Medindo a viabilidade de um projeto de DS

All rights reserved

Jul
2019

Guilherme Spadaccia Leme

- Analista de Sistemas na TOTVS por 6 anos
- Cientista de Dados na TOTVS Labs há 2 anos
- Mestrando em Inteligência de Sistemas na USP (aluno especial)

Motivação

Todo mundo quer usar Inteligência Artificial, mas nem sempre é possível...

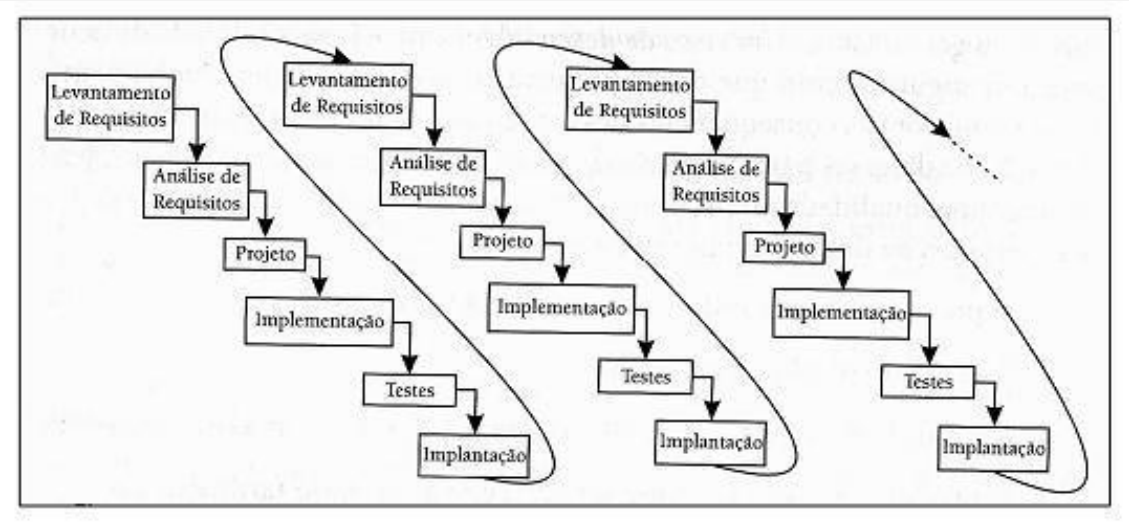
Desenvolvimento de Software Vs. Ciência de Dados

Projetos de Ciência de Dados são iguais a projetos de Desenvolvimento de Software?

Podemos assumir os mesmos paradigmas para definir a viabilidade do projeto?

Desenvolvimento de Software Vs. Ciência de Dados

Etapas de um projeto de desenvolvimento de software.



Fonte: https://wiki.sj.ifsc.edu.br/wiki/index.php/Ciclo_de_Vida_Iterativo_e_Incremental

Desenvolvimento de Software Vs. Ciência de Dados

Já um projeto de Ciência de Dados:

Desenvolvimento de Software Vs. Ciência de Dados

A diferença é clara: os **DADOS**!

- Ambos compartilham a complexidade de software.
- Desenvolvimento de Software tem **dados como consequência**.
- Ciência de Dados tem **dados como insumo**.

Desenvolvimento de Software Vs. Ciência de Dados

Vamos assumir que sabemos como desenvolver um software, então vamos focar no que interessa...

Primeiros passos

1. Aquisição dos dados (negócio / desenvolvimento)
2. Exploração inicial (negócio / desenvolvimento / matemática)
3. Construção da Baseline (negócio / desenvolvimento / matemática)

Aquisição dos dados

- Mapeamento do problema
  - Qual o problema?
  - Qual a origem?
  - Qual a consequência?
  - Exemplo simples churn:
    - O que é um churn para você?
    - Por que os clientes abandonam a empresa?
    - Como vocês revertem os clientes em churn e quais as consequências?
  - Define quais informações são relevantes

Aquisição dos dados

- Nem sempre os dados estão estruturados
  - Excel
  - Email
  - PDF
  - Post it
  - O cara lá sabe
- Existem os desafios técnicos: Para onde esses dados vão? Como extrair os dados da base do cliente?

