

PÓS - GRADUAÇÃO

DATA MINING



Coordenação:

Prof. Dr. Adolpho Walter
Pimazzi Canton

Profa. Dra. Alessandra de
Ávila Montini

ESTATÍSTICA APLICADA II

Disciplina: Estatística Aplicada II
Tema da Aula: Fundamentos de *Credit Scoring*

Prof. Marcelo Fernandes

MARCELO FERNANDES

Estatístico graduado na Universidade Federal do Paraná (1999) (registro 8597-A – CONRE 3a. região), pós-graduado em Engenharia Econômica (FAE Business School - 2000), pós-graduado em Pesquisa de Mercado, Opinião e Mídia (ESPM Business School - 2005) e mestre em Administração de Empresas (Universidade Presbiteriana Mackenzie - 2008), com mais de 23 anos de experiência na utilização de Analytics e Machine Learning, aplicados a Risco de Crédito, Cobrança, Fraude e Marketing, em grandes empresas como HSBC, Orbitall, Santander, SPSS, Redecard, Itaú-Unibanco, SAS, Ernst & Young e Paschoalotto. **Atualmente, é Analytic Science Manager na FICO – Fair Isaac Corporation.**



mpfaquila@gmail.com



[+55 \(11\) 9 9962-2701](https://wa.me/5511999622701)



<https://www.linkedin.com/in/marcelopiresfernandes/>



https://www.youtube.com/channel/UCOhh3V9RdkvZiKb9a9nEr0w?view_as=subscriber

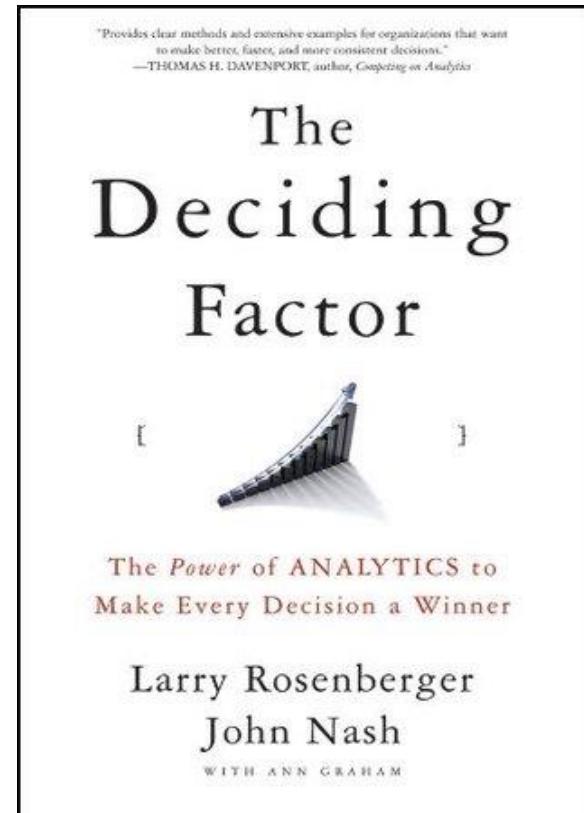
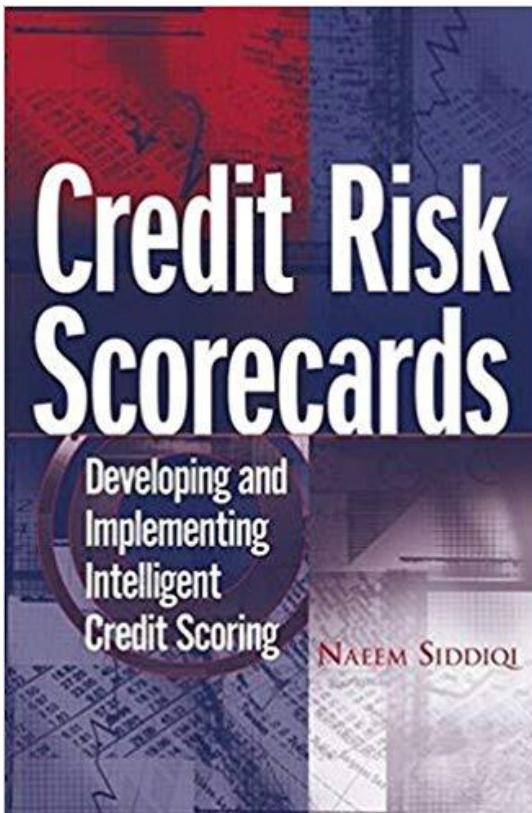
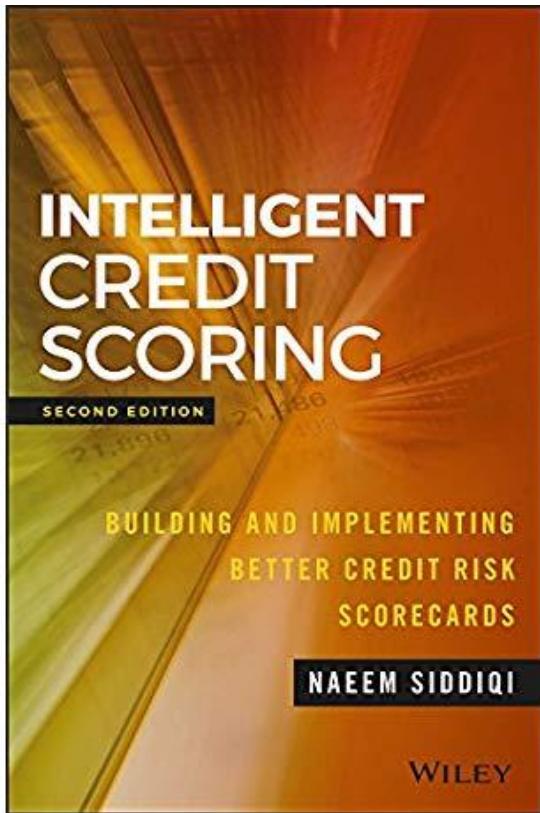
Conteúdo da Aula I de Credit Scoring

Item	Tema	Descrição
01	O que é Credit Scoring ou Escoragem de Crédito ?	Definição, utilização do método e diferenciação básica de outros processos clássicos de escoragem na indústria financeira.
02	Breve Histórico do Processo de Escoragem de Crédito	Como começou esse processo, de onde veio e as motivações de sua aplicação.
03	Resumo do CRISP-DM Aplicado ao Processo de Escoragem de Crédito	Como o lendário framework de mineração de dados, CRISP-DM (<i>Cross-Industry Standard Process for Data Mining</i>) se aplica ao processo de escoragem de crédito.
04	Desenho das Etapas de um Processo de Escoragem de Crédito	Explicação das etapas fundamentais do desenho, construção, aplicação e monitoramento de um modelo de escoragem de crédito.
05	A Definição de “ default ” em crédito: Definição de “bons” e “maus” clientes e a montagem das bases de treinamento e validação	O racional por trás do conceito de “ default ” e sua utilização no desenho de modelos de escoragem de crédito.
06	A Análise Exploratória de Dados – A análise bivariada, o mapa de correlação das variáveis (para avaliação de multicolinearidade) e a classificação de seu poder preditivo.	Uso de técnicas de Estatística Descritiva para identificação inicial de insights para construção de um modelo de escoragem de crédito.
07	O cálculo de métricas como %Bons/%Maus, Peso da Evidência (WOE) e Valor da Informação (IV)	Uso das métricas descritas para calcular o poder preditivo das variáveis criadas para construção de modelos preditivos de escoragem.
08	Processo de Categorização das Variáveis: <i>Fine and Coarse Classing and Optimal Binning</i>	O porquê da categorização de variáveis e o uso de técnicas para aumentar o valor gerado pela categorização.

Conteúdo da Aula II de Credit Scoring

Item	Tema	Descrição
09	Uso da Regressão Logística para Predição de Risco de Crédito	Regressão Logística como método simples e eficiente para criação de modelos preditivos para escoragem de crédito.
10	Avaliação e Validação do Poder Preditivo do Modelo	Métricas de avaliação do poder preditivo do modelo: K-S (<i>Kolmogorov – Smirnov</i>), AUC (<i>Area Under ROC Curve</i>), Matriz de Confusão ou Classificação. Validação do modelo para análise de consistência.
11	Uso de Segmentação Estatística para Refinamento do Processo de Escoragem	Uso de métodos de segmentação estatística para controle da variabilidade dos dados e melhor resultados do processo de escoragem.
12	Avaliação da Estabilidade de um Modelo de Escoragem de Crédito	USO do SPI (<i>Stability Population Index</i>) para definição e acompanhamento do nível de estabilidade das variáveis e do modelo ao longo do tempo.
13	Estratégias de Cálculo do Ponto de Corte para Tomada de Decisão	Alguns possíveis critérios para seleção do ponto de corte, a partir do qual a instituição financeira aceita/aprova o risco de crédito do proponente.
14	Apêndice I: Inferência de Rejeitados para Repescagem de Clientes não aprovados.	O quê fazer com os clientes que não foram aprovados pelo processo de decisão de crédito? Como considerá-los no processo de modelagem, se não temos sua performance?
15	Apêndice II: Uso de <i>Scorecards</i> para Melhor Interpretação de um Modelo de Escoragem de Crédito	Por conta das facilidades de interpretação, explicação e implementação de modelos de risco, o uso de <i>scorecards</i> faz parte das boas práticas de um processo de escoragem.

Referências Bibliográficas



01

O QUE É *CREDIT SCORING* OU ESCORAGEM DE CRÉDITO?

O que é *Credit Scoring* ou Escoragem de Crédito ?



Escoragem de risco, quando aplicada a fins de avaliação de risco de crédito, é chamada de ***Credit Scoring* ou Escoragem de Crédito**. Trata-se de uma ferramenta para avaliação de risco associado a proponentes ou clientes. O número, resultado do cálculo da escoragem de crédito, chama-se ***Credit Score* ou Score de Crédito**.

Diferentes tipos de Scores de Risco

Tipo de Score	Descrição
Credit Scoring	Avalia o risco na concessão, iniciação ou originação do crédito.
Behavior Score	Avalia o risco de um cliente “em dia” tornar-se inadimplente (ou gerar “default”).
Collection Score	Avalia a propensão de um cliente inadimplente pagar sua dívida.
Credit Rating	A propensão à inadimplência pode ser traduzida em letras: AAA, AA, A, BBB, BB, B, ..., D, em que AAA é a menor propensão e D é a maior (<i>default</i>).
FICO Score	Score, desenvolvido pela FICO – Fair Isaac Corporation, fortemente utilizado nos EUA para avaliar risco de inadimplência dos clientes. Contempla não somente os birôs negativos, como também os positivos. Varia entre 300 e 850 e reflete a chance de um cliente ser um bom pagador, em que 300 é o pior score e 850 é o melhor.
Score Serasa	Score desenvolvido pela Serasa Experian, que varia entre 0 e 1.000, em que 0 é o pior score e 1000 é o melhor.

Para quem quiser saber mais...

The screenshot shows the Experian website's header with links for Consumer, Small Business, Business, About Experian, Consumer Support, Credit Advice, and Global Sites. Below the header is a navigation bar with links for Reports & Scores, Identity Theft Protection, CreditMatch, Support, Education, Sign In, and a search icon. A black banner across the top has 'Categories' on the left and 'Subscribe' on the right. The main content area features a blurred background image of a person holding a document. The title 'Understanding Credit Scores' is centered above a section titled 'Types of Credit Scores'. Below this, there is explanatory text and two paragraphs detailing generic and custom credit scores. Social media sharing icons for Facebook, Twitter, and LinkedIn are visible on the left side of the main content.

Types of Credit Scores

A credit score is a number lenders use to help them decide how likely it is that they will be repaid on time if they give a person a loan or a credit card. Your personal credit score is built on your credit history. Your Experian credit score ranges from 330 to 830. A decent credit score is essential for your financial well-being because the higher it is, the less of a credit risk you are. There are primarily two types of credit scores, generic scores and custom scores:

Generic credit scores are used by many types of lenders and businesses to determine general credit risk. You can access your generic score as one score using the same formula across all three credit reporting agencies.

Custom credit scores are developed for use by individual lenders. They rely on credit reports and other information, such as account history, from the lender's own portfolio. They are unique to the specific business, or they may be used by specific types of lenders, such as credit unions. Custom credit scores can apply to specific types of lending, such as mortgage lending or auto lending.

Fonte: <https://www.experian.com/blogs/ask-experian/credit-education/score-basics/understanding-credit-scores/>

02

BREVE HISTÓRICO DOS PROCESSOS DE ESCORAGEM DE CRÉDITO

Histórico dos Processos de Escoragem de Crédito

A origem do score de crédito deriva de uma pergunta muito simples: ***Posso contar que vou emprestar dinheiro (conceder crédito) a um determinado indivíduo e vou receber o dinheiro de volta?***

DID YOU KNOW?

Credit scores started in the 1950s when Bill Fair, an engineer and Earl Isaac, a mathematician created an automated scoring system. They eventually found the Fair Isaac Corporation, developing and selling their credit scoring system to banks, retailers and corporations in the US and around the world.

MBOA MORTGAGE COMPANY | MBOAMTG.COM

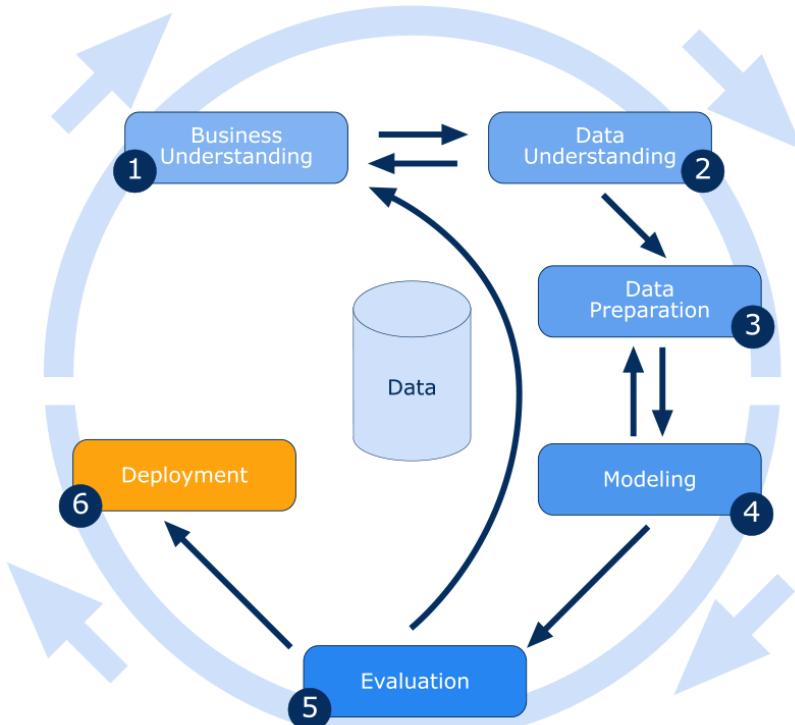
O engenheiro William Fair e o matemático Earl Isaac foram os precursores dos métodos de escoragem de risco, criados na década de 1950. Em 1956, eles criaram a Fair Isaac Corporation, que mais tarde, na década de 1980, deu origem ao **FICO Score**, o mais importante processo de escoragem do mundo.

FICO, um *rebranding* da Fair Isaac Corporation, é hoje uma empresa americana de software, dona do FICO Score e de ferramentas para gerenciamento dos processos de decisão nas empresas.

03

RESUMO DO CRISP-DM APLICADO AO PROCESSO DE ESCORAGEM DE CRÉDITO

CRISP-DM Aplicado ao Processo de Escoragem de Crédito



Etapa	Contexto
Entendimento do Problema de Negócio	Que problema de negócio a empresa quer resolver e qual o contexto por trás desse problema?
Entendimento dos Dados	Quais são os dados disponíveis, associados ao problema de negócio a ser estudado?
Preparação dos Dados	Tratamento, limpeza, consolidação, descrição, análise e insights sobre os dados associados ao problema de negócios
Modelagem	Uso de métodos quantitativos (regressão, redes neurais, árvores aleatórias, máquinas de vetor de suporte, etc.) para modelagem analítica dos dados.
Avaliação do Modelo	Quão preciso e estável é o modelo desenvolvido? Podemos confiar nele?
Implementação	Etapa em que o modelo passa a ser utilizado para auxiliar no processo de decisão e ajuda a empresa a melhor selecionar os clientes, de acordo com seu apetite de risco.

Fonte: <https://www.semantix.com.br/blog/como-explorar-e-gerenciar-dados-com-o-crisp-dm>

Para saber mais sobre o CRISP-DM...

Data Science Central® THE ONLINE RESOURCE FOR BIG DATA PRACTITIONERS

Home | AI | ML | DL | Analytics | Statistics | Big Data | DataViz | Hadoop | Podcasts | Webinars | Forums | Jobs | Membership | Groups | Search | Contact |

Explore all 10 trends

2019 Business Intelligence Trends

SEE THE TRENDS

+ + + + + + + + + +

Subscribe to DSC Newsletter

All Blog Posts My Blog

CRISP-DM – a Standard Methodology to Ensure a Good Outcome

Posted by William Vorhees on July 26, 2016 at 9:15am [View Blog](#)

Summary: To ensure quality in your data science group, make sure you're enforcing a standard methodology. This includes not only traditional data analytic projects but also our most advanced recommenders, text, image, and language processing, deep learning, and AI projects.

A Little History

In the early 1990's as data mining was evolving from toddler to adolescent we spent a lot of time getting the data ready for the fairly limited tools and limited computing power of the day. Seldom were there more than one or two 'data scientists' in the same room and we were much more likely to be called 'predictive modelers' since that modeling was state-of-the-art in its day.

As the 90's progressed there was a natural flow that drew us toward standardizing the lessons we'd learned into a common methodology. Efforts like this always start out by wondering aloud whether there even was a common approach given that the problems looked so dissimilar. As it turns out there was.

Two of leading providers of the day, SPSS and Teradata, along with three early adopter user corporations, Daimler, NCR, and CHRA convened a special interest group (SIG) in 1996 (also probably one of the earliest collaborative efforts over the newly available worldwide web) and over the course of less than a year managed to codify what is still today the CRISP-DM, Cross Industry Standard Process for Data Mining. I'm honored to say that I was one of the original contributors to that SIG.

CRISP-DM was not actually the first. SAS Institute that's been around longer than anyone can remember had its own version called SEMMA (Sample, Explore, Modify, Model, Assess) but within just a year or two many more practitioners were basing their approach on CRISP-DM.

What is CRISP-DM?

The process or methodology of CRISP-DM is described in these six major steps

- Business Understanding**

Focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan

RESOURCES

- Join DSC
- Free Books
- Forum Discussions
- Cheat Sheets

SEE THE TRENDS

+ + + + + + + + + +



Anúncio

stellar SQL Database Repair Software Stellar Data Recovery

Download

Download citation Share Download full-text PDF

Análise de Crédito Utilizando uma Abordagem de Mineração de Dados

Article (PDF Available) · September 2018 with 4 Reads

DOI: 10.25386/epa.v03.i97

Cite this publication

Joyce Maria

Iago Richard Rodrigues Silva

Raniel Gomes Da Silva

Gustavo Arcovado Luis

Show more authors

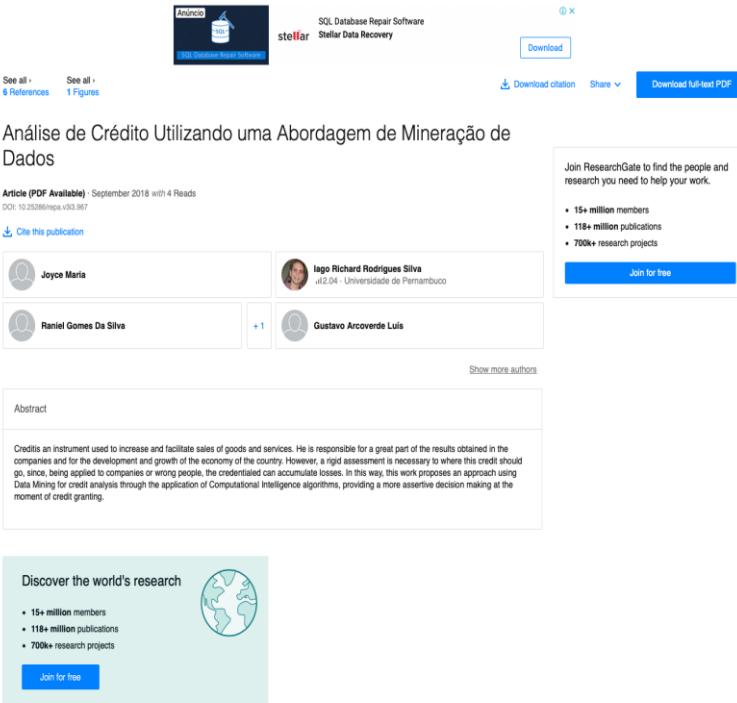
Abstract

Credit is an instrument used to increase and facilitate sales of goods and services. He is responsible for a great part of the results obtained in the companies and for the development and growth of the economy of the country. However, a rigid assessment is necessary to where this credit should go, since, being applied to companies or wrong people, the credential can accumulate losses. In this way, this work proposes an approach using Data Mining for credit analysis through the application of Computational Intelligence algorithms, providing a more assertive decision making at the moment of credit granting.

