# Previsão do sucesso de chamadas de telemarketing sobre vendas de depósito bancário a longo prazo

E. S. ITO *
*Ciência da Computação - Mestrado
E-mail: e159086@dac.unicamp.br
T. E. NAZATTO †
†Ciência da Computação - Mestrado
E-mail: t074388@dac.unicamp.br

***Resumo* – Aqui avaliamos a eficiência da promoção de vendas de depósitos bancários a longo prazo para um banco português por meio de chamadas de telemarketing. Os dados utilizados foram doados e agora publicamente disponíveis para pesquisa no sítio UCI Machine Learning Repository [1]. Os dados possuem 20 variáveis (features) que potencialmente poderiam influenciar a subscrição do cliente ao programa de depósito bancário. Os mesmos serão utilizados submetidos às fases da abordagem do aprendizado de máquina: extração dos dados, preparação dos dados, seleção de features, treinamento dos dados utilizando Regressão Logística. Esses 3 modelos foram treinados com uma amostra com 90% dos dados e treinado com uma amostra de teste com 10% dos dados, disponibilizados pela UCI. A acurácia para os 3 modelos foram de 85% (???). Apenas 7% (foram influenciados pela promoção (???)**

***Palavras-chave* – Machine Learning (ML), dataset (DS), Regressão Logística (RL).**

## I. INTRODUÇÃO

Regressão Logística em Machine Learning é uma técnica de aprendizado supervisionado que consiste na regressão de um modelo matemático que relaciona variáveis de entrada $X_i(i = 1, 2, ..., n)$ a diferentes grupos de classificação. Para isso, é usada a função *Sigmoid* para determinar a probabilidade de um determinado conjunto de variáveis a pertencerem a determinado grupo:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{1}$$

Como já visto anteriormente em Regressões Lineares, na Regressão Logística o melhor modelo de classificação é encontrado através da utilização do algoritmo de Gradiente Descendente, atualizando os valores de $\theta_j$ até encontrar o minímo da função custo $J$:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)})) \tag{2}$$

Quando o melhor modelo de classificação é encontrado, tal classificação está relacionado apenas a uma classe, sendo considerado um modelo de classificação binária por apenas determinar se um dado pode ser considerado da classe em questão ou não.

A escolha do tema de eficiência da campanha de vendas de depósitos bancários a termo foi baseado no artigo de Moro et al [2] e da disponibilidade dos dados no Repositório de Dados de Machine Learning da UCI [1]. O artigo refere-se à campanha de um banco português para obter mais clientes para um produto oferecido sobre depósitos bancários a termo, uma espécie de CDB do Brasil, cuja taxa de juro é baseada no euribor3m, que é uma taxa interbancária entre bancos da União Européia, com duração de 3 meses. A campanha foi realizada por meio de chamadas telefônicas ao telefone fixo residencial ou ao celular do potencial cliente. Uma abordagem ao cliente é realizada com uma certa duração onde são explicado o produto de venda em questão e anotados uma série de dados do perfil do cliente e também do momento sócio econômico. Alguns aceitam a subscrição bancária em troca do pagamento do juros após 3 meses, baseado na taxa euribor3m, outros simplesmente não aceitam.

Assim o banco gostaria de saber o que influencia o cliente na subscrição ao produto de venda em questão, se há possibilidade de predição de subscrição ao produto e quais dados do perfil do cliente ou quais dados do momento sócio econômico influenciam essa predição, porque se o custo-benefício de contratar uma empresa de telemarketing não compensar, poderia simplesmente diminuir o gasto com propaganda de acordo com Moro et al. [2].

O artigo de Moro et al [2] e a disponibilidade pública do dataset na UCI [1], também encontrado no sítio kaggle.com, junto com conteúdo dado nas aulas de Inteligência Artificial (MO416A), nos encorajaram a exercitar o tópico de machine learning.

Nos dados disponibilizados no sítio da UCI [1], observou-se que se tratava de uma caso típico de aprendizado supervisionado, onde a variável dependente é simplesmente aceitação ou não da subscrição bancária e os dados independentes eram valores categóricos (e.g. profissão, estado conjugal, educação, etc.) e dados numéricos (e.g. idade, número de contatos, valor do euribor3m, etc.).

