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 Feature Association
* 4.2 Data Source Overview and Appropriateness
* 4.3 Input Data Extraction, Preparation, and Quality & Completeness
* 4.4 Data Assumptions
* 5 Model Theoretical Framework and Methodology
* 5.1 Model Development Overview
* 5.2 Model Assumptions
* 5.3 Model Methodology
* 5.3.1 Ordinal scale converter of categorical features
* 5.3.2 WordGramEnetRegressor
* 5.3.3 Single Column Selector
* 5.3.4 Bind branches
* 5.3.5 Extreme Gradient Boosting Regressor with Grid Search and Early Stopping support
* 5.3.6 Arbitrary or median value-based numeric imputation (V4 with memory usage optimization)
* 5.4 Literature Review and References
* 5.5 Alternative Model Frameworks and Theories Considered
* 5.6 Variable Selection
* 5.6.1 DataRobot Quantitative Analysis
* 5.6.2 Expert Judgement and Variable Selection
* 5.6.3 Final Model Variables
* 5.6.3.1 Model Features and Summary Statistics
* 5.6.3.2 Data Quality Handling Report
* 6 Model Performance and Stability
* 6.1 Model Validation Stability
* 6.1.1 Cross Validation Scores
* 6.1.2 Data Partitioning Methodology
* 6.2 Model Performance (Sample Scores)
* 6.3 Sensitivity Testing and Analysis
* 6.3.1 Lift Chart
* 6.3.2 Key Relationships
* 6.3.3 Sensitivity Analysis (Partial Dependence)
* 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.

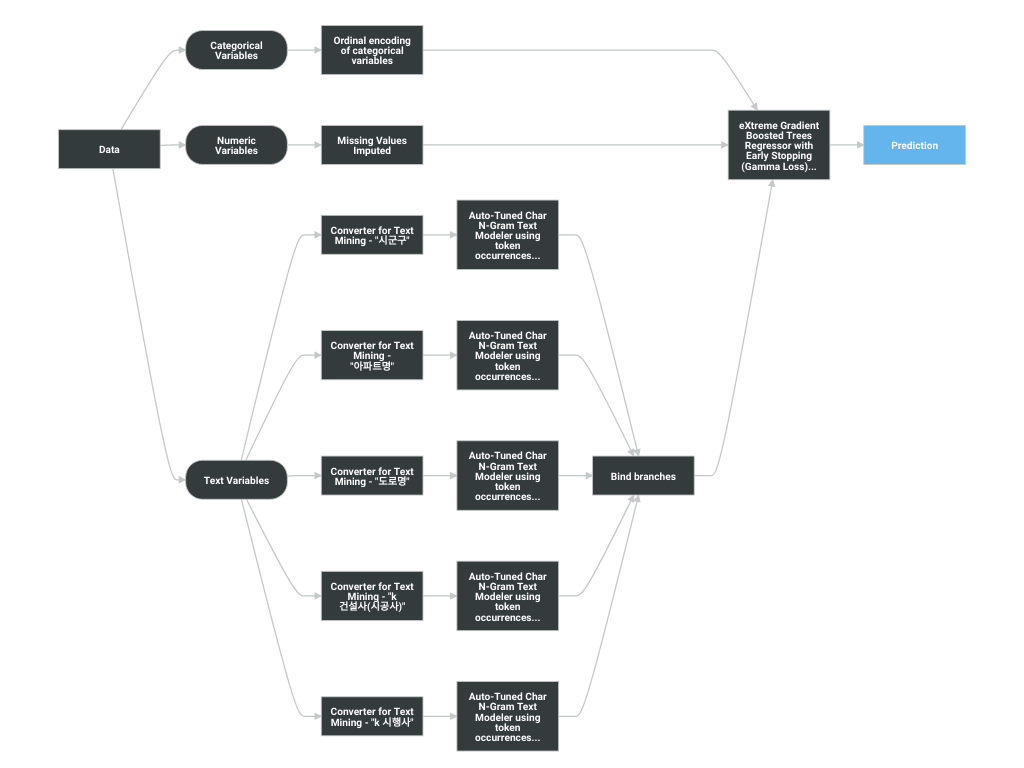
3.2 Model Development Purpose and Intended Use

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 (Gamma Loss) (Fast Feature Binning). This model was developed in a project created with veeed602d97362576 of DataRobot. This model is denoted within DataRobot by the Project ID: 668eea805ded71e21fc89fd1 and the Model ID: 668f273675e4f325515a2a64. The project was created on 2024-07-10 20:09:36.

