Tendencias em IA e Big data

* Big Data Não Existe
  + Mudança de paradigmas de dados
    - Antes tínhamos estoque de dados
    - Hoje temos fluxo de dados

Armazenar dados está cada vez mais barato

Coletamos muitos dados

1. Mais barato armazenar
2. Muita geração de dados

* O que é BIG data

É um conjunto de informações

Temos que ter um proposito

Necessário ter um dos 3 V’S

* + Volume
  + Variedade 🡪 Diversidade de formatos e fontes 🡪 Dados não estruturados
  + Velocidade

2 V’s alternativos

* + Veracidade 🡪 Integridade de dados, precisão confiabilidade
  + Valor 🡪 Vantagem competitiva insight estratégico
* Tipos de Dados
  + Não Estruturados
    - Post, Imagem, Áudio, Vídeo< Livros
  + Semi estruturados
    - Xml, Json
  + Dados estruturados
    - Tabelas linhas e colunas
    - Banco de dados
* Modelo preditivo

Uma função matemática capaz de identificar padrões.

* Simplificando IA

3 Tipos de IA

* + Super IA (ASI) Idealizada na ficção

Não existe atualmente mas existe estudo na academia para que seja segura

* + Inteligência artificial Geral (GAI)

Não existe AINDA!!! Mas estamos perto

Q\* é uma possível GAI

Ela procura soluções para problemas os para os quais não foi treinada.

* + Inteligência Artificial Restrita (NAI)

É A ia COM OBJETIVOS ESPECIFICOS. São todas as que conhecemos até agora

Machine learning

Análise preditiva

Prever está ficando mais barato e mais assertivo.

E isso possibilita novas aplicações

Autonomia não é algo novo mas normalmente é utilizada em ambientes controlados

Os condicionais são simples IF-ELSE

Autonomia de um carro dirigindo numa cidade com IF-ELSE é IMPOSSIVEL, a quantidade de condicionais é muito grande e são mutáveis, o ambiente não é controlado, é caótico.

Autonomia veicular só pode ser resolvido com previsão

R.P.A (Robotc Product Aplication)

Gente não é problema do RH é problema do executivo

Reduzir Turn-Over usando analise preditiva

\*Dataset: Conjunto de dados estruturados coluna target

p-Valor: Medida estatística

IA com dados não estruturados

Tem que ter uma maquina/server PARRUDO

Fim das anotações ao vivo

Anotações revisão:

* Big data não existe
* Modelo Preditivo

Para a maioria das pessoas e empresas I.A. é apenas uma ferramenta.

* Analise de IA: Basicamente é fazer previsão usando modelos matemáticos
  + Super Importante: Avaliar a qualidade desse modelo que esta sendo criado. A qualidade do modelo é a qualidade da I.A.
  + Também importante: Explicar os resultados do modelo.

API: (Aplication Program Interface)

* Big Data :

Mudança no conceito de como trabalhar com dados.

Antigamente tínhamos:

* + Estoque de Dados

Hoje Temos

* + Fluxo de Dados

Temos uma quantidade absurda de dados e em sua grande maioria dados perecíveis

Temos muitos dados por dois motivos:

* + Armazenagem de dados ficou cada vez mais barato
  + Coletamos cada vez mais informação
* O que é big data?

Big Data não é um conceito acadêmico, o conceito veio do mercado. Surgiu a partir das companhias.

* + Definição Segundo <https://www.gartner.com/en>

Big Data é um conceito de informações em alto volume, alta velocidade e/ou alta variedade, que demanda formas inovadoras e custo-efetivas de processamento de informações, permitindo uma visão aprimorada, tomada de decisões e automação de processos.

* Os Clássicos 3 v’s do BIG DATA

Com o tempo se solidificaram 3 grandes conceitos relacionados com big data: Volume, Variedade e Velocidade.

* + Volume
    - Volume Massivo de dados
    - Alinhado com o fato de termos acesso a cada vez mais dados
    - A característica que a maioria doas pessoas mais associa e em geral está relacionado.
  + Variedade
    - Diversidade de formatos e fontes
    - Dados heterogêneos
    - Conseguir trabalhar com dados semiestruturados e mão estruturados (texto livre, imagem, áudio e vídeo)
  + Velocidade
    - Rapidez de geração
    - Fluxo contínuo
    - Velocidade de processamento
    - Informação perecível
    - Informação não tem valor intrínseco
    - A Informação tem seu valor em um determinado momento e contexto
    - Perecível muito rapidamente
* Outros 2 v’s importantíssimos do BIG DATA
  + Veracidade
    - Integridade
    - Precisão dos dados
    - Confiabilidade das Informações
    - Validade das fontes
    - O dado precisa ser confiável
    - O que vale é a qualidade do dado
    - Dado deve servir para algo
    - Dado Puro não tem valor dado valioso e dado trabalhado
  + Valor
    - Insights estratégicos
    - Benefícios comerciais
    - Vantagem competitiva
* Informação =Dado+Contexto
* Big Data não é um aplicativo, biga data é um conceito
* Tipos de dados
  + Dados não estruturados

Qualquer tipo de dado:

* Posts, imagem, áudio, vídeo, livros
  + Dados semiEstruturados
    - XML, Json
  + Dados Estruturados
    - Tabelas do dia a dia
    - Linhas e colunas
* Quanto menos estruturado o dado mais complexo de trabalhar e mais recurso é consumido
* O que é um modelo preditivo?

