Processos de ML

- COMPLEXOS
- ALTO CUSTO HUMANO
- ALTO CUSTO COMPUTACIONAL
- RISCOS GRANDES

Auto ML

- Automatizar o processo
- Reduzir a interferência humana
- Melhor a performance computacional
- Melhorar a performance dos modelos

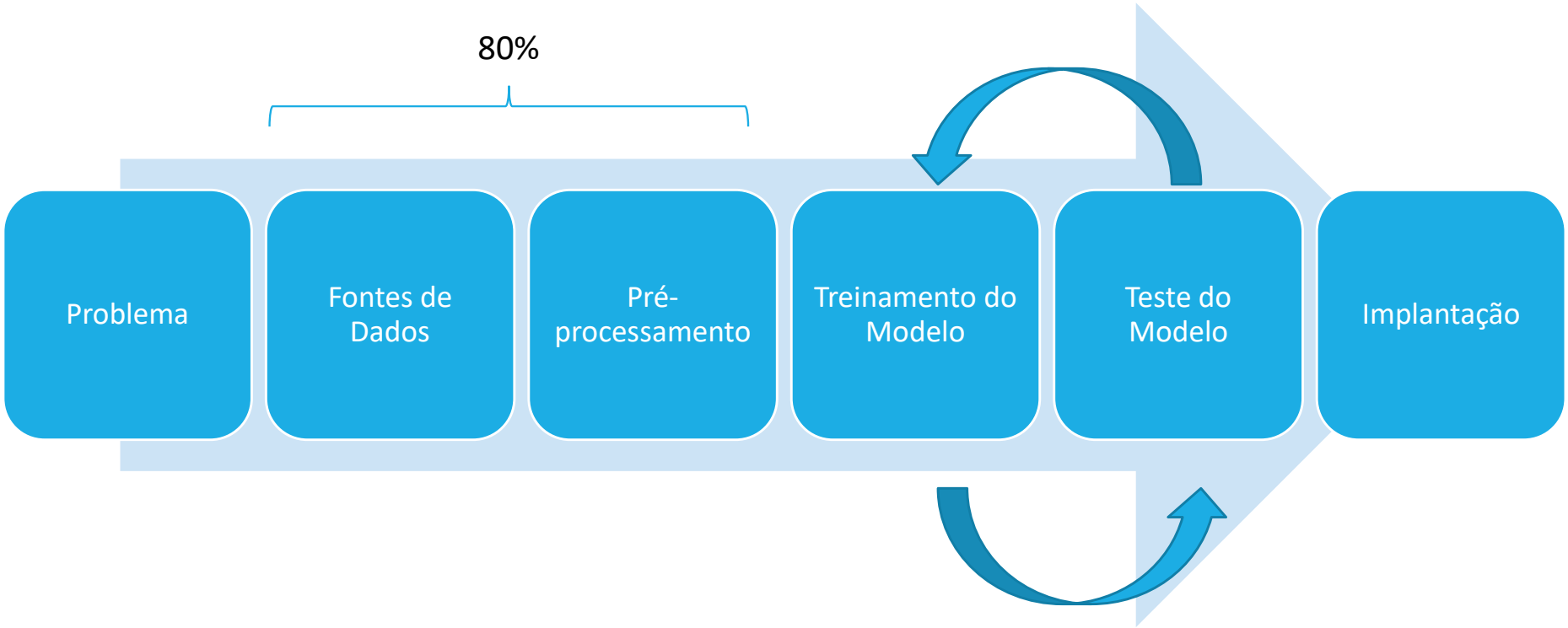
Auto ML: Por que?

