When Does Machine Learning Fail?

A Broad Stress Test of Common Models

Motivation

- ML models are evaluated under ideal assumptions.
- Real-world data is noisy and shifts over time.
- Accuracy hides brittleness.
- We study when models fail.

Background

- Benchmarks assume static data.
- Deployment failures are often silent.
- Reliability is underexplored in practice.

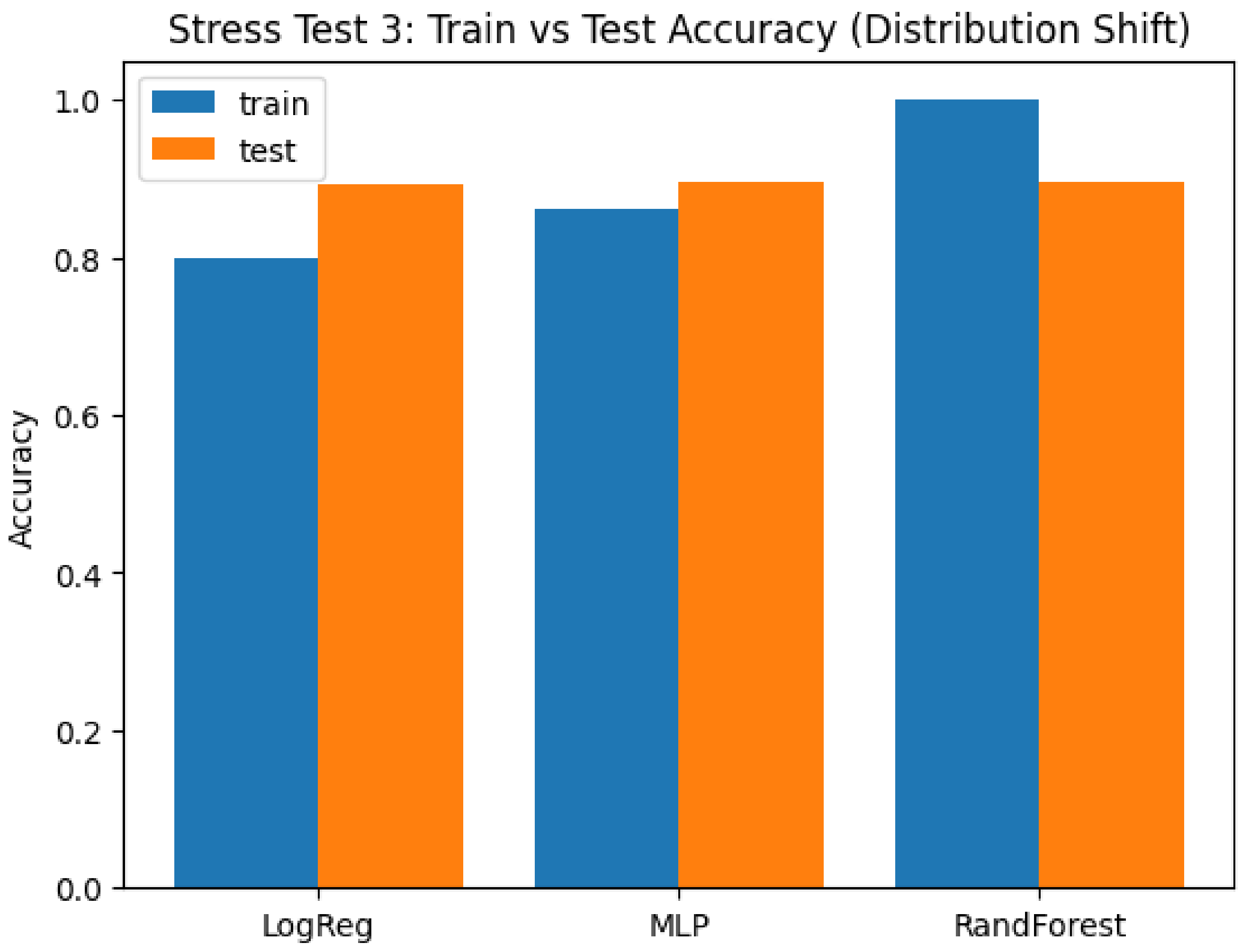
Related Work

- Robust ML
- Distribution shift
- Model calibration

Claim / Target Task

- Models fail in systematic ways.
- High-capacity models fail more sharply.
- Simple models degrade gracefully.