Discover the world's research

+ 15+ million members
+ 118+ million publications
+ 700k+ research projects

Join for free



Fonte: <https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>

Fonte: https://www.researchgate.net/publication/329922257_Analise_de_Credito_Utilizando_uma_Abordagem_de_Mineracao_de_Dados

04

DESENHO DAS ETAPAS DE UM PROCESSO DE ESCORAGEM DE CRÉDITO

Desenho das Etapas de um Processo de Escoragem de Crédito



05

A DEFINIÇÃO DE “*DEFAULT*” EM CRÉDITO E O CONCEITO DE “BONS” E “MAUS” CLIENTES

A definição de “*default*” em Crédito: A definição de “bons” e “maus” clientes para a construção de um Credit Scoring



- ❖ É uma definição que deve estar em linha com as políticas e estratégias do produto que será considerado para a modelagem (cartão de crédito, cheque especial, crédito pessoal, etc.).
- ❖ Via de regra, quando o cliente ultrapassa 60 dias de atraso, até 12 meses após o início da observação, pode ser classificado como “mau”. Clientes que não tiverem entrado em atraso no período podem ser classificados como “bons”. Os demais podem ser classificados como “indeterminados”.
- ❖ É fundamental analisar clientes em diversas safras de observação, para capturar diferentes efeitos, incluindo sazonais. Contudo, é importante que a base de análise não contemple um período muito extenso, cujas políticas de crédito ou características do portfolio tenham se alterado significativamente. Sugestão inicial poderia ser entre 18 e 24 safras de observação.
- ❖ O tempo de observação dependerá do horizonte de previsão a ser considerado. Usualmente, 12 meses é o período mais comum, podendo, em alguns casos, ser menor.

A definição de “*default*” em Crédito: A definição de “bons” e “maus” clientes para a construção de um Credit Scoring



- ❖ Nesse diagrama usado como exemplo, estamos considerando 24 safras de análise, cada uma delas com 12 meses de observação.
- ❖ Em cada safra, fazemos a marcação do cliente como “bom”, “mau” ou “indeterminado”, em função da evolução de seus níveis de atraso ao longo dos 12 meses, observando o status do atraso, mês a mês. Assim, se a definição de “mau” for atraso > 60 dias, aqueles clientes que ultrapassarem esse atraso ao longo desse período, independente de terem ou não retornado ao status “em dia”, serão classificados como “maus”, para efeito de modelagem.
- ❖ Um bom critério para compor as ABT's de Treinamento e Validação seria deixar as últimas 3 safras de concessão de crédito (mais recentes) como ABT's de validação (usadas somente para validação do modelo) e as demais, como ABT's de treinamento (usadas somente para o desenvolvimento do modelo).

ANÁLISE EXPLORATÓRIA DE DADOS:

ANÁLISE BIVARIADA, MAPA DE CORRELAÇÃO E AVALIAÇÃO DO PODER PREDITIVO DAS VARIÁVEIS (WOE E VALOR DA INFORMAÇÃO)

O Processo de Análise Exploratória de Dados – A Análise Bivariada

Um dos pontos fundamentais da análise exploratória de dados, sobretudo em problemas de predição, é a avaliação da relação entre as variáveis explicativas e a variável resposta. Uma maneira consagrada e utilizada ao longo dos anos tem sido a **análise de uma tabela cruzada** (na linha vai a variável explicativa e na coluna vai a variável resposta) que, popularmente, é chamada de **análise bivariada**.

O propósito da **análise bivariada** é identificar o quanto forte é a variável de análise para explicar a variável resposta. Por exemplo, um estudo pretende identificar as variáveis mais importantes para explicar se um cliente paga ou não paga. A variável em análise nesse caso é a “quantidade de contatos da assessoria de cobrança com o cliente certo (chamado de CPC – contato com a pessoa certa). A análise bivariada proporciona uma visão maravilhosa da relação dessa variável com o pagamento da dívida.

Quantidade de CPC (contatos com a pessoa certa)	QUANTIDADE DE CONTATOS COM O CLIENTE – ÚLTIMOS 3 MESES								
	Não		Sim		Total	%	B/M	WOE	INF. VALUE
	#	%	#	%					
Zero	1.366.039	66,3%	23.128	23,7%	1.389.167	64,4%	0,36	-103,0	0,4393
Um	289.587	14,1%	20.356	20,8%	309.942	14,4%	1,48	39,3	0,0266
Dois	139.243	6,8%	13.583	13,9%	152.826	7,1%	2,06	72,1	0,0515
Três	82.422	4,0%	8.202	8,4%	90.624	4,2%	2,10	74,1	0,0325
Quatro	103.220	5,0%	14.305	14,6%	117.525	5,4%	2,92	107,2	0,1032
Cinco	79.775	3,9%	18.157	18,6%	97.932	4,5%	4,80	156,8	0,2306
Total	2.060.286	100%	97.730	100,0%	2.158.015	100,0%	1,00	-	0,8838
								FORTE	

O Processo de Análise Exploratória de Dados – A Análise Bivariada

A análise dos percentuais revela que a incidência de pagamentos é bem mais baixa quando não se fala com o titular da dívida, enquanto que esse número se eleva consideravelmente à medida que a quantidade de CPC aumenta. Visualmente, é possível perceber isso. Contudo, para facilitar esse processo para um número maior de variáveis (dezenas, centenas de variáveis), a força da relação entre as variáveis pode ser traduzida por um único número gerado por um indicador chamado “**valor da informação (VI)**”, que contempla não apenas a diferença entre %bons e %maus para cada categoria, mas também considera, de algum modo, a penetração desse atributo na carteira.

$$VI = \sum_{i=0}^n \ln\left(\frac{\%Bons_i}{\%Maus_i}\right) * (\%Bons_i - \%Maus_i)$$

O indicador WOE (*weight of evidence*) ou “**peso da evidência**”, é um indicador da força do atributo e é parte do cálculo do VI, como $WOE = \ln\left(\frac{\%Bons_i}{\%Maus_i}\right)$.

Um artigo publicado no SAS Global Forum de 2013, denominado “**Variable Reduction in SAS by Using Weight of Evidence and Information Value**” traz, inclusive uma macro para automatizar no SAS esse processo de cálculo do valor da informação. Mais do que isso, traz uma sugestão de classificação da variável de análise, em função da dimensão do “**valor da informação**”.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

Valor da informação (VI)	Classificação
$\leq 0,02$	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

Nesse mesmo artigo, os autores sugerem que se o VI for superior a 0,50, a variável tem poder preditivo “suspeito” ou “alto demais” e é importante checar a consistência do dado, bem como a relevância prática dessa variável, assim como, eventualmente, o modo como foi gerada.

A partir do cálculo do VI gerado para todas as variáveis, essa lista pode ser ordenada e uma maior importância inicial pode ser dada às variáveis mais fortes (com VI mais elevado), reduzindo sensivelmente o esforço de processamento na modelagem, dedicando atenção para um conjunto menor de variáveis.

IMPORTANTE: O processo de análise bivariada gera o cálculo de WOE (peso da evidência) e VI (valor da informação) para cada atributo da variável de análise. O processo “ótimo” de agrupamento dos atributos (“*optimal binning*”) se dá pela junção de atributos com valores similares de peso da evidência, sobretudo em atributos próximos.

Fonte: <http://support.sas.com/resources/papers/proceedings13/095-2013.pdf>

CÁLCULO DAS MÉTRICAS DE AVALIAÇÃO DO PODER PREDITIVO DAS VARIÁVEIS E AVALIAÇÃO DE SINAIS DE MULTICOLINEARIDADE

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

Considere a tabela **Empréstimo Bancário.xlsx**, que traz dados de 5.000 propostas passadas de crédito, geradas para solicitação de um empréstimo. A base traz dados como idade, nível de instrução, tempo de experiência, tempo no endereço e renda, além da variável classif (0=bom, 1=mau). Vamos avaliar o potencial preditivo inicial dessas variáveis para predizer se um cliente será um mau pagador.

The screenshot shows the RStudio interface with two panes. The top pane displays R code for importing a dataset:

```
1 #Importação da tabela "Empréstimo Bancário.xlsx"
2 install.packages("readxl")
3 library(readxl)
4 empbanc<-read_excel("Emprestimo_Bancario.xlsx", sheet ="Plan1")
5 View(empbanc)
```

The bottom pane shows a data preview of the 'empbanc' dataset:

ID	idade	educacao	experiencia	tempo_endereco	renda	classif
1	41	3	17	12	35.9	0
2	30	1	13	8	46.7	0
3	40	1	15	14	61.8	0
4	41	1	15	14	72.0	0
5	57	1	7	37	25.6	0
6	45	1	0	13	28.1	0
7	36	1	1	3	19.6	1
8	39	1	20	9	80.5	0
9	43	1	12	11	68.7	0
10	34	3	7	12	33.8	0
11	26	1	1	2	22.2	0
12	37	2	17	10	78.3	1
13	44	1	8	15	77.8	0
14	36	2	8	1	48.1	1
15	27	2	1	8	34.7	0
16	35	1	7	13	42.8	0
17	40	1	12	7	41.2	0
18	23	1	5	3	20.6	0
19	61	1	84	7	144.7	0

Showing 1 to 19 of 5,000 entries

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

Existem pacotes prontos no R para calcular o valor da informação para cada variável. Contudo, a título de experiência, vamos fazer esse cálculo na mão, primeiro agrupando as variáveis em classes (nesse exemplo, vamos agrupar em 4 quartis), depois cruzando cada variável com a variável resposta (“classif”), para analisarmos os resultados:

```
7 #Agrupando cada variável em classes (nesse caso, 4 classes)
8 install.packages("arules")
9 library(arules)
10 idade_cl<-discretize(empbanc$idade, "frequency", breaks = 4)
11 experiencia_cl<-discretize(empbanc$experiencia, "frequency", breaks = 4)
12 tempend_cl<-discretize(empbanc$tempo_endereco, "frequency", breaks = 4)
13 renda_cl<-discretize(empbanc$renda, "frequency", breaks = 4)
14 empbanc<-data.frame(empbanc, idade_cl, experiencia_cl,tempend_cl,renda_cl)
15 View(empbanc)
```

ID	idade	educacao	experiencia	tempo_endereco	renda	classif	idade_cl	experiencia_cl	tempend_cl	renda_cl
1	1	41	3	17	12	35.9	0 [41,58] [13,38]	[12,37] [34,5,54,7]		
2	2	30	1	13	8	46,7	0 [29,35] [13,38]	[7,12) [34,5,54,7)		
3	3	40	1	15	14	61,8	0 [35,41] [13,38]	[12,37] [54,7,2,46e+03]		
4	4	41	1	15	14	72,0	0 [41,58] [13,38]	[12,37] [54,7,2,46e+03]		
5	5	57	1	7	37	25,6	0 [41,58] (7,13)	[12,37] [24,5,34,5)		
6	6	45	1	0	13	28,1	0 [41,58] [0,3)	[12,37] [24,5,34,5)		
7	7	36	1	1	3	19,6	1 [35,41] [0,3)	[3,7) [12,1,24,5)		
8	8	39	1	20	9	80,5	0 [35,41] [13,38)	[7,12) [54,7,2,46e+03]		
9	9	43	1	12	11	68,7	0 [41,58] (7,13)	[7,12) [54,7,2,46e+03]		
10	10	34	3	7	12	33,8	0 [29,35] (7,13)	[12,37] [24,5,34,5)		
11	11	26	1	1	2	22,2	0 [20,29] [0,3)	[0,3) [12,1,24,5)		
12	12	37	2	17	10	79,3	1 [35,41] (13,38)	[7,12) [54,7,2,46e+03]		

O comando “**discretize**”, bastante útil para a categorização de variáveis quantitativas, pode ser utilizado a partir do package “**arules**”.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

Agora, vamos cruzar as variáveis idade_cl, educação, experiencia_cl, tempend_cl e renda_cl com a variável classif.

```
17 #Cruzando cada variável com o target "classif"
18 install.packages("descr")
19 library(descr)
20 CrossTable(empbanc$idade_cl, empbanc$classif, prop.r = FALSE, prop.t = FALSE,
21             prop.chisq = FALSE)
22 CrossTable(empbanc$educacao, empbanc$classif, prop.r = FALSE, prop.t = FALSE,
23             prop.chisq = FALSE)
24 CrossTable(empbanc$experiencia_cl, empbanc$classif, prop.r = FALSE, prop.t = FALSE,
25             prop.chisq = FALSE)
26 CrossTable(empbanc$tempend_cl, empbanc$classif, prop.r = FALSE, prop.t = FALSE,
27             prop.chisq = FALSE)
28 CrossTable(empbanc$renda_cl, empbanc$classif, prop.r = FALSE, prop.t = FALSE,
29             prop.chisq = FALSE)
```

O pacote “**descr**” possui a função “**CrossTable**” para geração de tabelas cruzadas mais interessantes, se comparadas à utilização padrão do comando “**table**” para esse fim.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

empbanc\$idade_cl	0	1	Total
[20,29)	690 0.184	424 0.338	1114
[29,35)	966 0.258	370 0.295	1336
[35,41)	944 0.252	239 0.190	1183
[41,58]	1144 0.306	223 0.178	1367
Total	3744 0.749	1256 0.251	5000

Ao observar o resultado do cruzamento da variável “idade_cl” com a variável resposta, “classif (em que 0=bons, 1=maus)”, chegamos a algumas constatações interessantes:

- ❖ Na faixa de 20-29 anos, existe uma concentração maior de maus clientes do que de bons.
- ❖ À medida que a idade avança, essa relação passa a se inverter, tanto que, na classe de 41-58 anos, a concentração de bons é maior que a de maus clientes.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

		empbanc\$classif		
empbanc\$educacao		0	1	Total
1		2128	571	2699
		0.568	0.455	
2		1000	365	1365
		0.267	0.291	
3		388	178	566
		0.104	0.142	
4		199	125	324
		0.053	0.100	
5		29	17	46
		0.008	0.014	
Total		3744	1256	5000
		0.749	0.251	

Ao observar o resultado do cruzamento da variável “nível educacional (em que 1=fundamental, 2=médio, 3=superior, 4=mestrado e 5=doutorado)” com a variável resposta, “classif (em que 0=bons, 1=maus)”, chegamos ao seguinte:

- ❖ Pessoas com ensino fundamental apresentaram maior tendência a serem bons que maus.
- ❖ Para os demais níveis, houve uma inversão, indicando maior propensão ao grupo de maus clientes.
- ❖ Isso, de certa forma, contraria os padrões históricos de relevância para essa variável, à medida que pessoas com mais instrução teriam menor propensão à inadimplência. A causa pode ser tanto um comportamento distinto no produto “crédito pessoal”, quanto um viés causado pela potencial baixa qualidade desse dado cadastral.
- ❖ Mesmo assim, para quase todas as classes, a diferença entre os %maus e os %de bons não é tão relevante, o que sugere ser uma variável relativamente mais fraca.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

empbanc\$experiencia_cl	empbanc\$classif		
	0	1	Total
[0 ,3)	629	478	1107
	0.168	0.381	
[3 ,7)	844	385	1229
	0.225	0.307	
[7 ,13)	1090	254	1344
	0.291	0.202	
[13 ,38]	1181	139	1320
	0.315	0.111	
Total	3744	1256	5000
	0.749	0.251	

Ao analisar o resultado do cruzamento da variável “experiência_cl” com a variável resposta, “classif (em que 0=bons, 1=maus)”, observa-se que:

- ❖ Quanto menor o nível de experiência, maior a tendência de o cliente apresentar comportamento de inadimplência.
- ❖ Esse mesmo comportamento foi observado na variável idade_cl, de forma que, mesmo sem previamente termos calculado a correlação entre idade do cliente e tempo de experiência, já podemos suspeitar que existe uma alta associação entre essas 2 variáveis.
- ❖ Sobretudo nas pontas, existe uma diferença bastante relevante entre os % de maus e os % de bons, o que sugere que experiência_cl seja uma variável bastante forte.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

empbanc\$tempend_cl	empbanc\$classif		Total
	0	1	
[0,3)	741	424	1165
	0.198	0.338	
[3,7)	911	405	1316
	0.243	0.322	
[7,12)	903	251	1154
	0.241	0.200	
[12,37]	1189	176	1365
	0.318	0.140	
Total	3744	1256	5000
	0.749	0.251	

Ao analisar o resultado do cruzamento da variável “tempend_cl” com a variável resposta, “**classif (em que 0=bons, 1=maus)**”, observa-se que:

- ❖ Quanto menor o tempo no mesmo endereço, maior a tendência de o cliente apresentar comportamento de inadimplência.
- ❖ Esse mesmo comportamento foi observado na variável idade_cl e na variável experiência_cl, de forma que, mesmo sem previamente termos calculado a correlação entre essas variáveis, já podemos suspeitar que existe uma alta associação entre elas.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

		empbanc\$classif		
empbanc\$renda_cl		0	1	Total
[12.1,24.5)		824	410	1234
		0.220	0.326	
[24.5,34.5)		925	337	1262
		0.247	0.268	
[34.5,54.7)		992	262	1254
		0.265	0.209	
[54.7,2.46e+03]		1003	247	1250
		0.268	0.197	
Total		3744	1256	5000
		0.749	0.251	

Ao analisar o resultado do cruzamento da variável “renda_cl” com a variável resposta, “classif (em que 0=bons, 1=maus)”, observa-se que:

- ❖ Quanto menor a renda do cliente, maior sua tendência à inadimplência.
- ❖ Analogamente, o comportamento inverso também ocorre, já que para faixas mais altas de renda, existe uma maior concentração em bons clientes do que nos maus.

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

Para facilitar e acelerar a avaliação do poder preditivo da variável, podemos usar comando “`create_infotables`”, presente no pacote “`Information`” do R para calcular e classificar as 5 variáveis `idade_cl`, `experiencia_cl`, `educacao`, `tempend_cl`, `renda_cl`, com relação à sua importância de explicar a propensão à inadimplência.

```
31 #Calculando o nível de importância de cada variável para explicar "classif"
32 install.packages("Information")
33 library(Information)
34 empbanc_select<-empbanc[,c(3,7,8,9,10,11)]
35 View(empbanc_select)
36 IV <- create_infotables(data = empbanc_select, y = "classif", ncore = 2)
37 print(head(IV$Summary, 10), row.names = FALSE)
38 print(IV$Tables$experiencia_cl, row.names = FALSE) #Apenas para exemplificar
```

A ordem de importância apresentada a partir da execução do comando “`create_infotables`” bate com a análise que fizemos anteriormente?

O Processo de Análise Exploratória de Dados – A Análise Bivariada

EXERCÍCIO 01

```
Console Terminal ×  
~/Desktop/ ↗  
> print(head(IV$Summary, 10), row.names = FALSE)  
      Variable        IV  
experiencia_cl 0.44561702  
tempend_cl 0.24980947  
idade_cl 0.18452749  
renda_cl 0.07917799  
educacao 0.06685346
```

De acordo com a tabela que classifica o IV como suspeito, forte, médio, fraco e irrelevante, temos que:

- ❖ experiência_cl é **FORTE**, tempend_cl é **MÉDIA**, idade_cl é **MÉDIA**, renda_cl e educacao são **FRACAS**.

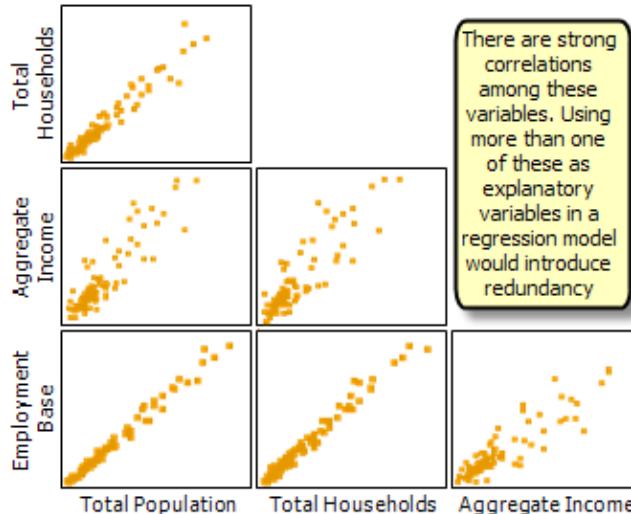
Percebam que, pela simples análise das percentagens, já se previa que a variável “experiência_cl” seria a mais forte e que “educacao” seria a mais fraca. Apenas a título de exemplo, vejamos o cálculo do valor da informação para a variável experiência_cl:

```
> print(IV$Tables$experiencia_cl, row.names = FALSE) #Apenas para exemplificar  
      experiencia_cl    N Percent       WOE        IV  
[0,3) 1107  0.2214  0.8177020  0.1738198  
[3,38] 1320  0.2640 -1.0474204  0.3882993  
[3,7) 1229  0.2458  0.3073133  0.4132228  
[7,13) 1344  0.2688 -0.3643762  0.4456170
```

$$\begin{aligned} IV_{\text{experiencia_cl}} &= \ln(0.168/0.381) * (0.168 - 0.381) + \\ &\quad \ln(0.225/0.307) * (0.225 - 0.307) + \\ &\quad \ln(0.291/0.202) * (0.291 - 0.202) + \\ &\quad \ln(0.315/0.111) * (0.315 - 0.111) = \\ &0,174 + 0,025 + 0,032 + 0,213 = \mathbf{0,445} \end{aligned}$$

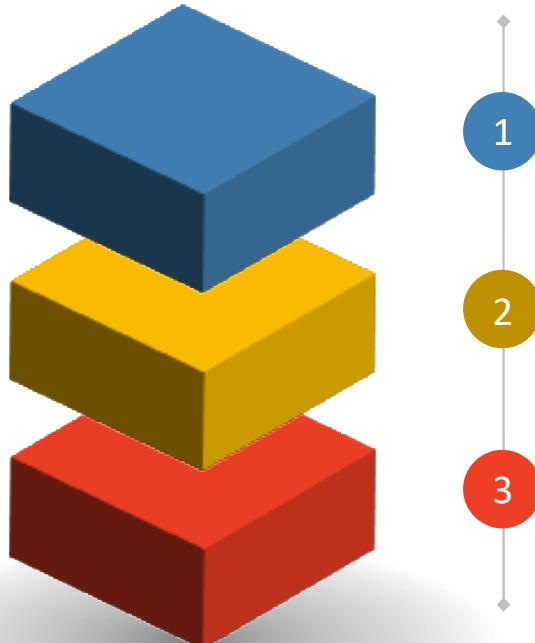
O Processo de Análise Exploratória de Dados – O Problema da Multicolinearidade

Quando trabalhamos com dados multivariados, como é o caso desse exercício, podem existir variáveis que carreguem informações similares. Esse tipo de redundância, quando falamos de análise de regressão, chama-se multicolinearidade.



