

Data Mining





Tema da Aula: Estudo de caso com a DataRisk

Coordenação:

**Prof. Dr. Adolpho Walter
Pimazzi Canton**

**Profa. Dra. Alessandra de
Ávila Montini**

Prof. Jhonata Emerick Ramos
jer@datarisk.io

Prof. Bruno de Paula Jacóia
[bruno.jacoia@datarisk.io](mailto;bruno.jacoia@datarisk.io)

Curriculum - Jhonata Emerick



Engenheiro aeronáutico pela Universidade de São Paulo (USP) com mestrado em Finanças Quantitativas pela Fundação Getúlio Vargas (FGV), iniciou sua carreira na Embraer e depois migrou para o mercado financeiro. Foi analista da Rio Bravo investimentos e do grupo Ambipar, além de Data Scientist no Itaú. Em outubro de 2014 fundou uma startup de logística urbana chamada 99motos que, em 2016 se fundiu com o Rapiddo e foi vendida ao iFood em setembro de 2018. Atualmente prepara-se para defender sua tese de doutorado na Poli-USP em IA. É co-fundador de duas startups com foco em IA: DataRisk (área de risco) e RadSquare(área médica).

Curriculum - Bruno Jacóia



Bacharel em Matemática pela Universidade de São Paulo (USP) com mestrado em Matemática Aplicada e doutorado em Matemática na área de Sistemas Dinâmicos pela USP. Atualmente é Data Scientist na DataRisk.

Objetivos da aula

- Apresentação da plataforma da DataRisk.
- Conceitos de *Machine Learning*.
- Como criar um bom modelo de *Machine Learning*.
- Exemplos de alguns casos.
- Resolução de problemas usando a plataforma.

Plataforma



datarisk.io

Problema



- Fluxo tradicional muito manual.
- Informações muito fragmentadas e dispersas.
- Tempo dedicado ao processo de análise e tratamento dos dados gera dificuldade de escalar o processo.
- *Time to market.*

Solução

Uma solução baseada em anos de pesquisa, com uma plataforma web e API, para criar modelos de *machine learning* e processamento de imagem que tem uma clara proposta de valor: ser rápido, seguro, escalável e fácil de usar.

The image displays two screenshots of the DataNok platform. The top screenshot shows the OCR (Optical Character Recognition) feature, where a document image is processed to extract data such as Name, RG number, Orgão Emissor RG, CPF number, Birth date, and Father name. The bottom screenshot shows the modeling dashboard, which includes tabs for ESTATÍSTICAS, PERFORMANCE, IMPORTÂNCIA DAS VARIÁVEIS, SCORE CARD, and ESTABILIDADE. The ESTATÍSTICAS tab is active, showing an ROC curve (Curva ROC) and a density histogram (Densidade).

Solução

Mining

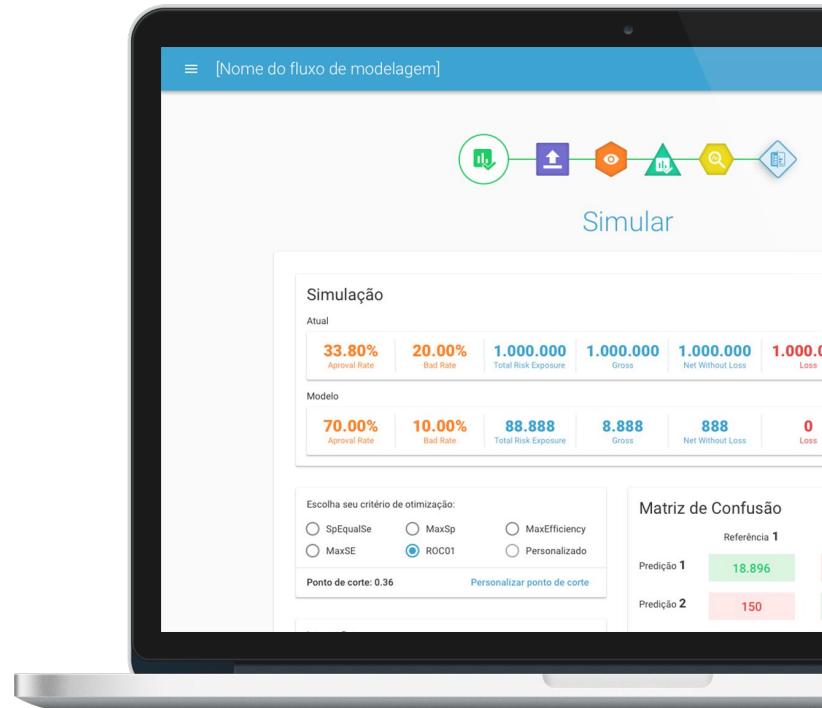
- Tratamento de *missings* e *outliers*
- Seleção de variáveis
- Otimizações
- Modelos e Relatórios

