

Analytics & Inteligência Artificial

Tema da aula
Tópicos Especiais em Regressão



BUSINESS SCHOOL

Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In Company e EAD



CONSULTING

Consultoria personalizada que oferece soluções baseada em seu problema de negócio



RESEARCH

Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino



Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais [escolas de negócios do mundo](#), possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 [projetos de consultorias](#) em organizações públicas e privadas.



Único curso de graduação em administração a receber as notas máximas



A primeira escola brasileira a ser finalista da maior competição de MBA do mundo



Única Business School brasileira a figurar no ranking LATAM



Signatária do Pacto Global da ONU



Membro fundador da ANAMBA - Associação Nacional MBAs



Credenciada pela AMBA - Association of MBAs



Credenciada ao Executive MBA Council



Filiada a AACSB - Association to Advance Collegiate Schools of Business



Filiada a EFMD - European Foundation for Management Development



Referência em cursos de MBA nas principais mídias de circulação

O Laboratório de Análise de Dados – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data, Analytics e Inteligência Artificial***.



Profª Drª Alessandra
Martin

O LABDATA é um dos pioneiros no lançamento dos cursos de ***Big Data e Analytics*** no Brasil

Os diretores foram professores de grandes especialistas do mercado
+10 anos de atuação
+1000 alunos formados



Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R\$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Conteúdo da Aula

- 1. Segmentação aplicada à Modelagem Preditiva
- 1. Otimização do Processo de Categorização de Variáveis
- 1. Ponto de Corte e Estratégia de Decisão
- 1. Validação e monitoramento de modelos preditivos
- 1. *Overfitting* de modelos preditivos

1. Segmentação aplicada à modelagem preditiva



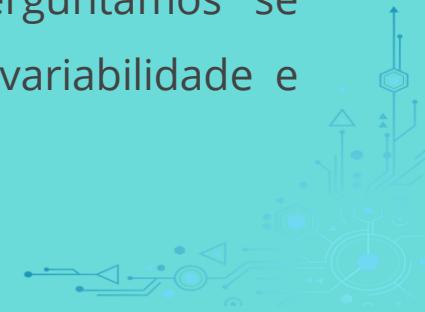
Um único modelo preditivo é suficiente?

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

6



Muitas vezes desenvolvemos modelos preditivos e nos perguntamos se deveríamos segmentar esses modelos, de forma a reduzir a variabilidade e aumentar a possibilidade de acerto dos modelos.



Créditos: <https://pixabay.com/pt/photos/laptop-computador-navegador-2562325/>

@2020 LABDATA FIA. Copyright all rights reserved.

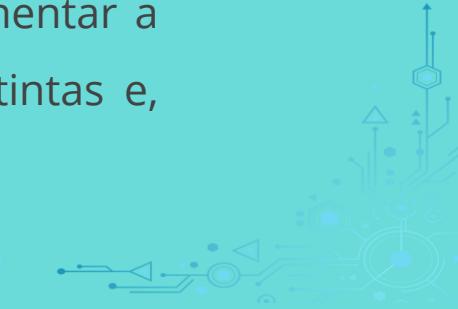
Um único modelo preditivo é suficiente?

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

7



Uma boa possibilidade de ação (não exaustiva e nem infalível) para isso é, antes de desenvolver modelos preditivos, tentar segmentar a base de dados, identificando grupos com características distintas e, depois, desenvolver modelos para cada grupo.



Créditos: <https://pixabay.com/pt/photos/gel%C3%A9ia-marmelada-doces-a%C3%A7%C3%BAcar-3032344/>

@2020 LABDATA FIA. Copyright all rights reserved.

Proposição: Segmentar para melhor separar/predizer

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

8



Não há garantia na melhoria de resultados, mas é uma boa iniciativa para reduzir a variabilidade nos dados e gerar modelos mais direcionados.

Créditos: Professor Marcelo Fernandes, FIA

@2020 LABDATA FIA. Copyright all rights reserved.



Case: People Analytics – Turnover de funcionários

CASE | REGRESSÃO LOGÍSTICA

9

Na base **SEGMENTACAO DE MODELOS HR ANALYTICS.xlsx**, estão disponíveis 14.999 registros de informações sobre funcionários de uma empresa.

Seja **left** o evento de interesse (1=sim, 0=não), que indica se o funcionário saiu ou não da empresa. Nosso interesse é, a partir das variáveis explicativas a seguir, tentar explicar o evento “saída da empresa”:



- **satisfaction_level**: Nível de satisfação do funcionário
- **last_evaluation**: Resultado da última avaliação
- **number_project**: Número de projetos desenvolvidos
- **average_montly_hours**: Média de horas trabalhadas por mês
- **time_spend_company**: tempo na empresa (em anos)
- **work_accident**: se teve ou não acidente de trabalho
- **promotion_last_5years**: se teve ou não promoção nos últimos 5 anos
- **sales**: área em que trabalha o funcionário
- **salary**: nível salarial
- **left**: 1=saiu, 0=não saiu



Case: People Analytics – Turnover de funcionários

CASE: BASE INTEIRA – SEM SEGMENTAÇÃO | REGRESSÃO LOGÍSTICA

10

Como ponto de partida, desenvolvemos o modelo de regressão logística para a base inteira, sem considerar a segmentação.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.4762862	0.1938373	-7.616	2.61e-14	***
satisfaction_level	-4.1356889	0.0980538	-42.178	< 2e-16	***
last_evaluation	0.7309032	0.1491787	4.900	9.61e-07	***
number_project	-0.3150787	0.0213248	-14.775	< 2e-16	***
average_montly_hours	0.0044603	0.0005161	8.643	< 2e-16	***
time_spend_company	0.2677537	0.0155736	17.193	< 2e-16	***
work_accident	-1.5298283	0.0895473	-17.084	< 2e-16	***
promotion_last_5years	-1.4301364	0.2574958	-5.554	2.79e-08	***
saleshr	0.2323779	0.1313084	1.770	0.07678	.
salesIT	-0.1807179	0.1221276	-1.480	0.13894	
salesmanagement	-0.4484236	0.1598254	-2.806	0.00502	**
salesmarketing	-0.0120882	0.1319304	-0.092	0.92700	
salesproduct_mng	-0.1532529	0.1301538	-1.177	0.23901	
salesRandD	-0.5823659	0.1448848	-4.020	5.83e-05	***
salessales	-0.0387859	0.1024006	-0.379	0.70486	
salessupport	0.0500251	0.1092834	0.458	0.64713	
salestechnical	0.0701464	0.1065379	0.658	0.51027	
salarylow	1.9440627	0.1286272	15.114	< 2e-16	***
salarymedium	1.4132244	0.1293534	10.925	< 2e-16	***

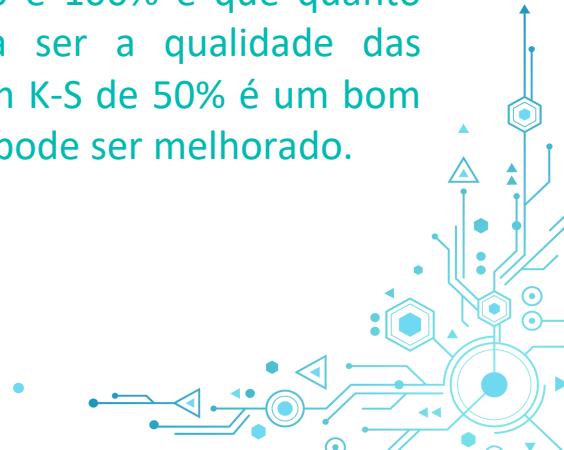
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

Two-sample Kolmogorov-Smirnov test

data: HR\$HR_pred[HR\$left == 0] and HR\$HR_pred[HR\$left == 1]
D = 0.50234, p-value < 2.2e-16
alternative hypothesis: two-sided



O modelo único para a base inteira gerou um K-S de 50,23%. Lembrem-se que o K-S varia entre 0 e 100% e que quanto maior esse valor, melhor tende a ser a qualidade das previsões do modelo. Nesse caso, um K-S de 50% é um bom resultado, mas que, potencialmente, pode ser melhorado.



Case: People Analytics – Turnover de funcionários

CASE: DESCRIPTIVA POR SEGMENTO | REGRESSÃO LOGÍSTICA

11

Agora, vamos usar a mesma planilha de **SEGMENTACAO DE MODELOS HR ANALYTICS.xlsx**, pasta **BASE_DADOS_CLUSTER**, que traz um campo adicional chamado “cluster”, que contém uma segmentação dos funcionários em 3 grupos distintos. Mais detalhes sobre esse processo de segmentação serão vistos mais adiante no curso, na disciplina de “Análise de clusters ou aprendizado não supervisionado”

ID	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary	cluster
1	1	0.38	0.53	2	157	3	0	1	0 sales	low	3
2	2	0.80	0.86	5	262	6	0	1	0 sales	medium	1
3	3	0.11	0.88	7	272	4	0	1	0 sales	medium	1
4	4	0.72	0.87	5	223	5	0	1	0 sales	low	2
5	5	0.37	0.52	2	159	3	0	1	0 sales	low	3
6	6	0.41	0.50	2	153	3	0	1	0 sales	low	3
7	7	0.10	0.77	6	247	4	0	1	0 sales	low	1
8	8	0.92	0.85	5	259	5	0	1	0 sales	low	2
9	9	0.89	1.00	5	224	5	0	1	0 sales	low	2
10	10	0.42	0.53	2	142	3	0	1	0 sales	low	3
11	11	0.45	0.54	2	135	3	0	1	0 sales	low	3
12	12	0.11	0.81	6	305	4	0	1	0 sales	low	1
13	13	0.84	0.92	4	234	5	0	1	0 sales	low	2
14	14	0.41	0.55	2	148	3	0	1	0 sales	low	3
15	15	0.36	0.56	2	137	3	0	1	0 sales	low	3
16	16	0.38	0.54	2	143	3	0	1	0 sales	low	3
17	17	0.45	0.47	2	160	3	0	1	0 sales	low	3
18	18	0.78	0.99	4	255	6	0	1	0 sales	low	2
19	19	0.45	0.51	2	160	3	1	1	1 sales	low	3
20	20	0.76	0.89	5	262	5	0	1	0 sales	low	2
21	21	0.11	0.83	6	282	4	0	1	0 sales	low	1
22	22	0.38	0.55	2	147	3	0	1	0 sales	low	3
23	23	0.09	0.95	6	304	4	0	1	0 sales	low	1
24	24	0.46	0.57	2	139	3	0	1	0 sales	low	3
25	25	0.40	0.53	2	158	3	0	1	0 sales	low	3
26	26	0.89	0.92	5	242	5	0	1	0 sales	low	2
27	27	0.82	0.87	4	239	5	0	1	0 sales	low	2
28	28	0.40	0.49	2	135	3	0	1	0 sales	low	3
29	29	0.41	0.46	2	128	3	0	1	0 accounting	low	3
30	30	0.38	0.50	2	132	3	0	1	0 accounting	low	3
31	31	0.09	0.62	6	294	4	0	1	0 accounting	low	1
32	32	0.45	0.57	2	134	3	0	1	0 hr	low	3
33	33	0.40	0.51	2	145	3	0	1	0 hr	low	3



A variável “cluster” contém 3 grupos distintos (1, 2 e 3). Na sequência, vamos avaliar o resultado de modelos construídos para cada grupo em específico.



Case: People Analytics – Turnover de funcionários

CASE: MODELO POR SEGMENTO | REGRESSÃO LOGÍSTICA

12

Agora, vamos montar um modelo para cada um dos 3 clusters construídos e avaliar o K-S de cada um dos modelos, bem como uma média ponderada de K-S.

```
> ks_cluster_I
```

Two-sample Kolmogorov-Smirnov test

Modelo específico para o cluster I apresentou um K-S de 79,26%.

```
data: HR_cluster_I$HR_pred_cluster_I[HR_cluster_I$left == 0] and HR_cluster_I$HR_pred_cluster_I[HR_cluster_I$left == 1]  
D = 0.79255, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

```
> ks_cluster_II
```

Two-sample Kolmogorov-Smirnov test

Modelo específico para o cluster II apresentou um K-S de 81,54%.

```
data: HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left == 0] and HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left == 1]  
D = 0.81541, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

```
> ks_cluster_III
```

Two-sample Kolmogorov-Smirnov test

Modelo específico para o cluster III apresentou um K-S de 87,60%.

```
data: HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left == 0] and HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left == 1]  
D = 0.87602, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

