Contratação de *term deposit*

Disciplina: PCS5787 - Tópicos Especiais em Ciência dos Dados e Big Data (2022)

Grupo 7
Marcos Paulo Pereira Moretti
NUSP 9345363
Curso: Mestrado
Engenharia de Computação

Objetivo

- Predizer se um cliente **vai contratar ou não** um *term deposit* após um contato telefônico
- **Comparar métodos** de *machine learning* nessa tarefa de classificação
- Identificar as **variáveis relevantes** na contratação
- Identificar o **público que mais contrata e menos contrata** com base nas predições

Fonte: https://www.flaticon.com/br/icone-gratis/operador-de-telemarketing_314306 (05/12/2022)

Motivação

- Tópico muito explorado na literatura
    - Artigo recente* cita **mais de 18 artigos** outros com técnicas de modelagem para o problema *Machine Learning for Bank Telemarketing Prediction*
    - Busca por "machine learning bank telemarketing" (feita em 05/12/2022) na WoS traz **16 resultados**, sendo 13 deles nos últimos 3 anos
- Apresentar **ofertas personalizadas** aos clientes de forma eficaz

** H. Toulni et. al. S. C. K. Tékouabou, S. C. Gherghina, "A machine learning framework towards bank telemarketing prediction," Journal of risk and Financial Management, no. 15, June 2022.*

Referências de **papers relevantes** sobre machine learning for bank telemarketing citados em [H. Toulni et. al. S. C. K. Tékouabou, S. C. Gherghina]

Table 1. Summary of the relevant papers dealing with direct bank telemarketing prediction using machine learning. SRAP = Scientific Research an Academy Publisher; CBWDJ = Class-based weighted decision jungle, JSCS = Japanese Society of Computational Statistics.

Ref.	Year	Nb_f	Tools	Algorithms	Metrics	Best Score (%)	Publisher	Type
Feng et al. (2022)	2022	21	Python	META-DES-AAP	Acc, AUC	89.39; 89.44	Elsevier	Article
Koumétio and Toulni (2021)	2021	13	Python	improved KNN	Acc, AUC, f_1	96.91	Springer	Chapter
Yan et al. (2020)	2020	21	-	S_Kohonen network	Acc	80	Elsevier	Article
Ghatasheh et al. (2020)	2020	21	-	CostSensitive-MLP	Acc	84.18	MDPI	Article
Selma (2020)	2020	21	-	ANN	Acc; f_1	98.93; 95.00	Waset	Article
Birant (2020)	2020	21	-	CBWDJ	(Acc; Arec; Rec)	(92.70; 84.92; 75.93)	IntechOpen	Chapter
Tekouabou et al. (2019)	2019	21	Python	DT C5.0	Acc, Prec, Rec, f_1	100	ACM	Conf
Farooqi and Iqbal (2019)	2019	21	WEKA	DT J48	Acc,Spe, Sen, prec, AUC, f_1	91.2; 95.9; 53.8; 62.7; 88.4; 58	IJRTE	Article
Mustapha and Alsufyani (2019)	2019	17	-	ANN	Info Gain, Entropy	-	The SAI	Article
Ilham et al. (2019)	2019	21	RapidMiner	SVM	Acc, AUC	97.7; 92.5	IOP	Chapter
Ładyżyński et al. (2019)	2019	21	H2O	RF,DL	prc, rec		Elsevier	Article
Koumétio et al. (2018)	2018	18	RapidMiner	DT C4.5	Acc, f_1	87.6; 81.4	IEEE	Conf
Moro et al. (2014)	2014	22	R/rminer	LR, DT, NN, SVM	AUC; ALIFT	80.0; 70.0	Elsevier	Article
Vajiramedhin and Suebsing (2014)	2014	8	-	C4.5	Acc, AUC	92.14; 95.60	Hikari	Article
Elsalamony (2014)	2014	17	SPSS	MLPNN, TAN, LR, C5.0	Acc, Sens, Spec	90.49; 62.20; 93.12	FCS	Article
Karim and Rahman (2013)	2013	21	WEKA	NB; C4.5	Acc, Prec, AUC	93.96; 93.34; 87.5	SRAP	Article
Elsalamony and Elsayad (2013)	2013	18	-	BC, RF, SC, GB (C5.0)	Acc; AUC; Kappa	96.11; 99.3; 91.70	SRP	Article
Moro et al. (2011)	2011	29	R/rminer	NB; DT; SVM	AUC; ALIFT	93.8; 88.7	EUROSIS-ETI	Article

Dataset escolhido

- Bank Marketing Data Set (UCI Repository)
  - Datasets sobre **ligações telefônicas de campanhas de marketing** de uma instituição bancária portuguesa
  - Ligações oferecendo *term deposit*
  - Variável resposta: contratou ou não o produto
- Dataset completo foi utilizado
  - 41.188 exemplos de ligações telefônicas
  - Ligações entre maio de 2008 e novembro de 2010
  - 21 atributos e 1 resposta

