MlOps Assignment 1

Arushi Makraria
24th October 2024

Accuracy and Privacy Comparison: Non-DP vs DP Models on Dataset v2

- The XGBoost model without DP performed the best, achieving an MSE of 21,067.47, highlighting its accuracy without privacy constraints.
- The first DP model, with an epsilon of 1.17, had a higher MSE of 67,546.30. This epsilon value reflects a lower level of privacy, allowing individual data points to have more influence on predictions, resulting in a trade-off between accuracy and privacy.
- In the second DP model with increased privacy parameters (epsilon = 0.58) reduced the influence of individual data on predictions but unexpectedly showed a slightly lower MSE of 64,368.23. This result likely stems from random initialization rather than actual improved performance, as increased privacy typically leads to greater error.

Observations:

XGBoost offers superior accuracy, while DP models trade accuracy for privacy, although results may vary slightly due to initialization randomness.

LakeFS vs DagsHub

Ease of Installation:

- DagsHub: Quick setup, intuitive documentation.
- LakeFS: This was also similar and easy to set up.

Data Versioning:

- DagsHub: User-friendly, similar to GitHub, making it familiar for users.
- LakeFS: Robust features, but less intuitive commands can be confusing.

Switching Between Versions:

- DagsHub: Smooth and fast version switching for testing.
- LakeFS: Needed more Python knowledge like context managers, leading to potential confusion.

Overall Preference:

I personally preferred **DagsHub** for its GitHub-like experience, making version management and privacy implementation smoother and more familiar.