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
* Converter for Text Mining
* Auto-Tuned Char N-Gram Text Modeler using token occurrences and tfidf
* eXtreme Gradient Boosted Trees Regressor with Early Stopping (Gamma Loss) (Fast Feature Binning)

3.4 Overview of Model Results

DataRobot runs performance testing during the model development process to evaluate model results and reliability. The validation, cross-validation, and holdout (if applicable) out-of-sample performance scores are presented below, as well as the number of observations for each partition. The performance metric used for this project was Gamma Deviance and the project included a total of 1,118,822 observations. An asterisk (\*) next to a score, whether validation or holdout, indicates that DataRobot used in-sample predictions to derive the score. (In-samples predictions are those that include data from the validation or holdout partitions due to sample size used to build the model.)

|  |  |
| --- | --- |
| Scoring Type | Score (Gamma Deviance) |
| cross\_validation | 0.0064\* |
| holdout | 0.0064\* |
| validation | 0.0063\* |

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

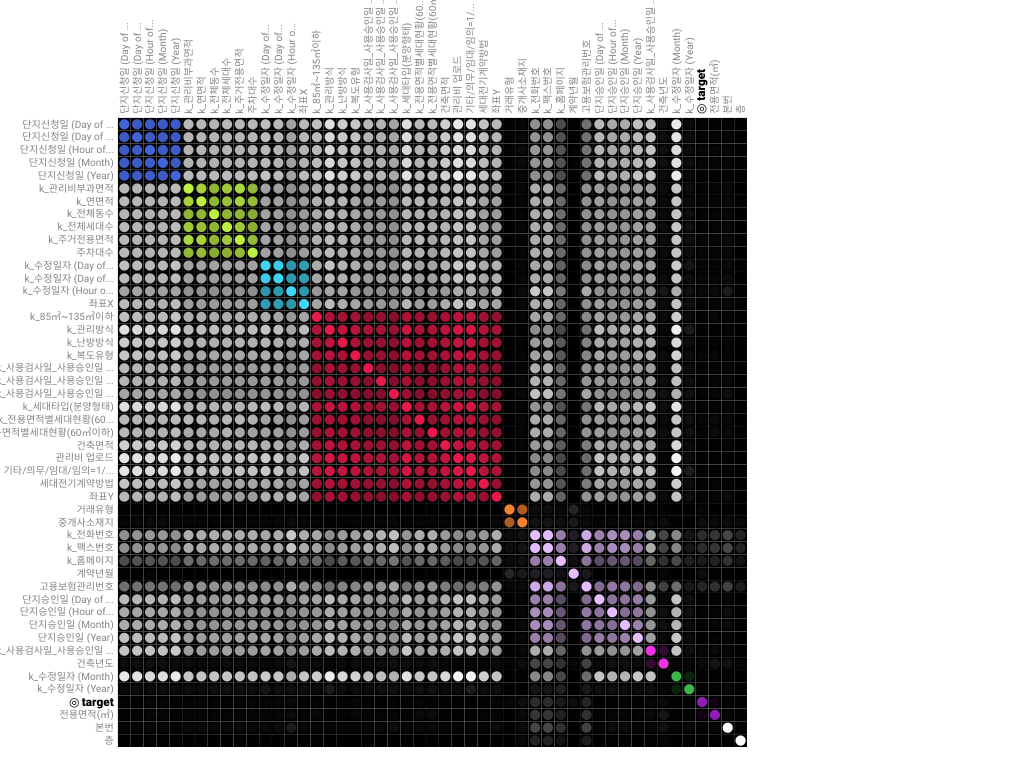
4.1 Feature Association

DataRobot’s Feature Association Matrix is populated by default by features from DataRobot’s Informative Features feature list. The Feature Association Matrix provides information on association strength between pairs of numeric and categorical features that are visually denoted by the opacity of the color (that is, num/cat, num/num, cat/cat, where lighter shades indicate weaker association and vice versa) and feature clusters. Clusters, families of features denoted by color on the matrix, are features partitioned into groups based on their association structure.

