

Data drift vs knowledge drift & RAG

- Knowledge drift is when facts or information in the real world change (eg new laws, events or discoveries) so the model's knowledge becomes outdated
- Data drift is when the statistical properties of the input data changes (eg different user behaviour or market trends) so the model's inputs (Features) no longer match what it was trained on.

* RAG cannot solve data drift

- RAG is a LLM-specific technique that solves only knowledge drift not numeric / data drift
- in RAG if a LLM is trained on data until 2024 if you want to know new info (2025 news, products etc) you don't retrain model inste² you retrieve fresh facts from an external database and feed them as context
- Data drift is a ML thing for Predictive models ie regression, classification not obs² LLMs and it's used to fix changes in input data distribution from last pages
- Fix 4 RAG is not one of the online methods inste² online adaptive learning = continuously or periodically updating the model with new data inste² retraining from scratch

Feature Selection

- is the process of choosing a subset of the most relevant Features from your original data set, the goal is to reduce dimensionality of data by removing irrelevant, noisy or redundant Features. This can lead to simpler models, improved model performance and a better understanding of underlying data