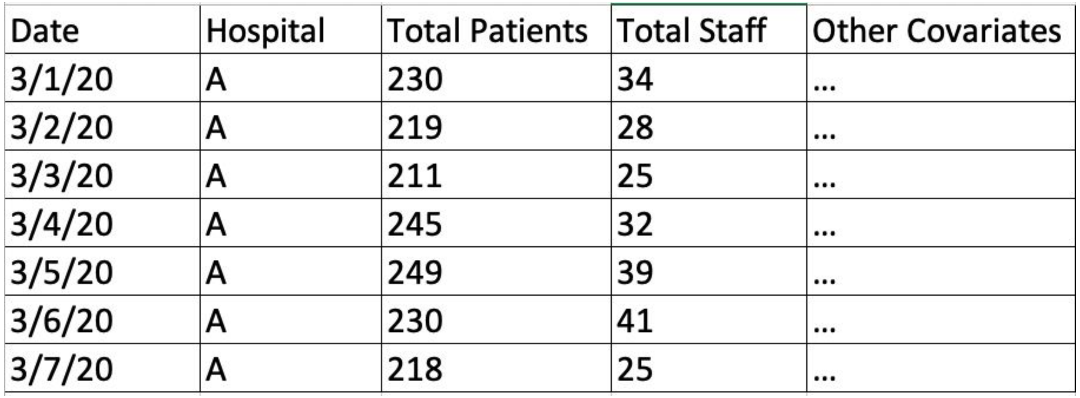
TIME-SERIES MODELING

A. Project Setup

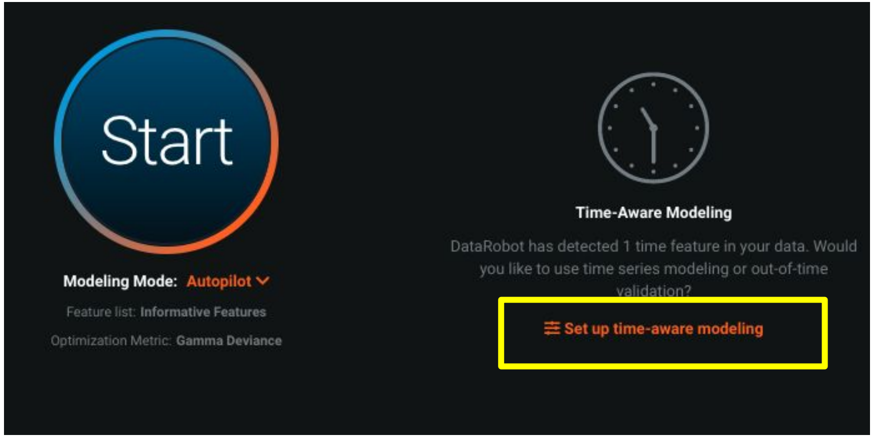
Data Requirements for Time Series



Date	Hospital	Total Patients	Total Staff	Other Covariates
3/1/20	A	230	34	...
3/2/20	A	219	28	...
3/3/20	A	211	25	...
3/4/20	A	245	32	...
3/5/20	A	249	39	...
3/6/20	A	230	41	...
3/7/20	A	218	25	...

- Each row must have unique timestamp; should be in regular intervals (days, hours, minutes, etc.).
- Aggregate rows to the level that makes sense for the use case.
- Multiple series are supported (hospital, department, clinic, etc.)
- Use time series when you want to predict a column into the future.
- Time Series supports covariates (like the AutoML product) - example: “total staff”

Setting up a DataRobot Time Series Project



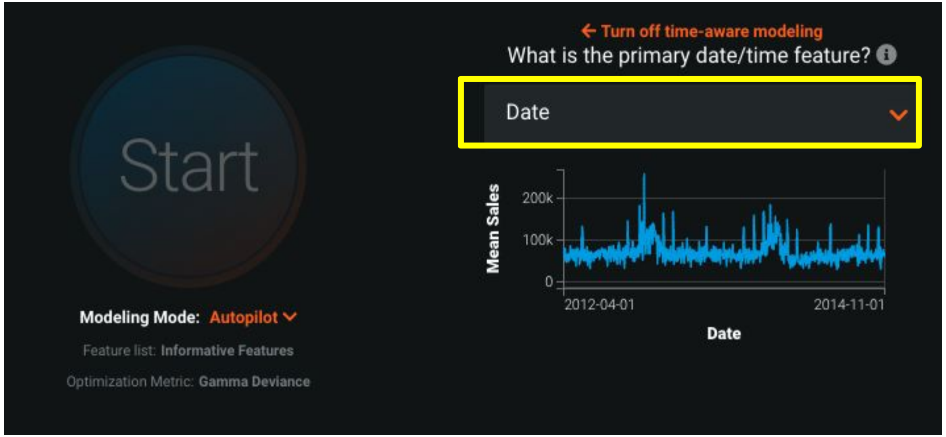
Modeling Mode: Autopilot

Feature list: Informative Features

Optimization Metric: Gamma Deviance

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

After you select your Target column, select "Set up time-aware modeling"



What is the primary date/time feature? Date

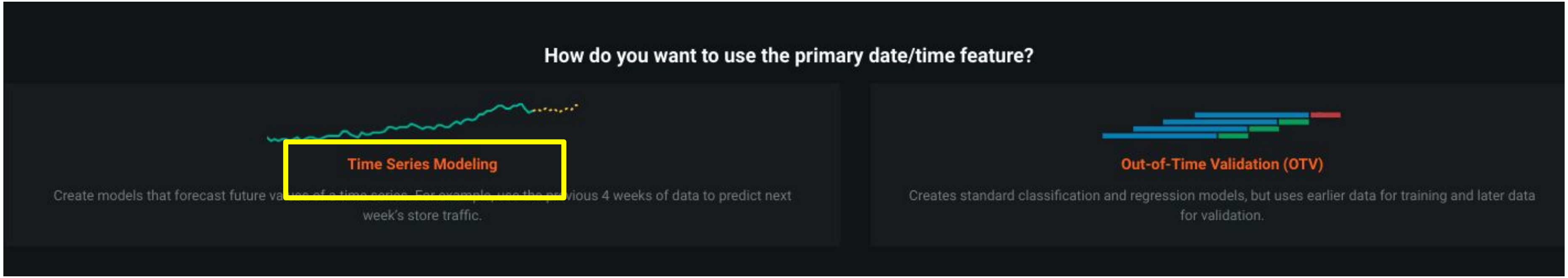
Modeling Mode: Autopilot

Feature list: Informative Features

Optimization Metric: Gamma Deviance

Supply your date/time column

Setup your Time Series project & specify series, if applicable



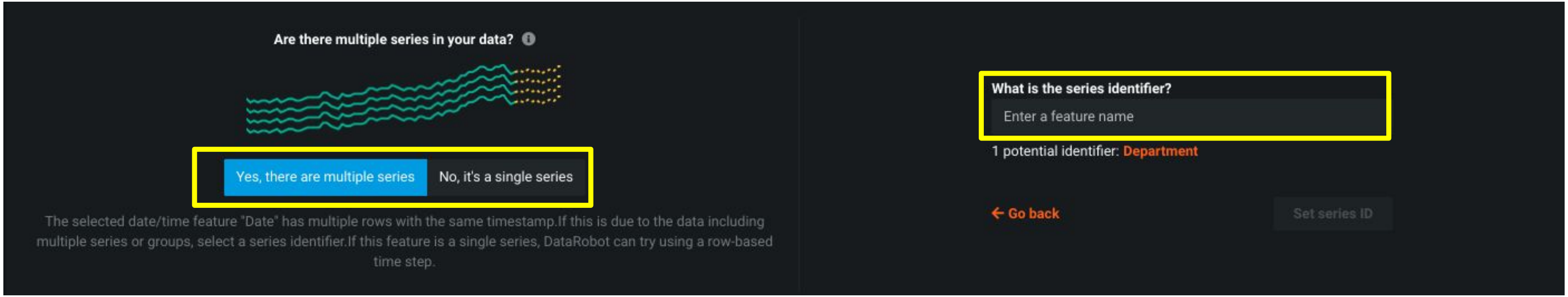
How do you want to use the primary date/time feature?