Why Failure Matters: Distribution Shift



Proposed Solution

- Controlled stress tests
- Noise, label corruption, distribution shift
- Compare LogReg, RandomForest, MLP

Implementation

- Dataset: UCI Adult Income
- Notebook-based pipeline
- Reproducible experiments

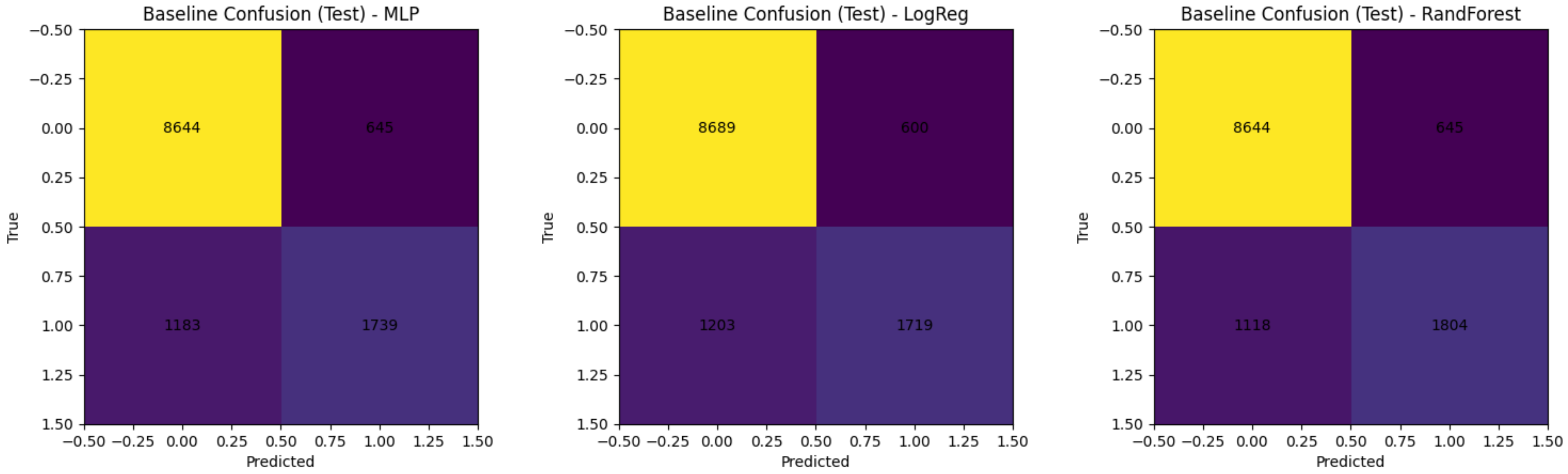
Data Summary

~48K samples

