

ANA RAQUEL

POSTECH

DATA ANALYTICS

DEPLOY DE APLICAÇÕES

AULA 02

SUMÁRIO

O QUE VEM POR AÍ?	3
HANDS ON.....	4
SAIBA MAIS	5
O QUE VOCÊ VIU NESTA AULA?	11
REFERÊNCIAS	12
PALAVRAS-CHAVE	13

O QUE VEM POR AÍ?

O próximo passo, após conhecer o problema de negócio e analisar os dados, é a identificação da coluna target do modelo. Mas será que a target sempre vem explícita na base de dados? Bem, vamos conversar e entender esse possível problema!

EMANIP

HANDS ON

Sabemos que a variável target é a resposta que estamos procurando para o problema de negócio a ser solucionado pelo algoritmo, mas **como identificamos a variável target na base de dados?** Nas aulas a seguir, exploraremos os dados e construiremos a variável target que irá definir se o cliente é um mau ou bom pagador. Você vai perceber, durante as aulas, que a nossa variável target de análise de crédito não está totalmente explícita. Vamos construí-la com base na informação de faixa de atraso de clientes.

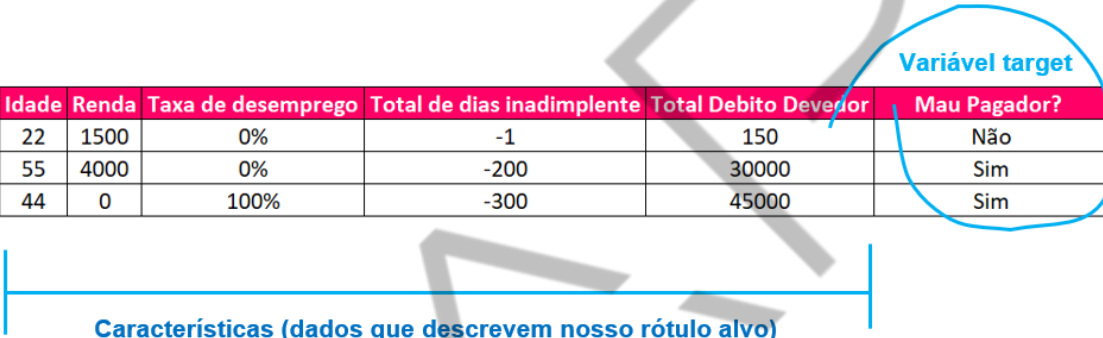
Acesse o link do [GitHub dessa aula](#) para visualizar os notebooks utilizados!

SAIBA MAIS

VARIÁVEL TARGET PARA CLASSIFICAÇÃO

Seguimos com a jornada da construção do pipeline de um modelo de aprendizado de máquina e chegou o momento de pensar na **variável target**.

Infelizmente, nem todas as bases de dados que você irá trabalhar em seus projetos conterá a variável target de forma explícita na base. Mas o que seria a target de forma explícita na base? Acompanhe comigo na figura a seguir:



Idade	Renda	Taxa de desemprego	Total de dias inadimplente	Total Debito Devedor	Mau Pagador?
22	1500	0%	-1	150	Não
55	4000	0%	-200	30000	Sim
44	0	100%	-300	45000	Sim

Características (dados que descrevem nosso rótulo alvo)

Figura 1 - Exemplo de variável target em base de dados.
Fonte: Elaborado pela autora (2023)

Perceba que, no exemplo da figura 1, temos uma base de dados que possui algumas características sobre os clientes e a informação se a pessoa é má ou boa pagadora. Mas e se essa base de dados não tivesse a target já criada?



Idade	Renda	Taxa de desemprego	Total de dias inadimplente	Total Debito Devedor
22	1500	0%	-1	150
55	4000	0%	-200	30000
44	0	100%	-300	45000

Figura 2 - Exemplo de variável target não explícita em base de dados.
Fonte: Elaborado pela autora (2023)

Observe, na figura 2, que não temos a coluna “Mau Pagador” criada, o que pode ser muito comum em projetos reais, pois nem todas as bases de dados estão


configuradas perfeitamente para um modelo de aprendizado de máquina. Você, como uma pessoa profissional em dados, deve criar estratégias com base nos dados disponíveis para criar a variável target (a resposta) para os dados do seu modelo.

