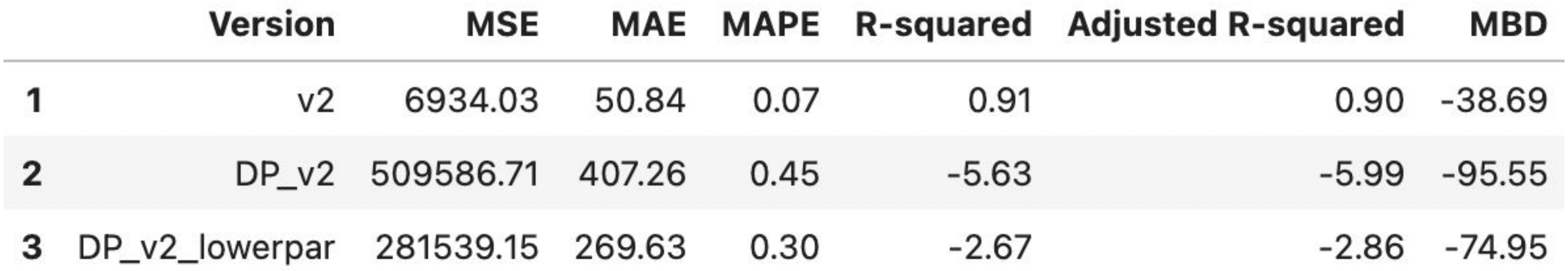
Linda Ji

Assignment 1 - ML Ops

Non-NP vs DP Models: Metric Comparison



	Version	MSE	MAE	MAPE	R-squared	Adjusted R-squared	MBD
1	v2	6934.03	50.84	0.07	0.91	0.90	-38.69
2	DP_v2	509586.71	407.26	0.45	-5.63	-5.99	-95.55
3	DP_v2_lowerpar	281539.15	269.63	0.30	-2.67	-2.86	-74.95

- DP model had a much higher error in metrics like MSE, MAE & MAPE, indicating it has more errors in terms of the model vs. non-DP model
- The R-squared is also very negative, indicating the DP model is a bad fit for the data, even with experimenting to lowering the noise multiplier from 1.1-> 0.5
- As DP introduces noise, it could be the cleaned dataset is too small, and it has a disproportionate effect on the data, reducing the accuracy of the model.

FS Lake vs. DVC: Ease of Use

Evaluation Criteria	LakeFS	DVC (Data Version Control)
Ease of Installation	Moderate: Requires setup on a cloud or local environment. Integrates with object stores (e.g., S3, GCS).	Easy: Installed via pip or package manager, integrates with Git. Lightweight and developer-friendly.
Ease of Data Versioning	High: Provides seamless data versioning on object stores with Git-like branching and commit operations.	Moderate: Works with large datasets through metadata tracking, but requires linking with Git.
Ease of Switching Between Versions for Same Model	Easy: Supports branching and commits, allowing users to easily switch between dataset versions through commands.	Moderate: Switching involves checking out specific tags or branches in Git. Requires careful management of .dvc files.
Effect of DP on Model Accuracy/Metrics	Limited Direct Support: Can store and version datasets with DP applied but may need to manually log DP parameters (like noise multiplier, clipping norm) within LakeFS metadata	Limited Direct Support: Can store and version datasets with DP applied but may need to manually log DP parameters (like noise multiplier, clipping norm) with git

LakeFS:

- Best suited for **large-scale cloud storage environments** with object stores.
- Ideal for **teams managing extensive datasets** with a need for collaborative version control at scale (similar to Git).

DVC:

- Lightweight and developer-friendly, perfect for **data scientists working with Git** workflows.
- Ideal for **individuals or small teams** needing **model reproducibility** with minimal setup.