H2O Training

1. Introduction to ML
2. Introduction to H2O
3. H2O Data Munging
4. Machine Learning Concepts
5. H2O Algorithms
6. Productionalizing Models

1. Train vs Test

Training Set vs. Test Set

- Partition the original data (randomly or stratified) into a **training** set and a **test** set. (e.g. 70/30)

Training Error vs. Test Error

- It can be useful to evaluate the training error, but you should not look at training error alone.
- Training error is not an estimate of **generalization error** (on a test set or cross-validated), which is what you should care more about.
- Training error vs test error over time is an useful thing to calculate. It can tell you when you start to overfit your model, so it is a useful metric in supervised machine learning.

Train vs Test

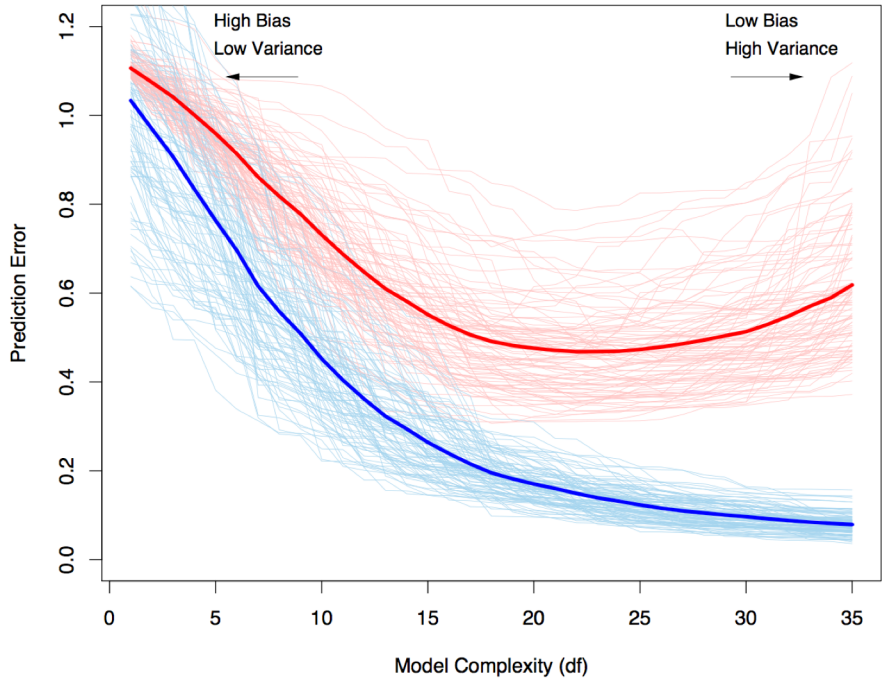


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

Source: Elements of Statistical Learning

2. Train vs Valid vs Test

Training Set vs. Validation Set vs. Test Set

- If you have “enough” data and plan to do some model tuning, you should really partition your data into three parts — Training, Validation and Test sets.
- There is **no general rule** for how you should partition the data and it will depend on how strong the signal in your data is, but an example could be: 50% Train, 25% Validation and 25% Test

Validation is for Model Tuning

- The validation set is used **strictly for model tuning** (via validation of models with different parameters) and the test set is used to make a final estimate of the generalization error.

3. Model Performance

Test Error

- Partition the original data (randomly) into a training set and a test set. (e.g. 70/30)
- Train a model using the training set and evaluate performance (a single time) on the test set.

K-fold Cross-validation

- Train & test K models as shown.
- Average the model performance over the K test sets.
- Report cross-validated metrics.

Performance Metrics

- Regression: R^2, MSE, RMSE
- Classification: Accuracy, F1, H-measure, Log-loss
- Ranking (Binary Outcome): AUC, Partial AUC

4. 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.

4. Class Imbalance Remedies

Artificial Balance

- You can **balance** the training set using sampling.

Potential Pitfalls

- Don't balance the test set! The test set should represent the true data distribution.
- The same goes for a hold-out validation set and cross-validation sets.
- Cross-validation will probably require custom coding.

Solutions

- H2O has a "balance_classes" argument that can be used to do this properly & automatically.
- You can manually **upsample (or downsample)** your minority (or majority) class(es) set either by duplicating (or sub-sampling) rows, or by using row weights.

5. Categorical Data

Real Data

- Most real world datasets contain categorical data.

Too Many Categories

- Problems can arise if you have **too many categories**:
  - Computational complexity during estimation
  - Infrequent categories can lead to overfitting

Solutions

- Use knowledge about hierarchical data to collapse categories.
- Use Cross-Validated Mean Target Encoding.
- Use Cross-Validated Weight of Evidence Encoding when modeling binary outcome.

Target Mean Encoding

What?

