Semantix - Desafio Data Science

Apresentação dos Resultados

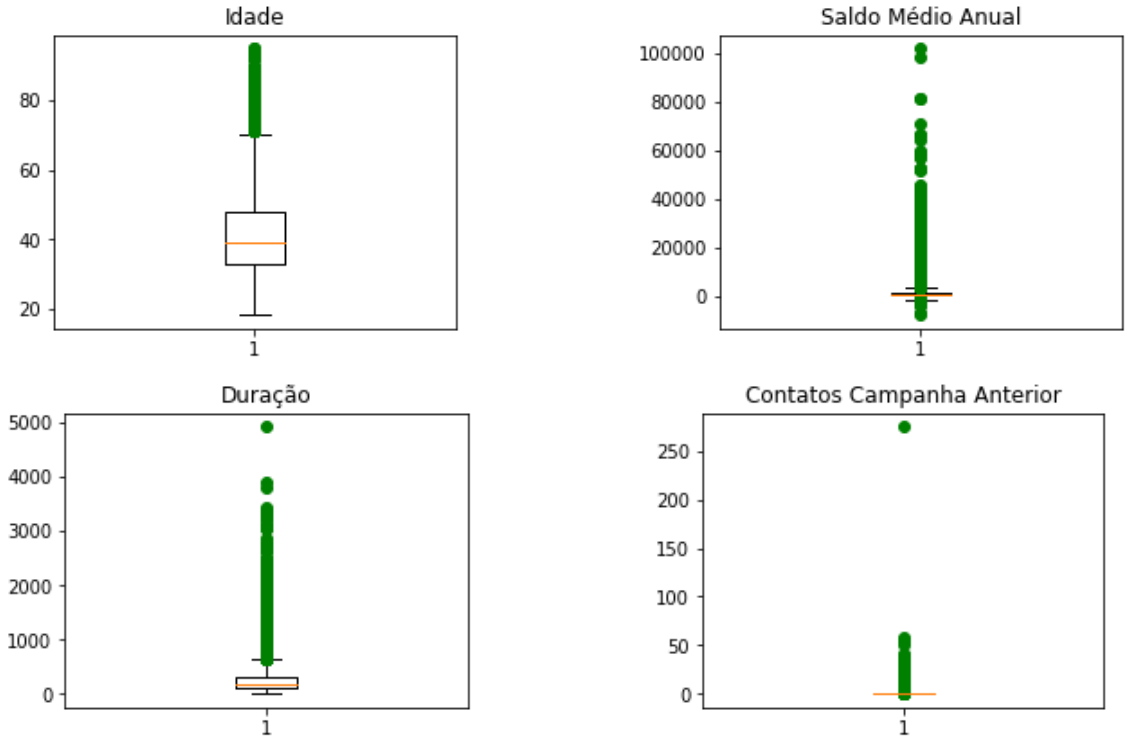
Felipe Alonso Martins

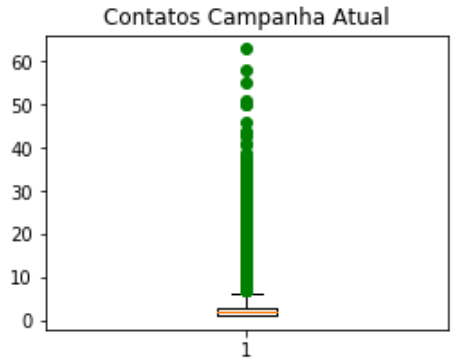
O presente relatório tem por finalidade apresentar a metodologia utilizada e os resultados obtidos na solução das questões relacionadas ao *dataset Bank Marketing,* disponível em https://archive.ics.uci.edu/ml/datasets/bank+marketing.

O documento apresenta 3 seções – primeiramente, é apresentada uma visão geral da abordagem utilizada para a solução dos problemas; em seguida, a ideia por trás da resolução de cada questão seguida pelos resultados; e, por fim, a última seção apresenta os comentários gerais, outras observações e conclusões.