- **Modelos de Machine Learning mais eficientes**
- **Produtos Orientados a Dados**
- **Participar de Competições de Machine Learning**

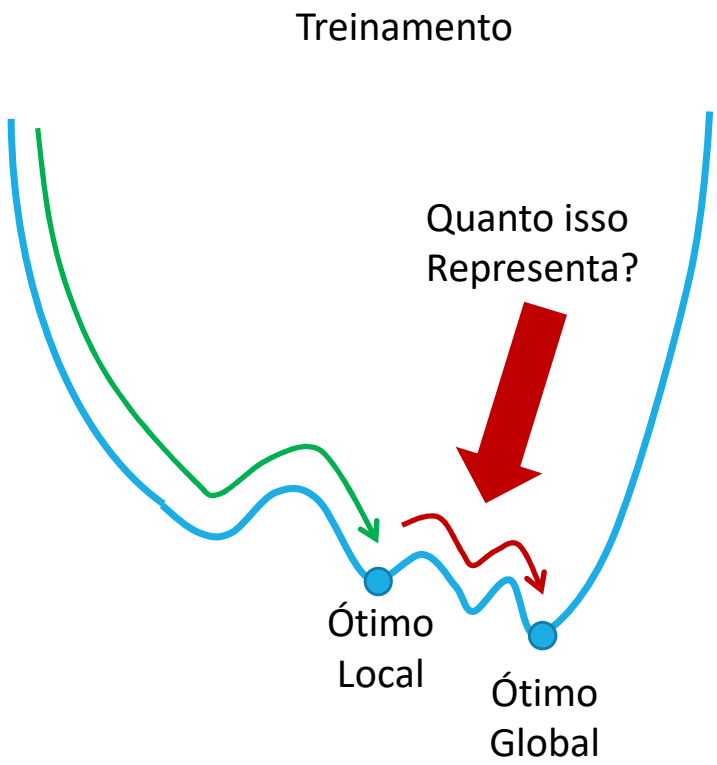
Auto ML: Tunning!

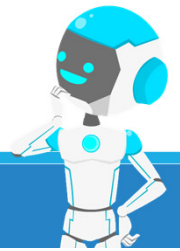
- Não se trata apenas de automatizar, mas automatizar buscando melhor performance
- O objetivo é que o Auto ML consiga performance melhor do que um humano

Processo de ML



Treinamento





Simulando. Identificar Fraudes em Varejo

10 Mil transações por dia

Valor médio R$ 100,00

Falsos Positivos: 1%

Falsos Negativos: 0,5%

Perda de Negócios

Perda de Capital

	Falsos Positivos	Falsos Negativos	Total
Dia	10.000	5.000	15.000
Mês	300.000	150.000	**450.000**

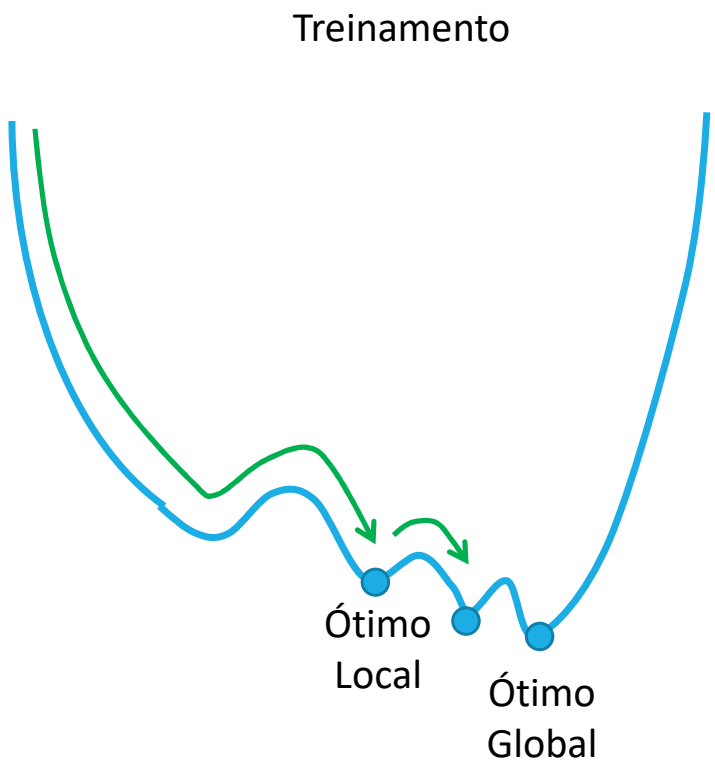
Restrições

Performance do modelo

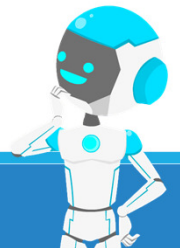
Tempo de treinamento

Recursos computacionais (memória, CPU)

Treinamento



Melhora nos Falsos Positivos em 0,5 %



Simulando...

