Before the Start Button: Best practices for setting up a DataRobot Project

Alexandre Gerbeaux
Matthew Cohen
Jack Jablonski
Begin a project by dragging data here or simply import files.
After Data Upload

DATAROBOT FEATURE LISTS

Informative Features

Raw Features

Explore My Dataset

[Feature Excluded]

1. Looking like an ID
   - [Reference ID] id
     - Var Type: Numeric
     - Unique: 50,000
     - Missing: 0
     - Mean: 1,918,444
     - Std Dev: 638,983
     - Median: 1,587,759
     - Min: 58,524
     - Max: 3,364,574

2. Does not contain enough values
   - [Reference ID] member_id
     - Var Type: Numeric
     - Unique: 50,000
     - Missing: 0
     - Mean: 2,283,786
     - Std Dev: 802,482
     - Median: 1,857,296
     - Min: 149,512
     - Max: 4,076,727

3. Is empty
   - [Empty] mths_sl_r_derog
     - Var Type: Numeric
     - Unique: 0
     - Missing: 10,000

4. Is duplicated
   - [Few values] loan_divided_by_income
     - Var Type: Numeric
     - Unique: 13,986
     - Missing: 0
     - Mean: 0.22
     - Std Dev: 0.11
     - Median: 0.21
     - Min: 2.08e-3
     - Max: 0.50

   - [Few values] collections_12_mths_ex_med
     - Var Type: Numeric
     - Unique: 3
     - Missing: 0
     - Mean: 0.00
     - Std Dev: 0.04
     - Median: 0
     - Min: 0
     - Max: 2

   - [Few values] mths_since_last_major_derog
     - Var Type: Numeric
     - Unique: 100
     - Missing: 42,880
     - Mean: 42.31
     - Std Dev: 20.91
     - Median: 40
     - Min: 0
     - Max: 152

   - [Few values] policy_code
     - Var Type: Categorical
     - Unique: 1
     - Missing: 0
     - Mean: 1
     - Std Dev: 1
     - Median: 1
     - Min: 1
     - Max: 1

   - [Few values] application_type
     - Var Type: Categorical
     - Unique: 1
     - Missing: 0

   - [Few values] acc_now_delinq
     - Var Type: Numeric
     - Unique: 4
     - Missing: 0
     - Mean: 8.20e-4
     - Std Dev: 0.03
     - Median: 0
     - Min: 0
     - Max: 4

   - tot_coll_amt
     - Var Type: Numeric
     - Unique: 980
     - Missing: 14,618
     - Mean: 52
     - Std Dev: 660
     - Median: 0
     - Min: 0
     - Max: 55,009
After Data Upload

Scan your data and explore it in DataRobot.

### Feature List

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Type</th>
<th>Value</th>
<th>Std Dev</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>loan_amnt</td>
<td>Percentage</td>
<td>7,468</td>
<td>9,600</td>
<td>500</td>
<td>35,000</td>
<td></td>
</tr>
<tr>
<td>funded_amnt</td>
<td>Numeric</td>
<td>7,192</td>
<td>9,250</td>
<td>500</td>
<td>35,000</td>
<td></td>
</tr>
<tr>
<td>term</td>
<td>Categorical</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>int_rate</td>
<td>Percentage</td>
<td>359</td>
<td>0</td>
<td>12.12%</td>
<td>3.73%</td>
<td>11.89%</td>
</tr>
<tr>
<td>installment</td>
<td>Numeric</td>
<td>6,381</td>
<td>0</td>
<td>321</td>
<td>211</td>
<td>275</td>
</tr>
</tbody>
</table>

Create a new feature list with the selected features.
After Data Upload

- Users can overwrite / add another feature type.

- This will impact the preprocessing steps applied on the feature.
What would you like to predict?

is_bad  CLASSIFICATION

Number of rows

0 1

is_bad

Modeling Mode: Autopilot

Feature list: Informative Features

Optimization Metric: LogLoss

Start

im gonna press the button

blue
What would you like to predict?

is_bad CLASSIFICATION

Number of rows:

Modeling Mode: Autopilot

Feature list: Informative Features
Optimization Metric: LogLoss

Time-Aware Modeling
DataRobot has detected 1 time feature in your data. Would you like to use time series modeling or out-of-time validation?

Set up time-aware modeling

Show Advanced Options

DON'T FORGET
Optimization metric: DataRobot’s recommendation for Regression

What would you like to predict?

Return

annual_inc

ClaimCount

Optimization Metric: RMSE

Optimization Metric: Gamma Deviance

Optimization Metric: Poisson Deviance
Optimization metric: DataRobot’s recommendation for Regression

What would you like to predict?

- **ClaimsCost**
  - Optimization Metric: **Gamma Deviance**

- **ClaimCount**
  - Optimization Metric: **Poisson Deviance**

- **IncurredClaims**
  - Optimization Metric: **Tweedie Deviance**
Optimization metric: DataRobot’s recommendation for Classification

What would you like to predict?

is_bad

Optimization Metric: LogLoss

grade

Optimization Metric: LogLoss
Optimization metric: where to change?

What would you like to predict?