I. Visão Geral

Primeiramente foram analisados os dados numéricos a fim de verificar *outliers.* A figura a seguir exibe diagramas de caixa para as variáveis *age* (Idade), *balance* (Saldo médio anual), *duration* (Duração), *previous* (Contatos da campanha anterior) e *campaign* (Contatos da campanha atual)*:*





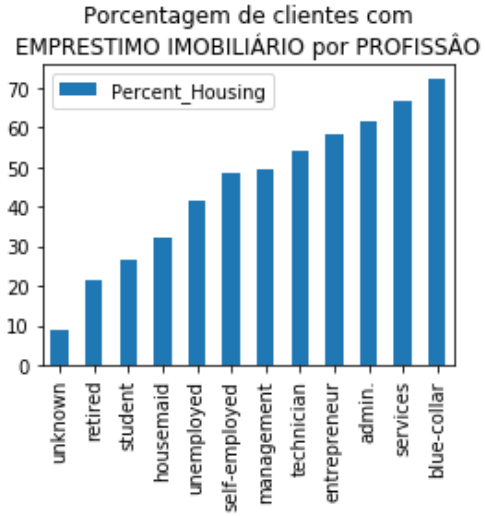
Considerando os diagramas, foi decidido discretizar as variáveis, de modo a encaixar os *outliers* em categorias extremas, aproveitando as informações contidas neles. Além disso a discretização permite dar um maior sentido para esses dados durante as análises posteriores. A tabela 1, no fim do documento, mostra os critérios e informações sobre as discretizações.

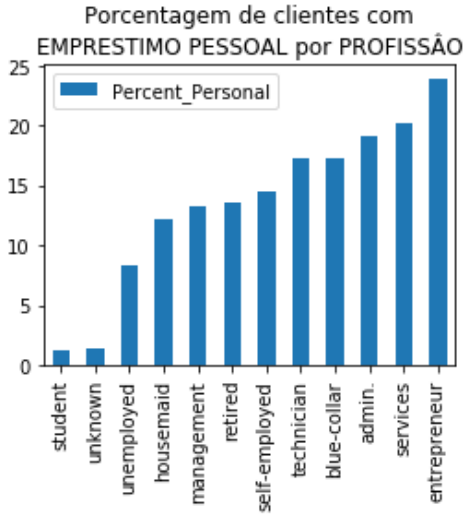
Observação: A variável *duration* não foi discretizada pois, ao analisar as questões, chegou-se à conclusão que ela não seria relevante para as respostas.

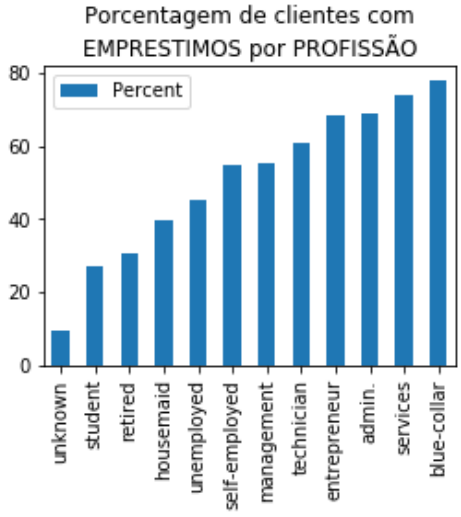
II. Métodos e Resultados

1. *Qual profissão tem mais tendência a fazer um empréstimo? De qual tipo?*

Para responder essa questão, foram calculadas as porcentagens de empréstimos realizados por profissão. A figura a seguir exibe três gráficos, empréstimos totais (imobiliário + pessoal) por profissão, empréstimo imobiliário por profissão e empréstimo pessoal por profissão:



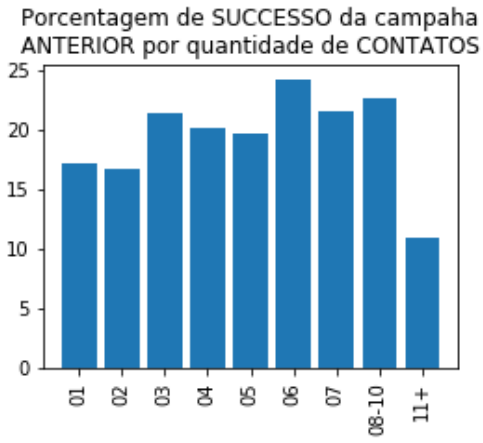


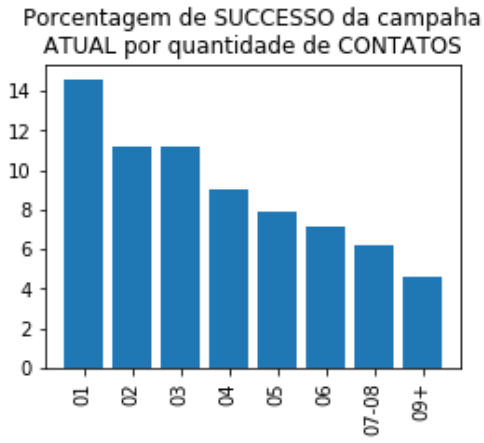


A análise nos mostra que **78,08%** dos profissionais de **colarinho azul** fazem algum tipo de empréstimo (note que existem casos em que um mesmo cliente faz os dois tipos simultaneamente), sendo também a profissão com maior tendência a fazer um empréstimo imobiliário (**72,42%** têm esse tipo de empréstimo). No caso dos empréstimos pessoais, a maior tendência fica com os **empreendedores**, já que **23,94%** deles obtém esse tipo de empréstimo.

2. *Fazendo uma relação entre número de contatos e sucesso da campanha quais são os pontos relevantes a serem observados?*

Para responder à questão, foi calculada a taxa de sucesso das campanhas anterior e atual (variáveis *poutcome* e *y*, respectivamente) em relação à quantidade de contatos realizados (*previous* e *campaign*), lembrando que essas quantidades foram discretizadas anteriormente. A seguir são apresentados dois gráficos contendo essas porcentagens:





Observando os dados, podemos notar que a taxa de sucesso na campanha **anterior oscila** em torno de **20%** entre **1 e 10 ligações**, e cai pela metade a partir de 11. Já no caso da campanha **atual**, a porcentagem de sucesso **cai gradativamente** ao passo que aumenta o número de contatos.

3. *Baseando-se nos resultados de adesão desta campanha qual o número médio e o máximo de ligações que você indica para otimizar a adesão?*

Considerando os dados da questão anterior, é possível verificar quem a taxa de sucesso já se encontra baixa a partir de 9 ligações, portanto seria indicado um número **médio de 3 ligações** e um **máximo de 8**.

4. *O resultado da campanha anterior tem relevância na campanha atual?*

Para verificar a relevância do resultado da campanha anterior, foi calculado a porcentagem de sucesso da campanha atual para 4 tipos de clientes:

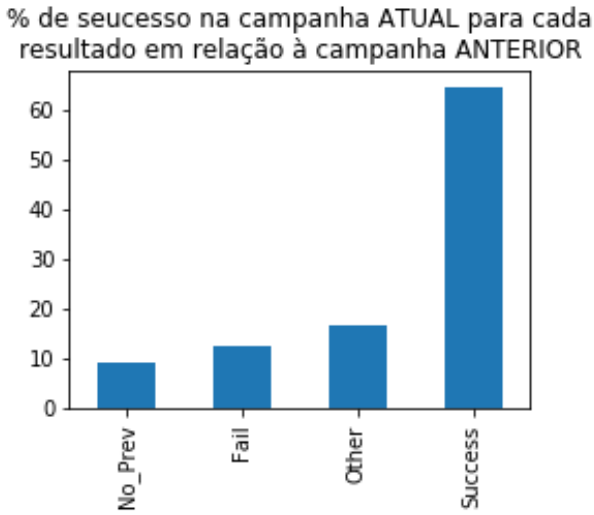
a) *No_Prev:* Não contatados na campanha anterior, (variável *previous* igual a 0)

b) *Fail*: Não aderiram à campanha anterior (variável *poutcome* igual a "fail")

c) *Other*: "Outro" resultado na campanha anterior (variável *poutcome* igual a "other")