10 Mil transações por dia

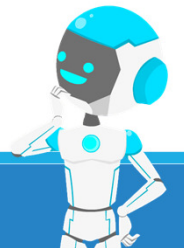
Valor médio R$ 100,00

Falsos Positivos: 0,5%

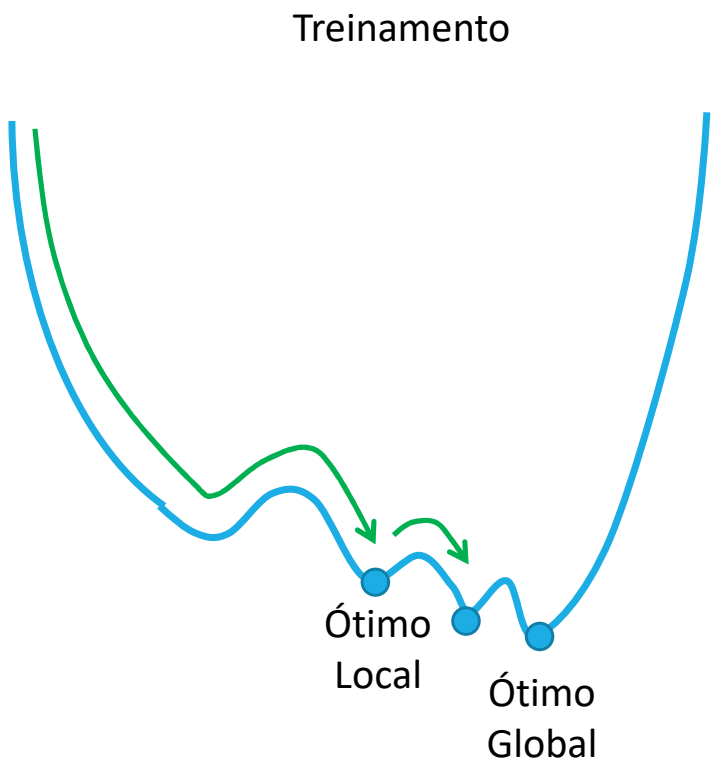
Falsos Negativos: 0,5%

	Falsos Positivos	Falsos Negativos	Total
Dia	10.000	5.000	15.000
Mês	300.000	150.000	450.000

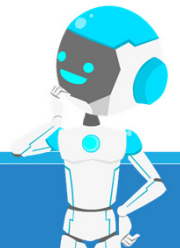
	Falsos Positivos	Falsos Negativos	Total
Dia	5.000	5.000	10.000
Mês	150.000	150.000	300.000



Treinamento



Melhora nos Falsos Positivos em 0,5 %



Classificadores

Batch size [10-1000]

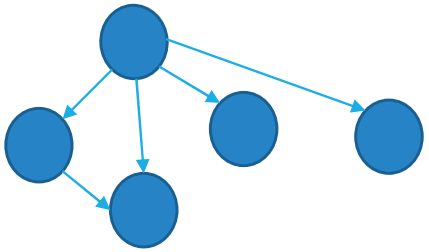
Estimator [4]

Search Algorithm [8]

Use ADTree [true/false]