Replace categorical variables with the mean of the response

Why?

Categorical variables increase the number of features (dummy encoding) and can cause us to overfit

Target Mean Encoding

Pay 1	Default Payment
Up To Date	0
Up To Date	0
Up To Date	0
Missed 1 Mo	1
Missed 1 Mo	0
Missed 1 Mo	0
Missed 5 Mo	1

Target Mean Encoding

Pay 1	Default Payment	Mean Target Encoding
Up To Date	0	0
Up To Date	0	0
Up To Date	0	0
Missed 1 Mo	1	0.33
Missed 1 Mo	0	0.33
Missed 1 Mo	0	0.33
Missed 5 Mo	1	1

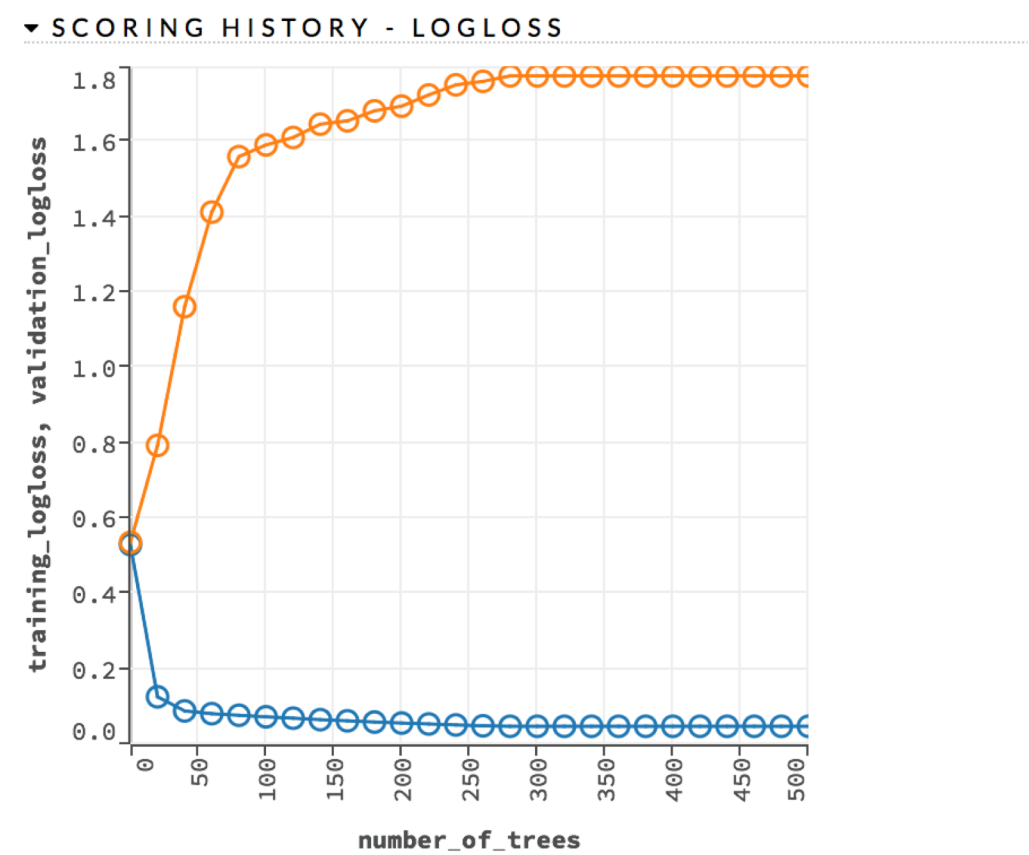
Target Mean Encoding

- Mean Target Encoding is based on the response column of the rows
- The lower the number of rows in the group, the more it reveals the response column value