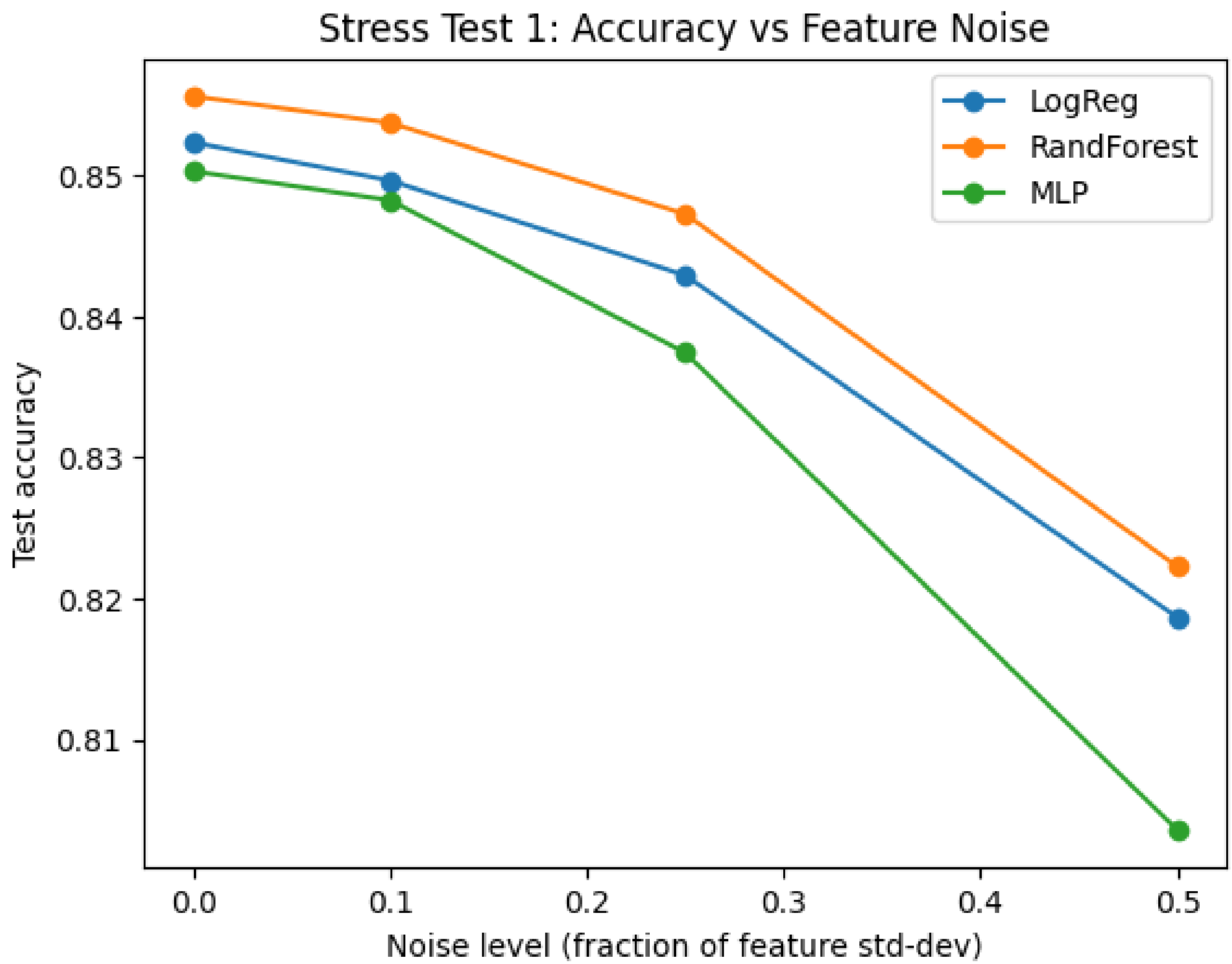
Numeric + categorical features

Binary income classification

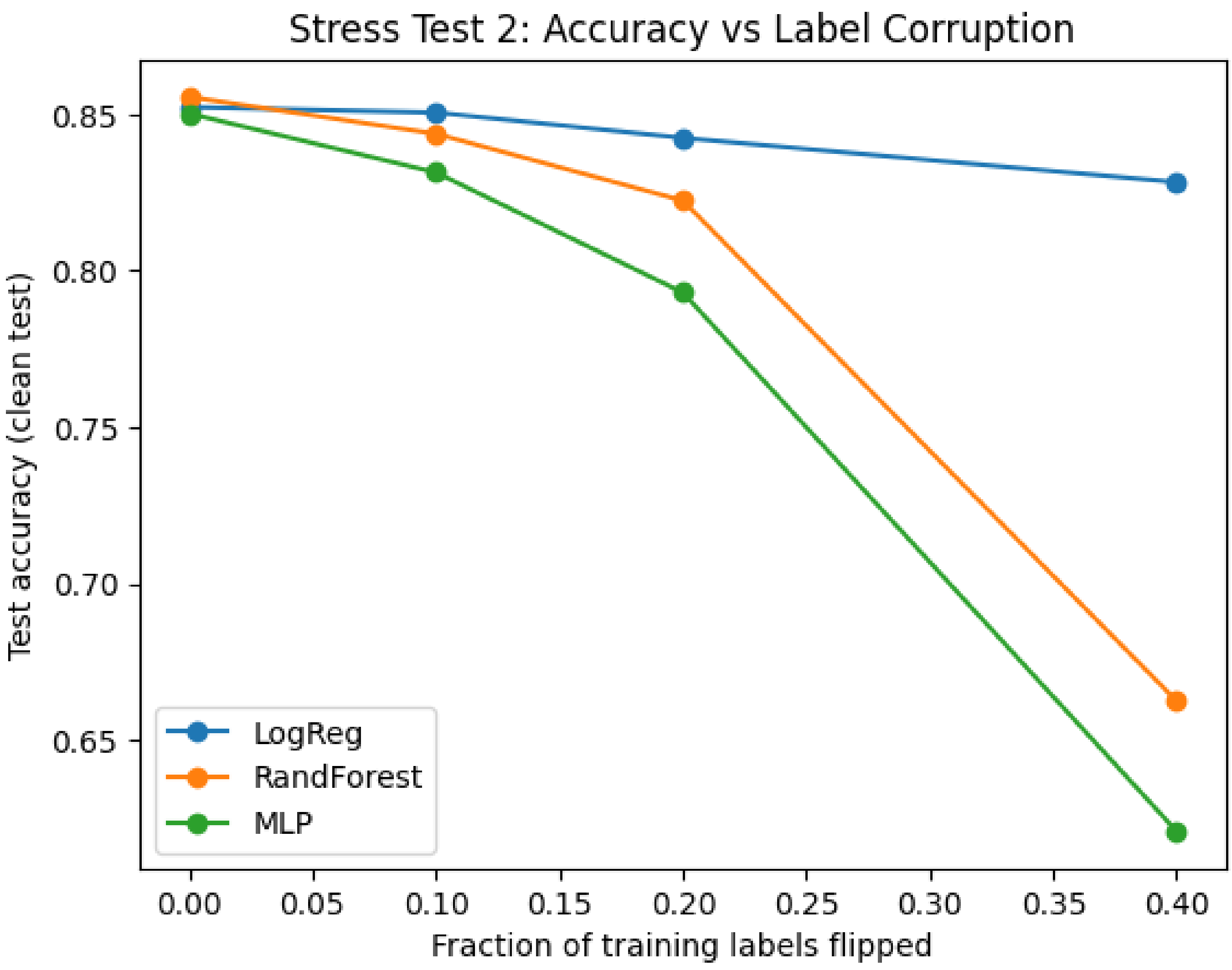
Confusion Matrices (Baseline)



