Podcast

Disciplina: Linguagens de programação para ciência de dados
Título do tema: Análise estatística de dados
Autoria: Yuri Vasconcelos de Almeida Sá
Leitura crítica: Henrique Salustiano Silva

Abertura:

Olá! No podcast de hoje vamos falar sobre volume de dados e prototipagem, vamos ver como criar e testar um modelo antes de testar ele no banco de dados completo.

Na era da informação, *Big Data* é a tendência de análise e o caminho mais seguido no contexto do *Data Science* para aquisição e organização dos dados.

Como o próprio nome diz, o volume de dados é imenso e pode trazer vários problemas na hora de desenvolvermos um modelo ou análise, ainda que usemos computadores potentes, testes rápidos e o desenvolvimento pode ser muito difícil se o volume de dados for grande o suficiente.

Uma prática usual que geralmente é adotada é o *sampling*, que é a prática de retirar amostras que representem o todo de forma fiel e integra.

Caso seus dados não sejam temporais, pode-se adotar uma estratégia mais direta de pegar uma porcentagem aleatória dos dados para que a análise possa ser desenvolvida e testada com hardware não especializado e local com uma performance aceitável e só então escalada para o *dataset* completo, similar ao de produção.

Caso seu *dataset* seja uma série temporal, simplesmente amostrar os dados aleatoriamente pode trazer prejuízos, pois não existe então a continuidade dos dados, tornando-os inúteis. Neste caso o melhor é selecionar pontos aleatórios ao longo do *dataset* e criar *DataFrames* com um número padrão de registros anteriores, a ideia aqui é simular um período, ainda que curto. Este tipo de estratégia cria pacotes de dados com começo, meio e fim e ajuda no passo inicial de prototipação da análise e modelo.

Estes são dois exemplos simplórios de *sampling*, mas ainda assim muito utilizados no mercado. Como tudo em Data Science, existem diversos outros métodos com seus casos de uso e melhores aplicações. Os principais cuidados a serem tomados são relacionamentos com outros *datasets* e agrupamentos.

O importante quando você for “samplear” seu *dataset* é garantir que você vai manter uma representação fiel dos seus dados, sem que haja, por exemplo um enviesamento da informação devido a amostragem, o que pode anular os esforços durante a fase de protótipo e não apresentar os mesmos resultados quando aplicados no *dataset* completo.

A prática, principalmente em *datasets* correlatos vai acabar trazendo maneiras que vão afiando seu senso de amostragem, mas nem por isso deve-se baixar a guarda. Podemos gastar muito tempo em uma análise fadada ao fracasso só porque não nos atentamos as especificidades do *dataset* e regras do negócio,

tendencias mudam e as estratégias também. O que nos leva a outro assunto, a melhoria contínua do modelo.

Mesmo que o modelo já esteja em produção, a amostragem surge como ferramenta de verificação da eficiência do modelo original, criando espaço para melhoria e adaptação aos novos tempos.

Embora só tenhamos citados poucos exemplos nesta edição do podcast, o *sampling* é muito rico e vasto e é feito através de algoritmos e scripts, então em um projeto até isso deve ser armazenado e catalogado junto as métricas originais, então você pode extrair métricas das amostras, como desvio e tendências, onde você pode fazer a comparação da evolução e eventualmente concluir que seu modelo está perdendo eficiência e deve ser ajustado.

Quando você for se aventurar na sua amostragem, preste atenção nestes detalhes e garanta a integridade dos dados que já é um ótimo começo.

Fechamento:

Este foi nosso podcast de hoje! Até a próxima!