Time Series Modeling

Create models that forecast future values of a time series. For example, use the previous 4 weeks of data to predict next week's store traffic.

Out-of-Time Validation (OTV)

Creates standard classification and regression models, but uses earlier data for training and later data for validation.



Are there multiple series in your data?

Yes, there are multiple series | No, it's a single series

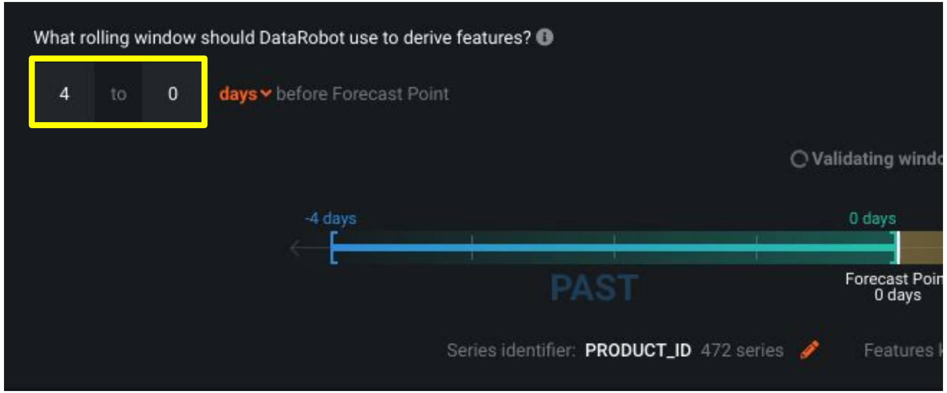
The selected date/time feature "Date" has multiple rows with the same timestamp.If this is due to the data including multiple series or groups, select a series identifier.If this feature is a single series, DataRobot can try using a row-based time step.

What is the series identifier?

Enter a feature name

1 potential identifier: Department

Feature derivation window and forecast distance...



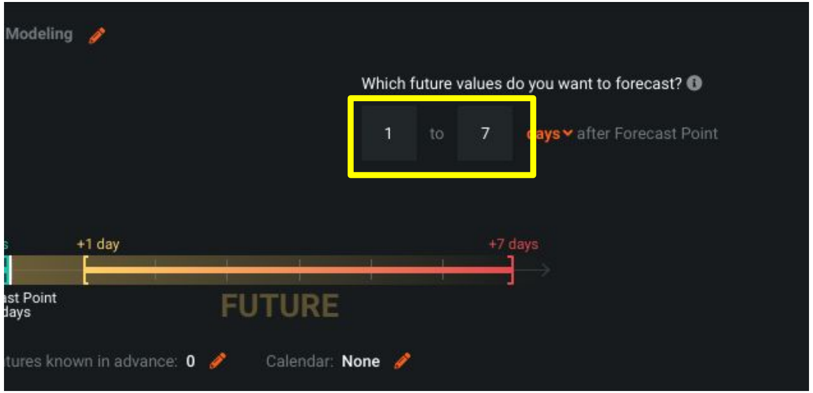
What rolling window should DataRobot use to derive features?

4 to 0 days before Forecast Point

Series identifier: PRODUCT_ID 472 series

Feature derivation window tells DataRobot how far back to look to derive features. Short windows will recover to changes quickly, but longer windows will be more stable (and not overreact to spikes).

Forecast distance tells DataRobot how far into the future it should make predictions. This should be based on what provides business value.



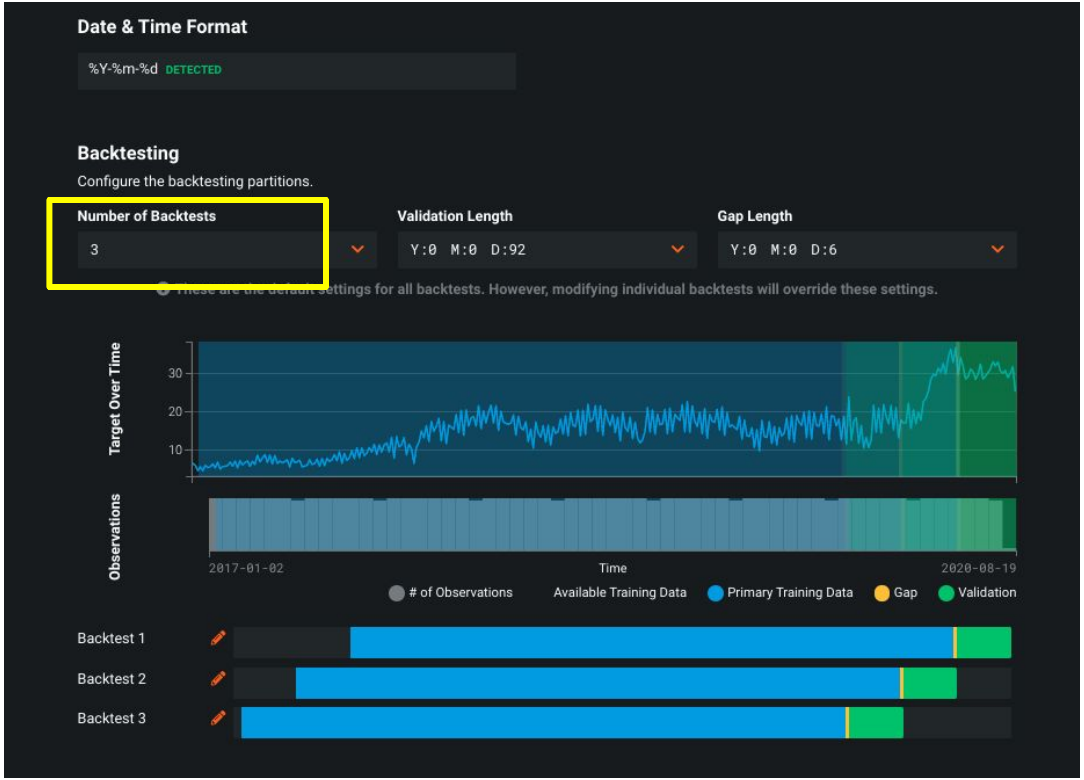
Which future values do you want to forecast?

