Estatística e Matemática Aplicadas a Data Science

Diógenes Justo
BM&FBOVESPA & Professor FIAP

Agenda

Modelagem para Data Science (Matemática e Estatística)

Detecção de Fraudes

Forecast (financeiro)

Conclusões

Modelagem para Data Science

"Modelagem matemática consiste na arte (ou tentativa) de se descrever matematicamente um fenômeno."

Modelagem para Data Science

Quanto a aleatoriedade

Modelo determinísticos: mesmas entradas produzem mesmas saídas; não se considera incerteza.

$$y = \alpha.x + \beta$$

Modelo probabilístico (estocástico): não se conhece sua saída, somente uma probabilidade e, portanto, uma incerteza.

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$

Modelagem para Data Science

Quanto aos dados de entrada

Modelo heurístico: provocam explicações sobre o que está sendo estudado (se algum variável aumenta, a outra também).

=> Análise exploratória (gráfico dispersão, distribuições/histograma)

Modelo empírico: baseado em observações diretas (experiência)

=> Modelagem do problema (dados de treino)

Modelagem para Data Science

Quanto a representação

Modelo qualitativo: promovem explicações e *insigths* sobre o que pode estar ocorrendo.

=> Análise exploratória: gráficos, descrições causais, inferências

Modelo quantitativo ou numérico: baseado em números ou classificações.

=> Modelagem do problema e geração de algoritmos (dados de treinamento, teste e previsão)

Modelagem para Data Science

Quanto a aplicação

Modelo descritivo ou simulatório: simular um problema.

Modelo de otimização: determinação de ponto ótimo (min / máx).

Modelo de controle: análise de uma variável específica (ex: modelo preditivo).

Modelagem para Data Science

Quanto ao tipo de saída

Regressão: variável de saída contínua.

Classificação: variável de saída discreta (finita) ou categórica.

Deteção de Fraudes

Gravação da apresentação:
https://www.infoq.com/br/presentations/estatistica-e-matematica-aplicadas-a-data-science/

Detecção de Fraudes

O Problema

Identificar fraudadores da forma mais eficaz possível, reduzindo o custo com a fraude, baseado em dados cadastrais.

O custo (t, $) da fiscalização pode inviabilizar 100% de fiscalização, em muitos casos.

Detecção de Fraudes

Abordagens

P1. Identificar fraudadores => Modelo de Controle

$$P(y=1|x,\alpha) = \alpha_i . x_i(w, h(x))$$

$P(y=1|x,\alpha)$ ← Característica ($h(x)$)

P2. Custo da fiscalização x Custo da fraude => Modelo de Otimização

$$Min(y) = \alpha_1 . x_1 + \alpha_2 . x_2$$

Minimização do custo (perda) ←

$\alpha_2 . x_2$ → Custo com fiscalização

$\alpha_1 . x_1$ → Custo com fraude

Detecção de Fraudes

Problema 1:

Características: dados cadastrais, geográficos e alguns tipos de comportamentos.

Possuo pré-classificação de fraudadores (observados) e não fraudadores.

Utilização de algoritmos de árvores para treinar o modelo.

Detecção de Fraudes

Problema 1:

Algoritmos:

- Árvores de Decisão
- RandomForest
- AdaBoost (com técnicas de boosting e bagging)
- Extreme Gradient Boosting.

C = 100*23 + 30*9

		Realidade	
		Fraude	Normal
Previsão	Fraude	7	2
	Normal	23	5

Detecção de Fraudes

Problema 1: Avaliação de modelos

1. Taxa de acerto dos fraudadores:

TAF = TP / (TP+FN) (*)

2. Quantidade de fiscalizações:

TQF = (TP+FP)/#TOT.POP. (**)