- ❖ De uma forma bem simples, multicolinearidade é um efeito que ocorre quando variáveis fortemente correlacionadas são consideradas no mesmo modelo matemático de regressão.
- ❖ Quando elas entram simultaneamente no modelo, a consequente redundância potencialmente eleva o nível de variabilidade dos pesos das variáveis correlacionadas, distorcendo sua interpretação e tornando o modelo instável e, muitas vezes, alterando a lógica esperada das variáveis envolvidas.

O Processo de Análise Exploratória de Dados – O Problema da Multicolinearidade



Seleção de Variáveis

Testar no modelo variáveis não fortemente correlacionadas entre si (sugestão: correlação<0,5, em módulo).

Análise de Componentes Principais

Criação de fatores ortogonais (não correlacionados), resultado da combinação linear das variáveis

Métodos de Regularização

Introdução de penalidades no valor dos parâmetros para estabilizar sua variabilidade, reduzir o VIF e minimizar a multicolinearidade. Como possíveis métodos, existem a Regressão Ridge, a Regressão LASSO e a Regressão Logística Regularizada.

O Processo de Análise Exploratória de Dados – O Problema da Multicolinearidade

VIF (Variance Inflation Factor) ou Fator de Inflação da Variância: Mede o quanto variância dos coeficientes de regressão (pesos) são inflacionados por problemas de multicolinearidade. Os autores mais conservadores consideram que **VIF>5 já revela sinais de multicolinearidade**. Esse é um bom parâmetro de referência para identificar sinais de multicolinearidade.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Regression Analysis: HeatFlux versus East, South, North

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	389.2	66.1	5.89	0.000	
East	2.12	1.21	1.75	0.092	1.12
South	5.318	0.963	5.52	0.000	1.21
North	-24.13	1.87	-12.92	0.000	1.09

Regression Equation

$$\text{HeatFlux} = 389.2 + 2.12 \text{ East} + 5.318 \text{ South} - 24.13 \text{ North}$$

Fonte: <http://blog.minitab.com/blog/statistical-software/what-in-the-world-is-a-vif>

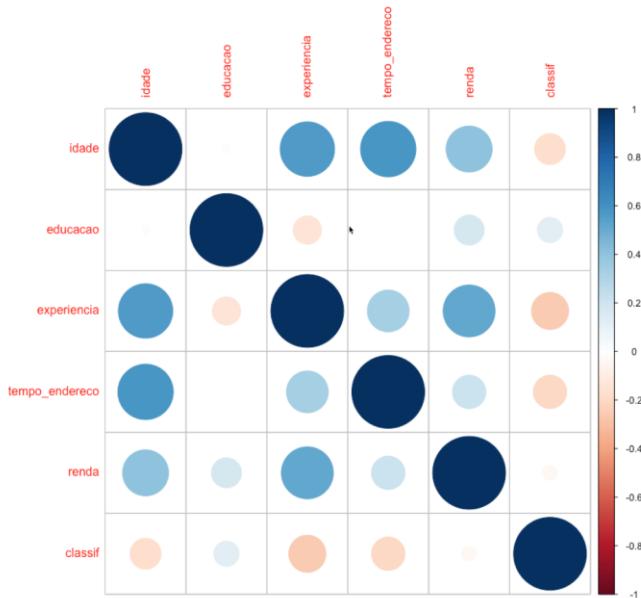
O Processo de Análise Exploratória de Dados – O Problema da Multicolinearidade

EXERCÍCIO 02

Continue considerando a tabela **Empréstimo Bancário.xlsx**, para realizarmos uma análise de correlação das variáveis nessa tabela. Vamos usar a matriz de correlação tradicional, assim como considerar o pacote, para mostrar a matriz de correlação de uma forma visual:

```
40 #Cálculo da correlação entre as variáveis
41 matriz_correl <- round(cor(empbanc[,2:7]), 2)
42 matriz_correl
43 install.packages("corrplot")
44 library(corrplot)
45 matriz_correl_I<-corrplot(matriz_correl, method = "circle")
46 matriz_correl_I

> matriz_correl
      idade educacao experiencia tempo_endereco renda classif
idade    1.00     0.01      0.56        0.58   0.40 -0.18
educacao  0.01    1.00     -0.15       0.00   0.17   0.12
experiencia 0.56   -0.15     1.00       0.33   0.51 -0.26
tempo_endereco 0.58     0.00      0.33       1.00   0.21 -0.21
renda      0.40     0.17      0.51       0.21   1.00 -0.04
classif    -0.18     0.12     -0.26      -0.21 -0.04   1.00
```



O Processo de Análise Exploratória de Dados – O Problema da Multicolinearidade

EXERCÍCIO 02

Vamos também identificar, por meio do VIF, se há algum sinal potencial de multicolinearidade. O pacote “HH” possui a função `vif`, que nos possibilita essa análise.

```
48 #Análise do VIF para detectar potencial multicolinearidade
49 install.packages("HH")
50 library(HH)
51 model <- lm(classif ~ idade_cl+educacao+experiencia_cl+tempend_cl+
52               renda_cl, data = empbanc)
53 summary(model)
54 vif(model)
```

```
Console Terminal ×
~/Desktop/ ▶

> vif(model)
      idade_cl[29,35]      idade_cl[35,41]      idade_cl[41,58]
      1.908814          2.221536          2.799530
      educacao      experiencia_cl[3,7]      experiencia_cl[7,13]
      1.190592          1.655418          2.008451
experiencia_cl[13,38]      tempend_cl[3,7]      tempend_cl[7,12]
      3.000558          1.587680          1.672787
      tempend_cl[12,37]      renda_cl[24.5,34.5]      renda_cl[34.5,54.7]
      2.058734          1.695835          2.149561
      renda_cl[54.7,2.46e+03]
      3.080268
```

Como o valor do VIF não foi superior a 5 para nenhuma das variáveis, não há sinais aparentes da presença de multicolinearidade e, portanto, nada a fazer nesse sentido.

08

PROCESSO DE CATEGORIZAÇÃO DAS VARIÁVEIS

Processo de Categorização das Variáveis: *Fine and Coarse Classing and Optimal Binning*

1

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

2

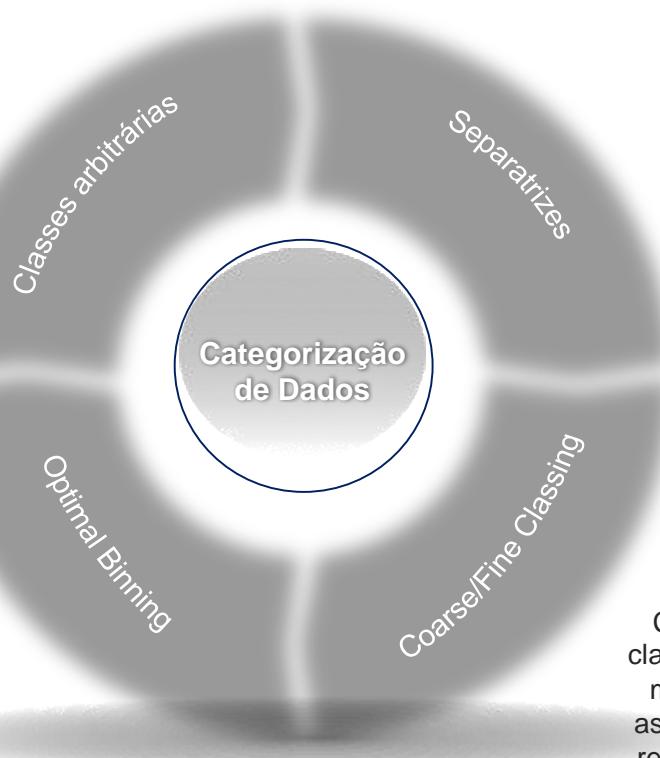
É bacana, pois é possível separar de forma equilibrada os dados em um número quaisquer de classes, baseando-se em quartis, decis ou percentis.

4

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

3

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



Por quê Categorizar Variáveis Pode Ser um Bom Negócio?

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

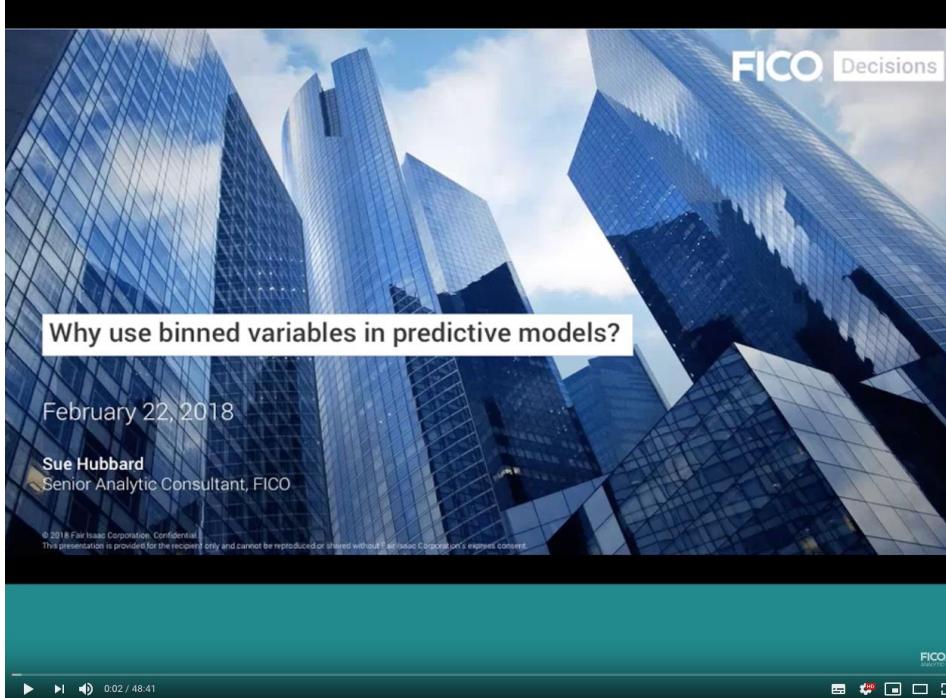


É mais fácil para explicar o modelo para não-modeladores (*C-level* ou outras pessoas de negócio)

Com variáveis categóricas, é mais fácil isolar o efeito de dados extremos ou faltantes.

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

Processo de Categorização das Variáveis: *Fine and Coarse Classing and Optimal Binning – Quais os benefícios de categorizar variáveis?*



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

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

A idéia por trás do processo de *coarse and fine classing* é criar um determinado número de faixas (ou classes, ou "bins"), ou seja, **aplicar o *fine classing***, de forma a, na sequência, **aplicar o *coarse classing***, agrupando aquelas classes adjacentes com níveis de risco parecidos e ter categorias com níveis de risco discriminantes entre si (por meio do WOE, ou peso da evidência).

Para exercitar esse conceito, vamos utilizar a tabela "**Loan Payment Data.xlsx**", que traz alguns dados do cliente sobre uma determinada operação de crédito.

```
56 #Importação da tabela "Loan Payments Data"
57 loan<-read_excel("Loan_Payments_Data.xlsx", sheet ="Planilha1")
58 View(loan)
59 loan$classif<-ifelse(loan$loan_status == "COLLECTION", 1, 0)
60 View(loan)
```

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

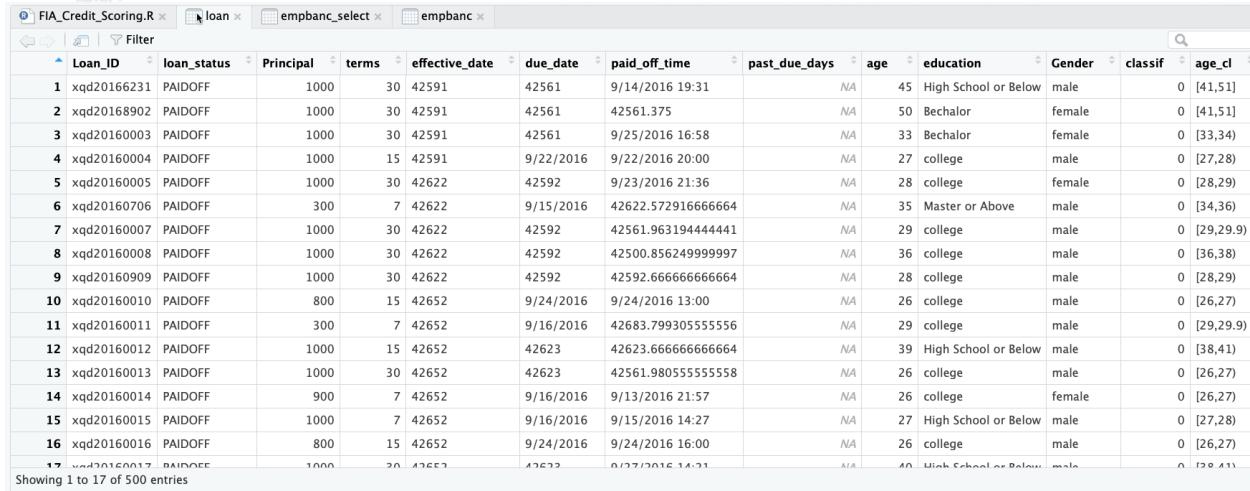
Loan_ID	loan_status	Principal	terms	effective_date	due_date	paid_off_time	past_due_days	age	education	Gender	classif
1 xqd20166231	PAIDOFF	1000	30	42591	42561	9/14/2016 19:31	NA	45	High School or Below	male	0
2 xqd20168902	PAIDOFF	1000	30	42591	42561	42561.375	NA	50	Bechelor	female	0
3 xqd20160003	PAIDOFF	1000	30	42591	42561	9/25/2016 16:58	NA	33	Bechelor	female	0
4 xqd20160004	PAIDOFF	1000	15	42591	9/22/2016	9/22/2016 20:00	NA	27	college	male	0
5 xqd20160005	PAIDOFF	1000	30	42622	42592	9/23/2016 21:36	NA	28	college	female	0
6 xqd20160706	PAIDOFF	300	7	42622	9/15/2016	42622.5729166666664	NA	35	Master or Above	male	0
7 xqd20160007	PAIDOFF	1000	30	42622	42592	42561.963194444441	NA	29	college	male	0
8 xqd20160008	PAIDOFF	1000	30	42622	42592	42500.856249999997	NA	36	college	male	0
9 xqd20160909	PAIDOFF	1000	30	42622	42592	42592.6666666666664	NA	28	college	male	0
10 xqd20160010	PAIDOFF	800	15	42652	9/24/2016	9/24/2016 13:00	NA	26	college	male	0
11 xqd20160011	PAIDOFF	300	7	42652	9/16/2016	42683.799305555556	NA	29	college	male	0
12 xqd20160012	PAIDOFF	1000	15	42652	42623	42623.6666666666664	NA	39	High School or Below	male	0
13 xqd20160013	PAIDOFF	1000	30	42652	42623	42561.98055555558	NA	26	college	male	0
14 xqd20160014	PAIDOFF	900	7	42652	9/16/2016	9/13/2016 21:57	NA	26	college	female	0
15 xqd20160015	PAIDOFF	1000	7	42652	9/16/2016	9/15/2016 14:27	NA	27	High School or Below	male	0
16 xqd20160016	PAIDOFF	800	15	42652	9/24/2016	9/24/2016 16:00	NA	26	college	male	0
17 xqd20160017	PAIDOFF	1000	30	42652	42623	9/27/2016 14:21	NA	40	High School or Below	male	0

Vamos categorizar a variável “age” em 15 diferentes *bins* (classes) e, depois tentar agrupar essas classes em um número menor de categorias.

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

```
62 #Categorização da variável "age" em 15 bins  
63 library(arules)  
64 loan$age_cl<-discretize(loan$age, "frequency", breaks = 15)  
65 View(loan)
```



The screenshot shows the RStudio interface with the following tabs: FIA_Credit_Scoring.R, loan, empbanc_select, and empbanc. The code in the FIA_Credit_Scoring.R script is as follows:

```
62 #Categorização da variável "age" em 15 bins  
63 library(arules)  
64 loan$age_cl<-discretize(loan$age, "frequency", breaks = 15)  
65 View(loan)
```

The 'View(loan)' command has been run, and the resulting data frame is displayed in the main pane. The columns shown are:

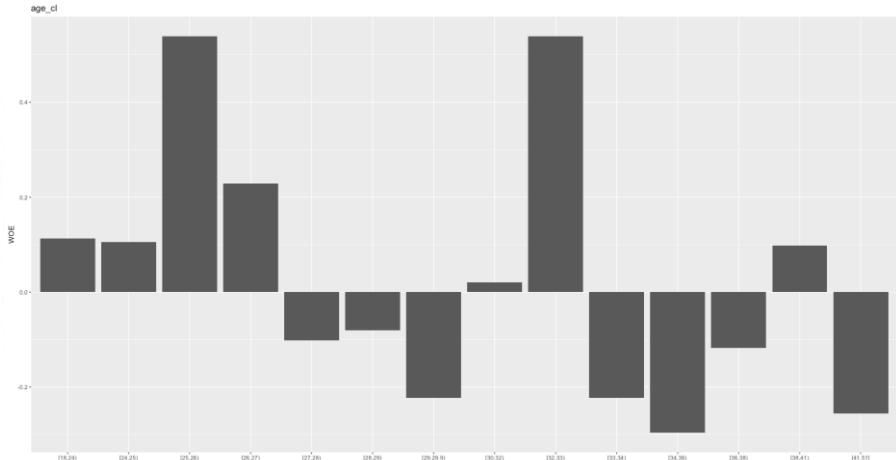
	Loan_ID	loan_status	Principal	terms	effective_date	due_date	paid_off_time	past_due_days	age	education	Gender	classif	age_cl
1	xqd20166231	PAIDOFF	1000	30	42591	42561	9/14/2016 19:31	NA	45	High School or Below	male	0	[41,51]
2	xqd20168902	PAIDOFF	1000	30	42591	42561	42561.375	NA	50	Bachelor	female	0	[41,51]
3	xqd20160003	PAIDOFF	1000	30	42591	42561	9/25/2016 16:58	NA	33	Bachelor	female	0	[33,34]
4	xqd20160004	PAIDOFF	1000	15	42591	9/22/2016	9/22/2016 20:00	NA	27	college	male	0	[27,28]
5	xqd20160005	PAIDOFF	1000	30	42622	42592	9/23/2016 21:36	NA	28	college	female	0	[28,29]
6	xqd20160706	PAIDOFF	300	7	42622	9/15/2016	42622.572916666664	NA	35	Master or Above	male	0	[34,36]
7	xqd20160007	PAIDOFF	1000	30	42622	42592	42561.963194444441	NA	29	college	male	0	[29,29.9]
8	xqd20160008	PAIDOFF	1000	30	42622	42592	42500.856249999997	NA	36	college	male	0	[36,38]
9	xqd20160909	PAIDOFF	1000	30	42622	42592	42592.666666666664	NA	28	college	male	0	[28,29]
10	xqd20160010	PAIDOFF	800	15	42652	9/24/2016	9/24/2016 13:00	NA	26	college	male	0	[26,27]
11	xqd20160011	PAIDOFF	300	7	42652	9/16/2016	42683.799305555556	NA	29	college	male	0	[29,29.9]
12	xqd20160012	PAIDOFF	1000	15	42652	42623	42623.666666666664	NA	39	High School or Below	male	0	[38,41]
13	xqd20160013	PAIDOFF	1000	30	42652	42623	42561.980555555558	NA	26	college	male	0	[26,27]
14	xqd20160014	PAIDOFF	900	7	42652	9/16/2016	9/13/2016 21:57	NA	26	college	female	0	[26,27]
15	xqd20160015	PAIDOFF	1000	7	42652	9/16/2016	9/15/2016 14:27	NA	27	High School or Below	male	0	[27,28]
16	xqd20160016	PAIDOFF	800	15	42652	9/24/2016	9/24/2016 16:00	NA	26	college	male	0	[26,27]
17	xqd20160017	PAIDOFF	1000	30	42653	42623	9/27/2016 14:31	NA	40	High School or Below	male	0	[30,41]

Showing 1 to 17 of 500 entries

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

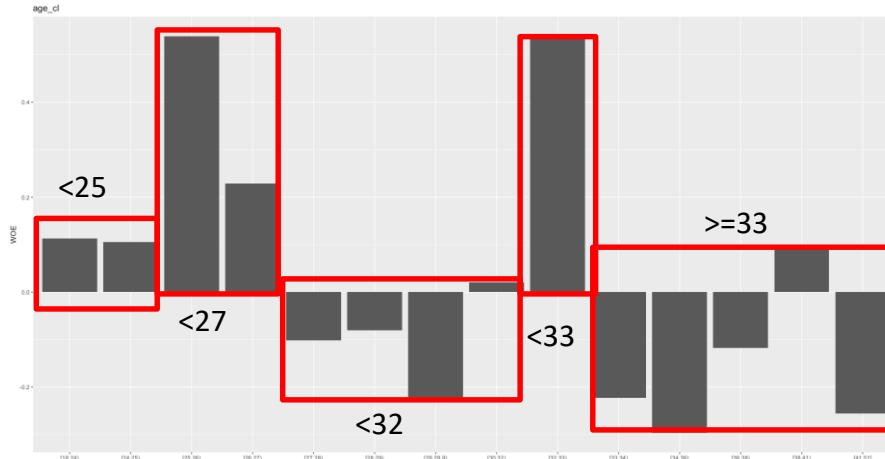
```
67 #Cálculo do Information Value e do WOE
68 library(Information)
69 IV <- create_infotables(data = loan, y = "classif", ncore = 2)
70 IV_valor<-data.frame(IV$Summary)
71 print(IV$Tables$age_cl, row.names=FALSE)
72 age_cl=data.frame(IV$Tables$age_cl)
73 View(age_cl)
74 plot_infotables(IV, "age_cl")
```



Agora, podemos agrupar a variável `age_cl` em um número menor de categorias, com níveis de WOE similares.