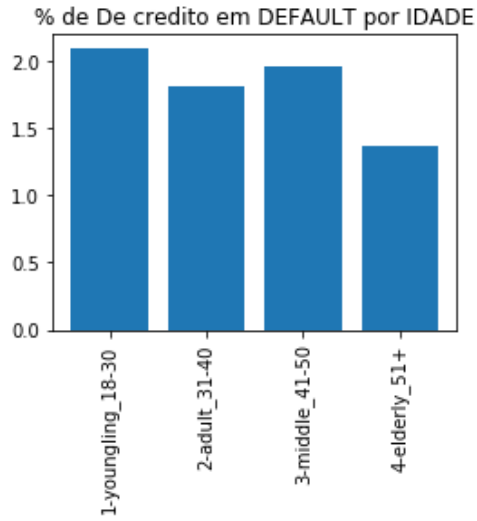
d) *Success*: Aderiram à campanha anterior (variável *poutcome* igual a "success")

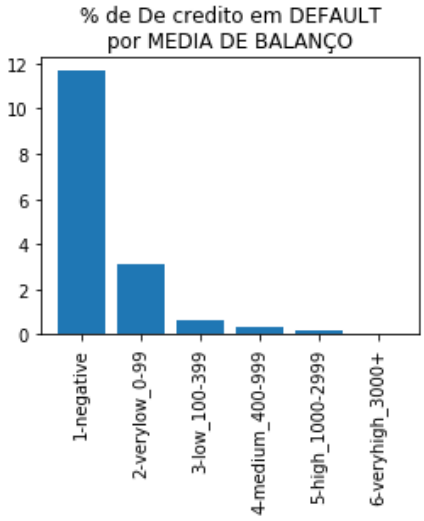
Pelo gráfico a seguir, podemos notar que o resultado da campanha anterior tem forte influência sobre a campanha atual, sendo que **64,73%** dos clientes que aderiram à campanha anterior também aderiram à atual, enquanto os outros não atingem nem **20%** de sucesso.

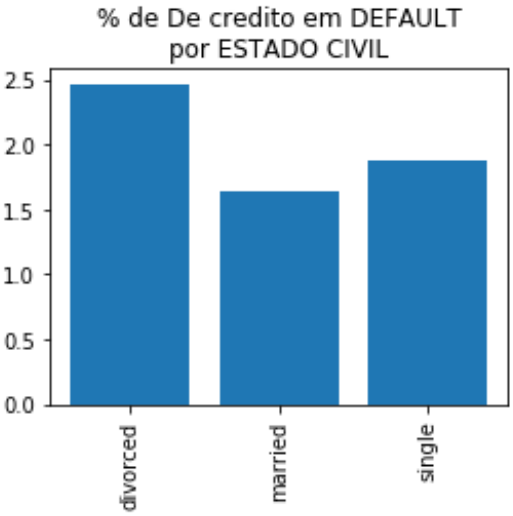


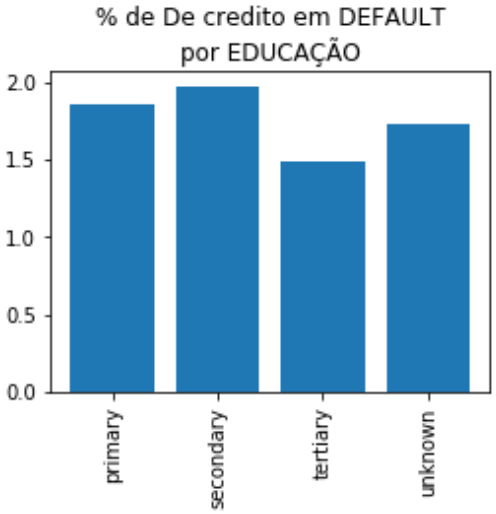
5. *Qual o fator determinante para que o banco exija um seguro de crédito?*