		Realidade	
		Fraude	Normal
Previsão	Fraude	TP	FP (E1)
	Normal	FN (E2)	TN

* Sensitivity ou Recall | ** Prevalence

Detecção de Fraudes

Problema 1: Avaliação de modelos

Modelos	TAF	TQF
Cenário 3.1: AD Com enriquecimento	37.52%	11.53%
Cenário 4.1: RF Com enriquecimento	39.69%	10.18%
Cenário 4.2: RF Sem enriquecimento	43.91%	11.62%
Cenário 5.1: AdaBgg Com enriquecimento	32.71%	9.27%
Cenário 5.2: AdaBgg Sem enriquecimento	42.25%	13.52%
Cenário 6.1: AdaBst Com enriquecimento	35.27%	10.42%
Cenário 6.2: AdaBst Sem enriquecimento	41.63%	13.26%
Cenário 7.1: XGBst Com enriquecimento	88.41%	29.94%

Detecção de Fraudes

Problema 2:

Custo da fiscalização x custo da fraude

Premissas:

- Amostra 100.000
- Tx. de fraude (observada): ~20%
- Custo da fraude: 100$
- Custo da fiscalização: 30$
- Fiscalização zera fraude

Detecção de Fraudes

Custo da fraude: 100$
Custo da fiscalização: 30$

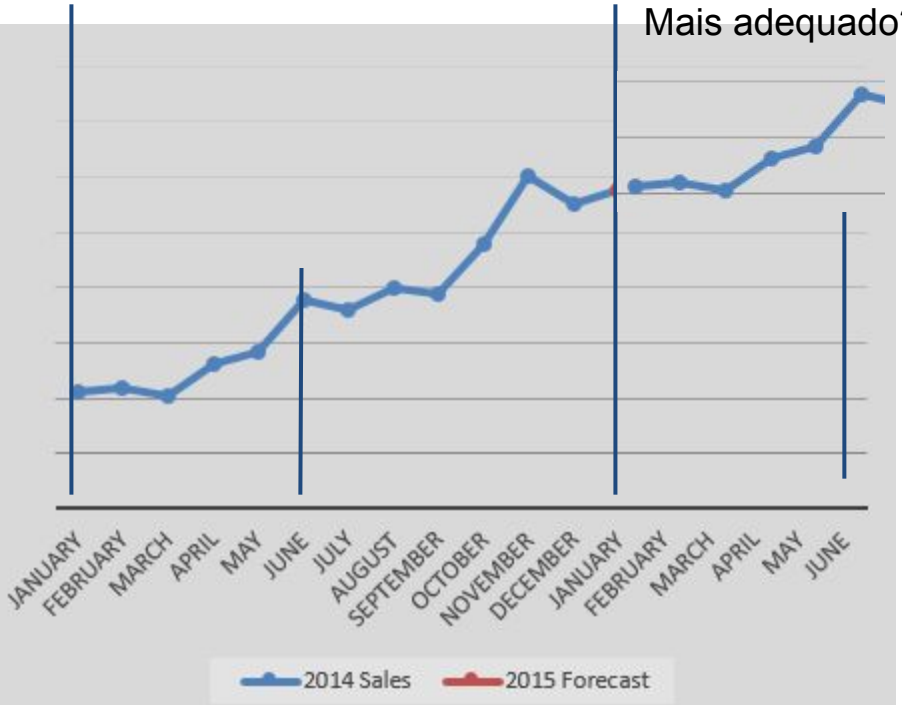
Problema 2: Avaliação de modelos

Modelos	# Fraude/ano	# Inspeção	Custo
Cenário ótimo - Fiscalizar só fraudadores	0	20000	600,000.00
Cenário 1 - Sem Fiscalização	20000	0	2,000,000.00
Cenário 2 - Fiscalizando todos	0	100000	3,000,000.00
Cenário 3.2 - Sem enriquecimento (56,38%)	10871	15480	1,551,536.93
Cenário 4.1: RF Com enriquecimento	12063	10180	1,511,686.84
Cenário 4.2: RF Sem enriquecimento	11217	11620	1,470,315.08
Cenário 5.1: AdaBgg Com enriquecimento	13458	9270	1,623,876.03
Cenário 5.2: AdaBgg Sem enriquecimento	11549	13520	1,560,510.10
Cenário 6.1: AdaBst Com enriquecimento	12947	10420	1,607,295.48
Cenário 6.2: AdaBst Sem enriquecimento	11674	13260	1,565,158.23
Cenário 7.1: XGBst Com enriquecimento	2318	29940	1,129,977.23

