

# Business Understanding:

## Bank Marketing

**Discentes:**

- \* Arthur Bezerra Calado
- \* Gabriel D'assumpção de Carvalho
- \* Pedro Henrique Sarmento de Paula

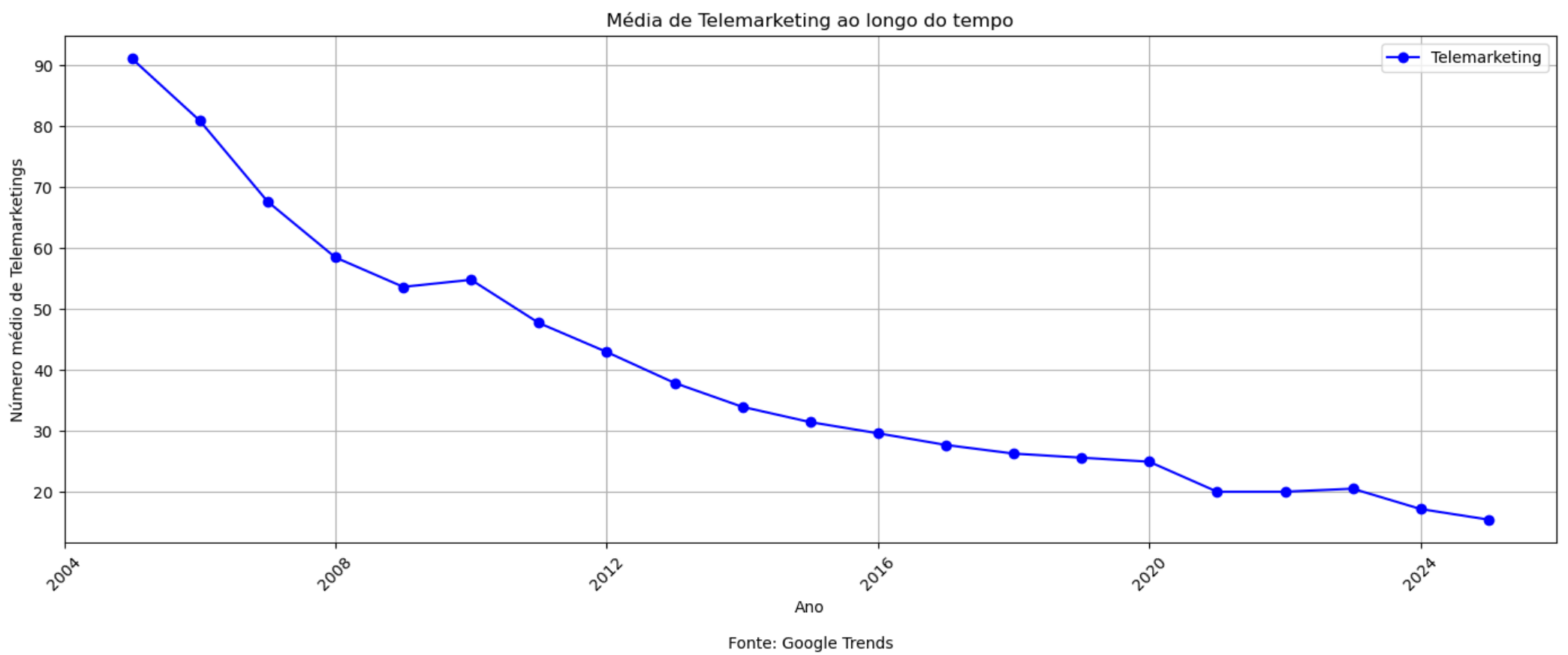
**Data:** 04/07/2024

### Introdução

O marketing é um conjunto de técnicas e métodos aplicados ao estudo das necessidades dos mercados, funcionando como um mecanismo essencial para potencializar os negócios. As organizações utilizam estratégias variadas para se comunicar diretamente com os clientes, seja por meio de redes sociais, outdoors ou telemarketing, visando alcançar objetivos específicos. No contexto do setor bancário, um dos desafios mais significativos é garantir a eficácia das campanhas de marketing para atrair novos clientes e reter os existentes. A base de dados disponível no UC Irvine Machine Learning Repository se concentra no marketing direto realizado por um banco português, com o propósito de avaliar o sucesso das campanhas de depósito a prazo.

O telemarketing, como forma de marketing direto, pode ser percebido como invasivo. Por isso, é crucial identificar se o cliente potencial realmente se enquadra no perfil desejado pela empresa, porque nos dias atuais os clientes "adoram compartilhar histórias, boas e ruins, sobre suas experiências de consumo" (KOTLER, 2017, p. 21). O objetivo do marketing é maximizar a retenção de clientes e incentivar a recompra de produtos ou serviços da empresa. Assim, as técnicas de marketing visam selecionar o melhor conjunto de clientes, aqueles mais propensos a se interessar pelo produto oferecido.

O telemarketing é amplamente reconhecido como um método invasivo de marketing, o que tem contribuído para uma diminuição gradual de sua popularidade ao longo dos anos, conforme indicado por uma análise do termo "telemarketing" no Google Trends.



Os dados apresentados ilustram claramente que o interesse pelo telemarketing tem declinado. Isso indica uma redução no número de pessoas interessadas em entender o funcionamento deste método ou em se envolver com ele. Esse cenário sugere que o telemarketing está perdendo espaço no mercado devido à sua natureza intrusiva.

O desafio de negócio aqui é identificar quais clientes têm maior probabilidade de aderir a um depósito bancário a prazo após uma campanha de marketing direto. Essa análise é crucial para que o banco possa direcionar eficientemente seus recursos e estratégias de marketing, concentrando-se nos clientes mais propensos à conversão. No setor bancário, o sucesso de campanhas de marketing não apenas amplia a base de clientes, mas também impulsiona significativamente os lucros.

Considerando o objetivo do marketing, é evidente que a análise de dados é uma ferramenta fundamental para a tomada de decisões, facilitando a compreensão do perfil de cada cliente. Com o avanço de métodos estatísticos e computacionais, é possível utilizar ferramentas de inteligência artificial que combinam conhecimentos de estatística multivariada e computação para a detecção de padrões. Nesse contexto, os modelos de classificação desempenham um papel crucial, permitindo obter conhecimento explicativo e preditivo a partir de dados brutos.

Existem diversos tipos de classificadores, como K-NN, Suport Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Neural Network (NN) e outros. Cada um desses modelos apresenta vantagens e desvantagens. Por exemplo, SVM e NN são robustos e capazes de lidar com dados lineares e não lineares, porém podem perder a capacidade explicativa. Portanto, a escolha do modelo depende do objetivo do banco: se deseja um modelo explicativo para entender por que certos clientes são mais propensos a fazer um depósito, ou se precisa apenas de um modelo preditivo que forneça uma resposta de sucesso ou fracasso ao inserir as características do cliente.

A análise dos dados coletados entre 2008 e 2010, durante uma das maiores crises financeiras dos Estados Unidos, permite identificar padrões e características dos clientes que levaram ao sucesso das campanhas. Essa análise não só ajudará a melhorar as estratégias atuais, mas também fornecerá insights valiosos para mitigar os impactos de crises financeiras futuras, garantindo a sustentabilidade e o crescimento do banco a longo prazo. O conjunto de dados a ser analisado contém 16 variáveis categóricas e numéricas, com 41.188 observações.

Nome da variável	Papel	Tipo	Demográfico	Descrição	Unidades	Valores ausentes
idade	Característica	Inteiro	Idade			Não
trabalho	Característica	Categórico	Ocupação	Tipo de emprego (categórico: 'Admin.', 'Blue-collar', 'Entrepreneur', 'Householdant', 'Management', 'Aposentado', 'Autônomo', 'Serviços', 'Estudante', 'Técnico', 'Desempregado', 'Desconhecido')		Não
conjugal	Característica	Categórico	Estado civil	estado civil (categórico: 'divorciado', 'casado', 'solteiro', 'desconhecido'; nota: 'divorciado' significa divorciado ou viúvo)		Não
educação	Característica	Categórico	Nível de escolaridade	(categórico: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'analfabeto', 'professional.course', 'university.degree', 'unknown')		Não
inadimplência	Característica	Binário		tem crédito inadimplente?		Não
equilíbrio	Característica	Inteiro		saldo médio anual	Euros	Não
habitação	Característica	Binário		Tem crédito habitação?		Não
empréstimo	Característica	Binário		Tem empréstimo pessoal?		Não
contato	Característica	Categórico		tipo de comunicação de contato (categórico: 'celular', 'telefone')		Sim
day_of_week	Característica	Data		último dia de contato da semana		Não
mês	Característica	Data		último mês do ano de contato (categórico: 'jan', 'feb', 'mar', ..., 'nov', 'dec')		Não
duração	Característica	Inteiro		Duração do último contato, em segundos (numérico). Observação importante: esse atributo afeta fortemente o destino de saída (por exemplo, se duration=0 e y='no'). No entanto, a duração não é conhecida antes de uma chamada ser realizada. Além disso, após o fim da chamada y é obviamente conhecido. Assim, esse insumo só deve ser incluído para fins de benchmark e deve ser descartado se a intenção for ter um modelo preditivo realista.		Não
campanha	Característica	Inteiro		número de contatos realizados durante esta campanha e para este cliente (numérico, inclui último contato)		Não
pdays	Característica	Inteiro		número de dias que se passaram após o cliente ter sido contatado pela última vez a partir de uma campanha anterior (numérico; -1 significa que o cliente não foi contatado anteriormente)		Sim
anterior	Característica	Inteiro		número de contatos realizados antes desta campanha e para este cliente		Não
presultado	Característica	Categórico		resultado da campanha de marketing anterior (categórico: 'fracasso', 'inexistente', 'sucesso')		Sim
y	Alvo	Binário		O cliente subscreveu um depósito a prazo?		

Conforme observado na tabela de variáveis do nosso banco de dados, as variáveis "contato", "pdays" e "presultado" apresentam observações faltantes. Antes de proceder com a seleção do classificador, é essencial realizar uma análise exploratória para tratar dados ausentes, identificar possíveis valores atípicos, ajustar a distribuição das variáveis para melhor simetria e transformar variáveis categóricas em variáveis dummy. Além disso, considerando que os dados são do período entre 2008 e 2010, pode haver desafios na aplicação de modelos para classificar novas observações com precisão.

Com mais de 41 mil observações, a busca por um modelo otimizado entre os classificadores mencionados no parágrafo anterior demanda recursos adequados. Isso inclui hardware e software robustos, além de suporte de TI para instalação de pacotes, manutenção preventiva e atualizações de software necessárias para análise de dados. Recursos adequados são cruciais para lidar com a quantidade volumosa de dados e garantir eficiência e precisão nas análises.

Além dos recursos mencionados, é fundamental decidir se o modelo será ajustado uma única vez com os dados de "bank marketing" ou se será desenvolvido um Data Warehouse. Um Data Warehouse permitirá armazenar todas as informações estruturadas dos clientes, facilitando atualizações e melhorias contínuas do modelo ao longo do tempo. Isso reduzirá a dependência de dados de terceiros e aumentará a precisão na classificação dos clientes recorrentes da empresa, proporcionando uma base sólida para estratégias futuras de marketing e análise de clientes.

Em resumo, este estudo visa utilizar o software Python para coleta de dados, análise e estimação de modelos de classificação, com o objetivo de maximizar os depósitos a prazo captados pelo telemarketing do banco. Portanto, é crucial definir métricas claras para avaliar o desempenho do projeto.

## Objetivos do Projeto

O objetivo principal deste projeto é explorar o dataset "Bank Marketing" para desenvolver um modelo preditivo robusto e preciso, capaz de identificar quais clientes têm maior probabilidade de subscrever um depósito a prazo. Isso visa aumentar a eficácia das campanhas de marketing direto do banco, melhorar a gestão de liquidez e a rentabilidade através de decisões mais informadas e personalizadas.

Para alcançar esse objetivo, o projeto se desdobra em objetivos específicos e metas mensuráveis, além de critérios claros para medir o sucesso do projeto.

### Objetivos Específicos e Metas Mensuráveis:

- Exploração e Compreensão dos Dados:** Realizar uma análise detalhada para compreender a estrutura dos dados e identificar padrões que possam influenciar a subscrição de depósitos a prazo.
- Preparação dos Dados:** Tratar valores ausentes, outliers e realizar transformações necessárias para garantir a qualidade e consistência dos dados.
- Modelagem Preditiva:** Desenvolver e comparar diferentes modelos de machine learning, como árvores de decisão, random forest, SVM e redes neurais, para prever com precisão a probabilidade de subscrição.
- Avaliação de Desempenho:** Avaliar o desempenho dos modelos utilizando métricas como acurácia, precisão, recall e F1-score, assegurando que o modelo selecionado seja capaz de generalizar bem para novos dados.
- Implementação Prática:** Implementar o modelo preditivo final em um sistema que possa ser integrado às estratégias de marketing existentes do banco, otimizando a alocação de recursos e melhorando a taxa de conversão das campanhas.

### Critérios de Sucesso:

O sucesso do projeto será medido pela capacidade do modelo em alcançar uma acurácia mínima de 80% e um F1-score superior a 0.75 na previsão de subscrição de depósitos a prazo. Além disso, espera-se uma melhoria significativa na taxa de conversão das campanhas de marketing direto, demonstrando assim o impacto positivo do modelo na estratégia de negócio do banco.

## Contexto do Negócio

O setor bancário enfrenta desafios significativos relacionados à atração e retenção de clientes, especialmente em um mercado competitivo onde a personalização das interações e a eficiência das campanhas de marketing são cruciais. A subscrição de depósitos a prazo representa não apenas uma oportunidade de estabilidade financeira para os clientes, mas também uma estratégia essencial para os bancos gerenciarem suas reservas de forma eficaz, otimizando o uso de recursos e aumentando a rentabilidade.

Neste contexto, é importante abordar os desafios atuais do setor e identificar os principais stakeholders e suas expectativas.

## Desafios Atuais do Setor:

- Segmentação de Clientes: Necessidade de segmentar de forma precisa os clientes com potencial interesse em depósitos a prazo, para maximizar a eficácia das campanhas direcionadas.
- Eficiência das Campanhas de Marketing: Melhoria na taxa de conversão das campanhas de marketing direto, garantindo um retorno maior sobre o investimento em marketing.
- Competitividade do Mercado: Capacidade de diferenciar-se no mercado através de estratégias de marketing mais eficazes e orientadas por dados.

## Principais Stakeholders e Expectativas:

- Departamentos de Marketing: Esperam que o projeto forneça insights acionáveis para melhorar a segmentação e personalização das campanhas, aumentando assim a eficácia na aquisição e retenção de clientes.
- Executivos do Banco: Buscam estratégias que não apenas aumentem a subscrição de depósitos a prazo, mas também fortaleçam a fidelidade dos clientes e impulsionem o crescimento sustentável da instituição financeira.
- Analistas de Dados: Focam na construção de modelos analíticos robustos que possam traduzir dados em informações estratégicas para tomada de decisão eficaz e otimização de processos.

## Questões de Negócio

Este projeto visa responder a questões chave que são fundamentais para o sucesso das estratégias de marketing do banco:

- Características Determinantes: Quais características dos clientes são mais indicativas de uma alta probabilidade de subscrever depósitos a prazo?
- Impacto dos Canais de Comunicação: Como diferentes canais de comunicação (telefone, celular) influenciam a taxa de sucesso das campanhas de marketing direto?
- Fatores Sazonais: Existe alguma sazonalidade ou período específico do ano que influencia significativamente na decisão dos clientes em subscrever depósitos a prazo?

## Justificativa para a Importância de Cada Questão:

Compreender essas questões é crucial para otimizar as estratégias de marketing do banco, garantindo uma abordagem mais direcionada e eficiente:

- Características Determinantes: Identificar as características-chave dos clientes permite uma segmentação mais precisa e personalizada, aumentando a eficiência das campanhas de marketing direto e melhorando a taxa de conversão.
- Impacto dos Canais de Comunicação: Compreender como diferentes canais afetam as taxas de conversão ajuda na alocação eficiente de recursos de marketing e na otimização das estratégias de comunicação, garantindo maior retorno sobre o investimento.
- Fatores Sazonais: Reconhecer padrões sazonais permite ajustar o timing das campanhas para maximizar o impacto e a resposta dos clientes durante períodos específicos do ano, aumentando assim a eficácia das estratégias de marketing sazonais.

## Fontes de Dados

O dataset "Bank Marketing" consiste em registros detalhados de campanhas de marketing direto realizadas por um banco português. Composto por 45.211 registros de clientes e 17 atributos, incluindo informações demográficas, financeiras e resultados de campanhas anteriores, os dados estão estruturados em formato CSV, facilitando a análise e manipulação através de ferramentas de ciência de dados.

Para garantir a eficácia do modelo preditivo, é crucial analisar a qualidade e a integridade dos dados, assim como considerar as limitações existentes.

## Qualidade e Integridade dos Dados:

Embora os dados sejam relativamente completos, é essencial realizar uma análise de qualidade para identificar e tratar valores ausentes, outliers e possíveis erros de entrada que possam comprometer a precisão dos modelos preditivos. Garantir a integridade dos dados é crucial para obter insights confiáveis e para a construção de modelos robustos.

## Limitações dos Dados Existentes:

As limitações incluem a exclusão de outros canais de comunicação além do telefone, o que pode limitar a generalização dos resultados para outras formas de interação cliente-banco. Além disso, vieses de amostragem podem estar presentes devido ao método de coleta dos dados, sendo importante considerar esses aspectos durante a análise e interpretação dos resultados.

## Restrições e Assunções

Este projeto enfrenta várias restrições e baseia-se em algumas premissas fundamentais para sua execução:

## Restrições de Tempo e Recursos:

- O projeto deve ser concluído dentro de um prazo de 6 meses, com recursos limitados de hardware e software, o que pode influenciar a complexidade e o escopo das análises realizadas.

## Dependência de Dados Históricos:

- A análise baseia-se em dados históricos que podem não capturar totalmente o comportamento atual dos clientes, exigindo adaptações e ajustes ao longo do projeto.

## Assunções de Representatividade dos Dados:

- Assume-se que os dados coletados são representativos da população alvo e que as métricas atuais são adequadas para avaliar o desempenho das campanhas de marketing.

## Abordagem Proposta

A abordagem adotada seguirá a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), estruturada nas seguintes fases:

## Entendimento do Negócio:

Nesta fase, serão definidos detalhadamente os objetivos específicos do projeto, os requisitos necessários para atingi-los e os critérios de sucesso que serão utilizados para avaliar o impacto das análises. É essencial alinhar as metas de negócio com as capacidades analíticas disponíveis, garantindo que todas as partes interessadas estejam alinhadas quanto aos resultados esperados.

## Entendimento dos Dados:

Aqui, será realizada a coleta inicial dos dados do dataset "Bank Marketing". Será feita uma descrição detalhada dos dados disponíveis, explorando suas características e identificando padrões iniciais e relacionamentos que possam ser relevantes para os objetivos do projeto. Esta fase é fundamental para uma compreensão profunda do contexto dos dados antes de qualquer análise ou modelagem.

## Preparação dos Dados:

Nesta etapa, os dados serão preparados para análise. Isso inclui a limpeza dos dados para lidar com valores ausentes, tratamento de outliers e possíveis erros de entrada. Além disso, serão realizadas transformações nos dados, como normalização ou codificação de variáveis categóricas, para criar um conjunto final de dados de alta qualidade e consistência, pronto para ser utilizado nas etapas subsequentes.

## Modelagem:

Aqui serão desenvolvidos, selecionados e ajustados diferentes modelos preditivos. Serão exploradas técnicas como árvores de decisão, random forest, SVM (Support Vector Machines) e redes neurais, com o objetivo de encontrar o modelo que melhor se adeque aos dados e maximize as métricas de desempenho estabelecidas. Esta fase envolve experimentação cuidadosa e avaliação rigorosa para garantir a escolha do modelo mais eficaz.

## Avaliação:

Após o desenvolvimento dos modelos, será realizada uma avaliação detalhada do desempenho de cada um deles. Métricas como acurácia, precisão, recall, F1-score e curvas ROC serão utilizadas para comparar e selecionar o modelo final. Ajustes serão feitos conforme necessário para melhorar a precisão e a robustez das previsões, garantindo que o modelo escolhido seja capaz de generalizar bem para novos dados.

## Implementação:

Na fase de implementação, o modelo preditivo final será integrado em um sistema de produção. Será realizado um monitoramento contínuo para acompanhar o desempenho do modelo em ambiente real e garantir sua eficácia a longo prazo. A implementação bem-sucedida garantirá que os insights obtidos sejam aplicados de maneira prática e contribuam diretamente para as estratégias de marketing do banco.

## Referência Bibliográfica

KOTLER, Philip; KARTAJAYA, Hermawan; SETIAWAN, Iwan. **Marketing 4.0: do tradicional ao digital** . Rio de Janeiro: Sextante, 2017.

Conroy, M. (2023, May 17). Business understanding for data professionals. DataLabNotes. Retrieved July 04, 2024, from <https://datalabnotes.com/business-understanding-for-data-professionals/>