```
> #K-S ponderado  
> round(((nrow(HR_cluster_I)*ks_cluster_I$statistic +  
+ nrow(HR_cluster_II)*ks_cluster_II$statistic +  
+ nrow(HR_cluster_III)*ks_cluster_III$statistic)/nrow(HR_cluster)),4)
```

D
0.8307

K-S ponderado dos 3 modelos foi de 83,07%, resultado bastante superior ao observado com um modelo único para a base toda.



Exercício: Empréstimo Bancário

CASE: MODELO POR SEGMENTO | REGRESSÃO LOGÍSTICA

13

A base **SEGMENTACAO DE MODELOS EMPBANC.xlsx** é composta por 5.000 empréstimos distintos que foram concedidos a clientes e o objetivo é testar se a segmentação tem um efeito positivo na geração de modelos preditivos mais precisos, quando comparado a um cenário de um único modelo para a base inteira.



idade - idade do cliente (em anos)

Tempo_endereco - tempo no mesmo endereço (em anos)

renda - renda mensal (em R\$)

cred_deb - Razão entre seus créditos e débitos (razão adimensional)

classif - Se pagou o não o empréstimo bancário (0=pagou, 1=não pagou)

cluster – Segmentação dos clientes bancários (3 grupos distintos, a serem usados para o desenvolvimento de modelos segmentados)

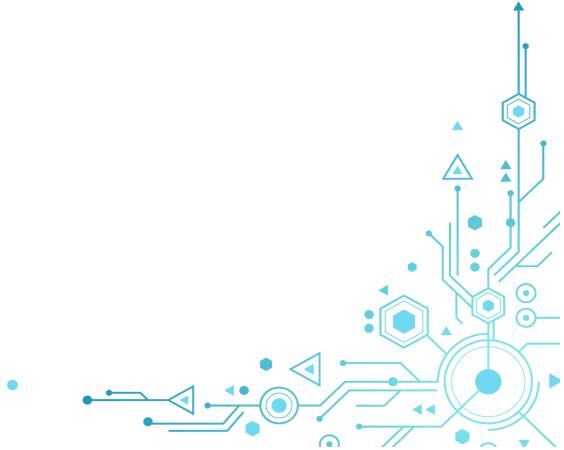


a) Desenvolva um modelo único (sem considerar a variável "cluster") e calcule o K-S

b) Faça um modelo para cada segmento da variável "cluster" , calcule o K-S de cada modelo e calcule o K-S ponderado. Quais as suas impressões?



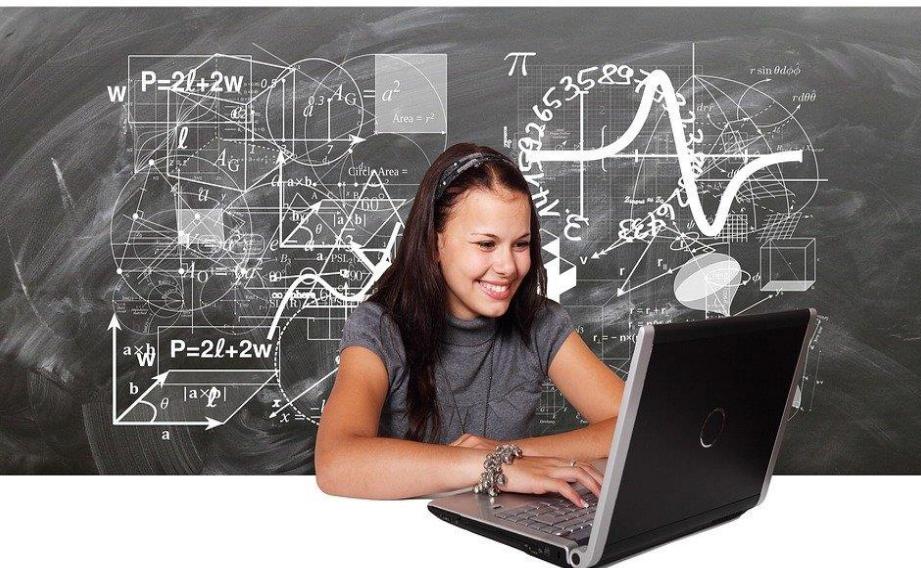
2. Otimização de Processos de Categorização de Variáveis



Como melhor categorizar uma variável quantitativa?

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

15



Ao desenvolver modelos, por vezes, uma forte tentação é querer já colocar as variáveis no modelo (do jeito mesmo que estão) e avaliar seu resultado.

Créditos: <https://pixabay.com/pt/photos/aprenda-escola-estudante-matem%C3%A1tica-1996845/>

@2020 LABDATA FIA. Copyright all rights reserved.

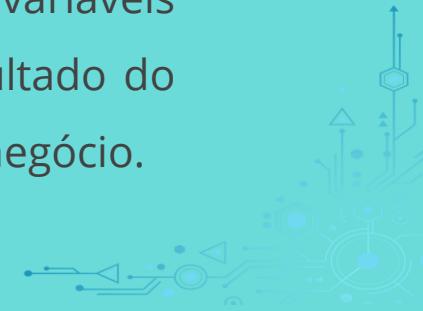
Como melhor categorizar uma variável quantitativa?

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

16



No entanto, algumas medidas específicas de tratamento das variáveis podem proporcionar experiências interessantes, seja no resultado do modelo, seja na interpretação dos resultados por pessoas do negócio.

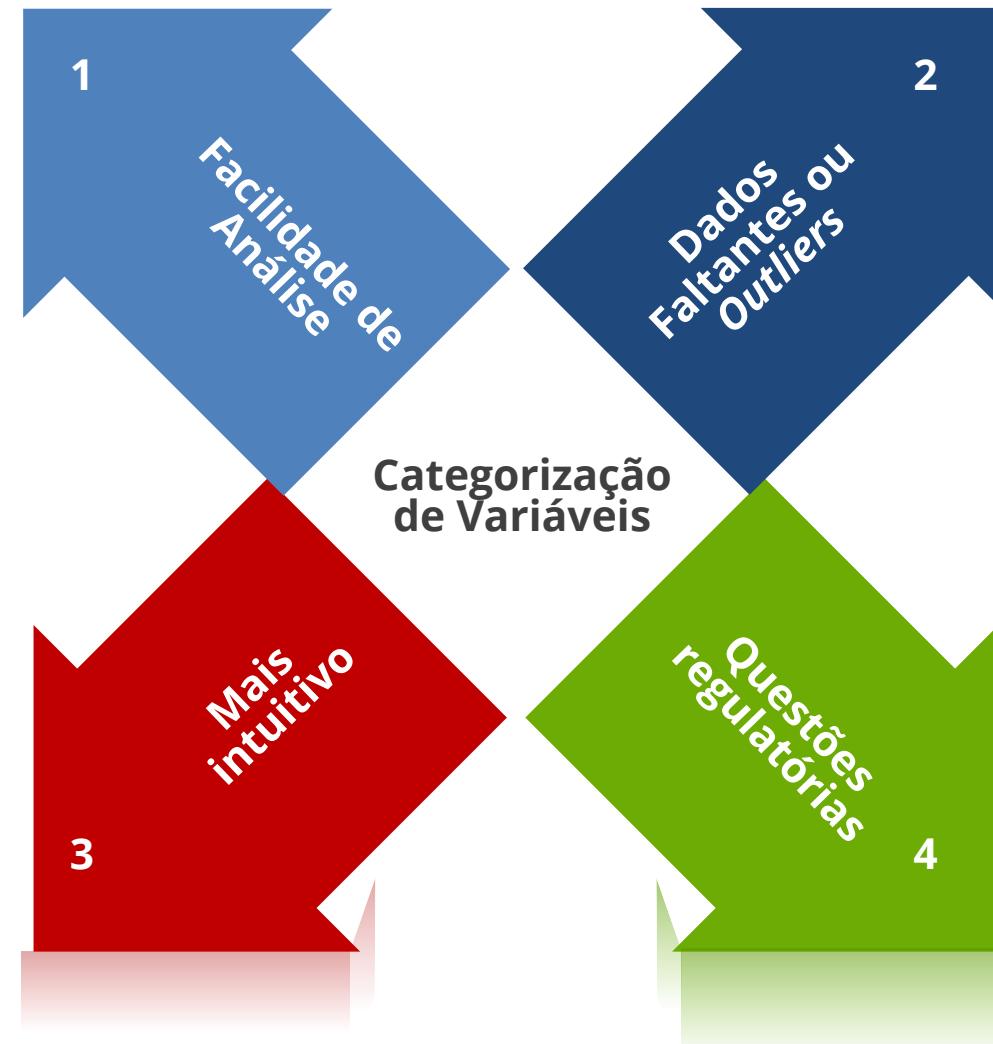


Créditos: <https://pixabay.com/pt/photos/mulheres-trabalho-em-equipe-equipe-1209678/>

@2020 LABDATA FIA. Copyright all rights reserved.

Por que categorizar pode ser uma boa estratégia?

Uma variável dividida em categorias (“*bins*”) é mais fácil de analisar e interpretar.



É mais fácil para explicar o modelo para não-modeladores (*C-level* ou outras pessoas de negócio)

Com variáveis categóricas, é mais fácil isolar o efeito de dados extremos (*outliers*) ou faltantes (*missing data*).

No caso de modelos de risco, é mais fácil estabelecer os chamados “*reason codes*”.

Tipos de categorização

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

18

1

A categorização subjetiva pode não ser bacana por concentrar demais os dados numa só classe ou não separar bem os dados

4

Identifica a quantidade e separação ideal entre as classes que minimiza perda de informação



2

É bacana, pois é possível separar de forma equilibrada os dados em um número quaisquer de classes, baseando-se em quartis, decis ou percentis.

3

Consegue identificar as classes com características mais parecidas (quando associadas a uma variável resposta como referência)



Por que categorizar pode ser uma boa estratégia?

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

19

The screenshot shows a presentation slide with a background image of modern skyscrapers. The title 'Why use binned variables in predictive models?' is displayed in white text. Below it, the date 'February 22, 2018' and the speaker's name 'Sue Hubbard' are shown. A small note at the bottom left states: '© 2018 Fair Isaac Corporation. Confidential. This presentation is provided for the recipient only and cannot be reproduced or shared without Fair Isaac Corporation's express consent.' The FICO Decisions logo is visible in the top right corner. The video player interface at the bottom includes a progress bar showing '0:02 / 48:41'.

Fonte: <https://www.youtube.com/watch?v=dA9ZK02eD9M>

@2020 LABDATA FIA. Copyright all rights reserved.

Nesse vídeo de cerca de 49min de duração, realizado por consultores da FICO (www.fico.com) , uma empresa americana de softwares de gestão de processos de decisão, responsável pelos primeiros modelos de *credit scoring* no mundo, são debatidos, de maneira abrangente diversos benefícios associados à categorização de variáveis quantitativas ("binning") , assim como seus impactos ao negócio.



Optimal Binning ou processo de otimização de categorias

CATEGORIZAÇÃO DE COVARIÁVEIS | REGRESSÃO LOGÍSTICA

20

Revolutions

Daily news about using open source R for big data analysis, predictive modeling, data science, and visualization since 2008

[« Tomorrow, 10AMPT: Live webinar on "checkpoint" package](#) | [Main](#) | [Participate in the 2015 Rexter Data Mining Survey »](#)

March 24, 2015

R Package 'smbinning': Optimal Binning for Scoring Modeling

by Herman Jopia

What is Binning?

Binning is the term used in scoring modeling for what is also known in Machine Learning as **Discretization**, the process of transforming a continuous characteristic into a finite number of intervals (the bins), which allows for a better understanding of its distribution and its relationship with a binary variable. The bins generated by this process will eventually become the attributes of a predictive characteristic, the key component of a Scorecard.

Why Binning?

Though there are some reticence to it [1], the benefits of binning are pretty straight forward:

- It allows missing data and other special calculations (e.g. divided by zero) to be included in the model.
- It controls or mitigates the impact of outliers over the model.
- It solves the issue of having different scales among the characteristics, making the weights of the coefficients in the final model comparable.

Unsupervised Discretization

Unsupervised Discretization divides a continuous feature into groups (bins) without taking into account any other information. It is basically a partition with two options: equal length intervals and equal frequency intervals.

Equal length intervals

- Objective: Understand the distribution of a variable.
- Example: The classic histogram, whose bins have equal length that can be calculated using different rules (Sturges, Rice, and others).
- Disadvantage: The number of records in a bin may be too small to allow for a valid calculation, as shown in Table 1.

Characteristic	:	Time since first account was open	→ Integer									
Target	:	Credit performance within the next 12 months (0:Bad / 1:Good)	→ Binary									
Binning Method	:	Equal Length Intervals										
Cutpoint	CntRec	CntGood	CntBad	CntCumRec	CntCumGood	CntCumBad	PctRec	BadRate	Odds	In Odds	NoE	IV
1	<= 50	6074	5209	865	6074	5209	865	0.5097	0.1424	6.0220	1.7954	-0.4251 -0.4608
2	<= 100	2783	2654	129	8057	7863	994	0.2335	0.0464	20.5736	3.0240	0.8035 0.4123
3	<= 150	1257	1213	44	10114	9076	1038	0.1055	0.0350	27.5682	3.3167	1.0962 0.2492
4	<= 200	513	502	11	10627	9578	1049	0.0430	0.0214	45.6364	3.8207	1.6002 0.1424
5	<= 250	86	85	1	10713	9663	1050	0.0072	0.0116	85.0000	4.4427	2.2222 0.0313
6	<= 300	4	4	0	10717	9667	1050	0.0003	0.0000	Inf	Inf	Inf
7	Missing	1200	1083	117	11917	10750	1167	0.1007	0.0975	9.2564	2.2253	0.0048 0.0011



Information

[About this blog](#)
[Comments Policy](#)
[About Categories](#)
[About the Authors](#)
[Local R User Group Directory](#)
[Tips on Starting an R User Group](#)

Search Revolutions Blog

 Search Blog

Got comments or suggestions for the blog editor?
Email [David Smith](#).

Follow David on Twitter: [@revodavid](#)
[+David Smith](#)

Get this blog via email with [Blogtrottr](#)

Categories

[academia](#)
[advanced tips](#)
[AI](#)
[airoundups](#)
[announcements](#)
[applications](#)
[beginner tips](#)
[big data](#)
[courses](#)
[current events](#)
[data science](#)
[developer tips](#)
[events](#)
[finance](#)
[government](#)
[graphics](#)

Consiste em determinar a quantidade e o intervalo das categorias, a partir de uma variável quantitativa, de forma a minimizar perda do valor da informação da variável (IV). Não é um método infalível, então, nem sempre o IV da categoria otimizada é maior que o da variável original.



Case: *People Analytics – Turnover de funcionários*

OPTIMAL BINNING | REGRESSÃO LOGÍSTICA

21

Na base **OTIMIZACAO_CATEGORIAS_HR_ANALYTICS.xlsx**, estão disponíveis 14.999 registros de informações sobre funcionários de uma empresa. Seja **left** o evento de interesse (1=sim, 0=não), que indica se o funcionário saiu ou não da empresa. Nossa interesse é, a partir das variáveis explicativas a seguir, tentar explicar o evento “saída da empresa”:



- **satisfaction_level**: Nível de satisfação do funcionário
- **last_evaluation**: Resultado da última avaliação
- **number_project**: Número de projetos desenvolvidos
- **average_montly_hours**: Média de horas trabalhadas por mês
- **time_spend_company**: tempo na empresa (em anos)
- **work_accident**: se teve ou não acidente de trabalho
- **promotion_last_5years**: se teve ou não promoção nos últimos 5 anos
- **sales**: área em que trabalha o funcionário
- **salary**: nível salarial
- **left**: 1=saiu, 0=não saiu



Vamos testar o processo de categorização ótima da variável quantitativa “**satisfaction level**” nessa base.



Case: *People Analytics – Turnover de funcionários*

OPTIMAL BINNING | REGRESSÃO LOGÍSTICA

22

Antes disso, vamos calcular o IV (valor da informação) original de todas as variáveis presentes na base, apenas como termos nossa referência em termos de valor preditivo das variáveis. É possível fazer isso por meio do **pacote "Information"**, disponível no R.

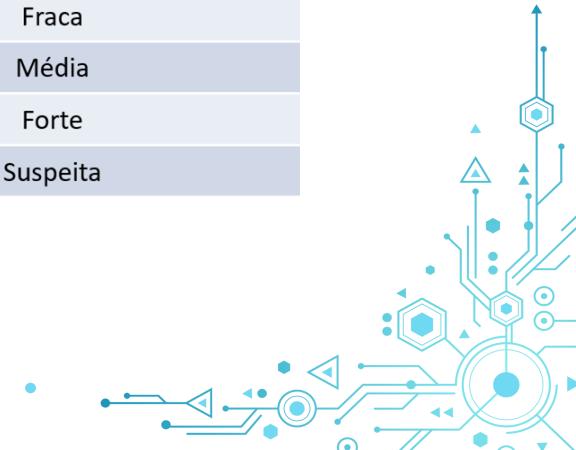
Os resultados são os que aparecem a seguir:

```
$Summary
      variable       IV
2 satisfaction_level 2.27130262
4      number_project 1.97240680
5 average_montly_hours 1.14238678
6 time_spend_company 0.92658691
3     last_evaluation 0.90652799
1             ID 0.42081312
7      work_accident 0.18535538
10            salary 0.17904981
9              sales 0.03561297
8 promotion_last_5years 0.03385306
```

O R já ordena as variáveis, da mais forte, para a mais fraca, de acordo com seu IV - Information Value (ou valor da informação). IV's maiores que 0.5 indicam variáveis muito fortes e abaixo de 0.1 indicam variáveis fracas. **Vejam que interessante o IV da variável “satisfaction_level”**.



Valor da informação (VI)	Classificação
<=0,02	Fraquíssima
Entre 0,02 e 0,10	Fraca
Entre 0,10 e 0,30	Média
Entre 0,30 e 0,50	Forte
> 0,50	Suspeita



Case: *People Analytics – Turnover de funcionários*

OPTIMAL BINNING | REGRESSÃO LOGÍSTICA

23

Usando o **pacote “smbinning”**, disponível no R, podemos testar se é possível uma categorização interessante da variável “satisfaction level”. A partir dela, determinamos o IV (valor da informação) correspondente.



Os resultados são os que aparecem a seguir:

```
> satlevel_optbin$ivtable
```

	Cutpoint	CntRec	CntGood	CntBad	CntCumRec	CntCumGood	CntCumBad	PctRec	GoodRate	BadRate	Odds	LnOdds	WoE	IV
1	<= 0.11	888	888	0	888	888	0	0.0592	1.0000	0.0000	Inf	Inf	Inf	Inf
2	<= 0.35	1283	94	1189	2171	982	1189	0.0855	0.0733	0.9267	0.0791	-2.5376	-1.3744	0.1068
3	<= 0.46	2012	1549	463	4183	2531	1652	0.1341	0.7699	0.2301	3.3456	1.2076	2.3709	0.9324
4	<= 0.71	4689	111	4578	8872	2642	6230	0.3126	0.0237	0.9763	0.0242	-3.7195	-2.5563	0.9446
5	<= 0.8	2125	380	1745	10997	3022	7975	0.1417	0.1788	0.8212	0.2178	-1.5243	-0.3611	0.0167
6	<= 0.91	2442	529	1913	13439	3551	9888	0.1628	0.2166	0.7834	0.2765	-1.2854	-0.1222	0.0024
7	> 0.91	1560	20	1540	14999	3571	11428	0.1040	0.0128	0.9872	0.0130	-4.3438	-3.1806	0.4108
8	Missing	0	0	0	14999	3571	11428	0.0000	NaN	NaN	NaN	NaN	NaN	NaN
9	Total	14999	3571	11428	NA	NA	NA	1.0000	0.2381	0.7619	0.3125	-1.1632	0.0000	2.4137

Vejam que interessante, o processo de categorização ótima sugeriu 7 classes distintas para a variável “satisfaction level” e, a partir dessa categorização, seu IV saiu de 2.2713 para 2.4137, uma melhora, em termos de IV, de 6,27%.



Case: People Analytics – Turnover de funcionários

OPTIMAL BINNING | REGRESSÃO LOGÍSTICA

24

Podemos incluir a variável “satisfaction_level” categorizada na base original e reprocessar o cálculo do IV, para mostrar o novo ranking das variáveis mais fortes da base:

ID	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary	satlevel_opt
1	1	0.38	0.53	2	157	3	0	1	0 sales	low	03 <= 0.46
2	2	0.80	0.86	5	262	6	0	1	0 sales	medium	05 <= 0.8
3	3	0.11	0.88	7	272	4	0	1	0 sales	medium	01 <= 0.11
4	4	0.72	0.87	5	223	5	0	1	0 sales	low	05 <= 0.8
5	5	0.37	0.52	2	159	3	0	1	0 sales	low	03 <= 0.46
6	6	0.41	0.50	2	153	3	0	1	0 sales	low	03 <= 0.46
7	7	0.10	0.77	6	247	4	0	1	0 sales	low	01 <= 0.11
8	8	0.92	0.85	5	259	5	0	1	0 sales	low	07 > 0.91
9	9	0.89	1.00	5	224	5	0	1	0 sales	low	06 <= 0.91
10	10	0.42	0.53	2	142	3	0	1	0 sales	low	03 <= 0.46
11	11	0.45	0.54	2	135	3	0	1	0 sales	low	03 <= 0.46
12	12	0.11	0.81	6	305	4	0	1	0 sales	low	01 <= 0.11
13	13	0.84	0.92	4	234	5	0	1	0 sales	low	06 <= 0.91
14	14	0.41	0.55	2	148	3	0	1	0 sales	low	03 <= 0.46
15	15	0.36	0.56	2	137	3	0	1	0 sales	low	03 <= 0.46
16	16	0.38	0.54	2	143	3	0	1	0 sales	low	03 <= 0.46
17	17	0.45	0.47	2	160	3	0	1	0 sales	low	03 <= 0.46
18	18	0.78	0.99	4	255	6	0	1	0 sales	low	05 <= 0.8
19	19	0.45	0.51	2	160	3	1	1	1 sales	low	03 <= 0.46
20	20	0.76	0.89	5	262	5	0	1	0 sales	low	05 <= 0.8
21	21	0.11	0.63	6	282	4	0	1	0 sales	low	01 <= 0.11
22	22	0.38	0.55	2	147	3	0	1	0 sales	low	03 <= 0.46
23	23	0.09	0.95	6	304	4	0	1	0 sales	low	01 <= 0.11
24	24	0.46	0.57	2	139	3	0	1	0 sales	low	03 <= 0.46
25	25	0.40	0.53	2	158	3	0	1	0 sales	low	03 <= 0.46
26	26	0.69	0.92	5	242	5	0	1	0 sales	low	06 <= 0.91
27	27	0.82	0.87	4	239	5	0	1	0 sales	low	06 <= 0.91
28	28	0.40	0.49	2	135	3	0	1	0 sales	low	03 <= 0.46

Showing 1 to 33 of 14,999 entries, 12 total columns

\$Summary

Variable	IV
11 satlevel_opt	2.41360025
2 satisfaction_level	2.27130262
4 number_project	1.97240680
5 average_montly_hours	1.14238678
6 time_spend_company	0.92658691
3 last_evaluation	0.90652799
1 ID	0.42081312
7 work_accident	0.18535538
10 salary	0.17904981
9 sales	0.03561297
8 promotion_last_5years	0.03385306

R Studio

Aqui, temos na mesma tabela, a variável “satisfaction_level” original e sua versão categorizada, com um aumento no nível de seu poder preditivo.

Exercício: *People Analytics – Turnover de funcionários*

OPTIMAL BINNING | REGRESSÃO LOGÍSTICA

25

Ainda com relação à base **OTIMIZACAO_CATEGORIAS_HR_ANALYTICS.xlsx**, a sugestão é avaliar outras variáveis quantitativas da base e testar a possibilidade de criação de variáveis categorizadas, bem como a comparação dos IV's antes e depois da categorização.



- a) Calcule o IV das variáveis `last_evaluation` e `average_montly_hours`, em seu formato original (sem categorização)
- b) Proponha uma categorização arbitrária para essas 2 variáveis (você pode definir as classes), calcule o IV de ambas e compare com o IV das variáveis originais. Comente seus achados.
- c) Proponha uma categorização ótima para essas variáveis **`last_evaluation`** e **`average_montly_hours`**. A quais conclusões você chega com relação a essa simulação nessas variáveis?
- d) Simule um modelo de regressão logística com as variáveis `satisfaction_level`, `last_evaluation` e `average_montly_hours`, com e sem variáveis categorizadas. Calcule o K-S e comente seus achados.



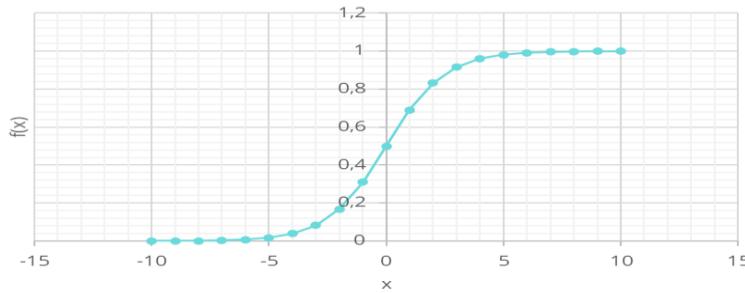
3. Ponto de Corte e Estratégia de Decisão



Probabilidade de ocorrência & decisão a ser tomada

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

27



$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)}}$$

De um lado, temos o nosso modelo de regressão logística, que vai soltar um número para nós, e sabemos que esse número vai variar entre 0 e 1, representando a probabilidade de ocorrência do evento de interesse.

Ok, tudo bem, mas o que fazer com esse número ou probabilidade?



Probabilidade de ocorrência & decisão a ser tomada

O CONTEXTO DE NEGÓCIO | REGRESSÃO LOGÍSTICA

28



Desenhar uma estratégia de decisão baseada nos insights gerados pela probabilidade é o que a empresa espera do cientista de dados. O modelo é apenas um detalhe. **É a estratégia de decisão que dirá o que a empresa deve fazer.**

Créditos: <https://pixabay.com/pt/photos/xequ-mate-xadrez-pedido-de-demiss%C3%A3o-1511866/>

@2020 LABDATA FIA. Copyright all rights reserved.



Relembrando a importância do ponto de corte

DEFINIÇÃO DO PONTO DE CORTE | REGRESSÃO LOGÍSTICA

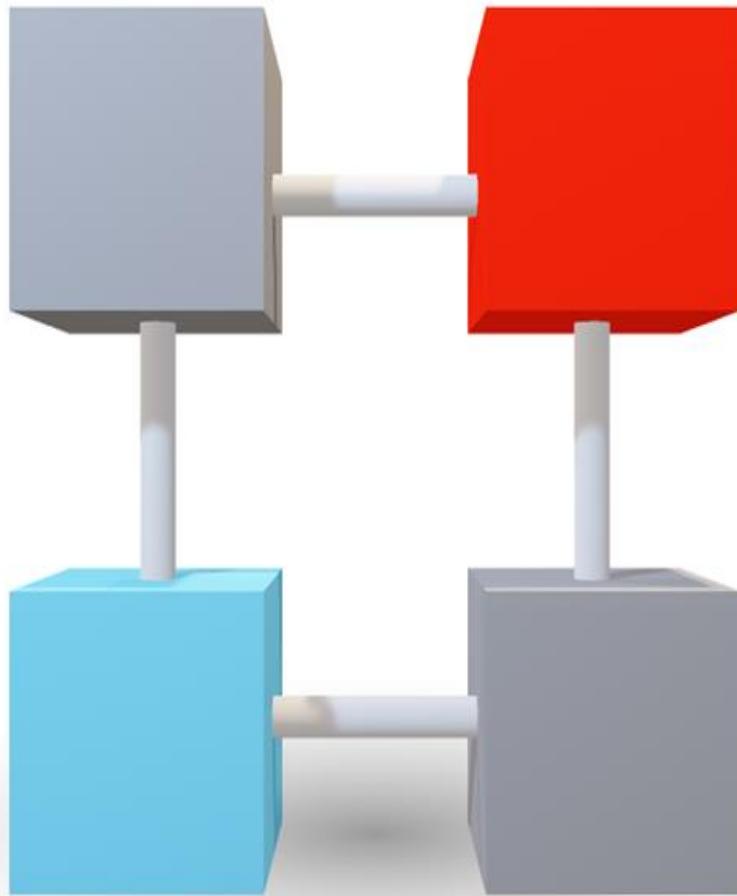
29

PROBABILIDADE

Modelos de regressão logística fornecem a probabilidade de ocorrência de um evento específico. Essa probabilidade varia entre 0 e 1.

QUAL PONTO DE CORTE?

Daí vem o questionamento: Qual ponto de corte pode ser o mais adequado?

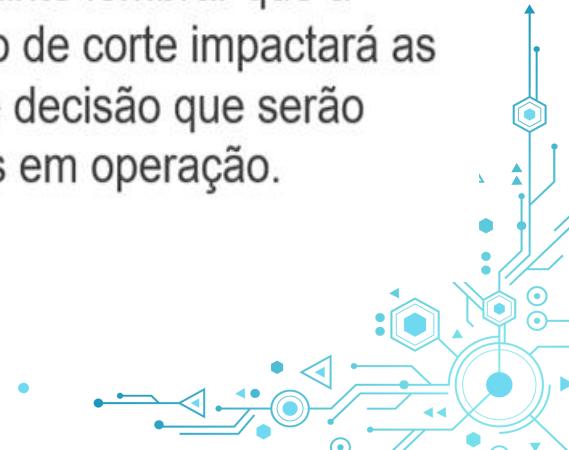


ALÉM DA PROBABILIDADE

A partir da probabilidade, se determina um ponto de corte, inicialmente, igual a 0.5 ou igual à incidência de ocorrência do evento na base.

IMPACTO NA ESTRATÉGIA

Super importante lembrar que a definição de ponto de corte impactará as estratégias de decisão que serão colocadas em operação.



Case: Doença cardíaca

DEFINIÇÃO DO PONTO DE CORTE | REGRESSÃO LOGÍSTICA

30

Na base **[Doenca_cardiaca.xlsx](#)**, estão disponíveis 299 prontuários de pacientes com doenças cardíacas, contendo diversas informações sobre seu estado de saúde, resultados de exames de sangue e a marcação se o paciente faleceu ou não. Essa variável “**óbito**” é a nossa variável de interesse e queremos entender quais dos fatores abaixo melhor explicam o óbito dos pacientes e entender, de forma antecipada, quem tem mais propensão a óbito. Os médicos podem usar essa informação para fazer uma triagem mais qualificada dos pacientes com risco mais alto de morte. As variáveis explicativas disponíveis são as seguintes:

- **idade**: idade do paciente
- **anemia**: 1=sim, 0=não
- **cpk**: nível da enzima cpk no sangue (mcg/L)
- **diabetes**: 1=sim, 0=não
- **pressao alta**: 1=sim, 0=não
- **plaquetas**: contagem de plaquetas no sangue, por mL
- **serum_sodio**: nível desse elemento no sangue (mEq/L)
- **sexo**: 0=mulher, 1=homem
- **fumante**: 1=sim, 0=não
- **obito**: 1=sim, 0=não

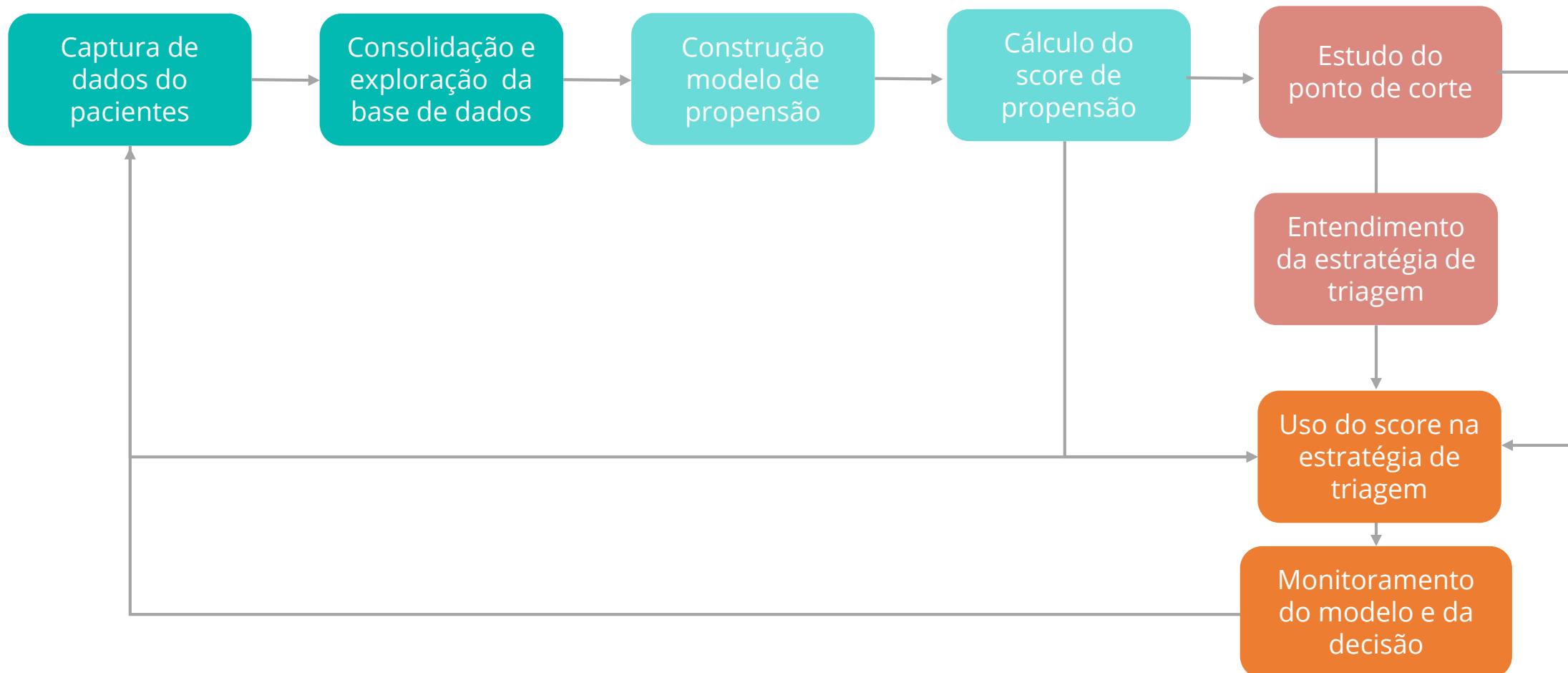


Case: Doença cardíaca

FLUXO SIMPLIFICADO DE DECISÃO | REGRESSÃO LOGÍSTICA

31

De forma bem simplificada, podemos ter um fluxo de decisão que começa com a captura dos dados dos pacientes, construção do modelo, execução do score, decisão de ponto de corte e uso do modelo para estratégias de ação e tomada de decisão.



Case: Doença cardíaca

DISCUSSÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

32

Vamos primeiro avaliar o % de óbitos na base e simular um primeiro modelo de regressão logística, assim como calcular o % de classificação correta.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	13.1315484	4.1556259	3.160	0.00158 **
anemia	0.4126120	0.2632590	1.567	0.11704
cpk	0.0002186	0.0001303	1.677	0.09346 .
pressao_alta	0.4169313	0.2645998	1.576	0.11509
serum_sodio	-0.1052130	0.0306595	-3.432	0.00060 ***

signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		



- O % de pacientes que vieram a óbito foi de 32,1%.
- Uma primeira simulação de ponto de corte poderia considerar o ponto de corte = 0,5 e, na sequência, comparar com o ponto de corte=0,321, avaliando impacto na **acurácia, sensibilidade e especificidade**.
- Em termos de composição do modelo inicial, a maioria das variáveis é não significativa. Apenas 4 variáveis tiveram níveis descritivos inferiores a 0,15, e essas foram mantidas no modelo.



Case: Doença cardíaca

DECISÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

33

Vamos simular os impactos em acurácia, sensibilidade e especificidade, com os 2 cenários de ponto de corte (cenário I, PC=0,5, cenário II < PC=0,32, que é a incidência de óbitos na base).

Ponto de corte = 0.5

Conteúdo das células			
N / Total da linha			
=====			
dc\$obito	dc\$classif	0	1
dc\$obito	dc\$classif	Total	
0	197	6	203
	0.970	0.030	0.679
1	83	13	96
	0.865	0.135	0.321
Total	280	19	299
=====			

Acurácia = $(197+13)/299 = 70,23\%$
Sensibilidade: $13/96 = 13,54\%$
Especificidade: $197/203 = 97,04\%$

Ponto de corte = 0.32

conteúdo das células			
N / Total da linha			
=====			
dc\$obito	dc\$classif	0	1
dc\$obito	dc\$classif	Total	
0	123	80	203
	0.606	0.394	0.679
1	37	59	96
	0.385	0.615	0.321
Total	160	139	299
=====			

Acurácia = $(123+59)/299 = 60,86\%$
Sensibilidade: $59/96 = 61,46\%$
Especificidade: $123/203 = 60,59\%$

Em bases desbalanceadas, fica visível que o ponto de corte=0,5 eleva a acurácia, mas derruba a sensibilidade que, nesse caso, é o indicador que nos interessa mais, pois queremos prever, com antecedência, quem tem mais chance de ir a óbito.

Essa decisão de ponto de corte é um elemento importante não apenas do ponto de vista estatístico, como também do ponto de vista da decisão.



Case: Doença cardíaca

DECISÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

34

Agora, em vez de apenas olharmos as taxas de acerto do modelo, vamos pensar em usar o modelo para decisão.

A equipe médica nos pede uma lista de pacientes propensos a óbito para que eles possam observá-los mais de perto. Supondo que a base histórica de pacientes sirva de parâmetro, **deveríamos escorar uma base nova (só com pacientes vivos)** e vamos separar os “mais propensos”, de acordo com os 2 cenários ($PC=0,5$ e $PC=0,32$).

Como só temos uma base histórica passada, vamos fazer essa comparação, considerando os casos que o modelo teria separado como “propenso” em cada cenário. Vamos ver a diante.



Créditos: <https://pixabay.com/pt/photos/plano-objectivo-estrat%C3%A9gia-objetivo-2372176/>

@2020 LABDATA FIA. Copyright all rights reserved.



Case: Doença cardíaca

DECISÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

35

Cenário I - Ponto de corte = 0,5

Conteúdo das células			
N / Total da linha			
dc\$obito	dc\$classif	0	1
0	197	6	203
	0.970	0.030	0.679
1	83	13	96
	0.865	0.135	0.321
Total	280	19	299

Quantos mandaremos para a equipe? Quem o modelo classificou como “propenso”, aqui, apenas como exemplo, 19 pacientes.

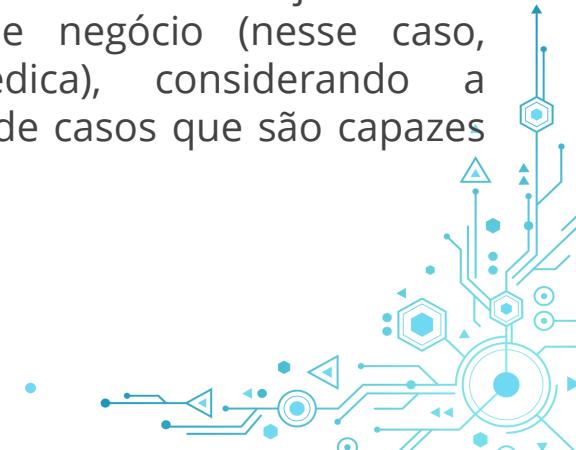
Cenário II - Ponto de corte = 0,32

Conteúdo das células			
N / Total da linha			
dc\$obito	dc\$classif	0	1
0	123	80	203
	0.606	0.394	0.679
1	37	59	96
	0.385	0.615	0.321
Total	160	139	299

Quantos mandaremos para a equipe? Quem o modelo classificou como “propenso”, aqui, apenas como exemplo, 139 pacientes.

No cenário I, teríamos mandado 19 pacientes para análise da equipe. Desses, **68,42% morreram**. Já no cenário II, teríamos mandado 139 pacientes. Desses, **42,44% morreram**.

Qual a conclusão ? Que a análise não pode apenas ser baseada em cálculos de acurácia do modelo. No fim das contas, a definição de ponto de corte é uma decisão não apenas estatística, mas também a ser definida junto com os times de negócio (nesse caso, equipe médica), considerando a quantidade de casos que são capazes de analisar.



Case: Doença cardíaca

DECISÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

36

Nessa análise, dividimos o score de propensão a óbito em 10 faixas distintas e trouxemos o % de óbito em cada faixa.

dc\$faixa_score	dc\$obito	Total
	0	1
[0.109, 0.186)	24 0.800	6 0.200 0.100
[0.186, 0.238)	26 0.867	4 0.133 0.100
[0.238, 0.255)	23 0.767	7 0.233 0.100
[0.255, 0.281)	24 0.800	6 0.200 0.100
[0.281, 0.304)	18 0.621	11 0.379 0.097
[0.304, 0.326)	20 0.667	10 0.333 0.100
[0.326, 0.353)	21 0.700	9 0.300 0.100
[0.353, 0.396)	15 0.500	15 0.500 0.100
[0.396, 0.47)	18 0.600	12 0.400 0.100
[0.47, 0.819]	14 0.467	16 0.533 0.100
Total	203	96 299

A partir das 3 últimas faixas, o % de óbitos já é acima da média histórica de óbitos e, assim, o ponto de corte 0.353 poderia ser uma opção de ponto de corte para a seleção.

Mandaríamos, baseado no padrão dessa base histórica, 90 casos para a equipe médica. Desses, novamente baseado nos padrões da base histórica, 43 faleceram, ou seja, cerca de 47,77%.

Se englobássemos a faixa anterior, com um ponto de corte de 0.326, o número de casos teria sido de 120 e o % de óbitos por volta de 43,33%. Perceba que esse número não conversa com a sensibilidade que, para o PC de 0.32, foi de 61,46%.



Case: Doença cardíaca

DECISÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

37

Nessa análise, dividimos o score de propensão a óbito em 10 faixas distintas e trouxemos o % de óbito em cada faixa.

dc\$faixa_score	dc\$obito		
	0	1	Total
[0.109, 0.186)	24	6	30
	0.800	0.200	0.100
[0.186, 0.238)	26	4	30
	0.867	0.133	0.100
[0.238, 0.255)	23	7	30
	0.767	0.233	0.100
[0.255, 0.281)	24	6	30
	0.800	0.200	0.100
[0.281, 0.304)	18	11	29
	0.621	0.379	0.097
[0.304, 0.326)	20	10	30
	0.667	0.333	0.100
[0.326, 0.353)	21	9	30
	0.700	0.300	0.100
[0.353, 0.396)	15	15	30
	0.500	0.500	0.100
[0.396, 0.47)	18	12	30
	0.600	0.400	0.100
[0.47, 0.819]	14	16	30
	0.467	0.533	0.100
Total	203	96	299

Vamos analisar os 2 indicadores e como eles são diferentes:

- **Sensibilidade:** De todos os pacientes que morreram, quantos o modelo conseguiu identificar? Foram 59 em 96 casos, ou seja, 61,46%. Esse número reflete não o acerto do modelo, mas a penetração do modelo no volume de óbitos.
- **Taxa de acerto da decisão:** Depois de definido o ponto de corte, o modelo selecionou os propensos. Foram 120 casos, dos quais, pela base histórica, 52 (16+12+15+9) morreram, ou seja, 43,33%.

A definição de ponto de corte mais adequada vai depender não apenas do modelo, mas também da quantidade de casos que podem ser absorvidos pela estratégia de decisão.



Exercício: Cancelamento de serviços (churn) em telecom

DECISÃO DE PONTOS DE CORTE | REGRESSÃO LOGÍSTICA

38

Na base **PONTO CORTE CANCELAMENTO SERVICOS.xlsx**, estão disponíveis dados de 10mil clientes sobre seu perfil e a informação se cancelaram ou não seu relacionamento com a empresa de telefonia. A idéia é montar uma estratégia ativa de retenção que possa ser efetiva e também atenda às necessidades/limitações do *call center* da empresa, já que o tamanho da operação é limitado. Os dados disponíveis para o desenho de *score* de *churn* são os seguintes:

score serasa: score do cliente no birô de crédito

sexo: Sexo do cliente

idade: idade do cliente, em anos

tempo_relacionamento: tempo de relacionamento, em anos

possui internet: 1=sim, 0=não

salario_anual: salário total do cliente ao longo do ano

cancelou: 1=sim, 0=não



4. Validação e monitoramento de modelos



Treino é treino, jogo é jogo!

O CONTEXTO DE NEGÓCIO | TÓPICOS DE MODELAGEM

40



Quando separamos uma amostra para treinamento, lançamos mão de muitos recursos para identificar padrões nos dados, seja criando novas variáveis, combinando variáveis existentes, seja utilizando diferentes métodos de análise.

Com qual objetivo? *Encontrar um modelo que consiga prever a realidade da forma mais precisa e estável possível.*

Créditos: <https://pixabay.com/pt/photos/homem-corrida-homem-correndo-1245658/>

@2020 LABDATA FIA. Copyright all rights reserved.



Treino é treino, jogo é jogo!

O CONTEXTO DE NEGÓCIO | TÓPICOS DE MODELAGEM

41



Contudo, a realidade não está só no laboratório, onde treinamos o modelo. Ele está na vida real, no momento em que a empresa decide colocar o modelo em operação. *É aí que ele será realmente testado.*

Créditos: <https://pixabay.com/pt/photos/corrida-pista-e-campo-execu%C3%A7%C3%A3o-801940/>

@2020 LABDATA FIA. Copyright all rights reserved.

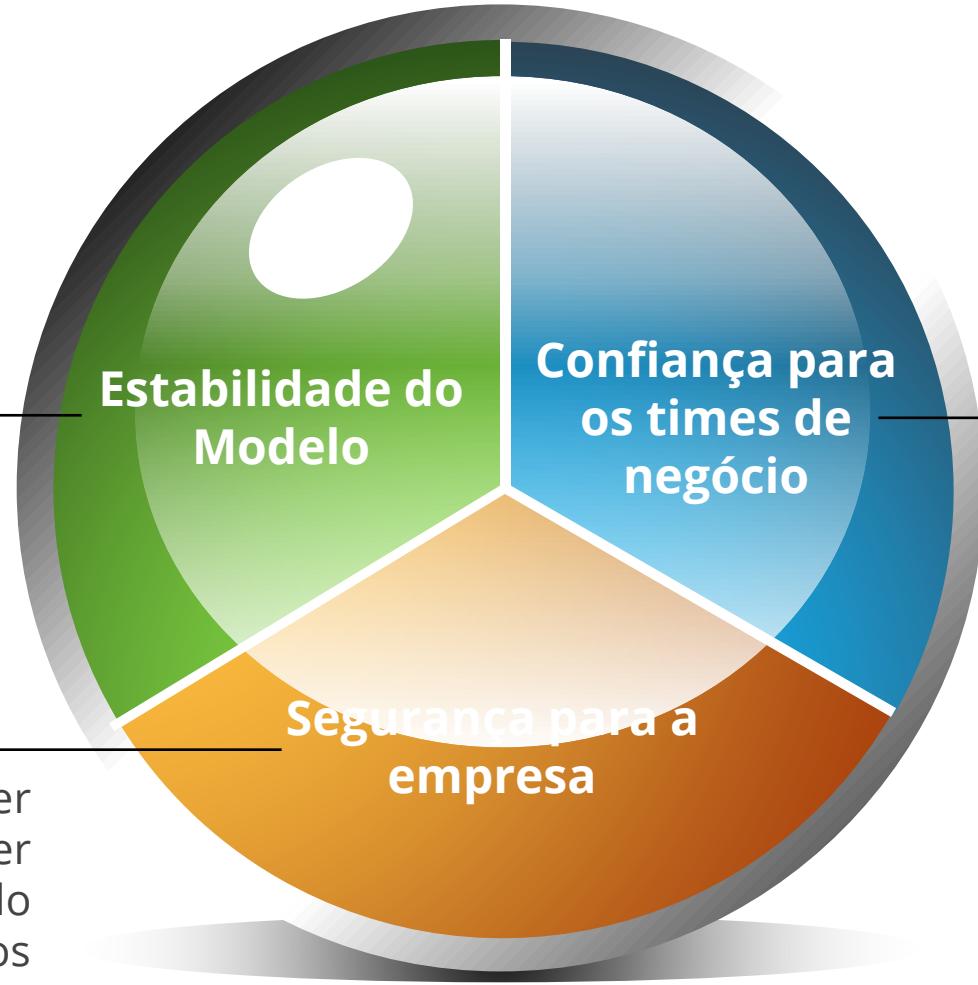


Utilidade das amostras de treino e teste

O CONTEXTO DE NEGÓCIO | TÓPICOS DE MODELAGEM

42

Assegurar que os resultados obtidos no treinamento do modelo serão também reproduzidos em uma amostra separada.



Consequentemente, trazer segurança que a estratégia a ser utilizada com base no modelo pode trazer resultados positivos para a empresa.

Transmitir confiança para os clientes internos e externos (via de regra, pessoas de negócios) de que os resultados da modelagem são confiáveis.



Validação dos modelos: O segredo do sucesso

O CONTEXTO DE NEGÓCIO | TÓPICOS DE MODELAGEM

43

 **Model Validation - Model Risk Management**

Credit Suisse ★★★★☆ 1,547 reviews - Raleigh, NC

Apply On Company Site 

this depends on the skills, experience and engagement of our employees. We offer a collaborative and ambitious environment that offers direct contact with senior management and encourages leadership at all levels.

The Model Risk Management (MRM) team has a mandate to validate the Bank's business-impactful models firm-wide and more generally to identify, measure, and handle model risk across Credit Suisse. The team is established in London, Zurich, Mumbai, Singapore, New York, Warsaw and now Raleigh.

As an entry level member of the MRM validation team you will get exposure to modeling in a wide variety of risk areas such as credit risk, market risk, operational risk etc. The current heightened regulatory focus on these areas and the team's broader model risk scope also guarantees a significant level of interest and visibility to the business and senior management.

Role Description:

- You will review, verify and validate risk models for theoretical soundness.
- You will test model design and identification of model weaknesses, ensuring ongoing monitoring as well as contribute in the firm-wide model risk and control assessment.
- You will be expected to demonstrate independence in testing.

Credit Suisse maintains a Working Flexibility Policy, subject to the terms as set forth in the Credit Suisse United States Employment Handbook.

You Offer

Fonte: <https://www.indeed.com/q-Quantitative-Model-Validation-jobs.html?vjk=c0debcca1c9939d3>

Validação de modelos é elemento chave para a estratégia das empresas, já que se o modelo não tiver um bom desempenho em produção, poderá seriamente impactar as estratégias de decisão das empresas.

Sobretudo em bancos, existem áreas fortes e bem consolidadas para validação de modelos de risco e outros modelos usados em políticas e processos de decisão.



Validação dos modelos: Algumas possibilidades

BASES DE TREINO, VALIDAÇÃO E TESTE | TÓPICOS DE MODELAGEM

44

Método I

Holdout

Separar 2-3 amostras aleatórias, uma para treino e outras para validação e teste.



É o mais comum, mais simples e mais limitado. Porém, por conta da facilidade, é amplamente usado.

Método II

Out of time

As amostras de validação/teste, além de serem aleatórias, são de períodos mais recentes.



É mais robusto, pois considera uma janela temporal distinta (e, em geral, mais atualizada). Contudo, nem sempre, esses dados estão disponíveis.

Método III

K-fold cross validation

Gera inúmeras simulações de distintas combinações de amostras para testar o modelo.



É também bastante robusto, demanda mais recursos computacionais e é muito utilizado após desenvolvimento de modelos “caixa preta”.



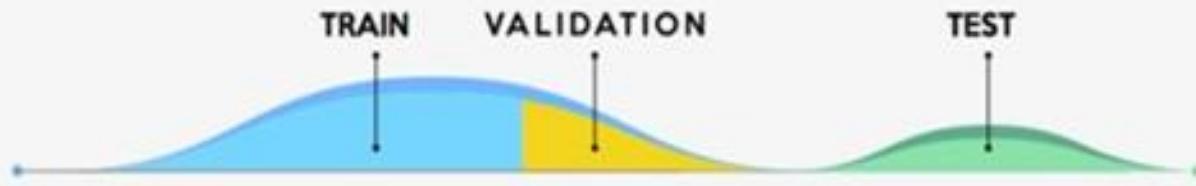
Validação dos modelos: Método *Holdout*

BASES DE TREINO, VALIDAÇÃO E TESTE | TÓPICOS DE MODELAGEM

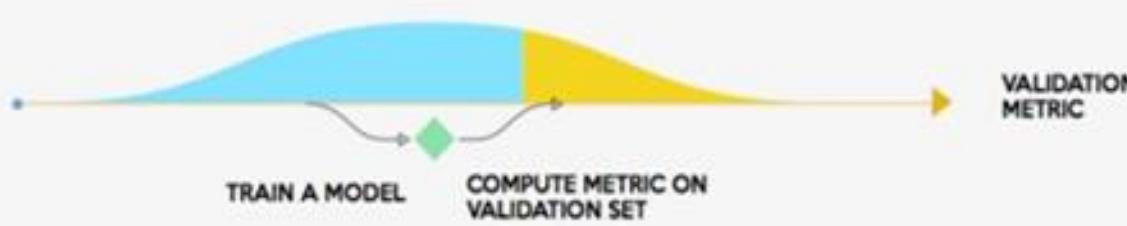
45

HOLDOUT STRATEGY

1 Split your data into train / validation / test

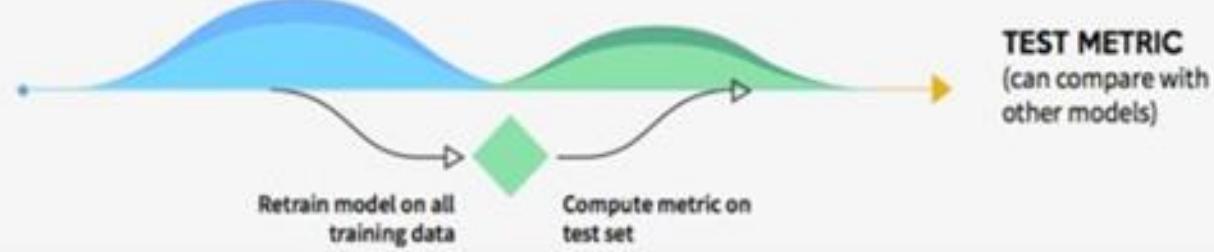


2 For each parameter combination



Parameter A (e.g., depth)
3 11 15 6 16 2 17
Parameter B (e.g., n trees)
1 11 15 6 16 2 17

3 Choose the parameter combination with the best metric



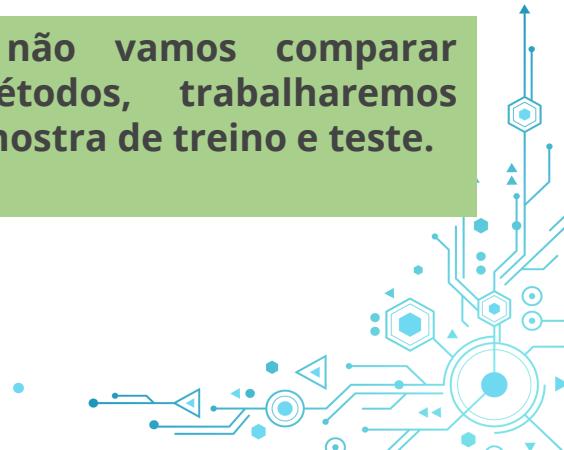
Fonte: <https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html>

Em geral, a partir da base inicial de análise, se definem percentuais como 70-30 (70% para treino e 30% para validação e teste) ou 80-20 (80% para treino e 20% para validação e teste), a depender da quantidade de dados disponíveis (sobretudo do evento de interesse).

Amostra de treinamento: onde o modelo é construído e os padrões são aprendidos.

Amostras de validação/teste: onde o modelo é escolhido (validação) é testado (teste).

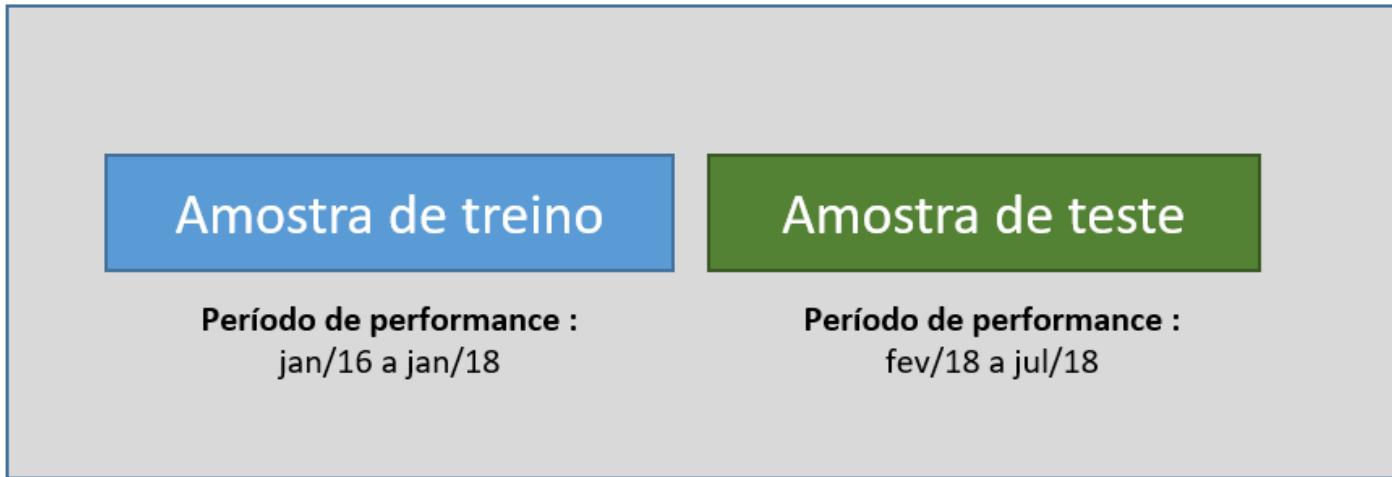
Como ainda não vamos comparar diferentes métodos, trabalharemos apenas com amostra de treino e teste.



Validação dos modelos: Método *Out of Time*

BASES DE TREINO, VALIDAÇÃO E TESTE | TÓPICOS DE MODELAGEM

46



Créditos: Professor Marcelo Fernandes, FIA

O método “**out of time**” é muito similar ao do “**holdout**”, com uma diferença crucial, que pode ser decisiva, a questão temporal:

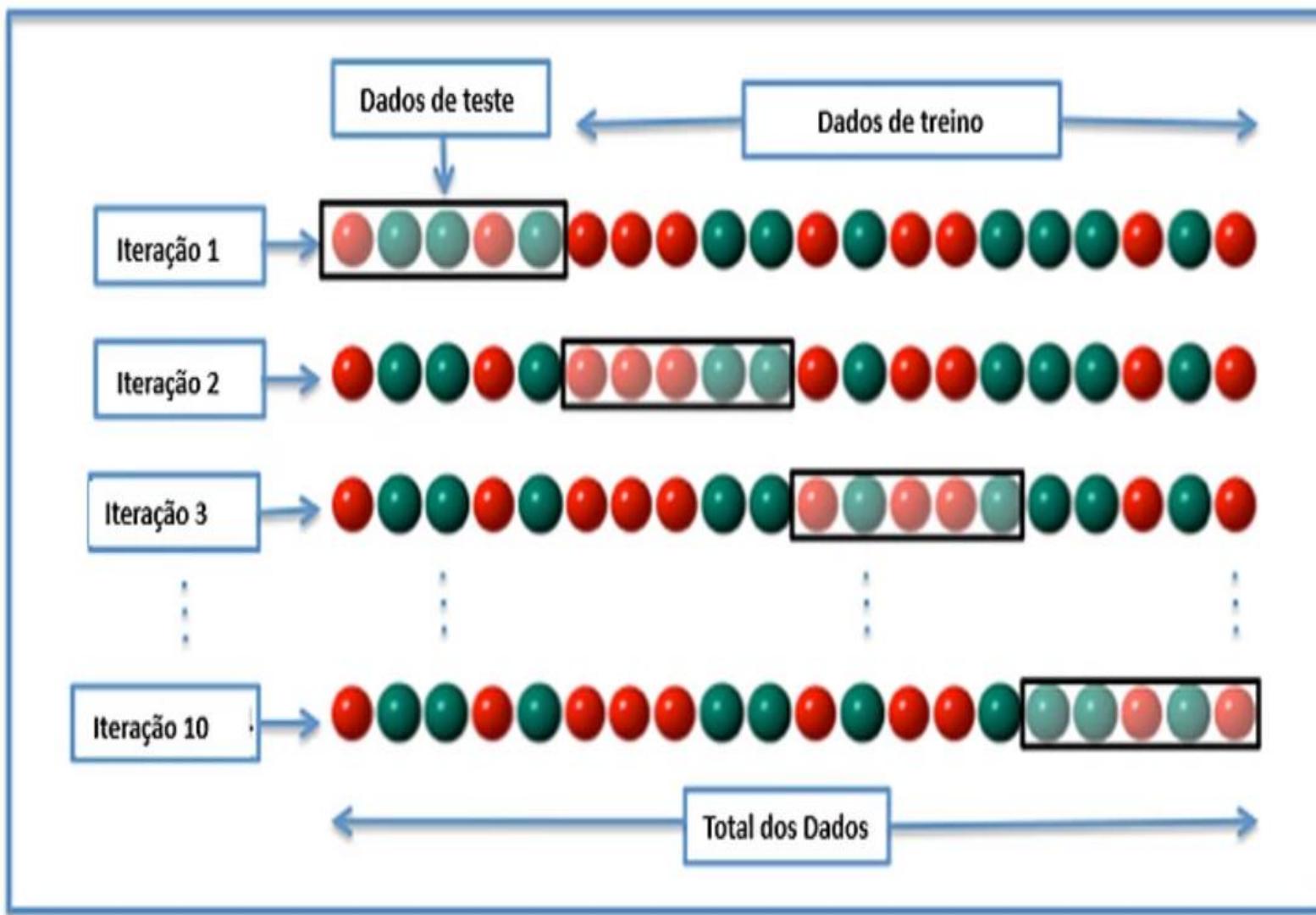
- A amostra de teste, além de não ter sido usada no treino/desenvolvimento do modelo, é de um período diferente, em geral, mais recente, de forma que é uma base com características mais próximas ao contexto real de implementação do modelo.
- **Exemplo:** Estamos em out/18. Imagine um modelo de risco sendo desenvolvido a partir de uma amostra de um período entre jan/16 e jan/18. A amostra de teste pode ser composta por safras de fev/18 a jul/18, que potencialmente reflete melhor a situação atual do que a base de treino.



Validação dos modelos: Método *K-Fold Cross Validation*

BASES DE TREINO, VALIDAÇÃO E TESTE | TÓPICOS DE MODELAGEM

47

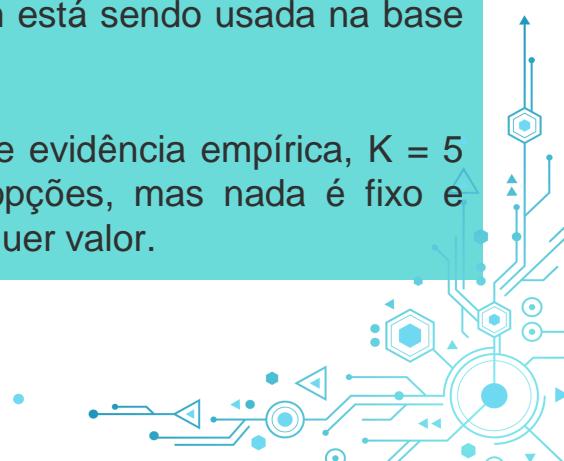


Fonte: <https://www.3dimensoes.com.br/post/valida%C3%A7%C3%A3o-de-modelos-em-machine-learning>

Neste método os dados são divididos em k subconjuntos (amostras). Em seguida, o método *holdout* é repetido k vezes, de tal forma que, a cada vez, um dos k subconjuntos é usado como base de teste e os outros subconjuntos $k-1$ são colocados juntos para formar uma base de treino. A estimativa de erro é calculada com base em todas as k tentativas para obter a eficácia total do nosso modelo.

Cada observação integra os dados de teste exatamente uma vez e integra a base de treino $k-1$ vezes. Isso reduz significativamente o *bias* (viés), já que estamos usando a maioria dos dados para treinamento e também reduz significativamente a variância, pois a maioria dos dados também está sendo usada na base de teste.

Como regra geral e evidência empírica, $K = 5$ ou 10 são boas opções, mas nada é fixo e pode-se usar qualquer valor.



Monitoramento de modelos: **PSI – Population Stability Index**

BASES DE TREINO, VALIDAÇÃO E TESTE | TÓPICOS DE MODELAGEM

48

Esse indicador, também chamado de PSI (ou em português, IEP – Índice de Estabilidade da População), é um forte aliado do processo de acompanhamento do modelo depois de sua implementação, de forma a avaliar se os resultados obtidos em produção refletem os padrões esperados, obtidos no momento da construção do modelo na amostra de treino. Sua composição é bastante simples:

$$PSI = \sum_{i=1}^n (\% Real_i - \% Esperado_i) * \ln\left(\frac{\% Real_i}{\% Esperado_i}\right)$$

PSI	Interpretação
<0,1	Nenhuma mudança significativa
Entre 0,1 e 0,25	Leve mudança
>0,25	Mudanças significativas

Fonte: <https://support.sas.com/resources/papers/proceedings10/288-2010.pdf>

Identificado algum desvio relevante, o caminho é avaliar em que variáveis do modelo essa mudança foi mais acentuada, de modo a direcionar focos de atuação para a atualização da fórmula do modelo.



Exercício sobre monitoramento de modelos

MONITORAMENTO DE MODELOS | TÓPICOS DE MODELAGEM

49

Vamos considerar a planilha “MONITORAMENTO MODELOS DISTRIBUICAO SCORE.xlsx”, que traz a distribuição esperada dos scores (que foi obtida no momento do treinamento) e a distribuição observada, que foi obtida depois da implementação. No Excel, vamos calcular o PSI, considerando esses dados:

Faixa de Score	Distribuição de Score (Qtde)					
	Distribuição Esperada			Distribuição Observada		
	Bons	Maus	Total	Bons	Maus	Total
<=100	25.412	52.515	77.927	29.425	59.522	88.947
101-200	28.844	41.251	70.095	23.412	51.225	74.637
201-300	29.855	32.412	62.267	21.254	50.112	71.366
301-400	35.451	19.525	54.976	29.825	26.522	56.347
401-500	38.455	14.221	52.676	32.412	13.541	45.953
501-600	48.541	9.522	58.063	38.665	11.254	49.919
601-700	52.514	6.855	59.369	48.954	9.225	58.179
701-800	59.522	6.224	65.746	61.254	9.211	70.465
801-900	75.425	3.541	78.966	65.224	8.554	73.778
>900	85.641	2.111	87.752	82.144	3.541	85.685
Total	479.660	188.177	667.837	432.569	242.707	675.276

A partir dessa tabela, calcule o PSI (ou IEP) e teça seus comentários sobre o nível de estabilidade das 2 distribuições (esperada, do modelo, vs observada, dos dados em produção).



Case – Predição de categorias de valor de veículos usados

50

Na base [Estimativa_Carros_Usados.xlsx](#), Estão disponíveis registros de 1711 anúncios de veículos extraídos de um revendedor. O objetivo é predizer a categoria de valor do veículo, **grupo I (veículos de até \$20.000), grupo II (veículos acima de \$20.000)**, a partir de suas características, bem como, antes de desenvolver o modelo, separar a base em treino e teste, deixando 80% para treino e 20% para teste.



Transmissao: tipo de transmissão do veículo

Cilindradas: quantidade de cilindradas

Litros: litros do motor

Kms: quilômetros rodados

Economia: quilômetros por litro

Tipo: tipo do veículo

Ano: ano do veículo

Combustivel: tipo de combustível

ID: ID do vendedor



Case – Predição de categorias de valor de veículos usados

51

Na base, não temos a variável “categoria de valor” do veículo (grupos I e II). Contudo, a partir da variável preço, podemos criar uma nova variável, chamada cat_valor, que será 1 caso o preço do carro seja > \$20.000 e 0, caso contrário:

```
#Criação da variável "categoria de valor" do veículo  
carros$cat_valor<-ifelse(carros$preco>20000,1,0)  
View(carros)
```

	transmissao	cilindradas	litros	kms	consumo	tipo	ano	preco	combustivel	SellerId	cat_valor
1	Manual		4	2.0	67169	6.0 2.0D	2011	26888	Diesel	AGC-SELLER-16890	1
2	Automatic		4	2.5	11000	8.1 2.5i-L	2015	30999	Petrol - Unleaded ULP	SSE-SELLER-2347077	1
3	Automatic		4	2.5	0	8.1 2.5i-L	2015	36844	Petrol - Unleaded ULP	AGC-SELLER-51471	1
4	Automatic		4	2.5	81000	9.6 XS Premium	2009	17100	Petrol - Unleaded ULP	SSE-SELLER-3267502	0
5	Automatic		4	2.0	9398	8.5 XT Premium	2015	44888	Petrol - Premium ULP	AGC-SELLER-10988	1
6	Manual		4	2.0	0	5.9 2.0D-L	2016	37184	Diesel	AGC-SELLER-18615	1
7	Automatic		4	2.5	1502	8.1 2.5i-S	2015	39990	Petrol - Unleaded ULP	AGC-SELLER-15108	1
8	Automatic		4	2.0	12	8.5 XT Premium	2015	46785	Petrol - Premium ULP	AGC-SELLER-29334	1
9	Automatic		4	2.5	91059	9.6 X	2008	14988	Petrol - Unleaded ULP	AGC-SELLER-10934	0
10	Automatic		4	2.5	11018	8.1 2.5i-L	2015	34990	Petrol - Unleaded ULP	AGC-SELLER-12886	1
11	Automatic		4	2.5	81966	9.7 XS	2008	18880	Petrol - Unleaded ULP	AGC-SELLER-2846	0
12	Automatic		4	2.5	108677	8.1 2.5i	2013	25990	Petrol - Unleaded ULP	AGC-SELLER-16437	1
13	Automatic		4	2.5	0	8.1 2.5i-S	2016	43490	Petrol - Unleaded ULP	AGC-SELLER-9452	1
14	Automatic		4	2.5	2166	8.1 2.5i-S	2015	40888	Petrol - Unleaded ULP	AGC-SELLER-17810	1
15	Manual		4	2.0	1582	5.9 2.0D-S	2015	39888	Diesel	AGC-SELLER-12320	1
16	Automatic		4	2.0	0	8.5 XT Premium	2015	53615	Petrol - Premium ULP	AGC-SELLER-17810	1
17	Automatic		4	2.5	5	8.1 2.5i-S	2016	43005	Petrol - Unleaded ULP	AGC-SELLER-12320	1

Essa variável “cat_valor” é nossa variável resposta e o modelo de regressão logística calculará a probabilidade do veículo pertencer ao grupo II, de valores superiores a \$20.000.



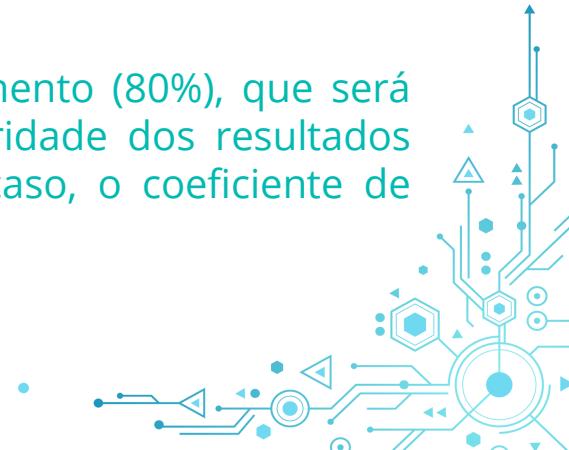
Case – Predição de categorias de valor de veículos usados

52

Vamos separar a base nas 2 amostras: treino e validação. Para isso, vamos considerar o método mais simples de validação de modelos, que é o *holdout*.