1 to 7 ays after Forecast Point

tures known in advance: 0

Calendar: None

Validation in a time series project (i.e. “Backtests”)



Date & Time Format

%Y-%m-%d DETECTED

Backtesting

Configure the backtesting partitions.

Number of Backtests: 3

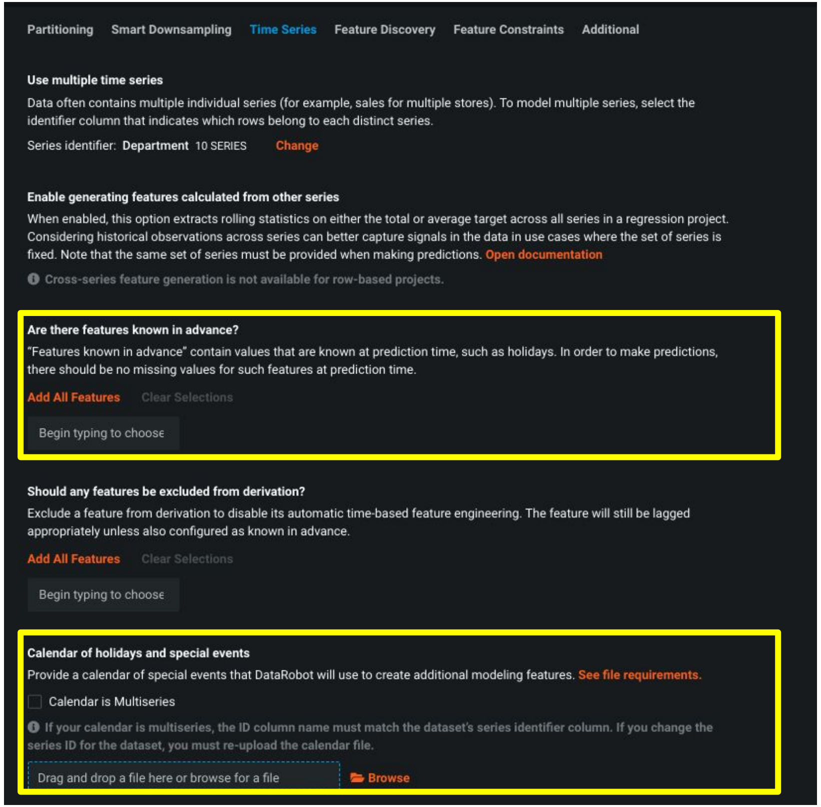
Validation Length: Y:0 M:0 D:92

Gap Length: Y:0 M:0 D:6

These are the default settings for all backtests. However, modifying individual backtests will override these settings.

- Validation in time series (green or red bars in image) is based on a time range after the training period (blue bars) to prevent leakage and provide accurate performance metrics.
- Project configuration will affect the validation windows somewhat - move these around to cover areas of interest (e.g. before, during, after COVID)

Additional Setup



Use multiple time series

Data often contains multiple individual series (for example, sales for multiple stores). To model multiple series, select the identifier column that indicates which rows belong to each distinct series.

Series identifier: **Department** 10 SERIES Change

Enable generating features calculated from other series

When enabled, this option extracts rolling statistics on either the total or average target across all series in a regression project. Considering historical observations across series can better capture signals in the data in use cases where the set of series is fixed. Note that the same set of series must be provided when making predictions. Open documentation

Cross-series feature generation is not available for row-based projects.

Are there features known in advance?

"Features known in advance" contain values that are known at prediction time, such as holidays. In order to make predictions, there should be no missing values for such features at prediction time.

Add All Features Clear Selections

Begin typing to choose

Should any features be excluded from derivation?

Exclude a feature from derivation to disable its automatic time-based feature engineering. The feature will still be lagged appropriately unless also configured as known in advance.

Add All Features Clear Selections

Begin typing to choose

Calendar of holidays and special events

Provide a calendar of special events that DataRobot will use to create additional modeling features. See file requirements.

□ Calendar is Multiseries

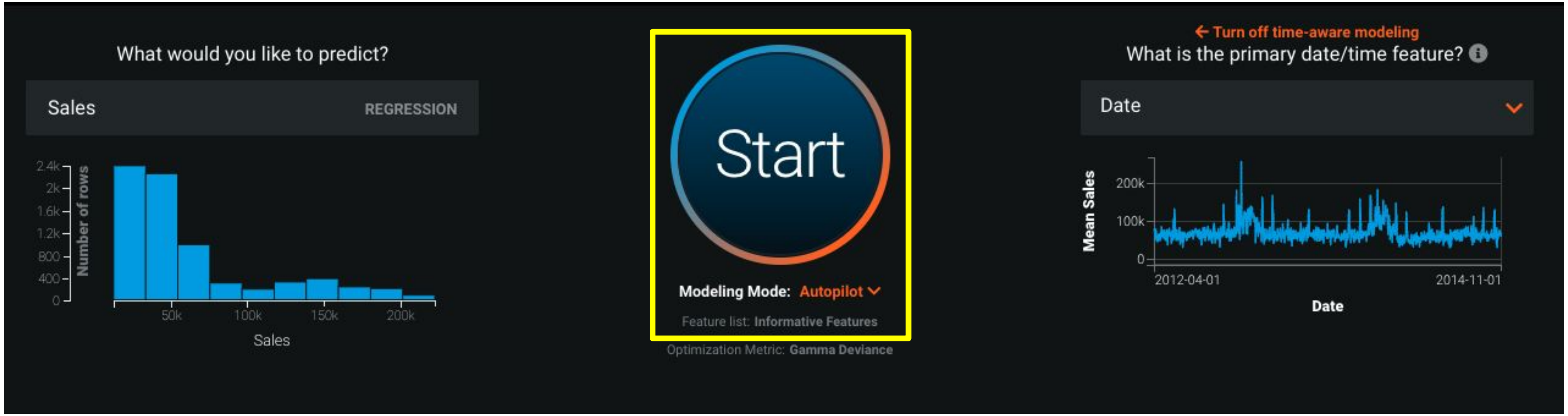
If your calendar is multiseries, the ID column name must match the dataset's series identifier column. If you change the series ID for the dataset, you must re-upload the calendar file.

Drag and drop a file here or browse for a file Browse

1) Known-in-Advance features are columns in your data that the model can use to make the current prediction, such as new product rollout or planned outages, etc.
If you're not sure, leave it off this list.

2) Holidays are also known in advance but can be provided as a separate csv file.

Ready to start building models?



What would you like to predict?

