Latest AI News

Three Subtle Examples of Data Leakage by abstractapplic

Source: Featured posts - LessWrong 2.0 viewer
Published: Tue, 01 Oct 2024 20:45:27 +0000

URL:
https://www.greaterwrong.com/posts/rzyHbLZHuqHq6KM65/three-subtle-examples-of-data-leakage

Summary: The article discusses the author's experiences with data science projects, focusing on the concept of 'Data Leakage'—the inappropriate use of information during training or evaluation that wouldn't be available in deployment. The author recounts several cases where they detected and addressed potential leakages. One instance involved modeling auction bids while excluding data above a certain price point, which led to flawed assumptions about the model's predictions. Another case emphasized the importance of chronological data splits over random sampling to avoid misleading results, revealing that treating data as time-travel led to inaccuracies. The final example highlighted a scenario where the author initially struggled with a Tobit model and how they eventually developed a leak-proof solution. Key takeaways from these experiences include acknowledging that leakages always have costs, the varying tolerability of leakages in different contexts, and the necessity of detecting leakages even when their full impact is unclear. The article concludes with a reflection on the challenges of identifying and quantifying the damage caused by leakages, drawing parallels between data leakages and general reasoning errors.