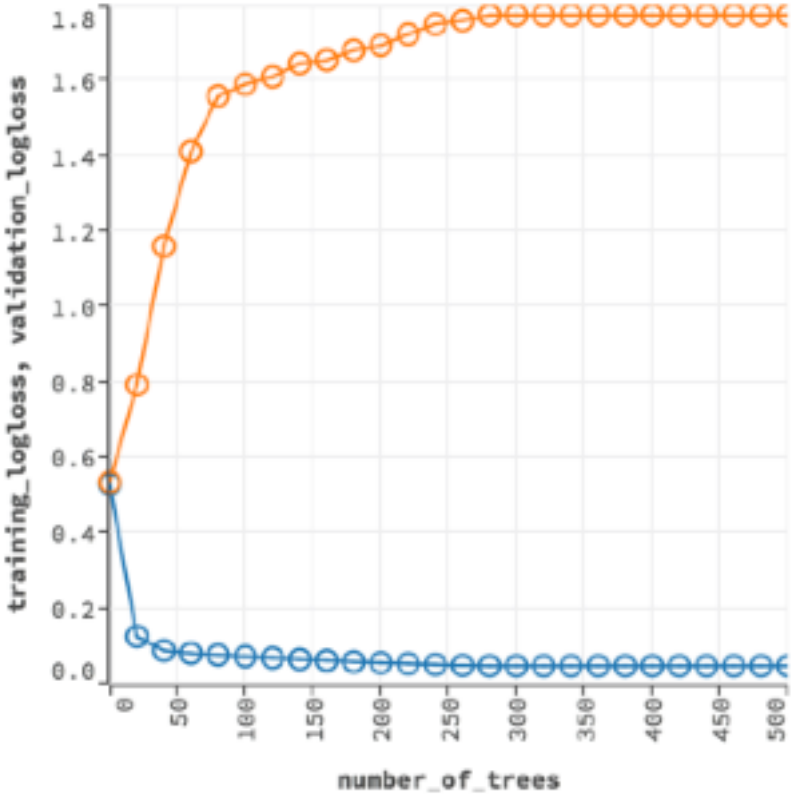
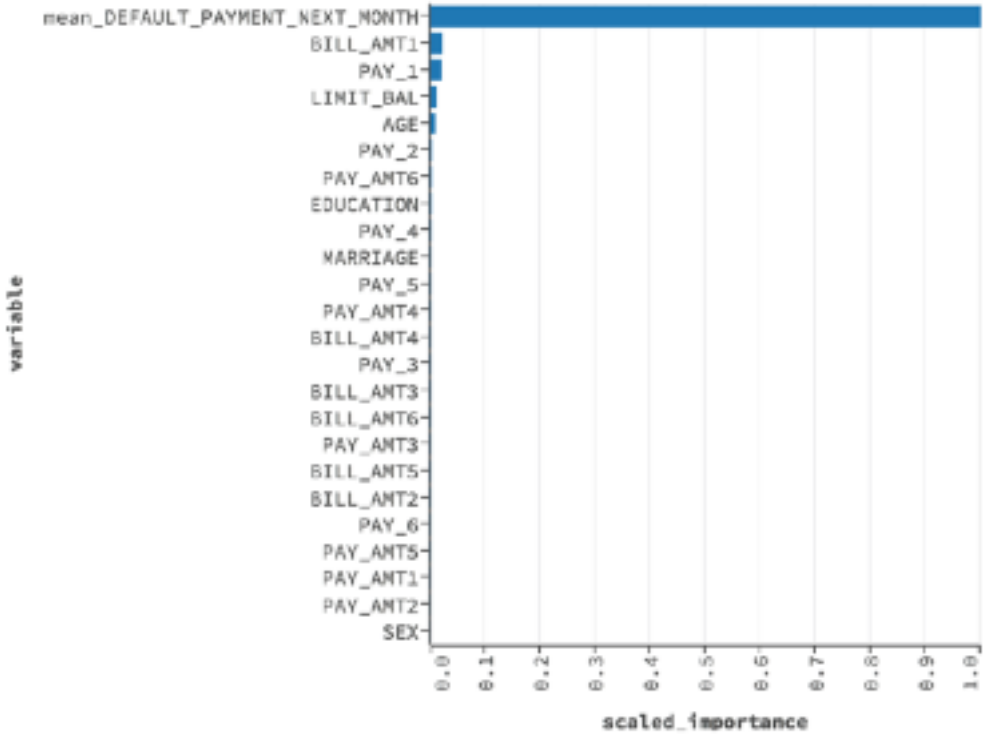