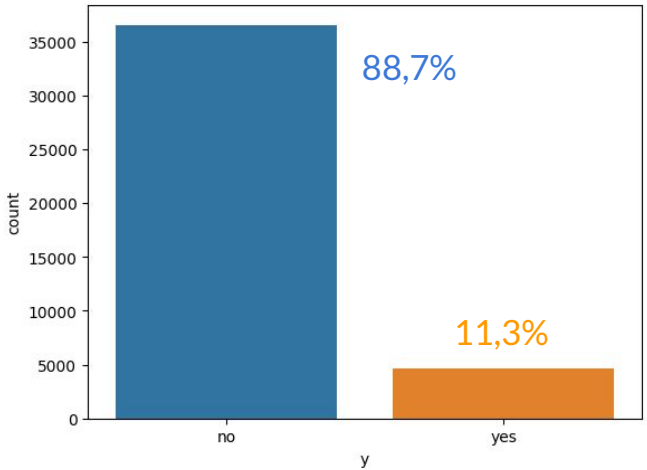
Fonte:
https://www.iconfinder.com/icons/2585818/bank_server_banking_database_bigdata_database_server_financial_database_icon
https://archive.ics.uci.edu/ml/index.php (06/12/2022)

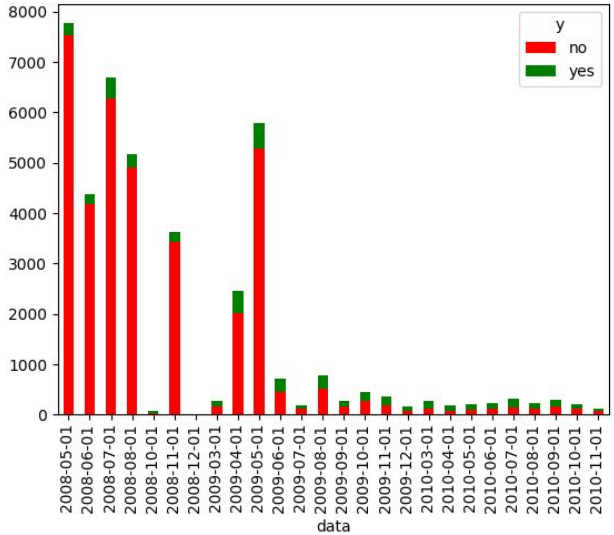
O que é *term deposit*?

- Investimento em renda fixa
- Emitido por instituição financeira
- Curto prazo
- Liquidado no vencimento do título
    - Maiores taxas de retorno
- Similar ao CDB (Certificado de Depósito Bancário)

Fonte: https://www.istockphoto.com/photo/license-gm1080435824-?utm_medium=organic&utm_source=google&utm_campaign=iptcurl (06/12/2022)

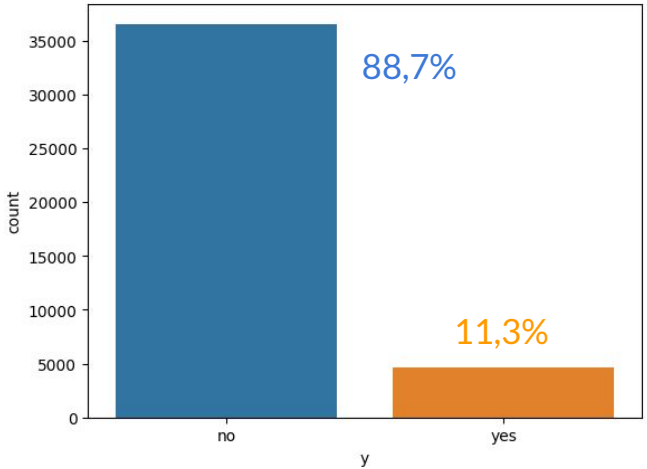
Análise exploratória - target

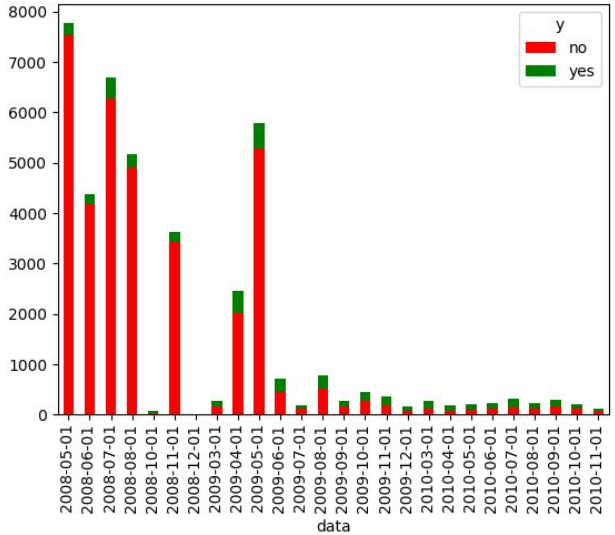




Análise exploratória - target

- Target "levemente" desbalanceada
- Referências faltantes
- Referências fortemente desbalanceadas