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

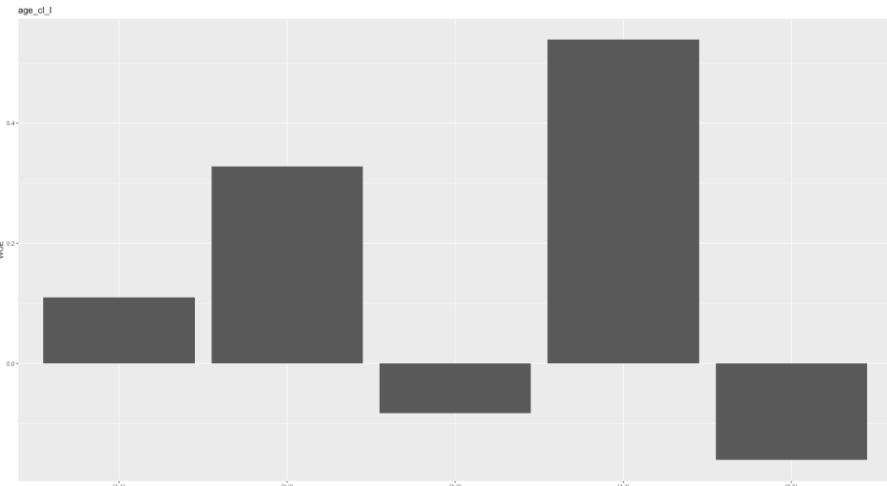


Observando o gráfico de WOE para as 14 faixas criadas, vemos que uma possibilidade é agruparmos em 5 novas categorias.

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

```
76 #Novo agrupamento de age_cl, agora em 5 grupos
77 loan$age_cl_I<-ifelse(loan$age<25, 1,
78           ifelse(loan$age<27, 2,
79           ifelse(loan$age<32, 3,
80           ifelse(loan$age<33, 4,
81           ifelse(loan$age>=33, 5, 0))))))
82 View(loan)
83 IV <- create_infotables(data = loan, y = "classif", ncore = 2)
84 IV$Summary
85 IV_valor<-data.frame(IV$Summary)
86 View(IV_valor)
87 print(IV$Tables$age_cl_I, row.names=FALSE)
88 age_cl_I=data.frame(IV$Tables$age_cl_I)
89 View(age_cl_I)
90 plot_infotables(IV, "age_cl_I")
```



Com base nos agrupamentos desenhados anteriormente, refizemos os agrupamentos, agora, com 5 categorias. Vamos avaliar se houve perda substancial de valor da informação com essa redução de quantidade de categorias.

Processo de Categorização de Variáveis – *Fine and Coarse Classing*

EXERCÍCIO 03

> IV\$Summary

	Variable	IV
6	due_date	1.1142406162
5	effective_date	1.0500492384
12	age_cl	0.0556118252
13	age_cl_I	0.0419181101
11	Gender	0.0400797927
10	education	0.0355088068
9	age	0.0334002888
8	past_due_days	0.0267521668
4	terms	0.0059566361
3	Principal	0.0006362407
1	Loan_ID	0.0000000000
2	loan_status	0.0000000000
7	paid_off_time	0.0000000000

O valor da informação da variável **age** pura é de 0,033. Após categorizá-la em 14 bins (*fine classing*), por meio da variável **age_cl**, seu IV foi para 0,056. Depois de agrupar algumas categorias, de acordo com o WOE, chegamos a 5 bins, na variável **age_cl_I** e seu IV caiu para 0,041. Apesar da queda, a quantidade de classes caiu de 14 para 5, facilitando tarefas de interpretação.

A seguir, vamos avaliar uma outra forma de agrupar categorias, menos manual e com bom potencial de resultados. É o que chamamos de ***optimal binning*** ou categorização ótima.

Processo de Categorização de Variáveis – *Optimal Binning*

EXERCÍCIO 04

O objetivo desse método de categorização é bastante claro: encontrar a quantidade ideal de classes, assim como a amplitude de cada classe, de forma a alavancar o valor da informação de determinada variável.

The screenshot shows a blog post from the Revolutions website. The header features the site's logo and navigation links for a webinar, main page, and survey. The post title is "R Package 'smbinning': Optimal Binning for Scoring Modeling" by Herman Jopia. The content discusses binning in scoring modeling, its benefits (like handling zero values and outliers), and its application in supervised and unsupervised discretization. It includes a table comparing equal length and frequency intervals. The sidebar contains links for information, a search bar, and social media links for Twitter and Bloglovin'. Categories listed include academia, advanced tips, AI, roundups, announcements, applications, beginner tips, bio data, business, current events, data science, developer tips, events, finance, government, and graphics.

Discretization Method	Open	Length	Number of Intervals	Target	Number of Records	Number of Bad	Number of Good	Number of Neutral	Min	Max
Equal length intervals	1	12 months	1	open <= 12 months	1200	400	400	400	1.7584	0.4051
1	120	120	10	open <= 120	1000	400	400	200	0.4051	0.4123
2	120	240	5	open <= 240	900	300	300	300	0.4051	0.4123
3	120	360	3.33	open <= 360	800	266.67	266.67	266.67	0.4051	0.4123
4	120	480	2.5	open <= 480	700	200	200	200	0.4051	0.4123
5	120	600	2	open <= 600	600	150	150	150	0.4051	0.4123
6	120	720	1.67	open <= 720	500	100	100	100	0.4051	0.4123
7	120	840	1.33	open <= 840	400	73.33	73.33	73.33	0.4051	0.4123
8	120	960	1	open <= 960	300	50	50	50	0.4051	0.4123
9	120	1080	0.83	open <= 1080	200	33.33	33.33	33.33	0.4051	0.4123
10	120	1200	0.67	open <= 1200	100	20	20	20	0.4051	0.4123
11	120	1320	0.5	open <= 1320	0	0	0	0	0.4051	0.4123

Fonte: <https://blog.revolutionanalytics.com/2015/03/r-package-smbinning-optimal-binning-for-scoring-modeling.html>

Processo de Categorização de Variáveis – *Optimal Binning*

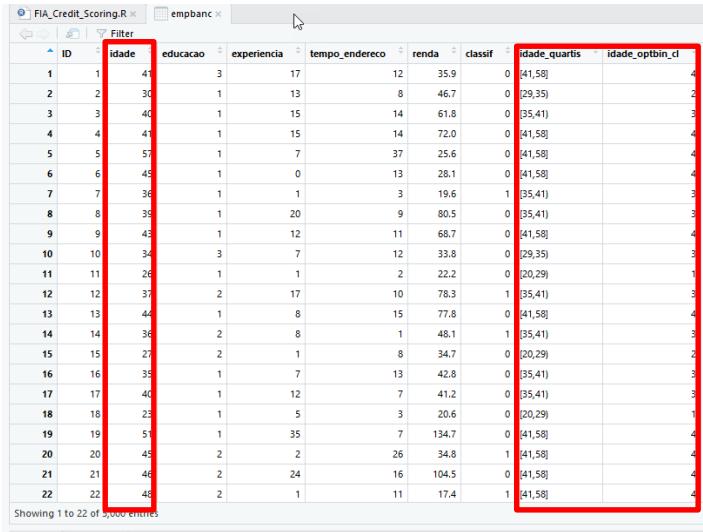
EXERCÍCIO 04

Vamos testar o conceito de *optimal binning* para a variável “**Idade**” da base de **Emprestimo Bancario.xlsx**, comparando com os resultados do valor da informação da variável pura e do agrupamento em quartis.

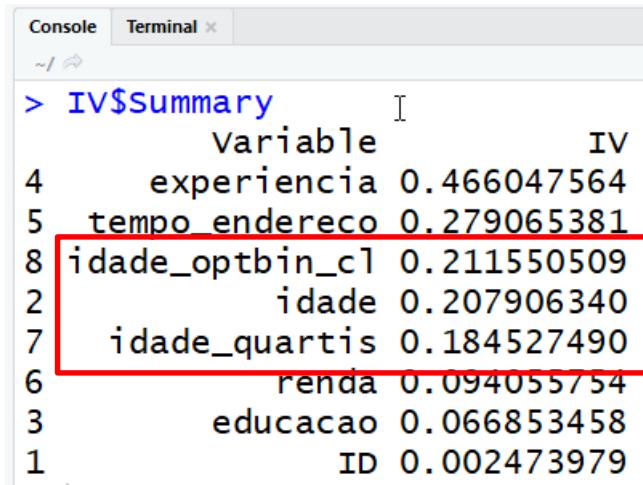
```
94 #Teste do optimal binning para a variavel Idade, da base Emprestimo Bancario
95 install.packages("smbinning")
96 library(smbinning)
97 empbanc<-as.data.frame(empbanc)
98 class(empbanc)
99 idade_optbin<-smbinning(df=empbanc, y="classif", x="idade", p=0.05)
100 idade_optbin
101 empbanc$idade_optbin_c1<-ifelse(empbanc$idade<=26,1,
102                                     ifelse(empbanc$idade<=31,2,
103                                         ifelse(empbanc$idade<=40,3,
104                                             ifelse(empbanc$idade>40,4,0))))
105
106 empbanc$idade_quartis<-discretize(empbanc$idade, "frequency", breaks = 4)
107 View(empbanc)
108 IV <- create_infotables(data = empbanc, y = "classif", ncore = 2)
109 IV$Summary
```

Processo de Categorização de Variáveis – *Optimal Binning*

EXERCÍCIO 04



ID	idade	educacao	experiencia	tempo_endereco	renda	classif	idade_quartis	idade_optbin_cl
1	1	41	3	17	12	35.9	0 [41,58]	4
2	2	30	1	13	8	46.7	0 [29,35)	2
3	3	40	1	15	14	61.8	0 [35,41)	3
4	4	41	1	15	14	72.0	0 [41,58]	4
5	5	57	1	7	37	25.6	0 [41,58]	4
6	6	45	1	0	13	28.1	0 [41,58]	4
7	7	36	1	1	3	19.6	1 [35,41)	3
8	8	39	1	20	9	80.5	0 [35,41)	3
9	9	43	1	12	11	68.7	0 [41,58]	4
10	10	34	3	7	12	33.8	0 [29,35)	3
11	11	26	1	1	2	22.2	0 [20,29)	1
12	12	37	2	17	10	78.3	1 [35,41)	3
13	13	44	1	8	15	77.8	0 [41,58]	4
14	14	36	2	8	1	48.1	1 [35,41)	3
15	15	27	2	1	8	34.7	0 [20,29)	2
16	16	35	1	7	13	42.8	0 [35,41)	3
17	17	40	1	12	7	41.2	0 [35,41)	3
18	18	23	1	5	3	20.6	0 [20,29)	1
19	19	51	1	35	7	134.7	0 [41,58]	4
20	20	45	2	2	26	34.8	1 [41,58]	4
21	21	46	2	24	16	104.5	0 [41,58]	4
22	22	48	2	1	11	17.4	1 [41,58]	4



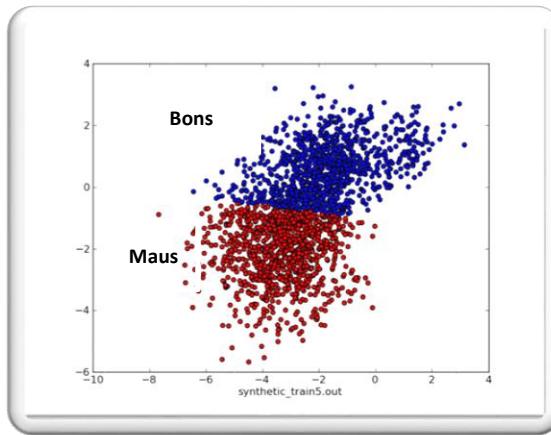
```
> IV$Summary
      Variable           IV
4     experiencia 0.466047564
5     tempo_endereco 0.279065381
8     idade_optbin_cl 0.211550509
2     idade 0.207906340
7     idade_quartis 0.184527490
6     renda 0.094055754
3     educacao 0.066853458
1     ID 0.002473979
```

O IV da variável idade, agrupada pelo método de optimal binning, foi o mais alto. Isso nem sempre vai ocorrer, vai depender bastante do nível de variabilidade da variável original e da quantidade distinta de combinações.

09

USO DA REGRESSÃO LOGÍSTICA PARA PREDIÇÃO DE RISCO DE CRÉDITO

Uso da Regressão Logística para Predição de Risco de Crédito



Pensem em um problema em que é necessário definir a **probabilidade de ocorrência de um evento**. Normalmente, esse evento é binário (como o caso ao lado, em que temos 2 grupos de “bons” e “maus”).

Nesse caso, quando temos uma variável resposta de interesse (**variável dependente ou variável alvo**), ela é **categórica e binária** e queremos calcular a probabilidade de ocorrência de cada um dos 2 grupos (“bons” e “maus”), a partir de variáveis explicativas (**independentes**), que podem ser tanto categóricas quanto quantitativas, podemos utilizar uma técnica chamada **Regressão Logística Binária**. Se a quantidade de categorias da variável resposta for superior a 2, temos o caso da **Regressão Logística Multinomial**.

Uso da Regressão Logística para Predição de Risco de Crédito

A regressão logística analisa dados distribuídos binomialmente da forma

$$Y_i \sim B(p_i, n_i),$$

em que **n** é a quantidade de tentativas e **p** é a probabilidade de sucesso. Essa probabilidade de sucesso, considerando a relação entre as variáveis dependentes e independentes, pode ser modelada da seguinte forma:

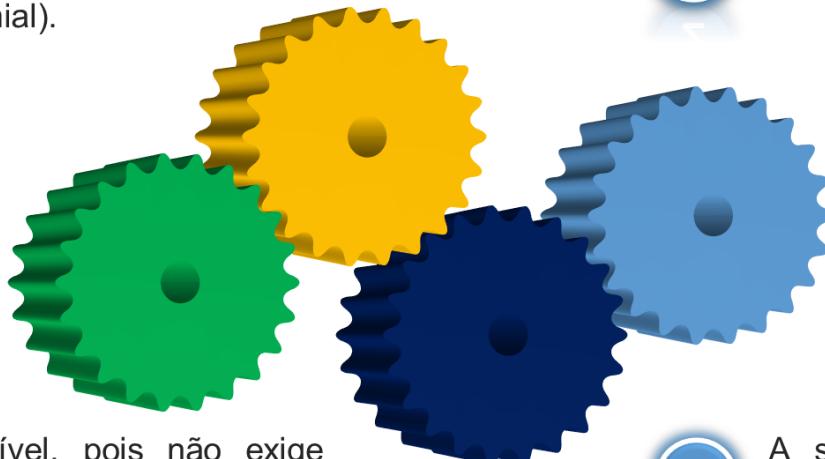
$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * X_1 + \dots + \beta_n * X_n)}}$$

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * X_1 + \dots + \beta_n * X_n$$

Uso da Regressão Logística para Predição de Risco de Crédito

1 Adequada para modelagem de problemas com resposta categórica (binária ou multinomial).

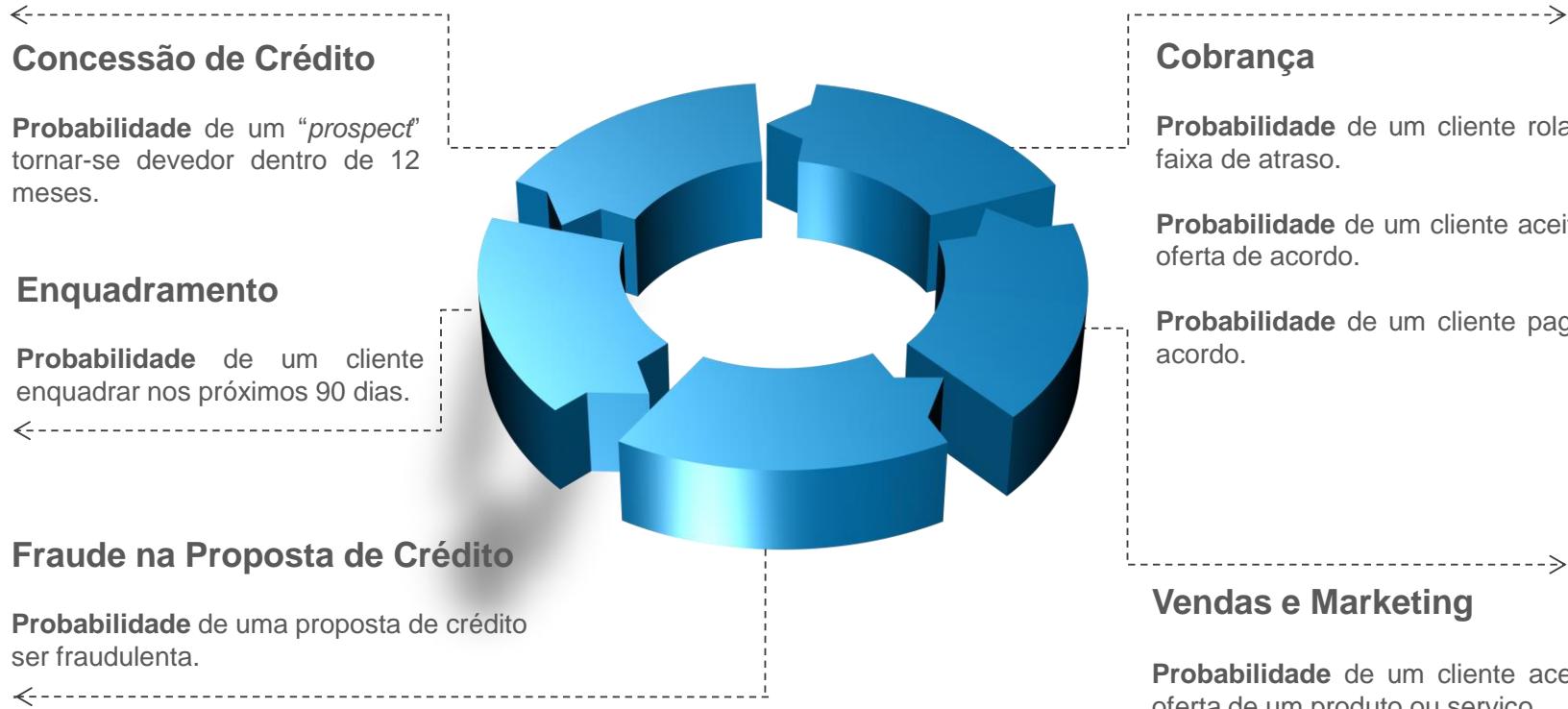
2 Trabalha com variáveis independentes categóricas e quantitativas.



3 Mais flexível, pois não exige normalidade das variáveis (como no caso da análise discriminante).

4 A saída do modelo logístico é uma probabilidade, que varia entre 0 e 1, o que torna a atividade de mensurar risco, mais fácil a partir dessa técnica.