Para solucionar o caso do exemplo da figura 2, poderíamos criar uma estratégia analisando, por exemplo, a coluna de taxa de desemprego e dias inadimplentes, para classificar se essa pessoa é uma boa ou má pagadora. Veja a função abaixo:

```
def classificar_mau_pagador(taxa_desemprego, dias_inadimplencia):
    if taxa_desemprego < 50 and dias_inadimplencia > 100:
        return "Sim"
    else:
        return "Não"
```

Se aplicarmos essa função sobre o dataset com a função **apply()**, teríamos a variável target criada e pronta para ser utilizada para treinar o modelo. Claro que, nessa situação, poderia até mesmo pensar em padronização e deixar essa coluna target em formato binário, ao invés de classes categóricas como “sim” e “não”. Essa é uma estratégia de **feature scaling** e comentamos sobre a sua importância na aula 1.

Idade	Renda	Taxa de desemprego	Total de dias inadimplente	Total Debito Devedor
22	1500	0%	-1	150
55	4000	0%	-200	30000
44	0	100%	-300	45000



Mau Pagador?
Não
Sim
Sim

Figura 3 - Aplicando a target na base de dados.
Fonte: Elaborado pela autora (2023)

Para criar a variável target, você pode construir medidas auxiliares que podem te ajudar na estratégia da construção da target. Pensaremos aqui no case que foi trabalhado nos vídeos dessa aula. Perceba que temos algumas **medidas auxiliares** dentro do mercado de análise de crédito, que realmente são utilizadas para a tomada de decisão de concessão de crédito:

- **Abertura:** para saber em quanto tempo o cliente estava com registros na base de dados.

- **Final:** para saber o último mês de registro (mês atual).

- **Janela:** para saber o período de registros (subtraindo Final de Abertura).

- **MOB: Month on book** - medida bem específica do mercado financeiro.

Quando alguém vai solicitar um crédito e o banco abre esse crédito, é contado no balanço mensal, lançando na parte contábil do banco que aquele crédito foi concedido. Esse crédito só sai do balanço do banco quando a pessoa termina de pagar. É uma informação bem útil para saber quem são os bons ou maus pagadores.

Além das medidas auxiliares acima, vamos criar uma análise vintage. Essa análise é bem específica em análise de crédito. Para realizá-la, além das colunas que utilizamos anteriormente, vamos agrupar as colunas **Abertura** e **MOB**. Com isso, nós unimos todos os clientes de acordo com a época que a conta foi aberta e o MOB (mês no livro).

O QUE É ANÁLISE VINTAGE?

O termo “Vintage” se refere ao mês ou trimestre em que a conta foi aberta, ou seja, o período em que o empréstimo ou cartão de crédito foi concedido. No caso do nosso dataset, temos informações até o período de 60 meses. Portanto, a análise vintage mede o desempenho de uma carteira em diferentes períodos de tempo, após a concessão do empréstimo ou cartão de crédito. O desempenho pode ser medido na forma de taxa de baixa cumulativa, proporção de clientes com atraso de 30/60/90 dias, taxa de utilização, saldo médio etc.

A análise vintage pode ser usada para uma variedade de propósitos. Alguns deles são:

- Identificar se as contas abertas em um determinado mês ou trimestre são mais arriscadas do que outras;
- Determinar o período ideal da janela de desempenho no desenvolvimento do credit scoring;
- Monitorar ou rastrear o risco de um portfólio.

COMO A ANÁLISE VINTAGE É USADA NA MODELAGEM DE RISCO DE CRÉDITO?

Ela é usada para determinar o número de meses que você deve considerar para a janela de desempenho. Se o cliente entrar em inadimplência, por exemplo, por 60 dias ou mais em atraso, durante a janela de desempenho, ele será considerado um cliente **mau pagador**.