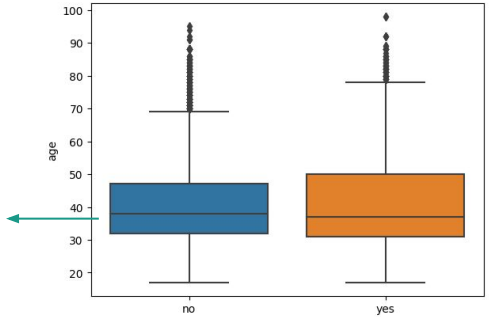




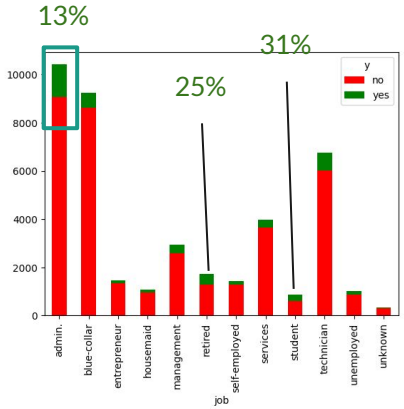
Análise exploratória - Dados do cliente (pessoais)

- Idade (anos)

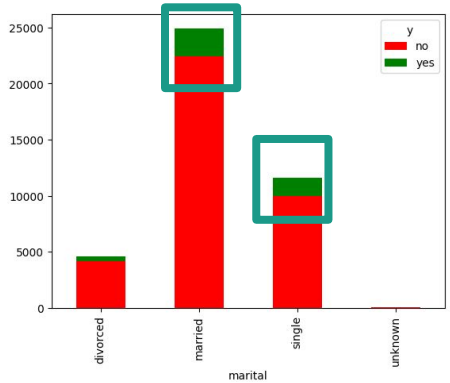
não discrimina y



- Profissão

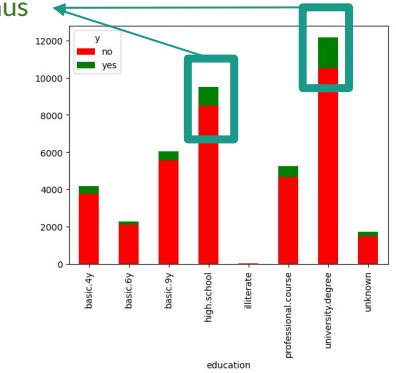


- Estado civil



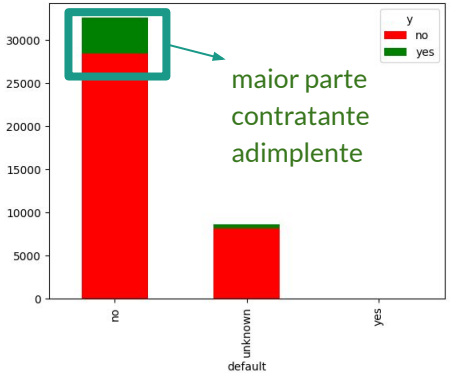
- Grau de escolaridade

maiores graus



Análise exploratória - Dados do cliente (crédito)

- Inadimplência

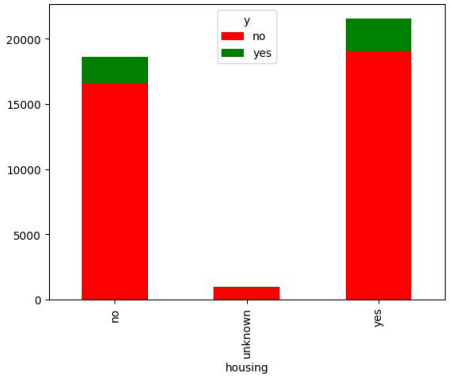


maior parte contratante adimplente

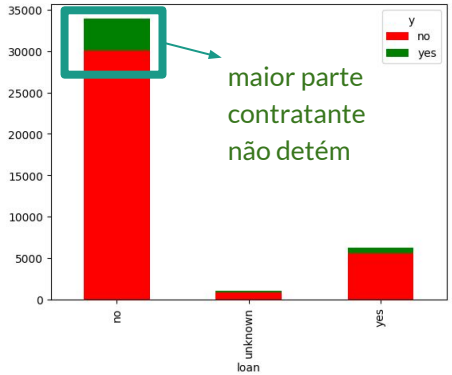
y	default	no	yes
0	no	28391.000	4197.000
1	unknown	8154.000	443.000
2	yes	3.000	NaN

somente 3 ocorrências de inadimplentes (vários *unknowns*)

- Detém crédito imobiliário



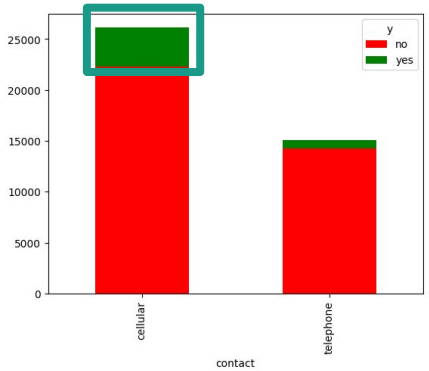
- Detém crédito pessoal



maior parte contratante não detém

Análise exploratória - Dados do contato

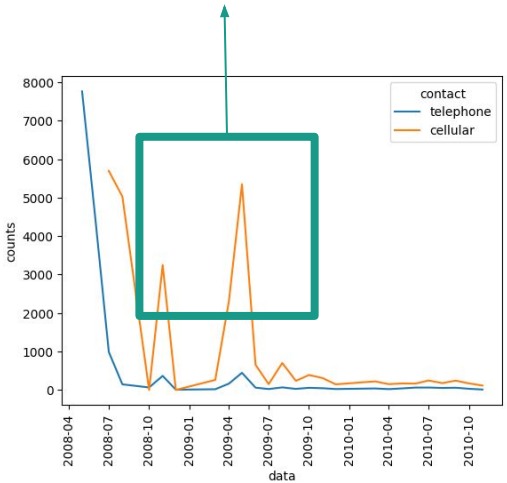
- Tipo do contato (telefone ou celular)



