Este capítulo aborda:

Elucidando as diferenças entre um cientista de dados e um engenheiro de ML

Focando na simplicidade em todo o trabalho do projeto para reduzir riscos

Aplicando os fundamentos do Agile ao trabalho de projeto de ML

Ilustrando as diferenças e semelhanças entre DevOps e MLOps

A simples verdade desta profissão é que o trabalho de projeto que se concentra em cada um desses elementos geralmente é bem-sucedido, enquanto aqueles que omitem um ou mais deles tendem ao fracasso. Esta é a razão das taxas de fracasso hiperbólicas e frequentemente citadas de projetos de ML na indústria (que considero bastante egoístas e alimentadas pelo pânico quando provenientes de materiais de marketing de fornecedores). Este capítulo aborda, em alto nível, esse trio de componentes para projetos bem-sucedidos. Empregar o equilíbrio apropriado de cada um, com foco na criação de soluções sustentáveis que sejam desenvolvidas em conjunto com clientes internos de forma colaborativa e inclusiva, aumentará muito as chances de construir soluções de ML duradouras. Afinal, o foco principal de todo o trabalho do DS é resolver problemas. A conformidade dos padrões de trabalho com uma metodologia comprovada focada na capacidade de manutenção e na eficiência se traduz diretamente na resolução de mais problemas com muito menos esforço.

**Aumentando uma profissão complexa com processos para aumentar o sucesso do projeto**

Em uma das primeiras definições do termo ciência de dados, conforme abordado em Data Science, Classification, and Related Methods (Springer, 1996), compilado por C. Hayashi et al., os três focos principais são os seguintes:

*Design para dados* – Especificamente, o planejamento em torno de como a informação deve ser coletada e em que estrutura ela precisará ser adquirida para resolver um problema específico

*Coleta de dados* – O ato de adquirir os dados

*Análise em dados* – Adivinhar insights a partir dos dados por meio do uso de metodologias estatísticas para resolver um problema.

Grande parte da ciência de dados moderna concentra-se principalmente no último desses três itens (embora, em muitos casos, uma equipe de DS seja forçada a desenvolver seu próprio ETL), já que os dois primeiros geralmente são gerenciados por uma equipe moderna de engenharia de dados. Dentro deste termo amplo, análise de dados, reside um grande foco do DS moderno: aplicação de técnicas estatísticas, atividades de manipulação de dados e algoritmos estatísticos (modelos) para obter insights e fazer previsões sobre os dados. A parte superior da figura 2.1 ilustra (de maneira intencionalmente breve e de alto nível) o foco do cientista de dados moderno de uma perspectiva técnica. Estes nosão os elementos da profissão em que a maioria das pessoas se concentra quando falam sobre o que fazemos: desde o acesso a dados até à construção de modelos preditivos complexos, utilizando uma variedade estonteante de abordagens algorítmicas e estatísticas avançadas. Não é uma avaliação particularmente precisa do que um cientista de dados realmente faz ao realizar um trabalho de projeto, mas concentra-se em algumas das tarefas e ferramentas empregadas na resolução de problemas. Pensar na ciência de dados dessa maneira é quase tão inútil quanto classificar o trabalho de um desenvolvedor de software listando linguagens, algoritmos, estruturas, eficiência computacional e outras considerações tecnológicas de sua profissão.

Podemos ver na figura 2.1 como o foco tecnológico do DS na parte superior (que muitos profissionais focam exclusivamente) é apenas um aspecto do sistema mais amplo mostrado na parte inferior. É nesta região, a engenharia de ML, que as ferramentas, processos e paradigmas complementares fornecem uma estrutura de orientação, apoiada fundamentalmente pelos aspectos centrais da tecnologia DS, para trabalhar de uma forma mais construtiva. A engenharia de ML, como conceito, é um paradigma que ajuda os profissionais a se concentrarem no único aspecto do trabalho do projeto que realmente importa: fornecer soluções para problemas que realmente funcionam. Por onde começar?