```
#Separar a amostra em treino e teste
set.seed(123) # para travar a amostra
sample.size <- floor(0.80 * nrow(carros))
train.index <- sample(seq_len(nrow(carros)), size = sample.size)
train <- carros[train.index, ]
View(train)
dim(train)
test <- carros[- train.index, ]
dim(test)
```

A partir da base inicial, foram geradas 2 amostras, aleatoriamente selecionadas, uma para treinamento (80%), que será usada para construir um modelo e outra para teste (20%), que será usada para avaliar a similaridade dos resultados encontrados na amostra de treinamento, sobretudo em termos de métricas de acurácia, neste caso, o coeficiente de determinação, ou R2.



Case – Predição de categorias de valor de veículos usados

53

Apenas a título de exemplo, foi gerado um modelo com apenas 3 variáveis (kms, consumo e litros).

```
#Montar um modelo de regressao na base de treino
modelo_carros<-glm(cat_valor ~ kms + consumo + litros, family = binomial(link = 'logit'), data=train)
summary(modelo_carros)
#Salvar a predicao na base de treino
train<-data.frame( train, pred=predict(modelo_carros, train, type="response"))
View(train)
#Calcular o KS para a amostra de treino
library(dgof)
ks.test(train$pred[train$cat_valor==0],
       train$pred[train$cat_valor==1])

call:
glm(formula = cat_valor ~ kms + consumo + litros, family = binomial(link = "logit"),
     data = train)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-2.84561 -0.22365  0.03892  0.13645  2.95814 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 1.322e+01 1.892e+00  6.984 2.87e-12 ***
kms        -5.562e-05 3.851e-06 -14.444 < 2e-16 ***
consumo    -5.601e-01 1.523e-01  -3.678 0.000235 ***
litros      -1.363e+00 1.051e+00  -1.298 0.194438  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Two-sample Kolmogorov-Smirnov test

data: train$pred[train$cat_valor == 0] and train$pred[train$cat_valor == 1]
D = 0.82621, p-value < 2.2e-16
alternative hypothesis: two-sided
```

O valor do K-S do modelo para a base de treino foi de 82,62%, um excelente resultado, lembrando que esse indicador pode variar entre 0 e 100% e que, quanto mais próximo de 100%, melhor o modelo conseguir separar as 2 distribuições de carros categoria I e categoria II.

Vamos agora avaliar se esse resultado também se confirmou na amostra de teste.



Case – Predição de categorias de valor de veículos usados

54

Vamos avaliar o resultado do modelo na amostra de teste.

