O processo de Data Science

Jorge Cristhian Chamby-Diaz

PhD em Ciência da Computação

May 2, 2024

E-mail: jchambyd@gmail.com

Sobre mim ...

JORGE CRISTHIAN CHAMBY DIAZ

- Engenheiro de Computação - **UNSA**
- Mestre e PhD em Ciência da Computação - **UFRGS**
- Cientista de Dados Especialista - **Samsung R&D**
- Professor de Ciência de Dados - **Ada Tech**
- Github: @jchambyd
- Instagram: @jorge.chamby

@jchambyd

Roteiro

Roteiro

Ciclo de vida de Data Science

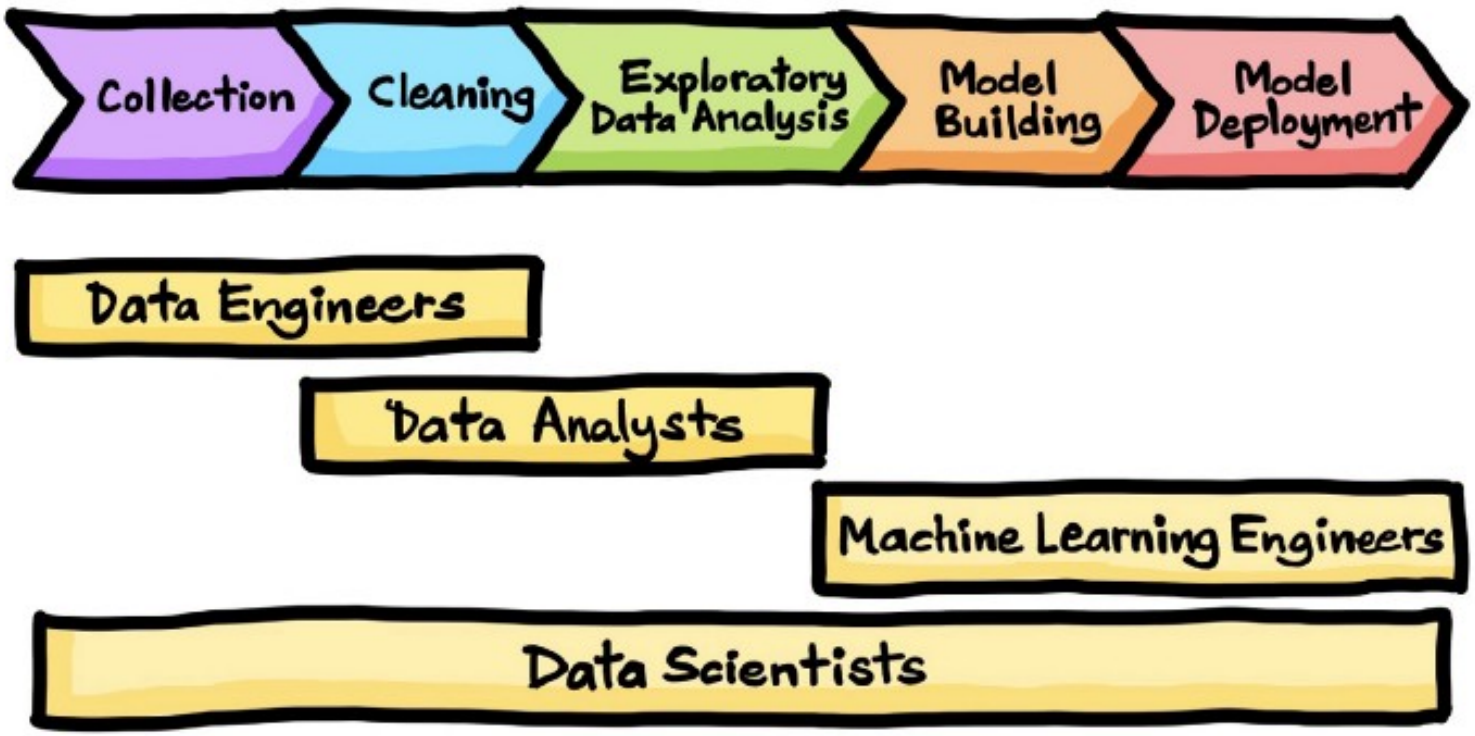


