Working with Small Datasets

Or, how not to overfit with only a few examples.
1. Issues from Limited Data
2. How Small Datasets Impact the Model Fit
3. Techniques for Working with Smaller Datasets
The Trouble With Limited Training Data
When You Don’t Have “Big Data”

From Terabytes .. to a couple hundred rows
Use Cases with Limited Data are Common

Consider a use case with clinical trial data, where only only a few hundred of patients were involved in a clinical trial.

You have many predictors from the medical records, or even genomic data, but which factors reliably predict your target outcome?

“Gather more data” may not be an option.
Sudden Changes Can Limit Relevant Data

You may have a long history of training data, but most of it could become irrelevant if there is a sudden change, like a crash in a commodity price.
How Models Fit
Overfitting

Machine learning models seek a balance between finding meaningful signal and finding spurious patterns in the training data.
Small Datasets and Model Selection

Classify These Dots - Red or Green?
Small Datasets and Outliers

For a very small dataset, there is risk that a single outlier could land in a validation or holdout fold, making the model appear to fall apart.

Take extra care to identify outliers prior to modeling, and determine whether the unusual example should be kept in the training data.
How many rows are enough?

Repeat the process of fitting your model to a subset of the data, and validating on a holdout.

Check Stability in out-of-sample error:

- Are errors in the Validation, Cross Validation & Holdout Consistent?
- How about the error in each individual fold?
How many rows are enough?

Predicting Loan Defaults: 10,000 rows, 38 Features

Curve shows change in out-of-sample error (y-axis) as amount of training data increases (x-axis).

Check Learning Curves:

- Are models consistently improving? Or is there “noise” in the out-of-sample error?
- Are curves leveling out as more training data is added?
How many rows are enough?

Predicting Loan Defaults: 1,000 rows, 38 Features

Curve shows change in out-of-sample error (y-axis) as amount of training data increases (x-axis).

Check Learning Curves:

- With fewer samples, some models are still improving substantially.
- Sudden increases in out of sample error indicate some models are still overfitting, and may not be reliable.
Techniques for Modeling with Small Data
Examining the Leaderboard

**Intuition for why some models do best on smaller datasets**
Models that work better on small datasets

Regularized Linear Regression (or Elastic Net)

Feature selection is key with smaller datasets, particularly as the ratio of columns to rows increases

- Lasso if only a subset of the predictors carry the signal
- Ridge when signal is carried by a larger fraction of the predictors
- Elastic Net balances the advantages of Lasso & Ridge

Support Vector Machines perform well on smaller datasets for many of the same reasons.
Models that work better on small datasets

- **Eureqa models fit an analytic expression to the training data.**

- **The expressions returned are limited by their “complexity” - penalizing complicated (and overfit) models.**

- **By trying millions of combinations of features in different formulas with penalized complexity, the algorithm is very effective at generating a robust model even with only a small number of training examples.**
Dynamic changes in diffusion measures improve sensitivity in identifying patients with mild traumatic brain injury


Problem: The goal of this study was to investigate patterns of axonal injury in the first week after mild traumatic brain injury (mTBI).

The available sample size only included 20 patients and 16 controls, for a total of 36 data points.

Eureqa in Action

“...For this study, we sought to evolve functions capable of predicting the sum of post-concussive symptoms at time 2 (S2)...”

<table>
<thead>
<tr>
<th>Eq. #</th>
<th>Evolved expression</th>
<th>Mean abs error</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$S_2 = 0.353S_1 + \Delta FA$</td>
<td>4.103</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>$S_2 = S_1*(\Delta FA)$</td>
<td>3.982</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>$S_2 = 93.73S_1 + \Delta genu$</td>
<td>2.936</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>$S_2 = 0.346S_1 + \Delta body + \Delta splenium$</td>
<td>2.513</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>$S_2 = 39.25\Delta FA + \Delta genu + \Delta splenium$</td>
<td>2.312</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>$S_2 = 52.05\Delta FA + 0.757\Delta body + 348.91\Delta splenium$</td>
<td>2.146</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>$S_2 = 0.282S_1 + 36.90\Delta body + 348.91\Delta splenium + \Delta body$</td>
<td>1.776</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>$S_2 = 0.356S_1 + 3.051\Delta FA + 0.868\Delta body$</td>
<td>1.491</td>
<td>21</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0178360.t007
Technique 1: Increase CV Folds

Impact of $K$ on the Bias and Variance of the MSE across $i$ datasets.