Sales REGRESSION

Modeling Mode: Autopilot

Feature list: Informative Features

Optimization Metric: Gamma Deviance

← Turn off time-aware modeling

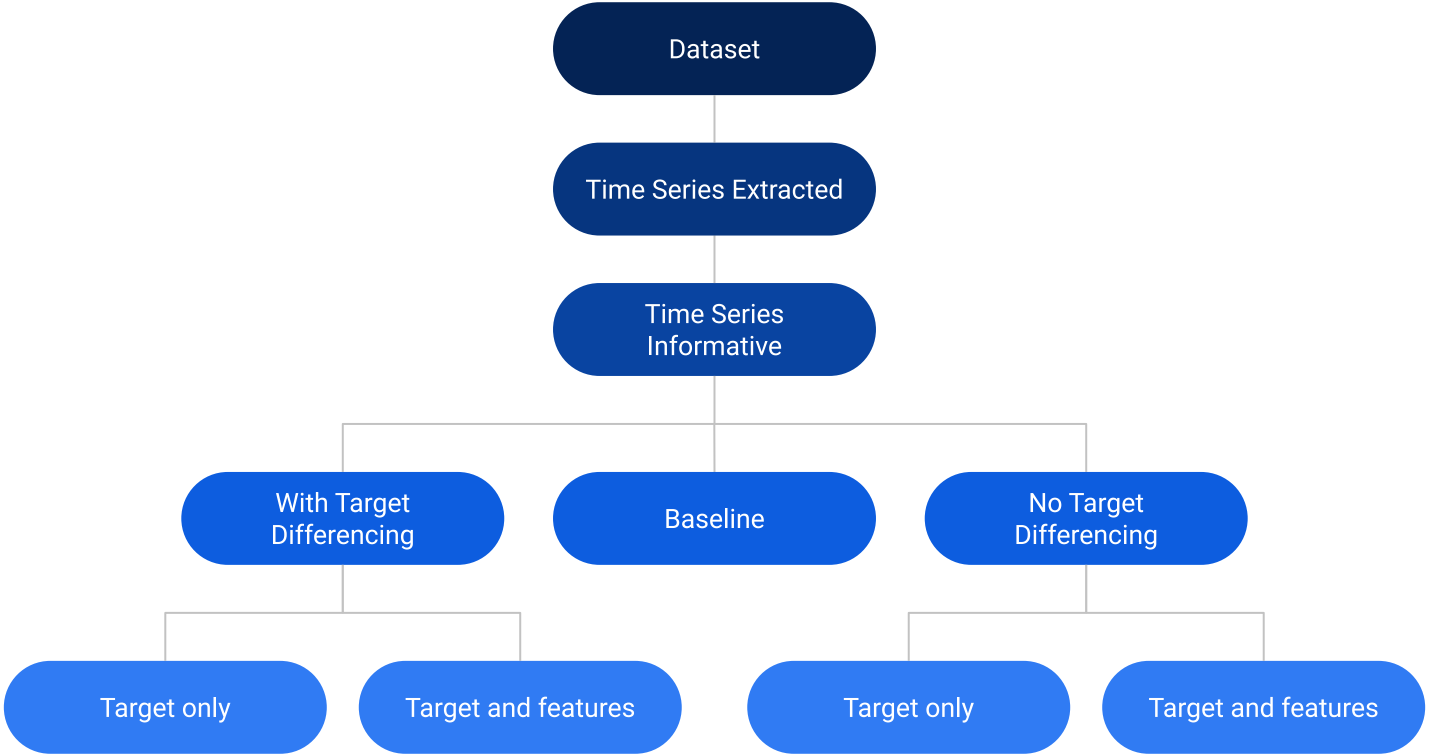
What is the primary date/time feature?

Date

- Correct any data issues that DataRobot highlighted during setup
- Verify your **feature derivation** windows & **forecast distance** ranges
- Adjust your **backtest** time ranges if needed (or just accept the defaults)
- Double-check your **known-in-advance** features + **holidays**
- Run Autopilot

B.Evaluation

Feature List Hierarchy



Dataset

Time Series Extracted

Time Series Informative

With Target Differencing

Baseline

No Target Differencing

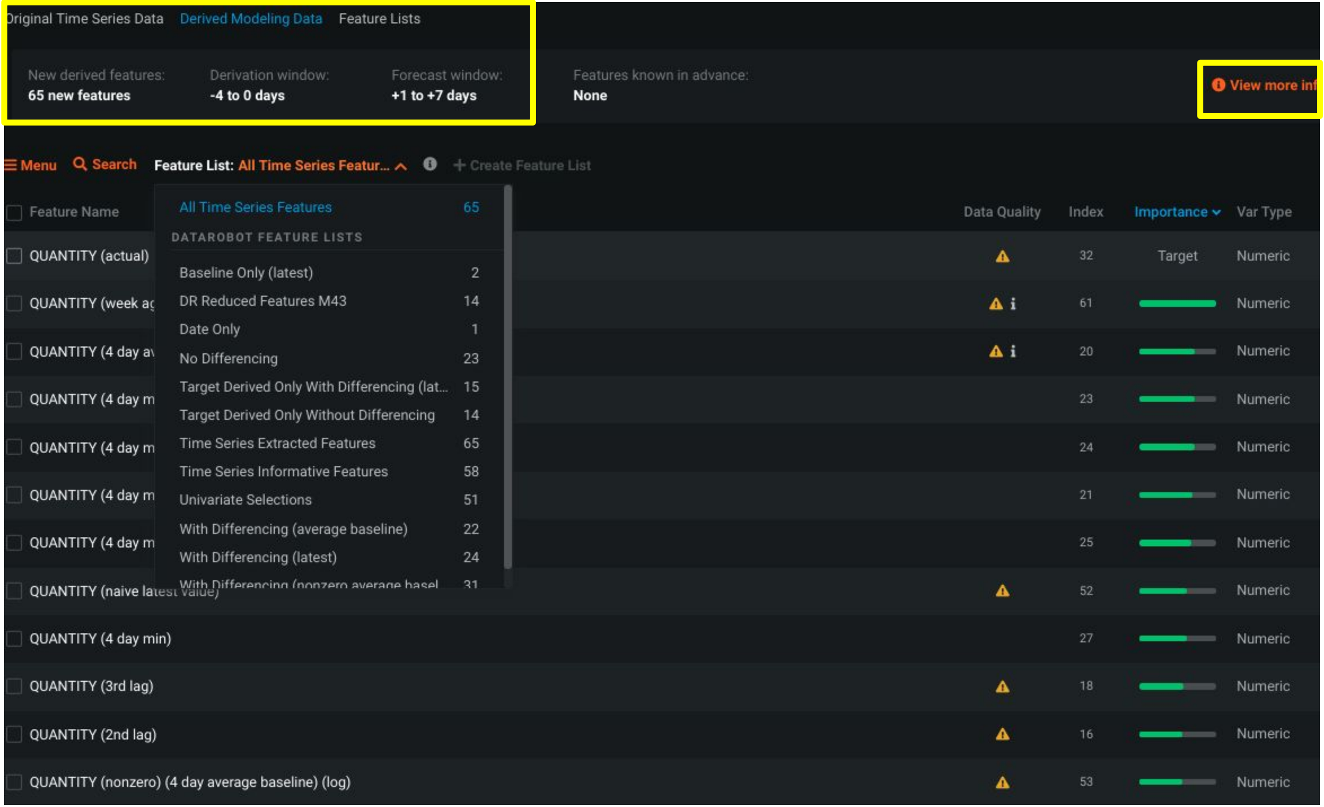
Target only

Target and features

Target only

Target and features

Check out the Derived Features



riginal Time Series Data   Derived Modeling Data   Feature Lists

New derived features: **65 new features**

Derivation window: **-4 to 0 days**

Forecast window: **+1 to +7 days**

Features known in advance: **None**

View more inf

Feature List: All Time Series Featur...