Some of the noted benefits of the Feature Association Matrix include:

* Understand the strength and nature of associations within the data;
* Detect families of pairwise association clusters; and,
* Identify clusters of high-association features prior to model building.

The Feature Association Matrix lists up to the top 50 features, selected by Importance Score, on both the X and Y axes, where the intersection of a feature pair provides an indication of their level of association. By default, the matrix displays by the Mutual Information values and sorts by the cluster.



The following are some general takeaways from looking at the matrix above:

* Each dot represents the association between two features (a feature pair), where the opacity of the color denotes the pair-wise strength of association.
* Each cluster is represented by a different color.
* The opacity of color indicates the level of association 0 to 1, between the feature pair. Levels are measured by the set metric, either mutual information or Cramer’s V.
* Shaded gray dots indicate that the two features, while showing some association, are not in the same cluster.
* White dots represent features that were not categorized into a cluster.
* The target feature, if present, is indicated by two small concentric circles next to the feature name.

4.2 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.

4.3 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.

4.4 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:

* Rapid Data Ingestion: User creates a modeling dataset that includes the prediction target and loads into DataRobot
* Target Selection: User selects the prediction target; DataRobot detects whether the target is categorical or continuous. If the target is categorical, DataRobot selects and builds classification blueprints. If the target is continuous, DataRobot selects and builds regression blueprints. DataRobot also selects an optimization performance metric based on the type of supervised learning problem, which can also be changed by the user
* Automated Data Preparation: DataRobot analyzes the input data and automatically performs advanced preprocessing steps that are discussed in detail in this document. DataRobot also automatically partitions the input dataset into learning, validation and holdout dataset; these can also be defined by the user.
* 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 an out-of-sample validation 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, such as AUC, Log-Loss and RMSE. DataRobot's Leaderboard actively tracks performance of candidate models using out-of-sample data for comparison purposes.
* Model Deployment and Monitoring: Once the final model is selected, DataRobot provides efficient solutions for deployment (i.e., model implementation) and monitoring. These features enable the model owner to effectively manage model controls in accordance with Model Risk Management standards and policies.

5.2 Model Assumptions

This section should include model limitations, potential effects, and any mitigating controls in place. Limitations come in part from weaknesses in the model due to its various shortcomings, approximations, and uncertainties. Regulatory guidance refers to limitations as "...a consequence of assumptions underlying a model that may restrict the scope to a limited set of specific circumstances and situations." This section should include model limitations, potential effects, and any mitigating controls in place. Also include details here about the implementation of the models, what data will be used for scoring and why it is reasonable to think that the training data and the scoring data will be similar.

Machine learning methods can produce more accurate predictive models than traditional statistical regression methods because they are more flexible and rely less on statistical assumptions. For instance, ordinary least squares regression requires that the Gauss Markov assumptions are supported, which ensures that the model is unbiased and efficient.

Traditional statistical regression techniques rely on formal hypothesis testing for variable significance and feature selection (e.g., t-test, p-value, standard error). These hypothesis tests tend to have distributional and independence assumptions that may not be supported by the data. Machine learning methods, on the other hand, offer more flexibility in defining the model structure, which typically results in better model performance. Because machine learning includes methods that do not rely on formal hypothesis testing to demonstrate model validity, and because heuristic-style feature selection methods (e.g., stepwise selection) are not used in most machine learning approaches, no such distributional assumptions are required. In this case, the only assumption being made is that the model training data is representative of the future scoring data. Of course, these assumptions must be closely monitored and tracked by the model's ongoing performance monitoring process.

A common limitation of machine learning methods is the potential for overfitting. Overfitting occurs when the model is trained too closely to the underlying training data and does not perform well out-of-sample. DataRobot utilizes a robust cross-validation and holdout methodology to ensure model performance is sound, reducing the risk of overfitting.