Dois artigos, em especial, nos serviram de guia para a elaboração deste artigo. O primeiro foi um artigo da Susan Li [3] que descreve como a preparação dos dataset de treinamento deve ser feito, balanceamento de amostra com a utilização da

técnica SMOTE (Synthetic Minority Over-sampling Technique) onde se mostra como balancear amostras com respostas positivas à subscrição e com respostas negativas à subscrição, bem como redução de features por meio da técnica RFE (Recursive Feature Elimination), técnicas segundo as quais diminuiria casos de Falsos Positivos. E o artigo Nelson Chris [4] que utiliza um técnica de Feature Engineering, onde se cria um novo campo a partir do campo pdays para representar se ouve contato anterior ou não. Este campo tem uma valor específico 999 que é usado para quando nunca ouve um prévio contato com o cliente, e em outras vezes os valores representam dias passados desde o último contato. Ambos os artigos mostram como tratar de variáveis dummy para variáveis categóricas.

Este artigo está dividido da seguinte forma. A Seção II descreve como será abordado os problemas que queremos tratar, sobre as bibliografias utilizadas como referência. A Seção III descreve a proposição do trabalho, de como será tratado a análise dos dados e como medir o desempenho da campanha de promoção de vendas de depósitos bancários. A Seção IV será descrito os materiais e métodos utilizados para aquisição, formatação dos dados, criação e teste do modelo. A Seção V mostrará os resultados dos experimentos e uma breve discussão dos resultados da análise. A Seção VI descreverá as principais conclusões do experimento.

## II. Abordagem do Problema

A abordagem do problema será por meio das fases do Machine Learning (ML approach), como descritos nas seguintes subseções.

### A. Extração dos Dados

Os dados serão extraídos da UCI [1], onde o ds bank-additional-full.csv contém 90% dos dados e será utilizado na fase de treinamento dos dados. O mesmo contém 41188 linhas e 20 colunas (features). E o ds bank-additional.csv contém 10% dos dados, com 4199 linhas e 20 features ,e será utilizado para teste do modelo. Há outros dois ds, com menos dados, que não serão utilizados para este projeto. São o bank-full.csv e bank.csv.

As variáveis dos datasets (ds) extraídos da UCI [1] são as seguintes:

- Dados bancários do cliente:
  1) **age:** idade (numérico)
  2) **job:** tipo de trabalho (categórico: 'admin.', 'blue-collar', 'entrepreneur','housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
  3) **marital:** estado conjugal (categórico: 'divorced', 'married', 'single', 'unknown'. Nota: 'divorced' significa divorciado ou viuvez).
  4) **education:** educação (categórico: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
  5) **default:** está insolvente? (categórico: 'no','yes','unknown')
  6) **housing:** tem empréstimo de habitação? (categórico: 'no','yes','unknown')
  7) **loan:** tem empréstimo pessoal? (categórico: 'no', 'yes', 'unknown') relativo ao último contato da campanha corrente.
  8) **contact** tipo de contato realizado (categórico: 'cellular', 'telephone').
  9) **month:** mês do ano do último contato (categórico: 'jan', 'feb', 'mar', ..., 'nov', 'dec').
  10) **day_of_week:** dia da semana do último contato (categórico: 'mon', 'tue', 'wed', 'thu', 'fri').
  11) **duration:** duração do último contato em segundos (numeric). Nota importante. Este atributo afeta altamente a variável dependente (e.g. se duration=0, então y='no'). A variável duration não é conhecida antes que a chamada seja concluída. Também, após o fim da chamada, "y"é obviamente conhecido. Dessa forma, esta variável poderia ser somente incluída para propósito de benchmark e poderia ser descartado se a intenção fosse para aplicar num modelo de predição realístico.
- Atributos do contexto social e econômico:
  13) **emp.var.rate**: indicador trimestral da taxa de variação do emprego (numérico).
  14) **cons.price.idx:** índice de preço mensal de preço ao consumidor (numérico) - semelhante ao inpc/ipca do Brasil.
  15) **cons.conf.idx:** índice de confiança do consumidor - indicador mensal (numérico) - semelhante ao ICC da FGV.
  16) **euribor3m:** Taxa euribor 3 meses - indicador diário (numérico).
  17) **nr.employed:** número de pessoas empregadas - indicator trimestral (numérico).
- Outros atributos:
  18) **campaign:** número de contatos realizados durante a campanha e para este cliente (numérico, inclui o último contato).
  19) **pdays:** número de dias que se passaram após o último contato com o cliente desde a última campanha (numérico; 999 significa que o cliente não foi previamente contactado).
  20) **previous:** número de contatos realizados antes desta campanha e para este cliente (numérico).
  21) **poutcome:** resultado da campanha de marketing prévia (categórica: 'failure', 'nonexistent', 'success')
- Variável dependente (saída do modelo/objetivo desejado):
  21) **y:** o cliente se subscreveu ao plano de depósito a termo (binário: 'yes', 'no').