Scoring

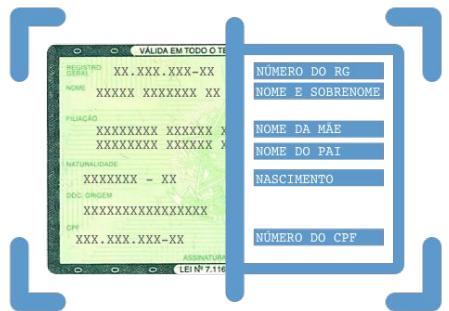
- Sistema ou API
- Batch ou online

Validação

- Avaliação dos modelos
- Relatórios periódicos da performance do modelo



Solução



OCR



Face Matching



Verificação de Assinatura

Fundadores



Jhonata Emerick Ramos

Co-founder & CEO



Gustavo Di Giovanni Bernardo

Co-founder & CTO

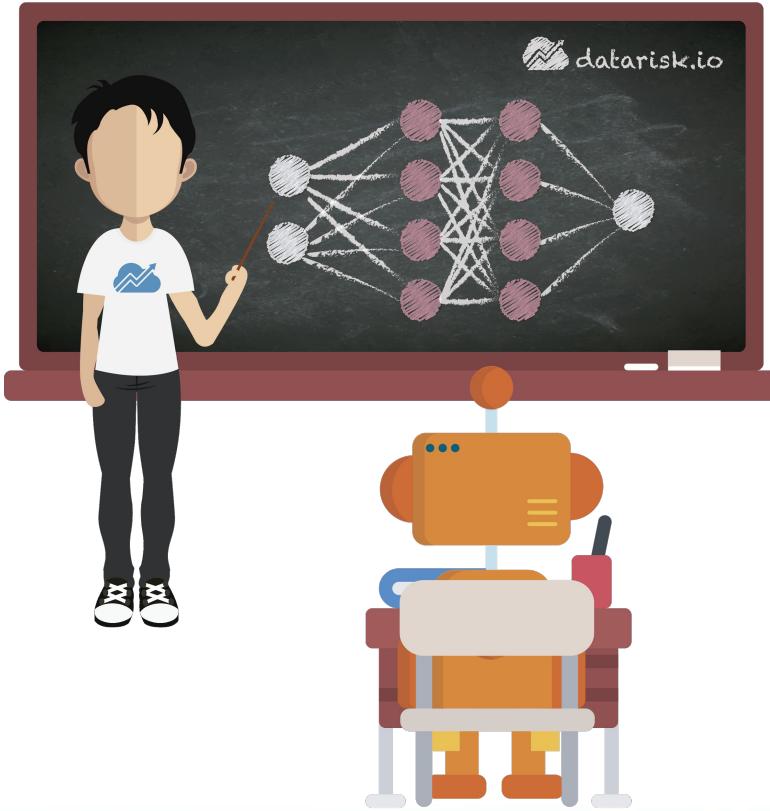


Carlos Eduardo M. Relvas

Co-founder & CDO



O que é Machine Learning?



Machine Learning

“Campo de estudo que fornece aos computadores a habilidade de aprender sem serem explicitamente programados”

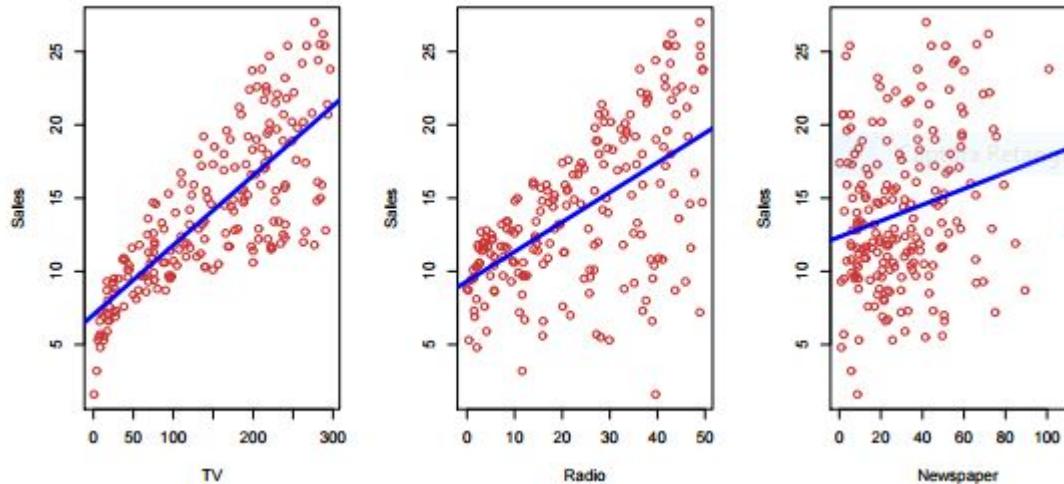
Arthur Samuel

“Um programa de computador aprende com experiência E com respeito a algumas tarefas T e medida de performance P, se sua performance para resolver as tarefas em T, medidas por P, melhora com a experiência E”

Tom M. Mitchell

Machine Learning

Prever quanto irei vender em função do quanto gastei em propaganda na TV, Rádio e Jornal.



$$\text{Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

Machine Learning

Classificar e-mails enviados para um mesmo indivíduo como Spam ou não Spam.

