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.

1. **Exploração**

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 instuição 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.

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

1. **Tratamento de valore desconhecidos:** Nos casos em que foram identificados valores desconhecidos ou ausentes em atributos qualitativos, foram realizadas análises de frequências para avaliar a viabilidade de inserção com técnicas de imputação em lacunas. Conforme apresentado na tabela 1 as variáveis job, education, contact e poutcome apresentam dados desconhecidos.

Dentro essas variáveis decidiu-se pela exclusão do atributo poutcome, considerando que há um grande volume de dados desconhecidos os quais pode interver no mal funcionamento do modelo de aprendizado. Para as demais variáveis, optou-se pela inserção utilizando o valor modal de cada atributo correspondente.

1. **Codificação de variáveis categóricas**: As features textuais (nominais e ordinais) foram convertidas para um formato numérico, para as variáveis nominais utilizou-se a técnica *One-Hot Encoding*, enquanto as ordinais aplicou-se *Label Encoding*. A variável dependente y foi devidamente transformada em valores binários (0/1).
2. **Tratamento de variáveis específicas:** A variável pdays com seu valor especial (-1) será analisada durante a EDA para definir a melhor forma de representála
3. **.Exclusão de duration para predição:** Para evitar vazamento de dados e garantir um cenário preditivo realista, a variável *duration* foi 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

.

1. **Escalonamento de variáveis numéricas:** Ao analisar a distribuição das variáveis *age*, *balance*, *campaign*, *pdays*, *previous* e *day*, foi identificado um padrão assimétrico. Considerando essas características, optou-se pela aplicação da técnica de normalização *MinMaxScaler, com o propósito de* reduzir o impacto de escalas distintas proporcionando um melhor desempenho no algoritmo.
2. **Identificação e analise de impacto de outliers:** Buscar técnicas para diminuir os impactos com LOG entre outros e analisar os impactos.
3. **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].
4. **Materiais e métodos**
   1. **Base de dados**

A base de dados será única, previamente coletada e devidamente organizada, sua divisão foi realizada por meio de uma separação estratificada, garantindo a preservação da proporção das classes nos conjuntos de treino e teste

* 1. **Ténicas de aprendizado**

Foram aplicadas técnicas de aprendizado supervisionado, escolhidas por sua reconhecida eficácia tanto em dados balanceados quanto em homogenios. Entre os modelos utilizados destacam-se: *XGBoost, Random Forest e Regresão Logística (Descrever o modelo escolhido)*

* 1. **Avaliação dos modelos:**

A avaliação dos modelos são realiadas de forma individual por meio de técnicas vinculadas a natureza de classificação binárias, são elas:

- Acurácia: representando a proporção das previsões corretas;

- Recall: Identifica a performance das classes minoritárias;

- F1 Score: Oferece um parâmetro de equilibrio entre o a precisão e o recall;

Com essa abordagemé possivel obter uma avaliação rigorosa e confiável, apoiando na decisão do modelo de aprendizados de maquina mais apropriado.

1. **Resultados**
2. **Discusão de resultados**