### B. Preparação dos Dados

O notebook Project3.ipynb [1] dá mais detalhes de como foi realizado a preparação dos dados.

[1] https://github.com/edbkei/MO416PROJ3/tree/master/Projeto3

A variável dependente y foi transformado em dados binários, em vez dos dados categóricos yes e no. Bem como feito também na variável independente contact, onde o cellular ficou 0, e o telephone ficou 1.

Foi criado uma nova variável independente pdays_no_contact derivado do pdays, de forma que o valor 999 ficou com o valor 1 (não houve contato) e 0 (houve contato), seguindo orientação do Nelson Chris [4].

Foi verificado inicialmente que houve 11.26% de subscrição e 88.72% de não subscrição no ds de treinamento. Se utilizado o ds de treinamento sem balanceamento, haveria o risco de o modelo fazer predição com maior número de FP. Seguindo a recomendação da Susan Li [3] , o dataset de treinamento foi balanceado utilizando o algoritmo SMOTE (Synthetic Minority Oversampling Tecnique).

As variáveis categóricas job, marital, education, default, housing, loan, month, day_of_week, poutcome foram transformadas em variáveis dummy, cujos valores viraram binários por meio da rotina get_dummies da módulo pandas. Assim como exemplo, a variável categórica marital, que tem valores married, single, unknown, viraram novas variáveis binárias marital_married, marital_single, marital_married. Tanto Nelson Chris como Susan Li utilizaram a técnica de criação de variáveis dummy para variáveis categóricas.

Fizemos a separação do ds de treinamento em ds da variável independente ("y") e variáveis independentes ("X") por meio do atributo loc do módulo pandas.

Realizamos também a normalização do ds das variáveis independentes ("X") por meio do algoritmo StandardScaler do módulo sklearn.

### C. Seleção das Features

Com a utilização do get_dummies, o número de variáveis aumentou de 20 para 54 variáveis independentes. Assim, Susan Li [3] utilizou a técnica RFE (Recursive Feature Elimination) para reduzir a quantidade de features, basicamente lista-se as variáveis independente com os seus pValue por meio do aplicativo summary2 do módulo Logit. Aquelas features que tiveram o pValue maiores que 5% seriam retirados manual da amostra. Durante o treinamento e no cross-validation (CRV), tiveram ótimo desempenho, com accuracy maior que 90%. Porém na utilização do modelo de Regressão Logística na amostra de teste, o resultado foi pífio. Houve aumento significativo de False Positive (FP), apenas 1 caso de True Positive (TP). Resolvemos fazer como Nelson Chris [4], i.e. manter todas as 54 variáveis independentes.

### D. Treinamento dos Dados

Os dados de treinamento foram treinado com o algoritmo LogisticRegression do módulo sklearn.linear_model. O modelo foi obtido a partir da função fit (fitness) entre os dados de treinamento balanceados e normalizados de X e y.