Uso da Regressão Logística para Predição de Risco de Crédito



Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 05

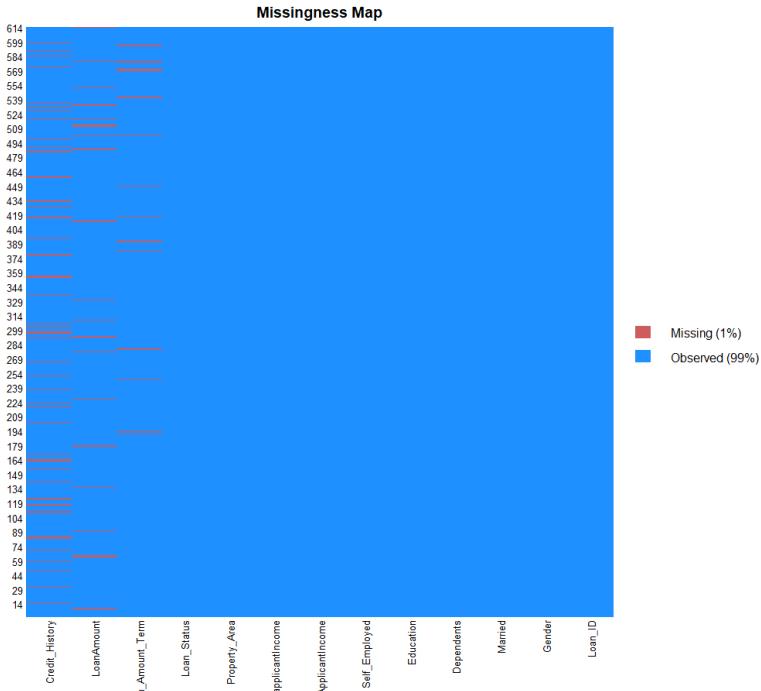
Vamos usar a base "Loan_Prediction_Data_train.csv" para exercitar conceitos associados à aplicação da Regressão Logística para avaliação de risco. Antes disso, porém, vamos checar o % de *missing values* na base de dados. Existe um pacote chamado "Amelia", que habilita uma função chamada "missmap", que mostra, de forma gráfica, a incidência dos valores faltantes na base.

```
111 #Importacao da base de dados de clientes do produto cartao de credito
112 loanpred<-read.csv("Loan_Prediction_Data_train.csv", sep = ",")
113 view(loanpred)
114 #checagem de missing values
115 install.packages("Amelia")
116 library(Amelia)
117 missmap(loanpred)
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	
1	LP001002	Male	No	0	Graduate	No	5849	0	NA	360	1	Urban	
2	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	
3	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	
4	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	
5	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	
6	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	
7	LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	
8	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 05



Os valores faltantes são mostrados em “vermelho”, enquanto que os válidos são apresentados em “azul”. De forma geral, o % de missing values é de 1%, mais concentrados nas variáveis “Credit History”, “LoanAmount” e “Loan_Amount_Term”.

Observando, por exemplo, a variável “Credit History”, o % de missing values é de 8,14%. Poderíamos, para essa variável, substituir os “NA” por “0”, usando o comando “`replace.value`” do pacote “`anchors`”.

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 05

```
119 #Substituição de missing values das variáveis Credit History, Loan Amount e Loan Amount Term por 0
120 install.packages("anchors")
121 library(anchors)
122 sum(is.na(loanpred$Credit_History))/nrow(loanpred) #% de missing values de Credit History
123 loanpred_I<- replace.value(loanpred,c("Credit_History", "LoanAmount", "Loan_Amount_Term"))
124 View(loanpred_I)
125 sum(is.na(loanpred_I$Credit_History))/nrow(loanpred) #% de missing values de Credit History
126 sum(is.na(loanpred_I$LoanAmount))/nrow(loanpred) #% de missing values de Loan Amount
127 sum(is.na(loanpred_I$Loan_Amount_Term))/nrow(loanpred) #% de missing values de Loan Amount Term
```

```
Console Terminal ×
~/
> sum(is.na(loanpred_I$Credit_History))/nrow(loanpred) #% de missing values de Credit History
[1] 0
> sum(is.na(loanpred_I$LoanAmount))/nrow(loanpred) #% de missing values de Loan Amount
[1] 0
> sum(is.na(loanpred_I$Loan_Amount_Term))/nrow(loanpred) #% de missing values de Loan Amount Term
[1] 0
```

A função “`replace.value`” basicamente substitui os “**NA**” das variável “`Credit_History`”, “`Loan Amount`” e “`Loan Amount Term`” por “**0**”.

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 05

Vamos agora calcular o valor da informação de cada variável da base "Loan_Prediction_Data_train.csv":

```
127 #Convertendo o target em numérico e calculando o IV das variáveis
128 loanpred_I$Loan_Status<-ifelse(loanpred_I$Loan_Status=="Y", 0, 1)
129 table(loanpred_I$Loan_Status)
130 library(Information)
131 IV<-create_infotables(data = loanpred_I, y = "Loan_Status", ncore = 2)
132 IV$Summary

> IV$Summary
      Variable          IV
11   Credit_History 0.831770359
12   Property_Area  0.096227947
9    LoanAmount     0.074651837
7    ApplicantIncome 0.039232101
3     Married       0.036281916
5     Education     0.033043685
4     Dependents    0.028513012
8 CoapplicantIncome 0.020869402
10  Loan_Amount_Term 0.007758638
2     Gender        0.004076742
6     Self_Employed  0.001234494
1     Loan_ID       0.000000000
```

A maioria das variáveis tem baixo poder preditivo para explicar o evento “default”, exceto a variável “Credit History”, que tem um IV bastante elevado e merece uma atenção especial.

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 05

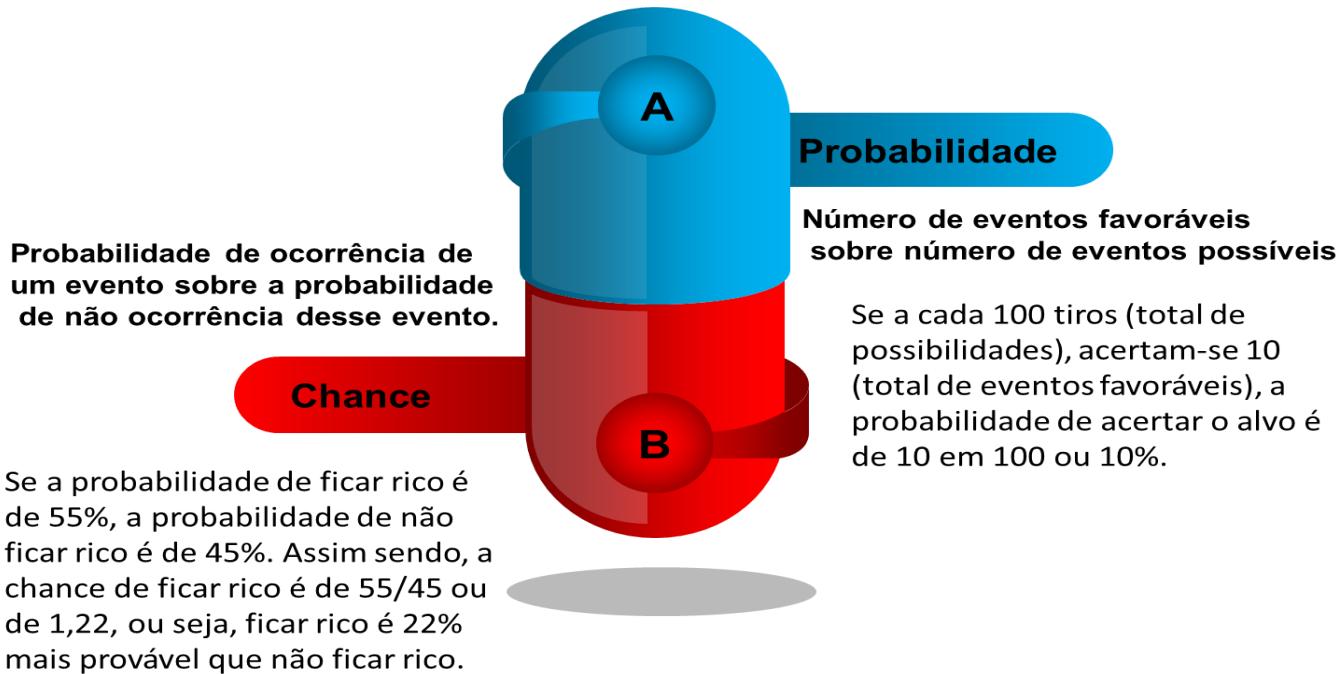
```
134 #Analisando a variável "Credit History"
135 library(descr)
136 CrossTable(loanpred_I$Credit_History, loanpred_I$Loan_Status, prop.r = FALSE,
137             prop.t = FALSE, prop.chisq = FALSE)
138
```

		loanpred_I\$Loan_Status		Total
loanpred_I\$Credit_History		0	1	
0	0	44	95	139
	1	0.104	0.495	
1	0	378	97	475
	1	0.896	0.505	
Total		422	192	614
		0.687	0.313	

Os clientes que não possuem histórico de crédito tem um nível de concentração de maus muito maior que de bons. Do lado oposto, aqueles que tem histórico de crédito tem uma tendência muto maior a serem bons clientes, sob a ótica de risco.

Uso da Regressão Logística para Predição de Risco de Crédito

Diferença entre Chance e Probabilidade



Uso da Regressão Logística para Predição de Risco de Crédito

O conceito de Odds Ratio

Essa medida reflete a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo.

ODDS RATIO

Por exemplo, pense no evento “inadimplência”. Pense agora em 2 grupos distintos: pessoas abaixo de 25 anos (“jovens”) e pessoas acima de 60 anos (“melhor idade”). Se pensarmos que a proporção histórica de pessoas “jovens” inadimplentes é de 15% (ou seja, a cada 100, 15 foram expostas à inadimplência e 85 não) e a proporção de pessoas da “melhor idade” inadimplentes é de 5% (a cada 100, 5 foram expostas à inadimplência, enquanto 95 não), podemos raciocinar o seguinte:

$$\text{ODDS (Jovens/Melhor Idade)} = \frac{\frac{\text{Jovens Expostos}}{\text{Jovens Não expostos}}}{\frac{\text{Melhor Idade Expostos}}{\text{Melhor Idade Não expostos}}} = \frac{\frac{0,15}{0,85}}{\frac{0,05}{0,95}} = \frac{0,1764}{0,0526} = 3,354$$

A **Odds Ratio** dos jovens com relação à melhor idade é de 3,354, ou seja, os jovens estão 3,354 mais expostos à inadimplência que as pessoas da melhor idade. Esse racional é o ponto central para a interpretação de variáveis em um modelo de regressão logística, sobretudo variáveis categóricas.

Uso da Regressão Logística para Predição de Risco de Crédito

O conceito de Odds Ratio

Agora, vejamos como esse conceito de “*odds ratio*” se aplica no caso da variável “Credit History”, da base “**Loan Prediction Data train.csv**”:

		loanpred_I\$Loan_Status		Total
loanpred_I\$Credit_History		0	1	
0	44	95	139	
	0.104	0.495		
1	378	97	475	
	0.896	0.505		
Total	422	192	614	
	0.687	0.313		

Dentre os clientes sem histórico de crédito, temos 44 que não foram expostos ao default e 95 que foram, totalizando 139 clientes. Dentre aqueles com histórico de crédito, 378 não foram expostos ao default e 97 foram, totalizando 475 clientes. Pensando em termos de *odds ratio*, temos que:

$$Odds(\text{sem histórico} / \text{com histórico}) = \frac{\frac{\text{Sem histórico expostos}}{\text{Sem histórico não expostos}}}{\frac{\text{Com histórico expostos}}{\text{Com histórico não expostos}}} =$$

$$\frac{\frac{95/139}{44/139}}{\frac{97/475}{378/475}} = \frac{2,159}{0,257} = 8,414$$

Com isso, baseado nesses dados, temos que o grupo de pessoas “sem histórico de crédito” apresentou uma chance 8 vezes maior de não pagar (ou ser exposto ao default) do que pessoas “com histórico de crédito”. Por isso, essa variável mostrou-se tão relevante.

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Vamos agora gerar um modelo de regressão logística, baseado nas variáveis existentes na base "Loan_Prediction_Data_train.csv":

```
139 #Modelo de regressão logística com as variáveis da base Loan Prediction, exceto Loand_ID  
140 modelo<-glm(Loan_Status ~ . , family = binomial (link='logit'), data = loanpred_I[,2:13])  
141 summary(modelo)
```

	Coefficients:	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.202e+01	5.056e+02	-0.024	0.98103	
GenderFemale	-4.430e-01	6.800e-01	-0.651	0.51514	
GenderMale	-3.614e-01	6.384e-01	-0.566	0.57131	
MarriedNo	1.381e+01	5.056e+02	0.027	0.97821	
MarriedYes	1.317e+01	5.056e+02	0.026	0.97923	
Dependents0	-6.252e-01	6.891e-01	-0.907	0.36427	
Dependents1	-2.424e-01	7.150e-01	-0.339	0.73458	
Dependents2	-7.039e-01	7.247e-01	-0.971	0.33142	
Dependents3+	-3.570e-01	7.584e-01	-0.471	0.63780	
EducationNot Graduate	3.335e-01	2.400e-01	1.390	0.16460	
Self_EmployedNo	2.568e-01	4.566e-01	0.562	0.57389	
Self_EmployedYes	2.138e-01	5.206e-01	0.411	0.68128	
ApplicantIncome	9.875e-06	2.104e-05	0.469	0.63884	
CoapplicantIncome	4.660e-05	3.743e-05	1.245	0.21317	
LoanAmount	-2.943e-04	1.440e-03	-0.204	0.83806	
Loan_Amount_Term	8.263e-04	1.242e-03	0.665	0.50588	
Credit_History	-2.152e+00	2.243e-01	-9.592	< 2e-16 ***	
Property_AreaSemiurban	-7.979e-01	2.478e-01	-3.220	0.00128 **	
Property_AreaUrban	-2.718e-01	2.452e-01	-1.108	0.26774	

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

Percebem que somente a variável “Credit_History” e um nível da variável “Property_Area”, como já era esperado, mostraram-se altamente significativas. As demais tiveram p-valor>0,15. Além de olhar pelo p-valor, a estatística “z value” também mostra a força da variável (em modulo).

Notem também que as variáveis categóricas tiveram seus respectivos fatores sendo pontuados no modelo, sempre com relação a uma categoria de referência. Mas se é assim, por que as variáveis “Married”, “Gender”, “Self-Employed”, só para citar algumas, apareceram com todos os níveis ? Vamos checar.

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

```
--  
143 #Estatísticas descritivas das variáveis Married, Self-Employed and Gender  
144 summary(loanpred$Gender)  
145 summary(loanpred$Self_Employed)  
146 summary(loanpred$Married)
```

Console Terminal ×

```
~/  
> summary(loanpred$Gender)  
   Female   Male  
      13    112    489  
> summary(loanpred$Self_Employed)  
  No Yes  
 32 500 82  
> summary(loanpred$Married)  
  No Yes  
 3 213 398  
>
```

Observando as frequências ao lado, vemos que todas as 3 variáveis analisadas tem valores faltantes. Podemos facilmente corrigir isso, seja excluindo as linhas com dados inválidos ou mesmo substituindo o dado faltante por algum valor. Nesse caso, vamos substituir pelo fator com maior representatividade:

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

```
148 #Substituindo os missing values das variáveis Married, Self-Employed and Gender pelo fator mais frequente
149 summary(loanpred_I$Married)
150 levels(loanpred_I$Married)[levels(loanpred_I$Married) == ""] <- "Yes"
151 summary(loanpred_I$Married)
152
153 summary(loanpred_I$Self_Employed)
154 levels(loanpred_I$Self_Employed)[levels(loanpred_I$Self_Employed) == ""] <- "No"
155 summary(loanpred_I$Self_Employed)
156
157 summary(loanpred_I$Gender)
158 levels(loanpred_I$Gender)[levels(loanpred_I$Gender) == ""] <- "Male"
159 summary(loanpred_I$Gender)
```

```
Console Terminal ×
> summary(loanpred_I$Married)
  No Yes
 3 213 398
> levels(loanpred_I$Married)[levels(loanpred_I$Married) == ""] <- "Yes"
> summary(loanpred_I$Married)
Yes No
401 213
`-
```

```
Console Terminal ×
> summary(loanpred_I$Self_Employed)
  No Yes
32 500 82
> levels(loanpred_I$Self_Employed)[levels(loanpred_I$Self_Employed) == ""] <- "No"
> summary(loanpred_I$Self_Employed)
No Yes
532 82
```

```
Console Terminal ×
~/
> summary(loanpred_I$Gender)
   Female   Male
      13    112    489
> levels(loanpred_I$Gender)[levels(loanpred_I$Gender) == ""] <- "Male"
> summary(loanpred_I$Gender)
Male Female
502    112
```

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Vamos agora gerar novamente o modelo de regressão logística, após a substituição dos valores faltantes das variáveis **Gender**, **Married** e **Self Employed**:

```
161 #Modelo I de regressão logística com as variáveis da base Loan Prediction, exceto Loand_ID  
162 modelo<-glm(Loan_Status ~ . , family = binomial (link='logit'), data = Loanpred_I[,2:13])  
163 summary(modelo)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)		
(Intercept)	8.182e-01	8.101e-01	1.010	0.31252		
GenderFemale	-9.769e-02	2.789e-01	-0.350	0.72617		
MarriedNo	6.501e-01	2.375e-01	2.738	0.00618 **		
Dependents0	-4.064e-01	6.403e-01	-0.635	0.52558		
Dependents1	-3.404e-02	6.671e-01	-0.051	0.95931		
Dependents2	-4.849e-01	6.769e-01	-0.716	0.47377		
Dependents3+	-1.202e-01	7.133e-01	-0.169	0.86614		
EducationNot Graduate	3.365e-01	2.392e-01	1.407	0.15943		
Self_EmployedYes	-7.617e-03	2.943e-01	-0.026	0.97935		
ApplicantIncome	1.007e-05	2.093e-05	0.481	0.63053		
CoapplicantIncome	4.677e-05	3.789e-05	1.234	0.21705		
LoanAmount	-2.576e-04	1.439e-03	-0.179	0.85792		
Loan_Amount_Term	8.212e-04	1.241e-03	0.662	0.50799		
Credit_History	-2.166e+00	2.241e-01	-9.666	< 2e-16 ***		
Property_AreaSemiurban	-7.950e-01	2.472e-01	-3.216	0.00130 **		
Property_AreaUrban	-2.728e-01	2.452e-01	-1.112	0.26595		

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

Vamos observar o efeito da variável “**Gender**”. Percebemos que só aparece o peso do fator “**Female**”. O fator “**Male**” não apareceu, por que ele entrou como categoria de referência. Mas, e se houvesse interesse em ter o fator “**Female**” como referência? O que aconteceria com o peso da variável “**Gender**”?

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Para alterar a categoria de referência de uma variável categórica no R, podemos usar o comando “`relevel`”. Depois de colocar o fator “Female” como referência e atualizar o modelo logístico, comparando-o com o anterior, temos que:

```
165 #Redefinindo categoria de referência da variável Gender e atualizando o modelo
166 loanpred_I$Gender <- relevel(loanpred_I$Gender, ref = "Female")
167 modelo<-glm(Loan_Status ~ . , family = binomial (link='logit'), data = loanpred_I[,2:13])
168 summary(modelo)
```

	Coefficients:	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	8.182e-01	8.101e-01	1.010	0.31252	
GenderFemale	-9.769e-02	2.789e-01	-0.350	0.72617	
MarriedNo	6.501e-01	2.375e-01	2.738	0.00618 **	
Dependents0	-4.064e-01	6.403e-01	-0.635	0.52558	
Dependents1	-3.404e-02	6.671e-01	-0.051	0.95931	
Dependents2	-4.849e-01	6.769e-01	-0.716	0.47377	
Dependents3+	-1.202e-01	7.133e-01	-0.169	0.86614	
EducationNot Graduate	3.365e-01	2.392e-01	1.407	0.15943	
Self_EmployedYes	-7.617e-03	2.943e-01	-0.026	0.97935	
ApplicantIncome	1.007e-05	2.093e-05	0.481	0.63053	
CoapplicantIncome	4.677e-05	3.789e-05	1.234	0.21705	
LoanAmount	-2.576e-04	1.439e-03	-0.179	0.85792	
Loan_Amount_Term	8.212e-04	1.241e-03	0.662	0.50799	
Credit_History	-2.166e+00	2.241e-01	-9.666	<2e-16 ***	
Property_AreaSemiurban	-7.950e-01	2.472e-01	-3.216	0.00130 **	
Property_AreaUrban	-2.728e-01	2.452e-01	-1.112	0.26595	

Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

	Coefficients:	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	7.205e-01	8.474e-01	0.850	0.39516	
GenderMale	9.769e-02	2.789e-01	0.350	0.72617	
MarriedNo	6.501e-01	2.375e-01	2.738	0.00618 **	
Dependents0	-4.064e-01	6.403e-01	-0.635	0.52558	
Dependents1	-3.404e-02	6.671e-01	-0.051	0.95931	
Dependents2	-4.849e-01	6.769e-01	-0.716	0.47377	
Dependents3+	-1.202e-01	7.133e-01	-0.169	0.86614	
EducationNot Graduate	3.365e-01	2.392e-01	1.407	0.15943	
Self_EmployedYes	-7.617e-03	2.943e-01	-0.026	0.97935	
ApplicantIncome	1.007e-05	2.093e-05	0.481	0.63053	
CoapplicantIncome	4.677e-05	3.789e-05	1.234	0.21705	
LoanAmount	-2.576e-04	1.439e-03	-0.179	0.85792	
Loan_Amount_Term	8.212e-04	1.241e-03	0.662	0.50799	
Credit_History	-2.166e+00	2.241e-01	-9.666	<2e-16 ***	
Property_AreaSemiurban	-7.950e-01	2.472e-01	-3.216	0.00130 **	
Property_AreaUrban	-2.728e-01	2.452e-01	-1.112	0.26595	

Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Percebam que nada mudou nos pesos das variáveis, exceto o peso da variável “Gender”. Portanto, a escolha da categoria de referência não altera o contexto do modelo, apenas da variável de maneira isolada.