Variáveis: porcentagem em que as 50 palavras ou pontuações mais frequentes aparecem em cada e-mail.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Essência

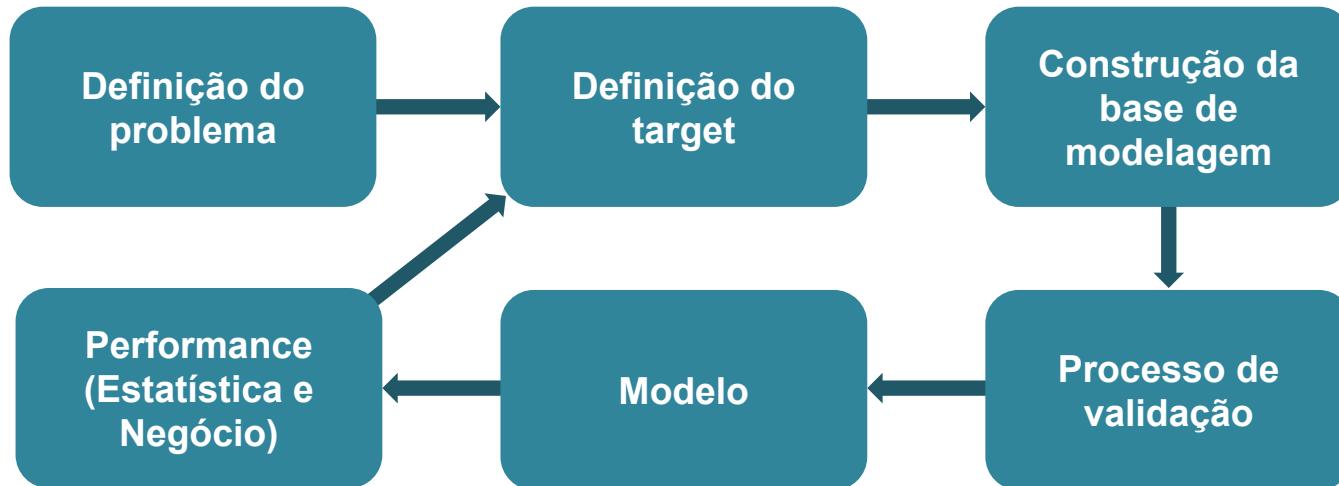
Existe um padrão?

Temos dados para explicar esse padrão?

Podemos expressá-lo matematicamente?



O que é preciso para criar bons modelos de Machine Learning?



Definição do problema

- Qual é o problema? Descreva o problema e todas as suposições que assumimos.
- Por que esse problema precisa ser resolvido? Quais são as motivações para resolver o problema, os benefícios que a solução trás e como a solução será usada.
- Como eu resolveria o problema?

Definição do problema

Exemplos:

- **Medir o risco de inadimplência de um cliente.**
- **Aumentar a taxa de conversão de uma campanha.**
- **Otimizar o retorno do valor cobrado de pessoas inadimplentes por telefone.**

Definição do target

Depois de definido o problema que queremos solucionar, temos a etapa de definição do target (ou variável resposta).

ESTE PASSO É EXTREMAMENTE IMPORTANTE NO PROCESSO!

Definição do target

Exemplos:

- O que representa o risco de inadimplência? Quanto depois depois do não pagamento é ruim? Cinco dias de atraso é OK? E dez dias de atraso? Quantos meses de observação olhar?
- Como aumentar a taxa de conversão de uma campanha? Envio e-mails para clientes que tem maior chance de engajar na campanha?
- Como otimizar o retorno de valor cobrado por telefone? Priorizando clientes mais propenso a atender a ligação? Devo priorizar algum horário do dia?

Construção da base de modelagem

- Devo usar todas os dados e variáveis disponíveis? Será que o signo de uma pessoa influencia na sua probabilidade de inadimplência?
- Todas as variáveis devem estar disponíveis no momento de usar o modelo para tomada de decisão. Ex: um modelo que irá decidir a aprovação ou não de crédito para um cliente só poderá usar variáveis disponíveis no momento da aplicação, não importa se a renda mensal atual é de 5000 reais se no momento da aprovação a renda era de 2000 reais.
- Uma arquitetura de bancos de dados apropriada ajuda na obtenção das variáveis históricas.

Construção da base de modelagem

Exemplo: Prever o quanto cada cliente irá gastar no próximo mês.

ID_CLIENTE	DATA	TARGET
1	20/01/2019	500,00
2	20/01/2019	450,00
3	20/01/2019	700,00
4	20/01/2019	300,00

Observado em
20/02/2019

Construção da base de modelagem

ID_CLIENTE	MERCHANT	RAMO	DATA	VALOR
1	XXX	Posto	10/01/2019	100,00
1	PPP	Posto	26/01/2019	150,00
1	RRR	Farmácia	07/01/2019	180,00
2	QQQ	Padaria	04/01/2019	80,00
2	WWW	Posto	09/01/2019	120,00
2	YYY	Padaria	30/01/2019	90,00
3	RRR	Farmácia	07/01/2019	50,00
3	OOO	Farmácia	08/01/2019	200,00
4	XXX	Posto	25/01/2019	180,00

Construção da base de modelagem

ID_CLIENTE	MERCHANT	RAMO	DATA	VALOR
1	XXX	Posto	10/01/2019	100,00
1	PPP	Posto	26/01/2019	150,00
1	RRR	Farmácia	07/01/2019	180,00
2	QQQ	Padaria	04/01/2019	80,00
2	WWW	Posto	09/01/2019	120,00
2	YYY	Padaria	30/01/2019	90,00
3	RRR	Farmácia	07/01/2019	50,00
3	OOO	Farmácia	08/01/2019	200,00
4	XXX	Posto	25/01/2019	180,00

ID_CLIE NTE	DATA	TARGET	# de compras 7 dias	# de compras 30 dias	R\$ de compras 7 dias	R\$ de compras 30 dias	Med R\$ de compras 30 dias
1	20/01/2019	500,00	0	2	0	280,00	140,00
2	20/01/2019	450,00	0	2	0	200,00	100,00
3	20/01/2019	700,00	1	2	200,00	250,00	125,00
4	20/01/2019	300,00	0	0	0	0	0

ID_CLIE NTE	DATA	TARGET	Var R\$ de compras 30 dias	# de compras 30 dias Posto	R\$ de compras 30 dias Posto	# de compras 30 dias Farmácia	...
1	20/01/2019	500,00	xx	1	100,00	1	...
2	20/01/2019	450,00	yy	1	120,00	0	...
3	20/01/2019	700,00	zz	0	0	2	...
4	20/01/2019	300,00	ww	0	0	0	...

