



Universidade Federal do Paraná
Laboratório de Estatística e Geoinformação - LEG



Introdução

Eduardo Vargas Ferreira

O que é Machine Learning?



Estatística

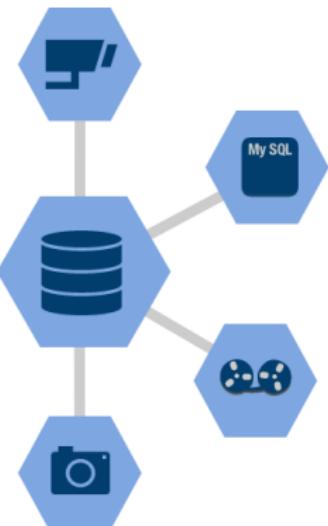


Machine Learning



aplicados a

Ciência da
computação



Métodos

problemas

- **Data Modeling Culture**

- ★ Domina a comunidade estatística;
- ★ O principal objetivo está na interpretação dos parâmetros;
- ★ Testar suposições é fundamental.

- **Algorithmic Modeling Culture**

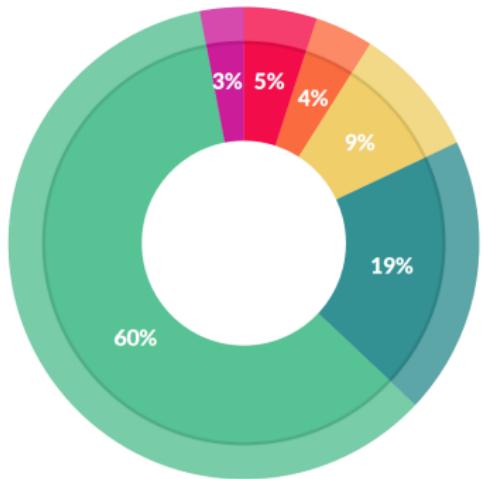
- ★ Domina a comunidade de Machine Learning;
- ★ O modelo é utilizado para criar bons algoritmos preditivos;
- ★ Interpretamos os resultados, mas esse - em geral - não é o foco.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199-231, 2001

Algoritmos de Machine Learning



Onde desprendemos mais tempo



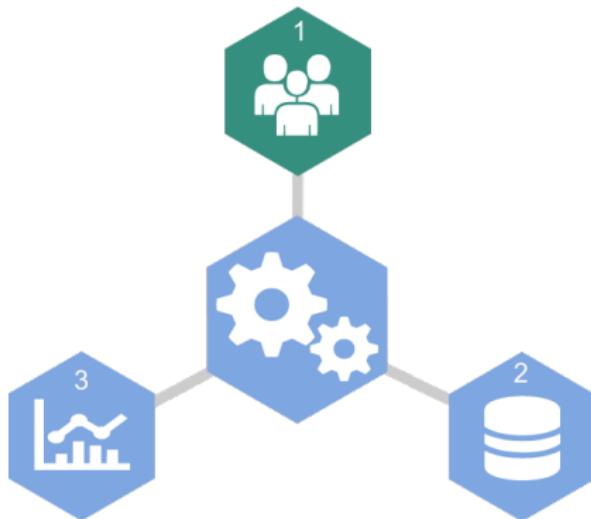
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

CrowdFlower

- 1 **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise. Devemos saber onde queremos chegar!
- 2 **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
- 3 **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
- 4 **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
- 5 **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
- 6 **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas.

Objetivos



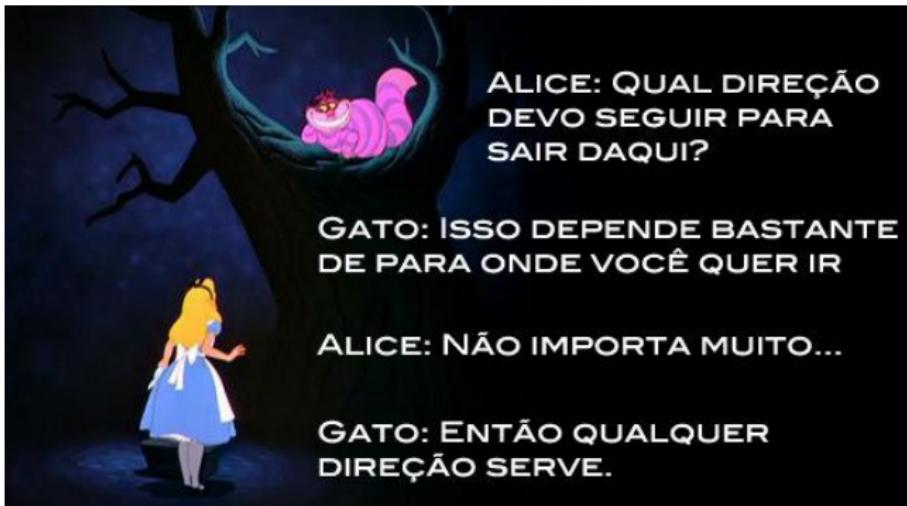
Modelos

Base de dados

Defina seus objetivos



- Para quem não sabe onde vai, qualquer direção serve!



- Qual o problema na foto ao lado?
- Sendo o animal de tração, troque-o por um avião!
- Se continuar, compre um mais potente;



Vamos procurar por “ideias fora da caixa”



- Com o passar do tempo, criamos padrões que ficam cada vez mais estabelecidos em nossa mente;



- Este pensamento reflete em toda organização. Notamos processos funcionando da mesma maneira, meses, até anos e não fazemos nada;



O homem criativo não é um homem comum ao qual se acrescentou algo. Criativo é o homem comum do qual nada se tirou

(Abraham Maslow)



- Isso não tem relação com lado do cérebro mais desenvolvido;
- E sim com técnica e vontade de fazer diferente;
- Soluções antigas não resolvem problemas novos.

① Criativo

- ★ Resulta em novas ideias e possibilidades;
- ★ Sem ele, em geral, ocorre “mais do mesmo”.

② Lógico positivo

- ★ Como fazer novas ideias funcionarem;
- ★ Sem ele mudanças não serão práticas e funcionais.

③ Lógico negativo (crítico)

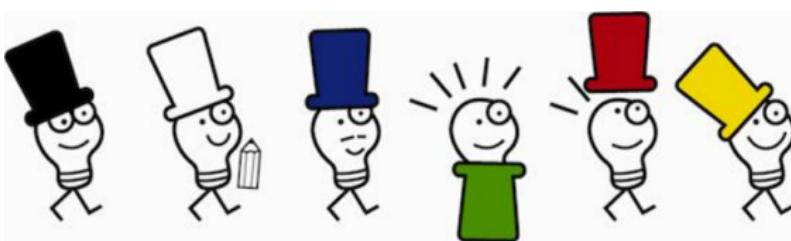
- ★ Busca por falhas na nova ideia;
- ★ Sem ele problemas podem não vir à tona.

Seis chapéus do pensamento

Seis chapéus do pensamento



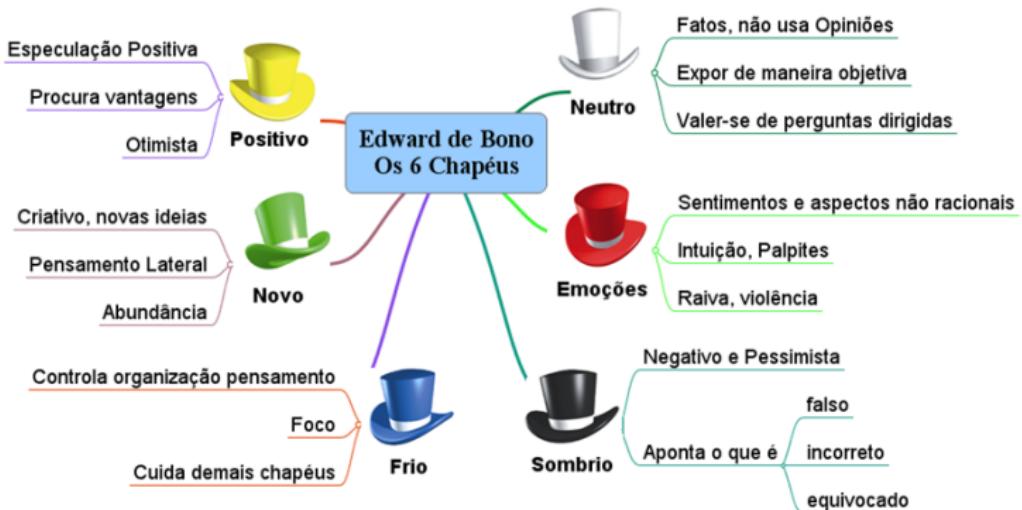
- Nos ajuda a analisar um problema, uma ideia ou situação de diversas perspectivas, permitindo uma visão mais abrangente da situação;



► Os 6 chapéus do pensamento

- De acordo com a cor do chapéu, nos focamos em apenas um aspecto do pensamento, deixando os demais de lado, até mudar do chapéu.