```
#Gerar e salvar predicoes na base de teste
test<-data.frame( test, pred=predict(modelo_carros, test, type="response"))
View(test)
#Calcular o KS para a amostra de teste
library(dgof)
ks.test(test$pred[test$cat_valor==0],
       test$pred[test$cat_valor==1])|
```

Two-sample Kolmogorov-Smirnov test

```
data: test$pred[test$cat_valor == 0] and test$pred[test$cat_valor == 1]
D = 0.81732, p-value < 2.2e-16
alternative hypothesis: two-sided
```

O K-S da amostra de teste foi de 81,73%, bastante próximo ao K-S observado na amostra de treino, que foi de 82,62%, o que mostra que o modelo gerado teve boa aderência em ambas as amostras.



Exercício – Predição da faixa de valor do salário de profissionais

55

Na base **"Fatores Impacto Salario.xlsx"**, estão disponíveis valores de salário de mais de 500 profissionais, além de dados sócio-demográficos do profissional. O propósito é estimar a probabilidade do profissional ganhar mais de 10 salários mínimos ao mês, em função das seguintes variáveis explicativas:

Idade : Idade do profissional

Educacao: Anos de estudo

Sexo: Sexo do profissional (0=Masculino, 1=Feminino)

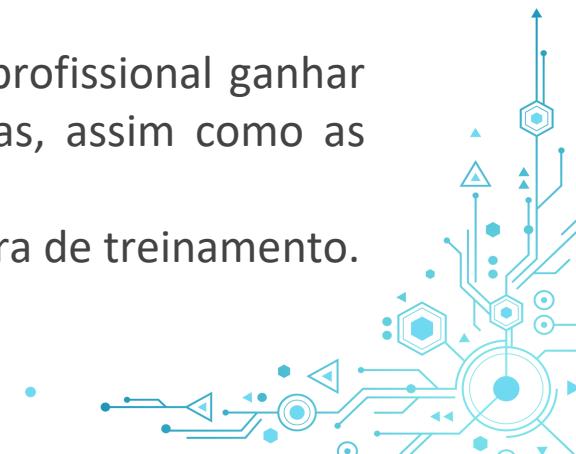
Sul: Se o profissional mora no Sul ou não

Anos_experiencia: Tempo de experiência do profissional

Salario: Quantidade de salários mínimos do profissional

Flag_casado: Se o profissional é casado ou não

- a) Divida a amostra em 2 partes: treinamento e teste.
- b) A partir da base de treinamento, desenvolva um modelo para estimar a probabilidade do profissional ganhar mais de 10 salários mínimos ao mês, em função de suas características sócio-demográficas, assim como as métricas de desempenho de K-S.
- c) Use a base de teste para também gerar o K-S e comparar com os resultados obtidos na amostra de treinamento.



5. O fenômeno chamado “*overfitting*”



Não confiar demais no pote de ouro no final do arco-íris!!

O CONTEXTO DE NEGÓCIO | OVERFITTING

57



Sonho de consumo do
cientista de dados

Gerar modelos altamente preditivos e grande capacidade de prever os eventos que estão para ocorrer. Esse é o verdadeiro “EL Dorado” ou desejo/sonho de consumo dos profissionais que trabalham com dados.

Fonte: <https://pixabay.com/pt/photos/seta-alvo-bullseye-objetivo-2886223/>



Não confiar demais no pote de ouro no final do arco-íris!!

O CONTEXTO DE NEGÓCIO | OVERFITTING

58



No entanto, é preciso tomar cuidados para que esse “super modelo” a ser criado possa ser generalizável e não reflita resultados apenas na base que está sendo usada para estudo.

Fonte: <https://pixabay.com/pt/photos/mouse-isca-perigo-risco-vermes-164751/>

@2020 LABDATA FIA. Copyright all rights reserved.



Não confiar demais no pote de ouro no final do arco-íris!!

O CONTEXTO DE NEGÓCIO | OVERFITTING

59

Forbes

Jan 9, 2018, 02:55pm EST

Management AI: Overfit, Why Machine Learning Isn't Trained to Perfection

F David A. Teich Senior Contributor
Tirias Research Contributor Group 
AI
B2B technology analyst, marketer, and consultant

This article is more than 2 years old.

The core of most modern Machine Learning (ML) systems is automated neural networks (ANNs). The training of ANN's requires large data sets. One misconception of those data sets is the idea that "if we get enough data, we can make the system 100% accurate." Yes, that can happen, but it's not what we really want.

Many methods can be used to group data into relevant categories. The analytics can be made more and more precise by the addition of variables that allow us to recognize items from each dataset. Let's see how that can work and how it can cause problems. Notice the figure below has two different lines to represent different algorithms for classifying data points.

The squiggly green line accurately groups every data point in the training set into the correct categories. The black line is a "best fit" algorithm with a smooth curve. What happens when we test the model on new data?

Fit versus Overfit CHABACANO
[HTTPS://COMMONS.WIKIMEDIA.ORG/W/INDEX.PHP?CURID=3610704](https://commons.wikimedia.org/w/index.php?curid=3610704)



Fonte: <https://www.forbes.com/sites/tiriasresearch/2018/01/09/management-ai-overfit-why-machine-learning-isnt-trained-to-perfection/?sh=32bad4fd4bc2>

O processo de desenvolvimento de modelos pressupõe a busca por estimativas e previsões que sejam, ao mesmo tempo, precisas e confiáveis.

A confiabilidade não está somente associada à precisão das previsões, mas o quanto elas podem continuar sendo precisas e relevantes em outras amostras ou bases, que não somente a base de treino.

Essa capacidade de generalização dos resultados do modelo é o que o torna, de fato, útil para processos de decisão.

Contudo, em alguns casos, geramos um modelo fantástico na amostra de treino, mas quando o validamos na amostra de teste, os resultados são decepcionantes.

O que está em jogo aí é o fenômeno chamado de “overfitting”.



Do que trata esse tema?

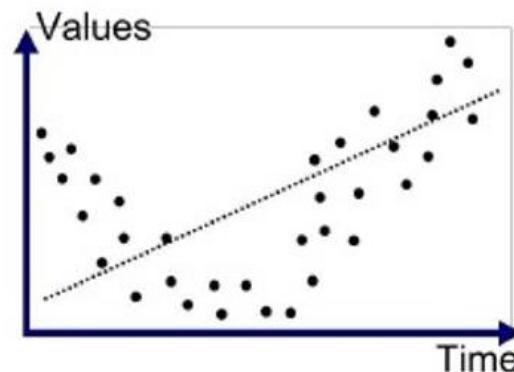
OVERFITTING | TÓPICOS DE MODELAGEM

60

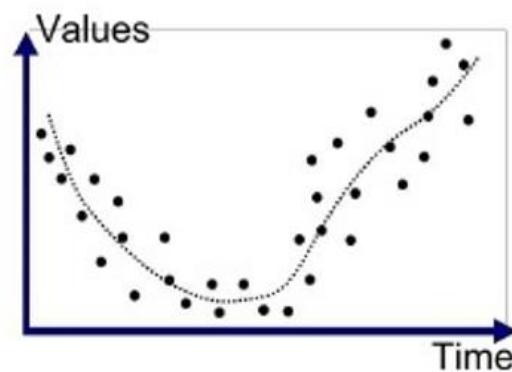
Muitas vezes, desenvolvemos modelos em que os resultados são muito bons, acima da média, altamente promissores, e isso nos deixa muito feliz, com a sensação que “descobrimos a roda”.

No entanto, é preciso ter um alto nível de consciência crítica para entender que, o que vale, não são apenas os resultados que encontramos em nossa base de estudos, mas, principalmente, o que obtemos quando o modelo é colocado em produção, em ambiente real de decisão.

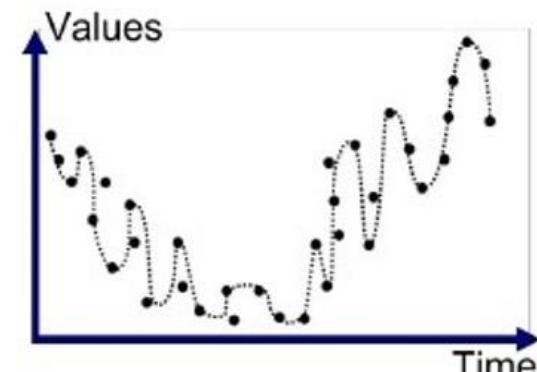
Portanto, vamos dar uns passos para trás e começar pelo básico. Veja o quadro a seguir:



Underfitted



Good Fit/R robust



Overfitted

Fonte: <https://medium.com/@cs.sabaribalaji/overfitting-6c1cd9af589>



Do que trata esse tema?

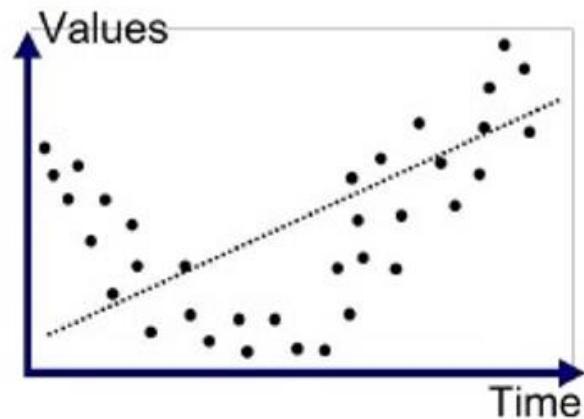
OVERFITTING | TÓPICOS DE MODELAGEM

61

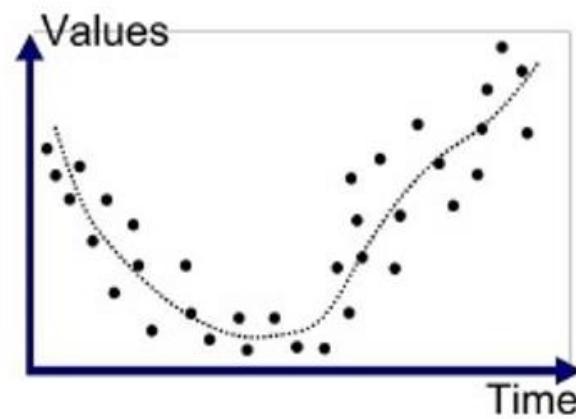
Underfitted: Nosso ajuste é muito pobre e nem ficamos felizes. Vamos tentar incluir outras variáveis e/ou outros métodos de análise para chegar a outros resultados.

Good fit/robust: Nosso ajuste razoável/bom e parece haver perspectiva, apesar de ainda não ser o mundo perfeito.

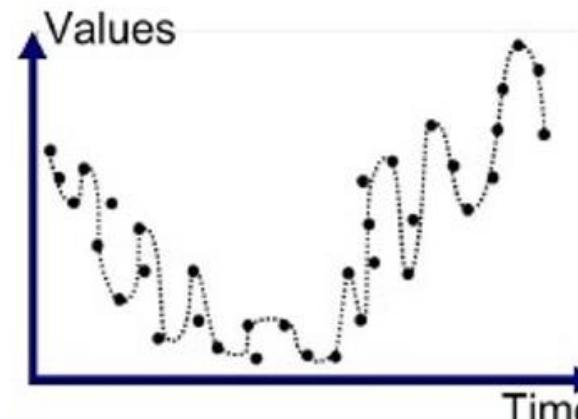
Overfitted: Nosso ajuste está sensacional e ficamos com a sensação boa de que o modelo vai ser um sucesso. É nesse momento que as precauções devem ser redobradas, de modo a assegurar que os resultados obtidos realmente são generalizáveis, ou seja, também ocorrerão em outras amostras.



Underfitted



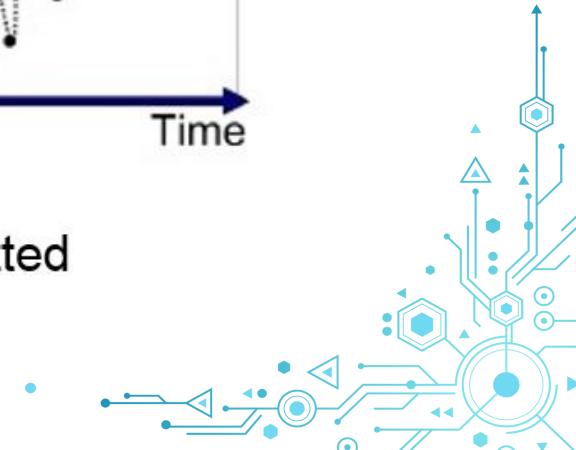
Good Fit/Robust



Overfitted

Fonte: <https://medium.com/@cs.sabaribalaji/overfitting-6c1cd9af589>

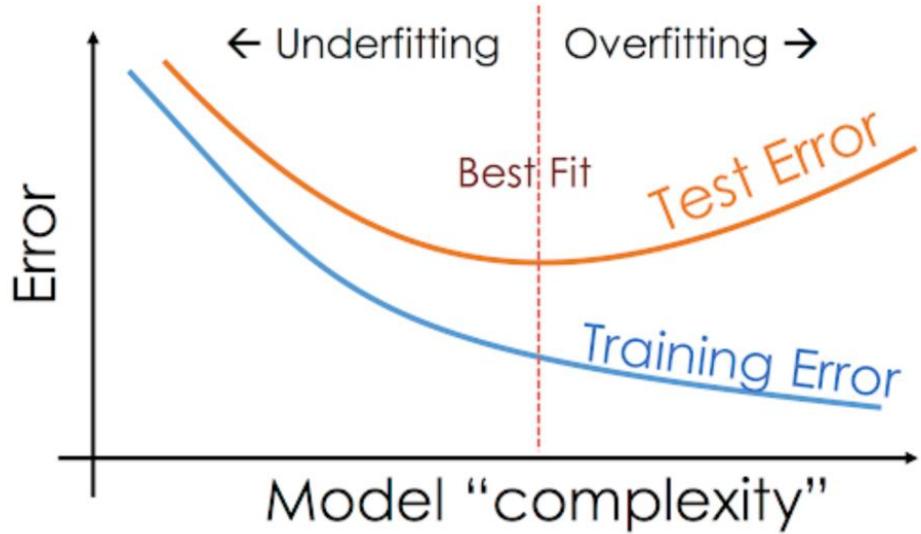
@2020 LABDATA FIA. Copyright all rights reserved.



Do que trata esse tema?

OVERFITTING | TÓPICOS DE MODELAGEM

62



Percebiam que existe uma certa relação entre “overfitting” e complexidade do modelo. Esse fenômeno é mais comum em modelos mais complexos ou com muitas variáveis.

A principal característica observada em processos envolvendo “overfitting” é a ocorrência de uma redução expressiva do erro de predição na base de treinamento e um aumento expressivo desse erro na base de teste, fazendo com que os resultados das 2 amostras (treino e teste) não se conversem.

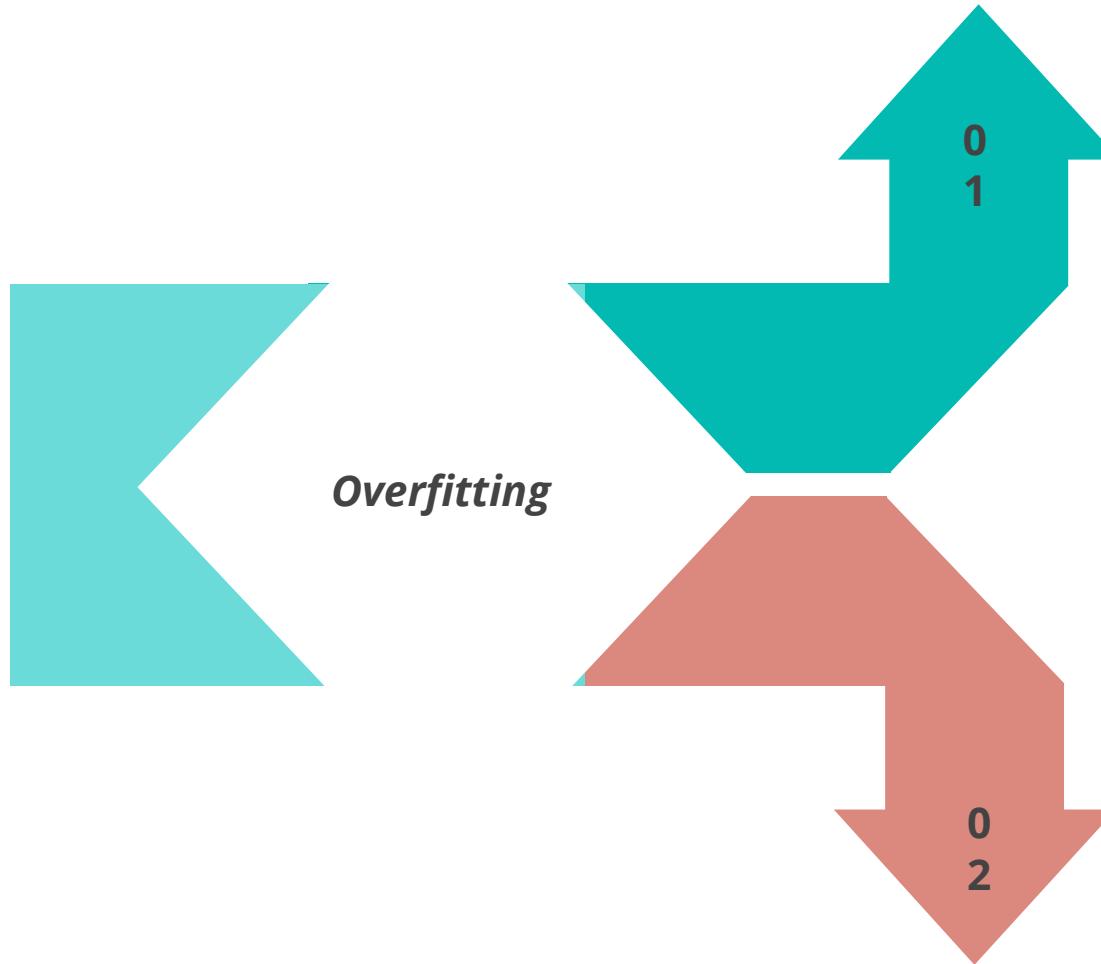
Se temos muitas variáveis (modelo complexo), as hipóteses aprendidas pelo modelo podem ajustar-se bem à base/amostra de treinamento, mas falhar em generalizar esse aprendizado para novos casos (por exemplo, predizer o risco de inadimplência para novos clientes).



Exemplo: Identificando *overfitting* nos dados

OVERFITTING | TÓPICOS DE MODELAGEM

63



Amostra de Treinamento

Resultado muito bom, altíssimo nível de acurácia do modelo.

Amostra de Teste

Ajuste muito ruim, indicadores de performance muito divergentes dos observados na amostra de treinamento.



Como minimizar a ocorrência de *overfitting* nos dados?

OVERFITTING | TÓPICOS DE MODELAGEM

64

Minimizando Ocorrência de Overfitting

01

Validação cruzada

Poderosa medida para evitar o overfitting, usando os dados de treinamento para gerar múltiplos splits de amostras de treino e teste, usando esses splits para tunar o modelo.

02

Modelar com mais dados

Muitas vezes, modelar com mais dados ajuda o modelo a detectar melhor os sinais.

03

Remover variáveis

Remover variáveis redundantes ou não relevantes ajuda a melhorar a generalização do modelo.

04

Regularização/ Ensembles

- Regularização introduz uma penalidade nos parâmetros, que ajuda a reduzir o overfitting.
- Ensembles como bagging ajudam a reduzir overfitting, por conta das combinações das previsões.



Case – Overfitting – Churn de Seguro de Automóvel

65

Na base **"Churn Seguro Auto.xlsx"**, pastas “churn_train” e “churn_test”, estão disponíveis dados de segurados do produto “seguro de automóvel”, com informações sócio-demográficas, assim como a marcação se o cliente cancelou ou não seu relacionamento com a seguradora. O objetivo é gerar um modelo na base de treino que prediga a probabilidade de o cliente cancelar seu seguro, assim como testar se as previsões do modelo continuam relevantes para a base de teste.

Renda : Faixa de renda do cliente

Reclamacoes: Faixa de quantidade de reclamações na central de atendimento

Educacao: Nível educacional do segurado

Tempo_cliente: Faixa de tempo como cliente da seguradora

Classe_idade: Faixa de idade do segurado

Idade_carro: Faixa de idade do veículo

Debito_autom: Flag de débito automático da parcela do seguro

Churn: 0=não cancelou, 1=cancelou

- Qual a qualidade de predição do modelo de churn para a base de treino?
- Algum sinal de *overfitting* nesses dados?



Case – Overfitting – Churn de Seguro Automóvel

66

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-9.6086	2.1883	-4.391	1.13e-05	***
Renda02_Entre R\$ 2001 e R\$ 5000	7.6025	1.8716	4.062	4.86e-05	***
Renda03_>R\$ 5000	7.1896	1.8202	3.950	7.82e-05	***
Reclamacoes02_Até 3	7.5638	1.9880	3.805	0.000142	***
Reclamacoes03_3 a 5	16.9063	4.1692	4.055	5.01e-05	***
Reclamacoes04_>5	17.3557	4.3401	3.999	6.36e-05	***
Educacao02_Ensino Médio	-2.4020	0.7889	-3.045	0.002330	**
Educacao03_Curso superior	5.0603	1.3914	3.637	0.000276	***
Educacao04_Pós-graduação	5.8573	1.6054	3.648	0.000264	***
Tempo_cliente02_1 a 3 anos	-5.7929	1.4681	-3.946	7.95e-05	***
Tempo_cliente03_>3 anos	-13.4140	3.2689	-4.104	4.07e-05	***
Classe_idade02_26 a 35	-1.7428	1.0640	-1.638	0.101435	
Classe_idade03_36 a 55	5.6446	1.5993	3.529	0.000416	***
Classe_idade04_>55	8.1364	2.1018	3.871	0.000108	***
Idade_carro02_1 a 3 anos	-1.2957	0.4949	-2.618	0.008849	**
Idade_carro03_4 a 6 anos	-3.0279	0.8249	-3.671	0.000242	***
Idade_carro04_7 a 9 anos	-3.4703	0.9174	-3.783	0.000155	***
Idade_carro05_>=10 anos	-4.5284	1.1666	-3.882	0.000104	***
Debito_autom02_Sim	-3.2632	0.8347	-3.909	9.25e-05	***
score	-7.9239	3.0762	-2.576	0.010000	**

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

Two-sample Kolmogorov-Smirnov test