Rede Bayesiana

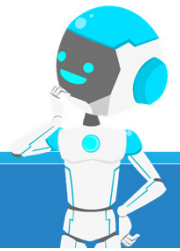
Grafo de Dependências



Tabelas de Distribuição de Probabilidade

Probability Distribution Table For property_magnitude

class	credit_amount	duration	other_parties	'real estate'	'life insura...	car	'no known ...
good	'(-inf-3913.5]'	'(-inf-15.5]'	none	0,422	0,224	0,261	0,094
good	'(-inf-3913.5]'	'(-inf-15.5]'	'co applicant'	0,423	0,346	0,192	0,038
good	'(-inf-3913.5]'	'(-inf-15.5]'	guarantor	0,722	0,241	0,019	0,019
good	'(-inf-3913.5]'	'(15.5-inf)'	none	0,229	0,229	0,437	0,104
good	'(-inf-3913.5]'	'(15.5-inf)'	'co applicant'	0,278	0,167	0,167	0,389
good	'(-inf-3913.5]'	'(15.5-inf)'	guarantor	0,536	0,179	0,179	0,107
good	'(3913.5-inf)'	'(-inf-15.5]'	none	0,219	0,219	0,281	0,281
good	'(3913.5-inf)'	'(-inf-15.5]'	'co applicant'	0,25	0,25	0,25	0,25
good	'(3913.5-inf)'	'(-inf-15.5]'	guarantor	0,167	0,5	0,167	0,167
good	'(3913.5-inf)'	'(15.5-inf)'	none	0,106	0,24	0,421	0,232
good	'(3913.5-inf)'	'(15.5-inf)'	'co applicant'	0,214	0,071	0,357	0,357
good	'(3913.5-inf)'	'(15.5-inf)'	guarantor	0,417	0,417	0,083	0,083
bad	'(-inf-3913.5]'	'(-inf-15.5]'	none	0,257	0,27	0,336	0,138
bad	'(-inf-3913.5]'	'(-inf-15.5]'	'co applicant'	0,375	0,125	0,375	0,125
bad	'(-inf-3913.5]'	'(-inf-15.5]'	guarantor	0,167	0,5	0,167	0,167
bad	'(-inf-3913.5]'	'(15.5-inf)'	none	0,272	0,203	0,361	0,163
bad	'(-inf-3913.5]'	'(15.5-inf)'	'co applicant'	0,389	0,167	0,167	0,278
bad	'(-inf-3913.5]'	'(15.5-inf)'	guarantor	0,188	0,438	0,188	0,188
bad	'(3913.5-inf)'	'(-inf-15.5]'	none	0,038	0,577	0,038	0,346
bad	'(3913.5-inf)'	'(-inf-15.5]'	'co applicant'	0,167	0,5	0,167	0,167
bad	'(3913.5-inf)'	'(-inf-15.5]'	guarantor	0,25	0,25	0,25	0,25
bad	'(3913.5-inf)'	'(15.5-inf)'	none	0,083	0,194	0,383	0,339
bad	'(3913.5-inf)'	'(15.5-inf)'	'co applicant'	0,05	0,15	0,35	0,45
bad	'(3913.5-inf)'	'(15.5-inf)'	guarantor	0,5	0,1	0,3	0,1



Parâmetros

Hiper parâmetros

Batch size [10-1000]

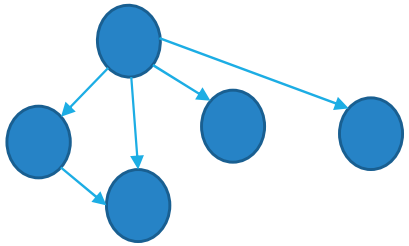
Estimator [4]

Search Algorithm [8]

