# IoT: Modelagem binária

Mestre Adilson Lopes Khouri

24 de agosto de 2019

### Sumário

- Processo de Data Science
- Engenharia de features
- Overview de modelos
- Treinamento de modelos
- Ferramentas
- Contato

#### Pessoal

 Adilson Khouri, jogador de Magic the Gathering, nerd, apaixonado por computação e machine learning.



Figura: Eu no Peru palestrando e na Argentina trabalhando

### Formação Acadêmica

- Bacharel em Sistemas de Informação (2011 USP)
- Mestre em Sistemas de Informação (2016 USP)
- Doutorando em Sistemas de Informação (cursando USP)

### Experiência de Mercado

- Programador na consultoria Arbit (2010-2011)
- Programador Itaú-Unibanco (2011-2013)
- Cientista de dados Sr. PagSeguro (2016 2018)
- Especialista em Modelagem NuvemShop (Atual)
- Professor carta convite SENAC (Atual)

### E os senhores?

- Nome
- Trabalho
- Tempo de experiência, área de atuação
- Conhecimento sobre o assunto da disciplina

## Expectativas

- Quais expectativas?
- O que deve ser evitado?
- (E-Mail: 0800dirso@gmail.com)

### Requisitos

- Computadores com MAC OS
- Computadores com internet
- Python, pandas, jupyter notebook, numpy, scipy e scikit-learn

#### Processo de Data Science

#### **Data Science Lifecycle Business** Start Understanding On-Premises vs Cloud Data Source Database vs Files Transform, Binning Feature Temporal, Text, Image Engineering Feature Selection Streaming vs Batch Pipeline low vs High Frequency Data Algorithms, Ensemble Model Modeling Acquisition & Parameter Tuning On-premises vs Cloud Retraining Training. Understanding Environment Database vs Data Lake vs .. Model management Small vs Medium vs Big Data Cross Validation Model Wrangling, Structured vs Unstructured Model Reporting Exploration & Data Validation and Cleanup Evaluation A/B Testing Cleaning Visualization Customer Deployment End Acceptance Scoring. Performance monitoring, etc

Figura: Obtido no link: Microsoft

### Engenharia de features

- Modelos usam muitas variáveis para tomar decisões
- Encontrar boas variáveis é parte fundamental para um modelo
- Citar exemplo de variáveis de transações financeiras
- Citar exemplo de variáveis de pagamento de assinaturas
- Citar exemplo de um classificador de brasileiros e peruanos
- Citar exemplo de um classificador de argentinos e peruanos

#### Modelos

- Modelos tomam decisões baseados em diversas variáveis para, entre outras coisas, classificar dados
- Há modelos para classificar em duas classes ou mais.

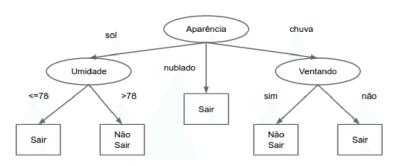


Figura: Exemplo de árvore de decisão

#### **Treinamento**

- O algoritmo de treinamento é único para cada modelo mas o processo de como se treinar um modelo é parecido
- Os dados são dividos em treino (70%) e teste (30%)
- O conjunto de treino é apresentado ao modelo com os rótulos de cada observação
- Tipicamente usa-se uma validação cruzada para treinar o modelo

### Validação Cruzada

#### 10-fold cross validation

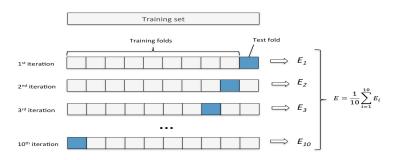


Figura: Obtido no link: python-machine-learning-book

### Validação

 O modelo é validado com o conjunto de teste, o qual não deve exibir os rótulos para o modelo

		PREDITO	
		Classe A	Classe B
VERDADEIRO	Classe A	VP	FN
	Classe B	FP	VN

Figura: Obtido no link: Scielo

### Validação - outras métricas

- Se usarmos a matriz de confusão acima podemos obter outras métricas
- Citar o problema das classes de seller (relacionar com F1)

Precision = 
$$\frac{TP}{TP + FP}$$
  
Recall =  $\frac{TP}{TP + FN}$   
F1 = 2 \*  $\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$ 

#### **Ferramentas**

- Na teoria pode-se usar qualquer linguagem de programação para trabalhar com Data Science
- Na prática usa-se, majoritariamente, a plataforma R e a linguagem Python (com alguns pacotes científicos)
- scikit-learn

### Hands on

• Treinar modelo em Python com a turma

## Trabalho para a aula

• Treinar um modelo de Random Forest usando o scikit-learn

### Fim!

Agradeço pela presença dos senhores :)

### Contato

- E-mail: 0800dirso@gmail.com
- Phone: +55119444 26191
- Linkedin
- Curriculum Lattes
- Código fonte GitHub