1 How To Use This Document

Highly regulated industries, such as banking and insurance, must comply with government regulations for model validation before a model can be put into production. This includes creating robust model development documentation. DataRobot automates the generation of model documentation, expediting the process required for regulatory compliance and following best practice for reducing model risk.

This document is split into two components: those sections that are automatically produced by DataRobot and those that require further input by the user. The sections in blue italicized font include specific instructions for the documenter and require additional user input of organization-specific information, such as business use cases, data sources, and implementation details. Once the sections are complete, remove the instructions. The remaining sections in non-blue italicized font are automatically populated by DataRobot and require no further input.

Copyright ©2024, DataRobot, Inc.



Table of Contents

* 1 How To Use This Document
* 2 DataRobot Model Development Documentation
* 3 Executive Summary and Model Overview
* 3.1 Model Stakeholders
* 3.2 Model Development Purpose and Intended Use
* 3.3 Model Description and Overview
* 3.4 Overview of Model Results
* 3.5 Model Interdependencies
* 4 Model Data Overview
* 4.1 Data Source Overview and Appropriateness
* 4.2 Input Data Extraction, Preparation, and Quality & Completeness
* 4.3 Data Assumptions
* 5 Model Theoretical Framework and Methodology
* 5.1 Model Development Overview
* 5.2 Model Methodology
* 5.2.1 Ordinal scale converter of categorical features
* 5.2.2 Arbitrary or median value-based numeric imputation (V4 with memory usage optimization)
* 5.2.3 Forecast Distance column extraction transformer
* 5.2.4 Extracts Naive Predictions
* 5.2.5 Extreme Gradient Boosting Regressor with Grid Search and Early Stopping support
* 5.2.6 Forecast Distance modeler using ElasticNet modelers for each value of forecast distance
* 5.3 Literature Review and References
* 5.4 Alternative Model Frameworks and Theories Considered
* 5.5 Variable Selection
* 5.5.1 DataRobot Quantitative Analysis
* 5.5.2 Expert Judgement and Variable Selection
* 5.5.3 Final Model Variables
* 5.5.3.1 Model Features and Summary Statistics
* 5.5.3.2 Data Quality Handling Report
* 6 Model Performance and Stability
* 6.1 Model Validation Stability
* 6.1.1 Data Partitioning Methodology
* 6.2 Model Performance (Backtesting)
* 6.3 Sensitivity Testing and Analysis
* 6.3.1 Lift Chart
* 6.3.2 Key Relationships
* 6.3.3 Sensitivity Analysis (Partial Dependence)
* 6.3.4 Accuracy Over Time
* 7 Model Implementation and Output Reporting
* 7.1 Version Control

2 DataRobot Model Development Documentation

A key component of effective model risk management is sufficiently detailed documentation for model development, implementation, and use, so that reasonable parties unfamiliar with a model can understand how the model operates, its limitations, and its key assumptions. Additionally, model documentation should contain enough detail for an independent party (e.g., independent model validation) to replicate all aspects of the underlying modeling process.

The purpose of this document is not to be prescriptive in format and content, but rather to serve as a guide in creating sufficiently rigorous model development, implementation, and use documentation. The documentation should provide enough evidence to show that the components of the model work as intended, the model is appropriate for its intended business purpose, and that it is conceptually sound.

3 Executive Summary and Model Overview

3.1 Model Stakeholders

Describe the model's purpose and its intended business use. Describe all stakeholders of this model, including their role, line-of-business, and team. This should include stakeholders of model ownership, model development, model implementation, and model risk management.

Model Owner(s): The individual who owns the business risk addressed by the model and provides approval for the model to be used within the line-of-business or enterprise function.

Model Developer(s): The individual responsible for building new models with DataRobot or maintaining existing models.

Model User(s): Those teams who will use the model output as part of their ongoing business operations.

Model Validator(s): The validators are responsible for independent model review and approval prior to its first use.

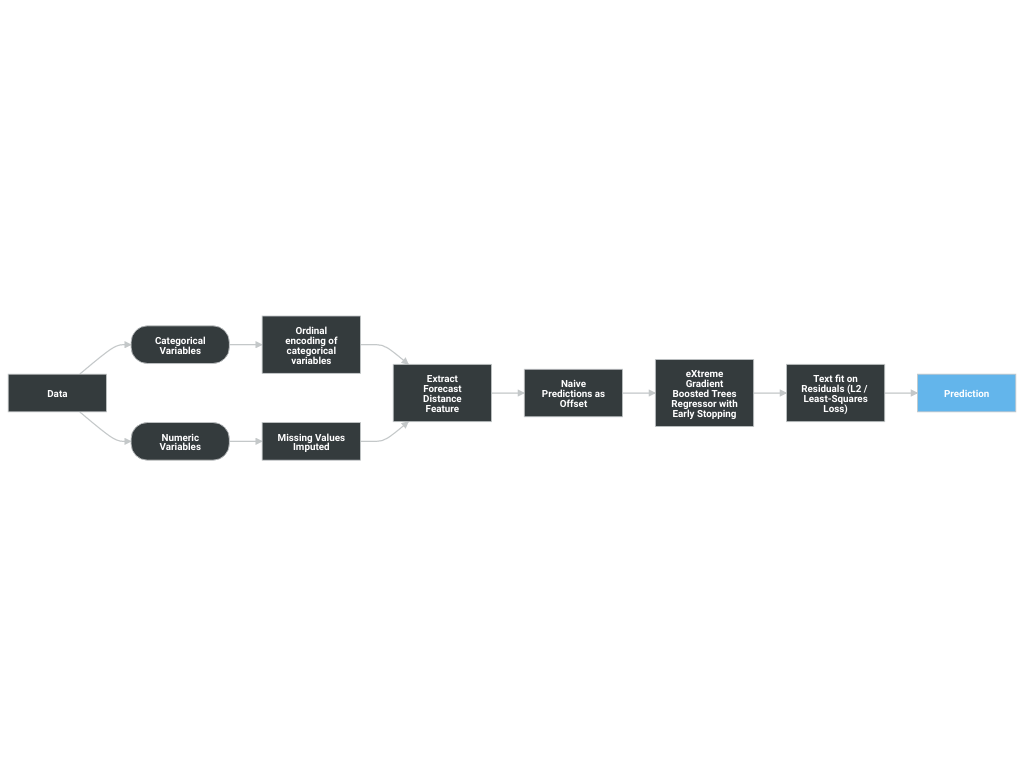
3.2 Model Development Purpose and Intended Use

Describe the model's purpose, including a summary of the business need for this particular model. Concisely describe how the model will be used to address this business problem. Furthermore, describe with great precision all model uses covered by this document. These descriptions will address this statement made in regulatory guidance, FRB SR-11-7, "Even a fundamentally sound model producing accurate outputs consistent with the design objective of the model may exhibit high model risk if it is misapplied or misused."

3.3 Model Description and Overview

The particular model referenced in this document: eXtreme Gradient Boosted Trees Regressor with Early Stopping. This model was developed in a project created with v11.0.0 of DataRobot. This model is denoted within DataRobot by the Project ID: 67602e59a2dead72473e24ba and the Model ID: 67602f4964721677fa700fa9. The project was created on 2024-12-16 13:42:49.

The model development workflow process (i.e., the model blueprint) is detailed in the figure below.



A Blueprint represents the high-level end-to-end procedure for fitting the model, including any preprocessing steps, algorithms, and post-processing. It illustrates the many steps involved in transforming input predictors and targets into a model. Each element (or, “node”) in a blueprint can represent multiple steps.

The following elements connect to create the blueprint:

* Ordinal encoding of categorical variables
* Missing Values Imputed
* Extract Forecast Distance Feature
* Naive Predictions as Offset
* eXtreme Gradient Boosted Trees Regressor with Early Stopping
* Text fit on Residuals (L2 / Least-Squares Loss)

3.4 Overview of Model Results

DataRobot runs performance testing during the model development process to evaluate model results and reliability. An overview of the out-of-sample performance scores are included below. The performance metric used for this project was RMSE and the project included a total of 2,334 observations.

|  |  |
| --- | --- |
| Scoring Type | Score (RMSE) |
| backtesting\_scores | 33246.1907, 35644.1867, 45137.219, 47146.8822, 39967.097, 62028.7678, 33292.6536 |
| holdout | 37362.439 |
| validation | 33246.1907 |

3.5 Model Interdependencies

Understanding interdependent relationships allows for enhanced understanding and improved ability to manage and aggregate model risk. Explain how this model is interconnected with other models in the model inventory--that is, the relative direction with regard to the model's position in the receiving input/sending output role. In addition to the directional relationship, also provide a brief description of each interconnected model.

4 Model Data Overview

4.1 Data Source Overview and Appropriateness

Explain how the data is suitable and relevant for the business problem and model use. For example:

Describe how, and from where, the data was obtained.

Provide a detailed description of the data source and its relevance to the business problem being addressed by this model.

Assess whether the data used for model development is appropriate given the populations to which the model will be applied.

If the model development and model implementation data sources differ, provide a detailed explanation justifying the use of different data sources.

4.2 Input Data Extraction, Preparation, and Quality & Completeness

Provide a detailed description of the data extraction and preparation process, and discuss any analysis conducted to confirm the data are complete and of sufficient quality (e.g., data validation). Include a detailed description of the data extraction process, hierarchical by extraction and preparation stage, and calling sequence. Provide data extraction code (e.g., SQL, Spark, etc.) in the Appendix.

Review and comment on any data weaknesses and limitations and their probable potential effects on the model. For example, data truncation, extraction timing, through-the-cycle data, and data exclusions could potentially cause unintended effects on the model.

4.3 Data Assumptions

Comment on data assumptions, the potential effects on the model, and any mitigating data controls. For example, assumptions related to data truncation, extraction timing, through-the-cycle data, reliability of source system, manual data overrides or imputation, and data exclusions could potentially cause unintended effects on the model.

5 Model Theoretical Framework and Methodology

5.1 Model Development Overview

DataRobot simplifies model development by performing a parallel heuristic search for the best model or ensemble of models, based on both the characteristics of the data and the prediction target. While some machine learning techniques tend to consistently outperform others, it is rarely possible to say in advance which will perform best for a given business problem. Therefore, during the modeling process, DataRobot develops dozens of independent challenger models, exposes the details of how these models were built and how they perform, and enables the user to select the best model for the particular business problem being addressed.

The fundamental workflow within DataRobot for model development is as follows:

* Data Ingestion: The user creates a modeling dataset that includes the prediction target and loads it into DataRobot for time-aware modeling.
* Target Selection: The user identifies the target prediction and selects the primary date/time feature for which the prediction target will vary over time. This will enable time-aware modeling using DataRobot. DataRobot detects whether the target is categorical or numeric. If the target is categorical, DataRobot selects and builds classification blueprints. If the target is numeric, DataRobot selects and builds regression blueprints. DataRobot also selects an optimization performance metric based on the type of supervised learning problem, which can be changed by the user.
* Automated Data Preparation: DataRobot automatically partitions the input dataset into date/time partitioning for validation, which can also be defined by the user.
* Stationarity Analysis & Feature Engineering: DataRobot's time series functionality works by encoding time-sensitive components (such as lags and moving averages) as features, transforming your original input dataset into a modeling dataset that can use conventional machine learning techniques. DataRobot automatically creates and selects time series features in the modeling data and automatically detects whether a project's target value is stationary (that is, whether the statistical properties of the target are constant over time). If the target is not stationary, DataRobot attempts to make it stationary by applying a differencing strategy prior to modeling. This improves the accuracy and robustness of the underlying models.
* DataRobot uses information about the selected target variable and predictors to define a set of candidate blueprints for analysis. It then trains models for each blueprint and ranks them on the model Leaderboard based on a validation and holdout accuracy score.
* Transparent Model Evaluation and Selection: DataRobot has built-in diagnostic tools to assess model accuracy and performance. Once DataRobot has trained and tested models, users can access them from the Leaderboard. From there, users can review model accuracy and, using built-in model diagnostic tools, understand how each independently built model performs. DataRobot provides many metrics for evaluating model accuracy and performance.
* Model Deployment and Monitoring: Once the final model is selected, DataRobot provides efficient solutions for deployment (i.e., model implementation) and monitoring.

5.2 Model Methodology

The modeling workflow consists of the following elements, which connect to create the blueprint:

* Ordinal encoding of categorical variables
* Missing Values Imputed
* Extract Forecast Distance Feature
* Naive Predictions as Offset
* eXtreme Gradient Boosted Trees Regressor with Early Stopping
* Text fit on Residuals (L2 / Least-Squares Loss)

The following subsections include details for each node of the modeling blueprint.

5.2.1 Ordinal scale converter of categorical features

For a categorical feature, convert categorical levels to an ordinal scale. The ordinal scale is 0 to (unique values of categorical\_var) - 1. Rare categories (=other) and missing values are encoded as -1 and -2, respectively. Mapping is based on the lexicographic ordering of the categorical values, the frequency of the levels, the response, or is done randomly.

Ordinal encoding is effective for tree-based models, as it usually performs as well as one-hot encoding but requires fewer computational resources (memory and cpu).

Ordinal encoding, however, does not work for linear methods.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| select | add\_cols\_metadata | If specified, add -cols to metadata. values: [False, True] | False |
| select | add\_maps\_metadata | If specified, add -maps to metadata. values: [False, True] | False |
| multi | card\_max | Maximum number of categorical feature levels allowed. If None, a feature with any number of levels is allowed. values: {'int': [1, 9999999], 'select': None} | None |
| select | method | Method used in the encoding. None: uses random\_scale. random: random ordering of levels, lex: lexicographical ordering by category level names, freq: frequency ordering from least frequent to most frequent, resp: response ordering. values: ['None', 'random', 'lex', 'freq', 'resp'] | freq |
| int | min\_support | Minimum number of levels required for a category to be represented on the ordinal scale. If a category level count is below the minimum, it will be grouped with other small cardinality levels or encoded as a missing value, depending on the value of other\_category. values: [1, 99999] | 5 |
| int | offset | Shift the ordinal scale of ordinal encoder values: [0, 99999] | 0 |
| bool | other\_category | If True, small cardinality values are mapped to a dedicated value (-1), otherwise they are encoded as missing values (-2). values: [False, True] | True |
| bool | random\_scale | Applies if method is None. If random\_scale is True, random ordering is used for the ordinal scale. If it is False, lexicographical ordering is used. values: [False, True] | True |
| int | seed | The RNG seed. values: [0, 99999] | 1234 |

5.2.2 Arbitrary or median value-based numeric imputation (V4 with memory usage optimization)

For a numeric feature, impute rows of missing values with an arbitrary (default: -9999) or median value. It also outputs the extra features (0, 1) indicating imputed rows.

This is effective for tree-based models, as they can learn a split between the arbitrary value (-9999) and the rest of the data (which ideally will not overlap this value). More advanced tree-based models usually use a method called “surrogate splits.” For models that don’t support this method, arbitrary-value imputation is a method that yields very similar results.

Imputation strategy:

A numeric feature is imputed with the arbitrary value (default: -9999) if it:

\* has enough rows with finite values (e.g., > t, default: 10).

\* has large number of rows with NaN (e.g., > min\_cna, default: 5).

\* is not configured as a monotonic-constrained feature.

Other numeric features will be imputed with the median value, if necessary. After imputation, the imputed numeric features will be scaled if the argument S is set to True.

Imputation indicator:

The indicator column (0, 1) is added to indicate imputed rows if the numeric feature:

\* is imputed with the median value.

\* has at least one row with nan.

\* has at least one unique value.

Example:

An imputation task is initialized with t=2 and min\_cna=2.

Input numeric features of this task:

feature0,feature1,feature2,feature3

1.0, 2.0, NaN, NaN,

2.0, 3.0, NaN, 18.0

3.0, 2.0, NaN, 16.0

4.0, 1.0, NaN, 14.0

5.0, 4.0, 2.0, 15.0

20.0, 1.0, 45.0, 46.0

Output numeric features of this task:

feature0, feature1, feature2, feature3, feature3-mi

1.0, 2.0, -9999.0, 16.0, 1.0

2.0, 3.0, -9999.0, 18.0, 0.0

3.0, 2.0, -9999.0, 16.0, 0.0

4.0, 1.0, -9999.0, 14.0, 0.0

5.0, 4.0, 2.0, 15.0, 0.0

20.0, 1.0, 45.0, 46.0, 0.0

\* Arbitrary value imputation is run on the feature 2 (num of nan rows >= min\_cna).

\* Median value imputation is run on the feature 3.

\* The feature3-mi is the indicator column for the imputation on feature 3.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| int | arbimp | Value to be used for imputing values: [-99999, 99999] | -9999 |
| int | min\_count\_na | Minimum number of missing values required for arbitrary imputation values: [0, 99999] | 5 |
| string | mono\_down | ID of the featurelist specifying the set of features to apply as monotonically decreasing in relation to the target | no |
| string | mono\_up | ID of the featurelist specifying the set of features to apply as monotonically increasing in relation to the target | no |
| bool | scale\_small | True if small values (the numeric variable is in a range of (0, 0.1]) are to be scaled values: [False, True] | True |
| int | threshold | Minimum number of finite elements required in a column for it to be considered for imputation (with arbitrary or median value) values: [0, 99999] | 10 |

5.2.3 Forecast Distance column extraction transformer

This transformer extracts the forecast distance column data from modeling data and returns that data as X.

5.2.4 Extracts Naive Predictions

Compared to BaselineExtractor, this handles multiple naive predictions.

This transformer extracts the Naive Prediction column data from modeling data. When the naive prediction is used for offset, the one with the longest periodicity will be selected. For non-stationary time series, it returns that data as a predictions\_to\_boost. For stationary time series, it returns that data as a feature. If output\_type is both, it returns that data as both offset and feature. The output is converted to the link function scale of the main modeler.

5.2.5 Extreme Gradient Boosting Regressor with Grid Search and Early Stopping support

Gradient Boosting Machines:

Gradient Boosting Machines (or Generalized Boosted Models, depending on who you ask to explain the acronym ‘GBM’) are an advanced algorithm for fitting extremely accurate predictive models. GBMs have won a number of recent predictive modeling competitions and are considered by many data scientists to be the most versatile and useful predictive modeling algorithm. GBMs require very little preprocessing, elegantly handle missing data, strike a good balance between bias and variance, and are typically able to find complicated interaction terms, making them a useful “Swiss army knife” of predictive models.

GBMs are a generalization of Freund and Schapire’s adaboost algorithm (1995) that handles arbitrary loss functions. They are very similar in concept to random forests, in that they fit individual decision trees to random re-samples of input data, where each tree sees a bootstrap sample of the rows of the dataset and N arbitrarily chosen columns, where N is a configurable parameter of the model. GBMs differ from random forests in a single major aspect: rather than fitting the trees independently, the GBM fits each successive tree to the residual errors from all the previous trees combined. This is advantageous, as the model focuses each iteration on the examples that are most difficult to predict (and therefore most useful to get correct).

Due to their iterative nature, GBMs are almost guaranteed to overfit the training data, given enough iterations. Therefore, the 2 critical parameters of the algorithm are the learning rate (or how fast the model fits the data) and the number of trees the model is allowed to fit. It is critical to tune one of these 2 parameters, and when done correctly, GBMs are capable of finding the exact point in the training data where overfitting begins, and halt one iteration prior to that point. In this manner GBMs are usually capable of squeezing every last bit of information out of the training set and producing a model with the highest possible accuracy without overfitting.

Extreme Gradient Boosting (XGBoost) is a very efficient, parallel version of GBM that has won a large number of Kaggle competitions. The base algorithm is very similar to GBM in R or in Python, but it has been heavily optimized and tweaked for faster runtimes and higher predictive accuracy.

Loss Function:

The XGBoost regressor uses least-squares loss by default, but can also use: tweedie loss for zero-inflated positive distributions, poisson loss for count problems, and gamma loss for right skewed positive distributions.

Grid Search Support:

Grid search is supported in this task. During training, grid search is run to estimate the optimal XGBoost parameter values that yield the best performance (evaluated by the configured loss function ). The grid search runs on a 70/30 train/test split within the training data; the estimated score uses 30% of the training data split. After the grid search completes and best tuning parameters are found, the final model is retrained on 100% of training data. Validation scores of the final model are different from the validation scores of the grid search.

Grid search is run on the task parameter with one of the following types: ‘intgrid’, ‘floatgrid’, ‘listgrid(int)’, ‘listgrid(float)’, ‘selectgrid’, or ‘multi’. Refer to the Parameters section for details of task parameter definitions.

For each grid search parameter, the search space is defined by the parameter values. Refer to the Parameters section for details of task parameter definitions.

Early Stopping Support:

The Early Stopping Extreme Gradient Boosting model also uses early stopping to determine the best number of trees. Early stopping is a method for determining the number of trees to use for a boosted trees model. The training data is split into a training set and a test set, and at each iteration the model is scored on the test set. If test set performance decreases for 200 iterations (tunable in Advanced Tuning), the training procedure stops and the model returns the fit from the best tree seen so far. The approach saves time by not continuing past the point where it is clear that the model is overfitting and further trees will not result in more accuracy.

Note that the early stopping test set uses a 90/10 train/test split within the training data for a given model. For example, a 64% model on the Leaderboard will internally use 57.6% of the data for training, and 6.4% of the data for early stopping. A 100% model on the leaderboard will internally use 90% of the data for training and 10% of the data for early stopping.

Since the early stopping test set was used for early stopping, it cannot be used for training.

This limitation also applies to grid search: within the grid search train/test split, the model will use a 90/10 train/test split for early stopping.

Please note: DataRobot DOES NOT use the early stoping logic from open-source XGBoost (also known as early\_stopping\_rounds). DataRobot instead uses custom early stopping logic with 2 parameters: smooth\_interval and interval. DataRobot does a moving average of the loss over the last smooth\_interval iterations, and then stops after interval rounds of increasing loss.

To replicate behavior similar to open-source XGBoost, you can set smooth\_interval=2 and then interval=desired\_early\_stopping\_rounds. This will still apply some smoothing to the loss function, resulting in it stopping a little bit later than open-source XGBoost.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| select | base\_margin\_initialize | If True, the intercept is initialized to the log odds of the target. values: [False, True] | False |
| floatgrid | colsample\_bylevel | Subsample of the features before each split in a tree. values: [0.1,1] | 1.0 |
| floatgrid | colsample\_bytree | Subsample ratio of columns when constructing each tree. By default, the value of colsample\_bytree for XGBoost classes is 1.0. However, based on the training data, DataRobot may choose a different initial value for this parameter. values: [0,1] | 0.2 |
| int | interval | Interval for early stopping. Once the model has hit “smooth\_interval” iterations, the early stopping logic checks for errors increasing at a rate of “interval” iterations in a row. For example, if smooth\_interval=200 and interval=10, XGBoost will run at least 200 iterations, and will early stop if the moving average of loss for the last 200 iterations increases for 10 iterations in a row. Higher values will make XGBoost run longer. values: [2, 500] | 10 |
| floatgrid | learning\_rate | Shrinks the contribution of each tree by learning\_rate. There is a trade-off between learning\_rate (lr) and n\_estimators(n). values: [5e-4,1] | 0.05 |
| bool | link\_transform | Whether the predictions from the previous step need to be link transformed at the link function scale before being used as an offset. values: [False, True] | False |
| select | loss | Loss function to be optimized. ‘ls’ refers to least squares regression. values: ['ls', 'labs', 'poisson', 'tweedie', 'gamma'] | ls |
| int | max\_bin | Used when tree\_method is set to ‘hist’. Maximum number of discrete bins to bucket continuous features. Increasing this number improves the optimality of splits at the cost of higher computation time. values: [16, 2048] | 256 |
| floatgrid | max\_delta\_step | Maximum delta step allowed for each tree’s weight estimation. If the value is set to 0, there is no constraint. Setting to a positive value makes the update step more conservative. Usually this parameter is not needed, but it might help in logistic regression when class is extremely imbalanced. Setting it to a value of 1-10 might help control the delta step update. values: [0,100] | 0.0 |
| intgrid | max\_depth | Maximum depth of the individual regression estimators. The maximum depth limits the number of nodes in the tree. Tune this parameter for optimal performance; the best value depends on the interaction of the input variables. The deeper the tree, the more variable interactions the model can capture. For frozen models that have larger sample sizes than the parent model, the max\_depth value is increased to retain similar accuracy. values: [1, 16] | 3 |
| floatgrid | min\_child\_weight | Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min\_child\_weight, the building process will give up further partitioning. In linear regression mode, this simply corresponds to the minimum number of instances needed to be in each node. The larger the value, the more conservative the algorithm will be. values: [0.01,float(1e5)] | 1.0 |
| floatgrid | min\_split\_loss | Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger the value, the more conservative the algorithm will be. values: [0,1e5] | 0.01 |
| float | missing\_value | Float value that should be treated as a missing value. When mono\_up or mono\_down is set, missing value will be set to -9999.0. values: [float(-1e5),float(1e5)] | -9999.0 |
| string | mono\_down | ID of the featurelist that defines the set of features with a monotonically decreasing relationship to the target. | no |
| string | mono\_up | ID of the featurelist that defines the set of features with a monotonically increasing relationship to the target. | no |
| int | n\_estimators | Number of boosting stages to perform. Gradient boosting is fairly robust to overfitting, so a larger number usually results in better performance. values: [1,20000] | 160 |
| intgrid | num\_parallel\_tree | Number of parallel trees created in each boosting stage. When this value is greater than 1, the model becomes a gradient-boosted random forest with (num\_parallel\_tree \* n\_estimators) trees. values: [1,16] | 1 |
| intgrid | random\_state | Seed used in the random number generator 'values': [0, int(1e9)] | 1234 |
| multi | reg\_alpha | L1 regularization term on weights; increasing this value will make the model more conservative. values: {'floatgrid': [0, 1e6], 'select': ['auto']} | 0.0 |
| multi | reg\_lambda | L2 regularization term on weights. Increasing this value will make the model more conservative. values: {'floatgrid': [0, 1e6], 'select': ['auto']} | 1.0 |
| float | scale\_pos\_weight | Scaling factor for examples in the positive class. values: [0,float(1e9)] | 1.0 |
| int | smooth\_interval | Moving average interval for early stopping. Loss over the last n intervals is averaged for determining early stopping. For example, if smooth\_interval=200, XGBoost will not stop early until it hits at least 200 iterations. After 200 iterations, it will use a moving average of loss over the last 200 iterations to determine early stopping. This helps remove noise in the loss function. Higher values will make XGBoost run longer. values: [2, 1000] | 200 |
| floatgrid | subsample | Subsample ratio of the training instance. Setting it to 0.5 means that XGBoost randomly collected half of the data instances to grow trees, which will prevent overfitting. | 0.333333333333 |
| select | tree\_method | Tree construction algorithm to use. ‘auto’: Heuristic to choose the faster algortihm. For small to medium datasets (<4M rows), exact greedy will be used. For large datasets (>=4M rows), approximate algorithm will be used. ‘exact’:Exact greedy algorithm. ‘approx’:Approximate greedy algorithm using sketching and histogram. ‘hist’:Fast histogram-optimized approximate greedy algorithm. It uses some performance improvements (e.g., bin caching). values: ['auto', 'exact', 'approx', 'hist'] | auto |
| float | tweedie\_p | Power parameter of the tweedie distribution. Only applicable when Tweedie loss is used. values: [1, 2] | 1.5 |

5.2.6 Forecast Distance modeler using ElasticNet modelers for each value of forecast distance

Parameters are passed as a single string and then separated into common parameters and parameters for a specific forecast distance. This allows for the modeler to be be re-tuned using different parameters for different forecast distances.

The target, target-derived features, and naive predictions are scaled by the average target value for each series prior to model fitting. This may allow the model to make better use of features across series with different scales.

ElasticNet is a linear regression model trained with L1 and L2 prior as regularizers. This combination allows for learning a sparse model where few of the weights are non-zero like Lasso, while still maintaining the regularization properties of Ridge.

The Model is named either Elastic-Net, Ridge or Lasso Regressor depending on the value of the alpha parameter (see: enet\_alpha).

ElasticNet is useful when there are multiple correlated features. While Lasso is likely to pick one feature at random, ElasticNet is likely to pick both.

A practical advantage of Ridge over Lasso is that it allows ElasticNet to inherit some of Ridge’s stability under rotation. ElasticNet allows for response variables that have error distribution models other than a normal distribution (poisson, gamma, tweedie).

Based on lightning CDRegressor.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| select | beta\_transform | Parameter used for blenders. If beta\_transform is set to ‘blender’, coefficients are non-negative and are all in the range [0, 1]. A very large weight of the penalty term will result in the average blender. values: ['id', 'blender'] | id |
| multi | enet\_alpha | ElasticNet mixing parameter, where 0 <= alpha <= 1. If alpha = 0, the penalty is an L2 penalty. If alpha = 1, the penalty is an L1 penalty. For 0 < alpha < 1, the penalty is a combination of L1 and L2. Results form an auto grid of 11 values, evenly spaced from 0.0 to 1.0. values: {{'floatgrid': [0, 1], 'select': ['auto']}} | 0.0 |
| multi | enet\_lambda | Weight for the penalty term. Results form an auto grid of 20 numbers, spaced evenly on a log10 scale from 3.16e-7 to 3.16e-0.1. Note that when fit\_alpha\_scaler=True (the default), enet\_lambda=0 will be the full model and enet\_lambda=1 will be the intercept-only model (enet\_lambda>1 will also be an intercept-only model). For example, when fit\_alpha\_scaler=True, enet\_lambda=.1 will result in a lightly penalized model; enet\_lambda=.9 will be a heavily penalized model. Using fit\_alpha\_scaler=True and 0 <= enet\_lambda <= 1 is highly recommended. values: {{'floatgrid':[1e-10,1e10],'select':['auto']}} | auto |
| bool | fit\_alpha\_scaler | It is highly recommended that you use fit\_alpha\_scaler=True (the default). Setting the value to False will result in unexpected behavior for some models. Also note that fit\_alpha\_scaler affects the enet\_lambda parameter. When fit\_alpha\_scale=True, enet\_lambda=0 will be the full model with no penalty; enet\_lambda=1 will be the intercept-only model (enet\_lambda>1 will also be an intercept-only model). This a useful parameterization, as it makes it easy to control the model’s penalty. For example, enet\_lambda=.1 will be a lightly-penalized mode; enet\_lambda=.9 will be a heavily penalized model. Ideally, use fit\_alpha\_scaler=True and 0 <= enet\_lambda <= 1. values: [False, True] | True |
| bool | fit\_intercept | Sets whether to calculate the intercept for this model. If set to false, no intercept will be used in calculations (i.e., data is expected to be centered). values: [False, True] | False |
| bool | fit\_tweedie\_p | If it is set to True, the tweedie power parameter is automatically set. values: [False, True] | False |
| bool | link\_transform | Sets whether predictions from the previous step require a link transform at the link function scale before being used as an offset. values: [False, True] | False |
| select | loss | Loss function to use. values: ['squared', 'poisson', 'gamma', 'tweedie'] | squared |
| int | max\_iter | Maximum number of iterations. values: [1, 1e6] | 100 |
| int | random\_state | The seed of the pseudo random number generator to use. values: [0, int(1e9)] | 1234 |
| float | tol | The tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol. values: [1e-10, 1e10] | 0.0001 |
| float | tweedie\_p | Tweedie power to pass to ElasticNet model. values: [1.0, 2.0] | 1.5 |

5.3 Literature Review and References

* [1] Feelders, Ad. “Handling missing data in trees: Surrogate splits or statistical imputation?” Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 1999. 329-334. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.36.7991&rep=rep1&type=pdf
* Chen, T, and He, T. Higgs Boson Discovery with Boosted Trees.” Cowan et al., editor, JMLR: Workshop and Conference Proceedings. No. 42. 2015. http://proceedings.mlr.press/v42/chen14.pdf
* Freund, Yoav, and Robert E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting.” Journal of computer and system sciences 55.1 (1997): 119-139. https://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic\_generalization.pdf
* Friedman, Jerome H. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of statistics (2001): 1189-1232. https://statweb.stanford.edu/~jhf/ftp/trebst.pdf
* Hastie, T., Tibshirani, R., and Friedman, J.. “Elements of Statistical Learning.” Springer, 2009. http://statweb.stanford.edu/~tibs/ElemStatLearn/
* Breiman, Leo. Arcing the edge. Technical Report 486, Statistics Department, University of California at Berkeley, 1997. https://statistics.berkeley.edu/sites/default/files/tech-reports/486.pdf
* Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent.” Journal of statistical software 33.1 (2010)

5.4 Alternative Model Frameworks and Theories Considered

As stated by regulatory guidance, comparison with alternative theories and approaches provides guidance for final model selection and is a fundamental component of a sound modeling process.

DataRobot develops dozens of alternative models, exposes the details of how these models were built and how they perform, and enables the user to select the best model for the particular business problem being addressed.

During the model development process, DataRobot considered the following alternative models. The final model was selected based on model performance as well as an analysis of model diagnostics and expert business judgment.

The performance metric used for this project was RMSE. The model types considered during the model selection process included the following models, which are sorted by the Holdout score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Backtest Score | All Backtests Score | Holdout Score | Training Length |
| Zero-Inflated Light Gradient Boosted Trees Regressor with Early Stopping (Gamma Loss) | 31311.1555 | 46561.6016 | 35893.5819 | 8 years 3 months 1 day |
| Ridge Regressor using Linearly Decaying Weights with Forecast Distance Modeling | 35619.0649 | N/A | 36774.4506 | 8 years 3 months 1 day |
| eXtreme Gradient Boosting on ElasticNet Predictions | 35572.3811 | N/A | 36976.0938 | 8 years 3 months 1 day |
| Ridge Regressor with Forecast Distance Modeling | 36801.2432 | N/A | 37041.5286 | 8 years 3 months 1 day |
| eXtreme Gradient Boosting on ElasticNet Predictions (learning rate =0.3) | 36413.3826 | N/A | 37437.0476 | 8 years 3 months 1 day |
| Performance Clustered eXtreme Gradient Boosting on Elastic Net Predictions | 36130.9812 | N/A | 37487.1637 | 8 years 3 months 1 day |
| Zero-Inflated eXtreme Gradient Boosted Trees Regressor with Early Stopping (Poisson Loss) | 34482.2217 | 46316.5434 | 37563.6034 | 8 years 3 months 1 day |
| Temporal Hierarchical Model with Elastic Net and XGBoost | 33176.703 | 45323.2518 | 37726.1687 | 8 years 3 months 1 day |
| eXtreme Gradient Boosted Trees Regressor with Early Stopping (learning rate =0.3) | 34505.1707 | 45150.0887 | 38402.8888 | 8 years 3 months 1 day |
| Eureqa Generalized Additive Model (250 Generations) | 35543.2064 | N/A | 38971.9329 | 8 years 3 months 1 day |
| Performance Clustered eXtreme Gradient Boosted Trees Regressor | 34926.0602 | 48114.218 | 41523.8708 | 8 years 3 months 1 day |
| Eureqa Regressor (Quick Search: 250 Generations) | 41369.8667 | N/A | 41913.8786 | 8 years 3 months 1 day |
| Performance Clustered Elastic Net Regressor with Forecast Distance Modeling | 36730.8371 | N/A | 43532.3351 | 8 years 3 months 1 day |
| Ridge Regressor with Forecast Distance Modeling and Series Scaling | 37401.2289 | N/A | 44881.1515 | 8 years 3 months 1 day |
| Light Gradient Boosting on ElasticNet Predictions (learning rate =0.3) | 36129.7397 | N/A | 46170.5327 | 8 years 3 months 1 day |
| Light Gradient Boosting on ElasticNet Predictions | 36394.7203 | N/A | 46545.1434 | 8 years 3 months 1 day |
| Per Series Elastic Net Regressor with Forecast Distance Modeling | 63895.6503 | N/A | 49481.485 | 8 years 3 months 1 day |
| Baseline Predictions Using Most Recent Value | 42366.0956 | N/A | 50546.8708 | 8 years 3 months 1 day |
| Keras Slim Residual Neural Network Regressor using Training Schedule (1 Layer: 64 Units) | 57095.6189 | N/A | 56752.547 | 8 years 3 months 1 day |

5.5 Variable Selection

The model's variable selection process includes a balance of quantitative analysis and key domain knowledge about the underlying business problem (i.e., expert judgment). The subsections below describe:

* DataRobot Quantitative Analysis: Key components related to variable selection that are automated by DataRobot.
* DataRobot automatically creates and selects time series features in the modeling data and will automatically detect whether or not a project's target value is stationary (that is, whether the statistical properties of the target are constant over time). If the target is not stationary, DataRobot attempts to make it stationary by applying a differencing strategy prior to modeling. This improves the accuracy and robustness of the underlying models.
* Expert Judgment and Variable Selection: Summary of the expert judgment used during the variable selection process.
* Final Model Variables: Final feature list chosen.

5.5.1 DataRobot Quantitative Analysis

A feature list is a defined set of features (variables) that DataRobot can use for modeling. DataRobot automatically creates three feature lists (described below) for each project. Users, however, can create customized feature lists that contain a subset of the total feature set, and use the new list to train new, alternative models. The default feature lists are described below:

* Informative Features (default): Features that pass a "reasonableness" check that determines whether they contain useful information. For example, DataRobot excludes features it determines are low information, such as a column containing all ones, duplicate columns, or a feature with too few values. The Informative Features list is sorted by each feature's correlation with the target variable.
* Raw Features: All features (variables) in the dataset, including those excluded from the Informative Features list.
* Univariate Selection: Features that meet a certain threshold for non-linear correlation with the selected target. DataRobot calculates, for each entry in the Informative Features list, the feature's individual relationship against the target.

Users also have the option to create user-defined feature transformations, which can then be included in a feature list for model exploration and to determine relative feature importance. Importance is measured using the information content of the variable; the calculation is done independently for each feature in the dataset. Features are then ranked on the Project Data from most to least important. This score represents a measure of predictive power using only that variable to predict the target. The score is measured using the project's accuracy metric that is defined by either the user (e.g., RMSE) or the default assigned by DataRobot.

5.5.2 Expert Judgement and Variable Selection

This section should include additional detail regarding the variable selection process and any expert judgment used during feature selection.

5.5.3 Final Model Variables

Below are two tables. The first contains a list of the final set of model feature variables, as well as summary statistics for the eXtreme Gradient Boosted Trees Regressor with Early Stopping model. The second table contains a detailed analysis of missing values.

The Model Features and Summary Statistics table provides a brief overview of the summary statistics of model features. This includes Feature Name, variable type (Var Type), number of unique values (Unique), Number of missing values (Missing), Mean, Standard Deviation (Std Dev), Median, Minimum Value (Min), Maximum Value (Max) and Assessment of target leakage risk (Target Leakage).

5.5.3.1 Model Features and Summary Statistics

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature Name | Var Type | Unique | Missing | Mean | Std Dev | Median | Min | Max | Target Leakage |
| %key\_company | Categorical | 3 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| platform no\_ | Categorical | 19 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| sales date | Date | 193 | 0 | 2019-04-28 | 1313.44 days | 2019-09-01 | 2008-11-01 | 2024-11-01 | N/A |
| plant country | Categorical | 3 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| %key\_DataRobot | Categorical | 22 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| customer group name | Categorical | 12 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| total net eur | Numeric | 2048 | 0 | 82575.89 | 76647.095 | 72848.22 | 0.0 | 426607.04 | N/A |
| OEM | Categorical | 4 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| First sales date | Date | 19 | 0 | 2014-07-29 | 1099.82 days | 2015-04-01 | 2008-11-01 | 2020-05-01 | N/A |
| last sales date | Date | 169 | 304 | 2019-10-18 | 1134.64 days | 2020-01-01 | 2008-11-01 | 2024-10-01 | N/A |
| revenue last 12 months | Numeric | 2027 | 304 | 1066469.3 | 777489.13 | 978008.71 | 0.0 | 4036453.0 | N/A |
| plant country unemployment rate | Numeric | 80 | 0 | 4.39 | 1.63 | 3.7 | 2.9 | 12.0 | N/A |
| plant country registration | Numeric | 280 | 390 | 176058.87 | 130107.404 | 241330.0 | 5.72 | 427111.0 | N/A |
| plant country Freight | Numeric | 178 | 515 | 8081.209 | 11837.067 | 27.02 | 17.62 | 30410.0 | N/A |
| plant country fuel | Numeric | 22 | 2245 | 0.48 | 0.051 | 0.49 | 0.26 | 0.58 | N/A |
| plant country fatalities | Numeric | 167 | 408 | 198.97 | 117.407 | 230.0 | 0.16 | 440.0 | N/A |
| plant country Order books | Numeric | 238 | 0 | -5.34 | 18.85 | -6.0 | -80.0 | 29.8 | N/A |
| plant country Business situation | Numeric | 300 | 0 | -7.93 | 22.24 | -7.1 | -67.0 | 52.8 | N/A |
| plant country Finished good stocks | Numeric | 183 | 0 | 5.31 | 9.44 | 4.9 | -15.3 | 30.0 | N/A |
| plant country selling prices | Numeric | 224 | 0 | 12.59 | 17.83 | 6.4 | -25.0 | 71.9 | N/A |
| plant country Demand evolution | Numeric | 230 | 0 | 15.9 | 10.56 | 16.7 | -61.3 | 37.7 | N/A |
| plant country Production | Numeric | 269 | 0 | 7.65 | 12.86 | 7.8 | -78.8 | 42.2 | N/A |
| plant country household savings rate | Numeric | 91 | 390 | 18.24 | 3.15 | 17.89 | 5.3 | 27.33 | N/A |
| plant country economic situation | Numeric | 281 | 0 | -20.99 | 13.57 | -19.2 | -72.7 | 15.5 | N/A |
| plant country consumer prices | Numeric | 296 | 0 | 30.903 | 12.75 | 30.2 | -16.9 | 81.9 | N/A |
| plant country consumer price index | Numeric | 378 | 16 | 2.66 | 2.94 | 1.708 | -1.0409 | 15.404 | N/A |
| plant country short\_term interest rate | Numeric | 193 | 0 | 0.54 | 1.51 | -0.308 | -0.58 | 4.24 | N/A |
| total production volume last 12 months | Numeric | 910 | 1312 | 150867.39 | 80479.46 | 155791.0 | 0.0 | 429794.0 | N/A |
| production volume \_9 months | Numeric | 856 | 1339 | 13018.42 | 8155.506 | 13255.0 | 0.0 | 44548.0 | N/A |
| production volume +3 months | Numeric | 816 | 1339 | 12917.47 | 8163.64 | 13587.0 | 0.0 | 59294.0 | N/A |
| production volume \_11 months | Numeric | 857 | 1339 | 13070.88 | 8155.63 | 13329.0 | 0.0 | 44548.0 | N/A |
| production volume +13 months | Numeric | 756 | 1339 | 12167.92 | 8968.16 | 12933.0 | 0.0 | 47143.0 | N/A |
| production volume +1 months | Numeric | 822 | 1339 | 12421.78 | 7767.97 | 13029.0 | 0.0 | 63467.0 | N/A |
| production volume +12 months | Numeric | 765 | 1339 | 12342.64 | 8972.87 | 13031.0 | 0.0 | 51192.0 | N/A |
| production volume +0 months | Numeric | 822 | 1339 | 12218.019 | 7593.82 | 12724.0 | 0.0 | 56935.0 | N/A |
| production volume \_12 months | Numeric | 856 | 1339 | 13149.85 | 8204.38 | 13358.0 | 0.0 | 44548.0 | N/A |
| production volume \_13 months | Numeric | 854 | 1339 | 13033.3 | 8209.36 | 13317.0 | 0.0 | 44548.0 | N/A |
| production volume +11 months | Numeric | 775 | 1339 | 12403.6 | 8934.76 | 12931.0 | 0.0 | 51459.0 | N/A |
| production volume \_1 months | Numeric | 832 | 1339 | 12836.65 | 8021.23 | 12895.0 | 0.0 | 49599.0 | N/A |
| production volume +5 months | Numeric | 807 | 1339 | 13047.92 | 8562.15 | 13514.0 | 0.0 | 58770.0 | N/A |
| production volume +4 months | Numeric | 818 | 1339 | 12938.43 | 8327.28 | 13467.0 | 0.0 | 59023.0 | N/A |
| production volume \_8 months | Numeric | 853 | 1339 | 12938.44 | 8141.083 | 13245.0 | 0.0 | 44548.0 | N/A |
| production volume +2 months | Numeric | 827 | 1339 | 12758.24 | 7927.606 | 13350.0 | 0.0 | 63094.0 | N/A |
| production volume \_10 months | Numeric | 857 | 1339 | 13016.44 | 8134.15 | 13294.0 | 0.0 | 44548.0 | N/A |
| production volume +10 months | Numeric | 786 | 1339 | 12496.17 | 8917.31 | 13034.0 | 0.0 | 51265.0 | N/A |
| production volume \_2 months | Numeric | 839 | 1339 | 12625.33 | 7974.8 | 13027.0 | 0.0 | 37722.0 | N/A |
| production volume \_4 months | Numeric | 847 | 1339 | 12740.67 | 7982.58 | 13245.0 | 0.0 | 37722.0 | N/A |
| production volume \_3 months | Numeric | 842 | 1339 | 12648.0 | 7972.54 | 13115.0 | 0.0 | 37722.0 | N/A |
| production volume +9 months | Numeric | 789 | 1339 | 12623.25 | 8867.6 | 13003.0 | 0.0 | 54488.0 | N/A |
| production volume +8 months | Numeric | 788 | 1339 | 12779.12 | 8865.54 | 13186.0 | 0.0 | 55640.0 | N/A |
| production volume \_5 months | Numeric | 847 | 1339 | 12877.97 | 8030.42 | 13317.0 | 0.0 | 37722.0 | N/A |
| production volume +7 months | Numeric | 799 | 1339 | 12929.83 | 8813.0 | 13400.0 | 0.0 | 53467.0 | N/A |
| production volume \_6 months | Numeric | 849 | 1339 | 13010.42 | 8108.89 | 13329.0 | 0.0 | 40520.0 | N/A |
| production volume +6 months | Numeric | 803 | 1339 | 13090.11 | 8754.53 | 13517.0 | 0.0 | 53069.0 | N/A |
| production volume \_7 months | Numeric | 850 | 1339 | 13028.203 | 8146.47 | 13324.0 | 0.0 | 41782.0 | N/A |
| design lead country | Categorical | 2 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| last actual | Date | 2 | 0 | 2024-01-09 | 0.48 days | 2024-01-09 | 2024-01-09 | 2024-01-10 | N/A |
| next facelift | Date | 16 | 0 | 2021-03-04 | 985.61 days | 2021-01-04 | 2016-01-05 | 2027-01-01 | N/A |
| global status | Categorical | 2 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| global body type | Categorical | 2 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| global size | Categorical | 4 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| global data plant | Categorical | 17 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| global platform | Categorical | 9 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| global production model | Categorical | 19 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| global make | Categorical | 7 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| manufacturer | Categorical | 5 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| sales group | Categorical | 4 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| Global Data country | Categorical | 10 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| trade region | Categorical | 3 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| local model line | Categorical | 17 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| Number months until series project EOP | Numeric | 228 | 0 | 69.105 | 43.14 | 69.0 | -45.0 | 182.0 | N/A |
| Number months since series project SOP | Numeric | 203 | 0 | 35.53 | 40.17 | 35.0 | -71.0 | 131.0 | N/A |
| Number months until Global EOP | Numeric | 229 | 0 | 68.85 | 43.74 | 68.0 | -44.0 | 184.0 | N/A |
| Number months since Global SOP | Numeric | 252 | 0 | -49.43 | 47.93 | -50.0 | -184.0 | 67.0 | N/A |
| Number months until next facelift | Numeric | 220 | 0 | 26.23 | 43.205 | 27.0 | -80.0 | 139.0 | N/A |
| Number months since last actual | Numeric | 193 | 0 | -64.49 | 43.061 | -60.0 | -190.0 | 2.0 | N/A |
| Number months since first sale | Numeric | 187 | 0 | 56.95 | 38.12 | 53.0 | 0.0 | 186.0 | N/A |
| Number months since last sale | Numeric | 15 | 304 | 1.14 | 1.68 | 1.0 | 1.0 | 48.0 | N/A |
| First sale | Categorical | 2 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| number months with revenue | Numeric | 148 | 0 | 43.63 | 34.39 | 40.0 | 0.0 | 147.0 | N/A |
| accumulated revenue | Numeric | 2031 | 0 | 4142673.66 | 4637265.54 | 3088061.12 | 0.0 | 24869691.99 | N/A |
| car set price since 2020 | Numeric | 947 | 1382 | 9.77 | 12.58 | 6.97 | 0.0 | 127.85 | N/A |
| car set price last 12 months | Numeric | 929 | 1405 | 11.96 | 86.35 | 7.0903 | 0.25 | 2620.207 | N/A |
| covid | Categorical | 2 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| revenue forecast +0 months | Numeric | 878 | 1405 | 88978.93 | 53016.4002 | 83001.19 | 0.0 | 528229.85 | N/A |
| revenue forecast +1 months | Numeric | 874 | 1405 | 90322.25 | 60993.49 | 82571.13 | 0.0 | 842110.94 | N/A |
| revenue forecast +2 months | Numeric | 871 | 1405 | 92266.011 | 72850.51 | 83519.74 | 0.0 | 1526942.13 | N/A |
| revenue forecast +3 months | Numeric | 865 | 1405 | 92785.71 | 73345.88 | 84668.77 | 0.0 | 1542370.11 | N/A |
| revenue forecast +4 months | Numeric | 859 | 1405 | 92977.87 | 83946.59 | 85368.35 | 0.0 | 1978532.15 | N/A |
| revenue forecast +5 months | Numeric | 853 | 1405 | 93060.34 | 89164.83 | 85242.24 | 0.0 | 2162489.47 | N/A |
| revenue forecast +6 months | Numeric | 846 | 1405 | 92448.24 | 84718.26 | 85331.39 | 0.0 | 1998995.66 | N/A |
| revenue forecast +7 months | Numeric | 838 | 1405 | 91163.33 | 85299.0059 | 83724.86 | 0.0 | 1975746.55 | N/A |
| revenue forecast +8 months | Numeric | 830 | 1405 | 89747.78 | 84605.98 | 83922.41 | 0.0 | 1908784.8 | N/A |
| revenue forecast +9 months | Numeric | 822 | 1405 | 88412.41 | 83946.92 | 82352.63 | 0.0 | 1845680.046 | N/A |
| revenue forecast +10 months | Numeric | 815 | 1405 | 87551.49 | 95053.85 | 81800.46 | 0.0 | 2328554.602 | N/A |
| revenue forecast +11 months | Numeric | 807 | 1405 | 86387.81 | 93044.54 | 81502.15 | 0.0 | 2255164.53 | N/A |
| revenue forecast +12 months | Numeric | 799 | 1405 | 85823.31 | 94869.21 | 81031.38 | 0.0 | 2297377.21 | N/A |
| revenue forecast +13 months | Numeric | 790 | 1405 | 83558.68 | 83201.54 | 79950.47 | 0.0 | 1826716.48 | N/A |

The last column in this table is an assessment of target leakage risk. DataRobot automatically tests for target leakage on a per-feature basis during the Autopilot process. Target leakage, sometimes called data leakage, occurs when a model is trained using a dataset that includes information that would not be available at the time of prediction. This can produce overly optimistic model performance results during training, given a feature will near-completely describe the target (e.g., the number of late payments on a loan as a predictor for loan default at loan application date.)

DataRobot tests for target leakage risk using Alternating Conditional Expectation (ACE) to measure the association between each feature and the target; the ACE score is normalized using the project optimization metric so that its value is in the range [0,1]. If above a certain threshold (see below), DataRobot will create a new feature list with those features flagged and possibly removed, and the user is notified by a banner in the user interface during modeling. Notably, because the definition of target leakage is directly tied with prediction time and not strength of association between a feature and the target, it's possible for DataRobot to not identify all sources of target leakage. Therefore, to reduce the risk for potential target leakage in the feature list, it's important to apply subject matter expertise.

The thresholds for target leakage risk are based on a normalized ACE score:

* High risk: > 0.975, flagged and removed
* Moderate risk: > 0.85, flagged but not removed
* Low risk: < 0.85, no action

The following table provides a summary of missing values. It includes the name of the feature, its type, a summary of the missing value count (both number of rows and as a percentage), and information on the type of imputation applied to the feature.

5.5.3.2 Data Quality Handling Report

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Name | Var Type | Missing Count | Missing Percentage | Imputation Name | Imputation Description |
| plant country fuel (12 month std) | Numeric | 5022 | 96 | Missing Values Imputed | Imputed value: -9999 |
| plant country fuel (24 month std) | Numeric | 5022 | 96 | Missing Values Imputed | Imputed value: -9999 |
| plant country fuel (6 month std) | Numeric | 5022 | 96 | Missing Values Imputed | Imputed value: -9999 |
| plant country fuel (12 month mean) | Numeric | 5019 | 96 | Missing Values Imputed | Imputed value: -9999 |
| plant country fuel (1st lag) | Numeric | 5019 | 96 | Missing Values Imputed | Imputed value: -9999 |
| plant country fuel (24 month mean) | Numeric | 5019 | 96 | Missing Values Imputed | Imputed value: -9999 |
| plant country fuel (6 month mean) | Numeric | 5019 | 96 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (12 month std) | Numeric | 3330 | 64 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (24 month std) | Numeric | 3330 | 64 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (6 month std) | Numeric | 3330 | 64 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (1st lag) | Numeric | 3300 | 63 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (12 month mean) | Numeric | 3273 | 63 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (24 month mean) | Numeric | 3273 | 63 | Missing Values Imputed | Imputed value: -9999 |
| car set price since 2020 (6 month mean) | Numeric | 3273 | 63 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (6 month std) | Numeric | 3264 | 62 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (1st lag) | Numeric | 3258 | 62 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (12 month std) | Numeric | 3249 | 62 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (24 month std) | Numeric | 3234 | 62 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (12 month std) | Numeric | 3210 | 61 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (24 month std) | Numeric | 3210 | 61 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (6 month std) | Numeric | 3210 | 61 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (6 month mean) | Numeric | 3180 | 61 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (12 month mean) | Numeric | 3165 | 61 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (6 month std) | Numeric | 3156 | 60 | Missing Values Imputed | Imputed value: -9999 |
| car set price last 12 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +0 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +1 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +10 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +11 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +12 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +13 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +2 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +3 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +4 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +5 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +6 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +7 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +8 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| revenue forecast +9 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (12 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (1st lag) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (24 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| total production volume last 12 months (6 month mean) | Numeric | 3153 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (12 month std) | Numeric | 3141 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (24 month std) | Numeric | 3129 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (1st lag) | Numeric | 3117 | 60 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (6 month mean) | Numeric | 3096 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (12 month mean) | Numeric | 3081 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +0 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +1 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +10 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +11 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +12 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +13 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +2 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +3 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +4 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +5 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +6 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +7 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +8 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume +9 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_1 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_10 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_11 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_12 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_13 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_2 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_3 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_4 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_5 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_6 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_7 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_8 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| production volume \_9 months (24 month mean) | Numeric | 3072 | 59 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (12 month std) | Numeric | 1272 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (24 month std) | Numeric | 1272 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (6 month std) | Numeric | 1272 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (12 month mean) | Numeric | 1243 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (1st lag) | Numeric | 1243 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (24 month mean) | Numeric | 1243 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country Freight (6 month mean) | Numeric | 1243 | 24 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (12 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (24 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (6 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (12 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (24 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (6 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (12 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (24 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (6 month std) | Numeric | 1065 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (12 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (1st lag) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (24 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country fatalities (6 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (12 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (1st lag) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (24 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country household savings rate (6 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (12 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (1st lag) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (24 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| plant country registration (6 month mean) | Numeric | 1033 | 20 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Month) (24 month most\_frequent) | Categorical | 975 | 19 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Day of Week) (24 month most\_frequent) | Categorical | 765 | 15 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Month) (12 month most\_frequent) | Categorical | 738 | 14 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Day of Week) (12 month most\_frequent) | Categorical | 636 | 12 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Month) (6 month most\_frequent) | Categorical | 564 | 11 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Day of Week) (6 month most\_frequent) | Categorical | 558 | 11 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Day of Week) (1st lag) | Categorical | 487 | 9 | Ordinal encoding of categorical variables | Imputed value: -2 |
| last sales date (Month) (1st lag) | Categorical | 487 | 9 | Ordinal encoding of categorical variables | Imputed value: -2 |
| Number months since last sale (1st lag) | Numeric | 487 | 9 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Month) (1st lag) | Numeric | 487 | 9 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (1st lag) | Numeric | 487 | 9 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (1st lag) | Numeric | 487 | 9 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last sale (6 month std) | Numeric | 435 | 8 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (6 month std) | Numeric | 435 | 8 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (6 month std) | Numeric | 435 | 8 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last sale (12 month std) | Numeric | 375 | 7 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (12 month std) | Numeric | 375 | 7 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (12 month std) | Numeric | 375 | 7 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last sale (24 month std) | Numeric | 297 | 6 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (24 month std) | Numeric | 297 | 6 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (24 month std) | Numeric | 297 | 6 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last sale (6 month mean) | Numeric | 286 | 5 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (6 month mean) | Numeric | 286 | 5 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (6 month mean) | Numeric | 286 | 5 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (5th lag) (diff 24 month mean) | Numeric | 264 | 5 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last sale (12 month mean) | Numeric | 235 | 4 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (12 month mean) | Numeric | 235 | 4 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (12 month mean) | Numeric | 235 | 4 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (4th lag) (diff 24 month mean) | Numeric | 219 | 4 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last sale (24 month mean) | Numeric | 199 | 4 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Year) (24 month mean) | Numeric | 199 | 4 | Missing Values Imputed | Imputed value: -9999 |
| revenue last 12 months (24 month mean) | Numeric | 199 | 4 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (3rd lag) (diff 24 month mean) | Numeric | 174 | 3 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (12 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (24 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (6 month std) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (2nd lag) (diff 24 month mean) | Numeric | 129 | 2 | Missing Values Imputed | Imputed value: -9999 |
| %key\_company (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| First sale (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| First sales date (Day of Week) (12 month most\_frequent) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| First sales date (Month) (12 month most\_frequent) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| Global Data country (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| OEM (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| covid (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| customer group name (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| design lead country (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global body type (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global data plant (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global make (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global platform (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global production model (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global size (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| global status (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| local model line (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| manufacturer (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| next facelift (Day of Week) (12 month most\_frequent) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| plant country (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| platform no\_ (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| sales group (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| trade region (1st lag) | Categorical | 85 | 2 | Ordinal encoding of categorical variables | Imputed value: -2 |
| First sale (12 month fraction equal yes) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| First sale (24 month fraction equal yes) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| First sale (6 month fraction equal yes) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| First sales date (Day of Week) (12 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| First sales date (Year) (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since Global SOP (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since first sale (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since first sale (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since first sale (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since first sale (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last actual (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last actual (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last actual (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since last actual (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since series project SOP (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since series project SOP (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since series project SOP (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months since series project SOP (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until Global EOP (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until Global EOP (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until Global EOP (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until Global EOP (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until next facelift (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until next facelift (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until next facelift (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until next facelift (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until series project EOP (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until series project EOP (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until series project EOP (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| Number months until series project EOP (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| accumulated revenue (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| covid (12 month fraction equal yes) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| covid (24 month fraction equal yes) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| covid (6 month fraction equal yes) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last actual (months from sales date) (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Month) (12 month fraction equal 1\_0) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Month) (24 month fraction equal 1\_0) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Month) (6 month fraction equal 1\_0) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Week) (12 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Week) (24 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Day of Week) (6 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Month) (12 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Month) (24 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| last sales date (Month) (6 month entropy) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| next facelift (Day of Month) (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| next facelift (Year) (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| number months with revenue (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Business situation (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Demand evolution (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Finished good stocks (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Order books (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country Production (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer price index (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country consumer prices (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country economic situation (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country selling prices (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country short\_term interest rate (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (12 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (1st lag) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| plant country unemployment rate (6 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (12 month max) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (12 month mean) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (12 month median) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (12 month min) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (1st lag) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (24 month max) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (24 month median) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (24 month min) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (6 month max) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (6 month mean) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (6 month median) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| total net eur (6 month min) (diff 24 month mean) | Numeric | 85 | 2 | Missing Values Imputed | Imputed value: -9999 |
| sales date (Day of Week) (actual) | Categorical | 0 | 0 | Ordinal encoding of categorical variables | Imputed value: -2 |
| sales date (Month) (actual) | Categorical | 0 | 0 | Ordinal encoding of categorical variables | Imputed value: -2 |
| sales date (Year) (actual) | Numeric | 0 | 0 | Missing Values Imputed | Imputed value: 2019 |

6 Model Performance and Stability

6.1 Model Validation Stability

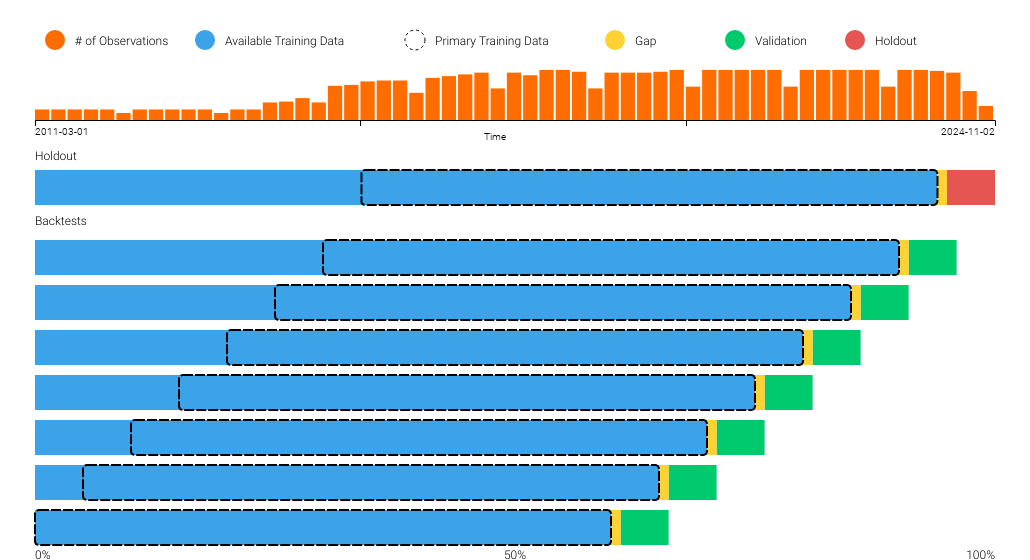
To find patterns in a dataset from which it can make predictions, an algorithm must first learn from a historical example – typically from a historical dataset that contains the output variable you want to predict. However, if a model is trained too closely on its training data then it may be overfit. Overfitting is a modeling error that occurs when a model is too closely fit to training data and therefore performs poorly on out-of-sample data (data that was not used to train the model). Overfitting generally results in an overly complex model that explains idiosyncrasies and random noise in the training data, rather than the underlying trends that the model was intended to capture. To avoid overfitting, the best practice is to evaluate model performance on out-of-sample data. If the model performs very well on in-sample data, (the training data) but poorly on out-of-sample data, that may be an indication that the model is overfit.

Out-of-time validation (OTV) allows the selection of specific time periods or durations to test the stability of your model rather than using random rows, creating “backtests” for your data. OTV ensures that there is no overfitting on prior periods of data on which the model was trained. Measuring backtest performance is a good way to assess model performance on new observations. In addition to the validation length partitioning (which is the length of time used for validation per each backtest), users can also adjust the gap length or the time between training and validation. In addition to the OTV partitioning, DataRobot uses a holdout sample to further test out-of-sample model performance and ensure the model is not overfit.

The following procedure was used during development to insure that overfitting did not occur:

* DataRobot used 7 backtests with a validation length of 8 months.
* A holdout fold with start date: 2024-03-02 and end date 2024-11-02 for additional testing. This dataset is used to verify that the final model performs well on data that has not been touched throughout the training process.

The following figure summarizes the CV process used by DataRobot, where the blue denotes available training data, green denotes the validation partition, and red denotes the holdout sample.



6.1.1 Data Partitioning Methodology

Time series projects, like OTV projects, use date/time partitioning, and all the workflow changes that apply to other date/time partitioned projects also apply to them. Unlike other projects, time series projects produce different types of models which forecast multiple future predictions instead of an individual prediction for each row.

DataRobot uses a general time series framework to configure how time series features are created and what future values the models will output. This framework consists of a Forecast Point (defining a time a prediction is being made), a Feature Derivation Window (a rolling window used to create features), and a Forecast Window (a rolling window of future values to predict).

Time series projects will automatically transform the dataset provided in order to apply this framework. During the transformation, DataRobot uses the Feature Derivation Window to derive time series features (such as lags and rolling statistics), and uses the Forecast Window to provide examples of forecasting different distances in the future (such as time shifts). After project creation, a new dataset and a new feature list are generated and used to train the models. This process is reapplied automatically at prediction time as well in order to generate future predictions based on the original data features.

6.2 Model Performance (Backtesting)

As an additional layer of model validity, DataRobot not only evaluated the statistical metrics underlying the model, but also performed testing on out-of-time records.

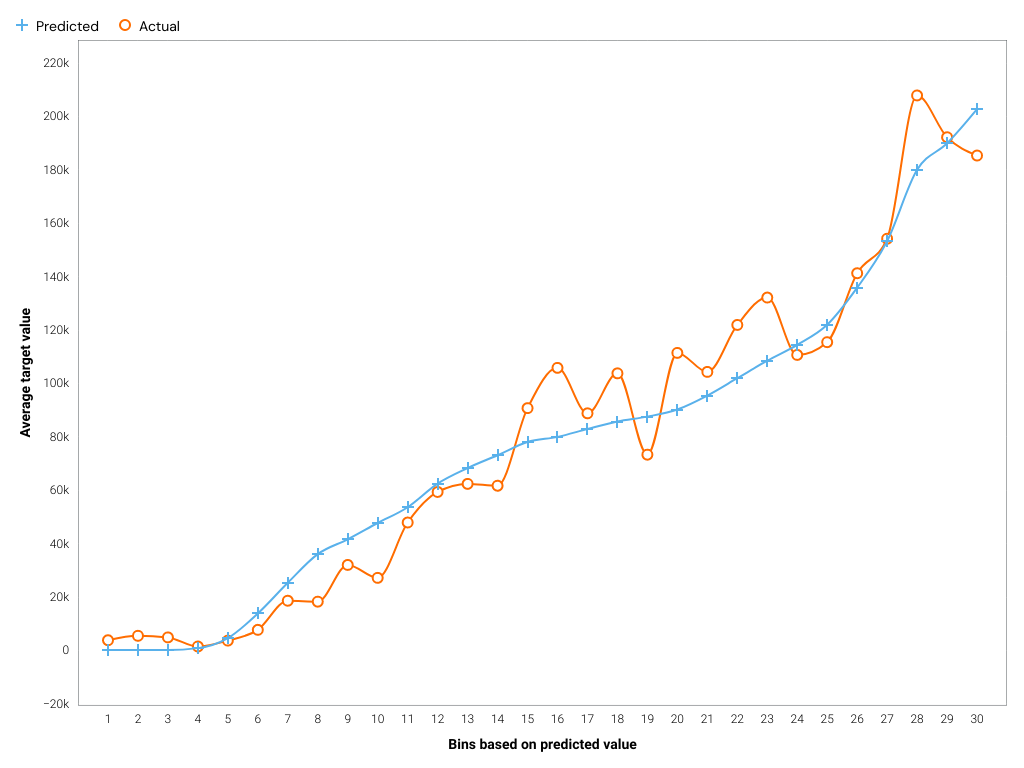
The performance metric used for this project was RMSE. The model performance results are presented below for out-of-time testing:

|  |  |
| --- | --- |
| Scoring Type | Score (RMSE) |
| backtesting\_scores | 33246.1907, 35644.1867, 45137.219, 47146.8822, 39967.097, 62028.7678, 33292.6536 |
| holdout | 37362.439 |
| validation | 33246.1907 |

6.3 Sensitivity Testing and Analysis

6.3.1 Lift Chart

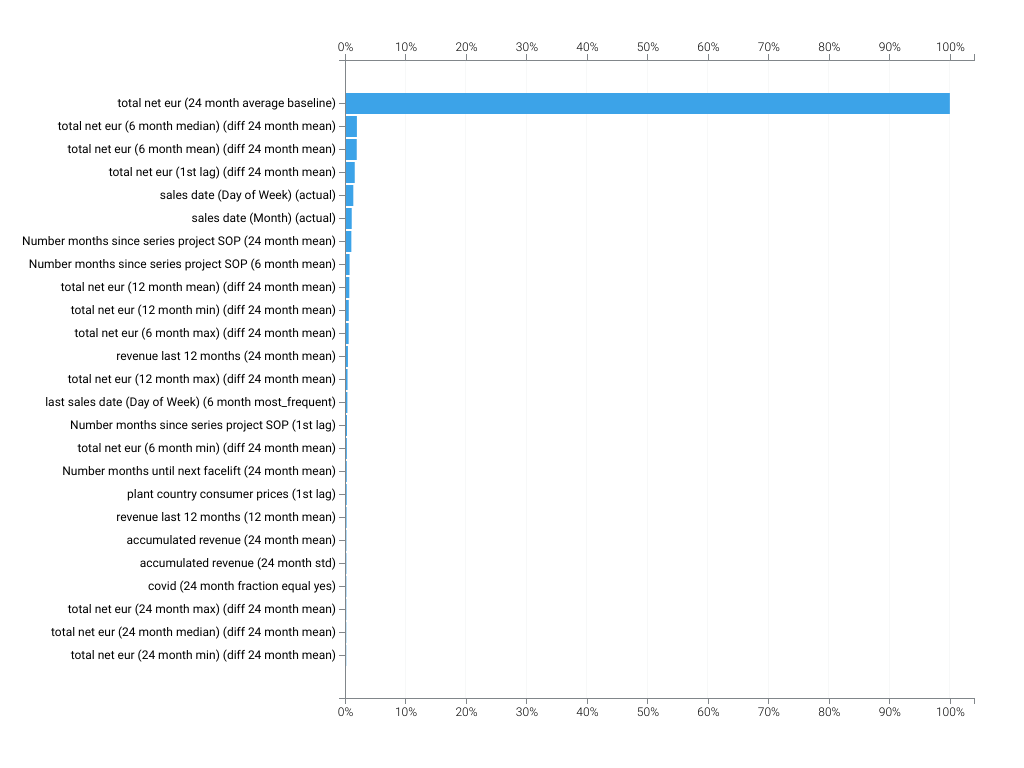
The Lift Chart sorts and groups numeric feature values into equal sized bins, depicting how well a model segments the target population and how capable it is of predicting the target. This helps the user to visualize model accuracy for each bin. The chart is sorted by predicted values -- lowest to highest predictions, for example -- which provides transparency to the model performance for different ranges of values of the target variable. Looking at the Lift Chart, the left side of the curve indicates where the model predicted a low score on one section of the population while the right side of the curve indicates where the model predicted a high score. The model Lift Chart is presented in the figure below.



The points on the Lift Chart indicate the average percentage in each bin. The "Predicted" blue line displays the average prediction score for the rows in that bin. The "Actual" orange line displays the actual percentage for the rows in that bin. In general, the steeper the Actual line is, and the more closely the Predicted line matches the actual line, the better the model. A close relationship between these two lines is indicative of the predictive accuracy of the model; a consistently increasing line is another good indicator of satisfactory model performance.

6.3.2 Key Relationships

Feature Impact, which is available for all model types, works by altering input data and observing the effect on a models score. This technique is sometimes called Permutation Importance. The Feature Impact for a given column measures how much worse a models error score would be if DataRobot made predictions after randomly shuffling that column (while leaving other columns unchanged). DataRobot normalizes the scores so that the value of the most important feature column is first and the other subsequent features are normalized to it.

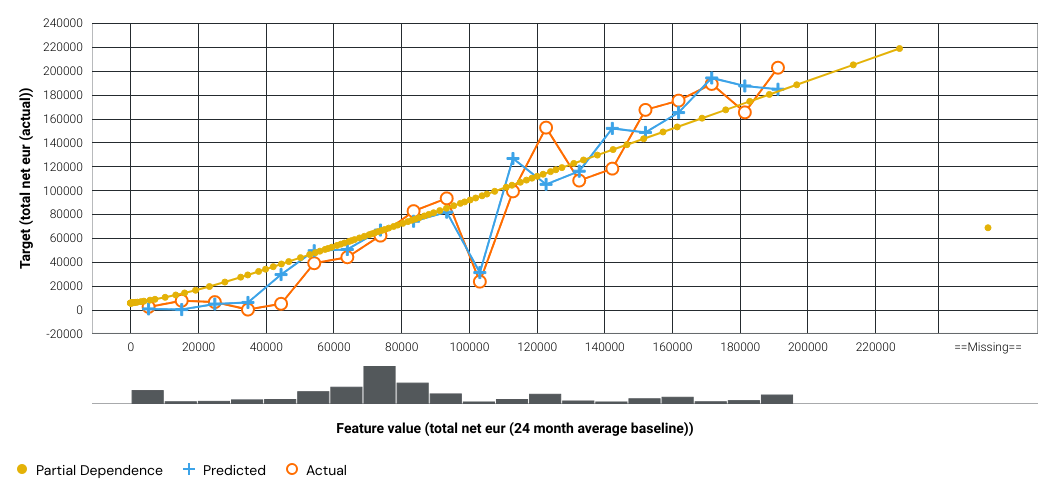


|  |  |  |
| --- | --- | --- |
| Feature Name | Impact Normalized | Impact Unnormalized |
| total net eur (24 month average baseline) | 1.0 | 63590.7331 |
| total net eur (6 month median) (diff 24 month mean) | 0.0197 | 1250.3695 |
| total net eur (6 month mean) (diff 24 month mean) | 0.0194 | 1235.3485 |
| total net eur (1st lag) (diff 24 month mean) | 0.0161 | 1023.7286 |
| sales date (Day of Week) (actual) | 0.0137 | 872.8904 |
| sales date (Month) (actual) | 0.0111 | 707.3604 |
| Number months since series project SOP (24 month mean) | 0.0105 | 666.6289 |
| Number months since series project SOP (6 month mean) | 0.0076 | 480.1196 |
| total net eur (12 month mean) (diff 24 month mean) | 0.0071 | 453.6782 |
| total net eur (12 month min) (diff 24 month mean) | 0.0063 | 397.4575 |
| total net eur (6 month max) (diff 24 month mean) | 0.006 | 382.6713 |
| revenue last 12 months (24 month mean) | 0.0048 | 306.8785 |
| total net eur (12 month max) (diff 24 month mean) | 0.0042 | 266.6721 |
| last sales date (Day of Week) (6 month most\_frequent) | 0.0039 | 250.5617 |
| Number months since series project SOP (1st lag) | 0.0033 | 208.6229 |
| total net eur (6 month min) (diff 24 month mean) | 0.003 | 192.824 |
| Number months until next facelift (24 month mean) | 0.0029 | 183.2537 |
| plant country consumer prices (1st lag) | 0.0028 | 177.6858 |
| revenue last 12 months (12 month mean) | 0.0025 | 161.776 |
| accumulated revenue (24 month mean) | 0.0023 | 146.6473 |
| accumulated revenue (24 month std) | 0.0023 | 146.3885 |
| covid (24 month fraction equal yes) | 0.0022 | 141.6139 |
| total net eur (24 month max) (diff 24 month mean) | 0.0021 | 135.3629 |
| total net eur (24 month median) (diff 24 month mean) | 0.0021 | 134.0102 |
| total net eur (24 month min) (diff 24 month mean) | 0.0021 | 130.5605 |
| total net eur (12 month median) (diff 24 month mean) | 0.002 | 129.8163 |
| total net eur (2nd lag) (diff 24 month mean) | 0.002 | 128.7006 |
| revenue last 12 months (6 month mean) | 0.0018 | 116.717 |
| revenue last 12 months (6 month std) | 0.0017 | 108.7973 |
| Number months until next facelift (1st lag) | 0.0017 | 108.2209 |
| Number months until series project EOP (12 month mean) | 0.0016 | 100.6023 |
| accumulated revenue (1st lag) | 0.0016 | 99.4451 |
| number months with revenue (24 month mean) | 0.0016 | 98.915 |
| plant country household savings rate (6 month std) | 0.0015 | 94.9941 |
| revenue last 12 months (12 month std) | 0.0015 | 94.846 |
| Number months until series project EOP (24 month mean) | 0.0014 | 91.7181 |
| plant country Demand evolution (24 month mean) | 0.0013 | 85.0959 |
| plant country Order books (12 month std) | 0.0013 | 82.2318 |
| plant country Order books (24 month std) | 0.0012 | 77.9085 |
| plant country Order books (6 month std) | 0.0012 | 77.7804 |
| plant country selling prices (24 month mean) | 0.0012 | 77.2155 |
| total net eur (3rd lag) (diff 24 month mean) | 0.0011 | 73.0858 |
| plant country fatalities (6 month mean) | 0.0011 | 70.5861 |
| plant country household savings rate (24 month std) | 0.0011 | 68.84 |
| revenue last 12 months (24 month std) | 0.001 | 66.1967 |
| Number months until next facelift (12 month mean) | 0.001 | 62.5137 |
| Number months since Global SOP (12 month mean) | 0.001 | 61.9542 |
| customer group name (1st lag) | 0.001 | 60.8846 |
| Number months since Global SOP (6 month mean) | 0.001 | 60.4482 |
| plant country Production (24 month mean) | 0.0009 | 59.8002 |
| revenue last 12 months (1st lag) | 0.0009 | 59.6575 |
| plant country Business situation (12 month std) | 0.0009 | 54.7687 |
| plant country economic situation (24 month std) | 0.0008 | 54.0029 |
| Number months until Global EOP (24 month mean) | 0.0008 | 52.0716 |
| Number months since first sale (12 month mean) | 0.0008 | 50.303 |
| last sales date (Day of Week) (24 month entropy) | 0.0008 | 49.9137 |
| plant country economic situation (24 month mean) | 0.0008 | 49.7495 |
| Number months until next facelift (6 month mean) | 0.0008 | 49.4119 |
| Number months since first sale (6 month mean) | 0.0007 | 47.278 |
| plant country short\_term interest rate (6 month std) | 0.0007 | 45.4969 |
| plant country Freight (1st lag) | 0.0007 | 44.9464 |
| production volume +6 months (6 month std) | 0.0007 | 44.755 |
| number months with revenue (12 month mean) | 0.0007 | 44.3981 |
| Number months since last sale (24 month mean) | 0.0007 | 43.6786 |
| accumulated revenue (6 month std) | 0.0007 | 43.5647 |
| plant country unemployment rate (6 month mean) | 0.0007 | 43.3932 |
| Number months since last sale (6 month std) | 0.0007 | 42.7237 |
| plant country short\_term interest rate (6 month mean) | 0.0007 | 41.822 |
| production volume \_4 months (12 month mean) | 0.0007 | 41.6737 |
| accumulated revenue (12 month std) | 0.0007 | 41.508 |
| plant country fatalities (6 month std) | 0.0006 | 40.1224 |
| plant country unemployment rate (1st lag) | 0.0006 | 39.884 |
| plant country consumer prices (24 month std) | 0.0006 | 39.5454 |
| production volume +13 months (12 month std) | 0.0006 | 39.1712 |
| Number months since Global SOP (1st lag) | 0.0006 | 38.8361 |
| plant country Business situation (12 month mean) | 0.0006 | 38.6092 |
| plant country Finished good stocks (24 month std) | 0.0006 | 38.4986 |
| plant country Production (12 month mean) | 0.0006 | 38.3378 |
| number months with revenue (1st lag) | 0.0006 | 38.0005 |
| plant country Demand evolution (24 month std) | 0.0006 | 37.848 |
| plant country Production (6 month std) | 0.0006 | 36.8758 |
| Number months since Global SOP (24 month mean) | 0.0006 | 36.7479 |
| revenue forecast +5 months (12 month std) | 0.0006 | 36.4079 |
| plant country Demand evolution (1st lag) | 0.0006 | 36.2952 |
| accumulated revenue (6 month mean) | 0.0006 | 36.1609 |
| Number months until Global EOP (6 month mean) | 0.0006 | 36.1507 |
| production volume +5 months (6 month std) | 0.0006 | 35.4038 |
| plant country Order books (6 month mean) | 0.0005 | 34.613 |
| plant country unemployment rate (24 month mean) | 0.0005 | 34.1899 |
| Number months since first sale (24 month mean) | 0.0005 | 33.9368 |
| plant country unemployment rate (12 month std) | 0.0005 | 33.9007 |
| total net eur (5th lag) (diff 24 month mean) | 0.0005 | 33.8792 |
| Number months until Global EOP (1st lag) | 0.0005 | 32.5265 |
| revenue forecast +9 months (1st lag) | 0.0005 | 32.2357 |
| production volume \_1 months (24 month mean) | 0.0005 | 32.0276 |
| plant country registration (12 month std) | 0.0005 | 31.2659 |
| plant country consumer prices (6 month mean) | 0.0005 | 30.2141 |
| First sales date (Day of Week) (12 month most\_frequent) | 0.0005 | 30.0509 |
| Number months since last actual (1st lag) | 0.0005 | 30.0105 |
| Number months until series project EOP (1st lag) | 0.0005 | 29.945 |
| sales date (Year) (actual) | 0.0005 | 29.794 |
| plant country fatalities (24 month mean) | 0.0005 | 29.3142 |
| plant country short\_term interest rate (12 month mean) | 0.0004 | 28.5235 |
| production volume \_4 months (1st lag) | 0.0004 | 28.1452 |
| plant country Finished good stocks (12 month std) | 0.0004 | 27.6501 |
| last sales date (Year) (12 month mean) | 0.0004 | 27.2267 |
| total net eur (4th lag) (diff 24 month mean) | 0.0004 | 26.5714 |
| last sales date (Day of Month) (24 month fraction equal 1\_0) | 0.0004 | 26.4934 |
| Number months since series project SOP (12 month mean) | 0.0004 | 26.2512 |
| plant country consumer price index (12 month std) | 0.0004 | 26.1473 |
| Number months until Global EOP (12 month mean) | 0.0004 | 24.5484 |
| plant country Freight (6 month mean) | 0.0004 | 24.2939 |
| revenue forecast +13 months (6 month std) | 0.0004 | 24.273 |
| production volume +1 months (1st lag) | 0.0004 | 23.8242 |
| plant country consumer price index (24 month mean) | 0.0004 | 23.726 |
| production volume +13 months (1st lag) | 0.0004 | 23.6736 |
| revenue forecast +6 months (12 month std) | 0.0004 | 23.5211 |
| plant country Order books (24 month mean) | 0.0004 | 23.3756 |
| plant country fatalities (12 month mean) | 0.0004 | 23.2826 |
| production volume \_2 months (12 month mean) | 0.0004 | 23.241 |
| production volume +11 months (24 month mean) | 0.0004 | 22.8334 |
| production volume +12 months (12 month std) | 0.0004 | 22.666 |
| production volume +3 months (6 month mean) | 0.0003 | 21.8786 |
| Number months since first sale (1st lag) | 0.0003 | 21.7064 |
| car set price last 12 months (6 month std) | 0.0003 | 21.6009 |
| plant country household savings rate (12 month mean) | 0.0003 | 21.4603 |
| revenue forecast +0 months (12 month mean) | 0.0003 | 19.8968 |
| plant country Production (1st lag) | 0.0003 | 19.8706 |
| plant country registration (6 month std) | 0.0003 | 19.7534 |
| plant country household savings rate (12 month std) | 0.0003 | 19.2217 |
| production volume \_4 months (24 month mean) | 0.0003 | 19.1528 |
| production volume \_4 months (6 month mean) | 0.0003 | 19.1001 |
| last sales date (Year) (12 month std) | 0.0003 | 18.3444 |
| production volume \_9 months (6 month mean) | 0.0003 | 18.13 |
| production volume +0 months (12 month mean) | 0.0003 | 17.578 |
| First sale (12 month fraction equal yes) | 0.0003 | 17.5702 |
| plant country selling prices (24 month std) | 0.0003 | 17.0251 |
| revenue forecast +2 months (1st lag) | 0.0003 | 16.8891 |
| plant country Finished good stocks (24 month mean) | 0.0003 | 16.626 |
| plant country consumer price index (12 month mean) | 0.0003 | 16.5436 |
| Number months until series project EOP (6 month mean) | 0.0003 | 16.5208 |
| production volume +0 months (12 month std) | 0.0003 | 16.3724 |
| plant country consumer prices (24 month mean) | 0.0003 | 16.253 |
| sales group (1st lag) | 0.0002 | 15.7705 |
| Number months since last sale (12 month mean) | 0.0002 | 15.3303 |
| production volume +10 months (1st lag) | 0.0002 | 15.2243 |
| plant country Demand evolution (12 month mean) | 0.0002 | 15.1936 |
| plant country economic situation (12 month std) | 0.0002 | 14.864 |
| production volume \_11 months (12 month mean) | 0.0002 | 14.7809 |
| production volume +9 months (6 month std) | 0.0002 | 14.6242 |
| trade region (1st lag) | 0.0002 | 14.5535 |
| revenue forecast +13 months (12 month mean) | 0.0002 | 14.505 |
| last actual (months from sales date) (1st lag) | 0.0002 | 14.3915 |
| plant country consumer prices (6 month std) | 0.0002 | 14.127 |
| plant country Order books (1st lag) | 0.0002 | 13.9898 |
| car set price since 2020 (6 month mean) | 0.0002 | 13.8569 |
| plant country registration (6 month mean) | 0.0002 | 13.5398 |
| car set price since 2020 (6 month std) | 0.0002 | 13.317 |
| accumulated revenue (12 month mean) | 0.0002 | 13.1551 |
| plant country Business situation (1st lag) | 0.0002 | 13.1029 |
| production volume +11 months (12 month mean) | 0.0002 | 13.0658 |
| next facelift (Day of Week) (12 month most\_frequent) | 0.0002 | 12.3628 |
| plant country short\_term interest rate (1st lag) | 0.0002 | 12.354 |
| plant country household savings rate (1st lag) | 0.0002 | 12.2735 |
| plant country consumer price index (1st lag) | 0.0002 | 11.7238 |
| plant country household savings rate (6 month mean) | 0.0002 | 11.7084 |
| covid (12 month fraction equal yes) | 0.0002 | 11.5438 |
| production volume +6 months (12 month mean) | 0.0002 | 10.9147 |
| revenue forecast +10 months (1st lag) | 0.0002 | 10.7176 |
| production volume +11 months (6 month mean) | 0.0002 | 10.4256 |
| production volume +5 months (24 month std) | 0.0002 | 10.4119 |
| plant country Freight (12 month mean) | 0.0002 | 10.3014 |
| number months with revenue (12 month std) | 0.0002 | 10.1504 |
| revenue forecast +0 months (24 month std) | 0.0002 | 10.1043 |
| production volume +13 months (6 month std) | 0.0002 | 10.075 |
| production volume +10 months (12 month mean) | 0.0002 | 10.0448 |
| Number months since last sale (6 month mean) | 0.0002 | 9.9184 |
| revenue forecast +12 months (6 month std) | 0.0002 | 9.8799 |
| production volume \_7 months (24 month std) | 0.0002 | 9.8541 |
| plant country consumer price index (6 month std) | 0.0002 | 9.8514 |
| Number months since last actual (12 month mean) | 0.0002 | 9.8409 |
| car set price last 12 months (12 month mean) | 0.0002 | 9.7145 |
| plant country fatalities (12 month std) | 0.0002 | 9.6987 |
| production volume +12 months (12 month mean) | 0.0002 | 9.6798 |
| Number months since last sale (12 month std) | 0.0002 | 9.6226 |
| production volume +8 months (12 month std) | 0.0001 | 9.4858 |
| revenue forecast +2 months (24 month std) | 0.0001 | 9.2356 |
| global make (1st lag) | 0.0001 | 8.9373 |
| plant country economic situation (6 month mean) | 0.0001 | 8.509 |
| production volume +2 months (24 month mean) | 0.0001 | 8.4618 |
| production volume \_1 months (6 month std) | 0.0001 | 8.402 |
| production volume +1 months (12 month std) | 0.0001 | 8.36 |
| production volume +8 months (6 month mean) | 0.0001 | 8.3047 |
| plant country economic situation (6 month std) | 0.0001 | 8.2854 |
| production volume +8 months (6 month std) | 0.0001 | 8.2696 |
| last sales date (Day of Week) (12 month most\_frequent) | 0.0001 | 8.1873 |
| last sales date (Year) (24 month mean) | 0.0001 | 8.1763 |
| production volume +8 months (24 month mean) | 0.0001 | 8.1397 |
| plant country economic situation (12 month mean) | 0.0001 | 8.1234 |
| production volume +2 months (6 month mean) | 0.0001 | 8.0953 |
| production volume \_4 months (12 month std) | 0.0001 | 7.8586 |
| revenue forecast +10 months (6 month std) | 0.0001 | 7.6349 |
| last sales date (Month) (1st lag) | 0.0001 | 7.5459 |
| plant country Freight (12 month std) | 0.0001 | 7.4988 |
| production volume \_6 months (6 month mean) | 0.0001 | 7.4815 |
| production volume \_5 months (1st lag) | 0.0001 | 7.3225 |
| revenue forecast +8 months (12 month mean) | 0.0001 | 7.2405 |
| revenue forecast +0 months (12 month std) | 0.0001 | 7.2321 |
| revenue forecast +3 months (1st lag) | 0.0001 | 7.2272 |
| plant country Finished good stocks (6 month std) | 0.0001 | 7.0641 |
| production volume +4 months (6 month std) | 0.0001 | 7.0087 |
| revenue forecast +6 months (6 month std) | 0.0001 | 6.8594 |
| revenue forecast +3 months (24 month mean) | 0.0001 | 6.7887 |
| plant country consumer price index (24 month std) | 0.0001 | 6.6513 |
| revenue forecast +0 months (1st lag) | 0.0001 | 6.3527 |
| production volume +6 months (12 month std) | 0.0001 | 6.3087 |
| plant country Demand evolution (12 month std) | 0.0001 | 6.2836 |
| plant country Demand evolution (6 month mean) | 0.0001 | 6.217 |
| revenue forecast +7 months (24 month mean) | 0.0001 | 5.9257 |
| revenue forecast +7 months (24 month std) | 0.0001 | 5.9145 |
| Global Data country (1st lag) | 0.0001 | 5.9112 |
| production volume \_7 months (6 month mean) | 0.0001 | 5.8148 |
| plant country short\_term interest rate (12 month std) | 0.0001 | 5.7006 |
| plant country registration (12 month mean) | 0.0001 | 5.5298 |
| revenue forecast +13 months (12 month std) | 0.0001 | 5.5134 |
| revenue forecast +8 months (6 month std) | 0.0001 | 5.4505 |
| production volume +11 months (1st lag) | 0.0001 | 5.4076 |
| First sales date (Year) (12 month mean) | 0.0001 | 5.3113 |
| production volume \_2 months (6 month mean) | 0.0001 | 5.2722 |
| production volume +13 months (12 month mean) | 0.0001 | 5.2503 |
| revenue forecast +8 months (12 month std) | 0.0001 | 5.1109 |
| plant country consumer price index (6 month mean) | 0.0001 | 5.0901 |
| car set price since 2020 (24 month std) | 0.0001 | 4.8844 |
| global production model (1st lag) | 0.0001 | 4.8115 |
| production volume +0 months (1st lag) | 0.0001 | 4.6601 |
| revenue forecast +5 months (24 month std) | 0.0001 | 4.6245 |
| production volume +10 months (12 month std) | 0.0001 | 4.2329 |
| global platform (1st lag) | 0.0001 | 4.1593 |
| production volume +13 months (24 month mean) | 0.0001 | 4.0352 |
| production volume \_1 months (24 month std) | 0.0001 | 3.9782 |
| production volume +2 months (1st lag) | 0.0001 | 3.8863 |
| revenue forecast +3 months (12 month mean) | 0.0001 | 3.8335 |
| plant country selling prices (12 month mean) | 0.0001 | 3.7905 |
| production volume +8 months (24 month std) | 0.0001 | 3.7816 |
| production volume \_11 months (6 month mean) | 0.0001 | 3.7785 |
| production volume +6 months (24 month mean) | 0.0001 | 3.7434 |
| First sale (6 month fraction equal yes) | 0.0001 | 3.6764 |
| production volume \_3 months (6 month mean) | 0.0001 | 3.5202 |
| revenue forecast +1 months (1st lag) | 0.0001 | 3.4813 |
| plant country registration (1st lag) | 0.0001 | 3.4802 |
| number months with revenue (6 month std) | 0.0001 | 3.4233 |
| production volume +11 months (6 month std) | 0.0001 | 3.381 |
| production volume \_9 months (24 month mean) | 0.0 | 3.095 |
| production volume +0 months (6 month std) | 0.0 | 3.0605 |
| plant country unemployment rate (24 month std) | 0.0 | 2.987 |
| plant country consumer prices (12 month mean) | 0.0 | 2.9371 |
| First sale (24 month fraction equal yes) | 0.0 | 2.9356 |
| production volume \_3 months (1st lag) | 0.0 | 2.8582 |
| revenue forecast +5 months (6 month std) | 0.0 | 2.7353 |
| production volume \_8 months (1st lag) | 0.0 | 2.7102 |
| number months with revenue (6 month mean) | 0.0 | 2.6866 |
| car set price since 2020 (12 month std) | 0.0 | 2.585 |
| production volume +10 months (6 month std) | 0.0 | 2.5274 |
| production volume \_1 months (1st lag) | 0.0 | 2.5125 |
| revenue forecast +8 months (24 month std) | 0.0 | 2.4357 |
| production volume +10 months (24 month std) | 0.0 | 2.3082 |
| production volume +9 months (24 month std) | 0.0 | 2.2963 |
| production volume +9 months (1st lag) | 0.0 | 2.2727 |
| revenue forecast +8 months (1st lag) | 0.0 | 2.211 |
| production volume +5 months (12 month mean) | 0.0 | 2.0787 |
| plant country Production (6 month mean) | 0.0 | 2.0186 |
| production volume +2 months (6 month std) | 0.0 | 1.9909 |
| car set price last 12 months (6 month mean) | 0.0 | 1.9813 |
| production volume +11 months (24 month std) | 0.0 | 1.9753 |
| revenue forecast +3 months (24 month std) | 0.0 | 1.799 |
| production volume +7 months (6 month mean) | 0.0 | 1.7457 |
| last sales date (Day of Week) (12 month entropy) | 0.0 | 1.5222 |
| production volume +7 months (24 month mean) | 0.0 | 1.5165 |
| revenue forecast +7 months (12 month mean) | 0.0 | 1.5017 |
| production volume \_9 months (1st lag) | 0.0 | 1.4933 |
| revenue forecast +10 months (6 month mean) | 0.0 | 1.4549 |
| revenue forecast +2 months (24 month mean) | 0.0 | 1.3378 |
| plant country selling prices (6 month std) | 0.0 | 1.1115 |
| last sales date (Month) (12 month most\_frequent) | 0.0 | 1.0308 |
| revenue forecast +2 months (6 month mean) | 0.0 | 0.8372 |
| plant country Freight (24 month mean) | 0.0 | 0.8356 |
| production volume +7 months (12 month std) | 0.0 | 0.7551 |
| plant country Business situation (6 month mean) | 0.0 | 0.728 |
| car set price since 2020 (12 month mean) | 0.0 | 0.6893 |
| plant country selling prices (12 month std) | 0.0 | 0.5604 |
| last sales date (Day of Week) (1st lag) | 0.0 | 0.2836 |
| last sales date (Day of Week) (24 month most\_frequent) | 0.0 | 0.1903 |
| revenue forecast +2 months (12 month std) | 0.0 | 0.0968 |
| production volume \_13 months (6 month mean) | 0.0 | 0.0948 |
| %key\_DataRobot (actual) | 0.0 | 0.0 |
| %key\_company (1st lag) | 0.0 | 0.0 |
| First sale (1st lag) | 0.0 | 0.0 |
| First sales date (Day of Week) (12 month entropy) | 0.0 | 0.0 |
| Number months since Global SOP (12 month std) | 0.0 | 0.0 |
| Number months since Global SOP (24 month std) | 0.0 | 0.0 |
| Number months since Global SOP (6 month std) | 0.0 | 0.0 |
| Number months since last actual (24 month mean) | 0.0 | 0.0 |
| Number months since last actual (6 month mean) | 0.0 | 0.0 |
| Number months since last sale (24 month std) | 0.0 | 0.0 |
| car set price last 12 months (24 month mean) | 0.0 | 0.0 |
| car set price since 2020 (1st lag) | 0.0 | 0.0 |
| car set price since 2020 (24 month mean) | 0.0 | 0.0 |
| covid (1st lag) | 0.0 | 0.0 |
| covid (6 month fraction equal yes) | 0.0 | 0.0 |
| design lead country (1st lag) | 0.0 | 0.0 |
| global data plant (1st lag) | 0.0 | 0.0 |
| global size (1st lag) | 0.0 | 0.0 |
| global status (1st lag) | 0.0 | 0.0 |
| last sales date (Day of Month) (12 month fraction equal 1\_0) | 0.0 | 0.0 |
| last sales date (Day of Month) (1st lag) | 0.0 | 0.0 |
| last sales date (Day of Month) (6 month fraction equal 1\_0) | 0.0 | 0.0 |
| last sales date (Day of Week) (6 month entropy) | 0.0 | 0.0 |
| last sales date (Month) (12 month entropy) | 0.0 | 0.0 |
| last sales date (Month) (24 month entropy) | 0.0 | 0.0 |
| last sales date (Month) (24 month most\_frequent) | 0.0 | 0.0 |
| last sales date (Month) (6 month entropy) | 0.0 | 0.0 |
| last sales date (Month) (6 month most\_frequent) | 0.0 | 0.0 |
| last sales date (Year) (1st lag) | 0.0 | 0.0 |
| last sales date (Year) (6 month mean) | 0.0 | 0.0 |
| last sales date (Year) (6 month std) | 0.0 | 0.0 |
| local model line (1st lag) | 0.0 | 0.0 |
| manufacturer (1st lag) | 0.0 | 0.0 |
| next facelift (Day of Month) (12 month mean) | 0.0 | 0.0 |
| next facelift (Year) (12 month mean) | 0.0 | 0.0 |
| number months with revenue (24 month std) | 0.0 | 0.0 |
| plant country (1st lag) | 0.0 | 0.0 |
| plant country Business situation (24 month mean) | 0.0 | 0.0 |
| plant country Business situation (24 month std) | 0.0 | 0.0 |
| plant country Demand evolution (6 month std) | 0.0 | 0.0 |
| plant country Finished good stocks (12 month mean) | 0.0 | 0.0 |
| plant country Finished good stocks (1st lag) | 0.0 | 0.0 |
| plant country Finished good stocks (6 month mean) | 0.0 | 0.0 |
| plant country Freight (24 month std) | 0.0 | 0.0 |
| plant country Order books (12 month mean) | 0.0 | 0.0 |
| plant country Production (24 month std) | 0.0 | 0.0 |
| plant country consumer prices (12 month std) | 0.0 | 0.0 |
| plant country economic situation (1st lag) | 0.0 | 0.0 |
| plant country fatalities (1st lag) | 0.0 | 0.0 |
| plant country fatalities (24 month std) | 0.0 | 0.0 |
| plant country fuel (12 month mean) | 0.0 | 0.0 |
| plant country fuel (12 month std) | 0.0 | 0.0 |
| plant country fuel (1st lag) | 0.0 | 0.0 |
| plant country fuel (24 month mean) | 0.0 | 0.0 |
| plant country fuel (24 month std) | 0.0 | 0.0 |
| plant country fuel (6 month mean) | 0.0 | 0.0 |
| plant country fuel (6 month std) | 0.0 | 0.0 |
| plant country household savings rate (24 month mean) | 0.0 | 0.0 |
| plant country registration (24 month mean) | 0.0 | 0.0 |
| plant country registration (24 month std) | 0.0 | 0.0 |
| plant country selling prices (6 month mean) | 0.0 | 0.0 |
| plant country short\_term interest rate (24 month std) | 0.0 | 0.0 |
| plant country unemployment rate (6 month std) | 0.0 | 0.0 |
| production volume +0 months (24 month mean) | 0.0 | 0.0 |
| production volume +0 months (6 month mean) | 0.0 | 0.0 |
| production volume +1 months (12 month mean) | 0.0 | 0.0 |
| production volume +1 months (24 month std) | 0.0 | 0.0 |
| production volume +1 months (6 month mean) | 0.0 | 0.0 |
| production volume +1 months (6 month std) | 0.0 | 0.0 |
| production volume +10 months (24 month mean) | 0.0 | 0.0 |
| production volume +11 months (12 month std) | 0.0 | 0.0 |
| production volume +12 months (1st lag) | 0.0 | 0.0 |
| production volume +12 months (24 month std) | 0.0 | 0.0 |
| production volume +12 months (6 month std) | 0.0 | 0.0 |
| production volume +13 months (24 month std) | 0.0 | 0.0 |
| production volume +2 months (12 month mean) | 0.0 | 0.0 |
| production volume +2 months (12 month std) | 0.0 | 0.0 |
| production volume +2 months (24 month std) | 0.0 | 0.0 |
| production volume +3 months (12 month mean) | 0.0 | 0.0 |
| production volume +3 months (12 month std) | 0.0 | 0.0 |
| production volume +3 months (1st lag) | 0.0 | 0.0 |
| production volume +3 months (24 month mean) | 0.0 | 0.0 |
| production volume +3 months (24 month std) | 0.0 | 0.0 |
| production volume +4 months (12 month mean) | 0.0 | 0.0 |
| production volume +4 months (12 month std) | 0.0 | 0.0 |
| production volume +4 months (1st lag) | 0.0 | 0.0 |
| production volume +4 months (24 month std) | 0.0 | 0.0 |
| production volume +4 months (6 month mean) | 0.0 | 0.0 |
| production volume +5 months (1st lag) | 0.0 | 0.0 |
| production volume +6 months (1st lag) | 0.0 | 0.0 |
| production volume +6 months (24 month std) | 0.0 | 0.0 |
| production volume +6 months (6 month mean) | 0.0 | 0.0 |
| production volume +7 months (12 month mean) | 0.0 | 0.0 |
| production volume +7 months (24 month std) | 0.0 | 0.0 |
| production volume +7 months (6 month std) | 0.0 | 0.0 |
| production volume +8 months (12 month mean) | 0.0 | 0.0 |
| production volume +8 months (1st lag) | 0.0 | 0.0 |
| production volume +9 months (24 month mean) | 0.0 | 0.0 |
| production volume +9 months (6 month mean) | 0.0 | 0.0 |
| production volume \_1 months (12 month mean) | 0.0 | 0.0 |
| production volume \_1 months (12 month std) | 0.0 | 0.0 |
| production volume \_1 months (6 month mean) | 0.0 | 0.0 |
| production volume \_10 months (12 month mean) | 0.0 | 0.0 |
| production volume \_10 months (12 month std) | 0.0 | 0.0 |
| production volume \_10 months (1st lag) | 0.0 | 0.0 |
| production volume \_10 months (24 month mean) | 0.0 | 0.0 |
| production volume \_10 months (24 month std) | 0.0 | 0.0 |
| production volume \_10 months (6 month mean) | 0.0 | 0.0 |
| production volume \_10 months (6 month std) | 0.0 | 0.0 |
| production volume \_11 months (12 month std) | 0.0 | 0.0 |
| production volume \_11 months (1st lag) | 0.0 | 0.0 |
| production volume \_11 months (24 month mean) | 0.0 | 0.0 |
| production volume \_11 months (24 month std) | 0.0 | 0.0 |
| production volume \_11 months (6 month std) | 0.0 | 0.0 |
| production volume \_12 months (12 month mean) | 0.0 | 0.0 |
| production volume \_12 months (12 month std) | 0.0 | 0.0 |
| production volume \_12 months (1st lag) | 0.0 | 0.0 |
| production volume \_12 months (24 month mean) | 0.0 | 0.0 |
| production volume \_12 months (24 month std) | 0.0 | 0.0 |
| production volume \_12 months (6 month mean) | 0.0 | 0.0 |
| production volume \_12 months (6 month std) | 0.0 | 0.0 |
| production volume \_13 months (12 month mean) | 0.0 | 0.0 |
| production volume \_13 months (12 month std) | 0.0 | 0.0 |
| production volume \_13 months (1st lag) | 0.0 | 0.0 |
| production volume \_13 months (24 month mean) | 0.0 | 0.0 |
| production volume \_13 months (24 month std) | 0.0 | 0.0 |
| production volume \_13 months (6 month std) | 0.0 | 0.0 |
| production volume \_2 months (12 month std) | 0.0 | 0.0 |
| production volume \_2 months (1st lag) | 0.0 | 0.0 |
| production volume \_2 months (24 month mean) | 0.0 | 0.0 |
| production volume \_2 months (6 month std) | 0.0 | 0.0 |
| production volume \_3 months (12 month mean) | 0.0 | 0.0 |
| production volume \_3 months (12 month std) | 0.0 | 0.0 |
| production volume \_3 months (24 month mean) | 0.0 | 0.0 |
| production volume \_3 months (24 month std) | 0.0 | 0.0 |
| production volume \_3 months (6 month std) | 0.0 | 0.0 |
| production volume \_4 months (24 month std) | 0.0 | 0.0 |
| production volume \_4 months (6 month std) | 0.0 | 0.0 |
| production volume \_5 months (12 month mean) | 0.0 | 0.0 |
| production volume \_5 months (12 month std) | 0.0 | 0.0 |
| production volume \_5 months (24 month mean) | 0.0 | 0.0 |
| production volume \_5 months (24 month std) | 0.0 | 0.0 |
| production volume \_5 months (6 month mean) | 0.0 | 0.0 |
| production volume \_5 months (6 month std) | 0.0 | 0.0 |
| production volume \_6 months (12 month mean) | 0.0 | 0.0 |
| production volume \_6 months (12 month std) | 0.0 | 0.0 |
| production volume \_6 months (1st lag) | 0.0 | 0.0 |
| production volume \_6 months (24 month mean) | 0.0 | 0.0 |
| production volume \_6 months (6 month std) | 0.0 | 0.0 |
| production volume \_7 months (12 month mean) | 0.0 | 0.0 |
| production volume \_7 months (12 month std) | 0.0 | 0.0 |
| production volume \_7 months (24 month mean) | 0.0 | 0.0 |
| production volume \_7 months (6 month std) | 0.0 | 0.0 |
| production volume \_8 months (12 month mean) | 0.0 | 0.0 |
| production volume \_8 months (12 month std) | 0.0 | 0.0 |
| production volume \_8 months (24 month mean) | 0.0 | 0.0 |
| production volume \_8 months (24 month std) | 0.0 | 0.0 |
| production volume \_8 months (6 month mean) | 0.0 | 0.0 |
| production volume \_8 months (6 month std) | 0.0 | 0.0 |
| production volume \_9 months (12 month mean) | 0.0 | 0.0 |
| production volume \_9 months (12 month std) | 0.0 | 0.0 |
| production volume \_9 months (24 month std) | 0.0 | 0.0 |
| production volume \_9 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +0 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +0 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +0 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +1 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +1 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +1 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +1 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +10 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +10 months (12 month std) | 0.0 | 0.0 |
| revenue forecast +10 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +10 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +11 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +11 months (12 month std) | 0.0 | 0.0 |
| revenue forecast +11 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +11 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +11 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +11 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +12 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +12 months (12 month std) | 0.0 | 0.0 |
| revenue forecast +12 months (1st lag) | 0.0 | 0.0 |
| revenue forecast +12 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +12 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +12 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +13 months (1st lag) | 0.0 | 0.0 |
| revenue forecast +13 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +13 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +13 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +2 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +2 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +3 months (12 month std) | 0.0 | 0.0 |
| revenue forecast +3 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +3 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +4 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +4 months (1st lag) | 0.0 | 0.0 |
| revenue forecast +4 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +4 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +4 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +4 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +5 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +5 months (1st lag) | 0.0 | 0.0 |
| revenue forecast +5 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +5 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +6 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +6 months (1st lag) | 0.0 | 0.0 |
| revenue forecast +6 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +6 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +6 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +7 months (12 month std) | 0.0 | 0.0 |
| revenue forecast +7 months (1st lag) | 0.0 | 0.0 |
| revenue forecast +7 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +7 months (6 month std) | 0.0 | 0.0 |
| revenue forecast +8 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +8 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +9 months (12 month mean) | 0.0 | 0.0 |
| revenue forecast +9 months (12 month std) | 0.0 | 0.0 |
| revenue forecast +9 months (24 month mean) | 0.0 | 0.0 |
| revenue forecast +9 months (24 month std) | 0.0 | 0.0 |
| revenue forecast +9 months (6 month mean) | 0.0 | 0.0 |
| revenue forecast +9 months (6 month std) | 0.0 | 0.0 |
| sales date (actual) | 0.0 | 0.0 |
| total net eur (quarter aggregation) (actual) | 0.0 | 0.0 |
| total production volume last 12 months (12 month mean) | 0.0 | 0.0 |
| total production volume last 12 months (12 month std) | 0.0 | 0.0 |
| total production volume last 12 months (1st lag) | 0.0 | 0.0 |
| total production volume last 12 months (24 month mean) | 0.0 | 0.0 |
| total production volume last 12 months (24 month std) | 0.0 | 0.0 |
| total production volume last 12 months (6 month mean) | 0.0 | 0.0 |
| total production volume last 12 months (6 month std) | 0.0 | 0.0 |
| production volume +12 months (6 month mean) | -0.0 | -0.2255 |
| plant country Freight (6 month std) | -0.0 | -0.4153 |
| production volume +5 months (6 month mean) | -0.0 | -0.6244 |
| Number months since last sale (1st lag) | -0.0 | -0.6992 |
| plant country short\_term interest rate (24 month mean) | -0.0 | -0.7228 |
| production volume +5 months (12 month std) | -0.0 | -0.7665 |
| global body type (1st lag) | -0.0 | -0.7671 |
| plant country Business situation (6 month std) | -0.0 | -0.8123 |
| production volume +13 months (6 month mean) | -0.0 | -0.8306 |
| car set price last 12 months (1st lag) | -0.0 | -0.8373 |
| revenue forecast +11 months (1st lag) | -0.0 | -1.1545 |
| production volume +3 months (6 month std) | -0.0 | -1.2388 |
| plant country Production (12 month std) | -0.0 | -1.2468 |
| production volume +0 months (24 month std) | -0.0 | -1.3586 |
| First sales date (Month) (12 month most\_frequent) | -0.0 | -1.3646 |
| car set price last 12 months (12 month std) | -0.0 | -1.6805 |
| production volume \_2 months (24 month std) | -0.0 | -2.0686 |
| revenue forecast +4 months (12 month std) | -0.0 | -2.3771 |
| platform no\_ (1st lag) | -0.0 | -2.7844 |
| production volume +10 months (6 month mean) | -0.0 | -2.82 |
| car set price last 12 months (24 month std) | -0.0 | -2.8205 |
| production volume \_7 months (1st lag) | -0.0001 | -3.7195 |
| production volume +1 months (24 month mean) | -0.0001 | -4.4343 |
| last sales date (Year) (24 month std) | -0.0001 | -4.5892 |
| production volume +4 months (24 month mean) | -0.0001 | -4.9295 |
| production volume +12 months (24 month mean) | -0.0001 | -5.273 |
| OEM (1st lag) | -0.0001 | -8.3653 |
| plant country selling prices (1st lag) | -0.0001 | -9.0268 |
| production volume \_6 months (24 month std) | -0.0001 | -9.0543 |
| revenue forecast +1 months (24 month mean) | -0.0001 | -9.2152 |
| revenue forecast +1 months (12 month std) | -0.0002 | -10.3595 |
| production volume +7 months (1st lag) | -0.0002 | -10.7938 |
| plant country unemployment rate (12 month mean) | -0.0002 | -12.3455 |
| production volume +9 months (12 month mean) | -0.0002 | -15.8089 |
| production volume +5 months (24 month mean) | -0.0003 | -16.6392 |
| production volume +9 months (12 month std) | -0.0003 | -17.5174 |

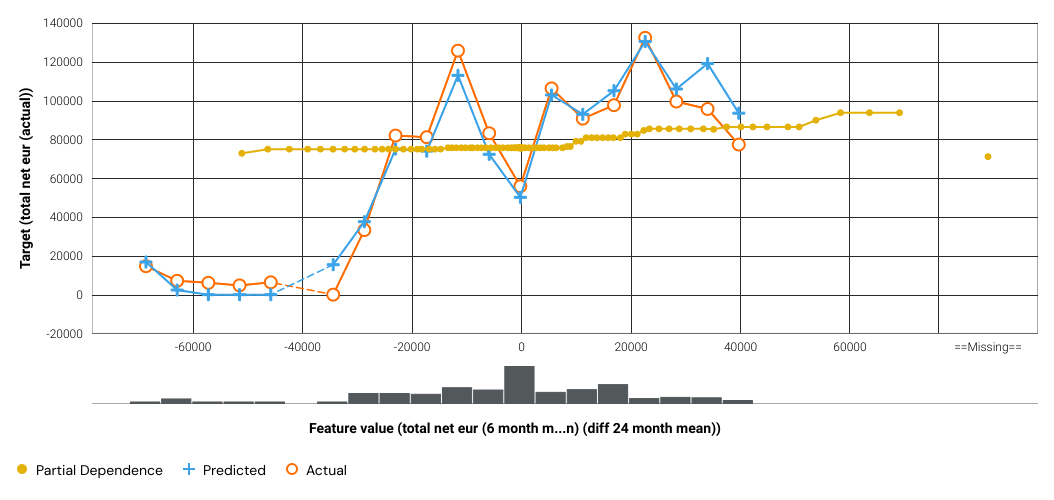
6.3.3 Sensitivity Analysis (Partial Dependence)

In the case of linear regression, users can gain considerable insight into the structure and interpretation of the model by examining its coefficients. For more complex models like support vector machines, random forests, or the blenders considered here, no comparably simple parametric description is available, making the interpretation of these models more difficult. To address this difficulty for his gradient boosting machine, Friedman (2001) proposed the use of partial dependence plots. Partial dependence plots show the average partial relationship between a set of predictors and the predicted response. The partial dependence plots below capture the top features in our model, as measured by Feature Impact.

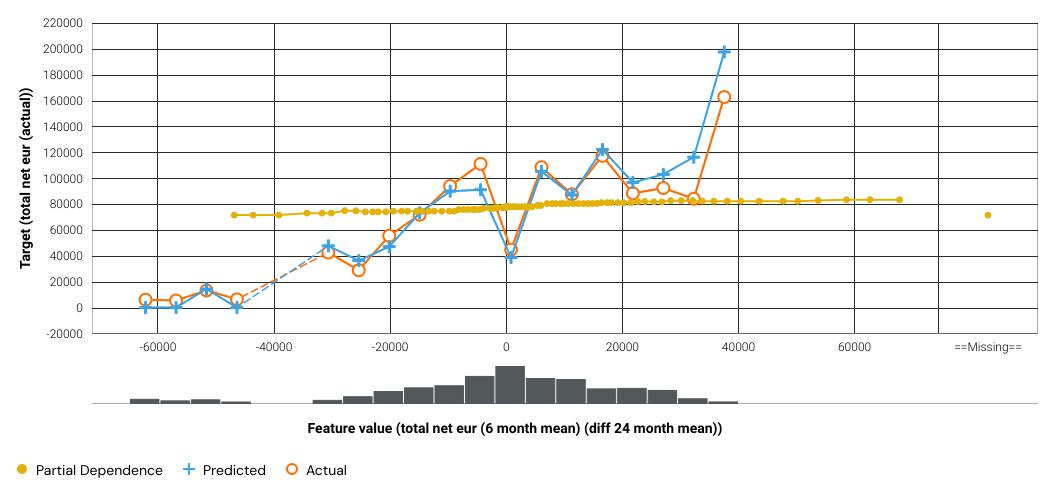
total net eur (24 month average baseline)



total net eur (6 month median) (diff 24 month mean)



total net eur (6 month mean) (diff 24 month mean)



The orange circles depict, for the selected feature, the average target value for the aggregated feature values. The blue crosses depict, for the selected feature, the average prediction for a specific value. From the graph you can see that DataRobot also averages the predicted feature values. Comparing the actual and predicted points can identify segments where model predictions differ from observed data. This typically occurs when the segment size is small. In those cases, for example, some models may predict closer to the overall average.

The yellow partial dependence data points depict the marginal effect of a feature on the target variable after accounting for the average effects of all other predictive features. It indicates how, holding all other variables constant, the value of this feature affects prediction. DataRobot holds constant the values of all columns in the sample except the feature of interest. The value of the feature of interest is then reassigned to each possible value, calculating the average predictions for the sample at each setting. These values help determine how the value of each feature affects the target. The shape of the yellow data points describes the model's view of the marginal relationship between the selected feature and the target.

6.3.4 Accuracy Over Time

The Accuracy Over Time chart is available using DataRobot’s data/time partitioning. It helps to visualize the predictive stability of a model by comparing how predictions vary over time with actual values. By default, the chart shows predicted and actual vs. time values for the training and validation data of the most recent (first) backtest. This is the backtest model DataRobot uses to deploy and make predictions. (In other words, the model for the validation set.). However, for multiseries projects, the chart presents an average of the series.

Bins within the Accuracy Over Time tab are equal width—that is, each bin spans the same time range—while bins in the Lift Chart are equal sized such that each bin contains the same number of rows. In the chart below, the blue line represents predicted values and the orange line represents actual values.

Accuracy Over Time chart has not been calculated for this model.

7 Model Implementation and Output Reporting

7.1 Version Control

DataRobot handles model and project version control automatically by tagging each model on the Leaderboard with a unique Model ID. The Model ID represents a single instance of a model type, feature list, sample size, and set of tuning parameter values. DataRobot also maintains unique Project IDs for each project, allowing accessibility to all models built for the project dataset. DataRobot's version control allows for reproducibility and traceability of the models it creates, which greatly increases the auditability of the model development process.

Users may also export Scoring Code for a DataRobot model in Java. You can download both a pre-compiled .jar file (with all dependencies included), plus the source code. Scoring Code is easy to deploy, test, and maintain on a variety of platforms, and you can inspect the generated Java code for complete transparency. DataRobot Scoring Code employs advanced features to ensure that predictions computed using generated Java code are the same as predictions computed inside DataRobot.