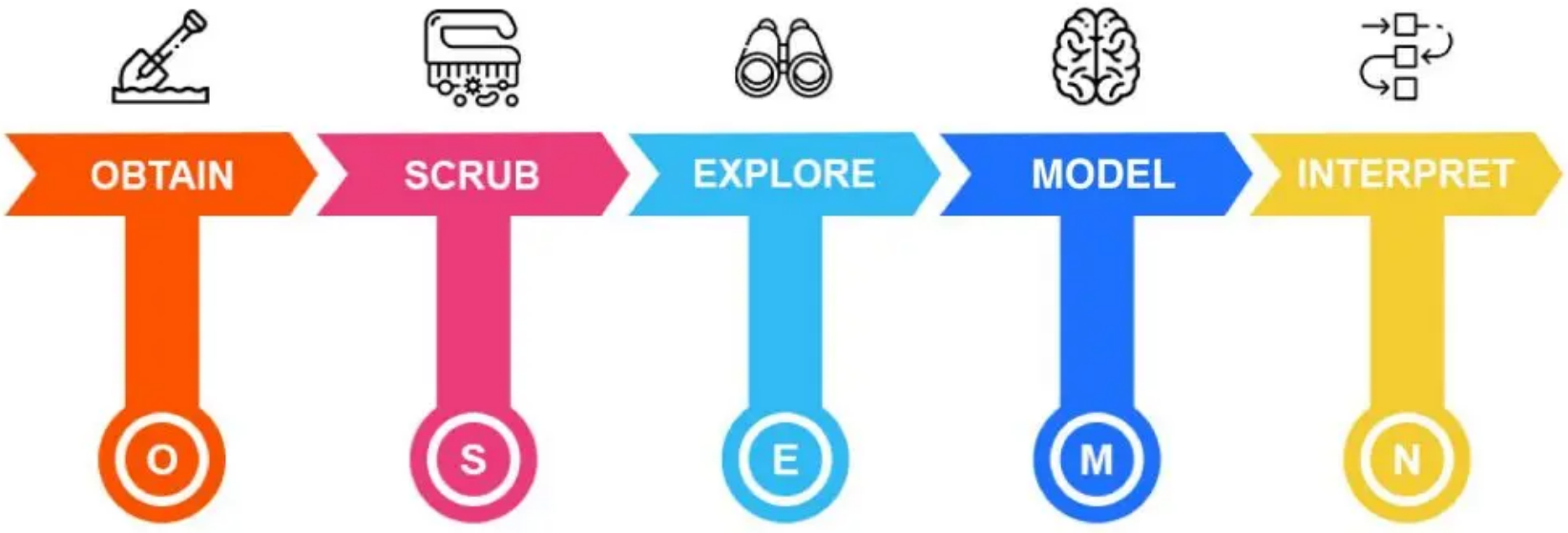
Figure: Ciclo de vida de Data Science (C. Nantasenamat)

Ciclo de vida de Data Science

- Muitas vezes, quando falamos de projetos de ciência de dados, resulta complicado apresentar uma explicação sólida de como todo o processo ocorre. Desde a coleta de dados, até a análise e apresentação dos resultados.
- Nesta apresentação vamos usar o framework OSEMN, que abrange todas as etapas do ciclo de vida do projeto de ciência de dados de ponta a ponta.