Processo de validação

Esta etapa é muito importante. Graças a ela que conseguimos testar qual será o efeito do modelo na prática. Queremos garantir que o nosso modelo terá o mesmo comportamento quando utilizarmos dados novos.

Algumas das estratégias mais comuns de validação envolvem separar a base de dados em treino e teste ou em treino, validação e teste.

Performance

A melhor estratégia é definir inicialmente qual métrica estatística será utilizada para verificar a performance do modelo.

Da mesma forma para as métricas de negócio. O modelo soluciona o problema inicial?

Modelo



Análise Descritiva

- Distribuição e correlação de cada variável com o target.
- Quantidade de valores de missings, valores estranhos, integridade dos dados.
- Bugs no processo de construção dos dados.

Missing e Outliers

- O que fazer com valores missing (Ex: nulo, strings vazias)?
Algumas técnicas de machine learning são capazes de lidar com isso automaticamente, mas a maioria não.
- Outliers podem impactar significativamente o modelo. O que fazemos com relação a isso? Esse pode ser um risco para o modelo?

Criação de novas variáveis

- Com as análises ou conhecimento do negócio podemos concluir que duas variáveis podem ser combinadas ou que podemos transformar aquela informação de alguma maneira.
- Essa etapa tende a ser ignorada por muitas pessoas , embora ela possa produzir grandes melhorias no desempenho do modelo.

Feature Engineering

- Qual a melhor forma de tratar cada variável?
- Categorizar é uma boa alternativa?
- Como normalizar?

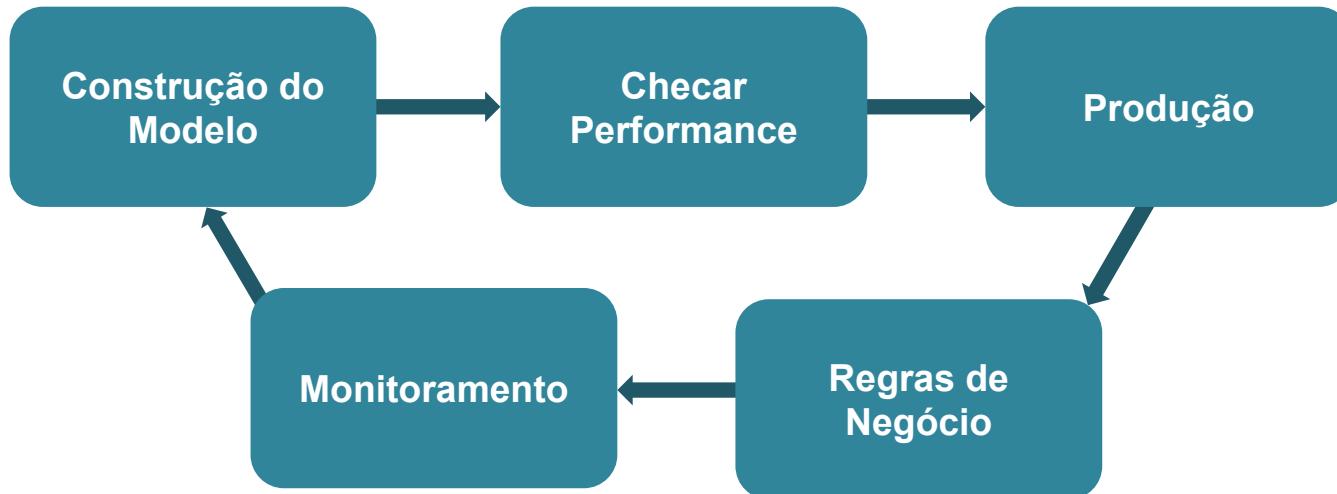
Feature Selection

- Tenho centenas de variáveis. Devo usar todas? Quais são as mais importantes?
- Alguns algoritmos se beneficiam de mais variáveis, outros podem apresentar problemas (Ex: instabilidade, multicolinearidade, tempo de execução, memória).
- Qual abordagem usar para escolher o melhor subconjunto de variáveis?

Modelo

- Qual técnica de modelagem usar (Ex: Linear, Random Forest, Boosting, Redes Neurais)?
- Trade-off: preciso interpretar o meu algoritmo (Ex: Linear), ou posso usar um algoritmo mais poderoso (Ex: Rede Neural)?
- Há alguma restrição regulatória (Ex: O Banco Central exige interpretação)?