Seis chapéus do pensamento



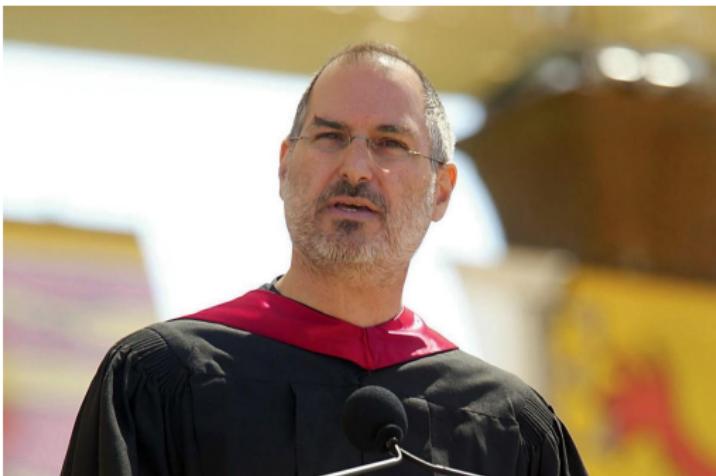
Mi criado Reinaldo Geraldo jun 2012

Ser criativo é ter um bom repertório



"Criatividade é apenas conectar coisas. Quando você pergunta a pessoas criativas como elas criaram algo, elas se sentem culpadas, pois não criaram algo de fato, apenas viram alguma coisa óbvio ali." **Steve Jobs**

► Steve Jobs, Connecting the dots



Técnica dos “Cinco Por quês”

Técnica dos “Cinco Por quês”



- Foi percebido que o monumento de Abraham Lincoln deteriorava-se mais rapidamente do que qualquer outro em Washington, D.C. Por quê?



- ➊ Porque é limpo com mais frequência que os outros monumentos. Por quê?
- ➋ Porque tem mais dejetos de pássaros que os outros monumentos. Por quê?
- ➌ Porque tem mais pássaros em volta deste monumento do que dos outros. Por quê?
- ➍ Porque tem mais insetos em torno deste monumento. Por quê?
- ➎ Porque a lâmpada que o ilumina é diferente das outras e atraí mais insetos.

- A solução para o problema é a troca da lâmpada. Poderiam trocar os produtos de limpeza ou colocar espantalho, mas o problema persistiria.

- O consumo de sorvete está correlacionado com o número de afogamentos de piscina;
- **O sorvete não causa afogamentos. Ambos estão correlacionados com o clima do verão;**
- Em 90% das brigas do bar que terminaram em uma morte, a pessoa que começou a briga morreu;
- **Claro, é a pessoa que sobreviveu contando a história;**
- Terapia de reposição hormonal está correlacionada com uma menor taxa de doença coronária;
- **Pessoas que realizam reposição hormonal, geralmente, pertencem à grupos socioeconômicos mais elevados, com hábitos mais saudáveis;**

Diagrama de causa e efeito

Diagrama de causa e efeito



- O Diagrama de causa e efeito ajuda a descobrir, organizar e resumir todo esse conhecimento atual, alinhando a equipe à respeito do problema;

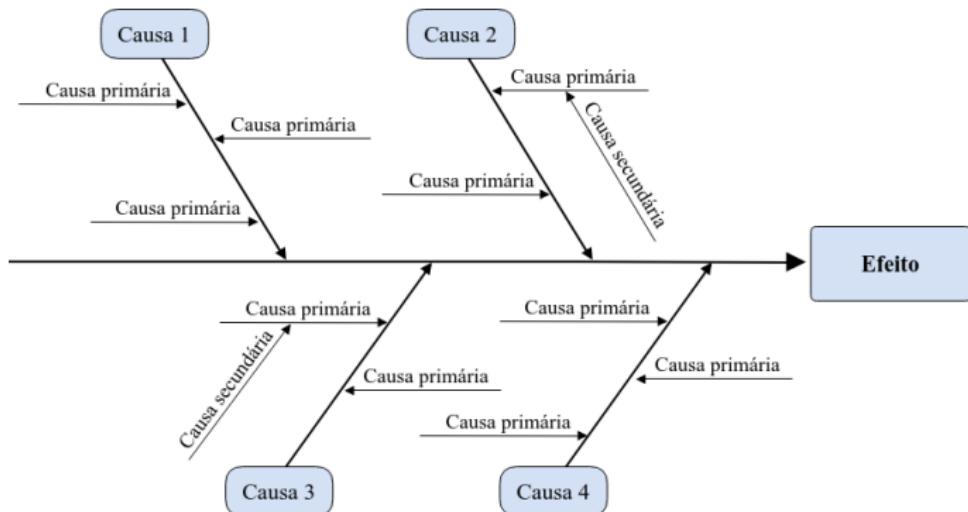


Diagrama direcionador

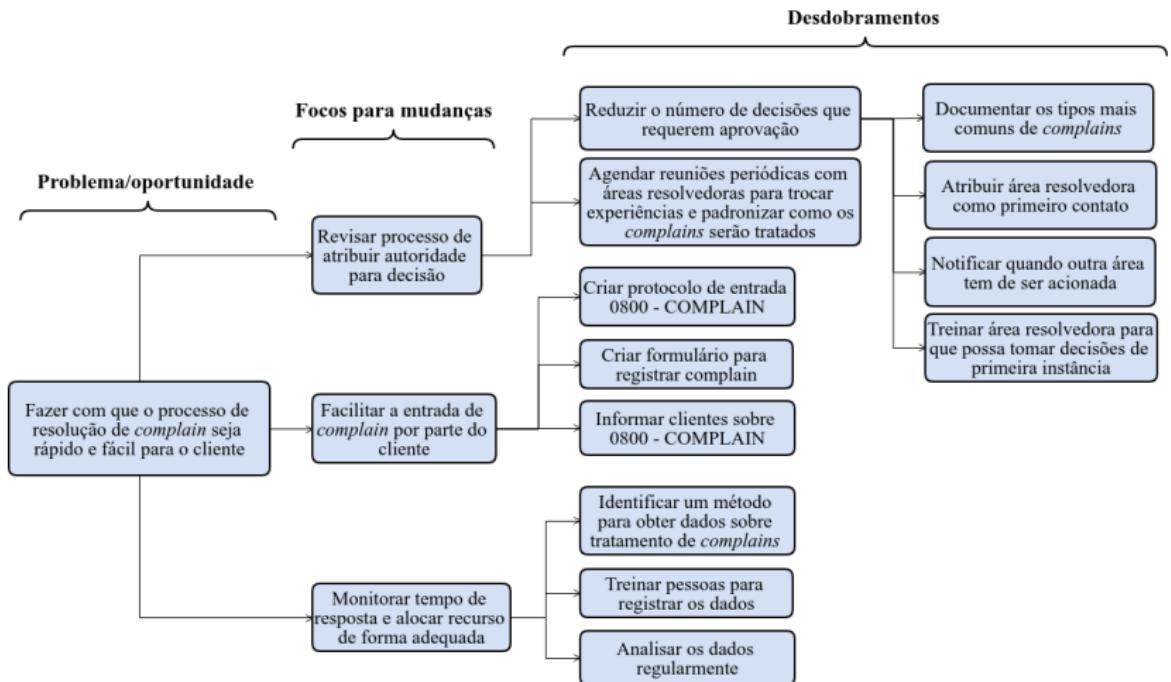
Diagrama direcionador

- Assim como placas e faixas auxiliam no trânsito, essa técnica contribui para a busca de soluções nas diversas fases das análises.

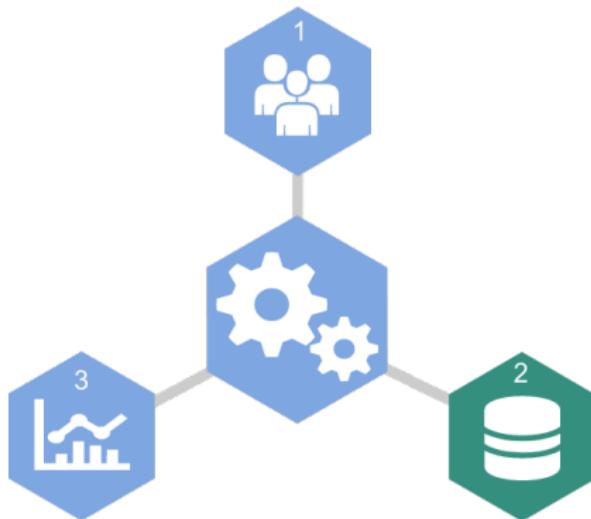


- Como uma espécie de mapa, ele aponta caminhos ou alternativas que podem ser tomados pelo grupo de trabalho do projeto.