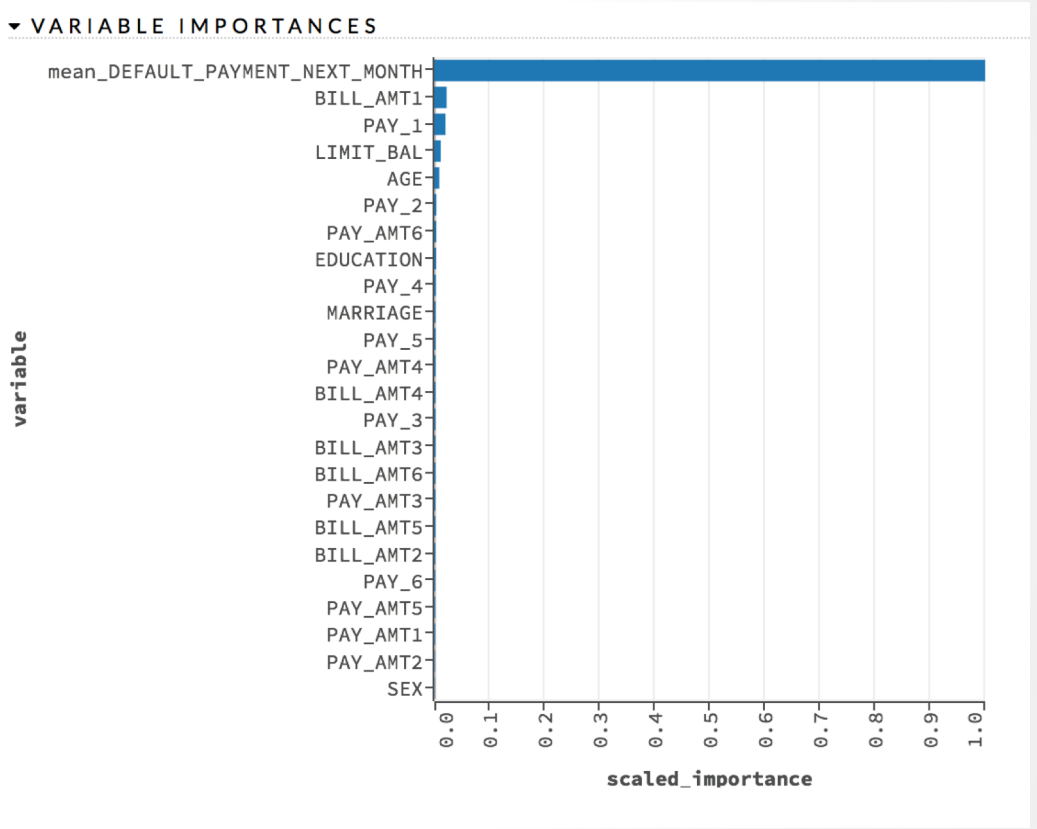
Pay 1	Default Payment	Mean Target Encoding
Missed 5 Mo	1	1

Worst Case Scenario: Response Column = Mean Target Encoding

Effects of Data Leakage



Scoring History: Training vs Testing



Data Leakage Feature is the only important feature

Cross Validation Target Encoding

Fold	Pay 1	Default Payment
1	Up To Date	0
2	Up To Date	0
3	Up To Date	0
2	Missed 1 Mo	1
1	Missed 1 Mo	0
3	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment
2	Up To Date	0
3	Up To Date	0
2	Missed 1 Mo	1
3	Missed 1 Mo	0

Fold	Pay 1	Default Payment
1	Up To Date	0
1	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment
1	Up To Date	0
1	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	
1	Missed 5 Mo	1	

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	NA

Weight Of Evidence Encoding

What?

In binary classification, replace categorical variables with

$$WOE_{ja} = \ln \frac{P(X_j = a|Y = 1)}{P(X_j = a|Y = 0)}$$

Why?

Leverage rich history in information theory and Bayesian statistics to manage overfitting of high cardinality variables

Weight Of Evidence Encoding

Pay 1	Default Payment	% 0s	% 1s	(% 1s) / (% 0s)	WOE
Up To Date	0	60 %	0 %	0	- Inf
Up To Date	0	60 %	0 %	0	- Inf
Up To Date	0	60 %	0 %	0	- Inf
Missed 1 Mo	1	40 %	50 %	1.25	0.223
Missed 1 Mo	0	40 %	50 %	1.25	0.223
Missed 1 Mo	0	40 %	50 %	1.25	0.223
Missed 5 Mo	1	0 %	50 %	Inf	Inf

6. Missing Data

Types of Missing Data

- Unavailable: Valid for the observation, but not available in the data set.
- Removed: Observation quality threshold may have not been reached, and data removed
- Not applicable: measurement does not apply to the particular observation (e.g. number of tires on a boat observation)

What to Do

- Ignore entire observation.
- Create an binary variable for each predictor to indicate whether the data was missing or not.
- Segment model based on data availability.
- Estimate missing values (Generalized Low Rank Models)
- Use alternative algorithm: decision trees accept missing values; linear models typically do not.

7. Outliers/Extreme Values

Types of Outliers

- Outliers can exist in response or predictors
- Valid outliers: rare, extreme events
- Invalid outliers: erroneous measurements

What Can Happen

- Outlier values can have a disproportionate effect.
- MSE will focus on handling outlier observations more to reduce squared error.
- Boosting will spend considerable modeling effort fitting these observations.

What to Do

- Remove observations.
- Apply a transformation to reduce impact: e.g. log skewed data, create categorical bins, impose cap a low/high end. Choose a more robust loss function: e.g. MAE vs MSE.
- Ask questions: Understand whether the values are valid or invalid, to make the most appropriate choice.

8. 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.