Feature List	Count
All Time Series Features	65
DATAROBOT FEATURE LISTS	
Baseline Only (latest)	2
DR Reduced Features M43	14
Date Only	1
No Differencing	23
Target Derived Only With Differencing (lat...	15
Target Derived Only Without Differencing	14
Time Series Extracted Features	65
Time Series Informative Features	58
Univariate Selections	51
With Differencing (average baseline)	22
With Differencing (latest)	24
With Differencing (nonzero average basel	31

□ Feature Name	Data Quality	Index	Importance	Var Type
□ QUANTITY (actual)	⚠	32	Target	Numeric
□ QUANTITY (week ag	⚠ i	61		Numeric
□ QUANTITY (4 day av	⚠ i	20		Numeric
□ QUANTITY (4 day m		23		Numeric
□ QUANTITY (4 day m		24		Numeric
□ QUANTITY (4 day m		21		Numeric
□ QUANTITY (4 day m		25		Numeric
□ QUANTITY (naive latest value)	⚠	52		Numeric
□ QUANTITY (4 day min)		27		Numeric
□ QUANTITY (3rd lag)	⚠	18		Numeric
□ QUANTITY (2nd lag)	⚠	16		Numeric
□ QUANTITY (nonzero) (4 day average baseline) (log)	⚠	53		Numeric

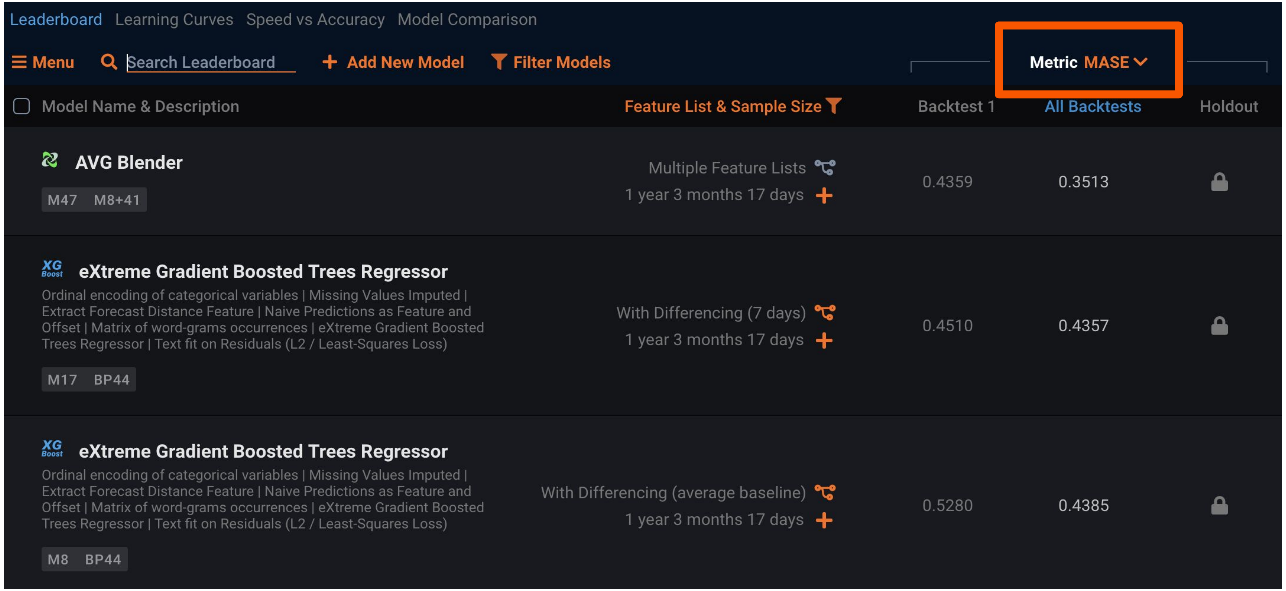
Explore the different feature lists and their distributions.

You can also get derivation log...

How do we assess quality? | Leaderboard

Mean Absolute Scaled Error (MASE)

$$\text{Mean Absolute Scaled Error (MASE)} = \frac{\text{Mean Absolute Error (MAE) of } \textbf{Model of Interest}}{\text{Mean Absolute Error (MAE) of } \textbf{Baseline}}$$