Diagrama direcionador



Objetivos



Modelos

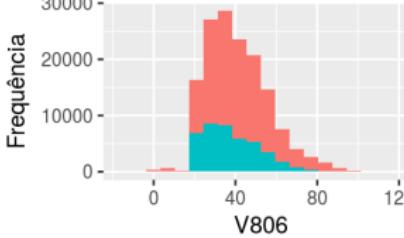
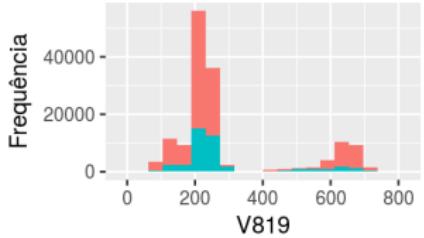
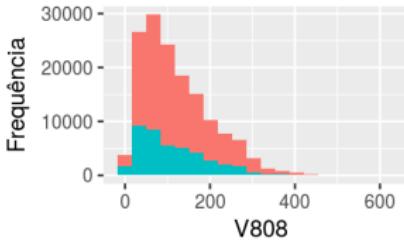
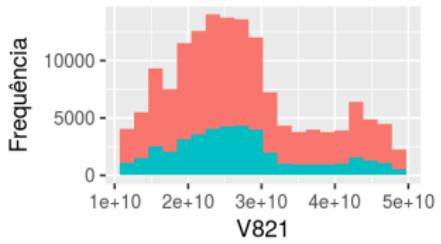
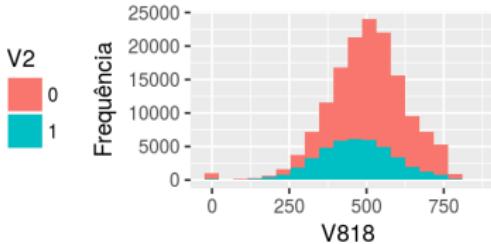
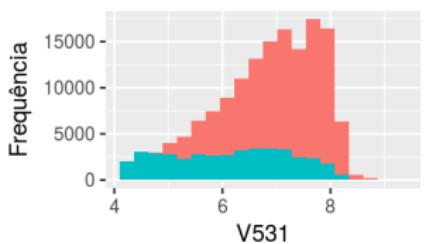
Base de dados

The Art of Feature Engineering



- **Feature engineering** é a arte de extrair informação dos dados já obtidos:
 - ➊ Criação de características;
 - ➋ Dados faltantes;
 - ➌ Dados desbalanceados;
 - ➍ Variáveis correlacionadas.

Exemplo: clientes em atraso



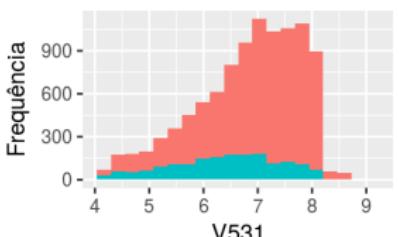
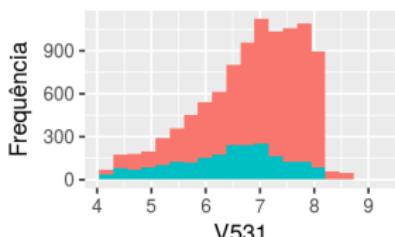
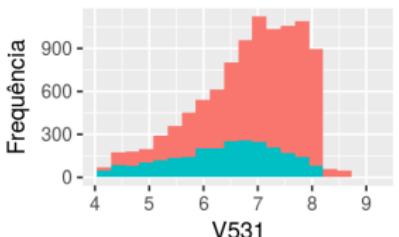
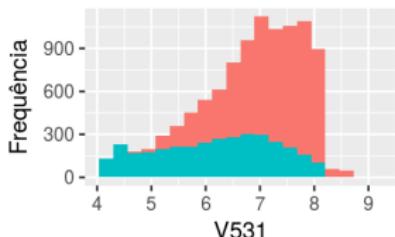
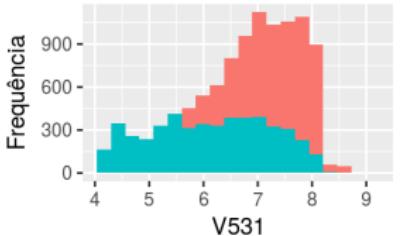
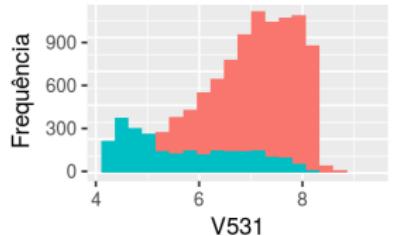
Pense um passo a frente



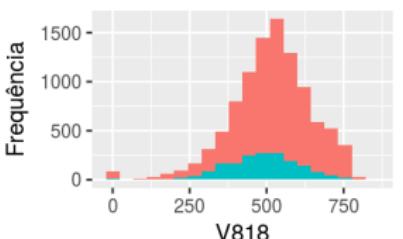
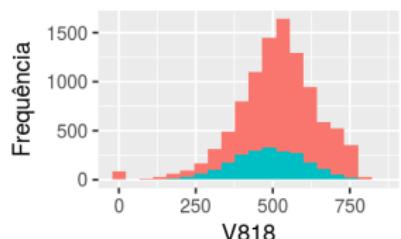
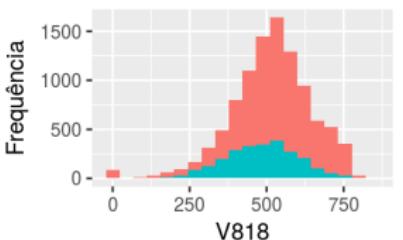
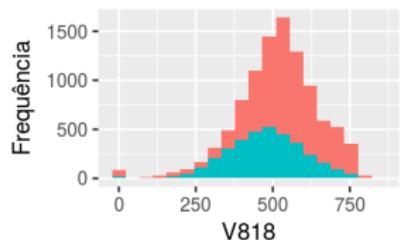
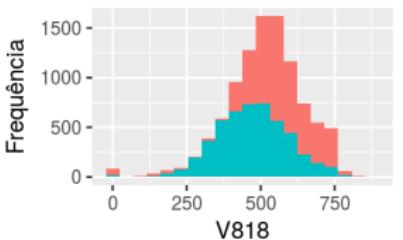
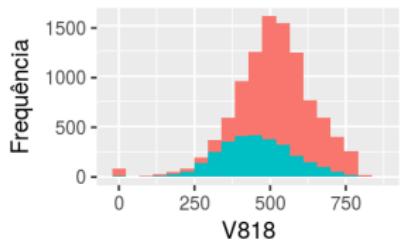
- A análise não pode parar no gráfico! Você deve extrair algo adicional que a máquina não é capaz de fazer.



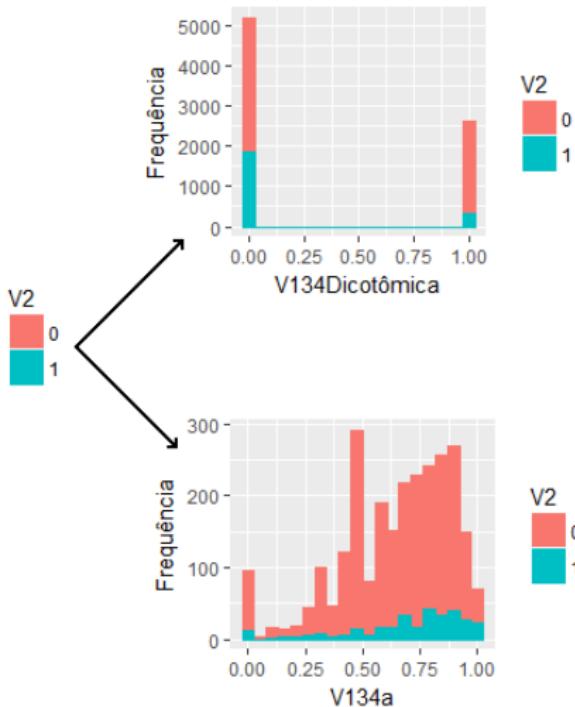
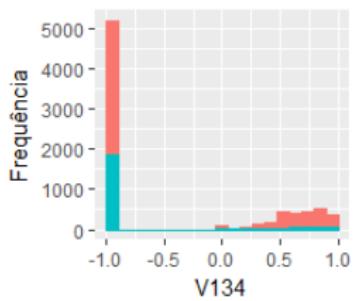
Exemplo: clientes em atraso