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

E se quisermos tentar reproduzir o cálculo do score do modelo logístico, fazendo o cálculo “na mão” ? Vamos usar o mesmo exemplo anterior, do modelo gerado para a base "Loan_Prediction_Data_train.csv":

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	7.205e-01	8.474e-01	0.850	0.39516
GenderMale	9.769e-02	2.789e-01	0.350	0.72617
MarriedNo	6.501e-01	2.375e-01	2.738	0.00618 **
Dependents0	-4.064e-01	6.403e-01	-0.635	0.52558
Dependents1	-3.404e-02	6.671e-01	-0.051	0.95931
Dependents2	-4.849e-01	6.769e-01	-0.716	0.47377
Dependents3+	-1.202e-01	7.133e-01	-0.169	0.86614
EducationNot Graduate	3.365e-01	2.392e-01	1.407	0.15943
Self_EmployedYes	-7.617e-03	2.943e-01	-0.026	0.97935
ApplicantIncome	1.007e-05	2.093e-05	0.481	0.63053
CoapplicantIncome	4.677e-05	3.789e-05	1.234	0.21705
LoanAmount	-2.576e-04	1.439e-03	-0.179	0.85792
Loan_Amount_Term	8.212e-04	1.241e-03	0.662	0.50799
Credit_History	-2.166e+00	2.241e-01	-9.666	< 2e-16 ***
Property_AreaSemiurban	-7.950e-01	2.472e-01	-3.216	0.00130 **
Property_AreaUrban	-2.728e-01	2.452e-01	-1.112	0.26595

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

$$p_{default} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_n * x_n)}} = \\ \frac{1}{1 + e^{-(7.205 * 10^{-1} + 9.769 * 10^{-2} * Male + \dots + (-2.728 * 10^{-1} * Urban))}}$$

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Imagine um cenário específico, vindo da tabela "Loan_Prediction_Data_train.csv":

▲	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
1	LP001002	Male	No	0	Graduate	No	5849	0	0	360	1	Urban

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)								
(Intercept)	7.205e-01	8.474e-01	0.850	0.39516								
GenderMale	9.769e-02	2.789e-01	0.350	0.72617								
MarriedNo	6.501e-01	2.375e-01	2.738	0.00618 **								
Dependents0	-4.064e-01	6.403e-01	-0.635	0.52558								
Dependents1	-3.404e-02	6.671e-01	-0.051	0.95931								
Dependents2	-4.849e-01	6.769e-01	-0.716	0.47377								
Dependents3+	-1.202e-01	7.133e-01	-0.169	0.86614								
EducationNot Graduate	3.365e-01	2.392e-01	1.407	0.15943								
Self_EmployedYes	-7.617e-03	2.943e-01	-0.026	0.97935								
ApplicantIncome	1.007e-05	2.093e-05	0.481	0.63053								
CoapplicantIncome	4.677e-05	3.789e-05	1.234	0.21705								
LoanAmount	-2.576e-04	1.439e-03	-0.179	0.85792								
Loan_Amount_Term	8.212e-04	1.241e-03	0.662	0.50799								
Credit_History	-2.166e+00	2.241e-01	-9.666	< 2e-16 ***								
Property_AreaSemiurban	-7.950e-01	2.472e-01	-3.216	0.00130 **								
Property_AreaUrban	-2.728e-01	2.452e-01	-1.112	0.26595								

Signif. codes:	0 ***	0.001 **	0.05 *	0.1 .	1							

$$\begin{aligned} &= 1 / (1 + \text{EXP}(-1 * (7.205 * 0.1 \\ &\quad + 9.769 * 0.01 * 1 \\ &\quad + 6.501 * 0.1 * 1 \\ &\quad - 4.064 * 0.1 * 1 \\ &\quad + 3.365 * 0.1 * 0 \\ &\quad - 7.617 * 0.001 * 0 \\ &\quad + 1.007 * 0.00001 * 5849 \\ &\quad + 4.677 * 0.00001 * 0 \\ &\quad - 2.576 * 0.0001 * 0 \\ &\quad + 8.212 * 0.0001 * 360 \\ &\quad - 2.166 * 1 * 1 \\ &\quad - 2.728 * 0.1 * 1))) = 0,2646 \end{aligned}$$

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Agora, gerando de forma automática as probabilidades na tabela "Loan_Prediction_Data_train.csv":

```
178 modelo_pred = predict(modelo, loanpred_I, type="response")
179 loanpred_I<-data.frame(loanpred_I, modelo_pred)
180 View(loanpred_I)
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CooapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status	modelo_pred
1	LP001002	Male	No	0	Graduate	No	5849	0	0	360	1	Urban	0	0.26461534

$$\begin{aligned} &= 1 / (1 + \text{EXP}(-1 * (7,205 * 0,1 \\ &\quad + 9,769 * 0,01 * 1 \\ &\quad + 6,501 * 0,1 * 1 \\ &\quad - 4,064 * 0,1 * 1 \\ &\quad + 3,365 * 0,1 * 0 \\ &\quad - 7,617 * 0,001 * 0 \\ &\quad + 1,007 * 0,00001 * 5849 \\ &\quad + 4,677 * 0,00001 * 0 \\ &\quad - 2,576 * 0,00001 * 0 \\ &\quad + 8,212 * 0,00001 * 360 \\ &\quad - 2,166 * 1 * 1 \\ &\quad - 2,728 * 0,1 * 1))) = 0,2646 \end{aligned}$$

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Vamos também avaliar potenciais sinais de multicolinearidade, usando tanto a matriz de correlação, quanto o VIF, fator de inflação da variância:

```
175 #Avaliando sinais de multicolinearidade
176 install.packages("sjpplot")
177 library(sjpplot)
178 loanpred_I[,c(2,3,4,5,6,12)] <- lapply(loanpred_I[,c(2,3,4,5,6,12)],as.numeric)
179 View(loanpred_I)
180 str(loanpred_I)
181 sjp.corr(loanpred_I[,2:13])
182 sjt.corr(loanpred_I[,2:13])
```

	Dependents	LoanAmount	CoapplicantIncome	ApplicantIncome	Education	Credit_History	Self_Employed	Loan_Status	Property_Area	Loan_Amount_Term	Married	Gender
Gender	-0.169***	-0.096*	-0.083*	-0.059	-0.045	-0.027	0.001	0.018	0.026	0.050	0.365***	
Married	-0.321***	-0.136***	-0.076	-0.052	-0.012	0.018	-0.004	0.091*	-0.004	0.077		
Loan_Amount_Term	-0.034	0.059	-0.050	-0.016	-0.109**	0.050	-0.029	-0.007	-0.082*			
Property_Area	-0.009	-0.057	0.011	-0.009	-0.065	-0.019	-0.031	-0.032				
Loan_Status	-0.014	0.011	0.059	0.005	0.086*	-0.433***	0.004					
Self_Employed	0.055	0.109**	-0.016	0.127**	-0.010	-0.005						
Credit_History	-0.011	-0.033	-0.059	0.007	-0.082*							
Education	0.051	-0.173***	-0.062	-0.141***								
ApplicantIncome	0.116**	0.538***	-0.117**									
CoapplicantIncome	0.034	0.190***										
LoanAmount	0.160***											
Dependents												

Uso da Regressão Logística para Predição de Risco de Crédito

EXERCÍCIO 06

Vamos também avaliar potenciais sinais de multicolinearidade, usando tanto a matriz de correlação, quanto o VIF, fator de inflação da variância:

```
184 #Cálculo do VIF  
185 install.packages("HH")  
186 library(HH)  
187 vif(modelo)
```

```
> vif(modelo)
      Gender           Married        Dependents       Education
      1.170795        1.269178        1.145637        1.074132
Self_Employed   ApplicantIncome CoapplicantIncome LoanAmount
      1.023005        1.546222        1.136309        1.626296
Loan_Amount_Term Credit_History Property_Area
      1.042371        1.016698        1.019613
```

Como o VIF de todas as variáveis é inferior a 5, nenhum sinal aparente de multicolinearidade foi detectado.

10

USO DE MÉTRICAS DE AVALIAÇÃO DE MODELOS PREDITIVOS

Uso de Métricas de Avaliação de Modelos Preditivos



A avaliação de modelos tem como finalidade testar padrões mínimos de qualidade a partir dos algoritmos utilizados, bem como criar mecanismos de acompanhamento desses modelos (após a implementação).

Há vários indicadores (à esquerda) que podem ser utilizados para avaliar e comparar resultado de modelos estatísticos.

No caso de modelos de regressão logística, alguns desses são especialmente úteis e tem maior aplicação:

- (a) K-S (Kolmogorov-Smirnov)
- (b) Matriz de confusão
- (c) PSI (Índice de estabilidade da população)
- (d) Lift
- (e) Gains Chart
- (f) Curva ROC

Uso de Métricas de Avaliação de Modelos Preditivos



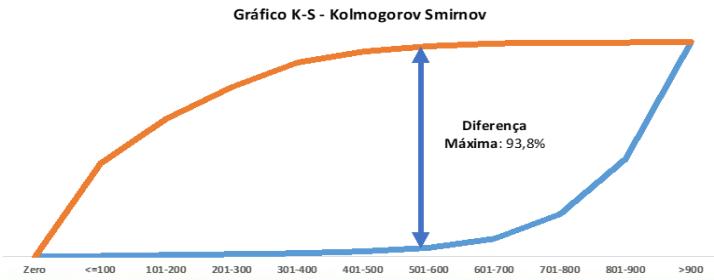
Vantagem: Indicador já consagrado e implementado na maioria dos softwares estatísticos.

Ponto de atenção: K-S pode mostrar-se alto, mas concentrado a uma única faixa de score.

Indicador K-S (Kolmogorov-Smirnov)

Referência aos matemáticos russos Andrey Kolmogorov e Vladimir Ivanovich Smirnov, inventores do teste. Bastante utilizado na indústria de crédito e cobrança. **Reflete a máxima separação (em termos absolutos) entre as curvas acumuladas de 2 grupos distintos quaisquer ("bons" e "maus", "pagou" e "não pagou", etc.).** Esse indicador varia entre 0% (nenhuma separação) a 100% (separação completa).

FAIXA DE SCORE	BONS			MAUS			DIF
	Bons	% Bons	% Acum Bons	Maus	% Maus	% Acum Maus	
<=100	1.254	0,3%	0,3%	124.121	43,5%	43,5%	43,2%
101-200	1.845	0,4%	0,6%	59.451	20,8%	64,3%	63,7%
201-300	2.412	0,5%	1,1%	42.112	14,8%	79,1%	77,9%
301-400	2.985	0,6%	1,8%	32.154	11,3%	90,3%	88,6%
401-500	3.895	0,8%	2,6%	14.211	5,0%	95,3%	92,8%
501-600	8.150	1,7%	4,2%	7.824	2,7%	98,1%	93,8%
601-700	19.787	4,1%	8,3%	3.211	1,1%	99,2%	90,9%
701-800	55.412	11,4%	19,8%	1.254	0,4%	99,6%	79,9%
801-900	124.588	25,7%	45,5%	655	0,2%	99,9%	54,4%
>900	264.111	54,5%	100,0%	412	0,1%	100,0%	0,0%
Total	484.439	100,0%	-	285.405	100,0%	-	93,8%



Uso de Métricas de Avaliação de Modelos Preditivos



		Valor Verdadeiro (confirmado por análise)	
		positivos	negativos
Valor Previsto (predito pelo teste)	positivos	VP Verdadeiro Positivo	FP Falso Positivo
	negativos	FN Falso Negativo	VN Verdadeiro Negativo

1. **Acurácia:** $(VP+VN)/(P+N)$. Proporção de valores corretos, sem considerar o que é negativo ou positivo.
2. **Sensibilidade:** $(VP)/(VP+FN)$. Proporção de verdadeiros positivos.
3. **Especificidade:** $(VN)/(FP+VN)$. Proporção de verdadeiros negativos.
4. **Eficiência:** $(SENS+ESPECIF)/2$. Média aritmética entre especificidade e sensibilidade, que, normalmente, caminham em direções opostas.
5. **Precisão:** $(VP)/(VP+FP)$. Proporção de acerto do modelo de predição.

Observação: É fundamental prestar atenção para o desbalanceamento entre as classes. A classe com menor proporção tende a apresentar piores taxas de classificação. Em casos como esse, avaliar metodologias para modelagem de eventos raros.

Uso de Métricas de Avaliação de Modelos Preditivos



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

Indicador do nível de mudança da distribuição de score, comparando a população de origem (utilizada para desenvolvimento do modelo) e a população atual (que reflete características atuais)

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

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

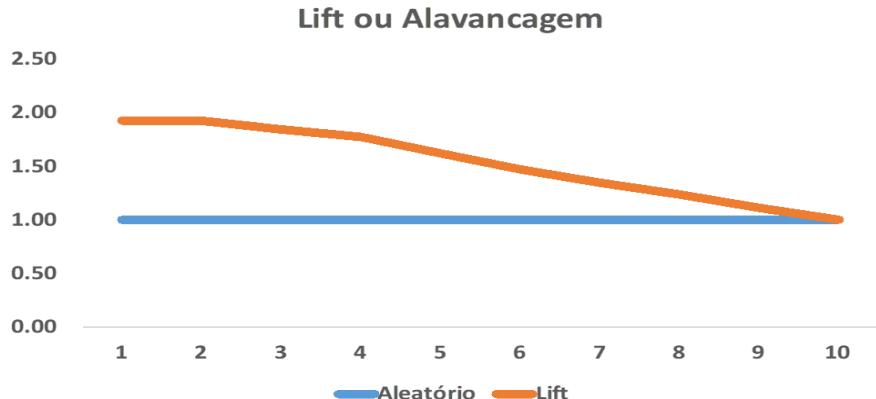
Identificada alguma mudança relevante, o caminho é avaliar em que variáveis do modelo essa mudança foi mais acentuada, de modo a direcionar focos de atuação para atualização da fórmula. O artigo colocado como “fonte” traz um processo de cálculo do PSI, utilizando SAS.

Uso de Métricas de Avaliação de Modelos Preditivos



A **alavancagem** (ou “*lift*”) representa quantas vez mais um modelo é capaz de detectar um evento, quando comparado a um modelo aleatório. Se pensarmos em um modelo aleatório e selecionarmos randomicamente 10% da base, é muitíssimo provável que capturemos 10% do nosso evento de interesse (ex.: os bons pagadores). Imaginem que desenvolvemos um modelo para capturar os bons, selecionamos os 10% mais propensos a pagar e identificamos 19% de bons pagadores. Para esse exemplo, o “*lift*” ou alavancagem do modelo foi de 19/10 ou de quase **2 vezes**. Essa medida é geralmente analisada por faixa de score e é bastante relevante para diferenciar as estratégias de atuação.

Grupo	Bons		Maus		Total	Total	Modelo Aleatório	Modelo Estatístico	Lift
	#	%	#	%					
1	48,541	19%	1,459	1%	50,000	10%	10%	19%	1.92
2	48,769	19%	1,231	0%	50,000	10%	20%	38%	1.92
3	42,412	17%	7,588	3%	50,000	10%	30%	55%	1.84
4	38,754	15%	11,246	5%	50,000	10%	40%	71%	1.76
5	25,641	10%	24,359	10%	50,000	10%	50%	81%	1.61
6	18,954	7%	31,046	13%	50,000	10%	60%	88%	1.47
7	15,411	6%	34,589	14%	50,000	10%	70%	94%	1.35
8	8,995	4%	41,005	17%	50,000	10%	80%	98%	1.22
9	4,214	2%	45,786	19%	50,000	10%	90%	100%	1.11
10	1,254	0%	48,746	20%	50,000	10%	100%	100%	1.00
Total	252,945	100%	247,055	100%	50,000	100%	-	-	-

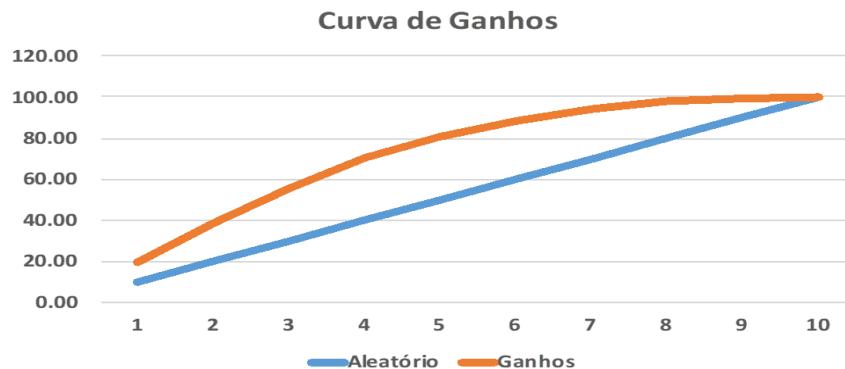


Uso de Métricas de Avaliação de Modelos Preditivos

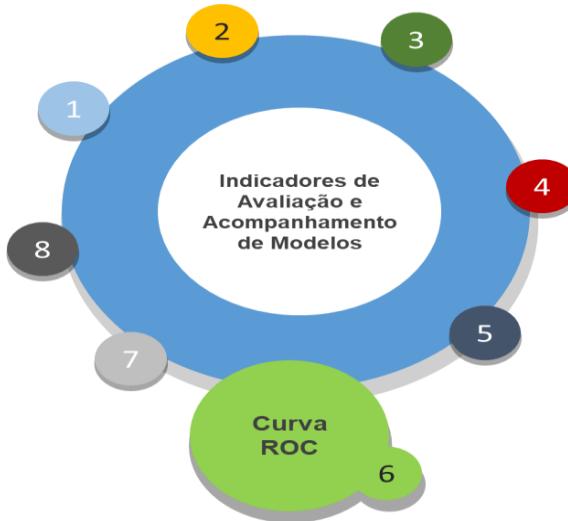


A **curva de ganhos** representa, por faixa de score, o % de resposta positiva com relação ao total. Quanto melhor o modelo, maior o % de respostas positivas nas primeiras faixas. Considerando 10% da base (aleatório), é provável que capturemos 10% do nosso evento de interesse (ex.: os bons pagadores). Imaginem um modelo para capturar os bons, selecionamos os 10% mais propensos a pagar e identificamos 19% de todos os bons pagadores. Para esse exemplo, o "gains" para essa faixa foi de 19%. Essa medida é acumulada por faixa. Assim, se considerarmos o grupo 2, é esperado que o modelo aleatório traga 20% dos bons pagadores. Se observarmos os resultados do modelo, as duas primeiras faixas (os 20% mais propensos) trouxeram 38%. O racional é análogo para as faixas subsequentes.

Grupo	Bons		Maus		Total	Total	Modelo Aleatório	Gains
	#	%	#	%				
1	48,541	19%	1,459	1%	50,000	10%	10%	19%
2	48,769	19%	1,231	0%	50,000	10%	20%	38%
3	42,412	17%	7,588	3%	50,000	10%	30%	55%
4	38,754	15%	11,246	5%	50,000	10%	40%	71%
5	25,641	10%	24,359	10%	50,000	10%	50%	81%
6	18,954	7%	31,046	13%	50,000	10%	60%	88%
7	15,411	6%	34,589	14%	50,000	10%	70%	94%
8	8,995	4%	41,005	17%	50,000	10%	80%	98%
9	4,214	2%	45,786	19%	50,000	10%	90%	100%
10	1,254	0%	48,746	20%	50,000	10%	100%	100%
Total	252,945	100%	247,055	100%	50,000	100%	-	-

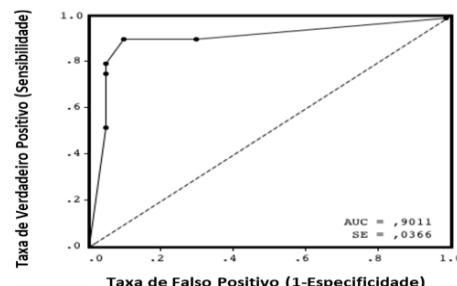


Uso de Métricas de Avaliação de Modelos Preditivos



Curva ROC (*Receiver Operating Characteristics ou Característica de Operação do Receptor*)

O uso dessa curva veio da teoria de detecção de sinais e foi criado por engenheiros na Segunda Guerra Mundial para detecção de objetos inimigos nas batalhas. Depois, foi para a Psicologia para uso na detecção de estímulos sensoriais. Por fim, passou a ser utilizada nas áreas de Aprendizado de Máquina e Mineração de Dados.



Valor Preditivo	Valor Observado (Valor Verdadeiro)	
	Mau	Bom
Mau	VP (verdadeiro positivo)	FP (Falso Positivo)
Bom	FN (Falso Negativo)	VN (Verdadeiro Negativo)

1. **Acurácia:** $(VP+VN)/(P+N)$. Proporção de valores corretos, sem considerar o que é negativo ou positivo.
2. **Sensibilidade:** $(VP)/(VP+FN)$. Proporção de verdadeiros positivos, ou, analogamente, proporção de "maus pagadores" identificados.
3. **Especificidade:** $(VN)/(FP+VN)$. Proporção de verdadeiros negativos, ou, analogamente, proporção de "bons pagadores" identificados.
4. **Eficiência:** $(SENS+ESPECIF)/2$. Média aritmética entre especificidade e sensibilidade, que, normalmente, caminham em direções opostas.
5. **Precisão:** $(VP)/(VP+FP)$. Proporção de acerto do modelo de predição.

Da curva ROC, deriva um indicador muito importante, utilizado na avaliação de modelos: **AUROC (Área sob a curva ROC) ou AUC (Área sob a curva)**. Esse índice pode variar entre 0 e 1. Quanto mais próximo de 1, mais o modelo em avaliação se diferencia de um modelo aleatório, representado pela linha diagonal. O artigo referenciado nesse link (www2.sas.com/proceedings/sugi31/210-31.pdf) traz uma abordagem para cálculo da área sob a curva ROC usando programação SAS).

Uso de Métricas de Avaliação de Modelos Preditivos

EXERCÍCIO 07

Vamos começar a avaliar o poder preditivo do modelo de risco gerado para a base "Loan_Prediction_Data_train.csv", iniciando pelo cálculo do K-S do modelo saturado (com todas as variáveis) e, depois só com algumas variáveis selecionadas:

```
189 #Cálculo de métricas preditivas do modelo para a base Loan Prediction
190 #Cálculo do K-S (Kolmogorov Smirnov)
191 modelo_sat<-glm(Loan_Status ~ . , family = binomial (link='logit'), data = loanpred_I[,2:13])
192 summary(modelo_sat)
193 install.packages("dgof")
194 library(dgof)
195 modelo_sat_pred = predict(modelo_sat, loanpred_I, type="response")
196 loanpred_I<-data.frame(loanpred_I, modelo_sat_pred)
197 View(loanpred_I)
198 ks.test(loanpred_I$modelo_sat_pred[loanpred_I$Loan_Status==0],
199         loanpred_I$modelo_sat_pred[loanpred_I$Loan_Status==1])
...

```

```
> ks.test(loanpred_I$modelo_sat_pred[loanpred_I$Loan_Status==0],
+           loanpred_I$modelo_sat_pred[loanpred_I$Loan_Status==1])
```

Two-sample Kolmogorov-Smirnov test

```
data: loanpred_I$modelo_sat_pred[loanpred_I$Loan_Status == 0] and loanpred_I$modelo_sat_
tstatus == 1]
D = 0.40277, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Lembrando que o K-S varia entre 0 e 1. Bons modelos de Credit Scoring geralmente variam entre 25% e 40%. Quando incorporam variáveis externas (redes sociais, birôs, dados judiciais, etc.), este K-S pode supercar os 50%.

