Engenharia de Recursos

Neste capítulo 5 a autora enfatiza que o sucesso dos sistemas de Machine Learning (ML) depende dos recursos utilizados, portanto, é fundamental que as pessoas e/ou organizações interessadas em implementar o ML dediquem tempo e esforço considerável à engenharia de recursos, pois a eficácia e desempenho dos modelos são afetados diretamente pela relevância e qualidade dos dados. Projetar bons recursos é uma tarefa complexa e se faz necessário uma experiência prática para executar tal função. Observar as técnicas utilizadas em projetos passados bem sucedidos pode ajudar nesse processo.

Engenharia de resursos envolve não apenas conhecimento técnico, mas também compreesão do domíno do problema, pois é essencial entender como os dados são gerados, coletados e processados, envolver especialistas sempre que possível pode facilitar essa tarefa. Assim é possível estabelecer boas práticas para a criação e seleção dos recursos, pautado em estratégias como:

1. Divisão de dados em conjunto de treinamento, validação e teste, ao invés de fazê-lo aleatoriamente, isso ajuda a evitar vazamento de informações e permite uma avaliação mais precisa do desempenho do modelo.
2. Se for necessário realizar amostragem dos dodos, é sugerido fazê-lo após a divisão dos conjuntos de dados. Isso ajuda a garantir que não ocorra o vazamento de informação entre os conjutos de treinamento e teste.
3. Recomenda-se lidar com dados ausentes de forma apropriada, ou seja, de forma característica ou conveniente, assim utilizando estatísticas apenas do conjunto de treinamento para dimensionar recursos e lidar com valores ausentes.
4. Dimensionar e normalizar os dados após a divisão dos conjuntos ajuda a evitar vazamento de informações e garante que os recursos tenham escalas comparáveis, o que deixa o processamento dos dados pela rede muito mais eficiente, e sem viesses.
5. Utilizar apenas os recusos mais relevantes e informativos para o modelo, removendo aqueles que não contribuem significativamente para a sua capacidade preditiva.

   5.1. Pois quando um modelo é “alimentado” com muitos recursos irrelevantes, ele pode sofre com a “maldição da dimensionalidade”, onde o espaço de busca se torna muito grande em relação ao tamanho do conjunto de dados. Isso pode levar a um aumento no tempo de treinamento do modelo e na complexidade computacional, sem necessariamente melhorar o desempenho predivo.

   5.2. Modelos de ML podem se ajustar demais aos dados de treinamento quando são fornecidos muitos recursos irrelevantes. Isso pode resultar em um modelo que se adapta excessivamente aos detalhes de uma forma incomum dos dados de treinamento, mas não generaliza bem para novos dados.

   5.3. Quando menos recursos um modelo possui, mais fácil é interpretar e entender como as decisões são tomadas. Modelos com muitos recursos podem ser mais difíceis de interpretar, o que pode ser problemático em contextos onde a explicabilidade do modelo é importante.
6. Reconhecer que o processo de engenharia de recursos é contínuo e não termina coma implementação do modelo, É necessário continuar refinando e ajustando os recursos com base em novos dados e insights.

Essas técnicas auxiliam na criação de modelos mais precisos, robustos e eficazes. Isso resulta em melhor desempenho do modelo, maior capacidade de generalização e, em última análise, melhores resultados para a organização que está utilizando o modelo.