Roteiro

O Framework OSEMN

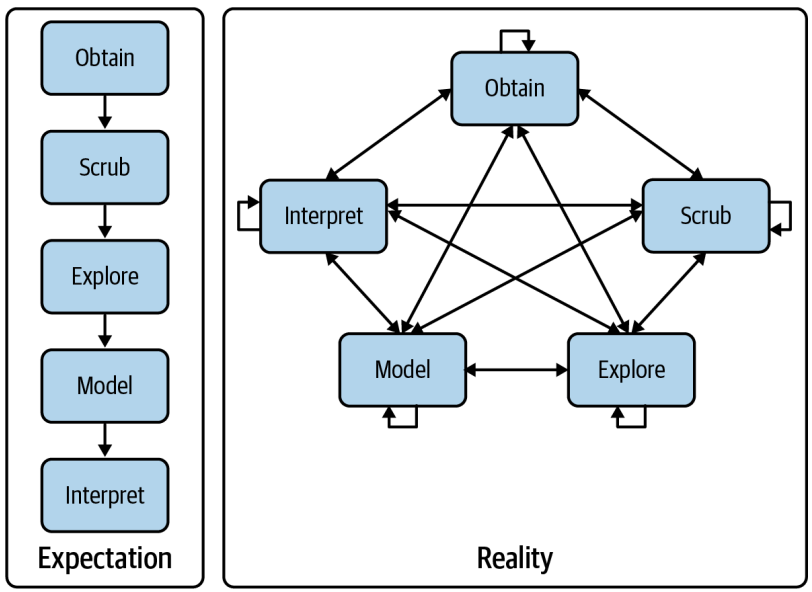


O Framework OSEMN

Embora as cinco etapas sejam discutidas de forma linear e incremental, na prática é muito comum alternar entre elas ou realizar várias etapas ao mesmo tempo.

O Framework OSEMN

- A Figura ilustra que fazer ciência de dados é um processo iterativo e não linear.
- Por exemplo, depois de modelar seus dados e observar os resultados, você pode decidir voltar à etapa de depuração para ajustar os recursos do conjunto de dados.



Etapa 1: Obter dados

- A primeira etapa de um projeto de ciência de dados é direta, obtemos os dados de que precisamos de fontes de dados disponíveis.

Etapa 1: Obter dados - Procedimentos

- Baixar dados de outro local (por exemplo, uma página da Web ou servidor)
- Consultar dados de um banco de dados ou API (por exemplo, MySQL ou Twitter)
- Extraia dados de outro arquivo (por exemplo, um arquivo HTML ou planilha)
- Gere dados você mesmo (por exemplo, lendo sensores ou fazendo pesquisas)

Etapa 1: Obter dados - Habilidades requeridas

- Para executar as tarefas acima, são necessárias certas habilidades técnicas.
- Para gerenciamento de banco de dados, é importante saber usar MySQL, PostgreSQL ou MongoDB (se estiver usando um conjunto de dados não estruturado).
- Para projetos em conjuntos de dados muito maiores ou big data, é necessário aprender a acessar usando armazenamento distribuído como Apache Hadoop, Spark ou Flink.

Etapa 2: Limpeza de dados

- Depois de obter os dados, a próxima coisa imediata a fazer é depurar os dados. Este processo é para nós "limpar" e filtrar os dados.
- Se os dados fossem não filtrados e irrelevantes, os resultados da análise não significarão nada.

Etapa 2: Limpeza de dados - Operações comuns

- Filtrando linhas
- Extraindo certas colunas
- Substituindo valores
- Extraindo palavras
- Lidando com valores ausentes e duplicatas
- Converter dados de um formato para outro

Etapa 2: Limpeza de dados - Habilidades requeridas

- Para esta etapa são necessárias ferramentas de código como Python ou R para ajudar a limpar os dados.
- Caso contrário, podemos usar uma ferramenta de código aberto como o OpenRefine ou adquirir um software corporativo como o SAS Enterprise Miner para facilitar esse processo.
- Para lidar com conjuntos de dados maiores, é necessário ter habilidades em Hadoop, Map Reduce ou Spark. Essas ferramentas podem ajudar a limpar os dados por meio de código.

Etapa 3: Exploração de dados

- Normalmente, em um ambiente corporativo ou de negócios, seu chefe apenas lança um conjunto de dados e cabe a você entendê-lo. Portanto, caberá a você ajudá-los a descobrir a questão de negócios e transformá-la em uma questão de ciência de dados.
- É aqui que fica interessante, porque quando você estiver explorando, você realmente conhecerá seus dados.

Etapa 3: Exploração de dados - Operações comuns

- Inspecionar os dados e suas propriedades. Diferentes tipos de dados como dados numéricos, dados categóricos, dados ordinais e nominais etc. requerem tratamentos diferentes.
- Computar estatísticas descritivas para extrair características e testar variáveis significativas.
- Visualização de dados para nos ajudar a identificar padrões e tendências significativos em nossos dados.