### E. Validação do Modelo (Cross Validation-CRV)

O modelo obtido a partir da amostra de de treinamento foi também aplicado tanto para a amostra de treinamento como para a amostra de verificação e medidos pela função score para obtenção do accuracy (acurácia) do modelo.

Foi também gerado uma matriz de confusão entre o y real e o y previsto, este foi obtido da função predict à amostra X.

A matriz de confusão mostra uma matriz de 2 x 2, onde estão registrados o número de Falsos Positivos (FP), Falsos Negativos (FN), Verdadeiros Positivos (TP) e Verdadeiros Negativos (TN). Respostas verdadeiras são TP+TN, respostas falsas são FP+FN.

### F. Teste do Modelo

Similar à metodologia aplicado ao CRV, mas aplicado à amostra de teste.

## III. Trabalho Proposto

O trabalho proposto será a busca do modelo de predição de subscrição ao depósito bancário por meio do processo explicado no modelo da Figura 1. Basicamente, essa busca ocorrerá de forma interativa tratando o dataset da UCI [1]. A aceitação do modelo no CRV e no teste de modelo será por meio dos valores das métricas. Espera-se um desempenho superior a 90% nas métricas de precision, AUC, F1.

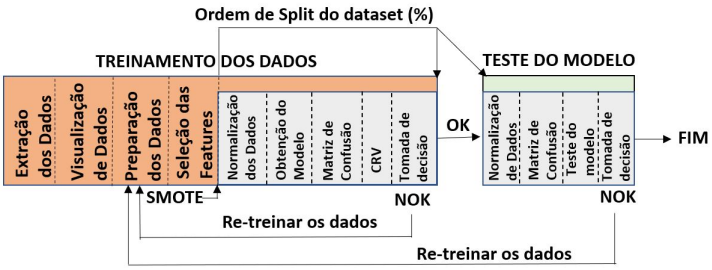


Figura 1. Estratégia de Treinamento, Cross-validation e Teste.

### A. Tabelas

Aqui a amostra de teste é submetida ao mesmo modelo obtido na fase de treinamento e validada no teste de CRV (Cross-Validation). Será o obtido a matriz de confusão, medido o accuracy, precision, F1-score, AUC e também será gerado o gráfico ROC.

Será verificado o número de respostas falsas e também será analisado os componentes principais que afetaram na decisão do y==1 (aceitação da subscrição). A tabela I contém os valores das métricas obtidas durante o CRV e teste do modelo.

Tabela I
Tabela de Desempenho.

| Modelo | Amostra | Score | AUC | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|
| LR(CRV) | 90% | 0.93 | 0.89 | 0.94 | 0.94 | 0.94 | 0.94 |
| LR(Teste) | 10% | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |

## IV. Materiais e Métodos

”Todo trabalho deve ser submetido a algum tipo de teste para que possa ser avaliado. Na verdade, buscamos aqui uma validação com um caráter mais científico de seu trabalho (validação de hipótese). Busca-se identificar quais os seus pontos fortes e fracos. Nesta seção você deve descrever claramente quais foram e como foram conduzidos os testes, quais os materiais e as metodologias empregadas.- Remover esse parágrafo depois.

Aqui avaliamos a eficiência de nossa metodologia, dado na Figura 1, para predição de subscrição de depósito bancário a termo. Inicialmente, extraímos o dataset da UCI [1], no entanto não foi possível garantir que os dados tivessem a mesma distribuição, pois foram registrados de forma cronológica, critérios de sazonalidade não foram considerados, nem houve nenhuma planejamento amostral, apenas a utilização do dataset da forma pública que está disponibilizada. Não está explícito no trabalho de Moro et al. [2], se houve algum critério para a escolha das variáveis independentes. Precisamos criar uma nova variável pdays_no_contact, como realizado por Nelson Chris [4], pois era uma forma de representar o valor 999, que se referia a nenhum contato. Mesmo assim, a amostra pode conter algum viés estatístico pela inclusão de atributos do contexto sócio-econômico, cuja variação é por vezes trimestral (emp.var.rate e nr.employed), mensal (cons.price.idx e cons.conf.idx) e diário (euribor3m). Isso pode resultar em variáveis com pValue alto, ou seja pouco significativo.