Exemplo: clientes em atraso

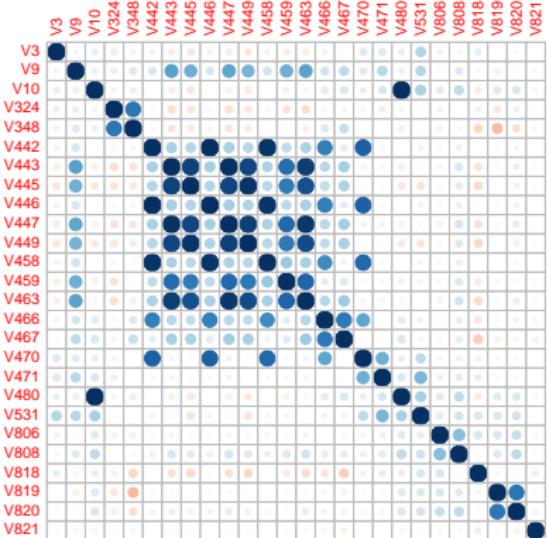


Exemplo: clientes em atraso

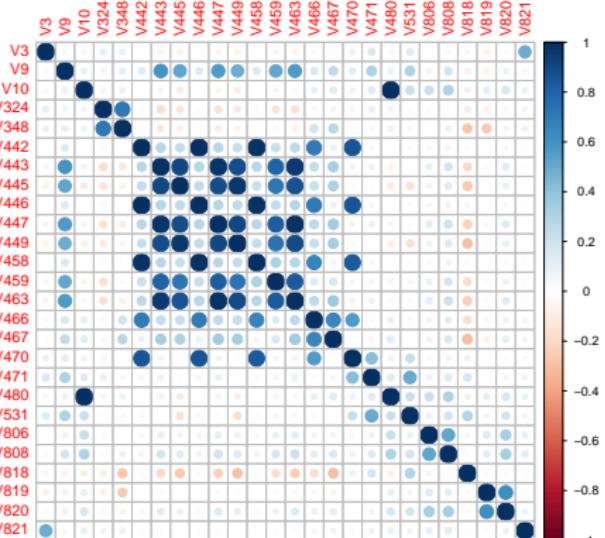


Exemplo: clientes em atraso

Dados completos



Somente adimplentes





- **Feature engineering** é a arte de extrair informação dos dados já obtidos.
 - ➊ Criação de características;
 - ➋ Dados faltantes;
 - ➌ Dados desbalanceados;
 - ➍ Variáveis correlacionadas.

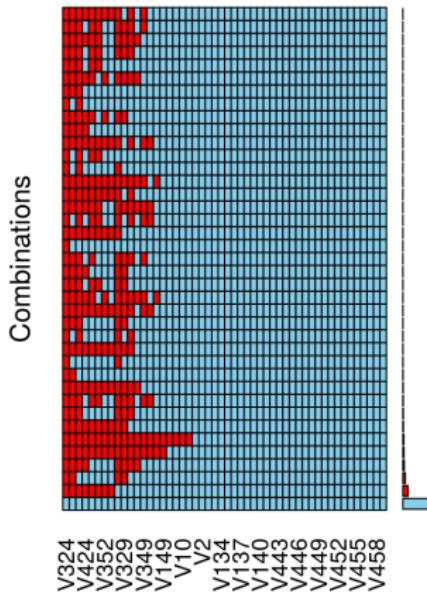
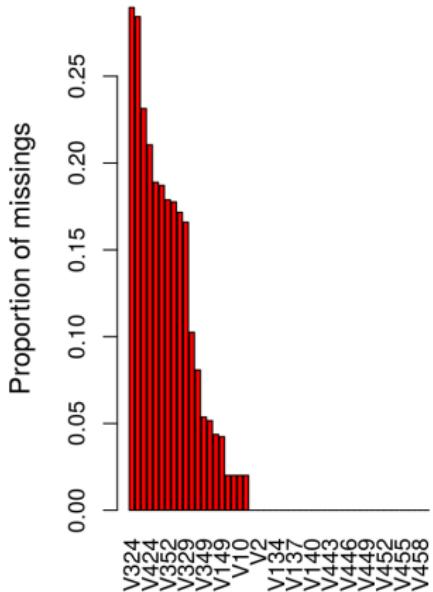
- **Missing completely at random:** quando a probabilidade dos dados faltantes é a mesma entre as observações, p. ex.:
 - ★ Dados perdidos por um *backup* incorreto;
- **Missing at random:** quando os dados faltantes variam de acordo com outras variáveis, p. ex.:
 - ★ *Missing* sobre idade pode ser diferente entre mulheres e homens;
- **Missing not at random:** quando a probabilidade de *missing* está relacionada com o *missing*, p. ex.,:
 - ★ Dependendo da renda do cliente, é mais provável que ele não responda sobre a renda;
 - ★ Indivíduo não comparece ao teste de droga, porque a utilizou na noite anterior.

Tratamento dos dados faltantes

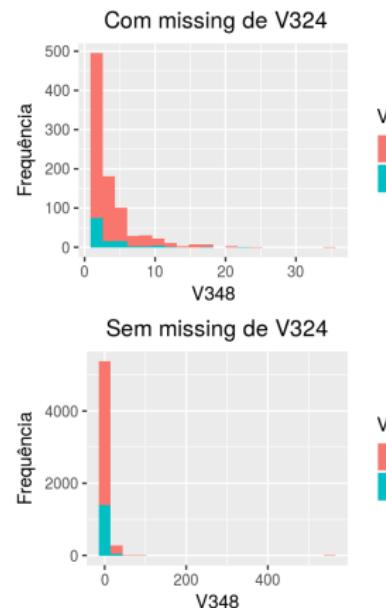
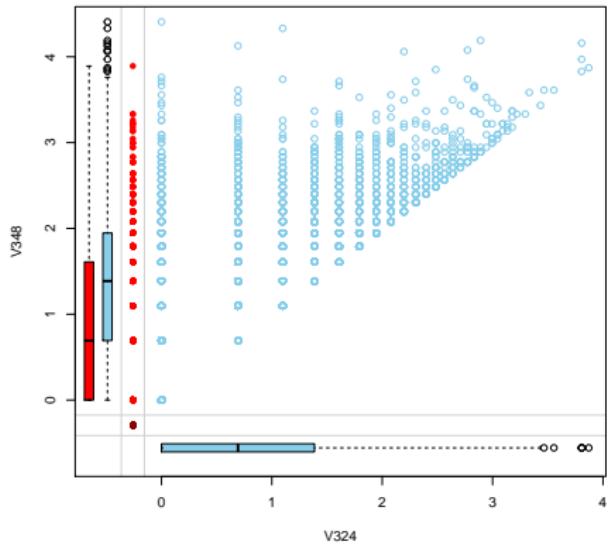


- **Deletar:** utilizado quando a natureza do “*missing*” é **completely at random**.
 - ★ Podemos eliminar a linha inteira. É uma abordagem simples, mas retira poder dos dados, devido à redução do tamanho da amostra;
 - ★ Ou utilizar os dados completos, de acordo - somente - com as variáveis de interesse.
- **Imputação:** utilizado quando trata-se de **missing at random** ou **missing not at random**.
 - ★ Média, mediana, moda;
 - ★ Modelo preditivo.

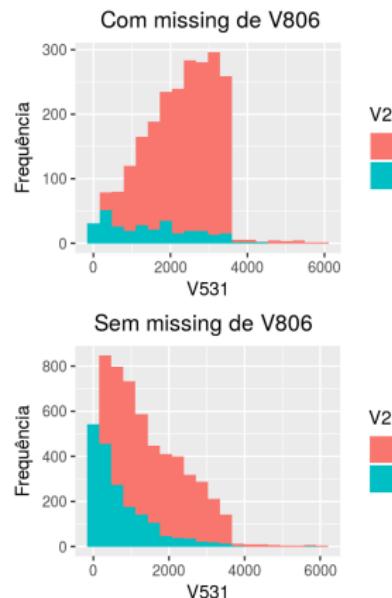
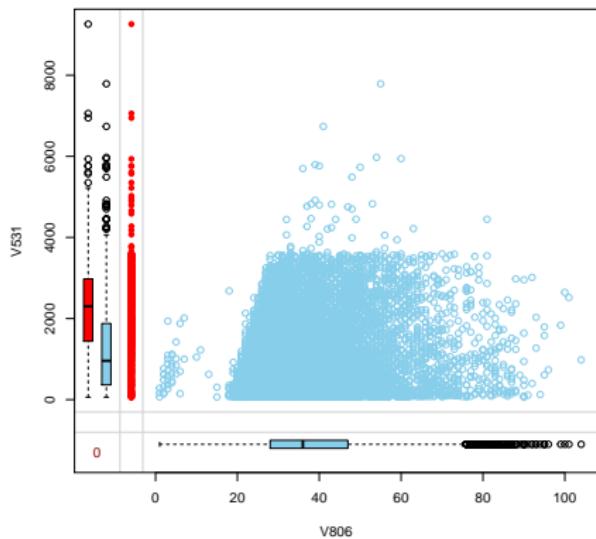
Exemplo: clientes em atraso