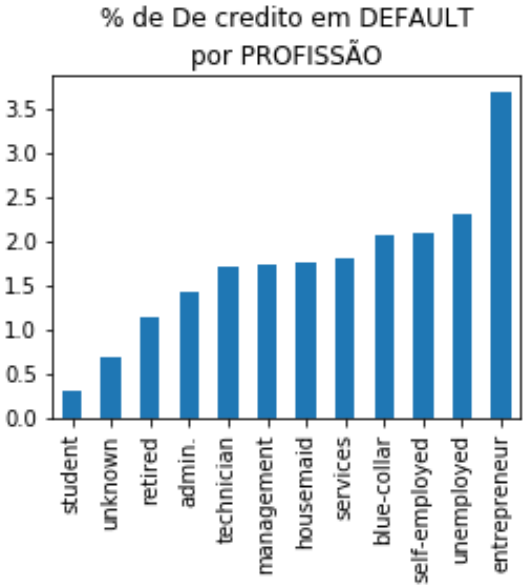
Foram selecionados 7 variáveis que podem ter relação com um possível *Default* do cliente: idade, saldo médio, estado civil, educação profissão, possuir empréstimo pessoal e possuir empréstimo imobiliário (as outras variáveis foram consideradas irrelevantes para o caso). A seguir são apresentados os gráficos com a porcentagem de clientes em *Default* em relação a cada uma das variáveis.

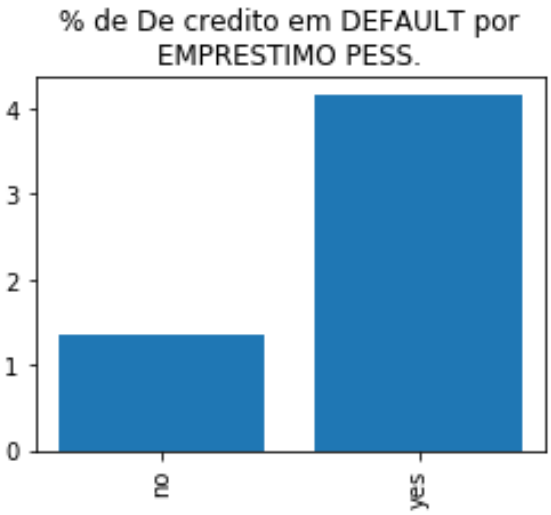

% de De credito em DEFAULT por IDADE
2.0
1.5
1.0
0.5
0.0
1-youngling_18-30
2-adult_31-40
3-middle_41-50
4-elderly_51+

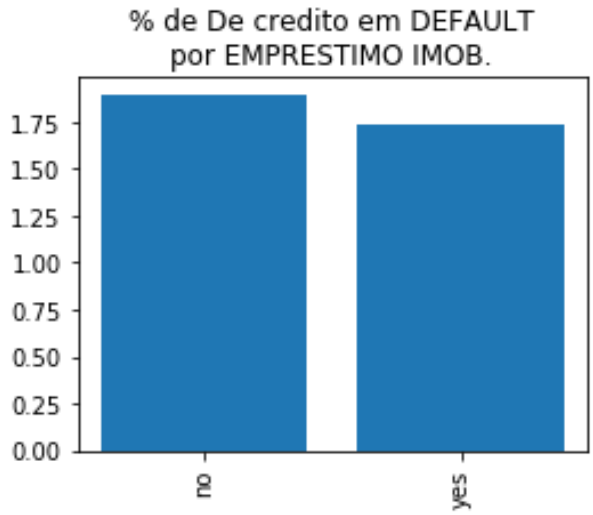

% de De credito em DEFAULT
por MEDIA DE BALANÇO
12
10
8
6
4
2
0
1-negative
2-verylow_0-99
3-low_100-399
4-medium_400-999
5-high_1000-2999
6-veryhigh_3000+


% de De credito em DEFAULT
por ESTADO CIVIL
2.5
2.0
1.5
1.0
0.5
0.0
divorced
married
single


% de De credito em DEFAULT
por EDUCAÇÃO
2.0
1.5
1.0
0.5
0.0
primary
secondary
tertiary
unknown


% de De credito em DEFAULT
por PROFISSÃO
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
student
unknown
retired
admin.
technician
management
housemaid
services
blue-collar
self-employed
unemployed
entrepreneur





Observando os dados verificamos que enquanto a frequência de *Default* é de 1,80% considerando todos os clientes, e os grupos de clientes analisados nos gráficos não chegam a 4% de *Default,* o grupo com **Saldo Médio Anual Negativo possui 11,63%** de clientes com crédito em *Default.* Portanto é recomendável que o banco exija um seguro de crédito para esses clientes.