Implementação



Automatização do processo



Automatização do processo

normalize_adult2_32.data.data					
<input checked="" type="checkbox"/>	Variável	Tipo	Linha 01	Linha 02	Linha 03
<input checked="" type="checkbox"/>	ID_DATARISK	Numérica ▾	1	2	3
<input checked="" type="checkbox"/>	X1	Numérica ▾	39	50	38
<input checked="" type="checkbox"/>	X2	Categórica ▾	STATE-GOV	SELF-EMP-NOT-INC	PRIVATE
<input checked="" type="checkbox"/>	X3	Numérica ▾	77516	83311	215646
<input checked="" type="checkbox"/>	X4	Categórica ▾	BACHELORS	BACHELORS	HS-GRAD
<input checked="" type="checkbox"/>	X5	Numérica ▾	13	13	9
<input checked="" type="checkbox"/>	X6	Categórica ▾	NEVER-MARRIED	MARRIED-CIV-SPOUS	DIVORCED
<input checked="" type="checkbox"/>	X7	Categórica ▾	ADM-CLERICAL	EXEC-MANAGERIAL	HANDLERS-CLEANER
<input checked="" type="checkbox"/>	X8	Categórica ▾	NOT-IN-FAMILY	HUSBAND	NOT-IN-FAMILY
<input checked="" type="checkbox"/>	X9	Categórica ▾	WHITE	WHITE	WHITE
Linhas por página <input type="button" value="10"/> 1-10 de 16					

Automatização do processo

Tipo de modelo *

Varejo

Escolha o tipo de modelo que deseja validar

Método do modelo *

GLM

Ensemble Model

Variável chave *

ID_DATARISK

Escolha o identificador único da sua tabela.

Variável resposta *

X15

Escolha uma variável binária que guiará o treinamento

VARIÁVEIS SELECIONADAS VARIÁVEIS NÃO SELECIONADAS

✓ ID_DATARISK ✓ X1 ✓ X2
✓ X3 ✓ X4 ✓ X5
✓ X6 ✓ X7
✓ X9 ✓ X10
✓ X12 ✓ X13
✓ X15

Treinar

Seu modelo está sendo treinado. Você pode acompanhar aqui as etapas do treinamento ou continuar suas tarefas no sistema.

1 Tratamento
Pré-seleção, tratamento de missings e outliers.

2 Seleção de variáveis
Análise de correlações e seleção de variáveis.

3 Ottimização
Refinamento e significância do modelo.

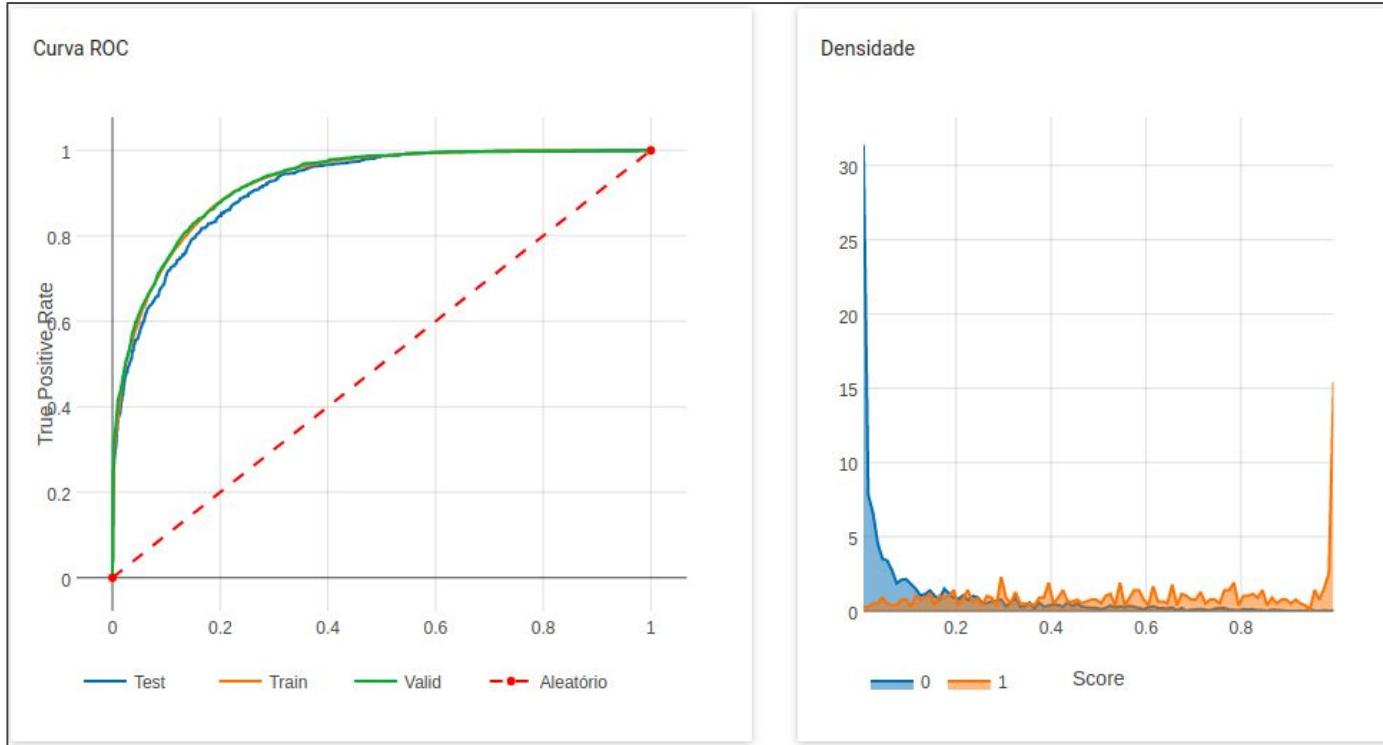
4 Modelo Final
Geração do modelo final e dos relatórios.