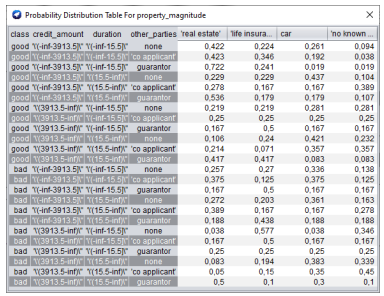
ADTree [true/false

Rede Bayesiana

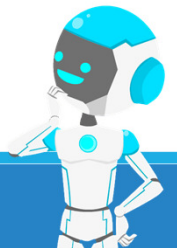
Grafo de Dependência



Tabelas de Distribuição de Probabilidade



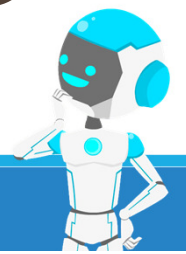
class	credit_amount	duration	other_parties	'real estate'	'life insura...	car	'no known ...
good	'(-inf-3913.5]'	'(-inf-15.5]'	none	0,422	0,224	0,261	0,094
good	'(-inf-3913.5]'	'(-inf-15.5]'	'co applicant'	0,423	0,346	0,192	0,038
good	'(-inf-3913.5]'	'(-inf-15.5]'	guarantor	0,722	0,241	0,019	0,019
good	'(-inf-3913.5]'	'(15.5-inf)'	none	0,229	0,229	0,437	0,104
good	'(-inf-3913.5]'	'(15.5-inf)'	'co applicant'	0,278	0,167	0,167	0,389
good	'(-inf-3913.5]'	'(15.5-inf)'	guarantor	0,536	0,179	0,179	0,107
good	'(3913.5-inf)'	'(-inf-15.5]'	none	0,219	0,219	0,281	0,281
good	'(3913.5-inf)'	'(-inf-15.5]'	'co applicant'	0,25	0,25	0,25	0,25
good	'(3913.5-inf)'	'(-inf-15.5]'	guarantor	0,167	0,5	0,167	0,167
good	'(3913.5-inf)'	'(15.5-inf)'	none	0,106	0,24	0,421	0,232
good	'(3913.5-inf)'	'(15.5-inf)'	'co applicant'	0,214	0,071	0,357	0,357
good	'(3913.5-inf)'	'(15.5-inf)'	guarantor	0,417	0,417	0,083	0,083
bad	'(-inf-3913.5]'	'(-inf-15.5]'	none	0,257	0,27	0,336	0,138
bad	'(-inf-3913.5]'	'(-inf-15.5]'	'co applicant'	0,375	0,125	0,375	0,125
bad	'(-inf-3913.5]'	'(-inf-15.5]'	guarantor	0,167	0,5	0,167	0,167
bad	'(-inf-3913.5]'	'(15.5-inf)'	none	0,272	0,203	0,361	0,163
bad	'(-inf-3913.5]'	'(15.5-inf)'	'co applicant'	0,389	0,167	0,167	0,278
bad	'(-inf-3913.5]'	'(15.5-inf)'	guarantor	0,188	0,438	0,188	0,188
bad	'(3913.5-inf)'	'(-inf-15.5]'	none	0,038	0,577	0,038	0,346
bad	'(3913.5-inf)'	'(-inf-15.5]'	'co applicant'	0,167	0,5	0,167	0,167
bad	'(3913.5-inf)'	'(-inf-15.5]'	guarantor	0,25	0,25	0,25	0,25
bad	'(3913.5-inf)'	'(15.5-inf)'	none	0,083	0,194	0,383	0,339
bad	'(3913.5-inf)'	'(15.5-inf)'	'co applicant'	0,05	0,15	0,35	0,45
bad	'(3913.5-inf)'	'(15.5-inf)'	guarantor	0,5	0,1	0,3	0,1



Diferenças

Hiper parâmetros são normalmente configurados pelos implementadores do classificador, antes do processo de treino. Ex: Cientista de Dados

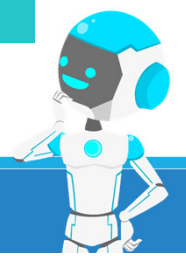
Parâmetros são configurados pelo algoritmo, durante o processo de treino



Hiper Parâmetros

Hiper Parâmetros de Modelo: Interferem na performance do modelo

Hiper Parâmetros de Algoritmo: Não interfere na performance do modelo, mas do processo de aprendizado



