# \*\*Introdução\*\*

O presente código tem como objetivo o desenvolvimento de um algoritmo de machine learning capaz de prever o resultado da campanha de marketing de um banco, identificando se o usuário realizou o depósito ou não após a campanha.

Portanto o algoritmo deve ser capaz de classificar de maneira binária o resultado da campanha.

Com base nisto, será definido os seguintes objetivos:

- Identificar quais características são mais importantes para a realização da campanha de marketing, visando modificar futuramente a campanha focando nos aspectos que geram mais resultados.

- O modelo desenvolvido só será aceito com o valor de acurácia maior que 80%;

# \*\*Visualização e pré-processamento \*\*

<img src="/Bank\_marketing\_dataset/Project\_prints/data\_desc.png ">

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_01.png ">

Inicialmente não se pode notar a ausência de dados ou inconsistência em alguma variável.

Com base nisto, será feito uma análise exploratória dos dados e será analisado se alguma feature possui uma tendência ou correlação com a target (deposit).

# \*\*Análise Exploratória de Dados (EDA)\*\*

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_02.png ">

Analisando as variáveis categóricas temos:

- 'Poutcome' e 'contact' apresentam mais categorias do que as listadas pela referência do dataset, sendo para o 'poutcome' adicionado duas novas categorias e para a variável 'contact' existe mais uma nova categoria.

- A maioria dos contatos é feito pelo celular.

- O dataset parece balanceado em relação a variável target 'deposit'.

- A maioria dos entrevistados não possui personal loan.

Analisando as variáveis 'poutcome' e 'contact' vemos que para 'poutcome' existem duas novas categorias são 'unknown' e 'other', já para a variável 'contact' a nova categoria é 'unknown'.

Isto pode ter sido gerado pela perda de dados no armazenamento do dataset em seu banco de dados original, mas de alguma forma ao invés de ser identificado como um dado ausente ele foi categorizado desta forma.

Portanto, para uma melhor classificação destes dados, eles serão convertidos em formato 'NaN' e serão tratados como dados ausentes.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_03.png ">

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_04.png ">

Analisando os dados numéricos tem-se:

- A média dos entrevistados está entre 30 a 50 anos.

- O valor de 'balance' está concentrado entre 0 e 4.000, porém possui pontos de valores muito elevados, chegando à valores de 80.000.

- A duaração das ligações possui em média entre 200 a 600 segundos.

- As variáveis 'duration' e 'balance' possuem valores muito distante da distribuição dos dados mas ainda podem ser considerados valores plausíveis para essas variáveis.

- Foram realizadas mais de 60 campanhas, mas a maioria dos dados pertence as campanhas iniciais (1 ao 5).

Com base nessas vizualizações será feito uma investigação para verificar se os valores elevados de 'balance' e 'duration' tem uma relação maior com o 'deposit'

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_05.png ">

Pelo gráfico de boxplot não se observa uma relação entre os valores altos de balance e a taxa de depósito. Com base nisto pode ser realizado a remoção destes outliers, já que eles não agregam em alguma classe em específico. Mas primeiro é necessário verificar o quanto de informação será perdida com a remoção destes dados

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_06.png ">

Observando a distribuição nota-se que será perdido cerca de 7% dos dados ao realizar a remoção destes outliers. Eu considero isto uma quantidade relevante de dados para ser retirada da análise. Com base nisto será removido os outliers que passam de 20.000 de balance.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_07.png ">

Obsevando o gráfico vemos que a média e o range de duração da ligação para as pessoas que fizeram 'deposit' é maior do que as pessoas que não fizeram o 'deposit'. Com base nisto não será feito a remoção dos outliers já que estes outliers possuem uma certa relação com a variável target.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_08.png ">

Analisando os gráficos de densidade categorizados pela variável target vemos que não existe uma distinção de distribuição expressiva entre eles. Isto indica que estes atributos podem ter pouca correlação com a variável target ou pouco impacto na hora de implementar o algoritmo de machine learning.

