qualitativo ordinal): Nível de escolaridade.
- `default` (texto, qualitativo nominal): Indicador de inadimplência por parte do cliente.
- `balance` (inteiro, quantitativo de razão): Saldo médio anual em euros.
- `housing` (texto, qualitativo nominal): Indicador de posse de empréstimo habitacional por parte do cliente.
- `loan` (texto, qualitativo nominal): Indicador de posse de empréstimo pessoal por parte do cliente.
- `contact` (texto, qualitativo nominal): Tipo de comunicação do contato.
- `day` (inteiro, qualitativo ordinal): Dia do mês do último contato (1-31).
- `month` (texto, qualitativo ordinal): Mês do último contato no ano.
- `duration` (inteiro, quantitativo de razão): Duração do último contato, em segundos.
- `campaign` (inteiro, quantitativo de razão): Número de contatos realizados para o cliente durante a campanha.
- `pdays` (inteiro, quantitativo de razão): Número de dias que passaram após o cliente ter sido contatado pela última vez de uma campanha anterior (-1 significa que o cliente não foi contatado anteriormente).
- `previous` (inteiro, quantitativo de razão): Número de contatos realizados para o cliente antes da campanha.
- `poutcome` (texto, qualitativo nominal): Resultado da campanha de marketing anterior.

Table 1. Contagem de valores faltantes identificados na amostra inicial (20% do dataset).

Variável	Contagem de valores 'unknown' ou vazios ('')
`age`	0
`job`	56
`marital`	0
`education`	364
`default`	0
`balance`	0
`housing`	0
`loan`	0
`contact`	2574
`day`	0
`month`	0
`duration`	0
`campaign`	0
`pdays`	0
`previous`	0
`poutcome`	7358
`y`	0

Com base na caracterização e nos problemas identificados na amostra, o seguinte pipeline de pré-processamento será implementado e ajustado exclusivamente no conjunto de treino, e depois aplicado aos conjuntos de validação e teste:

1. **Valores ausentes/desconhecidos:** Conforme observado na Tabela 1, as variáveis `job`, `education`, `contact` e `poutcome` apresentam valores faltantes. A alta frequência, especialmente em `poutcome`, requer uma estratégia de tratamento. Durante a EDA, será investigado o significado e a distribuição desses valores. Possíveis abordagens a serem exploradas incluem: tratar 'unknown' como uma categoria separada, usar técnicas de imputação (ex: moda, ou modelos mais sofisticados se justificado), ou avaliar a remoção da feature (especialmente para `poutcome` devido à alta frequência). A decisão será baseada no impacto na análise e nos modelos.
2. **Codificação de variáveis categóricas:** As features textuais (nominais e ordinais) precisarão ser convertidas para um formato numérico. Técnicas como One-Hot Encoding, Label Encoding (para ordinais), ou outras serão consideradas e avaliadas na execução do planejamento quanto à sua adequação para os diferentes algoritmos de ML e para a interpretabilidade. A variável dependente `y` será mapeada para 0/1.
3. **Tratamento de variáveis específicas:** A variável `pdays` com seu valor especial (-1) será analisada durante a EDA para definir a melhor forma de representá-la.
4. **Exclusão de `duration` para predição:** Para evitar vazamento de dados e garantir um cenário preditivo realista, a variável `duration` será excluída do conjunto de features usado para treinar os modelos preditivos finais, uma vez que os valores dessa variável somente são

obtidos após o contato telefônico. Sua relação com y será investigada na EDA.

5. **Escalonamento de variáveis numéricas:** Variáveis numéricas (`age`, `balance`, etc.) podem necessitar de escalonamento, por exemplo, padronização ou Normalização, para otimizar o desempenho de algoritmos sensíveis à escala. A necessidade e o método serão avaliados conforme EDA.
6. **Investigação de outliers:** A presença e o impacto de outliers em variáveis numéricas serão investigados durante a EDA. A decisão sobre o tratamento (se necessário) será tomada com base nessa análise.
7. **Prevenção de vazamento de dados:** A regra fundamental de ajustar qualquer etapa do pré-processamento apenas nos conjunto de treino e validação e aplicar consistentemente no conjunto de teste será rigorosamente seguida [Roman, 2025a].

2.3 Análise Exploratória de Dados (EDA) Planejada

Após a divisão dos dados completos e a aplicação do pré-processamento ajustado no treino, a Análise Exploratória de Dados será conduzida **exclusivamente no conjunto de treino** para:

- **Análise Univariada:**
  - Verificar a distribuição da variável alvo `y` para identificar possível desbalanceamento de classes.
  - Analisar a distribuição de cada variável numérica (pré e pós-escalonamento) usando diferentes técnicas gráficas como histogramas, gráficos de densidade, box plots, etc., para entender sua forma, tendência central, dispersão e presença de outliers.
  - Analisar a frequência de cada categoria nas variáveis categóricas (pré e pós-codificação, onde aplicável).
- **Análise Bivariada e Multivariada:**
  - Investigar a relação entre cada variável independente e a variável dependente `y`.
  - Calcular e visualizar a matriz de correlação entre as variáveis numéricas para identificar multicolinearidade potencial.
- **Refinamento das Questões de Pesquisa:** Com base nos padrões observados, refinar a questão de pesquisa inicial e formular hipóteses mais específicas a serem testadas na fase de confirmação (modelagem). Por exemplo: 'Clientes com maior saldo (`balance`) e sem empréstimo habitacional são significativamente mais propensos a aderir?', 'O mês do contato (`month`) tem impacto na taxa de sucesso, mesmo controlando por outras variáveis?'.