```
data: train$score[train$churn == 0] and train$score[train$churn == 1]
D = 0.87495, p-value < 2.2e-16
alternative hypothesis: two-sided
```

O modelo de churn , gerado a partir da base de treino, foi bastante interessante, todas as variáveis foram bastante significativas. Interessante se isso se refletir também na base de treino.

O K-S do modelo na base de treino foi de 87,50%, um excelente nível de separação e excelentes perspectivas de predição do churn.

Vamos agora avaliar o mesmo modelo gerado na base de treino, na base de teste, para ver como se comporta.



Amostra de Treino

Two-sample Kolmogorov-Smirnov test

```
data: train$score[train$churn == 0] and train$score[train$churn == 1]
D = 0.87495, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Amostra de Teste

Two-sample Kolmogorov-Smirnov test

```
data: test$score[test$churn == 0] and test$score[test$churn == 1]
D = 0.51805, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Comparando o K-S obtido nas bases de treino e teste, encontramos diferenças muito grandes em termos de acurácia. O mesmo modelo que foi top para a base de treino, não teve a mesma performance na base de teste. Não chega a ser um modelo ruim com um K-S de 51,81%, mas essa diferença pode atribuir descrédito ao modelo e, com isso, ele nem ser implementado.

Medidas como validação cruzada, uso de mais variáveis ou, até mesmo, uso de *ensembles* como *random forests*, pode ser benéfico para reduzir esse *overfitting* do modelo.



Exercício – *Overfitting* – Churn em Telecom

68

Na base “[**Telco Customer Churn.xlsx**](#)”, estão disponíveis dados de mais de 7mil clientes de uma operadora de telecom. O objetivo é gerar um modelo na base de treino que prediga a probabilidade de o cliente cancelar sua assinatura, assim como testar se as previsões do modelo continuam relevantes para a base de teste.

Idcliente: ID do cliente

Sexo: Sexo do cliente

Idoso: Se é idoso ou não

Dependentes: Tem dependentes ou não

Tempo_conta: Tempo de conta (em meses)

Internet: Tipo de internet

Streaming_TV: Se o cliente tem serviço de streaming na TV

Streaming_Movies: Se o cliente tem serviço de streaming para filmes

Contrato: Tipo de contrato

Fatura_Digital: Se o cliente tem fatura digital ou não

Método_Pagamento: Método de pagamento

Fatura_Média: Valor médio da fatura

Churn: Cancelou (1), não cancelou (0)

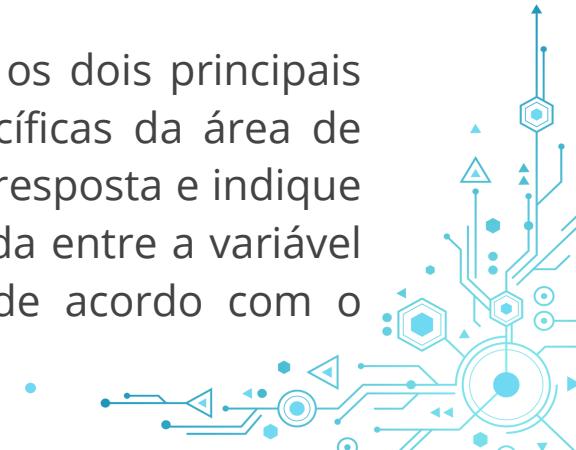
- Separar a base geral em base de treino (80%) e teste (20%) e analise quais são as variáveis que melhor explicam o churn, assim como um modelo de churn na base de treino. Qual o K-S do modelo e como vc o interpreta?
- Algum sinal de *overfitting* nesses dados? Qual a disparidade entre a acurácia do modelo na base de treino e de teste?



Após identificar um alto índice de *turnover* voluntário, a área de RH de uma empresa de serviços decidiu usar a abordagem de People Analytics para tentar identificar fatores que possam retratar os principais motivos pelo qual um colaborador deixa a empresa. Para isso, forneceu os dados de seus funcionários a uma empresa de Consultoria esperando receber inputs sobre esse cenário de modo que possa atuar para reduzir esta taxa de *turnover* observada.

Turnover: são os pedidos de demissão de funcionários em uma empresa. Existem diversos motivos que contribuem para o *turnover*, como: salários baixos, más condições de trabalho e, muitas vezes, a própria insatisfação do funcionário.

- (a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados na visão do negócio. Analise a consistência dos dados e existência de *missing values*.
- (b) Faça uma análise do % de *turnover*. Faz sentido na visão do RH da empresa de serviços?
- (c) Na reunião inicial com a área solicitante do estudo, o diretor de RH indicou acreditar que os dois principais motivos para um funcionário deixar a companhia são: baixo salário e as atividades específicas da área de Vendas. Para avaliar essa afirmação: (c.1) Calcule as medidas resumo da variável Salário vs a resposta e indique se a suspeita do diretor de RH tem fundamento. (c.2) Faça uma tabela de freqüências cruzada entre a variável Departamento e o evento de interesse. Faz sentido dizer que o % de *turnover* muda de acordo com o departamento, e que o departamento de Vendas é o que tem maior pedido de demissão?



Case final – People Analytics (continuação)

70

Após identificar um alto índice de *turnover* voluntário, a área de RH de uma empresa de serviços decidiu usar a abordagem de People Analytics para tentar identificar fatores que possam retratar os principais motivos pelo qual um colaborador deixa a empresa. Para isso, forneceu os dados de seus funcionários a uma empresa de Consultoria esperando receber inputs sobre esse cenário de modo que possa atuar para reduzir esta taxa de *turnover* observada.

Turnover: são os pedidos de demissão de funcionários em uma empresa. Existem diversos motivos que contribuem para o *turnover*, como: salários baixos, más condições de trabalho e, muitas vezes, a própria insatisfação do funcionário.

(d) Ainda na conversa inicial com o diretor de RH, houve a indicação de que aumentar o salário de toda a companhia é totalmente inviável, mas que expandir o benefício de estoque de ações (*stock option*) para todos os funcionários seria uma possibilidade. Qual o percentual de funcionários que possuem este benefício atualmente? Este benefício, isoladamente, parece ser um fator favorável à retenção do funcionário? Avalie a relação entre a posse de ações e o *turnover*.

(e) Faça a análise bivariada das variáveis explicativas (covariáveis) vs. a variável resposta. Quais variáveis discriminam o evento resposta? Como você poderia tratar as categorias com missing values na análise bivariada?

(f) Rode o modelo de Regressão Logística e analise a significância das variáveis explicativas ao nível 5%. Selecione um modelo final no qual a interpretação dos parâmetros esteja de acordo com a análise bivariada.



Após identificar um alto índice de *turnover* voluntário, a área de RH de uma empresa de serviços decidiu usar a abordagem de People Analytics para tentar identificar fatores que possam retratar os principais motivos pelo qual um colaborador deixa a empresa. Para isso, forneceu os dados de seus funcionários a uma empresa de Consultoria esperando receber inputs sobre esse cenário de modo que possa atuar para reduzir esta taxa de *turnover* observada.

Turnover: são os pedidos de demissão de funcionários em uma empresa. Existem diversos motivos que contribuem para o *turnover*, como: salários baixos, más condições de trabalho e, muitas vezes, a própria insatisfação do funcionário.

- (g) Faça a análise de multicolinearidade entre as variáveis explicativas. Reajuste o modelo caso seja necessário, garantindo que as estimativas dos parâmetros fiquem condizentes com a análise exploratória bivariada.
- (h) Qual o perfil, a probabilidade e a representatividade do funcionário mais propenso a apresentar o *turnover*? E do menos propenso?
- (i) Analise a sensibilidade, especificidade e acurácia pela tabela de classificação.
- (j) Como você classifica o desempenho do modelo?
- (k) Pelo valor do KS, você indicaria para área de RH utilizar o modelo?
- (l) Qual o percentual da base de dados que seria classificado como propenso ao *turnover*?