y	contact	no	yes	%yes
0	cellular	22291	3853	0.147
1	telephone	14257	787	0.052

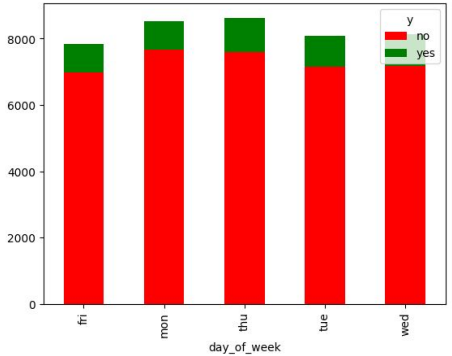
celular tem mais sucesso

evolução do uso do celular no tempo



Análise exploratória - Dados do contato

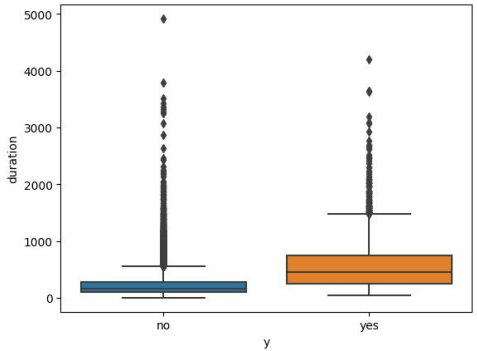
- Dia da semana



y	day_of_week	no	yes	%yes
2	thu	7578	1045	0.121
3	tue	7137	953	0.118
4	wed	7185	949	0.117
0	fri	6981	846	0.108
1	mon	7667	847	0.099

aproximadamente mesma proporção de contratantes

- Duração

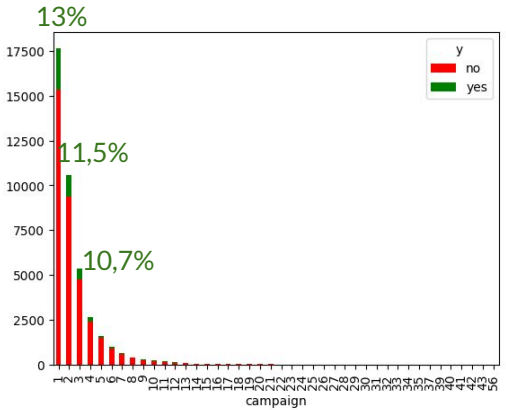


parece discriminar, pois ligações mais longas geram mais contratações

não é possível utilizar essa variável em tempo de predição

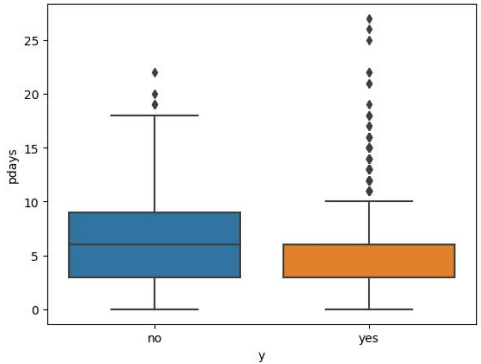
Análise exploratória - Campanhas

- Contatos feitos na campanha atual



> 10% de contratações quando fazemos até 3 ligações

- Dias passados desde a última chamada



Não parece discriminar y

- Contatos feitos **antes** da campanha atual

previous	no	yes	%yes
5	5.000	13.000	0.722
6	2.000	3.000	0.600
3	88.000	128.000	0.593
4	32.000	38.000	0.543
2	404.000	350.000	0.464
1	3594.000	967.000	0.212
0	32422.000	3141.000	0.088
7	1.000	0.000	0.000

Depois de ser ligado 1 vez, 20% das ligações levaram à contratação

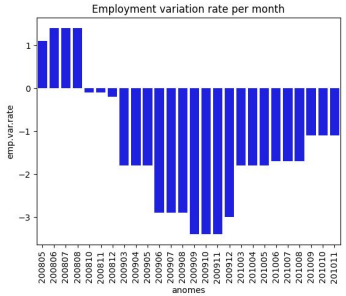
- Resultado da última campanha

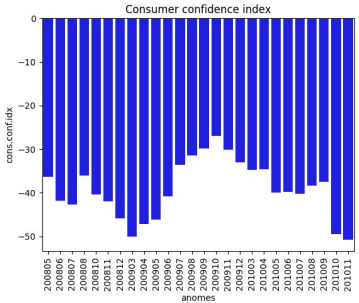
poutcome	no	yes	%yes
success	479	894	0.651
failure	3647	605	0.142
nonexistent	32422	3141	0.088