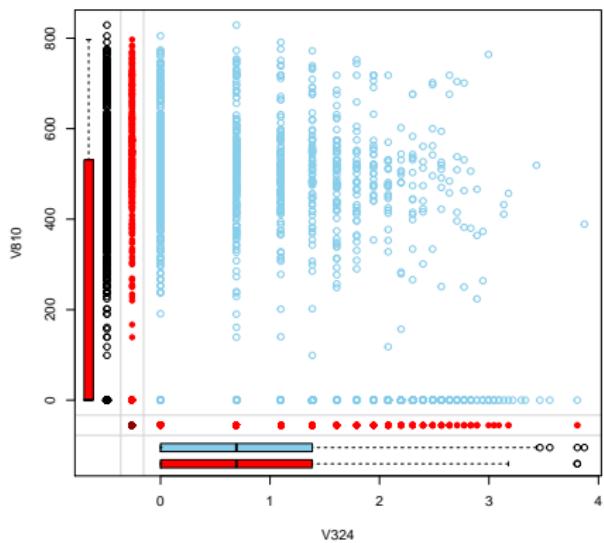
Exemplo: clientes em atraso



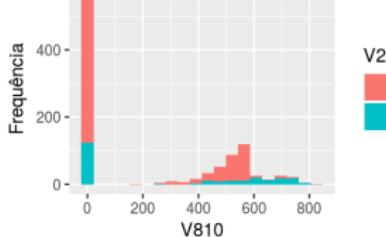
Exemplo: clientes em atraso



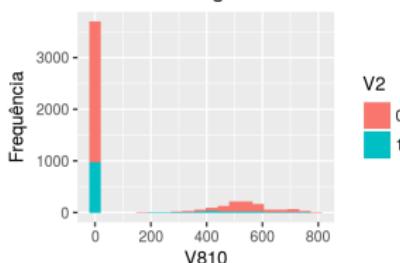
Exemplo: clientes em atraso



Com missing de V324



Sem missing de V324

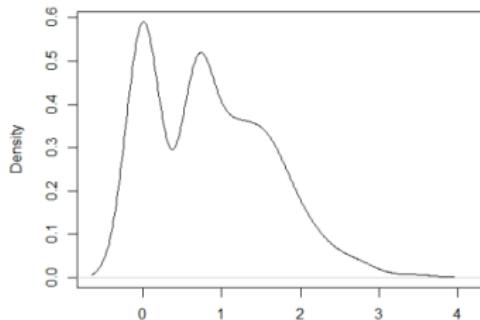


Seja cauteloso na imputação dos dados!

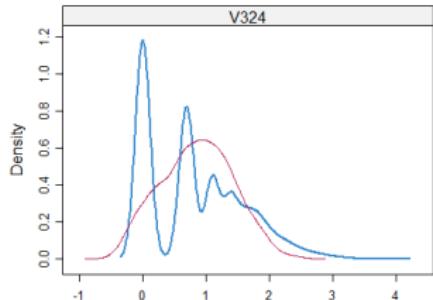


"The idea of imputation is both seductive and dangerous." **D.B. Rubin**

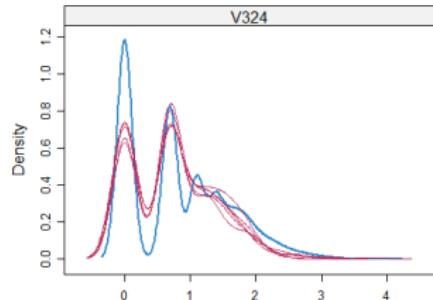
Exemplo: clientes em atraso



Regressão



Random Forest



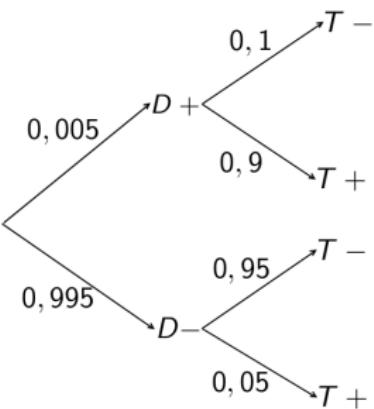
The Art of Feature Engineering



- **Feature engineering** é a arte de extrair informação dos dados já obtidos.
 - ➊ Criação de características;
 - ➋ Dados faltantes;
 - ➌ Dados desbalanceados;
 - ➍ Variáveis correlacionadas.

Exemplo: Regra de Bayes

- Consider a routine triage test for a disease. Suppose that the prevalence of the disease in the population (basic rate) is 0.5%.



- The test is highly accurate:
 - $P(\text{False positive}) = P(T+|D-) = 0,05$;
 - $P(\text{False negative}) = P(T-|D+) = 0,1$.

Exemplo: Regra de Bayes



- Qual a probabilidade de ter a doença, dado que o teste foi positivo?

$$P(D+ | T+) = \frac{P(D+) \times P(T+ | D+)}{P(T+)}$$

- Calculamos o denominador utilizando a lei da probabilidade total:

$$\begin{aligned}P(T+) &= P(D+) \times P(T+ | D+) + P(D-) \times P(T+ | D-) \\&= .005 \times .9 + .995 \times .05 \\&= .05425\end{aligned}$$

- Assim,

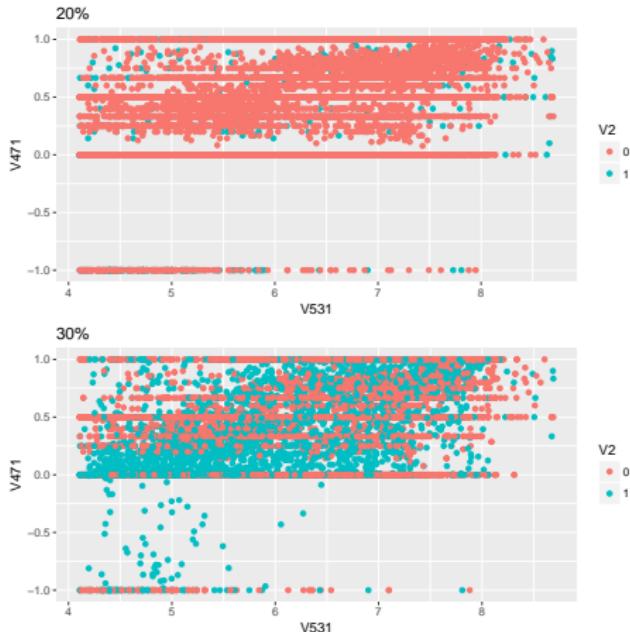
$$\begin{aligned}P(D+ | T+) &= \frac{P(D+) \times P(T+ | D+)}{P(T+)} = \frac{0,005 \times 0,9}{0,05425} \\&\approx 8,3\%.\end{aligned}$$

“Realizar um simples exame de urina pode levar a um falso positivo, o que poderia desencadear uma cascata de outros testes, apenas para descobrir ao final que não há nada de errado com você”, diz Mehrotra. ▶ [Link do artigo](#)



Dados desbalanceados

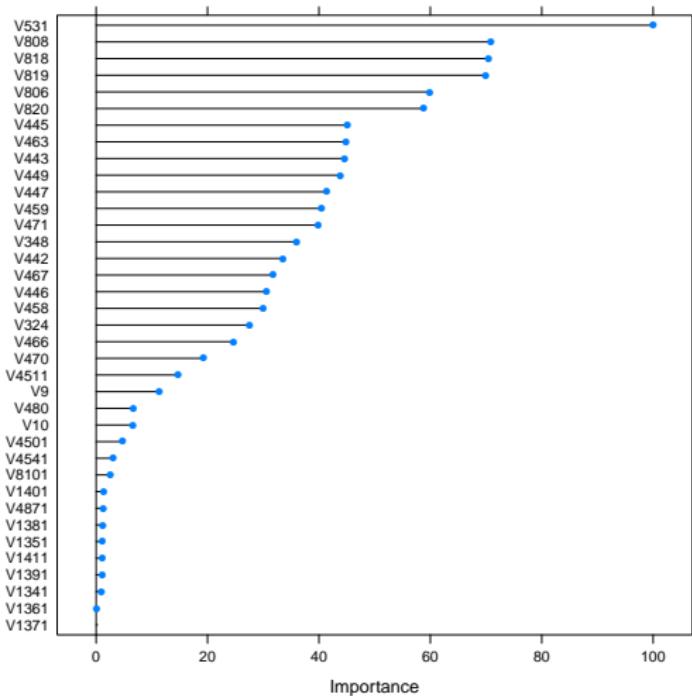
- ① **SMOTE (Synthetic Minority Over-sampling Technique):** remostrar o conjunto de dados original por superamostragem da classe minoritária;



Exemplo: clientes em atraso



SMOTE: 20% sim e 80% não (dados originais)



Confusion Matrix and Statistics

		Reference	
		0	1
Prediction	0	4168	890
	1	142	238

Accuracy : 0.8102

95% CI : (0.79, 0.82)

Sensitivity : 0.9671

Specificity : 0.2110

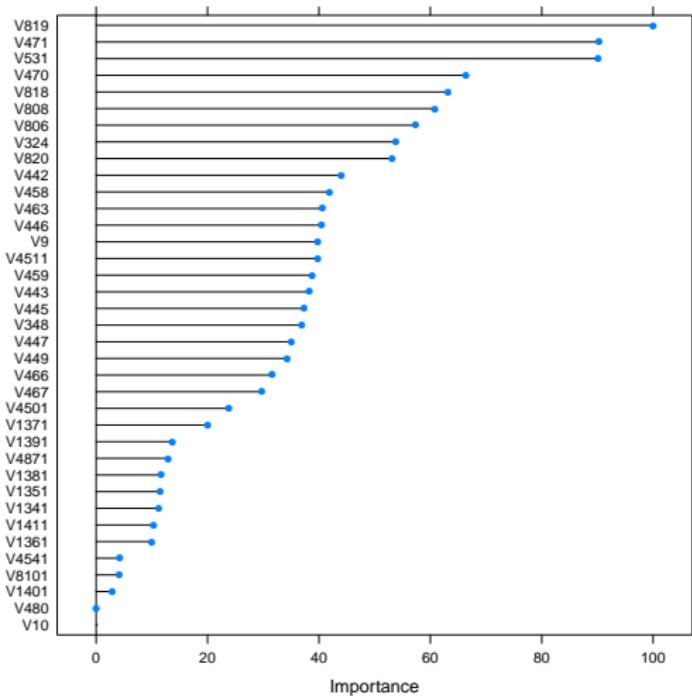
Pos Pred Value : 0.8240

Neg Pred Value : 0.6263

Exemplo: clientes em atraso



SMOTE: 25% sim e 75% não



Confusion Matrix and Statistics

		Reference	
		Prediction	0
Prediction	0	1	
0	4258	729	
1	52	399	

Accuracy : 0.8564
95% CI : (0.84, 0.86)