CANCELAR TREINAMENTO

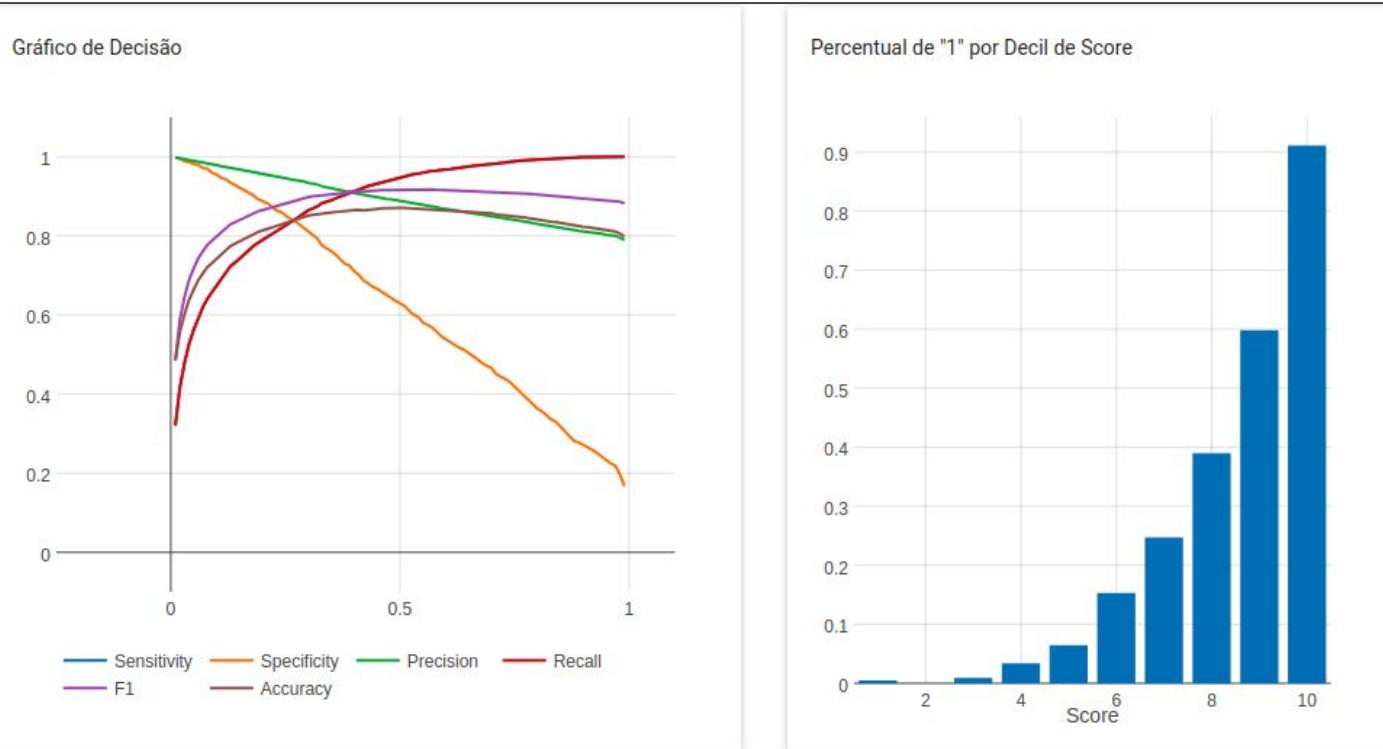
Automatização do processo

ESTATÍSTICAS	PERFORMANCE	IMPORTÂNCIA DAS VARIÁVEIS	SCORE CARD																								
<p>Modelo de varejo</p> <p>Técnica: GLM</p> <p>Tempo de processamento: 17m 43s</p> <p>Linhas: 32561</p> <p>Variáveis totais: 14</p> <p>Variáveis finais: 13</p>	<p>Amostragem</p> <table><thead><tr><th></th><th>Perc.</th></tr></thead><tbody><tr><td>Treino</td><td>70%</td></tr><tr><td>Validação</td><td>20%</td></tr><tr><td>Teste</td><td>10%</td></tr></tbody></table>		Perc.	Treino	70%	Validação	20%	Teste	10%	<p>Estatísticas</p> <table><thead><tr><th></th><th>Treino</th><th>Validação</th><th>Teste</th></tr></thead><tbody><tr><td>KS</td><td>0.683</td><td>0.68</td><td>0.653</td></tr><tr><td>GINI</td><td>0.848</td><td>0.852</td><td>0.827</td></tr><tr><td>AUC</td><td>0.924</td><td>0.926</td><td>0.914</td></tr></tbody></table>		Treino	Validação	Teste	KS	0.683	0.68	0.653	GINI	0.848	0.852	0.827	AUC	0.924	0.926	0.914	
	Perc.																										
Treino	70%																										
Validação	20%																										
Teste	10%																										
	Treino	Validação	Teste																								
KS	0.683	0.68	0.653																								
GINI	0.848	0.852	0.827																								
AUC	0.924	0.926	0.914																								

Automatização do processo



Automatização do processo



Automatização do processo

Simulação

Atual

1	0.2409	1,177,500	32,580	117,750	1,092,330	85,170
Taxa de Redução	Taxa de Conversão	Valor Total	Custos Totais	Valor Retido	Perdas Totais	Retorno

