INTRODUÇÃO À CIÊNCIA DE DADOS

Predição de Dados

PREDIÇÃO

- Todas as vezes que se lê, assiste ou escuta uma abordagem sobre DS, rapidamente o termo "Predição" é colocado em pauta.
- É também esse termo que cria uma relação de diferença entre estudos de BI e de DS, visto que tanto BI quanto DS têm processos muito parecidos, inclusive no contexto de Big Data, entretanto, entende-se que o BI fomenta a tomada de decisão baseado no que já aconteceu e o DS projeta a tomada decisão baseado no que aconteceu, e no que PREVÊ que acontecerá.

PREDIÇÃO - DICIONÁRIO

predizer

predizer | *v. tr.*

pre·di·zer lêl - **Conjugar**

(latim *predico, -ere*, dizer antecipadamente)

verbo transitivo

Anunciar antecipadamente o que deve acontecer, seja pelo cálculo (ex.: *predizer um eclipse*), seja por alegada magia (ex.: *predizer o futuro*), seja por conjectura (ex.: *predizer um acontecimento*). = PREVER, PROFETIZAR, PROGNOSTICAR, VATICINAR

Palavras relacionadas: antedizer, antemostrar, augurar, bem-fadar, fatídico, fatiloquente, preconizar.

PREDIÇÃO

- A análise preditiva, ou capacidade de prever resultados e gerar *insights* futuros, com base em dados do passado (e do presente), com o uso de machine learning e modelagem estatística, é um dos principais recursos e objetos de desejo de quem estuda DS.
- A análise preditiva tem sido utilizada pelas maiores empresas do mundo para tomar decisões nos mais variados segmentos e está atrelada a uma diversidade de aplicações e de possibilidades, que torna impossível enumerá-las.

PREDIÇÃO

- Novas perspectivas nunca são demais. Quanto mais você consegue enxergar, melhores se tornam suas decisões — e ficar no escuro não é uma boa opção.
- É preciso saber o que o espera, preferencialmente antes das outras pessoas.
  - Exemplo: É como participar de um programa de TV em que você precisa escolher uma das portas para ganhar um prêmio surpresa. Elas são todas iguais, então tudo o que pode fazer é arriscar seu melhor palpite — a escolha depende de você e da sorte. Mas e se você tivesse uma vantagem — a capacidade de espiar pela fechadura? A análise preditiva proporciona essa vantagem.

PREDIÇÃO

- É importante que você, cientista de dados, compreenda o poder de um modelo de predição com alto índice de acerto.
- O que você faria em um mundo em que pudesse saber qual é a probabilidade de se casar com um colega de faculdade? Em que fosse possível prever qual é a profissão mais adequada para você? Em que pudesse prever a melhor cidade ou país para morar? Em que pudesse prever qual casa comprar, quando tem a intenção de vender no futuro com lucro?

PREDIÇÃO

- Resumindo, imagine um mundo em que seja possível maximizar o potencial de cada momento de sua vida. A vida seria produtiva, eficiente e poderosa.
- Você teria (de certo modo) superpoderes — e pouparia muito tempo.
- Esse mundo pode parecer um pouco chato para pessoas que gostam de assumir riscos não calculados, mas não para uma organização que visa lucros.
- As empresas gastam milhões em gerenciamento de riscos. E se existir algo que as ajude a gerenciar riscos, otimizar operações e maximizar lucros, com certeza você deveria conhecer.
- Esse é o mundo da análise preditiva!

PREDIZENDO

- Podemos modelar comportamentos e predizer ações:
  - Quais produtos os clientes visualizaram antes de comprar?
  - Quais páginas os clientes visualizaram antes de comprar?
  - Clientes que compraram olham descrições?
  - Clientes que compraram leram comentários?
  - Clientes que compraram leram comentários positivos e negativos?
  - Compraram algum produto além do que estavam buscando?
  - As casas vendidas em um determinado bairro têm algo em comum que as que não foram vendidas não têm?

PREDIZENDO

- Clientes que cancelaram um plano (de algum produto ou serviço) tiveram que comportamento antes de solicitar o cancelamento?
- Produtos que venceram (ou chegaram perto disso) no estoque foram comprados em que período? Eles atendem uma demanda sazonal? Eles atendem ao fluxo de caixa da empresa?
- Produtos que foram comprados na "Promoção" do fornecedor realmente valeram a pena?
- O aumento da visibilidade de um produto ou serviço foi impactado diretamente pelo marketing pago nas redes sociais?

OPORTUNIDADES

- Organizações de todo o mundo esforçam-se para se aprimorar, competir e economizar.
- Buscam tornar seu processo de planejamento mais ágil, investigam como administrar inventários e otimizar as alocações de recursos para ter mais benefícios, e buscam agir nas oportunidades conforme surgem em tempo real.
- A análise preditiva pode tornar todos esses objetivos mais atingíveis. Os domínios em que a análise preditiva pode ser aplicada são ilimitados; o campo está aberto, e vale tudo!

COMPREENSÃO PARA PREDIZER

- Quando se quer prever um evento com alguma precisão, é necessário conhecer o passado e entender a situação atual. Isso requer diversos processos:
  - Extrair os fatos que estão acontecendo no momento.
  - Distinguir os fatos passados daqueles que acabaram de acontecer.
  - Deduzir possíveis cenários (criar hipóteses) que poderiam ocorrer.
  - Classificar os cenários (modelos de ML) de acordo com a probabilidade de acontecerem.

COMBINAR PARA PREDIZER

- As análises preditivas geralmente nascem de três ingredientes principais:
  - Conhecimento de negócios.
  - Equipe de ciência de dados e tecnologia.
  - Os dados.

- Embora a proporção desses três ingredientes varie de um negócio para outro, todos são necessários para uma solução de análise preditiva bem-sucedida que resulte em insights acionáveis!

REALIZAR A PREDIÇÃO

- Já vimos várias etapas de um projeto de DS. Para executar o processo do ML, efetivar a predição, os passos são:
  - Carregar os dados.
  - Escolher um algoritmo que realize a tarefa de predição.
  - Treinar o modelo.
  - Visualizar o modelo.
  - Testar o modelo.
  - Avaliar o modelo.

COMO ESCOLHER O ALGORITMO?

- Essa é uma das principais tarefas do cientista de dados, compreender a necessidade e as características dos projetos para poder decidir qual é o melhor algoritmo a ser utilizado.
- Existe uma regra, uma tabela, uma dica?
  - Pode até existir, mas geralmente é a experiência com o uso e aplicação dos algoritmos que fortalece o conhecimento do cientista de dados sobre caminhos a serem seguidos.
  - Saber escolher o algoritmo (sim, precisa testar mais de um), é uma das tarefas árduas do cientista de dados.

COMO ESCOLHER O ALGORITMO?

- É importante saber que os algoritmos podem realizar tarefas diversas.
- A escolha dos algoritmos depende do tipo de modelo adequado (supervisionado ou não supervisionado), que depende dos dados que se têm, e da hipótese que se tem em mente.
- É importante também compreender se é preciso fazer uma regressão, uma classificação, um agrupamento.
- Há algoritmos que têm capacidade de implementar todos esses tipos de modelos.

COMO ESCOLHER O ALGORITMO?

- Uma boa dica é começar os primeiros estudos e projetos com algum tipo de algoritmo, e identificar se ele atende às necessidades, lembrando que mudanças nos atributos (dados de entrada), uso de diferentes técnicas de regularização, mudança nos parâmetros dos algoritmos (e muitos deles têm vários parâmetros) podem representar resultados muito diferentes.
- Observe: estamos tratando aqui apenas de mudanças para um algoritmo, em várias situações estarão disponíveis uma gama de algoritmos que realizam tarefas semelhantes com resultados diferentes.

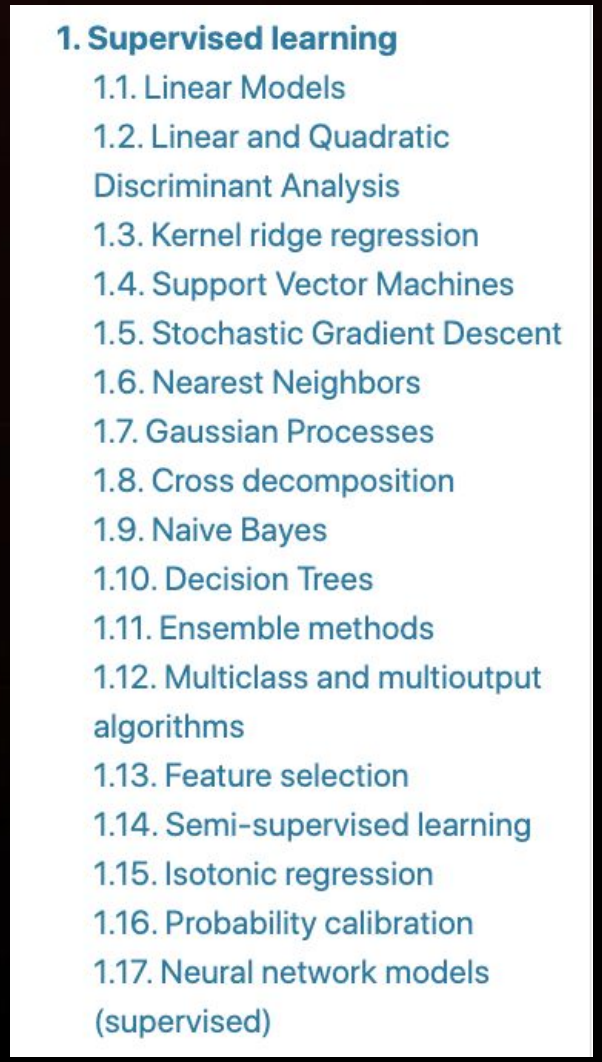
COMO ESCOLHER O ALGORITMO?

- A experiência e o aprofundamento no conhecimento sobre o funcionamento de cada algoritmo vai tornar a tarefa da escolha mais fácil, entretanto, a prática sempre mostra que alguns algoritmos, com customizações adequadas (com a experiência do profissional), vai sempre levar aos melhores resultados.

- Conhecer com profundidade tudo que um algoritmo pode entregar e quais são as características que impactam diretamente nos resultados é muito importante.

COMO ESCOLHER O ALGORITMO?

- Veja esse exemplo:
  - A imagem ao lado representa a disponibilidade de algoritmos para trabalhar com ML supervisionado na biblioteca Scikit-Learn.
  - Há muitas possibilidades para se aplicar modelos que possam gerar os melhores resultados.



COMO ESCOLHER O ALGORITMO?

- Agora veja esse exemplo:
  - Algoritmo SVM, da Scikit-Learn, vamos comentar no slide seguinte.

1.4. Support Vector Machines

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

COMO ESCOLHER O ALGORITMO?

- O algoritmo serve para classificação, regressão e também para detecção de outliers (pode ser usado para pré-processamento).

- Veja que ele indica vantagens de ser usado para conjuntos de dados com dimensões grandes (muitos atributos) e ainda sugere eficácia para conjuntos de dados que a dimensão é maior que a amostra (mais atributos que linhas de dados).

COMO ESCOLHER O ALGORITMO?

- Vejam que isso é uma característica da estrutura dos dados e não do tipo de negócio ou do tipo de dados que se está trabalhando.

- Na imagem também é possível ver que na descrição são apresentadas desvantagens para o algoritmo.

FINALIZANDO

- Predição é um dos objetivos a serem alcançados em projetos de DS.
- A predição pode vir por meio de uma regressão, de uma classificação, de um algoritmo x ou y.
- Há uma infinidade de algoritmos, e muitas características que determinam sua escolha (mesmo antes de se iniciar uma modelagem).
- O objetivo aqui não foi descrever cada um dos algoritmos, mas facilitar a compreensão de que é necessário experimentar e testar algoritmos para os projetos.
- Na próxima aula vamos nos dedicar a um algoritmo específico para trabalhar, o KNN (K-Nearest Neighbors).

REFERÊNCIAS

- Parte do conteúdo dessa aula é baseado nos livros:

Bari, Anasse, et al. Análise Preditiva Para Leigos. Disponível em: Minha Biblioteca, Editora Alta Books, 2019.

Grus, Joel. Data Science do Zero. Disponível em: Minha Biblioteca, (2nd edição). Editora Alta Books, 2021.

INTRODUÇÃO À CIÊNCIA DE DADOS

Predição de Dados