A EDA guiará possíveis ajustes finos no pré-processamento como tratamento de outliers e transformação de variáveis e na seleção do conjunto de variáveis para a modelagem.

2.4 Plano de modelagem preditiva

O objetivo principal da modelagem preditiva é construir e avaliar modelos de classificação binária para prever a variável dependente `y`, respondendo às questões específicas sobre a viabilidade preditiva. O plano envolve os seguintes itens:

- **Algoritmos candidatos:** Regressão Logística, Árvore de Decisão, Random Forest, SVM, Gradient Boosting (XGBoost/LightGBM/CatBoost), MLP.
- **Divisão dos dados:** Divisão (estratificada ou não) do dataset completo em treino, validação e teste, evitando a amostra utilizada no planejamento.
- **Validação e ajuste:** Validação cruzada nos conjunto de treino e validação para seleção de modelo e otimização de hiperparâmetros.
- **Tratamento de desbalanceamento (se necessário):** Avaliação de técnicas como SMOTE ou ajuste de pesos de classe.
- **Avaliação final:** O desempenho do(s) melhor(es) modelo(s) será medido no conjunto de teste, utilizando Acurácia, Precisão, Recall, F1-Score e AUC, além da análise da matriz de confusão.

2.5 Cronograma de Execução Proposto

A execução da método descrita está planejada para ocorrer em aproximadamente um mês (quatro semanas), conforme as diretrizes do exercício [Roman, 2025b]. Este cronograma visa fornecer uma estrutura para garantir a conclusão das etapas essenciais dentro do prazo.

Um resumo das principais fases e atividades por semana é apresentado na Tabela 2. Este cronograma conciso destaca os focos principais de cada semana.

Um detalhamento completo das atividades planejadas para cada semana, incluindo subtarefas específicas, pode ser encontrado no Apêndice A. A aderência a este cronograma será monitorada, permitindo ajustes conforme necessário durante a execução do planejamento, com quaisquer desvios sendo devidamente justificados no relatório final.

3 Resultados esperados

A execução da método descrita na Seção 2 deverá produzir os seguintes tipos de resultados:

1. **Conhecimentos de domínio:** A EDA fornecerá uma compreensão quantitativa e visual dos fatores associados à adesão de depósitos a prazo. Espera-se identificar perfis de clientes (demográficos, financeiros) e características de campanha (timing, histórico) que se destacam pela maior ou menor probabilidade de conversão. A análise dos indicadores socioeconômicos pode revelar a influência do contexto macroeconômico nas decisões individuais. 2. **Modelos preditivos avaliados:** Serão obtidos resultados de desempenho (Acurácia, F1-Score, AUC, etc.) para os diferentes algoritmos de classificação testados, tanto na validação cruzada quanto no conjunto de teste final. Antecipa-se que modelos baseados em ensemble provavelmente apresentarão

Table 2. Cronograma Conciso da Execução Planejada.

Semana	Foco Principal / Atividades Chave
1	**Preparação final dos dados:** Limpeza do dataset completo, implementação do pré-processamento, divisão treino/validação/teste estratificada.
2	**EDA aprofundada e engenharia de features:** Análise detalhada no conjunto de treino, visualizações, investigação de relações, possível criação de novas features.
3	**Treinamento e ajuste de modelos:** Treinamento dos algoritmos candidatos com validação cruzada, otimização de hiperparâmetros, comparação inicial de modelos, tratamento de desbalanceamento (se necessário).
4	**Avaliação final e relatório:** Treinamento do(s) modelo(s) final(is), avaliação no conjunto de teste, análise dos resultados (métricas, importância), compilação do relatório final.

o melhor desempenho preditivo bruto, uma vez que o problema a ser modelado não se mostra de cunho essencialmente linear, embora modelos mais simples como Regressão Logística possam oferecer maior interpretabilidade. 3. **Identificação de variáveis-chave:** Através de técnicas como análise de importância de variáveis (disponível em modelos como árvores e ensembles), espera-se identificar quais variáveis têm maior poder preditivo no modelo final. 4. **avaliação da viabilidade preditiva:** O desempenho no conjunto de teste indicará a viabilidade prática de usar ML para prever o sucesso da campanha com informações disponíveis realisticamente. Uma comparação entre o desempenho com e sem a variável 'duration' (realizada apenas para fins analíticos, não para o modelo final) quantificará o impacto do vazamento de dados.

A discussão dos resultados abordará as implicações práticas dos achados (por exemplo, como os fatores identificados podem informar estratégias de segmentação), as limitações do estudo (contexto específico, importância dos dados faltantes) e a robustez dos modelos (comparação entre algoritmos, impacto do desbalanceamento).