O próximo passo é calcular a porcentagem cumulativa de clientes mau pagadores em relação aos meses no livro (MOB - Months on book), que é o número de meses concluídos desde a data de concessão do crédito.

ANÁLISE DE TENDÊNCIAS

Também podemos analisar tendências utilizando a análise vintage para, por exemplo, checar o saldo médio para clientes abertos em trimestres diferentes e analisar sua tendência nos meses subsequentes, após a data de abertura da conta.

No gráfico da figura 4, mostramos um dos resultados da nossa análise vintage realizada em aula. Temos o MOB no eixo X e a porcentagem acumulada de maus pagadores no eixo Y. Cada linha colorida representa o período de abertura do crédito. Nesse caso, para facilitar, estamos visualizando apenas os 10 últimos períodos de abertura. Nos primeiros meses do MOB, temos uma subida e grande variação nos dados, que vão se estabilizando ao longo do tempo. Nesse cenário, vemos que a abertura de 55 meses foi aquela que mostrou uma maior porcentagem acumulada de maus pagadores.

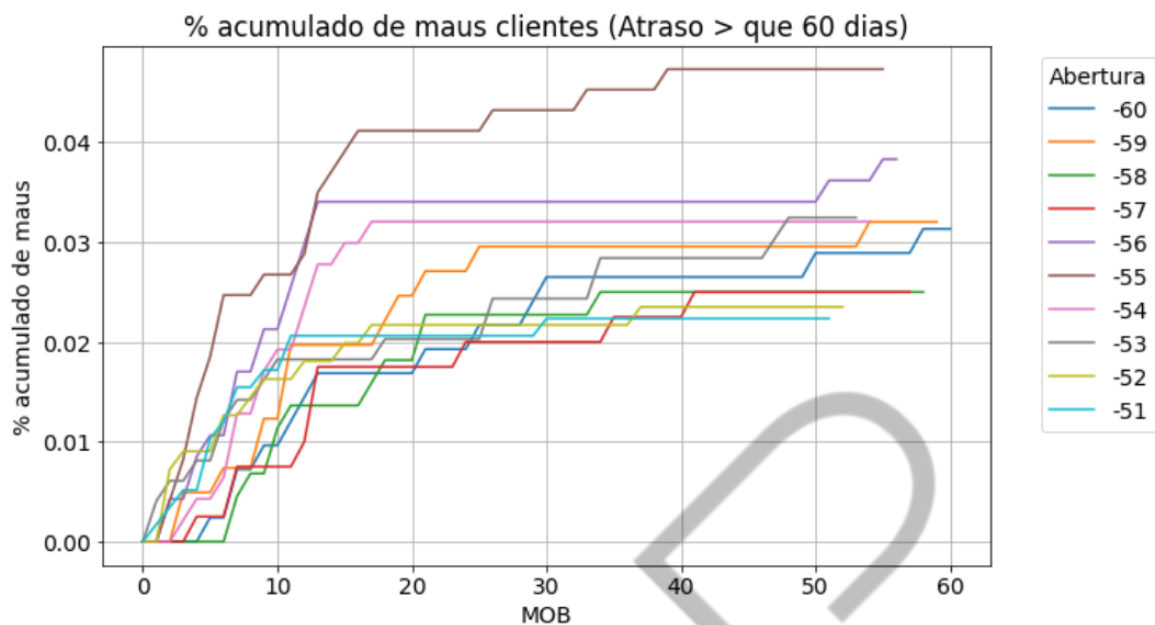


Figura 4 - MOB.
Fonte: Elaborado pela autora (2023)

Com isso em mente, podemos concluir que a análise vintage é um método amplamente utilizado para gerenciar o risco de crédito, pois ilustra o comportamento após a abertura de uma conta. Então, com base no mesmo período de criação, calcula o índice de baixa de uma carteira de crédito.