Sensitivity : 0.9879

Specificity : 0.3537

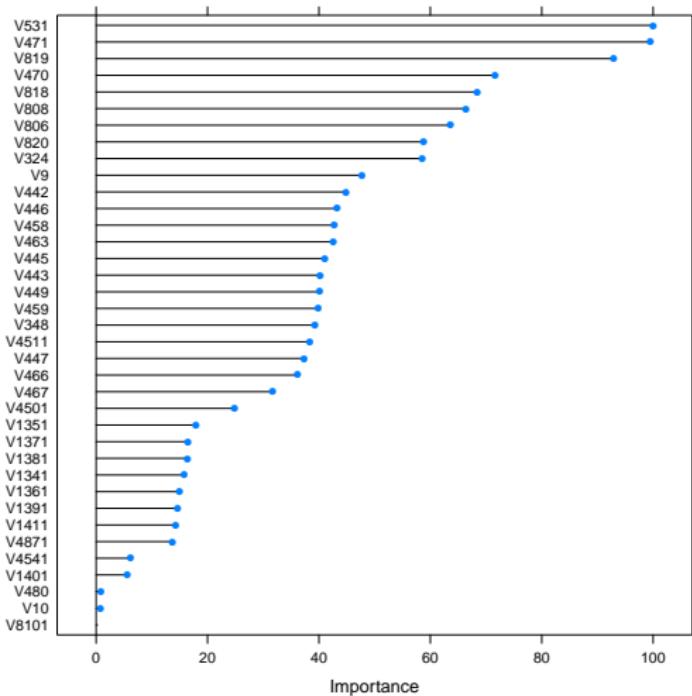
Pos Pred Value : 0.8538

Neg Pred Value : 0.8847

Exemplo: clientes em atraso



SMOTE: 30% sim e 70% não



Confusion Matrix and Statistics

		Reference	
		0	1
Prediction	0	4189	612
	1	121	516

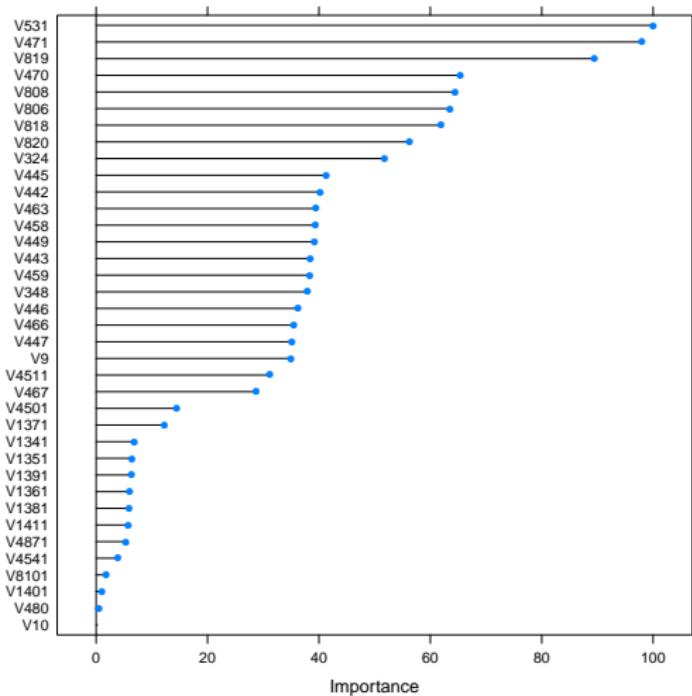
Accuracy : 0.8652
95% CI : (0.85, 0.87)

Sensitivity : 0.9719
Specificity : 0.4574
Pos Pred Value : 0.8725
Neg Pred Value : 0.8100

Exemplo: clientes em atraso



SMOTE: 50% sim e 50% não



Confusion Matrix and Statistics

		Reference	
		Prediction	0
		0	3861 368
		1	449 760

Accuracy : 0.8498

95% CI : (0.84, 0.85)

Sensitivity : 0.8958

Specificity : 0.6738

Pos Pred Value : 0.9130

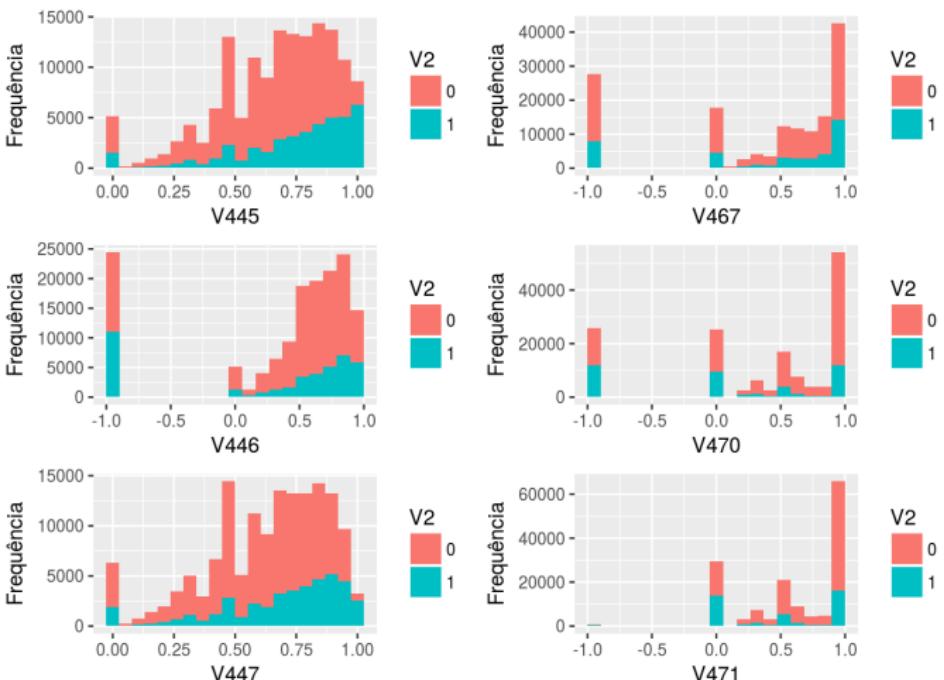
Neg Pred Value : 0.6286

- 2 **Utilizar diferentes algoritmos:** abordagens simples como Árvores, geralmente, apresentam um bom desempenho em dados desbalanceados;
- 3 **Modelos penalizados:** existem várias versões de algoritmos penalizados como *penalized-SVM* e *penalized-LDA*.
- 4 **Conceitos em outras perspectivas:** há vários campos dedicados a dados desbalanceados. P. ex., [detecção de anomalias](#), [detecção de alterações](#);
- 5 **Ser criativo:** busque inspirações, por exemplo, em respostas do Quora: "[in classification, how do you handle an unbalanced training set?](#)"
 - ★ "Decomponha a classe maior em pequenas outras classes".
 - ★ "Reamostre os dados desbalanceados em não somente um conjunto balanceado, mas vários".



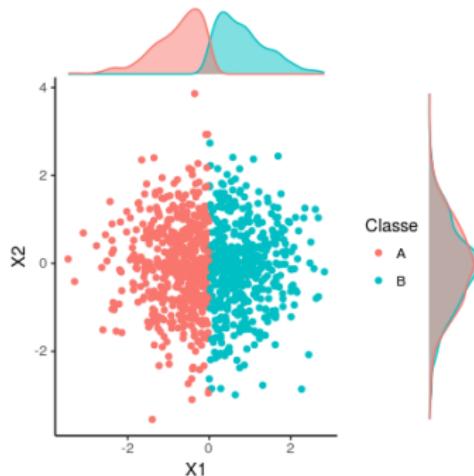
- **Feature engineering** é a arte de extrair informação dos dados já obtidos.
 - ➊ Criação de características;
 - ➋ Dados faltantes;
 - ➌ Dados desbalanceados;
 - ➍ Variáveis correlacionadas.