Análise / Exploração inicial

- Avaliar os dados
  - Volume
    - Precisamos de muitos dados, mas balanceie quantidade e necessidade
    - Mudança de controle ou de métrica durante o tempo. O que era medido de uma forma hoje pode ser medido de outra.
    - Algumas soluções
      - Sampling (muitos dados)
      - Over-sampling (poucos dados)
      - Dados sintéticos (poucos dados)

Análise / Exploração inicial

- Avaliar os dados
  - Variedade
    - Os dados abrangem cenários diversos?
    - Dados balanceados
    - Algumas soluções
      - É possível obter mais dados?
      - Under-sampling
      - Over-sampling
      - Dados sintéticos
      - Modelos com penalização

Análise / Exploração inicial

- Avaliar os dados
  - Veracidade
    - Os dados refletem a realidade?
    - Podemos descrever os acontecimentos da empresa através dos dados?
    - “Nós temos o controle mas não usamos”
    - Algumas soluções
      - Relacionar origem vs. consequência

Construção da Baseline

É aqui que vamos realmente colocar nossos dados a prova.

Construção da Baseline

- O que é uma baseline?
  - Um modelo simples que visa reproduzir a assertividade que o cliente já possui e que servirá como base para avaliar seus modelos futuros.

- Qual a importância da baseline para um projeto?
  - Quando falamos de CIÊNCIA de dados, métrica é muito importante. Precisamos garantir a melhora dos resultados.
  - O método* é importante, ele pode definir se sua medição de resultado está ou não correta.

Entenda por método todo o processo que envolve o modelo: Do tratamento das features até o cálculo da acurácia do resultado.

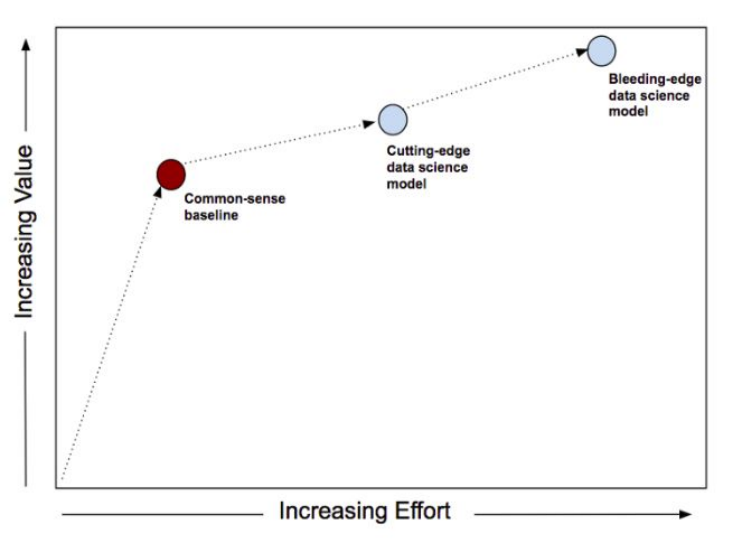
Construção da Baseline

- “Meu modelo está ótimo, tenho 95% de acurácia.”
  - *“If you torture the data long enough, it will confess to anything.” - Ronald Coase*
- Existem diversas formas de se medir assertividade:

Fonte: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Construção da Baseline

- Usar ou não usar Machine Learning na baseline?
  - Opte por não usar machine learning na baseline ou use um modelo muito simples
  - Muitas vezes com pouco esforço conseguimos bons resultados



Fonte: https://towardsdatascience.com/first-create-a-common-sense-baseline-e66dbf8a8a47

Construção da Baseline

- Usar ou não usar Machine Learning na baseline?
  - Exemplo esdrúxulo:
    - Para um problema de classificação binária, balanceado, uma solução “flip coin” atinge, teoricamente, 50% de assertividade.
  - Exemplo menos esdrúxulo:
    - Modelo de Auditoria Médica: de todos os pedidos médicos 85% são sempre aceitos. Calcular a probabilidade dos últimos 3 meses daquele tipo de pedido ser aceito. Garantimos 85% de assertividade.