Uso de Métricas de Avaliação de Modelos Preditivos

EXERCÍCIO 07

Na sequência, vamos usar a matriz de confusão para calcular o % de classificação correta do modelo, partindo de uma premissa, por exemplo, que , se a probabilidade for > 0.5 (ou 50%), o cliente é mau, caso contrário, ele é bom.

```
201 #Cálculo do % de classificação correta
202 loanpred_I$Loan_Status_pred<-ifelse(loanpred_I$modelo_sat_pred>0.5,1,0)
203 View(loanpred_I)
204 CrossTable(loanpred_I$Loan_Status_pred, loanpred_I$Loan_Status, prop.r = FALSE, prop.t = FALSE,
205             prop.chisq = FALSE)
206
```

		loanpred_I\$Loan_Status		Total
		0	1	
loanpred_I\$Loan_Status_pred	0	378 0.896	96 0.500	474
	1	44 0.104	96 0.500	140
Total		422 0.687	192 0.313	614

A taxa de acerto dos maus, que é o que mais nos interessa, é de 50%, enquanto que a taxa de acerto dos bons é de quase 90% (via de regra, é muito mais fácil acertar os bons).

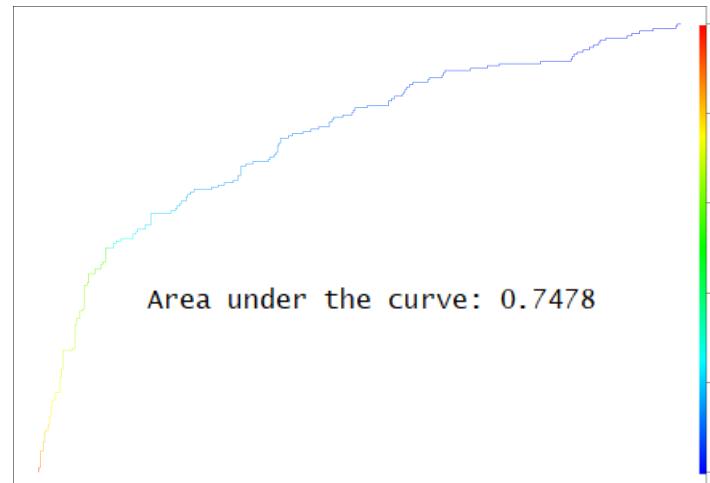
Uso de Métricas de Avaliação de Modelos Preditivos

EXERCÍCIO 07

A curva ROC também é uma forma interessante de avaliar a acurácia de um modelo, por conta do cálculo da área sob a curva, denominado de AUC (Area Under Curve). Essa estatística varia entre 0,5 e 1. Quanto mais próximo de 1, melhor a acurácia do modelo.

```
215 #Cálculo do AUC (Area Under Curve) e desenho da curva ROC
216 install.packages("ROCR")
217 install.packages("pROC")
218 library(ROCR)
219 library(pROC)
220 ROCRpred <- prediction(modelo_sat_pred, loanpred_I$Loan_Status)
221 ROCRperf <- performance(ROCRpred, 'tpr', 'fpr')
222 plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
223 roc_obj <- roc(loanpred_I$Loan_Status, loanpred_I$modelo_sat_pred)
224 auc(roc_obj)
```

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)



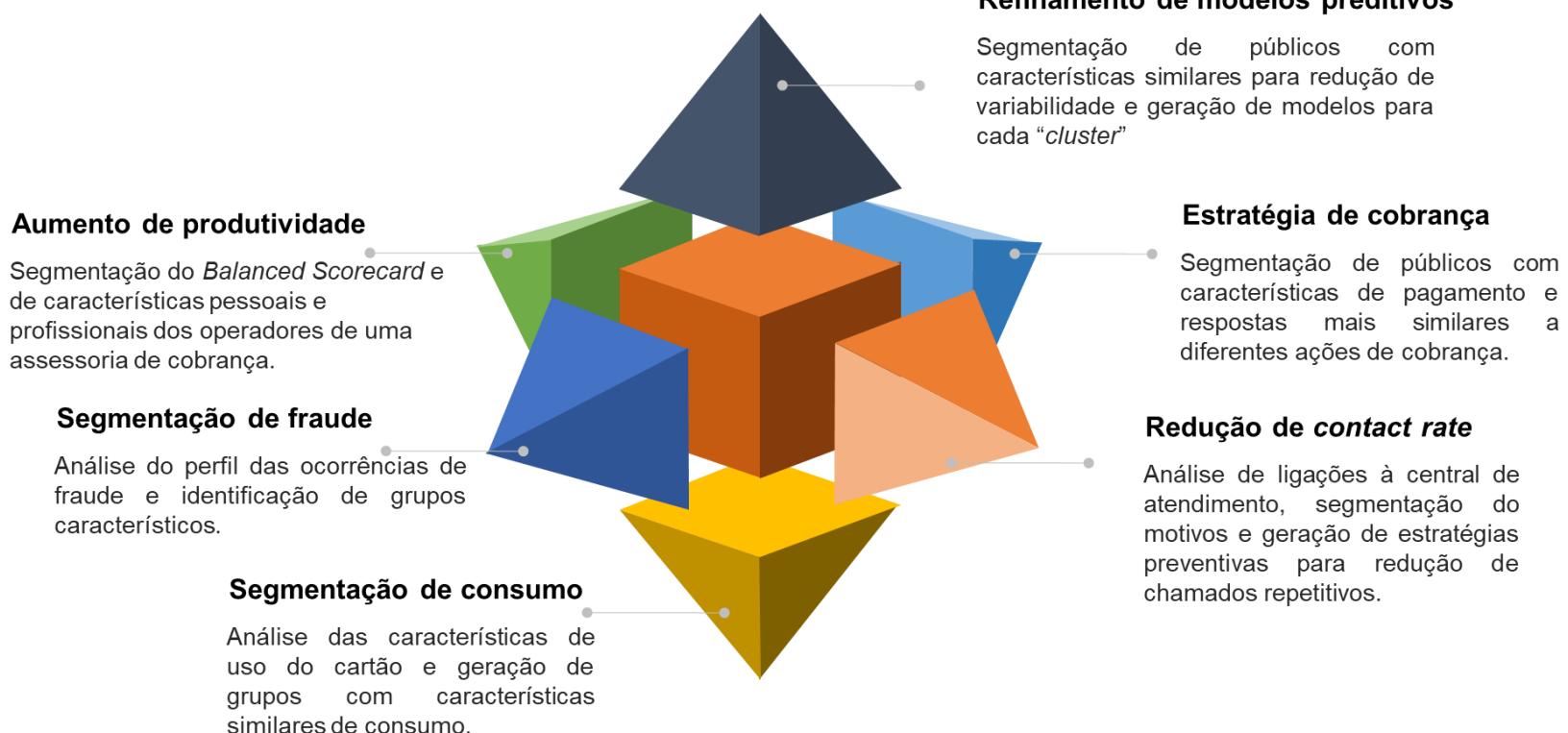
USO DE SEGMENTAÇÃO ESTATÍSTICA PARA REFINAMENTO DO PROCESSO DE ESCORAGEM

Segmentação Estatística para Refinamento do Processo de Escoragem

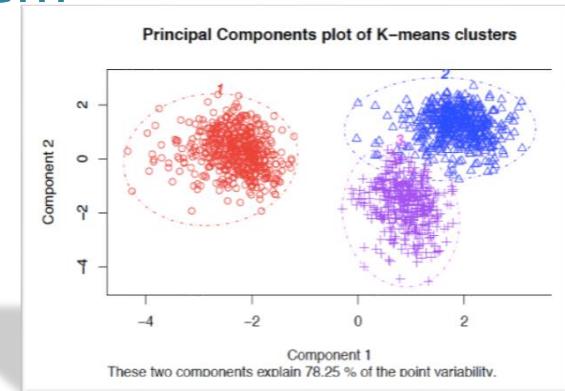
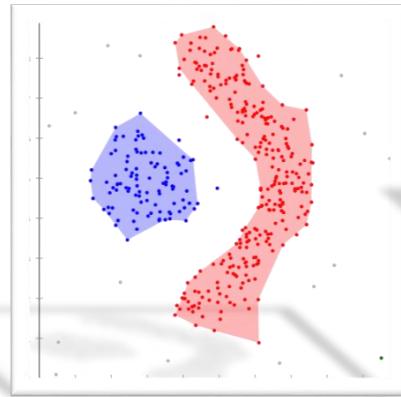
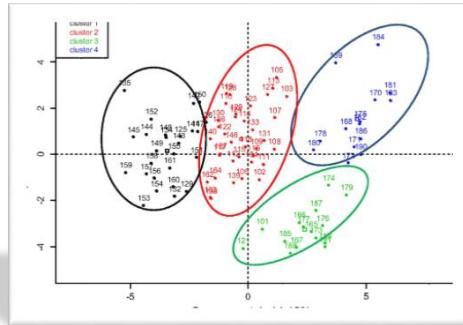
A análise de **clusters**, segmentação estatística, ou como queiram chamar, só existe, por conta da **variabilidade**. Ela mesma, se não houvesse variações, se não existissem diferenças, não haveria porquê buscar as semelhanças. Elas seriam lugar comum. Parece óbvio? Que bom!



Segmentação Estatística para Refinamento do Processo de Escoragem



Segmentação Estatística para Refinamento do Processo de Escoragem



Por quê sonho de consumo? Pois o resultado dessas análises conseguiu identificar agrupamentos (ou segmentos) bem separados e claramente identificáveis. Esse é o mundo ideal quando falamos em análise de clusters ou análise de agrupamentos.

Contudo,...

Segmentação Estatística para Refinamento do Processo de Escoragem



Nem sempre, uma separação tão clara entre os grupos é possível e o desafio aumenta. São várias as razões:

- (a) Falta de informação relevante
- (b) Qualidade do dado é precária ou insuficiente
- (c) Realmente não existe diferença que consiga ser explicitada de forma tão clara (ou seja, as semelhanças são muito maiores)

Segmentação Estatística para Refinamento do Processo de Escoragem

EXERCÍCIO 08

Para esse exemplo, vamos trabalhar com uma base de RH do Kaggle, chamada HR_Analytics.xlsx, que traz dados de funcionários de uma empresa, inclusive se saíram ou não da empresa. Nesse caso, o objetivo é desenhar um modelo de predição de saída de funcionários (*churn*) e, mais do que isso, identificar o impacto de uma segmentação sobre a capacidade de predizer a saída dos funcionários:

```
220 #Segmentação de uma base de funcionários de uma empresa
221 HR<-read_excel("HR_Analytics_Cluster.xlsx", sheet = "HR_Analytics_Cluster")
222 View(HR)
```

The screenshot shows a data grid in RStudio with 14,999 entries. The columns are labeled: ID, Satisfaction_level, last_evaluation, number_project, average_montly_hours, time_spend_company, Work_accident, left, promotion_last_years, sales, and salary. The data includes various numerical values and categorical labels like 'low', 'medi', and 'high' for the salary column.

ID	Satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_years	sales	salary
1	0.38	0.53	2	157	3	0	1	0	sales	low
2	0.80	0.86	5	262	6	0	1	0	sales	medi
3	0.11	0.88	7	272	4	0	1	0	sales	medi
4	0.72	0.67	5	223	5	0	1	0	sales	low
5	0.37	0.52	2	159	3	0	1	0	sales	low
6	0.41	0.50	2	153	3	0	1	0	sales	low
7	0.10	0.77	6	247	4	0	1	0	sales	low
8	0.92	0.85	5	259	5	0	1	0	sales	low
9	0.89	1.00	5	224	5	0	1	0	sales	low
10	0.42	0.53	2	142	3	0	1	0	sales	low
11	0.45	0.54	2	135	3	0	1	0	sales	low
12	0.11	0.81	6	305	4	0	1	0	sales	low
13	0.84	0.92	4	234	5	0	1	0	sales	low
14	0.41	0.55	2	148	3	0	1	0	sales	low
15	0.36	0.56	2	137	3	0	1	0	sales	low
16	0.38	0.54	2	143	3	0	1	0	sales	low

Segmentação Estatística para Refinamento do Processo de Escoragem

EXERCÍCIO 08

Inicialmente, vamos montar um modelo saturado único, tentando identificar o K-S desse modelo:

```
224 #Criação modelo de risco único  
225 modelo<-glm(left ~ . , family = binomial (link='logit'), data = HR[,2:11])  
226 summary(modelo)  
227 HR_pred = predict(modelo, HR, type="response")  
228 HR<-data.frame(HR, HR_pred)  
229 View(HR)  
230 ks.test(HR$HR_pred[HR$left==0],  
231           HR$HR_pred[HR$left==1])  
---
```

```
> ks.test(HR$HR_pred[HR$left==0],  
+          HR$HR_pred[HR$left==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
```

**Modelo único:
K-S = 50,23%**

Segmentação Estatística para Refinamento do Processo de Escoragem

EXERCÍCIO 08

Agora, vamos estudar a utilidade de uma **variável chamada QCL_1, resultado da aplicação do método de segmentação *k-means*, considerando apenas as variáveis quantitativas, gerando 3 distintos clusters. Vamos desenvolver um modelo específico para cada cluster:**

```
233 #Modelos específicos para cada um dos 3 clusters
234
235 HR_cluster_I<-subset(HR, HR$QCL_1==1)
236 HR_cluster_II<-subset(HR, HR$QCL_1==2)
237 HR_cluster_III<-subset(HR, HR$QCL_1==3)
238 |
239 modelo_cluster_I<-glm(left ~ . , family = binomial (link='logit'), data = HR_cluster_I[,2:11])
240 modelo_cluster_II<-glm(left ~ . , family = binomial (link='logit'), data = HR_cluster_II[,2:11])
241 modelo_cluster_III<-glm(left ~ . , family = binomial (link='logit'), data = HR_cluster_III[,2:11])
242
243 HR_pred_cluster_I<-predict(modelo_cluster_I, HR_cluster_I, type = "response")
244 HR_cluster_I<-data.frame(HR_cluster_I,HR_pred_cluster_I )
245 View(HR_cluster_I)
246 str(HR_cluster_I)
247
248 HR_pred_cluster_II<-predict(modelo_cluster_II, HR_cluster_II, type="response")
249 HR_cluster_II<-data.frame(HR_cluster_II,HR_pred_cluster_II )
250 View(HR_cluster_II)
251
252 HR_pred_cluster_III<-predict(modelo_cluster_III, HR_cluster_III, type="response")
253 HR_cluster_III<-data.frame(HR_cluster_III,HR_pred_cluster_III )
254 View(HR_cluster_III)
255
```

Segmentação Estatística para Refinamento do Processo de Escoragem

EXERCÍCIO 08

```
256 #K-S modelo único  
257 ks.test(HR$HR_pred[HR$left==0],  
258     HR$HR_pred[HR$left==1])  
259  
260 #K-S modelo cluster I  
261 ks.test(HR_cluster_I$HR_pred_cluster_I.1[HR_cluster_I$left==0],  
262     HR_cluster_I$HR_pred_cluster_I.1[HR_cluster_I$left==1])  
263
```

```
> #K-S modelo único  
> ks.test(HR$HR_pred[HR$left==0],  
+           HR$HR_pred[HR$left==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  
  
> #K-S modelo cluster I  
> ks.test(HR_cluster_I$HR_pred_cluster_I.1[HR_cluster_I$left==0],  
+           HR_cluster_I$HR_pred_cluster_I.1[HR_cluster_I$left==1])  
  
Two-sample Kolmogorov-Smirnov test  
  
data: HR_cluster_I$HR_pred_cluster_I.1[HR_cluster_I$left == 0] and HR_<br/>cluster_I.1$left == 1]  
D = 0.88046, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

```
264 #K-S modelo cluster II  
265 ks.test(HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left==0],  
266     HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left==1])  
267  
268 #K-S modelo cluster III  
269 ks.test(HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left==0],  
270     HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left==1])  
271
```

```
> #K-S modelo cluster II  
> ks.test(HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left==0],  
+           HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left==1])  
  