Analisando os plots de pares de variáveis tem-se as seguintes obeservações:

- No plot de 'balance' e 'duration' percebe-se que quanto menor é o balance maior é o valor da duração da chamada.

- Nota-se que no gráfico 'balance' e 'campaign', a tendencia do decorrer das campanhas é focar em valores de 'balance' menores.

- Nota-se que no plot 'campaign' e 'age' com o decorrer das campanhas o foco foi dado para idades entre 25 e 60 anos.

- Existe uma correlação entre 'campaign' e 'duration' onde com o passar das campanhas a tendência é a diminuição da duração da chamada.

- No plot de 'campaing' e 'pdays' nota-se que as campanhas finais foram direcionadas a clientes novos, com pdays = 0.

- Existe uma correlação entre 'previous' e 'duration' onde com o aumento de previous corre a diminuição da duração da chamada.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_09.png ">

Observando esses gráficos é possível fazer algumas inferências:

- job:

\* Blue-collar tem maior tendência a ter rejeição da proposta de deposit;

\* Student e retire tem maior tendência a ter aceitação da proposta de deposit;

- education:

\* Conforme o nível de educação vai subindo maior é a tendencia a realizar o depósito;

- housing:

\* Se a pessoa já tiver crédito de habitação ela tem uma maior chance de não realizar o depósito;

- contact

\* Ligações por celular tem uma maior taxa de aceitação da proposta de deposit;

- month

\* Campanhas feitas no mês de 'may' tem maior tentencia de resultarem em fracasso;

\* Campanhas feitas no mês de 'sep', 'mar' e 'abr' tem maior tentencia de resultarem em sucesso;

# \*\*Feature Engineering\*\*

## Tratamento dos dados ausentes

Antes de começar aplicar técnicas de tratamento de dados ausente é necessário saber qual é o tipo de missing data está sendo tratada, para que assim a técnica aplicada seja a mais eficiente e evite aumentar o viés dos resultados futuramente.

Existem três tipos de dados ausentes:

- Missing completely at random (MCAR):

Nesse caso, não há relação entre os dados faltantes e quaisquer outros valores observados ou não observados em determinado conjunto de dados.

Ou seja, os valores faltantes são completamente independentes de outros dados, sem tipo de padrão.

No caso do MCAR, os dados podem estar faltando devido a erro humano, alguma falha do sistema/equipamento, perda de amostra, etc...

- Missing at random (MAR):

Neste caso, MAR significa que o motivo dos valores faltantes pode ser explicado por variáveis sobre as quais você tem informações completas, pois existe alguma relação entre os dados faltantes e outros valores/dados.

Neste caso, os dados não faltam para todas as observações. Está faltando apenas nas subamostras dos dados e há algum padrão nos valores ausentes.

Por exemplo, se você verificar os dados de uma pesquisa, poderá descobrir que todas as pessoas responderam ao seu 'Sexo', mas os valores de 'Idade' estão faltando principalmente para as pessoas que responderam ao seu 'Sexo' como 'feminino'.

- Missing not at random (MNAR):

Se houver alguma estrutura/padrão nos dados faltantes e outros dados observados não puderem explicá-la, então a falta não é aleatória. Se os dados em falta não se enquadrarem no MCAR ou MAR, então podem ser categorizados como MNAR.

Isso pode acontecer devido à relutância das pessoas em fornecer as informações necessárias. Um grupo específico de pessoas pode não responder a algumas perguntas de uma pesquisa. Por exemplo, suponha que o nome e o número de livros vencidos sejam solicitados na enquete de uma biblioteca. Portanto, a maioria das pessoas que não têm livros vencidos provavelmente responderá à enquete. Pessoas com mais livros atrasados têm menos probabilidade de responder à enquete.

Então, neste caso, o valor faltante da quantidade de livros vencidos depende das pessoas que têm mais livros vencidos.

Com base nessas definições será feita uma avaliação dos missing datas por meio de um teste de hipótese para verificar se o mesmo pertence a categoria MCAR. Caso não pertença ele poderá ser MAR ou MNAR.