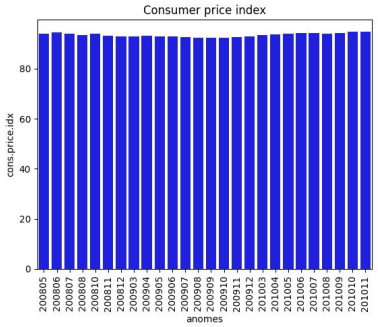
Sucesso na última campanha parece discriminar y

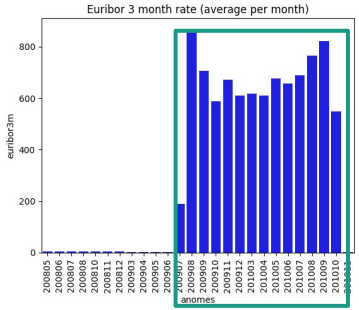
Análise exploratória - Índices externos

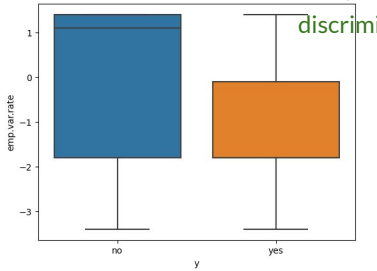
- Variação no índice de empregabilidade (trimestral)
- Índice de Confiança do Consumidor (CCI) (mensal)
- Índice de Preços do Consumidor (CPI) (mensal)
- Euribor (Euro Interbank Offered Rate) (trimestral)



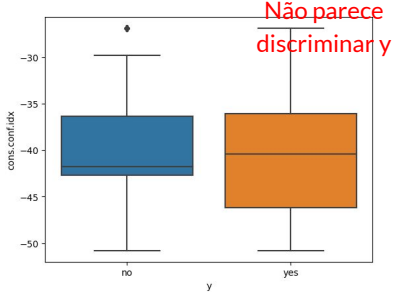




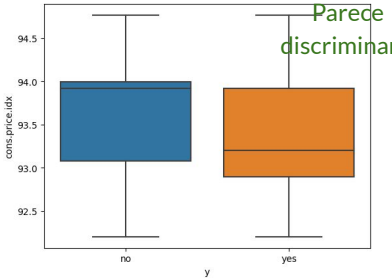




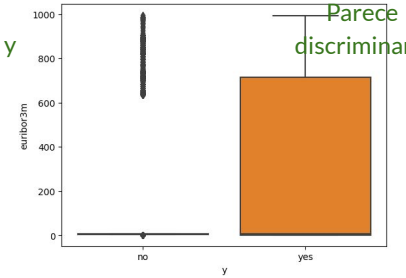
Parece discriminar y



Não parece discriminar y



Parece discriminar y



Parece discriminar y

Análise exploratória - Correlações com a target

Variáveis selecionadas para modelagem

Não foram selecionadas:

- Variáveis temporais (alto desbalanceamento)
- duração (indisponível no tempo de predição)
- Dados pessoais do cliente (idade, profissão, estado civil, profissão)
- Correlação "muito baixa"

duration	0.405
anomes	0.352
year	0.348
euribor3m	0.319
poutcome_success	0.316
previous	0.230
contact_cellular	0.145
month_mar	0.144
month_oct	0.137
month_sep	0.126
default_no	0.099
job_student	0.094
job_retired	0.092
month_dec	0.079
month_apr	0.076
cons.conf.idx	0.055
marital_single	0.054
education	0.044
month_nb	0.037
poutcome_failure	0.032
job_admin.	0.031
age	0.030
job_unemployed	0.015
day_of_week_thu	0.014
housing_yes	0.012
day_of_week_tue	0.008
day_of_week_wed	0.006
marital_unknown	0.005
loan_no	0.005

loan_unknown	-0.002
default_yes	-0.003
loan_yes	-0.004
job_self-employed	-0.005
job_technician	-0.006
job_housemaid	-0.007
day_of_week_fri	-0.007
month_aug	-0.009
month_jun	-0.009
marital_divorced	-0.011
housing_no	-0.011
month_nov	-0.012
job_entrepreneur	-0.017
day_of_week_mon	-0.021
month_jul	-0.032
job_services	-0.032
marital_married	-0.043
campaign	-0.066
job_blue-collar	-0.074
default_unknown	-0.099
month_may	-0.108
cons.price.idx	-0.136
contact_telephone	-0.145
poutcome_nonexistent	-0.194
emp.var.rate	-0.298
pdays	-0.325
nr.employed	-0.355