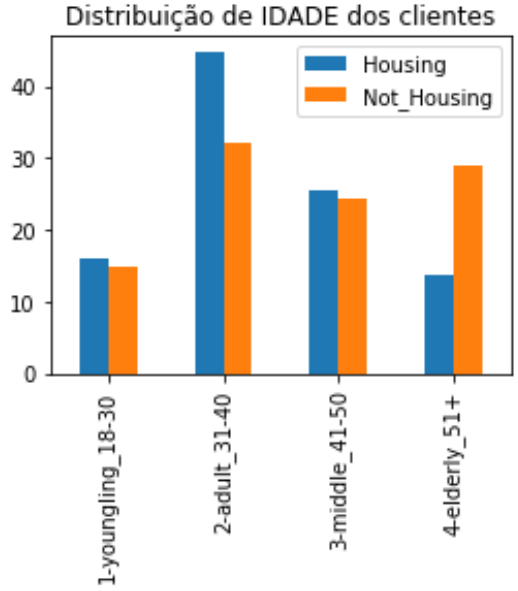
6. *Quais são as características mais proeminentes de um cliente que possua empréstimo imobiliário?*

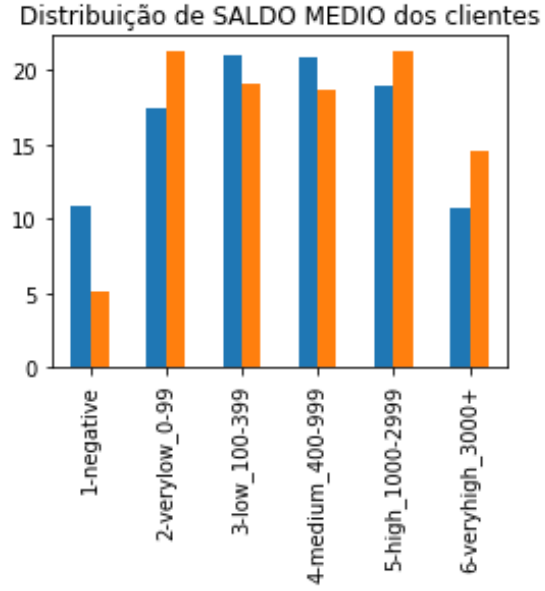
Para verificar as características mais proeminentes de clientes com empréstimo imobiliário, foram comparadas as distribuições desse grupo com as distribuições do grupo de clientes sem empréstimo imobiliário para as seguintes variáveis: idade, saldo médio, estado civil, educação, profissão, possuir empréstimo pessoal, contatos realizados pela campanha anterior e pela atual, estar em *default*, sucesso da campanha anterior e da atual (as outras variáveis foram consideradas irrelevantes para o caso). Cada comparação gerou um gráfico. Porém, como não são todos relevantes, são exibidos apenas os que evidenciam alguma característica dos clientes com empréstimo imobiliário. A análise dos gráficos nos permitem chegar às seguintes conclusões:

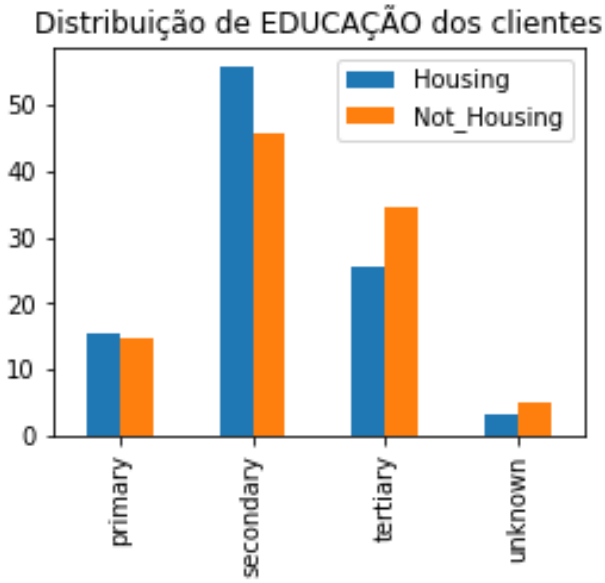
- Em relação a idade, clientes que possuem empréstimo tendem a ser mais jovens do que os que não possuem: enquanto **86,20%** dos que possuem empréstimo tem **menos que 51 anos**, apenas **71,17%** dos que não possuem empréstimo estão nessa faixa.
- Aqueles que possuem **saldo médio anual negativo**, são duas vezes mais frequentes no grupo dos clientes que possuem empréstimo imobiliário, correspondendo a **10,90%** dos clientes, enquanto apenas **5,12%** dos clientes sem empréstimo tem saldo negativo.
- Clientes com empréstimo imobiliário tendem a possuir menos formação, com **25,38%** possuindo terceiro grau, enquanto **34,48%** dos que não tem empréstimo possuem essa formação.
- No tocante à profissão, temos 3 três com grande tendência a possuir empréstimo: Trabalhadores de **colarinho azul** correspondem a **28,05%** dos empréstimos imobiliários, contra **13,37%** dos clientes sem empréstimo; Profissionais da área de administração correspondem a **12,66%** dos empréstimos imobiliários, contra **9,90%** dos clientes sem

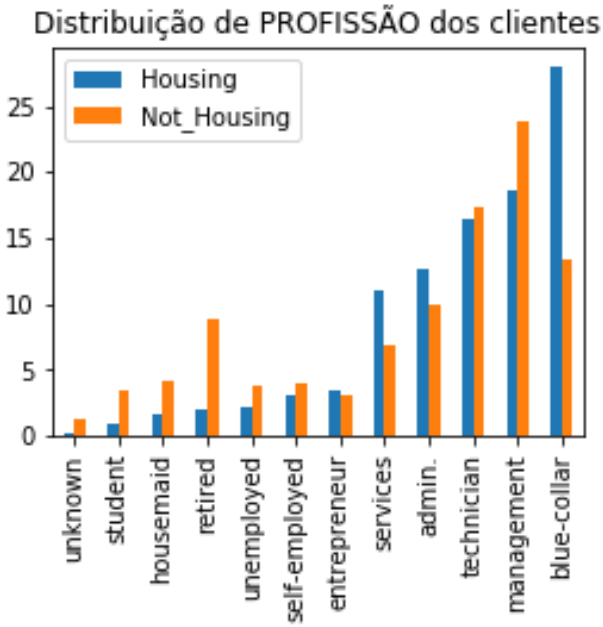
empréstimo. Por fim, profissionais da área de serviços correspondem a **11%** dos empréstimos imobiliários, contra **6,91%** dos clientes sem empréstimo.

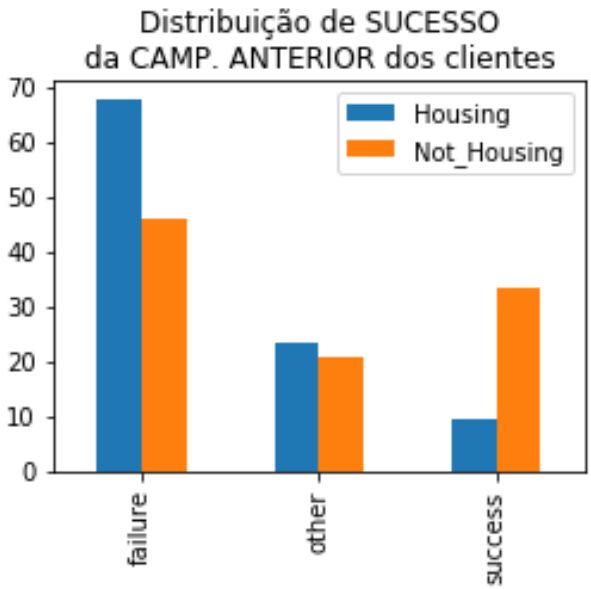
- Esse grupo de clientes tendem a **aderir menos às campanhas**. A campanha anterior atingiu um índice de apenas **9,16%** de sucesso com clientes possuidores de empréstimo imobiliário, contra **33,42%** de sucesso com clientes sem empréstimo imobiliário. Já a campanha atual teve **7,70%** de sucesso com clientes com empréstimo contra **16,70%** de clientes sem.

