
H2O.ai

Data Leakage

- Leakage is allowing your model to use information that will not be available in a production setting.
- Example: using the Dow Jones daily gain/loss as part of a model that predicts stock performance.

- Understand the nature of your problem and data.
- Scrutinize model feedback, such as relative influence or linear coefficient.

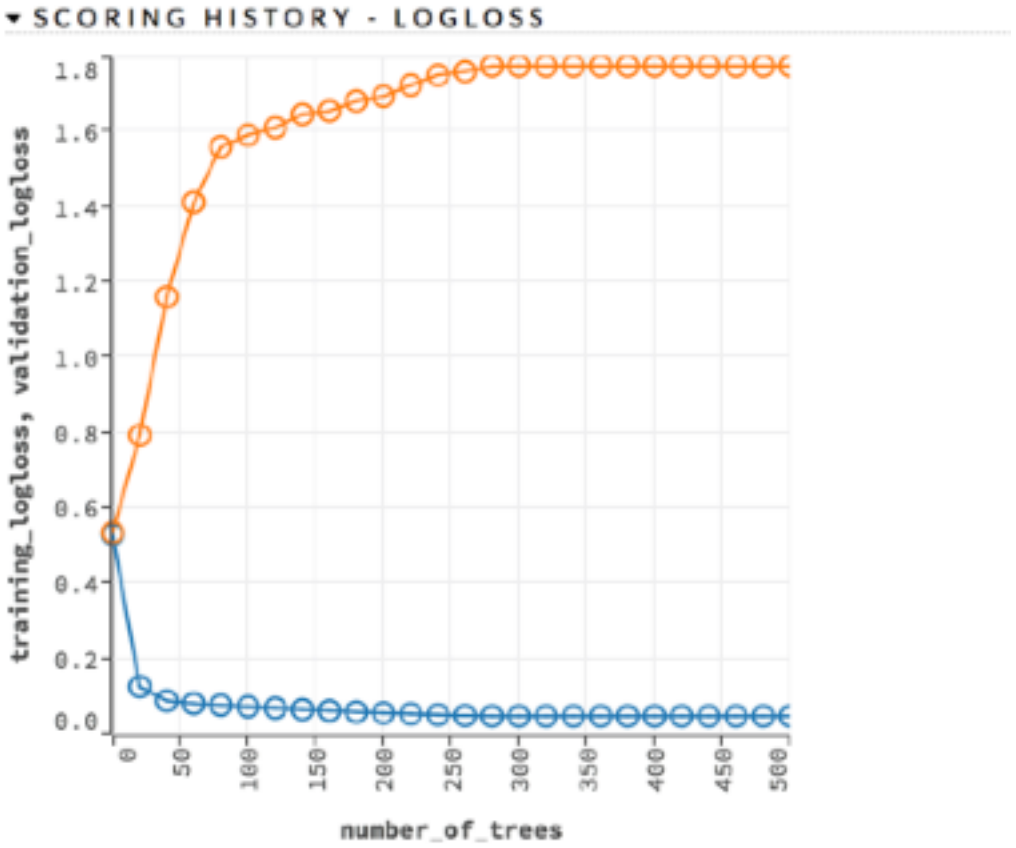
What Is It

What to Do

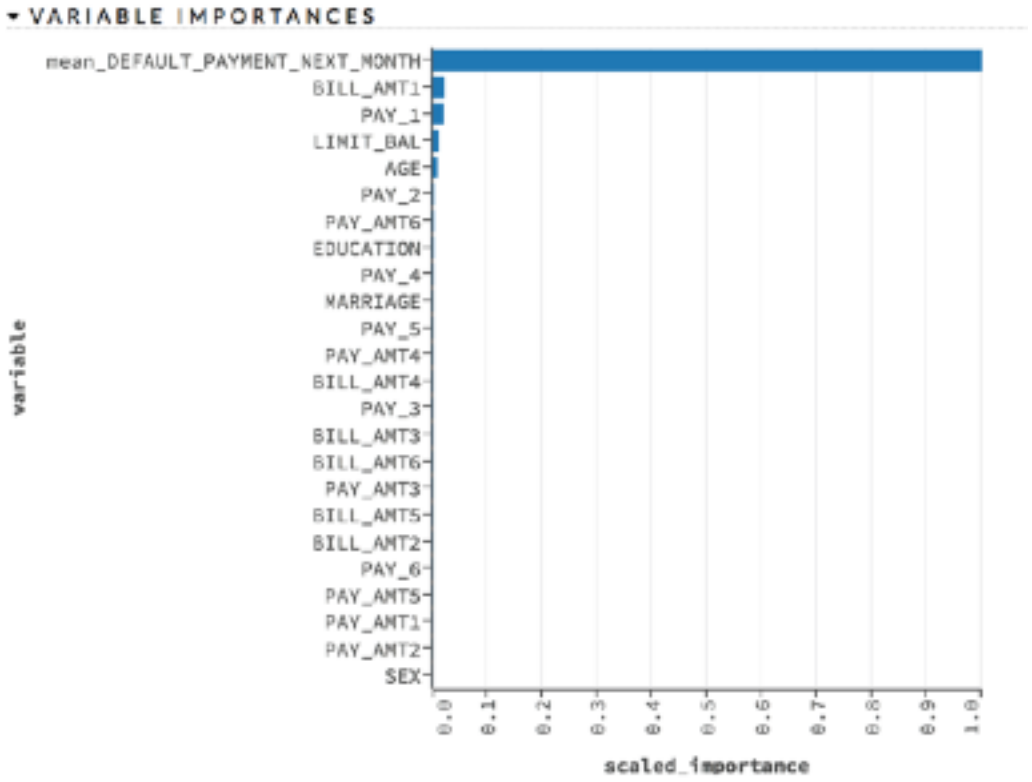
What Happens

- Model is overfit.
- Predictions will be inconsistent with those scored during model training (even with a validation set).
- Insights derived from the model will be incorrect.

Data Leakage



Scoring History: Training vs Testing



Data Leakage Feature is the only important feature

H2O.ai

Data Leakage

What Is It

- Leakage is allowing your model to use information that will not be available in a production setting.
- Example: using the Dow Jones daily gain/loss as part of a model that predicts stock performance.

What Happens

- Model is overfit.
- Predictions will be inconsistent with those scored during model training (even with a validation set).
- Insights derived from the model will be incorrect.

What to Do

- Understand the nature of your problem and data.
- Scrutinize model feedback, such as relative influence or linear coefficient.

H2O.ai