Modelagem - Feature engineering

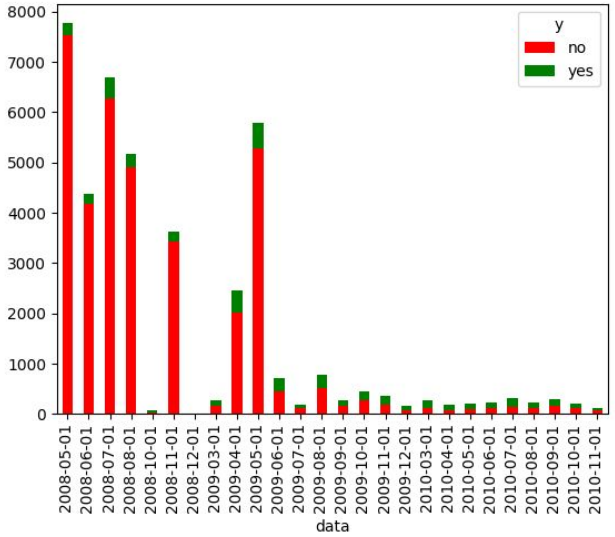
- Encoding das variáveis categóricas
  - yes/no/unknown ⇒ 1/0 (ex.: *y*, *loan*)
  - success/failure/unknown ⇒ 1/0 (ex.: *poutcome*)
  - cellular/telephone ⇒ 1/0 (ex.: *contact*)
- OneHotEncoding de variáveis categóricas
  - Dia da semana

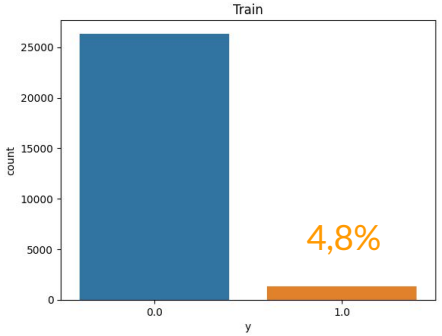
day_of_week	day_of_week_fri	day_of_week_mon	day_of_week_thu	day_of_week_tue	day_of_week_wed
mon	0	1	0	0	0
tue	0	0	0	1	0
wed	0	0	0	0	1
thu	0	0	1	0	0
fri	1	0	0	0	0

Modelagem - Divisão em treino e teste

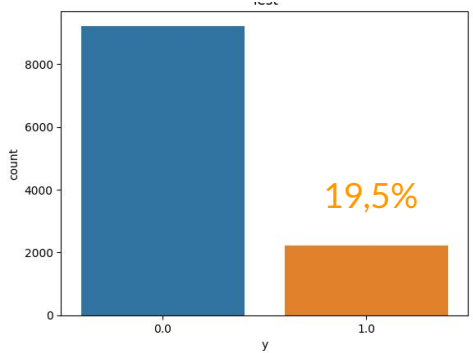
- Out-of-time
  - Treino: 05/2008 a 12/2008
  - Teste: 01/2009 a 12/2009

- Removemos 2010, pois tinha muito desbalanceamento na quantidade de exemplos





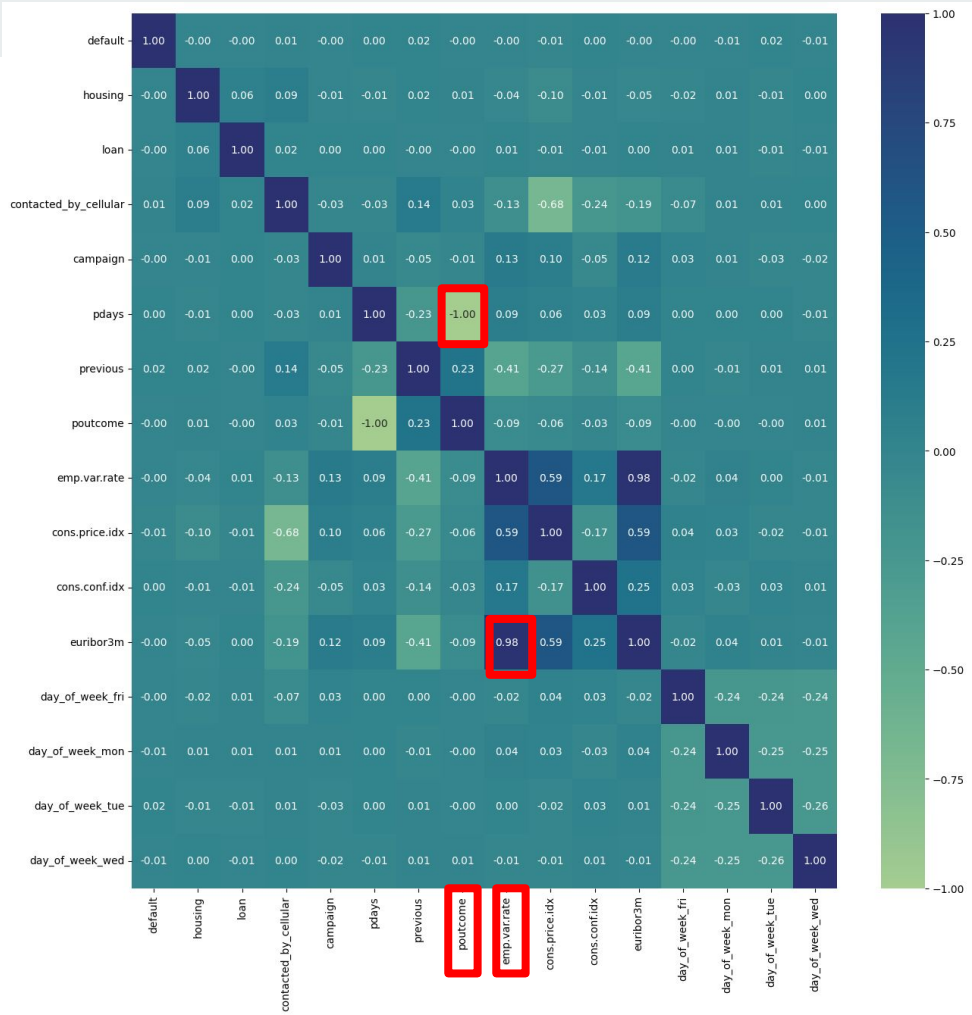
(27690, 17)