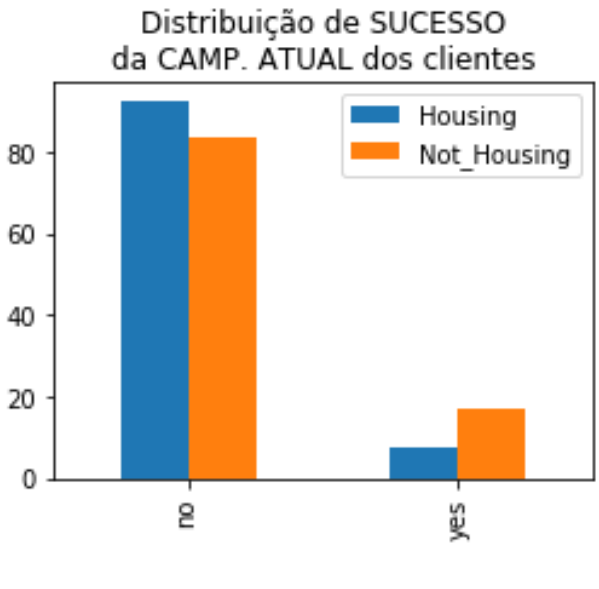












III. Conclusão

Os exercícios foram praticamente todos resolvidos utilizando-se noções básicas de correlação e estatística. Já que as questões não abordavam a predição de resultados nem análises numéricas dependentes de regressão linear, concluiu-se que técnicas de *machine learning* não seriam adequadas para a resolução dos problemas (com exceção da questão 6), por isso foi decidido o uso uma abordagem mais direta.

Em tempo, o candidato avalia que a questão 6 tem potencial de ser melhor desenvolvida utilizando-se *machine learning*, porém, por não possuir ainda experiência suficiente com utilização da técnica para análise de dados (Possui experiências do uso da técnica na área de reconhecimento de padrões em imagens), considerou melhor não utilizá-la.

É importante salientar que no caso de uso de *machine learning* a abordagem de discretização poderia ter sido completamente diferente, já que algoritmos de regressão fazem bom uso de varáveis numéricas contínuas (caso da idade, saldo, duração e quantidade de contatos). Por outro lado, poderia ter sido utilizada regressão logística, que faria uso das versões categorizadas, porém, antes elas deveriam passar por um estágio de *one-hot-encoding*.

Tabela 1. Categorização das variáveis numéricas.

Variável	Significado	Critério	Intervalos	Categorias	Qtd. de dados
age	Idade	Priorizar uma divisão de categorias que faça sentido, mas visando equilibrar ao máximo a quantidade de registros em cada categoria	18 – 30	Jovem	7030
			31 – 40	Adulto	17687
			40 – 50	Meia-idade	11239
			> 50	Idoso	9255
balance	Saldo médio anual (Euros)	Equilibrar ao máximo a quantidade de registros em cada categoria	< 0	Negativo	3766
			0 – 99	Muito Baixo	8671
			100 – 399	Baixo	9113
			400 – 999	Médio	9019
			1000 – 2999	Alto	9027
			> 2999	Muito Alto	5615
campaign	Quantidade de contatos realizados na campanha atual	Manter os valores que se encontram na faixa “normal” (os *não-outliers*) e agrupar os *outliers* de maneira a equilibrar a quantidade de dados do último valor dentro da faixa	1	01	17544
			2	02	12505
			3	03	5521
			4	04	3522
			5	05	1764
			6	06	1291
			7-8	07-08	1275
			>8	09+	1789
previous	Quantidade de contatos realizados na campanha anterior	Manter os valores que se encontram na faixa “normal” (os *não-outliers*) e agrupar os *outliers* de maneira a equilibrar a quantidade de dados do último valor dentro da faixa	0	00 *	36954
			1	01	2772
			2	02	2106
			3	03	1142
			4	04	714
			5	05	459
			6	06	277
			7	07	205
			8-10	08-10	288
			>10	11+	294
** Não utilizado nas análises*					