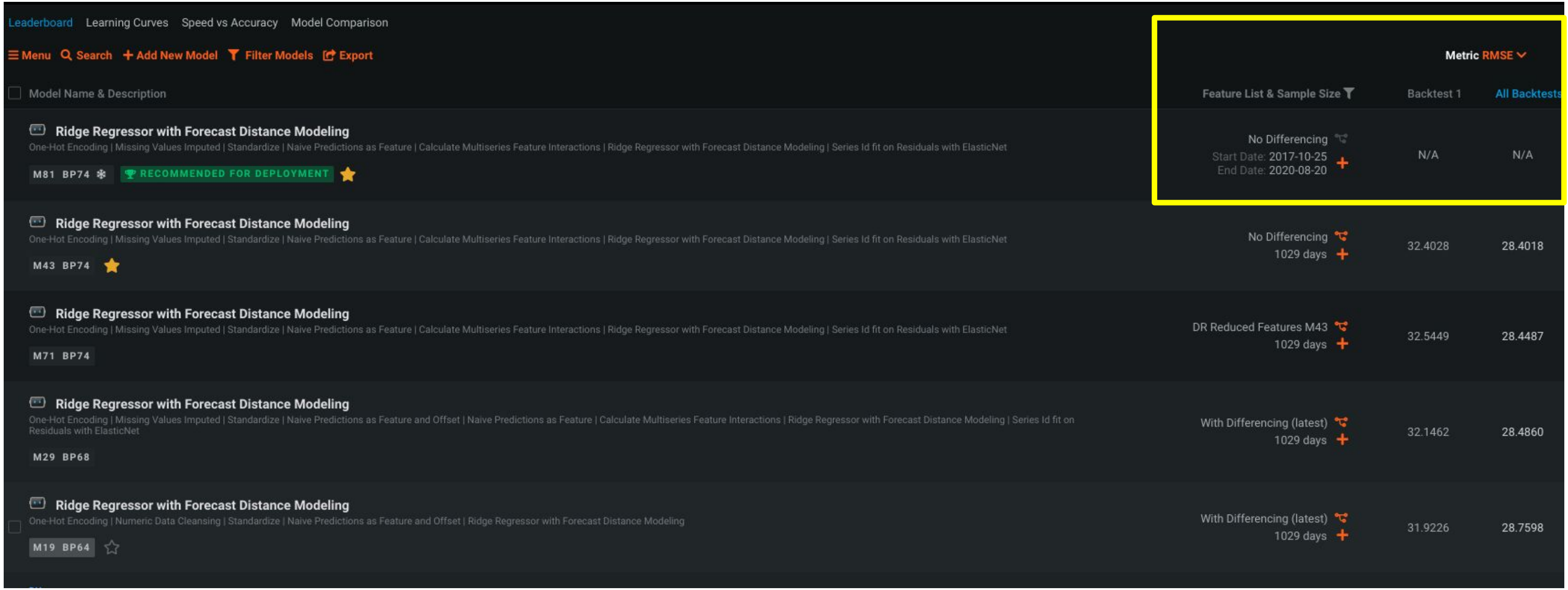


Leaderboard Learning Curves Speed vs Accuracy Model Comparison

☰ Menu | Search Leaderboard | + Add New Model | Filter Models | Metric MASE

□ Model Name & Description	Feature List & Sample Size	Backtest 1	All Backtests	Holdout
AVG Blender M47 M8+41	Multiple Feature Lists 1 year 3 months 17 days	0.4359	0.3513	🔒
eXtreme Gradient Boosted Trees Regressor Ordinal encoding of categorical variables | Missing Values Imputed | Extract Forecast Distance Feature | Naive Predictions as Feature and Offset | Matrix of word-grams occurrences | eXtreme Gradient Boosted Trees Regressor | Text fit on Residuals (L2 / Least-Squares Loss) M17 BP44	With Differencing (7 days) 1 year 3 months 17 days	0.4510	0.4357	🔒
eXtreme Gradient Boosted Trees Regressor Ordinal encoding of categorical variables | Missing Values Imputed | Extract Forecast Distance Feature | Naive Predictions as Feature and Offset | Matrix of word-grams occurrences | eXtreme Gradient Boosted Trees Regressor | Text fit on Residuals (L2 / Least-Squares Loss) M8 BP44	With Differencing (average baseline) 1 year 3 months 17 days	0.5280	0.4385	🔒

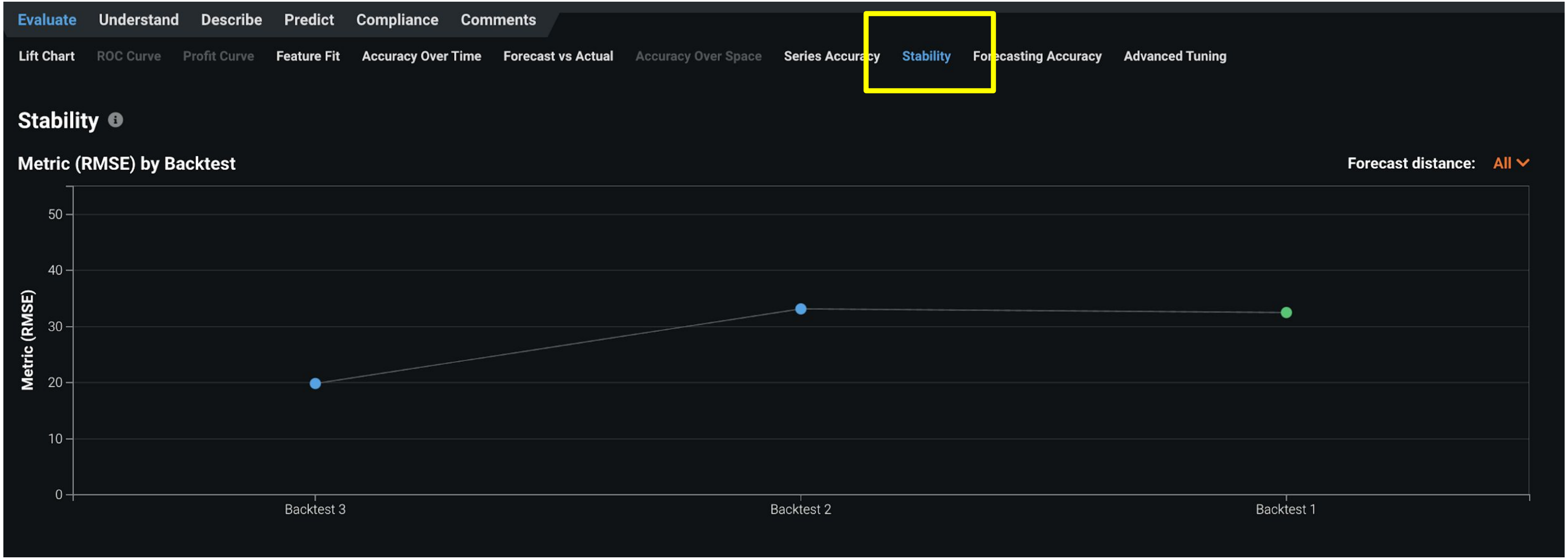
Find the Recommended Model on Leaderboard



Model Name & Description	Feature List & Sample Size	Backtest 1	All Backtests
Ridge Regressor with Forecast Distance Modeling One-Hot Encoding | Missing Values Imputed | Standardize | Naive Predictions as Feature | Calculate Multiseries Feature Interactions | Ridge Regressor with Forecast Distance Modeling | Series Id fit on Residuals with ElasticNet M81 BP74 ❄ RECOMMENDED FOR DEPLOYMENT ★	No Differencing Start Date: 2017-10-25 End Date: 2020-08-20	N/A	N/A
Ridge Regressor with Forecast Distance Modeling One-Hot Encoding | Missing Values Imputed | Standardize | Naive Predictions as Feature | Calculate Multiseries Feature Interactions | Ridge Regressor with Forecast Distance Modeling | Series Id fit on Residuals with ElasticNet M43 BP74 ★	No Differencing 1029 days	32.4028	28.4018
Ridge Regressor with Forecast Distance Modeling One-Hot Encoding | Missing Values Imputed | Standardize | Naive Predictions as Feature | Calculate Multiseries Feature Interactions | Ridge Regressor with Forecast Distance Modeling | Series Id fit on Residuals with ElasticNet M71 BP74	DR Reduced Features M43 1029 days	32.5449	28.4487
Ridge Regressor with Forecast Distance Modeling One-Hot Encoding | Missing Values Imputed | Standardize | Naive Predictions as Feature and Offset | Naive Predictions as Feature | Calculate Multiseries Feature Interactions | Ridge Regressor with Forecast Distance Modeling | Series Id fit on Residuals with ElasticNet M29 BP68	With Differencing (latest) 1029 days	32.1462	28.4860
□ **Ridge Regressor with Forecast Distance Modeling** One-Hot Encoding | Numeric Data Cleansing | Standardize | Naive Predictions as Feature and Offset | Ridge Regressor with Forecast Distance Modeling M19 BP64 ☆	With Differencing (latest) 1029 days	31.9226	28.7598

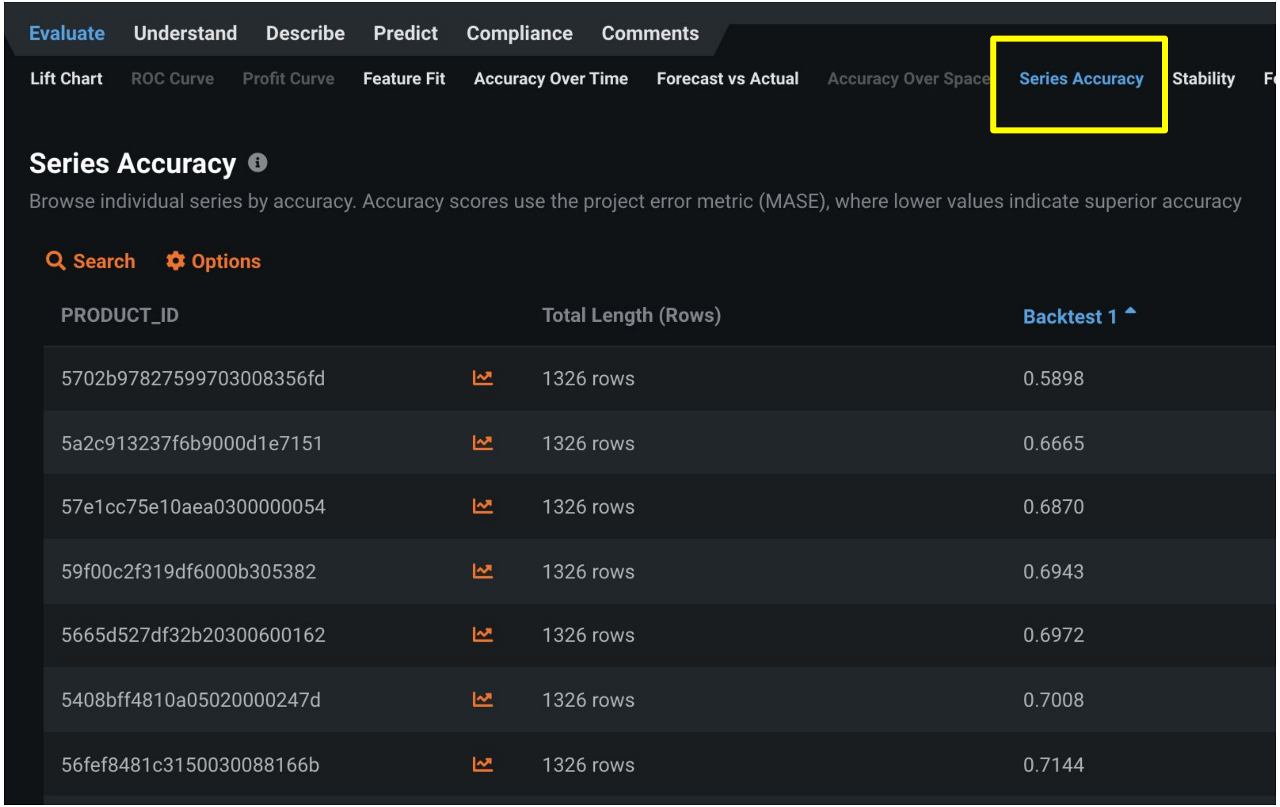
You can adjust the Metrics to re-rank the models. There's no accuracy on the recommended model - it's trained to end of the data,

Stability - How does model perform over time?



For RMSE, lower is better - accuracy is significantly better in backtest 3

Series Accuracy - which series has best/worst accuracy?



Series Accuracy

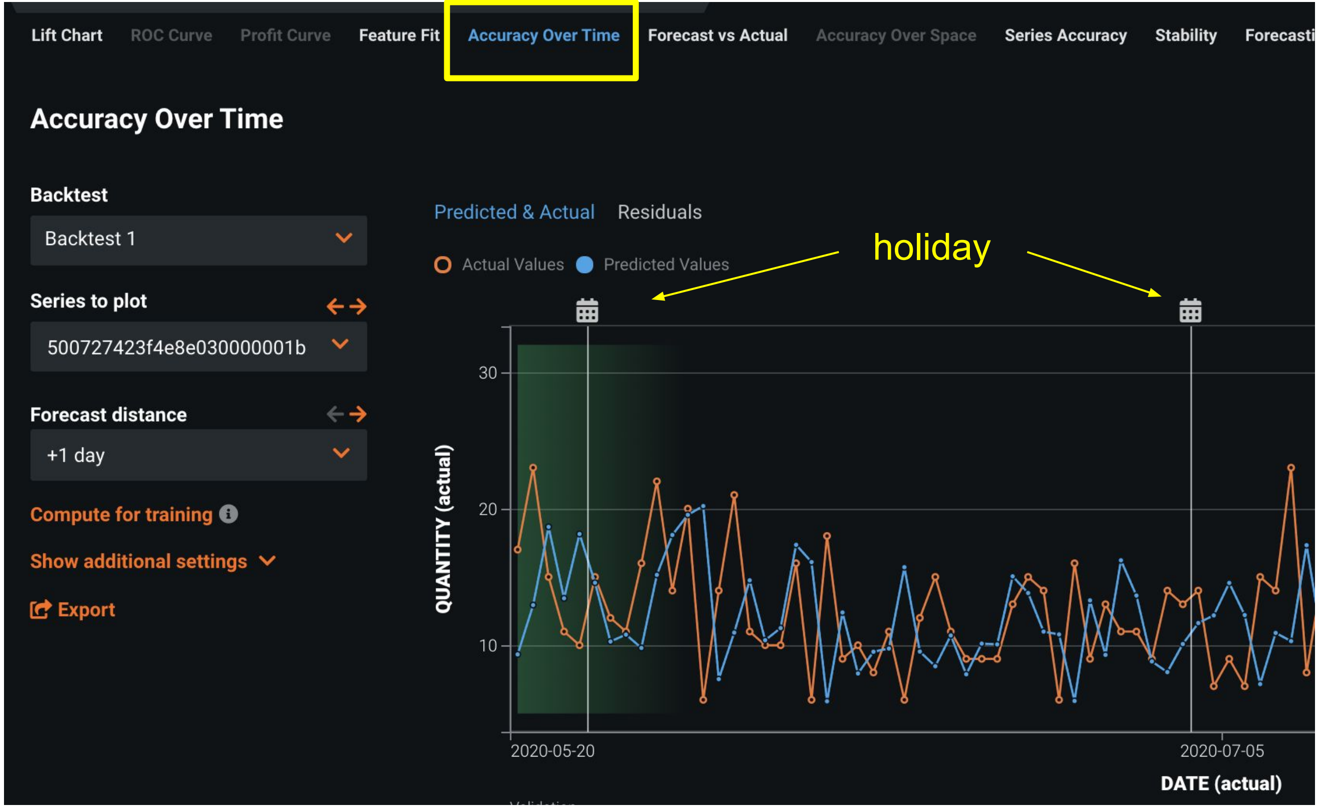
Browse individual series by accuracy. Accuracy scores use the project error metric (MASE), where lower values indicate superior accuracy

PRODUCT_ID	Total Length (Rows)	Backtest 1
5702b97827599703008356fd	1326 rows	0.5898
5a2c913237f6b9000d1e7151	1326 rows	0.6665
57e1cc75e10aea0300000054	1326 rows	0.6870
59f00c2f319df6000b305382	1326 rows	0.6943
5665d527df32b20300600162	1326 rows	0.6972
5408bff4810a05020000247d	1326 rows	0.7008
56fef8481c3150030088166b	1326 rows	0.7144

Which series had the best accuracy? The worst?

From here you can identify potential areas for improvement or exclusion.

Accuracy over Time - evaluate predictions



Backtest: Backtest 1

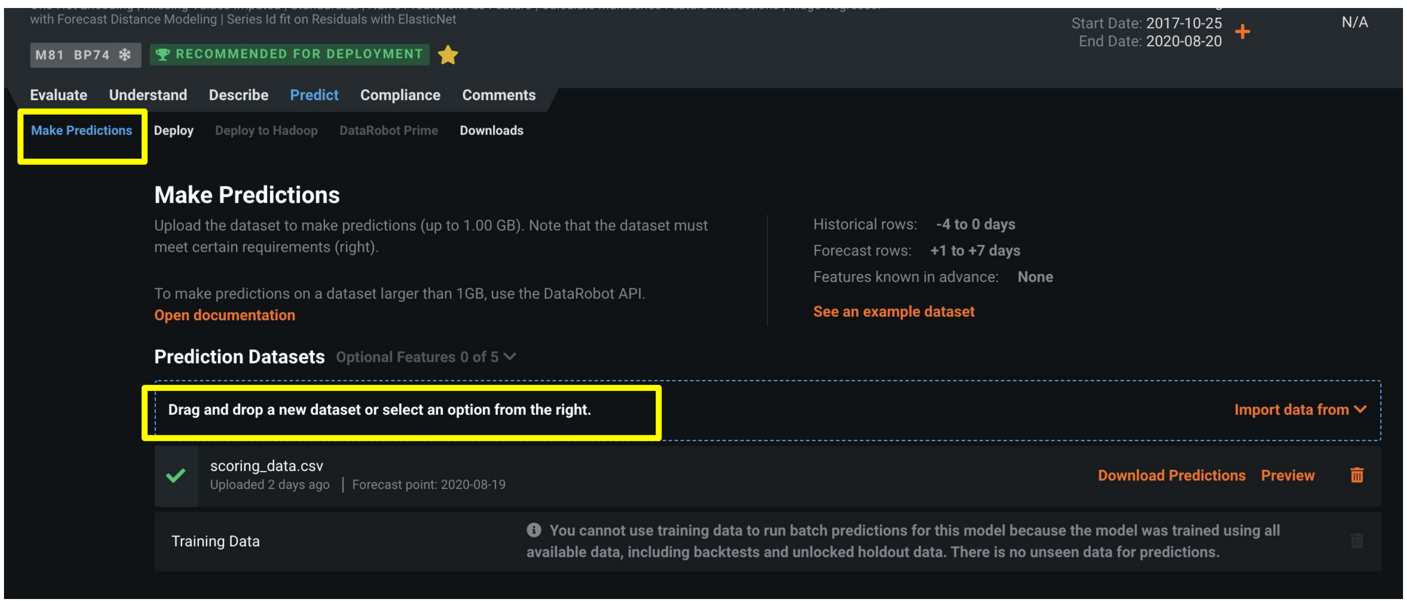
Series to plot: 500727423f4e8e030000001b

Forecast distance: +1 day

holiday

Can display for each backtest, series, and forecast distance.

Making Predictions



with Forecast Distance Modeling | Series Id fit on Residuals with ElasticNet

M81 BP74 ❄ RECOMMENDED FOR DEPLOYMENT ★

Start Date: 2017-10-25
End Date: 2020-08-20 +

N/A

Evaluate | Understand | Describe | Predict | Compliance | Comments

Make Predictions | Deploy | Deploy to Hadoop | DataRobot Prime | Downloads

Make Predictions

Upload the dataset to make predictions (up to 1.00 GB). Note that the dataset must meet certain requirements (right).

To make predictions on a dataset larger than 1GB, use the DataRobot API.
Open documentation

Historical rows: -4 to 0 days
Forecast rows: +1 to +7 days
Features known in advance: None

See an example dataset

Prediction Datasets Optional Features 0 of 5

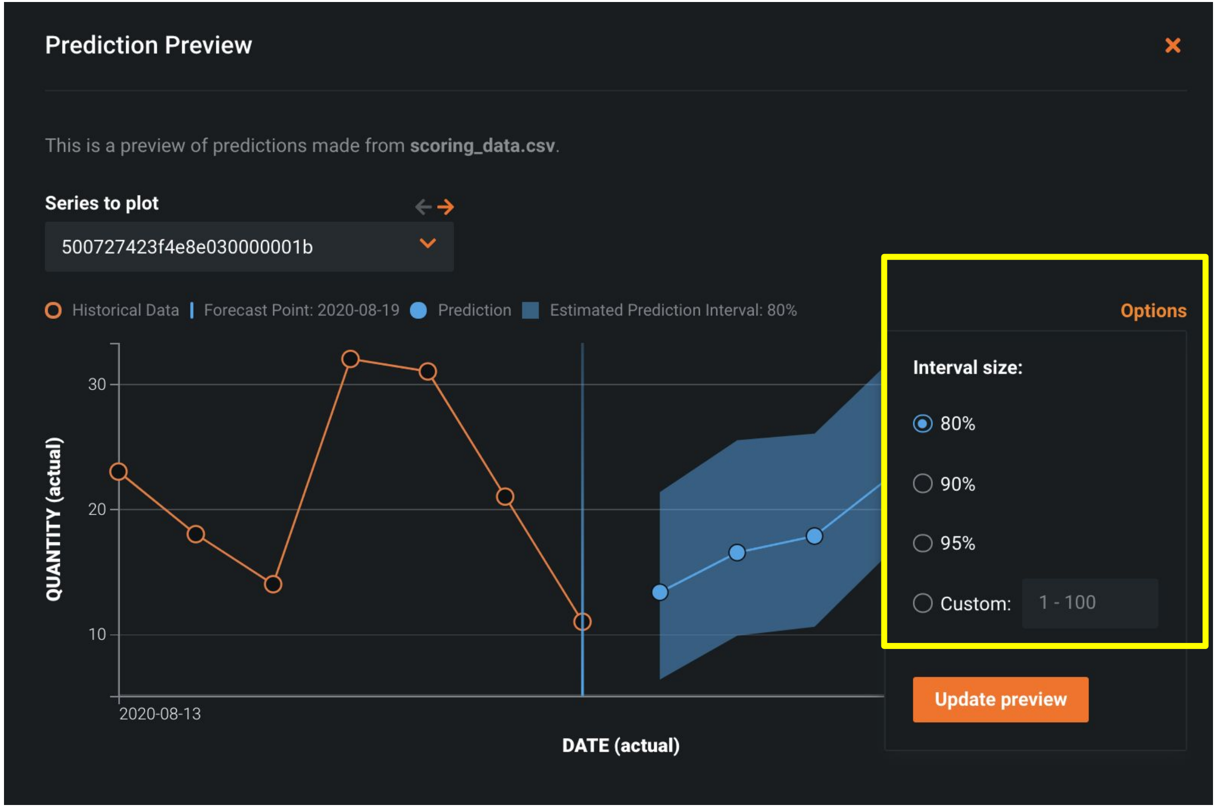
Drag and drop a new dataset or select an option from the right. Import data from

✓ scoring_data.csv
Uploaded 2 days ago | Forecast point: 2020-08-19
Download Predictions Preview

Training Data
You cannot use training data to run batch predictions for this model because the model was trained using all available data, including backtests and unlocked holdout data. There is no unseen data for predictions.

You must prepare a file with a record for each series and each desired forecast date (i.e. next 7 days…) and any known-in-advance features.

Prediction Previews & Intervals



Prediction Preview

This is a preview of predictions made from **scoring_data.csv**.

Series to plot

500727423f4e8e030000001b

○ Historical Data | Forecast Point: 2020-08-19 ● Prediction ▨ Estimated Prediction Interval: 80%

Options

Interval size:

● 80%

○ 90%

○ 95%

○ Custom: 1 - 100

The intervals* estimate the range of values DataRobot expects actual values of the target to fall within. They are based on the residual errors measured during the model's backtesting.

* to generate intervals, change and update preview, then download