Modelo

0.4153	0.5248	1,065,000	13,530	106,500	972,030	92,970
Taxa de Redução	Taxa de Conversão	Valor Total	Custos Totais	Valor Retido	Perdas Totais	Retorno

Escolha seu critério de otimização:

SpEqualSe MaxSp MaxEfficiency
 MaxSe ROC01 Personalizado

Ponto de corte: 0.16 Personalizar ponto de corte

Custos

Ticket Médio

Matriz de Confusão

	Referência 0	Referência 1
Predição 0	1830	75
Predição 1	643	710

Acurácia: 0.77

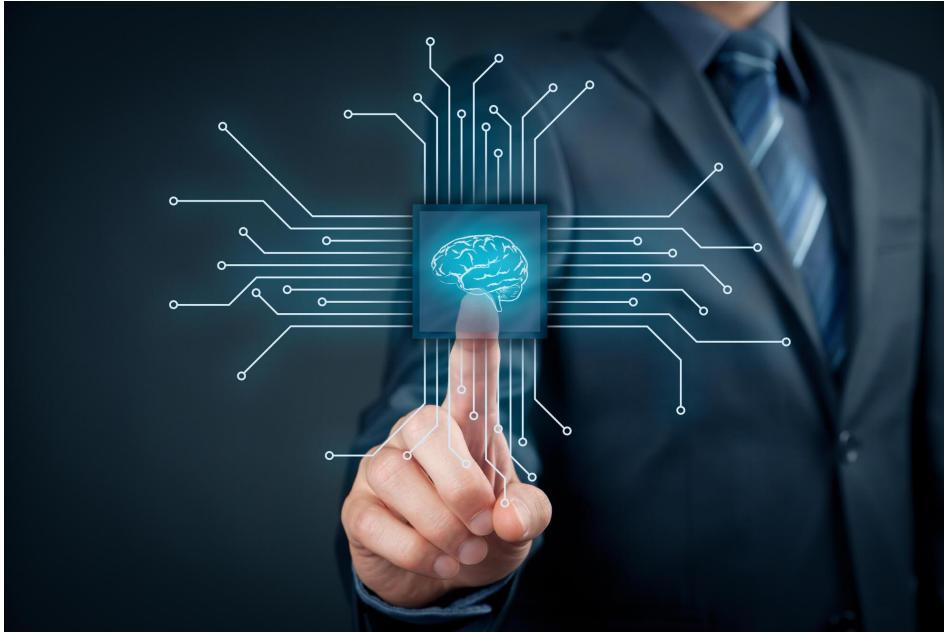
Sensibilidade: 0.74

Especificidade: 0.9045

Positive Predict Value: 0.9606

Item	Valor
X11	0.045
X1	0.012
X12	0.010
X13	0.007
X7	0.006
X8	0.004
X10	0.002
X2	0.0015
X6	0.001
X9	0.0005
X3	0.0002
X5	0.0001
X4	0.00005

Cases

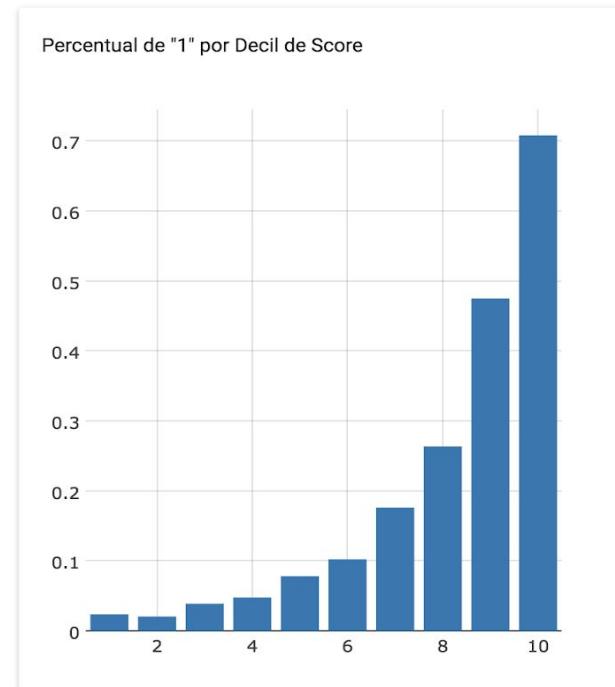
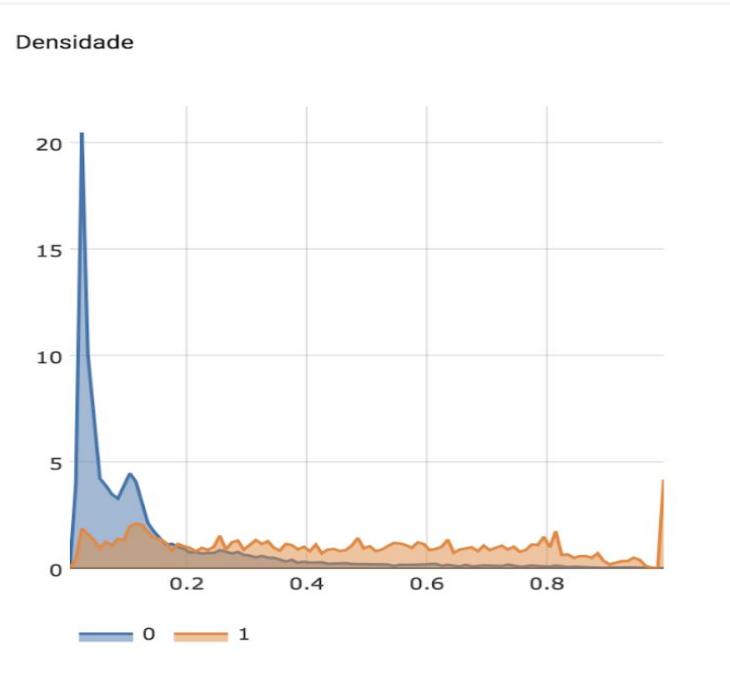


Case - Empresa de Cobrança

PROBLEMA: Como faço a priorização de ligações a fim de maximizar o retorno do valor cobrado?