Variáveis correlacionadas



Ranking de características individuais

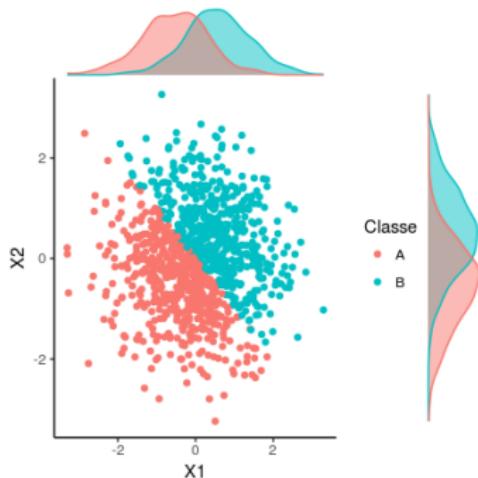
- Neste caso, a variável x_1 é relevante individualmente, e x_2 não ajuda a obter uma melhor separação.



- O ranking de características individuais funciona bem. A característica que proporciona uma boa separação de classe será escolhida.

Rotações no espaço de características

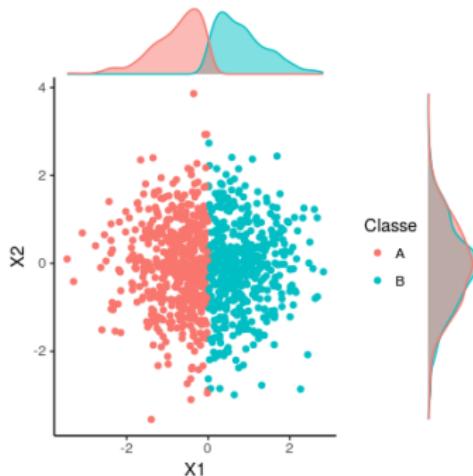
- A figura anterior foi obtida da figura abaixo após rotação de 45 graus. Agora, para alcançar a mesma separação, são necessárias x_1 e x_2 .



- Vários métodos de pré-processamento, como a análise de componentes principais (PCA), realizam transformações lineares (como a rotação).

Rotações no espaço de características

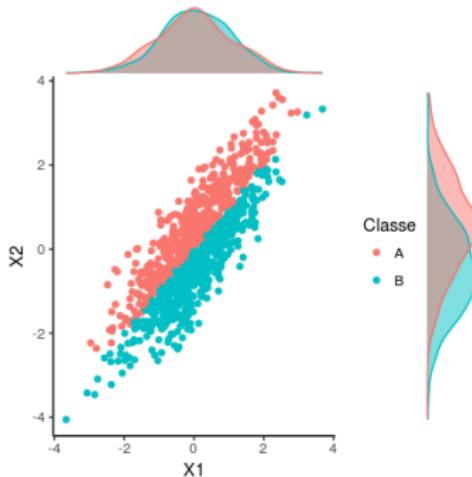
- **Pergunta:** no caso abaixo, você manteria x_1 e x_2 ou eliminaria x_2 ? Note que a noção de relevância está relacionada ao objetivo perseguido;



- $P(Y|X)$ não é independente de x_2 , mas a taxa de erro do classificador Bayes ideal é a mesma se x_2 é mantida ou descartada.

Características individualmente irrelevantes

- Nota-se uma separação linear, em que as características individualmente irrelevantes ajudam a obter uma melhor separação quando em conjunto;

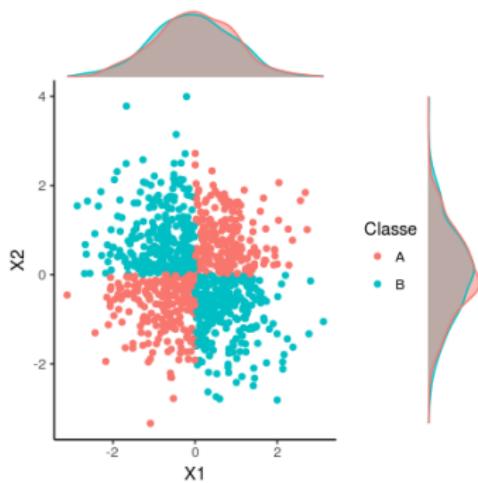


- Neste caso, é justificável o uso de métodos multivariados, que utilizam o poder preditivo das características em conjunto.

Características individualmente irrelevantes



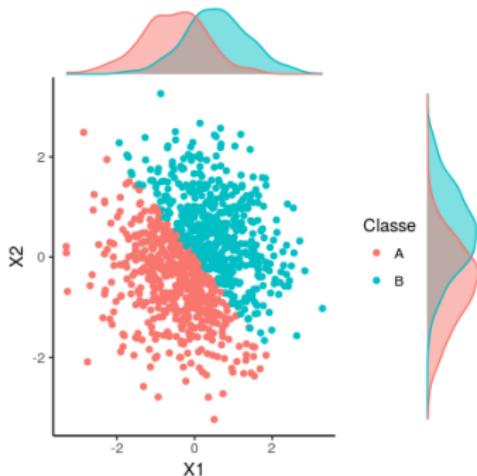
- O que podemos dizer sobre o problema abaixo, alguma característica é individualmente irrelevante?



- Este caso é conhecido como problema do tabuleiro de xadrez. As características são conjuntamente relevantes.

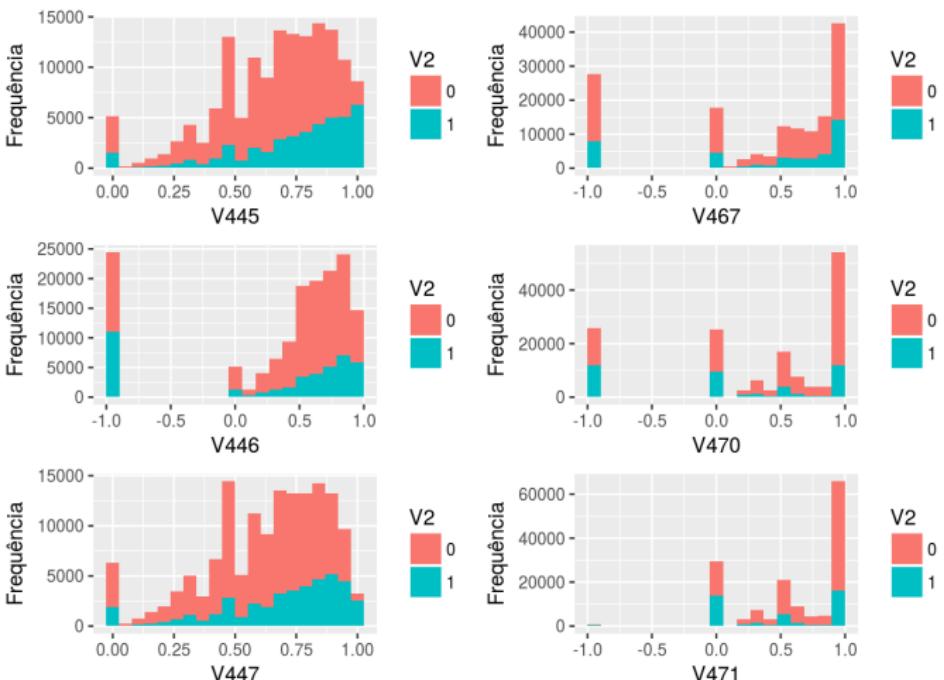
Características aparentemente redundantes

- A redução do ruído pode ser alcançada quando características com distribuições projetadas idênticas.



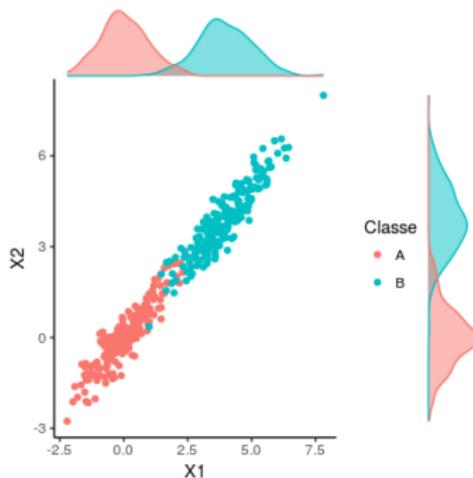
- A distribuição bidimensional mostra uma separação de classe melhor, quando comparado com qualquer característica individual.

Variáveis correlacionadas



Correlação não implica redundância

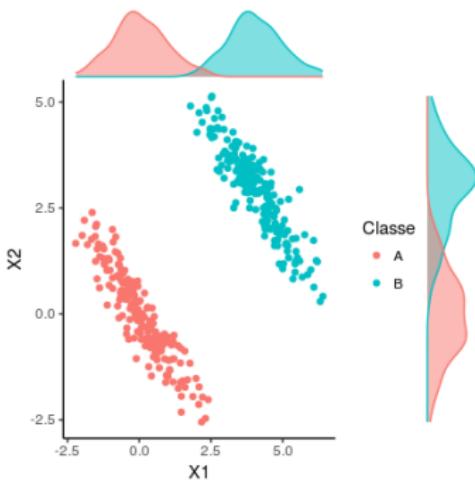
- Geralmente, se pensa que correlação de característica significa redundância de recurso.



- As características são redundantes. I. e., a separação de classe não é aprimorada ao se considerar as variáveis conjuntamente.