(11450, 17)

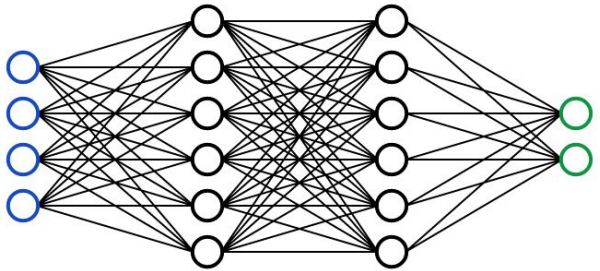
Modelagem - Seleção de variáveis

- Remoção de variáveis muito correlacionadas
  - *pdays*
  - *emp.var.rate*



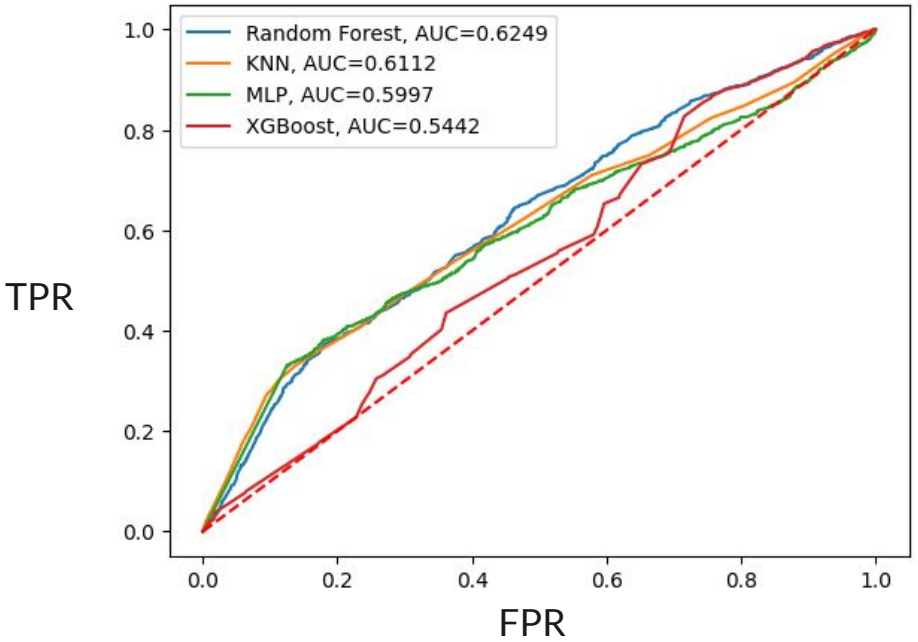
Modelagem - Treino dos modelos

- 15 variáveis utilizadas
- Grid Search para encontrar hiperparâmetros
  - Validação cruzada de 5 folds, avaliados com ROC-AUC
- 4 modelos treinados
  - Random Forest
  - XGBoost
  - KNN
  - Multi-Layer Perceptron

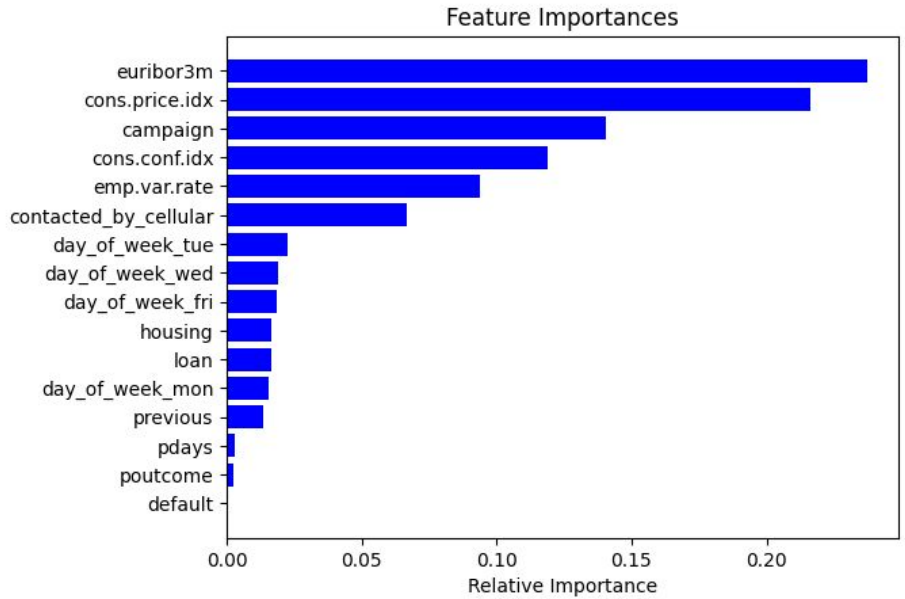


Fonte: https://victorzhou.com/series/neural-networks-from-scratch/ (06/12/2022)