Forecast

Forecast

	2014 Sales	2015 Forecast
January	420	
February	440	
March	410	
April	524	
May	567	
June	755	
July	720	
August	800	
September	780	
October	960	
November	1 200	
December	1 100	
January	1 150	1 150
February		1 240
March		1 280
April		1 310
May		1 360
June		1 370

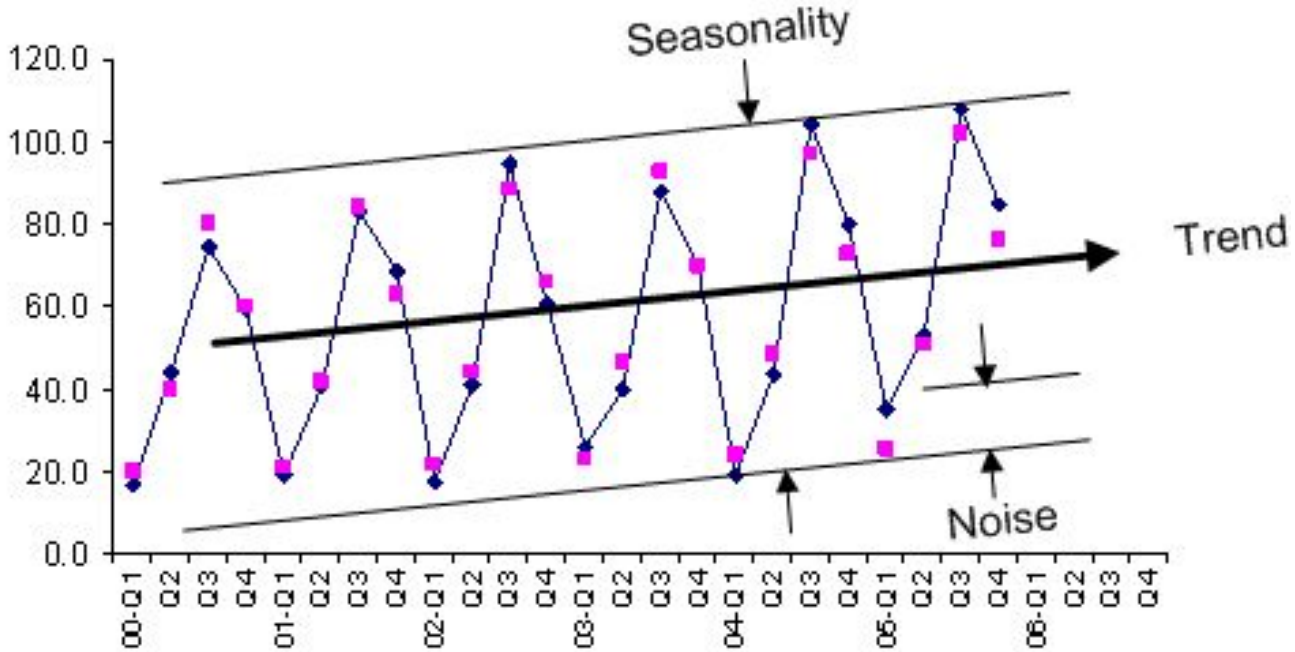


Forecast

Interesse por Churrasco

Forecast

Componentes



Forecast

Problema:

Prever o comportamento de uma variável agregada ao longo do tempo (projeção ou previsão).

Modelo de controle =>

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$

Forecast

Abordagens

Alternativa 1. Existe uma varíavel que explique o comportamento (relação causa e efeito ou correlação)

=> Regressão Linear: $$y_i = \alpha . x_i + \beta + \varepsilon_i$$

Alternativa 2. A variável tem um histórico que a explica ao longo do tempo e uma característica de tendência.

=> Modelo autoregressivo: $$y_i = \alpha . y_{i-1} + \beta + \varepsilon_i$$

Conclusões

diogenes.justo@.com