Uma função matemática capaz de identificar padrões por meio de uma série de dados previamente coletados.

* Simplificando IA

Atualmente há um consenso que temos 3 tipos de I.A. (não é um consenso final pode ser mudado)

* Os 3 tipos
  + Super Inteligência Artificial (ASI-Artificial Super Inteligence)

Máquinas autoconscientes. Ainda não existem, mas já é estudada para que seja ética e segura. É a IA da ficção. Não a um consenso se é uma tecnologia viável que um dia poderá ser alcançada.

* + Inteligência artificial Geral (GAI- General Artificial Inteligence)

Ia que procura soluções para problemas para os quais ela não foi especificamente treinada.

Também não existe ainda, porém é consenso que é viável e estima-se que estamos perto de alcançar. Q\* pode vir a ser uma

* + Inteligência artificial restrita (NAI -Narrow Artificial Inteligence)

É a IA com objetivos específicos. A maior parte dos modelos baseados em IA que conhecemos e estão no nosso dia a dia são desse grupo. Sempre é criada para um objetivo bem específico. IA generativa estra dentro de NAI. É a Ia que nos interessa no futuro Próximo.

* Machine learning = Análises preditivas

“Prever” está mais rápido, mais barato e mais assertivo, e possibilita novas aplicações.

* O termo mais correto é predição, por simplificação usamos prever.

Autonomia veicular nunca no passado nunca foi pensado como um problema de previsão.

Autonomia não é algo novo, mas tradicionalmente é utilizada em ambientes controlados, como por exemplo armazéns.

* Tradicionalmente se usa algoritmos condicionais (IF ... ELSE ...)

A Nova autonomia não pode usar condicionais if esle, as variáveis são infinitas e mutáveis, não há como mapear as possibilidades. O mundo real é imperfeito.

Para resolvermos esse problema temos que usar análise preditiva.

* + Análise preditiva:
    - Prediz uma ação ou consequência



* + - Erra Ou acerta
    - Aprende com o erro ou com o acerto



* + - Atualiza os parâmetros
    - Prediz uma ação ou consequência...

Fim Bloco 1

Bloco 2

Qual a apólice de seguro para não perder para a IA ?

* Criatividade
* Inteligência emocional (Lidar com pessoas)

Engenheiro de software vai ser a parte pensante.

* Ia com dados estruturados

Exemplo: Reduzir Turn Over usando técnicas de IA machine learning e analise preditiva.

* + Dataset = Conjunto de dados, é uma informação necessária para se trabalhar com I.A. em formato estruturado (Linhas e Colunas)

Coluna Target = característica mais relevante do Dataset

p-Valor = é uma medida estatística com fórmula de intervalo de confiança e nível de confiança.

Quanto mais próximo de 2 mais a variável ajuda a identificar as características mais relevantes do meu DATASET

Forma simples do que se pode chamar de um machine learning ou de IA, ou analise preditiva, Ela se chama regressão Linear e é baseada em LOGITE

* Modelo Preditivo: é um modelo que “Aprendeu” com os dados, não esquecendo que modelo é uma função matemática.
* Treinar o modelo é basicamente fornecer os dados para que o modelo possa ser gerado.
* Os dados podem ser diferentes para o mesmo modelo
* Usando IA quem cria os condicionais é o computador
* Ia com dados não estruturados
* Testar <https://cloud.google.com/vision>
* Data-driven – orientada a dados

Fim Bloco 2

Bloco 3

Qualidade dos dados

* The Signal and the Noise: Why So Many Predictions Fail-But Some Don't

Nate Silver

Signal = informação Util

Noise = Ruido, informação errada ou que atrapalha

* Analise exploratória de dados
  + Procurar por missing values

Quando achar decidir por estratégias como:

Usar a média

Inferir o valor a partir de outro atributo

* Ia Tem viés e depende do cientista de dados

Trilogia laurentino gomes: escravidão

Machine learning = Aprendizado de maquina

LLM = Large Language Models = ChatGPT etc . . .

NPL = Processamento de Linguagem Natural = precursor LLM

SOTA = State Of The Art

* Qualidade e explicação de modelos preditivos

1. Não existe modelo bom, quando os dados de treinamento são ruins.
2. Porem, mesmo com dados de qualidade, não necessariamente podemos criar um modelo.

Quanto mais dados e de melhor qualidade mais fácil criar um modelo viável.

Parte importante na criação do modelo é conseguir aferir a capacidade preditiva que o modelo fornece.

O quando que o modelo consegue acertar.

* + Sempre é possível criar um modelo para uma analise preditiva, a questão é se ele é útil ou não, ou seja se a previsão dele tem uma confiança aceitável ou não.
* Matriz de confusão

Ao rodar um modelo eu não rodo com o volume total dos dados, eu reservo uma parte para conferir a confiança do meu modelo, conferindo a previsão do modelo com os dados reais montando a seguinte matriz:

Tabela, Calendário

Descrição gerada automaticamente com confiança média

No exemplo acima conseguimos tirar as seguintes métricas:

* Sensibilidade (Recall) = VP/(VP+FN)

1949 / (1949+473) = 0,80470 **≅ 80%**

* Especificidade = VN/(VN+FP)

1816 / (1816+860) = 0,67862 **≅ 68%**

* Acuracidade = (VP+VN) / N

(1949+1816) / 5098 = 0,7385 **≅ 74%**

* VPN (Valor preditivo Negativo) = VN/(VN+FN)

1816 / (1816+473) = 0,79335 **≅79%**

* VPP (Valor Preditivo Positivo) = VP/(VP+FP) = Precisão

1949/ (1949+860) = 0,693841 **≅ 69%**

* F1\_score = 2\*(Precisão (VPP)\*Recall) / (Precisão (VPP)+Recall)

2\*(0,693841\*0,8047) / (0,693841+0,8047) = 0,745169 **≅75%**

Acuracidade é uma métrica fraca mas muito cobiçada pelo mercado

No caso acima as métricas mais relevantes são o VPN e a precisão VPP

Explicando em palavras:

O VPN = 79% Significa que 1 em cada 5 é negativo, no caso inadimplência de **≅ 20%.**

O VPP = 69% significa que 2/3 é positivo no caso 1/3 de risco de perca de receita

O cientista de dados fornece os dados. Quem toma a decisão é o dono do negocio.

* Explicação do modelo

Az vezes a capacidade de explicar o modelo é um ponto que ajuda mais na decisão do que uma previsão mais precisa.

Ex:

Em uma regressão logística uma taxa de 75% de acerto é considerada uma boa taxa asserto para o modelo, já uma rede neural com menos de 95% de acerto é considerada uma taxa ruim para o modelo Ruim. Portanto a rede neural é mais precisa.

No entanto a rede neural é muito complexa e pode ser considerada uma black-box pelo cliente. Já a regressão logística pode ser plenamente explicada para alguém que conheça o modelo de negócios, por isso pode ser que esse modelo faça mais sentido para ele.

Os dados contam a verdade do mundo.

Toda predição por IA/machine learning deve poder ser explicada

Pensar no chatgpt como se fosse um assistente, delegar trabalho braçal **que possa ser conferido.**

Outros modelos de ia opensource:

* + Stanford Alpaca 🡪 <https://crfm.stanford.edu/2023/03/13/alpaca.html>
  + 22-hours cabrita 🡪 <https://github.com/22-hours/cabrita>
  + Mistral AI 🡪 <https://mistral.ai/>
* Ia generativa não fornece probabilidade de quanto acerta
* GPT não é “melhor” que todos os métodos de machine learning já consolidados
* GPT é muito bom em estruturar dados
* GPT( e não o chat-gpt) será amplamente utilizado em tarefas de automação
* Automação precisa de supervisão adequada
* LLM -Alucina
* Aonde precisamos chegar para que a IA seja segura e confiável?
  + Quando eles forem como uma calculadora, portável precisa simples e barata.
* Ia não é a nova moda, é uma ferramenta para reduzir custos ou aumentar a eficiência.
* Cultura de dados é uma jornada.

Não se vira uma entidade(pessoa/empresa) datadriven do dia para a noite, é uma forma de pensar, uma forma de agir,

* O que faz diferença não é o software, antes de mais nada pessoas é que fazem a diferença. Pessoas sabem definir problemas.
* O Ouro são os dados
* Ia é o meio não é o fim

O que interessa são os resultados obtidos.

* Se existe um caminho mais simples esse é o caminho a solução para ser boa não necessariamente precisa ser complicada.
* Por onde começar?
  + Os dados
* Não existe um modelo bom à partir de tanto, quem determina se o nível de confiança é bom ou aceitável é o dono do negócio.
* Não existe modelo que acerta 100%