Etapa 3: Exploração de dados - Habilidades requeridas

- Se você estiver usando Python, precisará saber como usar Numpy, Matplotlib, Pandas ou Scipy; se você estiver usando R, precisará usar o GGplot2 ou o canivete suíço de exploração de dados Dplyr.
- Além disso, você precisa ter conhecimento e habilidades em estatística inferencial e visualização de dados.
- Por mais que você não precise de mestrado ou doutorado, para fazer ciência de dados, essas habilidades técnicas são cruciais para conduzir um projeto experimental, para que você possa reproduzir os resultados.

Etapa 4: Modelagem de dados

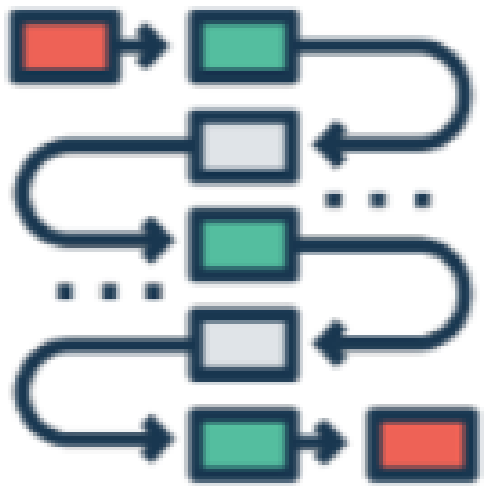
- Esta é a fase em que a maioria das pessoas considera interessante. Como muitos chamam de "onde a mágica acontece".
- Se você deseja explicar os dados ou prever o que acontecerá, provavelmente deseja criar um modelo estatístico de seus dados. As técnicas para criar um modelo incluem agrupamento, classificação, regressão e redução de dimensionalidade.

Etapa 4: Modelagem de dados - Habilidades requeridas

- Em Machine Learning, as habilidades necessárias são algoritmos supervisionados e não supervisionados. Para bibliotecas, se você estiver usando Python, precisará saber como usar o Scikit Learn; e se você estiver usando R, precisará usar CARET.
- Após o processo de modelagem, você precisará calcular pontuações de avaliação, como precisão, recall e pontuação F1 para classificação. Para regressões, você precisa estar familiarizado com R^2 para medir a qualidade do ajuste e usar pontuações de erro como MAE (Erro médio médio) ou RMSE (Erro quadrado médio médio) para medir a distância entre os pontos de dados previstos e observados.

Etapa 5: Interpretação de dados

- Interpretar dados refere-se à apresentação de seus dados a um leigo não técnico.
- Entregamos os resultados para responder às perguntas de negócios que fizemos quando iniciamos o projeto, juntamente com os insights acionáveis que encontramos por meio do processo de ciência de dados.



Etapa 5: Interpretação de dados - Operações comuns

- Tirar conclusões de seus dados
- Avaliar o que seus resultados significam
- Comunicar seus resultado

Etapa 5: Interpretação de dados - Habilidades requeridas

- Nesse processo, as principais habilidades a serem adquiridas estão além das habilidades técnicas. Você precisará de um forte conhecimento do domínio de negócios para apresentar suas descobertas de uma maneira que possa responder às perguntas de negócios que você se propôs a responder e traduzi-las em etapas acionáveis.
- Além de ferramentas necessárias para visualização de dados, como Matplotlib, ggplot, Seaborn, Tableau, d3js etc., você precisará de habilidades sociais, como habilidades de apresentação e comunicação, combinadas com um talento para relatórios e habilidades de escrita, definitivamente o ajudarão nesta fase do projeto ciclo da vida.

Resumo

- Se for um projeto totalmente novo, geralmente gastamos cerca de 60 a 70% do nosso tempo apenas coletando e limpando os dados.
- OSEMN, como é um framework, você pode usá-lo como guia com suas ferramentas favoritas.
- O verdadeiro norte são sempre aquelas questões de negócios que definimos, antes mesmo de iniciar o projeto de ciência de dados.
- Lembre-se sempre de que questões de negócios sólidas, dados limpos e bem distribuídos sempre superam modelos sofisticados.

Perguntas & Respostas