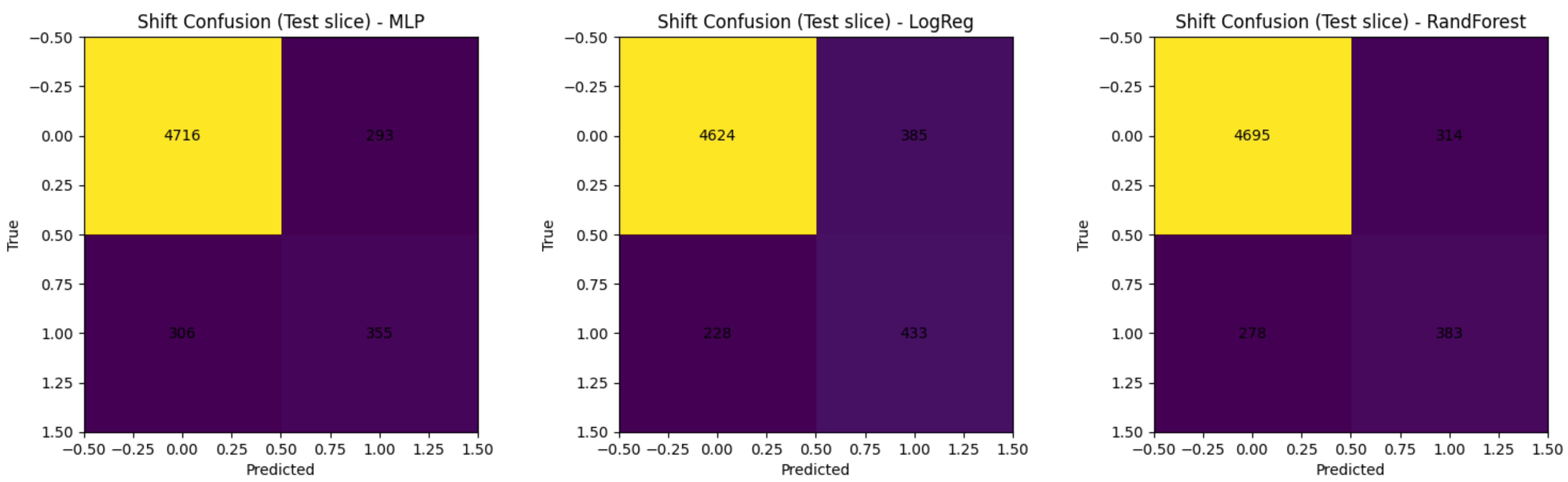
Experimental Results: Feature Noise



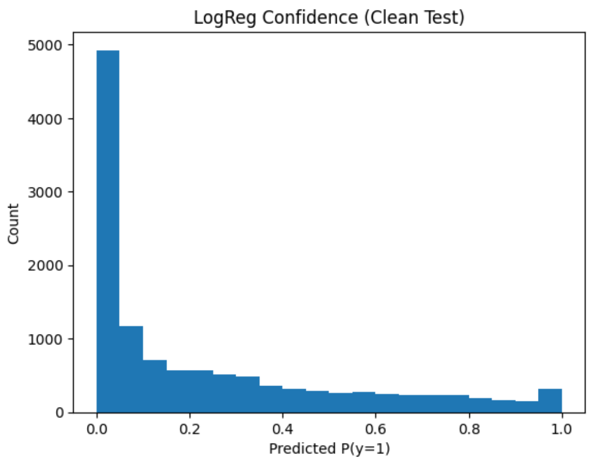
Experimental Results: Label Corruption

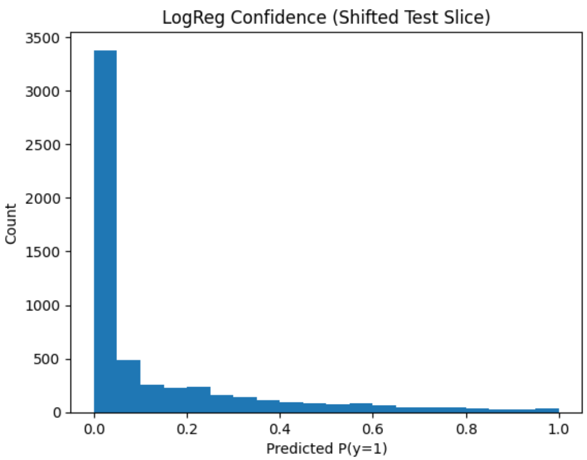


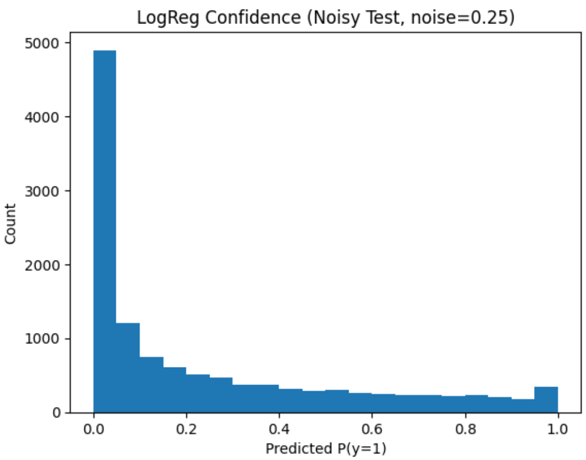
Confusion Matrices (Test)