2 Uma base de simplicidade

Quando se trata de explicar verdadeiramente o que os cientistas de dados fazem, nada pode ser mais sucinto do que “Eles resolvem problemas através da aplicação criativa da matemática aos dados”. Por mais amplo que seja, reflete a ampla gama de soluções que podem ser desenvolvidas a partir de informações (dados) registradas. Nada é prescrito em relação às expectativas do que um DS faz em relação a algoritmos, abordagens ou tecnologia enquanto busca resolver um problema de negócios. Muito pelo contrário, na verdade. Somos solucionadores de problemas, utilizando uma ampla gama de técnicas e abordagens. Infelizmente para os recém-chegados à área, muitos cientistas de dados acreditam que estão agregando valor a uma empresa apenas quando usam a tecnologia mais recente e “melhor” que surge. Em vez de se concentrar nas últimas novidades em torno de uma nova abordagem catalogada em um whitepaper seminal ou fortemente anunciada em uma postagem de blog, um DS experiente percebe que a única coisa que realmente importa é o ato de resolver problemas, independentemente da metodologia. Por mais empolgantes que sejam as novas tecnologias e abordagens, a eficácia de uma equipe de DS é medida pela qualidade, estabilidade e custo de uma solução que ela fornece. Como mostra a figura 2.2, uma das partes mais importantes do trabalho de ML é navegar pelo caminho da complexidade ao enfrentar qualquer problema. Ao abordar cada nova solicitação de uma empresa com essa mentalidade como a verdadeira pedra angular dos princípios de ML (focando na solução mais simples possível que resolva o problema da empresa), a solução em si pode ser focada, em vez de uma abordagem específica ou um novo algoritmo sofisticado. Ter um foco construído em torno deste princípio – de buscar a implementação mais simples possível para resolver um problema – é a base sobre a qual todos os outros aspectos da engenharia de ML são construídos. É de longe o aspecto mais importante da engenharia de ML, pois informará todos os outros aspectos do trabalho do projeto, escopo e detalhes de implementação. Esforçar-se para sair do caminho o mais cedo possível pode ser o maior fator determinante para determinar se um projeto irá falhar.

**Mas não é trabalho de ciência de dados se a solução não usar IA”**

Nunca entrei nesta carreira com expectativas de usar tecnologia, um algoritmo, estrutura ou metodologia específica. Conheci muitas pessoas que o fizeram, e muitas que conheci ao longo de suas jornadas profissionais, acabaram ficando surpresas com o quão pouco acabaram usando uma estrutura ou biblioteca específica frequentemente mencionada para seu trabalho. A maioria deles ficou especialmente surpresa com quanto tempo gastou escrevendo SQL, realizando análises estatísticas de seus dados e limpando dados confusos para resolver um problema. Suponho que nunca tive aquela experiência aparentemente desmoralizante que muitos de meus colegas tiveram em relação à aplicação pouco frequente de abordagens de ponta no “mundo real”, porque comecei na análise antes de passar para o ML muito mais tarde. Aprendi no início da transição para esta área que as soluções mais simples para os problemas sempre eram a melhor abordagem. A razão pouco sofisticada para isso é bastante simples: eu tive que manter a solução. Seja mensalmente, diariamente ou em tempo real, minha solução e código eram coisas que eu precisaria para depurar, melhorar, solucionar inconsistências e, francamente, apenas continuar executando. Quanto mais sofisticada for uma determinada solução, mais tempo levará para diagnosticar falhas, mais difícil será solucionar problemas e mais frustrante será alterar sua lógica interna para obter recursos adicionais. O objetivo de buscar a simplicidade nas soluções (isto é, o design e a abordagem mais simples que ainda resolvem o problema) se traduz diretamente em menos tempo gasto na manutenção de soluções para problemas que você resolveu. Isso libera você para resolver mais problemas, agregar mais valor à sua empresa e, geralmente, lhe dá exposição a mais problemas. Já vi a paixão que as pessoas têm pelo uso de algoritmos interessantes funcionar mal muitas vezes. Um dos mais notáveis ​​foi um GAN para aumento de resolução de imagem que levou 10 meses para uma equipe de 12 cientistas de dados chegar a um estado pronto para produção e escalonável. Ao conversar com sua equipe de nível C, eles disseram que estavam contratando consultores para construir um modelo de rotatividade, um modelo de fraude e um modelo de previsão de receitas. Eles sentiram que precisavam contratar consultores externos para fazer o importante trabalho crítico de modelagem porque sua equipe interna estava muito ocupada trabalhando em um projeto de P&D. Nas 12 semanas de trabalho com aquela empresa, toda a equipe do DS foi demitida e o projeto de imagem abandonado. Às vezes, trabalhar nas coisas básicas que agregam valor incrível a uma empresa pode ajudá-lo a manter seu emprego (o que não quer dizer que a previsão, a rotatividade e a modelagem de fraude sejam simples, mesmo que não pareçam particularmente interessantes).