Resultado da análise

- Cada um dos 3 passos descritos (aquisição dos dados, análise e baseline) pode demonstrar algum impeditivo para o projeto
- Possíveis resultados:
  - Positivo (é possível seguir com o projeto)
  - Negativo (algo foi identificado que inviabiliza o projeto)

Resultado da análise

- Positivo
  - Evolução da baseline
  - Criação de modelos mais complexos
  - Todo novo modelo deve superar a baseline e a partir desse momento esse modelo se torna a nova baseline

Resultado da análise

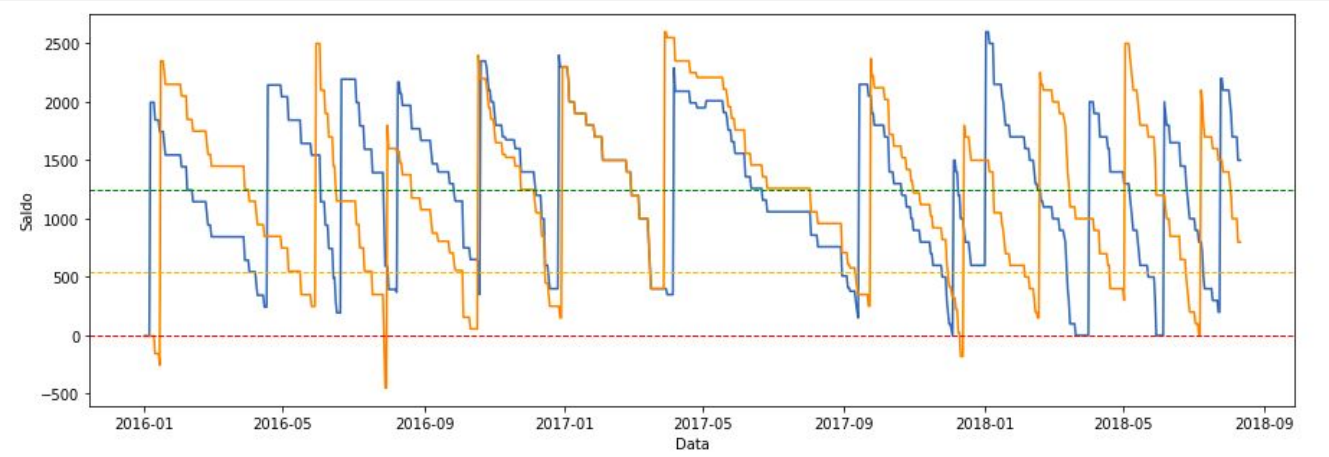
- Negativo
  - Houve algum impeditivo?
    - A análise já pode agregar valor ao cliente
    - Visualizações podem gerar novos insights
    - Sugestões de novos controles e melhorias de processos
    - Benefício da infraestrutura de Big Data
  - Não foi possível superar a baseline?
    - A própria baseline (usando ou não ML) pode melhorar os processos do cliente

Resultado da análise

- Casos reais
  - Projeto de previsão de refugo
    - Objetivo: Prever quanto que a produção de um determinado produto por determinada máquina geraria de refugo
    - Status: Negativo
    - Motivo: Não existiam processos bem definidos e os dados do maquinário não refletiam a realidade

Resultado da análise

- Casos reais
  - Projeto de previsão de data e quantidade de compra
    - Objetivo: Prever qual seria a melhor data e qual a melhor quantidade a ser comprada de determinado produto
    - Status: Em andamento
    - Resultado: A baseline foi tão bem aceita que o cliente queria colocá-la em produção (não foi utilizado Machine Learning)



Resultado da análise

- Mensagens
  - O objetivo de um projeto deve ser agregar valor.
  - Machine Learning é uma ferramenta e deve ser usada quando necessário e da forma correta!!!
  - Ciência de Dados é uma CIÊNCIA, se preocupe com métrica, método, processos e resultados.

Obrigado

Guilherme Spadaccia Leme

guilherme.spadaccia@totvs.com.br

Tecnologia e conhecimento é nosso DNA

O sucesso do cliente é nosso sucesso

#SOMOSTOTVS

totvs.com

company/totvs

@totvs

fluig.com