**A Importância da Estatística para o Cientista de Dados**

O Cientista de Dados é a versão adulta do garoto que não para de perguntar: “ Por que? “. Na prática a profissão de Cientista de Dados engloba diversas áreas de conhecimento e este profissional precisa ter o perfil de alguém capaz de aprender sob demanda e dominar diferentes técnicas e ferramentas.

E uma destas ferramentas de trabalho do Cientista de Dados é a Estatística, uma parte crucial da Ciência de Dados. A verdade é que o Big Data foi o combustível que ajudou a inflamar esta importante área, que envolve Matemática e Análise. Os Estatísticos precisam de dados para tornar suas análises precisas e o Big Data oferece esta matéria-prima em abundância.

Existem três grandes etapas principais em um projeto de Ciência de Dados: você coleta e organiza os dados, analisa e em seguida comunica os resultados. Mas é raro que este processo siga em uma única direção: muitas vezes a sua análise irá revelar que você precisa de mais dados ou dados diferentes, ou quando apresentar resultados você vai descobrir uma falha em seu modelo. Acredite, isso vai acontecer. A Estatística possui diversas técnicas para coleta de dados: amostragem e planejamento de experimentos são campos bem estabelecidos e apoiados por décadas de pesquisa. Essas ferramentas ajudam a coletar e compreender os dados.

Após a coleta, os dados precisam ser organizados (ou normalizados) em um formato que permita a análise. Organizar os dados no formato adequado é essencial para a análise de dados e se feito de forma incorreta, você vai passar a maior parte do seu tempo lutando contra suas ferramentas analíticas, não analisando os dados.

Qualquer análise de dados envolve a manipulação de dados (Data Wrangling ou Data Munging), visualização e modelagem. Visualização e modelagem são complementares. Utilizar as visualizações ao longo do processo, ajuda a refinar sua análise. No entanto, visualizações dependem de interpretação humana e a capacidade de escala é fundamentalmente restrita. A Estatística pode ajudar em diversas etapas do processo e os algoritmos de Machine Learning são modelos preditivos baseados em Estatística.

O produto final de uma análise não é um modelo: é uma solução. Uma análise é sem sentido a menos que você convença alguém a agir. Nos negócios, isso normalmente significa apresentar soluções para problemas, o que pode significar aumento de vendas ou mesmo sobrevivência de uma empresa.

Vejamos 6 perguntas de negócio que o Cientista de Dados pode responder com a ajuda da Estatística:

**1. Por que o anúncio A tem uma taxa 5% maior de cliques do que anúncio B em nosso web site?**

Os Cientista de Dados podem ajudar a determinar se essa diferença é significativa o suficiente para justificar maior atenção, foco e investimento. Eles podem ajudar a entender os resultados experimentais e isso é especialmente útil quando você tem muitas métricas, realizando experimentos que afetam uns aos outros.

Vamos dizer que você é um varejista nacional e que você está tentando testar o efeito de uma nova campanha de marketing. O Cientista de Dados pode ajudar a decidir que dados devem ser obtidos, de quais fontes e como atribuir grupos experimentais para obter um bom equilíbrio entre o tamanho da amostra que você deve selecionar para obter resultados claros, e como executar experimentos com o menor custo possível.

Técnicas Estatísticas utilizadas: Design Experimental, Testes de Hipóteses e Intervalos de Confiança.

**2. Por que as vendas em dezembro aumentaram em 5%?**

Os Cientistas de Dados podem usar Estatística para descobrir possíveis razões pelas quais as vendas aumentaram em 5% e podem ainda ajudar a compreender o que impulsiona as vendas, qual a previsão de vendas para o próximo mês e potenciais tendências.

Técnicas Estatísticas utilizadas: Regressão, Classificação, Análise de Séries Temporais, Análise Causal.

**3. Por que alguns clientes compram apenas alimentos saudáveis, enquanto outros só compram quando há promoção?**

Qualquer um pode observar que uma empresa tem 100.000 clientes que compram 10.000 itens em sua loja. Os Cientistas de Dados podem ajudar a identificar cada cliente, agrupá-los com os clientes semelhantes e entender seus hábitos de compra. Isto permite ver como o desenvolvimento de novos negócios pode afetar certos grupos da população e assim direcionar a criação de novos produtos com maior probabilidade de acerto.

Técnicas Estatísticas utilizadas: Clustering, Redução de Dimensionalidade

**4. Várias pessoas estão se inscrevendo em nosso site e nunca mais voltam. Por que?**

Por que seus clientes compram itens de seu site? Como você faz com que seus clientes retornem? Por que os usuários abandonam seu web site? Que tipos de e-mails da sua empresa engajam os usuários? Quais são alguns dos principais indicadores de engajamento, atividade ou sucesso? Quais são algumas boas oportunidades de vendas? Todas são perguntas que podem ser respondidas com a ajuda da Estatística.

Técnicas Estatísticas utilizadas: Regressão, Análise de Causa e Efeito

**5. Dada uma matriz de usuários (clientes) e suas interações (cliques, compras, classificações) com seu web site de e-commerce (anúncios, produtos, filmes), você pode sugerir quais itens os usuários vão querer comprar?**

Web sites são fontes abundantes de dados, com todos os cliques e interações dos usuários. Cada clique representa um comportamento. E se pudermos obter estes cliques por região geográfica, por estado, por faixa etária, por grau de escolaridade? E se juntarmos tudo isso, cruzarmos os dados com dados de vendas e criarmos modelos preditivos? Uma empresa poderia prever com precisão quanto será o faturamento ou quantos funcionários terão que ser contratados, criando políticas e processos internos muito mais precisos. Tudo isso já está sendo feito e com a ajuda da Estatística.

Técnicas Estatísticas utilizadas: Modelagem Preditiva, Redução de Dimensionalidade, Filtragem Colaborativa, Clustering

**6. Por que nós temos um banner com 100 impressões e 0 cliques?**

0% é uma boa taxa de cliques em um banner publicitário? Com certeza não. Os Cientistas de Dados podem reunir dados para obter uma estimativa desejável, prever as propriedades dessa estimativa e compreender qual a probabilidade de cliques. Facebook e Google são bons exemplos.

Técnicas Estatísticas utilizadas: Análise Bayesiana

O papel do Cientista de Dados é servir como embaixador entre os dados e a empresa. A comunicação é fundamental e o Cientista de Dados deve ser capaz de explicar suas ideias de uma forma que a empresa possa obter os benefícios que o Big Data Analytics pode trazer. O Cientista de Dados não resume simplesmente os números, mas explica por que os números são importantes e que insights se pode obter a partir deles.

Sabemos que Ciência de Dados não é apenas Estatística. Ciência da Computação e Técnicas de Apresentação tem um papel fundamental ao longo do processo. Mas sem a Estatística, o Cientista de Dados terá seu trabalho limitado e sua capacidade analítica comprometida.

Por isso este e o próximo capítulo são tão importantes. Abra sua mente, aprenda coisas novas e descubra como se tornar um profissional diferenciado. O mercado está a sua procura.