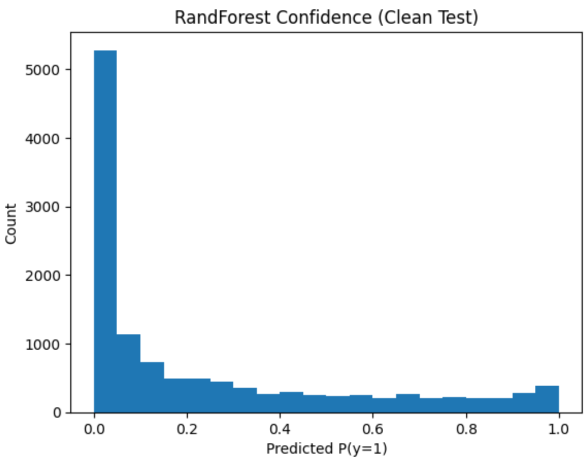
Experimental Analysis: Confidence

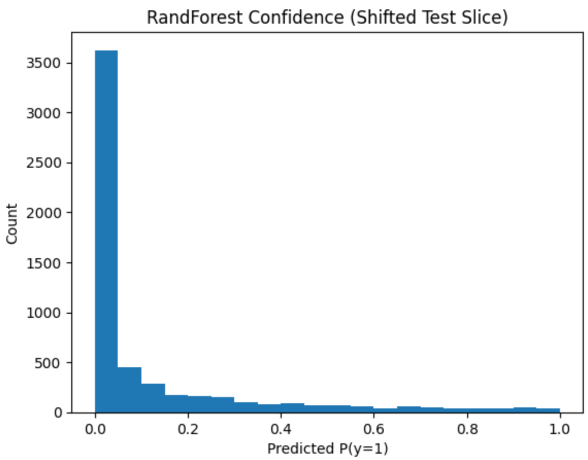


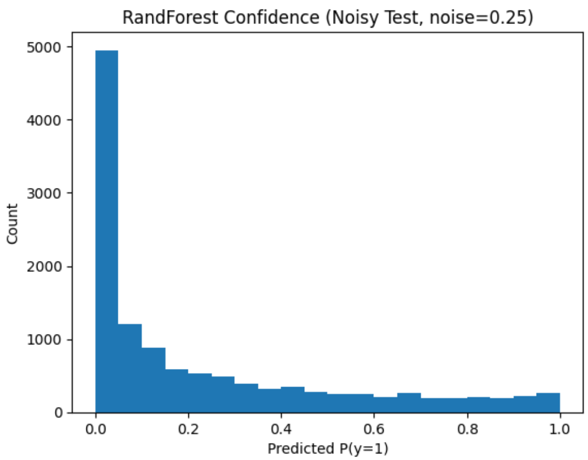




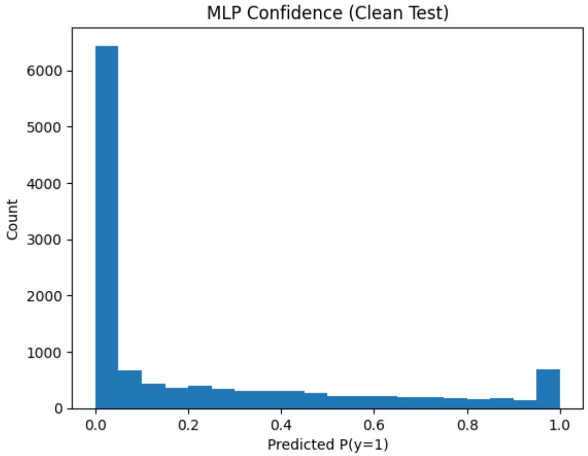
Experimental Analysis: Confidence

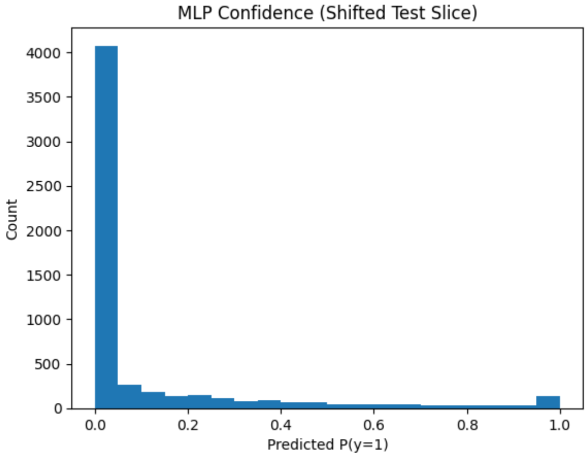


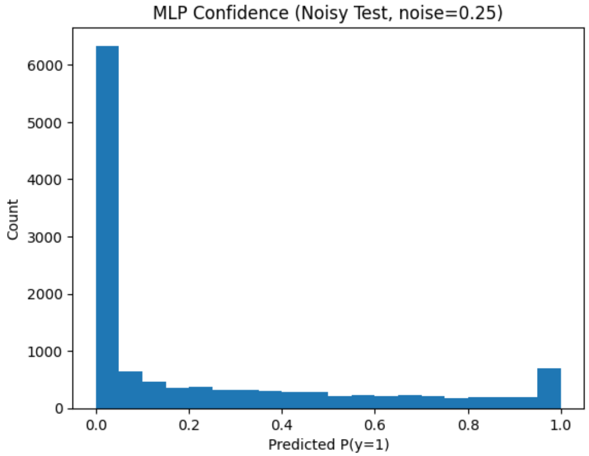




Experimental Analysis: Confidence







Experimental Analysis

- Random Forest most robust to noise
- MLP most sensitive to label corruption
- Large gaps under distribution shift
- Confidence remains high despite errors

Conclusion and Future Work

- Stress testing reveals hidden brittleness
- Reliability requires more than accuracy
- Future: calibration and fairness tests

References

- UCI Adult Income Dataset
- Scikit-learn
- Course materials