4 Conclusão e próximos passos

Este artigo propôs uma investigação sobre os fatores que determinam o sucesso das campanhas de telemarketing bancário para depósitos a prazo, utilizando o dataset UCI Bank Marketing. Apresentamos uma método detalhada, baseada nos princípios de experimentação em aprendizado de máquina, que combina análise exploratória de dados e modelagem de classificação supervisionada. O plano inclui etapas específicas de pré-processamento, com atenção a desafios como valores ausentes e potencial vazamento de dados, e define uma estratégia clara para treinamento, validação e avaliação de múltiplos algoritmos de ML.

Espera-se que a execução deste plano forneça insights valiosos sobre o comportamento do cliente neste contexto e avalie realisticamente a capacidade preditiva dos modelos de ML. Os resultados têm o potencial de informar práticas de marketing mais eficientes no setor bancário.

O próximo passo imediato é a execução rigorosa do método aqui descrito, culminando na análise dos resultados e na elaboração de um relatório detalhado das descobertas. Trabalhos futuros poderiam envolver a aplicação de técnicas mais avançadas, a exploração de outros datasets ou a investigação de aspectos não cobertos, como a análise de texto em notas de campanha, se disponíveis.

Declarações

Acknowledgements

Os autores agradecem ao instrutor da disciplina SIN5032, Norton Trevisan, por fornecer o arcabouço e a orientação para este exercício de planejamento.

Disponibilidade dos dados e materiais

O dataset utilizado (Bank Marketing) está publicamente disponível no Repositório de Machine Learning da UCI em `https://archive.ics.uci.edu/dataset/222/bank+marketing`. Todo o código desenvolvido tanto para EDA quanto para a modelagem dos experimentos está disponível em `https://colab.research.google.com/drive/11284NdymZS4KXoOluwiCsYrbAQD_Nbdx?usp=sharing`

Referências

Moro, S., Cortez, P., and Rita, P. (2014a). Bank Marketing Dataset.

Moro, S., Cortez, P., and Rita, P. (2014b). A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.*, 62:22–31.

Roman, N. T. (2025a). Aula 03 – Passos para Experimentação. Material de aula, SIN5032 - Experimentação em Aprendizado de Máquina Supervisionado.

Roman, N. T. (2025b). Primeiro Exercício Prático. Guia do exercício, SIN5032 - Experimentação em Aprendizado de Máquina Supervisionado.

A Cronograma Detalhado da Execução Planejada

A fase de execução (EP2) está planejada para ocorrer ao longo de aproximadamente quatro semanas, com as seguintes atividades principais detalhadas:

- **Semana 1: Preparação dos Dados e Configuração Final**
  - Carregar e realizar verificação final de integridade do dataset.
  - Finalizar limpeza (tratamento definitivo de 'unknown', verificação de outras inconsistências).

  - Implementar e testar o pipeline de pré-processamento planejado (codificação, transformação ‘pdays‘, exclusão ‘duration‘, escalonamento).
  - Realizar a divisão estratificada treino-validação-teste (ex: 70%-15%-15%), garantindo que o conjunto de teste permaneça separado até a avaliação final.

- **Semana 2: EDA Aprofundada e Engenharia de Features**
  - Conduzir Análise Exploratória de Dados detalhada exclusivamente no conjunto de **treino**.
  - Gerar visualizações planejadas (distribuições, relações com alvo, correlações) e outras que surjam como relevantes.
  - Realizar análises estatísticas para abordar as questões exploratórias refinadas.
  - Avaliar a necessidade e implementar (se justificável) engenharia de features (ex: criação de interações, agrupamento de categorias raras). Testar o impacto no conjunto de validação.
- **Semana 3: Treinamento e Ajuste de Modelos**
  - Implementar e treinar os algoritmos de ML candidatos (Reg. Logística, Árvore, RF, SVM, Boosting, etc.) usando validação cruzada nos dados combinados de treino e validação (ou apenas treino, dependendo da estratégia de validação final).
  - Realizar ajuste de hiperparâmetros para os modelos mais promissores usando Grid Search ou Randomized Search com validação cruzada.
  - Comparar modelos com base nas métricas de desempenho escolhidas (Acurácia, F1, AUC, etc.) obtidas durante a validação cruzada.
  - Se o desbalanceamento for um problema significativo (confirmado na EDA e nos resultados preliminares), implementar e avaliar técnicas de tratamento (ex: SMOTE, pesos de classe).
- **Semana 4: Avaliação Final e Relatório**
  - Selecionar o(s) melhor(es) modelo(s) com base nos resultados da validação e ajuste.
  - Treinar o(s) modelo(s) final(is) usando todo o conjunto de treino+validação.
  - Avaliar o desempenho final e imparcial do(s) modelo(s) no conjunto de teste mantido separado.
  - Analisar os resultados finais: calcular métricas finais, gerar matriz de confusão, analisar erros, extrair importância das features (para modelos aplicáveis).
  - Compilar o relatório final, documentando rigorosamente toda a método executada, os achados da EDA, o processo de modelagem, os resultados da avaliação e as conclusões da investigação, incluindo discussão das limitações e possíveis trabalhos futuros.