Isto pode ser comprovado com a análise de componentes principais, que corroborou que esses variáveis x, y, z, ...., não foram significativas para o modelo. ”Atualizar ....”

Utilizamos também a técnica SMOTE, como recomendado por Susan Li [3], para balancear casos com respostas positivas (yes) com as negativas (no) e dessa forma, mitigar as predições com respostas falsas, aquelas referidas como Falsos Positivos ou Falsos Negativos.

## V. Resultados e Discussão

”Nesta seção você deve apresentar claramente os resultados obtidos para os testes efetuados. Procure organizar os dados utilizando uma linguagem científica. Algumas opções são o uso de tabelas e gráficos, para que a compreensão seja fácil e rápida. - remover depois.

## VI. Conclusões

Nesta seção, faça uma análise geral de seu trabalho, levando em conta todo o processo de desenvolvimento e os resultados. Quais os seus pontos fortes? Quais os seus pontos fracos? Quais aspectos de sua metodologia de trabalho foram positivas? Quais foram negativas? O que você recomendaria (ou não recomendaria) a outras pessoas que estejam realizando trabalhos similares aos seus?

A campanha de promoção de vendas de depósito bancário à prazo, por meio de telemarketing, obteve subscrição de apenas 11 % dos potenciais clientes abordados. Possivelmente, devido ao fato que a abordagem para determinadas classes ou condições sociais econômicas sejam impeditivas ou não entenderem o tipo de investimento que lhes são propostos. Depósito bancário à prazo pode ser interessante se os juros pagos no resgate forem também interessantes. O euribor3m é a é taxa interbancária contra a qual um grupo representativo de bancos europeus contrai empréstimos mutuamente cuja duração é de 3 meses, muito similar o CDB do Brasil, o ajuste de juros de empréstimos, conta poupança, hipoteca, etc seguem essa taxa. O nosso modelo indicou o euribor3m como feature importante, como esperado que fosse. Algumas pessoas veem valor com esse tipo de transação bancária, pois pode emprestar o dinheiro por um tempo para o banco em troca de juros, como é o caso de aposentados (job_retired) e estudantes (job_student), pois tem nível universitário (education_university.degree) e portanto é de se esperar que entendam o mecanismo da aplicação financeira. O clima econômico parece favorecer potenciais clientes para aquisição da aplicação como indica a média trimestral do número total de cidadãos empregados (nr.employed) e a taxa de variação de empregabilidade trimestral (emp.var.rate). A enquete durante a campanha mostra que potenciais clientes escondem se os mesmos tem algum problema com débitos pendentes (default_unknown). Contato por meio de telefone ou celular (contact), ou mais de um contato (campaign), com o cliente, também se já houve um contato prévio bem sucedido (poutcome_success) ou não existente (poutcome_nonexistent), com uma certa duração (duration), parecem influir na aquisição do plano do depósito à prazo. Alguns meses no ano tiveram melhor aceitação da campanha como março, maio, julho, agosto, setembro, novembro e dezembro. O melhor dia da semana para abordar clientes foi quarta-feira.

+————-+

## Referências


[1] M. L. Repository, *Bank Marketing Data Set*, 2014 (accessado Julho 18, 2020). [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing 1, 2, 3, 4

[2] P. Moro, Sergio; Cortez and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” *Decision Support Systems*, vol. 62, pp. 22–31, 2014. 1, 4

[3] S. Li, *Building A Logistic Regression in Python, Step by Step*, 2019 (acessado Julho 24, 2020). [Online]. Available: https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8 1, 3, 4

[4] N. Chris, *Bank Marketing campaign prediction using logistic regression*, 2019 (acessado Julho 24, 2020). [Online]. Available: https://medium.com/@ogbeide331/bank-marketing-campaign-prediction-using-logistic-regression-d3a1072ac1552, 3, 4


# Submissão

Seu trabalho deve ser submetido via Google ClassRoom.

**PRAZO: 09/08/2020**