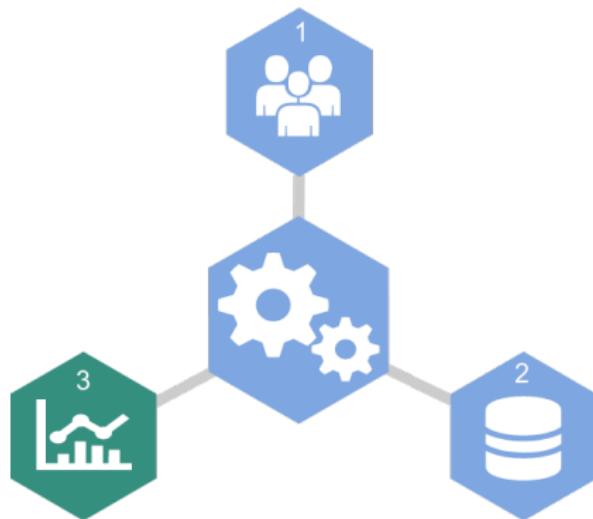
Correlação não implica redundância

- Geralmente, se pensa que correlação de característica significa redundância de recurso.



- Apesar das projeções serem semelhantes à anterior (e correlacionadas), elas não são redundantes.

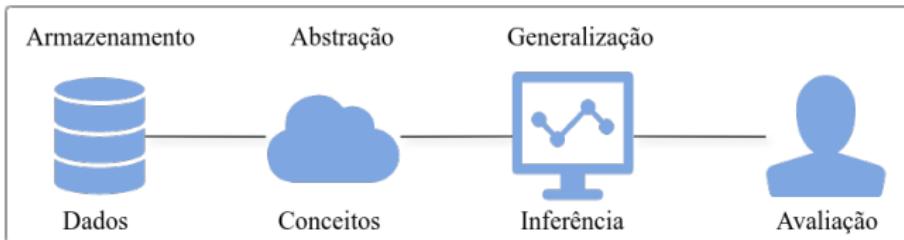
Objetivos



Modelos

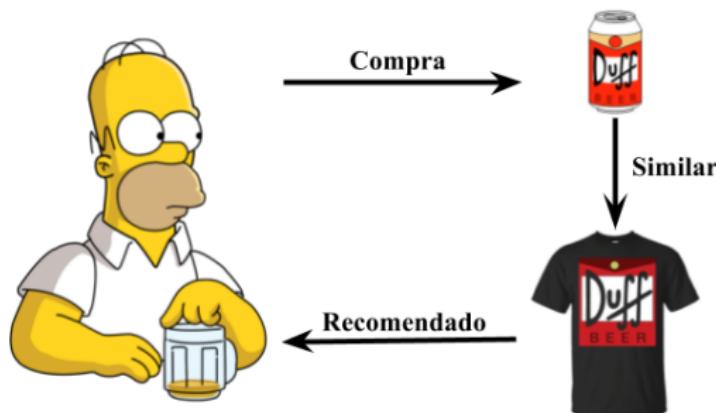
Base de dados

Como as máquinas aprendem?



- **Armazenamento dos dados:** utiliza a observação para fornecer uma base para o raciocínio adicional;
- **Abstração:** envolve a tradução dos dados em representações e conceitos;
- **Generalização:** cria conhecimento e inferência que direcionam ações em novos contextos;
- **Avaliação:** fornece um mecanismo de *feedback* para medir a utilidade do conhecimento adquirido e informar potenciais melhorias.

Exemplo



Todos os aspectos são importantes



Dados bons



+

Modelo ruim



=

Resultado ruim



Dados ruins



+

Modelo perfeito



=

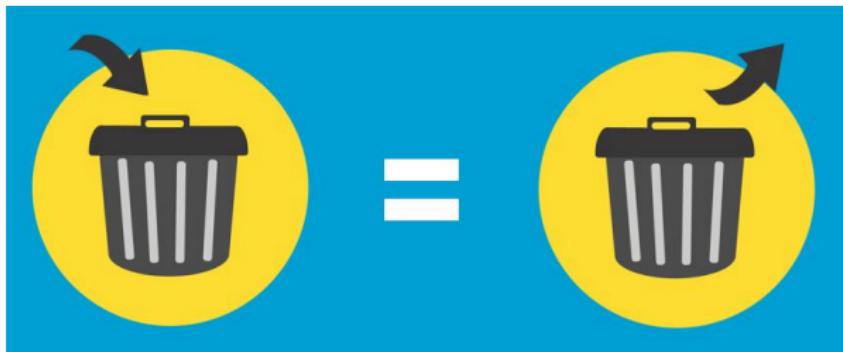
Resultado ruim



Os limites do Machine Learning



- Machine Learning tem pouca flexibilidade para extrapolar os parâmetros de aprendizagem e não conhece o senso comum!

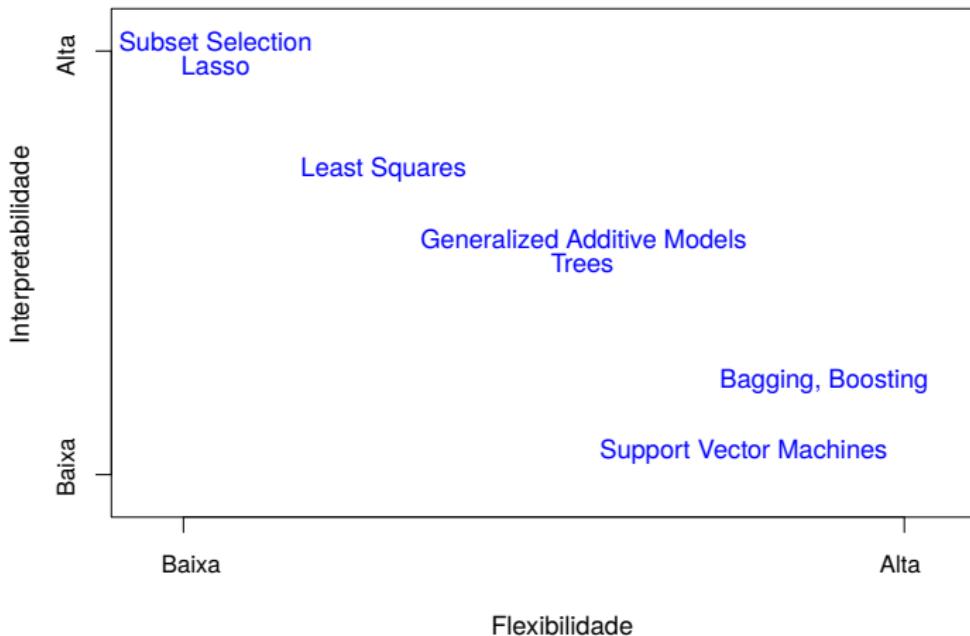


- Ele é tão bom quanto os dados são para ensinar. É um paradigma “Garbage in, garbage out!”

Os limites do Machine Learning



Interpretabilidade vs Flexibilidade



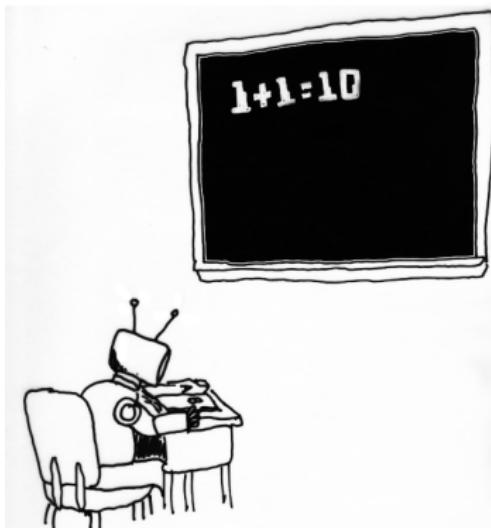
Tipos de aprendizado



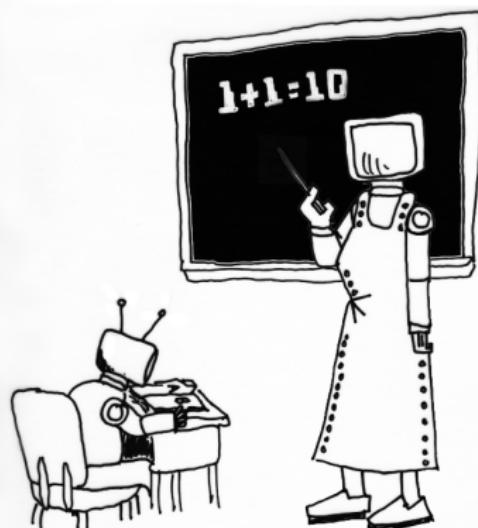
Tipos de aprendizado



UNSUPERVISED MACHINE LEARNING



SUPERVISED MACHINE LEARNING



Fonte: Proofreader's Whimsy

- James, G., Witten, D., Hastie, T. e Tibshirani, An Introduction to Statistical Learning, 2013;
- Hastie, T., Tibshirani, R. e Friedman, J., The Elements of Statistical Learning, 2009;
- Lantz, B., Machine Learning with R, Packt Publishing, 2013;
- Tan, Steinbach, and Kumar, Introduction to Data Mining, Addison-Wesley, 2005;
- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani