Processos da ciência de dados

Os cientistas de dados seguem um processo semelhante para concluir seus projetos:

1. **Definição do problema de negócios**

O cientista de dados trabalha com os stakeholders para definir claramente o problema que eles desejam resolver ou a pergunta que precisam responder, juntamente com os objetivos e os requisitos de solução do projeto.

2. **Definição da abordagem analítica**
Com base no problema de negócios, o cientista de dados decide qual abordagem analítica seguir:

- Descritiva para obter mais informações sobre o status atual.
- Diagnóstica para entender o que está acontecendo e por quê.
- Preditiva para prever o que vai acontecer.
- Prescritiva para entender como resolver o problema.

3. **Obtenção dos dados**

O cientista de dados identifica e adquire os dados necessários para alcançar o resultado desejado. Isso pode envolver a consulta de bancos de dados, a extração de informações de sites (extração da Web) ou a obtenção de dados de arquivos. Os dados podem estar disponíveis internamente ou talvez seja preciso que a equipe compre os dados. Em alguns casos, as organizações talvez precisem coletar novos dados para poder executar um projeto com êxito.

4. **Limpeza dos dados, também conhecida como depuração**

Normalmente, esta etapa é a mais demorada. Para criar o conjunto de dados para modelagem, o cientista de dados converte todos os dados no mesmo formato, organiza os dados, remove o que não é necessário e substitui os dados ausentes.

5. **Exploração dos dados**

Depois que os dados são limpos, um cientista de dados explora os dados e aplica técnicas analíticas estatísticas para revelar as relações entre os recursos de dados e as relações estatísticas entre eles e os valores que eles predizem (conhecidos como rótulo). O rótulo previsto pode ser um valor quantitativo, como o valor financeiro de algo no futuro ou a duração de um atraso de voo em minutos.

Exploração e preparação costumam envolver uma grande quantidade de visualização e análise de dados interativa, geralmente, usando linguagens como o Python e o R em ferramentas interativas e ambientes projetados especificamente para essa tarefa. Os scripts usados para explorar os dados normalmente são hospedados em ambientes especializados, como o Jupyter Notebooks. Essas ferramentas permitem que os cientistas de dados explorem os dados de forma programática, enquanto documentam e compartilham os insights encontrados.

6. **Modelar os dados**

O cientista de dados cria e treina modelos prescritivos ou descritivos e, em seguida, testa e avalia o modelo para garantir que ele responda à pergunta ou corrija o problema de negócios. Em sua

forma mais simples, um modelo é um trecho de código que usa uma entrada e produz uma saída. Criar um modelo de machine learning envolve selecionar um algoritmo, fornecer dados a ele e ajustar hiperparâmetros. Os hiperparâmetros são parâmetros ajustáveis que permitem controlar o processo de treinamento do modelo. Por exemplo, com redes neurais, o cientista de dados decide o número de camadas ocultas e o número de nós em cada camada. Oajuste de hiperparâmetro, também chamado de otimização de hiperparâmetro, é o processo de localizar a configuração de hiperparâmetros que resulta no melhor desempenho.

Uma pergunta comum é "Qual algoritmo de machine learning devo usar?" Um algoritmo de machine learning transforma um conjunto de dados em um modelo. O algoritmo que o cientista de dados seleciona depende principalmente de dois aspectos diferentes do cenário de ciência de dados:

- Qual é a pergunta comercial que o cientista de dados deseja responder aprendendo com os dados anteriores?
- Quais são os requisitos do cenário de ciência de dados, incluindo precisão, tempo de treinamento, linearidade, número de parâmetros e número de recursos?

Para ajudar a responder as perguntas. o Azure Machine Learning fornece um portfólio abrangente de algoritmos, como floresta de decisão multiclasse, sistemas de recomendação, regressão de rede neural, rede neural multiclasse e cluster K-Means. Cada algoritmo foi projetado para atender a um tipo diferente de problema de machine learning. Além disso, a Folha de referências do algoritmo de machine learning ajuda os cientistas de dados a escolher o algoritmo certo para responder à pergunta de negócios.

7. **Implantar o modelo**

O cientista de dados entrega o modelo final com documentação e implanta o novo conjunto de dados em produção após o teste, para que ele possa desempenhar um papel ativo em uma empresa. Previsões de um modelo implantado podem ser usadas para decisões de negócios.

8. **Visualizar e comunicar os resultados**

Ferramentas de visualização como Microsoft Power BI, Tableau, Apache Superset e Metabase facilitam a exploração dos dados pelo cientista de dados e geram belas visualizações que mostram as descobertas de uma maneira simples para o público não técnico entender.

Os cientistas de dados também podem usar notebooks de ciência de dados baseados na Web, como Notebooks Zeppelin, durante todo o processo de ingestão, descoberta, análise, visualização e colaboração de dados.

Métodos de ciência de dados

Os cientistas de dados usam métodos estatísticos, como teste de hipóteses, análise fatorial, análise de regressão e agrupamento para descobrir insights estatisticamente sólidos.

Documentação da ciência de dados

Embora a documentação de ciência de dados varie de acordo com o projeto e o setor, ela geralmente inclui a documentação que mostra de onde vêm os dados e como eles foram modificados. Isso ajuda outros membros da equipe de dados a usar efetivamente os dados no

futuro. Por exemplo, a documentação ajuda os analistas de negócios a usar ferramentas de visualização para interpretar o conjunto de dados.

Os tipos de documentação da ciência de dados incluem:

- **Planos do Project** para definir os objetivos de negócios do projeto, métricas de avaliação, recursos, cronograma e orçamento.
- **Histórias de usuários de ciência de dados** para gerar ideias para projetos de ciência de dados. O cientista de dados escreve a história do ponto de vista do stakeholder, descrevendo o que o stakeholder gostaria de alcançar e o motivo pelo qual o stakeholder está solicitando o projeto.
- **Documentação do modelo de ciência de dados** para documentar o conjunto de dados, o design do experimento e os algoritmos.
- **Documentação de sistemas de suporte**, incluindo guias de usuário, documentação de infraestrutura para manutenção do sistema e documentação de código.