Resultados - Avaliação dos modelos



Resultados - Variáveis mais importantes



- Euribor indica o valor da **taxa de ganho** do investimento
- CPI é a **inflação**, ou seja, melhor se for menor
- **Qtde. de chamadas** na campanha também é importante

Resultados - Perfil mais propenso

pouco público

score_classification	Qtt. of calls	age	job	marital	education	default	housing	loan	contact	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m
3.00	2	58.50	admin.	divorced	university.degree	no	yes	no	cellular	70	20.50	NaN	0.00	-1.80	92.84	-50.00	1.66
4.00	91	38.33	blue-collar	married	high.school	no	yes	no	cellular	7	6.80	2.64	0.95	-1.85	92.90	-45.90	1.30
5.00	3263	38.31	blue-collar	married	high.school	no	yes	no	cellular	11	3.01	4.82	0.65	-1.86	92.92	-45.89	1.57
6.00	7180	39.22	admin.	married	university.degree	no	yes	no	cellular	104	1.71	5.57	0.25	-2.13	92.85	-43.13	133.07
7.00	844	42.93	admin.	married	university.degree	no	no	no	cellular	207	1.76	6.47	0.15	-3.08	92.54	-33.11	583.84
8.00	70	45.16	admin.	married	university.degree	no	yes	no	telephone	120	1.30	13.00	0.01	-3.17	92.40	-30.00	788.50

pouco público

Conclusões

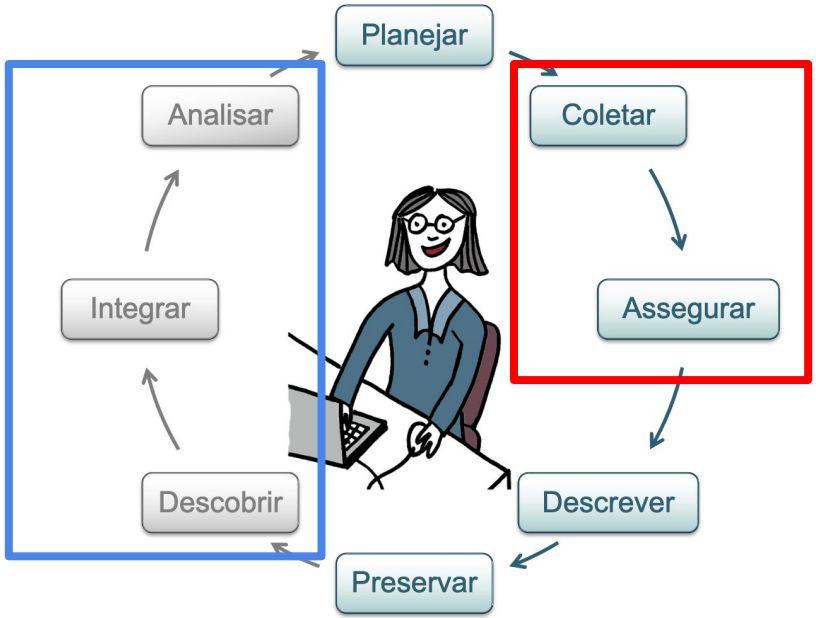
- Treinamos um classificador **melhor que aleatório**
- Descobrimos que os **índices externos são importantes** para determinar a contratação do investimento (Euribor, CPI, CCI)
- Também é importante a **variável de contatos realizados na campanha**
- Descobrimos um perfil de **alta escolaridade**, **baixo tomador de crédito** e **boa qualidade de contatos** (alta duração, mais dias para *recall*, menos recontatos na campanha) como potenciais contratantes

Fonte: https://www.stockio.com/free-icon/business-icons-idea (06/12/2022)

E o ciclo dos dados?

Executado aqui:

Análise de dados e modelagem



A fazer:

Deploy e escoragem com bases futuras ⇒ campanhas

Fonte: Aula 1 - Introdução à Ciência dos Dados e Big Data (PCS5787 - 2022) (06/12/2022)

Muito obrigado!

Dúvidas?

Fonte: https://www.freeiconspng.com/images/question-mark-icon (06/12/2022)

Referências

- https://www.investopedia.com/terms/t/termdeposit.asp
- https://archive.ics.uci.edu/ml/index.php
- https://archive.ics.uci.edu/ml/datasets/bank+marketing
- H. Toulni et. al. S. C. K. Tékouabou, S. C. Gherghina, “A machine learning framework towards bank telemarketing prediction,” Journal of risk and Financial Management, no. 15, June 2022
- Abbas, “Recognition of spoken languages from acoustic speech signals using fourier parameters,” International Journal of Computer Applications, , no. 3, 2015.
- A. Alqaddoumi S. E. Saeed, M. Hammad, “Predicting customer’s subscription response to bank telemarketing campaign based on machine learning algorithms,” International Conference on Decision Aid Sciences and Applications (DASA), , no. 3, 2022.
- G. Casey D. Dheeru, “UCI machine learning repository,” 2017.
- P. Cortez S. Moro and P. Rita, “A data-driven approach to predict the success of bank telemarketing. decision support systems,” 2014.