Left Hand Side: Kfolds for 200 data points, Right Hand Side: Kfolds for 40 data points

Technique 2: Use Nested CV

Fig 3. Gaussian noise classification accuracy distributions with different validation approaches. K-Fold, Nested, Train/Test Split and two types of partially nested validation methods used. Thick lines show mean validation accuracy and dash-dot lines show 95% confidence intervals for 50 runs. A: SVM-RFE feature selection and SVM classification. B: t-test feature selection and logistic regression classification.

https://doi.org/10.1371/journal.pone.0224365.g003
Rerun Autopilot with Different Random Seeds

Reshuffle the cross validation folds, and check for variation in the top model & leaderboard scores.

Project V1

Project V2

Project V3
Wide but short datasets
When there are far more columns than rows

Your model risks breaking down even with regularization other feature selection techniques:

- Apply domain knowledge to remove irrelevant columns and identify known, natural groupings
- Reshuffle the CV folds a large number of times to identify features that only correlate to the target due to chance
- Compare Feature Impact across projects

Common Examples
- Genomic data
- IOT Sensor Data
- Natural Language Processing (NLP)
What about data augmentation?

Avoid duplicating rows.

- If a duplicated row appears in multiple cross validation folds, the error metric could be meaningless.
- This can create an overconfident model, even if you are careful to keep duplicate rows in separate cross validation folds.

Synthetic data is risky (using techniques like SMOTE), because it is unclear if the new data added is representative of the dataset you are trying to model.
Time Series: Cold Starts
What is a Cold Start?

Different products have been around for different lengths of time.

- As long as we have recent history, we can make a forecast.
- But, what about a product that was just launched?

Each line represents the history of a product, or a series.
What is a Cold Start?

A sudden change can cause series with a long history to still become cold starts.

Past training data is only going to confuse the model.

Each line represents the history of a product, or a series.
What is a Time Series Forecast?

- **Past**
  - Feature Derivation Window (FDW): -28 days to -7 days

  - Blind History Gap (BHG)

- **Forecast Point**
  - Now

- **Future**
  - Forecast Window (FW): +1 day to +7 days

  - Can’t Operationalize Gap (COG)
Define True Cold Starts

Determine What is a Cold Start, Warm Start, or Ongoing Series

Warm Starts:
- Temporarily run forecasts with a shorter feature derivation window.

Cold Starts:
- Temporarily work with a shorter time step
- Build a temporary ordinary regression model with out-of-time validation
- Train time series once enough history is re-established
Questions & Answers
Join your peers today at community.datarobot.com
Engage, learn, and accelerate your AI/ML journey
Connect with peers to find solutions to AI challenges
Explore helpful content to take your AI to the next level
Build your brand as an AI expert & thought leader
Covid: https://www.datarobot.com/lp/covid-19-research/
Questions: aisuccess-webinars@datarobot.com
1. Anomaly detection to identify outliers

Data Quality Checks

Models fit to smaller datasets are more sensitive to outliers

- If outlier rows aren’t obvious, consider running unsupervised mode to identify unusual rows to remove from the training dataset
Check Variation in Models

Notes

- You are comparing feature selection across multiple types of models
- Instead of comparing different models, consider re-running a top model on different

Aggregated Feature Impact Across Top Models
Technique 3: Blenders to Smooth Out Signal