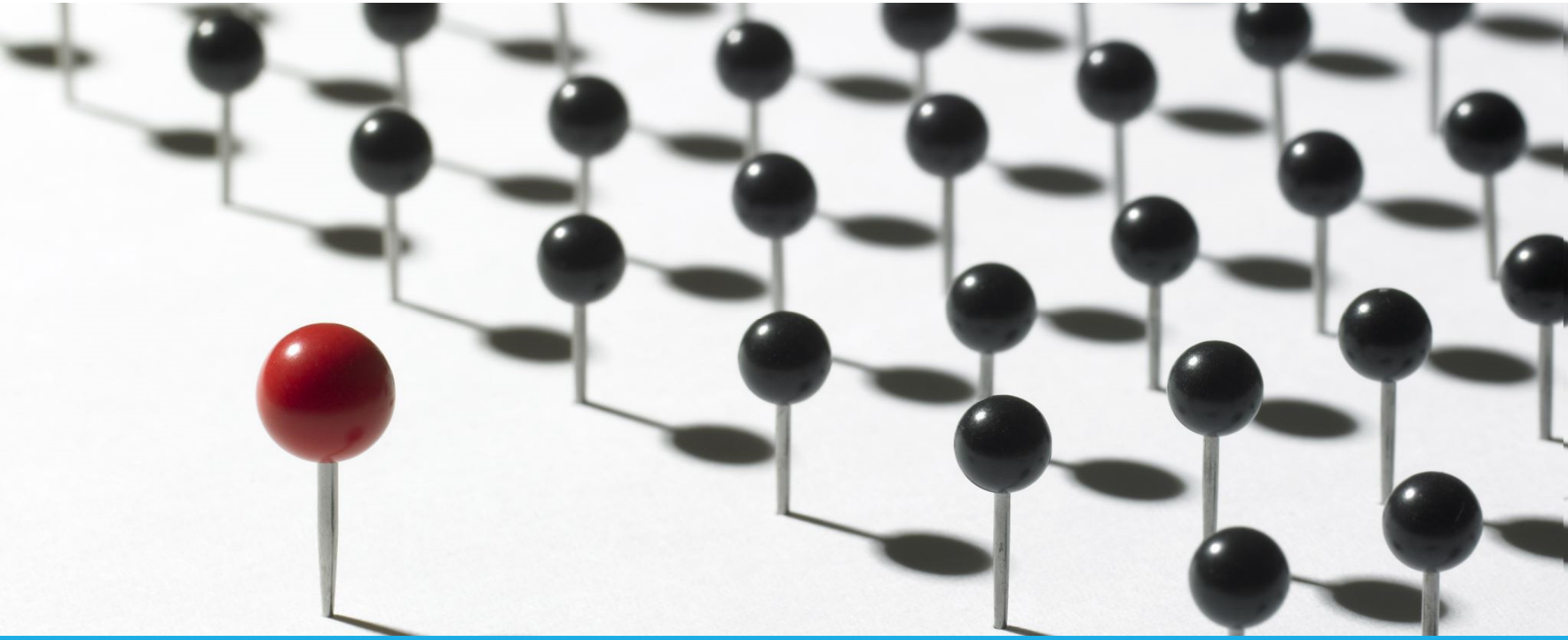
Domínio

Inteiros: Epochs

Valores Reais: Learning Rate

Binários: Normalizar Atributos

Categóricos: Estimador



Hiper parâmetros condicionais

A escolha de um depende ou invalida outro

Um método de busca depende do avaliador de atributo

Parâmetros

Hiper parâmetros

Batch size [10-1000]

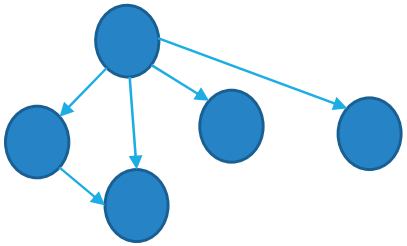
Estimator [4]

Search Algorithm [8]

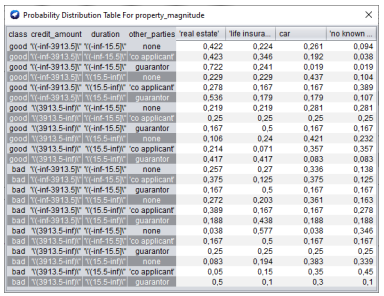
ADTree [true/false

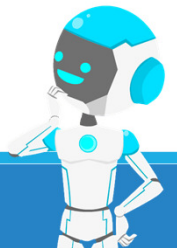
Rede Bayesiana

Grafo de Dependência



Tabelas de Distribuição de Probabilidade





Conclusão:

A definição ótima dos hiper parâmetros é vital para a performance do modelo!

Mas quais valores para escolher para os hiper parâmetros?

E será que o classificador que eu estou usando é o melhor?

Hiper Parâmetros

Existem boas práticas, regras gerais, valores default

Exemplo, para topologia de uma RNA:

$$t = \frac{a + c}{2}$$

Porém:

- O número de configurações é muito grande
- O custo computacional é muito alto