Como o dataset utilizado é uma parcela de um dataset maior, que pode ser obtido neste link: [Bank marketing dataset](https://archive.ics.uci.edu/ml/datasets/bank+marketing) , todos os dados que não forem MCAR será considerado MAR visto que não está sendo utilizado todos os dados para comprovar se o missing data seria do tipo MNAR ou foi uma amostra do dataset que pegou uma grande quantidade de missing data daquela feature.

## Teste de hipótese MCAR

Para comprovar se o missing value é do tipo MCAR ou não será feito o teste de chi², que é realizado entre duas variáveis categóricas, onde tem a seguinte hipótese nula:

- H0: (null hypothesis) the two variables are independent, in other words, the missing values of the first variable do not depend on the second variable.

- H1: (alternative hypothesis) The two variables are not independent .

Se para todas as variáveis categóricas a hipotese nula se provar verdadeira indica que o tipo de missing value é o MCAR, caso uma variável invalide a hipotese nula será considerado como MAR.

Como parâmetro para validação da hipótese nula o p-value deve ser maior que 0.05.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_10.png ">

Analisando a tabela de p-value obtida vemos que em todos os casos ocorre a invalidação da hipótese nula, comprovando assim que o tipo de missing value de todas as features é do tipo MAR.

Como este teste só é valido para duas variáveis categóricas, se todas as hipoteses nulas se provarem válidas deve ser feito o teste com as variáveis quantitativas por meio do T-test para comprovar que o missing value é do tipo MCAR.

Caso o dataset utilizado fosse o completo, a feature 'poutcome' seria classificada como MNAR, visto a sua ausência elevada e os baixos valores de p-value e caso o t-test também apresenta-se baixa correlação.

## Tratamento dos dados ausentes

Para o tratamento de dados faltantes do tipo ‘MAR’, são utilizados dois métodos comuns: remoção dos dados faltantes e imputação dos dados faltantes.

A remoção de dados faltantes é uma técnica simples, mas muito radical e geralmente é empregada nos casos em que não há impacto significativo no modelo e quando a quantidade de dados faltantes é inferior a 1% dos dados. Neste caso, é removida a linha onde não há dados ou é removida toda a coluna, que não possui modelo.

A imputação de dados faltantes é um dos métodos mais utilizados, diversas técnicas que podem ser utilizadas e existem para tratar os dados faltantes. As técnicas mais comuns são a substituição da média em características numéricas e a substituição por características categóricas. Deve ser aplicado quando a quantidade de dados faltantes for baixa em relação ao conjunto de dados.

Outra forma de técnica é a implementação de recursos para imputação desses dados. as organizações mais utilizadas para esse fim são: imputação KN, imputação RandomForest e MICE. Esses tipos de animais são geralmente aplicados em casos onde há uma grande parcela de dados faltantes em relação ao conjunto de dados. Garantem maior precisão na aplicação de valores implementando técnicas de substituição pela média ou por uma moda, mas devido ao seu maior custo computacional geralmente não compensam alguns dados faltantes.

Como existe uma grande quantidade de dados faltantes que devem ser tratados neste caso, será implementada uma técnica MICE para imputar os dados faltantes. MICE significa algoritmo de imputação multivariada de equação encadeada, uma técnica pela qual podemos imputar facilmente valores ausentes em um conjunto de dados observando dados de outras colunas e tentando estimar a melhor previsão para cada valor ausente.

Como o modelo a ser aplicado é baseado em equações encadeadas, é necessário realizar a codificação dos traços categóricos.

## Codificando as variáveis

Para realizar o enconding das variáveis categoricas é necessario saber algumas informações, a primeira e mais importante é se a variável é ordinal ou nominal.

Caso ela seja ordinal o enconding será levando em consideração a hierarquida desta variável, sendo a classe mais baixa classificada com o menor valor como o 1, e a mais alta com o maior valor, para que no momento de implementação do algoritmo o maior valor tenha mais impacto no modelo.

Caso ela seja nominal o enconding será feito utilizando o one-hot-enconding, que simplificando seria a transformação das categorias em novas colunas do dataframe de forma binária.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_11.png ">

No total temos 10 variáveis categóricas.

Analisando o gráfico de distribuição de dados feito anteriormente nota-se que as variáveis 'job', 'marital', 'education' e 'month' possuem mais de duas categorias.

No caso da variável 'job' os tipos de emprego não possuem uma ordem hierárquica, sendo classificado como uma variável nominal. Neste caso foi classificado como não-hierárquico pois este dataset foi feito com base em pessoas não relacionadas, mas caso fosse feito com base em uma empresa em específico poderia ser considerado hieráriquico com base na estrutura de classificação de cargo da empresa.

Analisando a variável 'marital' não é possível observar nenhuma hierarquia em relação as classes, sendo classificada como uma variável nominal.

Para a variável 'education' o nível de escolaridade possui uma ordem hierárquica onde a classe 'tertiary' é acima de 'secondary' por assim em diante.

No caso da variável 'month' ocorre um caso mais específico, onde é uma variável que possui ordem, onde o mês de 'feb' vem antes de 'may', possuindo assim uma hierarquia, porém é ciclico, onde após o mês de 'dec' vem o mês de 'jan' novamente. Com isso será feito um tratamento específico para a feature 'month' visando tornar os dados mais ciclicos.

As demais variáveis serão convertidas em formato binário de 0 e 1.

# \*\*Correlação\*\*

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_12.png ">

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_13.png ">

Observando as matrizes de correlação não existem que tenha uma correlação expressiva com outra. A que possui a maior correlação é a variável 'marital\_married' com 'marital\_single' possuindo 0.78 de correção em pearson e spearman. Como o valor de corte inicialmente estabelecido foi de 0.85 para remoção de features não foi atingido nenhuma variável será removida.

# \*\*Implementação do modelo\*\*

Para a seleção do modelo a ser aplicado no desenvolvimento da solução será utilizado o pacote \*\*'LazyClassifier'\*\* onde ele realiza a aplicação de diversos modelos de classificação e os ranqueia baseado nas métricas selecionadas. A métrica utilizada para verificar o desempenho do modelo será o \*\*acurácia\*\*.

Após isto será realizado o tunning dos hiperparâmetros do melhor modelo que obteve a melhor performance durante a aplicação do LazyClassifier.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_14.png ">

Observando os resultados do LazyClassifier, os melhores modelos foram o \*\*RandomFlorestClassifier\*\*, \*\*LGBMClassifier\*\* e o \*\*XGBClassifier\*\* que tiveram pontuações iguais. Portanto será selecionado o XGBClassifier para realizar a otimização.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_15.png ">

Observando o histórico de otimização é possível ver uma melhora na performance da acurácia do modelo, alcançando cerca de 86,6% de acurácia.

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_16.png ">

Avaliando os dados de teste nota-se um resultado similar ao encontrado na otimização. Com base nesse modelo, serão investigadas quais são as características mais importantes para a construção desse modelo.

#\*\*Conclusão\*\*

<img src="/Bank\_marketing\_dataset/Project\_prints/img\_17.png ">

Olhando para o gráfico, as características mais importantes são 'housing', 'duration' e 'poutcome' seguidas por 'marital\_married' e 'month'.

Com base nisso, temos algumas orientações que podem ser dadas para serem desenvolvidas nas próximas campanhas de marketing:

- Procure aumentar o tempo dos atendimentos, oferecendo novos planos e vantagens;

- Concentre-se na divulgação das campanhas para estudantes e desempregados, que tendem a aceitar mais as propostas de depósito;

- Foco em pessoas que não possuem crédito habitação;

- Realizar campanhas por volta dos meses de setembro, abril e março, que são os meses em que as campanhas anteriores tiveram mais sucesso;

- Concentre-se em pessoas que não são casadas, pois são mais propensas a fazer o depósito.