H2O.ai

Data Leakage

▼ SCORING HISTORY - LOGLOSS



VARIABLE IMPORTANCES



Data Leakage Feature is the only important feature

Scoring History: Training vs Testing

Target Class Imbalance

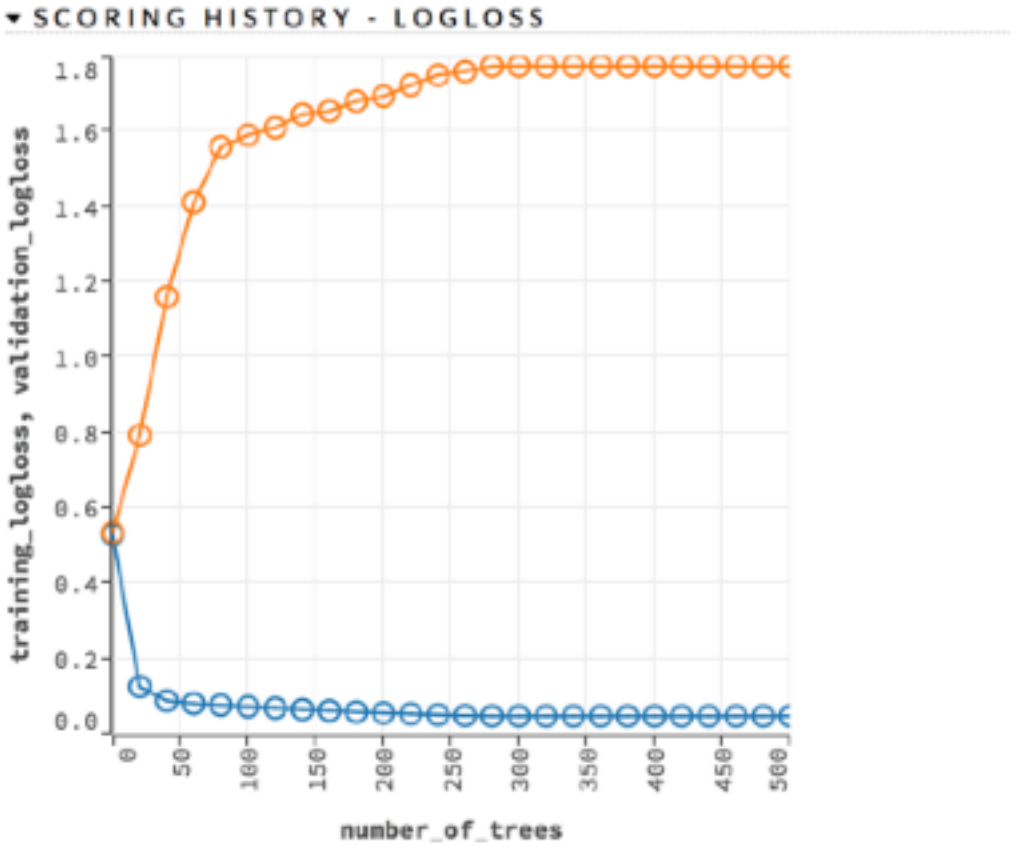
Imbalanced Response Variable

- A dataset is said to be **imbalanced** when categorical responses occur at widely varying rates.
- Standard optimizations by machine learning algorithms may favor majority classes.
- Rule of thumb for binary response: If the minority class makes < 10% of the data, this can cause issues.

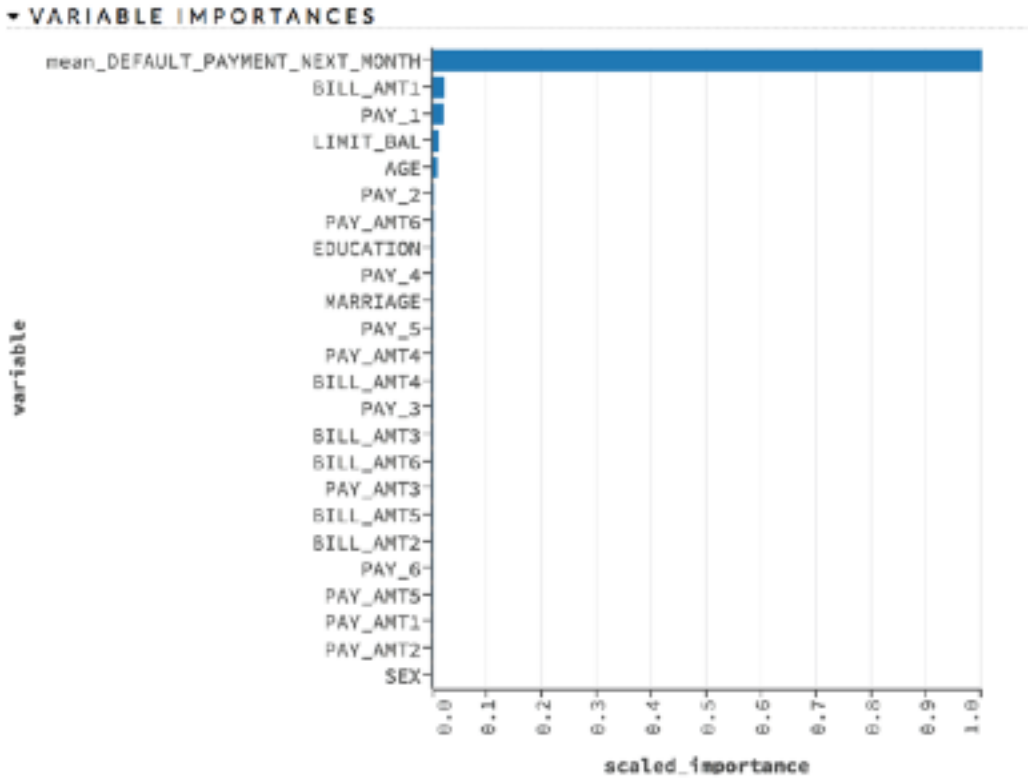
Common Examples Across Industries

- Advertising – Probability that someone clicks on ad is very low… very very low.
- Healthcare & Medicine – Certain diseases or adverse medical conditions are rare.
- Fraud Detection – Insurance or credit fraud is rare.

Data Leakage



Scoring History: Training vs Testing



Data Leakage Feature is the only important feature