Two-sample Kolmogorov-Smirnov test  
  
data: HR_cluster_II$HR_pred_cluster_II[HR_cluster_II$left == 0] and HR_<br/>cluster_II$left == 1]  
D = 0.62831, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

```
> #K-S modelo cluster III  
> ks.test(HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left==0],  
+           HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left==1])  
  
Two-sample Kolmogorov-Smirnov test  
  
data: HR_cluster_III$HR_pred_cluster_III[HR_cluster_III$left == 0] and HR_<br/>cluster_III$left == 1]  
D = 0.7862, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

K-S
Único:
0,5023

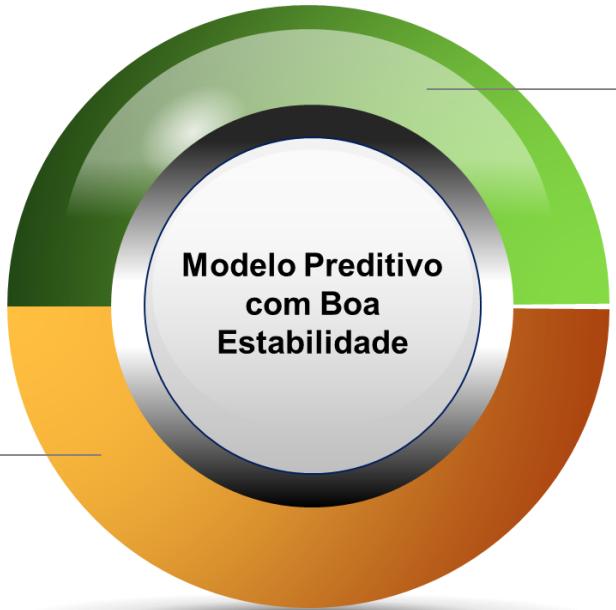
K-S
Ponderado:
0,7745

12

AVALIAÇÃO DA ESTABILIDADE DE UM MODELO DE ESCORAGEM DE CRÉDITO

Avaliação da Estabilidade do Modelo

O que significa ter um modelo estável, considerando sua implementação em produção:



O perfil dos clientes no momento da execução do modelo é similar ao perfil considerado no treinamento do modelo.

A distribuição dos scores dos clientes depois da implementação do modelo é similar àquele obtido na ocasião do treinamento do modelo.

Avaliação da Estabilidade do Modelo



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

Indicador do nível de mudança da distribuição de score, comparando a população de origem (utilizada para desenvolvimento do modelo) e a população atual (que reflete características atuais)

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

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

Identificada alguma mudança relevante, o caminho é avaliar em que variáveis do modelo essa mudança foi mais acentuada, de modo a direcionar focos de atuação para atualização da fórmula. O artigo colocado como “fonte” traz um processo de cálculo do PSI, utilizando SAS.

Avaliação da Estabilidade do Modelo

EXERCÍCIO 09

Quando falamos em estabilidade do modelo, nos referimos tanto à estabilidade da distribuição de score, quanto à distribuição de cada uma das características (variáveis) presentes no modelo. Vamos considerar a planilha “Distribuição 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. Vamos calcular o PSI,c 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

Avaliação da Estabilidade do Modelo

EXERCÍCIO 09

Faixa de Score	Distribuição de Score (Qtd)		
	Distribuição Esperada		
	Bons	Maus	Total
<=100	25.412	52.515	77.927
101-200	28.844	41.251	70.095
201-300	29.855	32.412	62.267
301-400	35.451	19.525	54.976
401-500	38.455	14.221	52.676
501-600	48.541	9.522	58.063
601-700	52.514	6.855	59.369
701-800	59.522	6.224	65.746
801-900	75.425	3.541	78.966
>900	85.641	2.111	87.752
Total	479.660	188.177	667.837

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

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

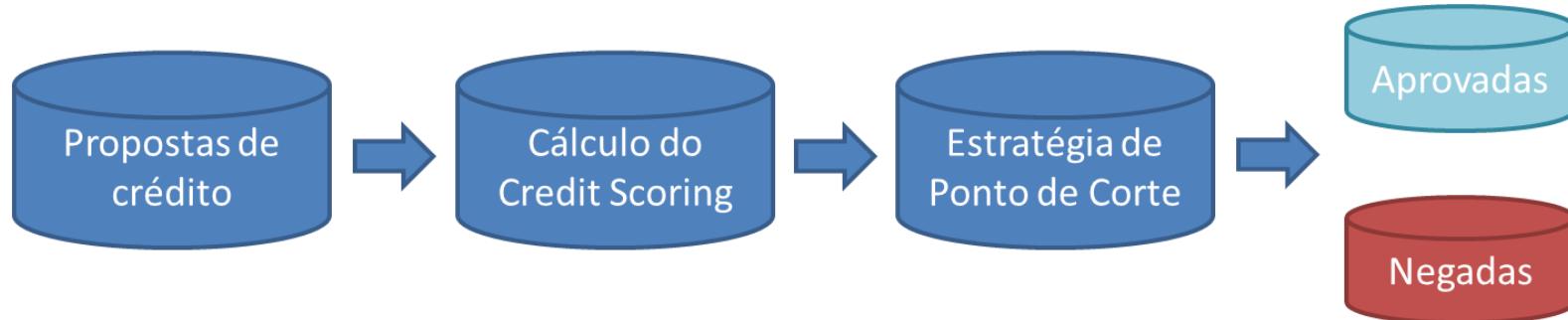
Como o valor obtido de PSI=0,0087 é inferior a 0,1, podemos concluir que o score está estável e não causa nenhuma preocupação aparente. Se ele estivesse acima de 0,25, seria necessário avaliar, separadamente, cada uma das características, para identificar as potenciais fontes de distorção.

PSI	Distribuição de Score (%)		
	Distribuição Esperada		
	Bons	Maus	Total
0,0018	6,8%	24,5%	13,2%
0,0003	5,4%	21,1%	11,1%
0,0016	4,9%	20,6%	10,6%
0,0000	6,9%	10,9%	8,3%
0,0016	7,5%	5,6%	6,8%
0,0021	8,9%	4,6%	7,4%
0,0001	11,3%	3,8%	8,6%
0,0003	14,2%	3,8%	10,4%
0,0007	15,1%	3,5%	10,9%
0,0002	19,0%	1,5%	12,7%
0,0000	100,0%	100,0%	100,0%
0,0087	100,0%	100,0%	100,0%

ESTRATÉGIAS DE CÁLCULO DO PONTO DE CORTE PARA TOMADA DE DECISÃO

Estratégias de Cálculo de Ponto de Corte para Tomada de Decisão

Quando pensamos em utilizar um modelo de Credit Scoring para tomada de decisão, deparamo-nos com o seguinte cenário, sobretudo para sua aplicação mais clássica, que é a concessão de crédito:



A estratégia para decisão de ponto de corte pode ser decidida em razão de inúmeros fatores, como por exemplo: **(a) Taxa de aprovação desejada**, **(b) Bad rate esperada**, **(c) Margem de contribuição desejada**. Muitos podem associar ponto de corte do modelo à estratégia padrão que os softwares estatísticos normalmente consideram, que é selecionar como 0,5 (50%), o ponto divisório entre bons e maus. Isso não se aplica quando estamos falando de escolher o ponto de corte para decidir se vamos aprovar ou negar o crédito.

Estratégias de Cálculo de Ponto de Corte para Tomada de Decisão

EXERCÍCIO 10

Vamos considerar a seguinte tabela de distribuição de score e considerar a taxa de aprovação como critério para definição do ponto de corte:

Faixa de Score	Distribuição de Score de Propostas Avaliadas em Dez/2018		
	Qtde de Propostas	% de Propostas	% Acumulado de Propostas
<=100	21.251	2,9%	100,0%
101-200	26.521	3,6%	97,1%
201-300	36.412	4,9%	93,6%
301-400	52.121	7,0%	88,7%
401-500	78.521	10,6%	81,7%
501-600	71.124	9,6%	71,1%
601-700	98.645	13,3%	61,5%
701-800	99.011	13,3%	48,3%
801-900	107.521	14,5%	35,0%
>900	152.321	20,5%	20,5%
Total	743.448	100,0%	-

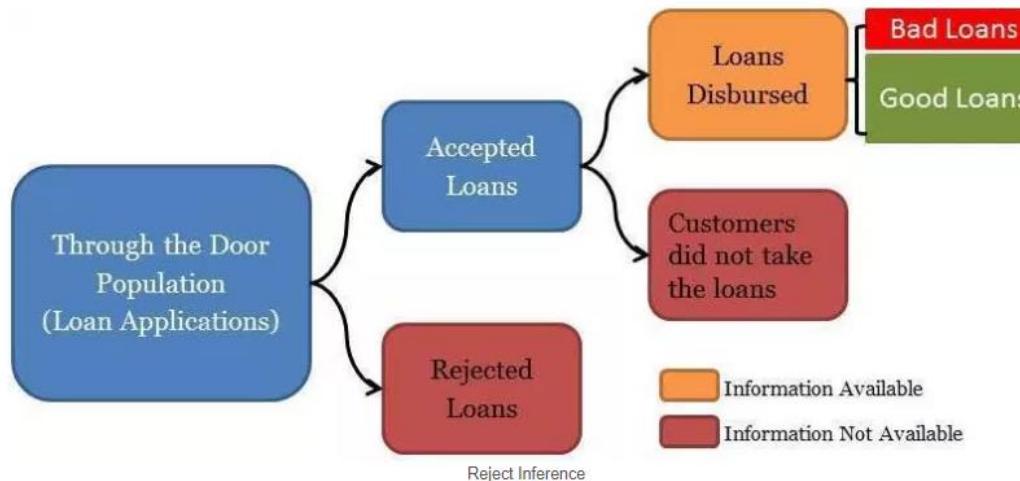
Analisando a tabela e considerando que queremos ter por volta de 89% de aprovação das propostas, nosso ponto de corte deveria ser estabelecido como 301, ou seja, todos aqueles clientes que tiverem score superior a 300 seriam aprovados. Os demais, por volta de 11% seriam negados. Logicamente, essa estratégia pode ser insuficiente se desejarmos ser mais rigorosos com as futuras taxas de default, devendo considerar outros elementos, como taxa de maus ("bad rate"), margem de contribuição, entre outras métricas de mercado.

14

INFERÊNCIA DE REJEITADOS

Inferência de Rejeitados

Em um processo tradicional de análise de crédito, o score de crédito (**Credit Score**) desempenha função estratégica. Ele ajuda a determinar quais clientes serão **aprovados** e terão a concessão do crédito e quais serão **rejeitados (ou negados)** e não o receberão (salvo por políticas de exceção). Como os clientes rejeitados, via de regra, não receberam crédito, não foi possível obter e acompanhar sua performance e, portanto, é difícil considerá-los em um processo de modelagem de risco de crédito. Para endereçar esse problema, existe uma abordagem chamada “**Inferência dos Rejeitados**”.



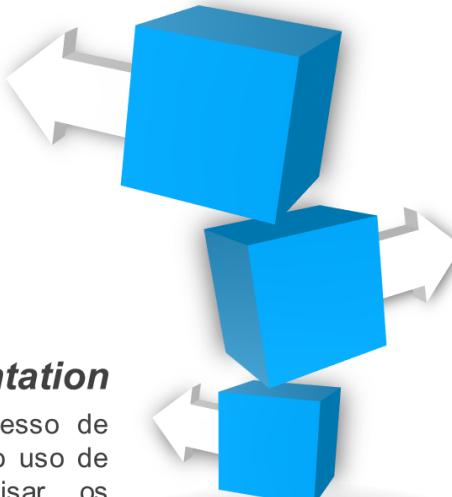
Fonte: <http://ucanalytics.com/blogs/reject-inference-scorecards-banking-case-part-5/>

Inferência de Rejeitados

Em um processo tradicional de análise de crédito, o score de crédito (**Credit Score**) desempenha função estratégica. Ele ajuda a determinar quais clientes serão **aprovados** e terão a concessão do crédito e quais serão **rejeitados (ou negados)** e não o receberão (salvo por políticas de exceção). Como os clientes rejeitados, via de regra, não receberam crédito, não foi possível obter e acompanhar sua performance e, portanto, é difícil considerá-los em um processo de modelagem de risco de crédito. Para endereçar esse problema, existe uma abordagem chamada “**Inferência dos Rejeitados**”.

Credit Bureaus

Captura de informações de performance dos rejeitados nos birôs de crédito, que concentram dados desses clientes em outras instituições.



Fuzzy Augmentation

É uma extensão do processo de *parceling*, que contempla o uso de lógica fuzzy para analisar os rejeitados, sendo considerado um método mais sofisticado de análise.

Augmentation Through Parceling

Essa é a forma mais conhecida para criar um modelo KGB (*known-good-bad model*) e escorar a base de rejeitados.

USO DE SCORECARDS PARA MELHOR INTERPRETAÇÃO DE UM MODELO DE ESCORAGEM DE CRÉDITO

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

A forma como um modelo de risco é mostrado e apresentado ao público executivo e aos reguladores faz com que seu entendimento e aceitação sejam mais fáceis. Vejamos o caso da saída tradicional de um modelo de regressão logística, gerado no R:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	7.205e-01	8.474e-01	0.850	0.39516
GenderMale	9.769e-02	2.789e-01	0.350	0.72617
MarriedNo	6.501e-01	2.375e-01	2.738	0.00618 **
Dependents0	-4.064e-01	6.403e-01	-0.635	0.52558
Dependents1	-3.404e-02	6.671e-01	-0.051	0.95931
Dependents2	-4.849e-01	6.769e-01	-0.716	0.47377
Dependents3+	-1.202e-01	7.133e-01	-0.169	0.86614
EducationNot Graduate	3.365e-01	2.392e-01	1.407	0.15943
Self_EmployedYes	-7.617e-03	2.943e-01	-0.026	0.97935
ApplicantIncome	1.007e-05	2.093e-05	0.481	0.63053
CoapplicantIncome	4.677e-05	3.789e-05	1.234	0.21705
LoanAmount	-2.576e-04	1.439e-03	-0.179	0.85792
Loan_Amount_Term	8.212e-04	1.241e-03	0.662	0.50799
Credit_History	-2.166e+00	2.241e-01	-9.666	< 2e-16 ***
Property_AreaSemiurban	-7.950e-01	2.472e-01	-3.216	0.00130 **
Property_AreaUrban	-2.728e-01	2.452e-01	-1.112	0.26595

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

Os pesos das variáveis são estabelecidos, mas, muitas vezes, a interpretação dos resultados não é tão óbvia ou transparente para um executivo do *C-level* ou mesmo para pessoas não habituadas ao dia a dia da construção e análise de modelos de risco.

Usualmente, o coeficiente de uma variável (“beta”) é interpretado como o **impacto na odds ratio** gerado pelo crescimento de 1 unidade na variável em questão.

É consenso que essa forma de interpretação não é nada intuitiva e, muitas vezes, torna mais complexa e trabalhosa o processo de “venda” do valor do modelo preditivo.

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

Por essa e por outras razões que foi criado o conceito e aplicação do scorecard, ou seja, uma grande tabela que atribui pontos (positivos, neutros ou negativos) a diversos atributos, em que, no final, o score é a soma dessas pontuações. Essa forma de apresentar o modelo é amplamente intuitiva e fácil de ser assimilada pelos mais diversos públicos.

Decision from Scorecard 38

Characteristic	Attribute	Scorecard Points
AGE	<2	100
AGE	22<=AGE<26	120
AGE	26<=AGE<30	185
AGE	30<=AGE<32	200
AGE	32<=AGE<37	210
AGE	37<=AGE<42	225
AGE	>=42	250
HOME	OWN	225
HOME	RENT	110
INCOME	<10000	120
INCOME	10000<=INCOME<17000	140
INCOME	17000<=INCOME<20000	180
INCOME	28000<=INCOME<35000	200
INCOME	35000<=INCOME<42000	225
INCOME	42000<=INCOME<58000	230
INCOME	>=58000	260

Let cutoff=600
So, a new customer applies for credit.....

AGE	35	210 points
INCOME	\$38K	225 points
HOME	OWN	225 points
Total		660 points
Decision: GRANT CREDIT		

Note: A scorecard is scaled with the **Odds**, **Scorecard Points** and **Points to Double the Odds** properties.

Fonte: <https://www.slideshare.net/TuhinChattopadhyay/credit-scorecard>

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

Software “Decision Modeler”, da FICO.

Variables		Bins				
Age Type real Baseline score 0	Description	Values	Label (optional)	Partial score	Unexpected	Reason code
	< 18			5	<input type="checkbox"/>	ACAD05
	18 <= .. < 21			20	<input type="checkbox"/>	ACAD03
	21 <= .. < 60			30	<input type="checkbox"/>	ACAD02
	> = 60			50	<input type="checkbox"/>	ACAD01
	All Other			0	<input checked="" type="checkbox"/>	UEXP
> Marital Status		Description				

Fonte: <https://www.youtube.com/watch?v=hBh2fET0H5A>

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

Software Estatístico “Statistica”, da Tibco.

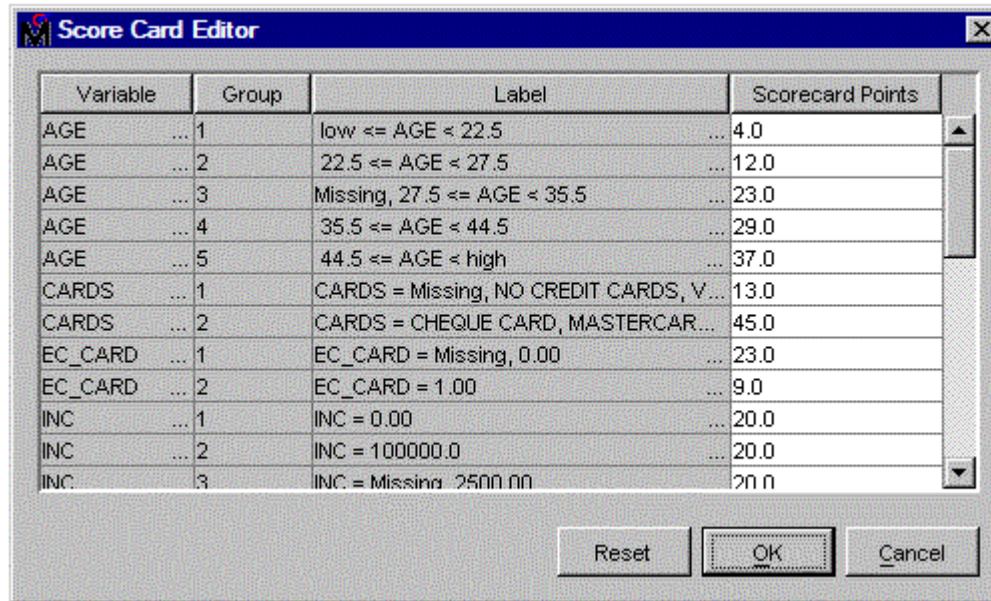
The screenshot shows a software window titled "Scorecard". On the left is a table with columns: Variable, Value/Range, WoE, Estimate, Wald stat., p value, Scoring, and Rounded scoring. The table lists various credit-related variables like "Balance of Current Account" and "Duration of Credit" with their respective ranges and statistical values. On the right side of the window, there is a vertical toolbar with buttons for "Close", "Back", and options for "Report", "Save as Excel", "Save", "Script", and "XML Script". A message box also appears stating "Scoring values can be changed".

Variable	Value/Range	WoE	Estimate	Wald stat.	p value	Scoring	Rounded scoring
Balance of Current Account	no running account	-81.810	0.00932	51.19893	0.00000	20.575	21
Balance of Current Account	no balance	-40.139	0.00932	51.19893	0.00000	31.781	32
Balance of Current Account	<= \$300	104.229	0.00932	51.19893	0.00000	70.604	71
Balance of Current Account	>\$300	104.229	0.00932	51.19893	0.00000	70.604	71
Balance of Current Account	Neutral value	-	-			47.062	47
Duration of Credit	[-,inf[75.377	0.00277	1.20626	0.27207	48.600	49
Duration of Credit	[9;15[38.549	0.00277	1.20626	0.27207	45.656	46
Duration of Credit	[15;30[-10.834	0.00277	1.20626	0.27207	41.709	42
Duration of Credit	[30;36[-61.368	0.00277	1.20626	0.27207	37.670	38
Duration of Credit	[36;inf[-91.629	0.00277	1.20626	0.27207	35.252	35
Duration of Credit	Neutral value	-	-			42.491	42
Payment of Previous Credits	paid back	73.374	0.00750	14.59009	0.00013	58.454	58
Payment of Previous Credits	hesitant	-123.407	0.00750	14.59009	0.00013	15.869	16
Payment of Previous Credits	problematic running accounts	-123.407	0.00750	14.59009	0.00013	15.869	16
Payment of Previous Credits	no previous credits	-8.787	0.00750	14.59009	0.00013	40.674	41
Payment of Previous Credits	no problems with current credits	-8.787	0.00750	14.59009	0.00013	40.674	41
Payment of Previous Credits	Neutral value	-	-			43.541	44
Purpose of Credit	other	-35.920	0.01100	17.13579	0.00003	31.174	31
Purpose of Credit	new car	77.384	0.01100	17.13579	0.00003	67.136	67
Purpose of Credit	furniture	41.006	0.01100	17.13579	0.00003	55.590	56
Purpose of Credit	repair	-60.614	0.01100	17.13579	0.00003	23.337	23
Purpose of Credit	retraining	-23.052	0.01100	17.13579	0.00003	35.258	35
Purpose of Credit	used car	-10.286	0.01100	17.13579	0.00003	39.310	39

Fonte: <http://www.statsoft.com/textbook/credit-scoring>

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

Software Estatístico “SAS Credit Scoring for Enterprise Miner”, do SAS



The screenshot shows the "Score Card Editor" dialog box from SAS. It contains a table with four columns: Variable, Group, Label, and Scorecard Points. The table lists rules for variables AGE, CARDS, EC_CARD, and INC. The rules are ordered by Group (1 to 5) and then by Label. The "Scorecard Points" column shows the points assigned to each rule. The dialog box has standard Windows-style buttons at the bottom: Reset, OK, and Cancel.

Variable	Group	Label	Scorecard Points
AGE	1	low <= AGE < 22.5	4.0
AGE	2	22.5 <= AGE < 27.5	12.0
AGE	3	Missing, 27.5 <= AGE < 35.5	23.0
AGE	4	35.5 <= AGE < 44.5	29.0
AGE	5	44.5 <= AGE < high	37.0
CARDS	1	CARDS = Missing, NO CREDIT CARDS, V...	13.0
CARDS	2	CARDS = CHEQUE CARD, MASTERCAR...	45.0
EC_CARD	1	EC_CARD = Missing, 0.00	23.0
EC_CARD	2	EC_CARD = 1.00	9.0
INC	1	INC = 0.00	20.0
INC	2	INC = 100000.0	20.0
INC	3	INC = Missing, 2500.00	20.0

Fonte: <http://documentation.sas.com/?docsetId=emref&docsetTarget=n181v13wdwn89mn1pfpqm3w6oaz5.htm&docsetVersion=14.3&locale=en>

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

Package “scorecard” no R

Package ‘scorecard’

February 10, 2019

Version 0.2.3

Title Credit Risk Scorecard

Description The ‘scorecard’ package makes the development of credit risk scorecard easier and efficient by providing functions for some common tasks, such as data partition, variable selection, woe binning, scorecard scaling, performance evaluation and report generation. These functions can also be used in the development of machine learning models.

The references including:

1. Refaat, M. (2011, ISBN: 9781447511199). Credit Risk Scorecard: Development and Implementation Using SAS.
2. Siddiqi, N. (2006, ISBN: 9780471754510). Credit risk scorecards. Developing and Implementing Intelligent Credit Scoring.

Depends R (>= 3.1.0)

Imports data.table (>= 1.10.0), ggplot2, gridExtra, foreach, doParallel, parallel, openxlsx

Suggests knitr, rmarkdown, pkgdown, testthat

License MIT + file LICENSE

URL <https://github.com/ShichenXie/scorecard>

BugReports <https://github.com/ShichenXie/scorecard/issues>

LazyData true

VignetteBuilder knitr

RoxygenNote 6.1.1

Fonte: <https://cran.r-project.org/web/packages/scorecard/scorecard.pdf>

Uso de Scorecards para Melhor Interpretação de um Modelo de Risco

EXERCÍCIO 11

Demonstração do processo de construção de
um *scorecard* no R, a partir da base
“Empréstimo Bancário.xlsx”.