5.3 Model Methodology

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

* Ordinal encoding of categorical variables
* Missing Values Imputed
* Converter for Text Mining
* Auto-Tuned Char N-Gram Text Modeler using token occurrences and tfidf
* eXtreme Gradient Boosted Trees Regressor with Early Stopping (Gamma Loss) (Fast Feature Binning)

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

5.3.1 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 |

5.3.2 WordGramEnetRegressor

Fit a single-word n-gram model to each text feature in the input dataset, then use the predictions from these models as inputs to an ElasticNet regressor.

Word N-Gram:

A word n-gram model is a special case of n-gram modeling, where the n-gram is a contiguous sequence of n items from a given sequence of text (an “item” is a single word). Word n-gram models first construct a “vocabulary” of all sequences of words in the input dataset and then make a matrix with counts of n-gram occurrences in each row.

Auto-Tuned N-Gram:

Auto-tuned n-grams automatically decide the size of n-gram to use, trying unigrams (or a bag-of-words model), bigrams, and trigrams, and then selecting the method with the best score on an out-of sample dataset. Unigrams, in this case, are individual words in the corpus. Bigrams will be any pair of words, such as “thank you” or “new york”. Trigrams are trios of sequential words, for example, “how are you”, “old brown dog”, “long, long ago”.

Further Text Processing:

This task can also use the TF-IDF transformation, which weights a word or sequence more highly if it appears more frequently within a document. It downweights words that appear more frequently between documents. This transformation performs well in a wide variety of downstream problems.

The task combines a text-vectorizer with an ElasticNet Regression estimator to enable gridsearches that select the best text-vectorization parameters. It includes the advanced settings option of running Singular Value Decomposition (SVD) on the vectorized text features.