Number of rows: 0, 1

is_bad

Big Start Button

Advanced Options

Partitioning  Smart Downsampling  Time Series  Feature Constraints  Additional

Optimization Metric

The optimization metric determines how your models are scored.

- **LogLoss (Accuracy)**  (RECOMMENDED)
- Logarithmic Loss: Measures the inaccuracy of predicted probabilities
Optimization metric: for Classification

LogLoss (Accuracy)  **RECOMMENDED**
Logarithmic Loss: Measures the inaccuracy of predicted probabilities

**AUC  SUGGESTED**
Area Under the (ROC) Curve: Measures the ability to distinguish the ones from the zeros

**Gini Norm  SUGGESTED**
Normalized Gini Coefficient: Measures the ability to rank

**RMSE  SUGGESTED**
Root Mean Squared Error: Measures the inaccuracy of predicted mean values

**Rate@Top10%**
Response rate in the top 10% highest predictions

**Rate@Top5%**
Response rate in the top 5% highest predictions
Optimization metric: for Regression

**RMSE (Accuracy)**  [RECOMMENDED]
Root Mean Squared Error: Measures the inaccuracy of predicted mean values

**MAE  SUGGESTED**
Mean Absolute Error: Measures the inaccuracy of predicted median values

**Gini Norm  SUGGESTED**
Normalized Gini Coefficient: Measures the ability to rank

**R Squared  SUGGESTED**
Measures the proportion of total variation of outcomes explained by the model

**MAPE**
Mean Absolute Percentage Error: Measures the percent inaccuracy of the mean values

**RMSLE**
Root Mean Log Squared Error: Measures the inaccuracy of predicted mean values when the target is skewed and lognormally distributed
Partitioning strategy: DataRobot

What would you like to predict?

Number of rows

is_bad

Big Start Button

Advanced Options

Select partitioning method:
- Random
- Partition Feature
- Group
- Date/Time
- Stratified

Partitioning | Smart Downsampling | Time Series | Feature Constraints | Additional
Partitioning strategy: Random Sampling

The most popular method for assigning observations/rows to a partition.

We create a holdout that is completely hidden from the models during the training process. Models are evaluated on the Validation partition.

The validation partition is only used for ranking models. For example, hyperparameter tuning is not done on the validation partition.

In DataRobot, look for:

Default for above 800mb
Partitioning strategy: Cross Validation

In DataRobot, look for:

- CV Folds
- Holdout

Default for below 800mb
Partitioning strategy: Stratified Cross Validation

What would you like to predict?

In DataRobot, look for:

- Stratified

Number of rows

CV Folds | Holdout
---|---
0% | 100%
50% |
Partitioning strategy: Group Partitioning

What would you like to predict?

How likely my patient is going to come back for an unexpected visit within the next 30 days?

1 year of visits

<table>
<thead>
<tr>
<th>Visit</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Partitioning strategy: Group Partitioning

What would you like to predict?

How likely my patient is going to come back for an unexpected visit within the next 30 days?

1 year of visits

<table>
<thead>
<tr>
<th>Visit</th>
<th>Patient</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Partitioning strategy: Group Partitioning

What would you like to predict?

How likely my patient is going to come back for an unexpected visit within the next 30 days?

<table>
<thead>
<tr>
<th>Visit</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Training**: If the same patient falls into the training and validation set, the algorithm might just memorize the patient himself instead of its underlying condition.
- **Validation**: The model will be deployed on a daily basis, therefore only one patient will be predicted at a time.
Partitioning strategy: Group Partitioning

Select partitioning method:

- Random
- Partition Feature
- Group
- Date/Time
- Stratified

All rows with the same single value of the selected feature are in the same partition. Each partition can contain multiple values of the feature.

Group ID feature:

This must be a feature from the dataset.

patient_id
Partitioning strategy: Out of time validation
Partitioning strategy: Out of time validation

What would you like to predict?

What will be the selling price of my car in 1 year?

5 years of historical contracts

Today
Partitioning strategy: Out of time validation

What would you like to predict?

What will be the selling price of my car in 1 year?

5 years of historical contract

Contract

1
2
3
4
5

Today

$\text{p} : \text{Targeted price}$

$\text{t} : \text{Start date}$

$\text{p} : \text{1 year period}$

$\text{1 year blind period}$
Partitioning strategy: Out of time validation

Target Over Time

Observations

2012-07-01

Time

2014-06-14

# of Observations

Available Training Data

Primary Training Data

Gap

Validation

Holdout

Advanced

Holdout

Backtest 1

Backtest 2
Questions & Answers
Downsampling

- Faster Model
- Similar Accuracy
Weight

Differentially weight some observations when calculating error metrics.

Begin typing to choose feature
Partitioning strategy: Group Partitioning

Select partitioning method:

- Random
- Partition Feature
- Group
- Date/Time
- Stratified

A partition is created for each unique value of the selected feature.

Partition Feature:
This must be a feature from the dataset with a cardinality between 2 - 100.

Begin typing to choose feature
DataRobot Community

- **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
- **Recommended Learning Session to view:**
  - March 10: Model Building--Autopilot Done, Now What?
- **Questions:** aisuccess-webinars@datarobot.com