VARIÁVEL TARGET PARA REGRESSÃO

Exploramos as situações para as variáveis targets para classificação, mas como funciona para regressão? Para as regressões, como nesse caso, em que estamos prevendo valores, basicamente podemos encontrar a mesma situação que a classificação: o valor já pode existir na base de dados ou esse valor pode ser uma combinação de “n” fatores nos dados (várias colunas). Vale lembrar que não existe uma receita de bolo para a criação da target e problemas de negócios diferentes vão trazer soluções diferentes, dependendo da natureza dos dados e do objetivo a ser solucionado com o algoritmo.

A IMPORTÂNCIA DA TARGET NO APRENDIZADO DE MÁQUINA

Quando estamos falando de modelos supervisionados, a variável target é a **resposta do que queremos encontrar com o algoritmo**. Ela será utilizada na etapa de **treinamento** do algoritmo, para “ensinar” a máquina (o algoritmo) a aprender os

padrões contidos nos dados. Após realizar a aplicação dos algoritmos, a variável target também tem muita importância na etapa de validação e teste da efetividade do algoritmo.

Podemos encontrar vários tipos de variáveis target dentro da modelagem de algoritmos, tais como:

- **Variável target numérica contínua:** variável que pode assumir qualquer valor numérico em um intervalo contínuo. Nesse exemplo, podemos citar a regressão linear, onde estamos buscando prever o valor de um imóvel com base em algumas características.
- **Variável target numérica discreta:** variável que assume valores inteiros. Podemos assumir aqui, por exemplo, a predição da idade de uma pessoa com base em algumas características de comportamento. Nesse exemplo, as regressões também costumam solucionar esse tipo de problema.
- **Variável target binária:** uma das variáveis mais conhecidas, ela possui apenas dois tipos de classificação, codificados por 0 e 1. Exemplo: criar um modelo para prever se um estudante é evadido ou não, com base nas características acadêmicas (Evadido =0 e Ativo =1).
- **Variável target categórica:** quando falamos de variáveis categóricas, podemos ter dois cenários: categóricas binárias ou multiclases. As binárias vão ter sempre a possibilidade entre duas classes, como **sim e não**, por exemplo, onde você consegue aplicar técnicas de feature scaling para transformar os dados e trabalhar com as categorias. A variável categórica de multiclases é o tipo onde sua target pode ter mais de duas classes contidas dentro das possibilidades de classes.
- **Variável target categórica ordinal:** variável target onde a disposição das categorias possui uma ordenação natural dos dados, como níveis de educação, classe social, entre outros. Esse tipo de target pode ser desafiador, pois encontrar o equilíbrio entre as multiclases nos modelos nem sempre é uma tarefa fácil, por conta do possível desequilíbrio de classes da sua amostra de dados.

O QUE VOCÊ VIU NESTA AULA?

Nessa aula você aprendeu a importância e os tipos de variáveis targets nos modelos de aprendizado de máquina. Esse passo é fundamental para a decisão da resposta que queremos encontrar com os dados.

Tem alguma dúvida ou quer conversar sobre o tema desta aula? Entre em contato conosco pela comunidade do Discord! Lá você pode fazer networking, receber avisos, tirar dúvidas e muito mais.

EMANDA

REFERÊNCIAS

BRUCE, A. **Estatística Prática para Cientistas de Dados**. Sebastopol: O'Reilly Media, 2019.

DOCUMENTAÇÃO SCIKIT-LEARN. **Scikit-learn**. Disponível em: <<https://scikit-learn.org/stable/>>. Acesso em: 04 Ago 2023.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. Sebastopol: O'Reilly Media, 2019

PALAVRAS-CHAVE

Variável Target, Target, Deploy de aplicações.

EMENDAS

The background is a dark blue field filled with numerous small, light blue dots. Overlaid on this are several large, wavy, translucent lines in shades of blue and yellow. A vertical line with a small 'x' at the bottom is on the left. A circle containing the number '7' is in the upper center. A hexagon is in the lower right.

POSTECH