SOLUÇÃO: Criação de um modelo preditivo que retorna a probabilidade da pessoa atender um telefonema e realizar o pagamento. Utilizamos um ano de histórico de ligações e pagamento para a criação do modelo (10 milhões de ligações).

Case - Empresa de Cobrança



Case - Empresa de Cobrança

Teste A/B: fila teste (priorização feita pela DataRisk), fila concorrente (melhor priorização da empresa).

Data	Modelo V2									
	28/01/2019			29/01/2019			30/01/2019			Var(%)
	Fila_teste	Fila_concorrente	Var(%)	Fila_teste	Fila_concorrente	Var(%)	Fila_teste	Fila_concorrente	Var(%)	
Carteira	5.323	5.468	✖ -2,7%	5.323	5.468	✖ -2,7%	5.323	5.468	✖ -2,7%	✖ -2,7%
% Penetração	69,06%	51,94%	✔ 24,8%	45,50%	37,38%	✔ 17,8%	44,82%	13,24%	✔ 70,5%	✔ 70,5%
Funil Unique Trabalhado	2.676	2.840	✔ 6,7%	2.422	2.614	✔ 7,6%	2.286	2.704	✔ 60,7%	✔ 60,7%
Alô	297	122	✔ 58,9%	236	73	✔ 69,1%	197	66	✔ 66,5%	✔ 66,5%
CPC	20	6	✔ 70,0%	17	9	✔ 47,1%	12	6	✔ 50,0%	✔ 50,0%
Acordo	2	0	✔ 100,0%	1	1	✔ 0,0%	1	0	✔ 100,0%	✔ 100,0%

Case - Empresa de Fidelidade

PROBLEMA: Baixa taxa de conversão de e-mails para resgate de pontos. Mandavam e-mails para 6 milhões de clientes e a maioria ou não lia ou não trocava os pontos.

SOLUÇÃO: Modelo preditivo utilizando dados históricos para calcular a probabilidade de um cliente resgatar os pontos. Mais de 2 mil variáveis criadas. Calculamos essa probabilidade para todos os clientes no próximo mês.

RESULTADOS PRÁTICOS: Reduzimos o envio de e-mails de 6 milhões para 500 mil e obtivemos 90% da conversão anterior.

Case – Grande Banco

PROBLEMA: Cada modelo de crédito era feito por 2 analistas e demorava mais de 2 semanas.

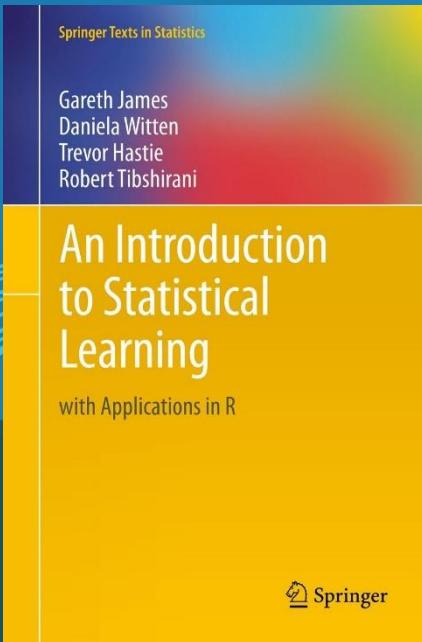
SOLUÇÃO: Uso da plataforma de modelagem da DataRisk para automatizar o processo.

RESULTADOS: Modelo construído em 2 horas. Performance preditiva igual ao modelo construído pelos analistas.

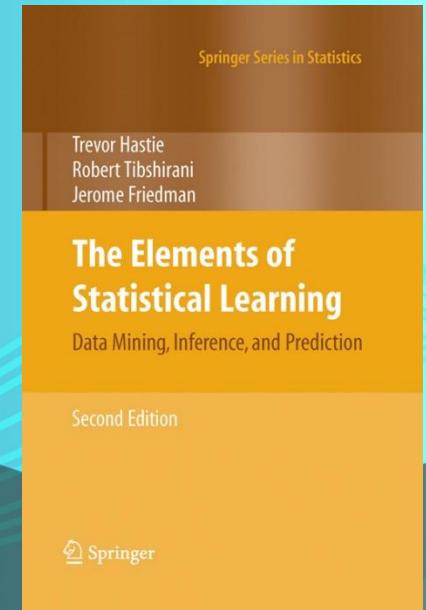
Laboratório



Referências Bibliográficas



“The elements of Statistical Learning”
- Hastie, Tibshirani e Friedman



“An Introduction to Statistical Learning”
- Hastie, Tibshirani, James e Witten