If Japanese text is detected, the task uses MeCab to tokenize the text prior to finding word n-grams.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| select | binary | When True, indicates a more general presence or absence (using 1) of counts rather than the actual count or term frequency. Default is to report counts/frequencies. values: [False, True] | True |
| select | decode\_error | Desired behavior when an encoding error is encountered. values: ['strict', 'ignore', 'replace'] | replace |
| select | encoding | Unicode encoding method use for decoding files. values: ['utf-8', 'latin-1'] | utf-8 |
| multi | enet\_alpha | ElasticNet mixing parameter, where 0 <= alpha <= 1. When alpha = 0, an L2 penalty is assigned; when alpha = 1, an L1 penalty is assigned. When 0 < alpha < 1, the penalty is a combination of L1 and L2. The ‘auto’ selection assigns a grid of 11 values, evenly spaced from 0.0 to 1.0. values: {'floatgrid': [0, 1], 'select': ['auto']} | 0.0 |
| int | enet\_gs\_folds | Number of gridsearch folds for elastic net tuning values: [1, 100] | 1 |
| float | enet\_gs\_fraction | Fraction to holdout for gridsearch for elastic net tuning values: [1e-5, 1 - 1e-5] | 0.15 |
| select | enet\_gs\_light | Whether to run light gridsearch for elastic net tuning This saves time, but might not yield quite as predictive a model. values: [False, True] | True |
| int | enet\_n\_lambda | Number of terms use in the grid for tuning the penalty term. The grid used is a grid of enet\_n\_lambda numbers, spaced evenly on a log10 scale from 3.16e-5 to 3.16e-0.1. values: [2, 100] | 50 |
| select | enet\_stk\_seq | Elastic Net Stack Sequential: Whether to produce out-of-sample predictions in parallel values: [False, True] | True |
| bool | fit\_tweedie\_p | When True, the Tweedie power parameter (tweedie\_p) is automatically set. values: [False, True] | False |
| select | include\_lower | When True, the default, includes all terms down to unigrams in results. When False, use num\_ngram. values: [False, True] | True |
| select | language | Language to use in stemmers (when available) and stop words. If ‘japanese’ is selected, the default word tokenizer used will be japanese. values: ['arabic', 'english', 'chinese', 'danish', 'dutch', 'finnish', 'french', 'german', 'hindi', 'hungarian', 'italian', 'norwegian', 'portuguese', 'romanian', 'russian', 'spanish', 'swedish', 'japanese', 'turkish', 'other'] | other |
| select | lemmatizer | (Early Release, contact your DataRobot representative to enable). Lemmatizer implementation to use. Lemmatization uses part-of-speech and context to determine word lemmas, and provides different results than stemming. Lemmatization is mutually exclusive with stemming and part-of-speech tagging. There are two lemmatization libraries used: the WordNet corpus from NLTK and a spaCy model. Note that the wordnet lemmatizer option differs from the wordnet stemmer option in that it uses part-of-speech tagging for more accurate lemmas at the cost of increased runtime. Currently only English is supported. values: [None, 'wordnet', 'spacy'] | None |
| select | loss | Loss function to use. values: ['squared', 'poisson', 'gamma', 'tweedie'] | gamma |
| select | lowercase | If True, converts all characters to lowercase before tokenizing. By default, False, no conversion is applied. values: [False, True] | False |
| float or int | max\_df | Sets the document frequency maximum thresholds using either raw counts or percentages. If float, only keep words that occur less than max\_df percent of the time. If int, only keep words that occur fewer than max\_df times. values: '{'floatgrid': [0.0, 1.0],'intgrid': [0, 1e30]} | 0.4 |
| multi | max\_features | If a value is specified, build a vocabulary that only considers the top max\_features, ordered by term frequency, across the corpus. values: {'select': [None], 'intgrid': [1, 1e30]} | 20000 |
| int | max\_iter | The maximum number of iterations values: [1, 1e6] | 100 |
| multi | max\_samples | Maximum sample size, in megabytes, used for training. values: {'float': [0, 1000], 'select': ['None']} | 500 |
| select | metric\_ace | Metric to base the ACE calculation on, used for stacked predictions. If None, no ACE score is calculated. | Gamma Deviance |
| float or int | min\_df | Sets the document frequency minimum thresholds using either raw counts or percentages. If float, only keep words that occur greater than min\_df percent of the time. If int, only keep words that occur more than min\_df times. values: {'floatgrid': [0.0, 1.0],'intgrid': [0, 1e30]} | 5 |
| int | min\_tc | Minimum term count for TF-IDF. values: [0, int(1e9)] | 0 |
| int | n\_features | Number of features (buckets) the hashing algorithm will use. Used only if use\_hashing is True. values: [1, 2\*\*20] | 262144 |
| select | norm | Norm used to normalize term vectors. If None, no normalization occurs. values: [None, 'l1', 'l2'] | l2 |
| intgrid | num\_ngram | Number of ngrams to use in the vectorizer. If include\_lower == True than also includes all values down to 1 values: [1, 1e30] | [2, 3] |
| select | segmenter | Legacy parameter, superseded by tokenizer. If set to ‘japanese’, tokenizer and language are set appropriately. values:[None, 'japanese'] | None |
| select | smooth\_idf | When True, smooth weights by adding 1 to all document frequencies. This has the effect of acting as if a document containing every term exactly once was added to the corpus. values: [False, True] | True |
| select | stemmer | NLTK word stemmer to use when analyzer=’word’. values: [None, 'snowball', 'lancaster', 'porter', 'wordnet'] | None |
| multi | stop\_words | If True, removes stop words in the vectorizer results (defaults to False). The language parameter specifies which stop words to consider. values: {'select': [False, True]} | None |
| select | sublinear\_tf | Apply sublinear term frequency scaling (replace tf `with `1 + log(tf)). values: [False, True] | False |
| select | svd\_algorithm | Deprecated. SVD solver to use for finding singular vector. We always use ‘randomized’, regardless of this value. values: [None, 'randomized', 'arpack'] | None |
| intgrid | svd\_max\_components | Maximum number of components (to manage memory in the metablueprint) to keep. values: [1, int(1e5)] | 1000 |
| intgrid | svd\_n\_components | Maximum number of singular values/vectors to keep. values: [10, 500] | 100 |
| int | svd\_n\_iter | Number of iterations for the randomized SVD solver. values: [0, 10] | 2 |
| float | svd\_tol | Deprecated. Tolerance of ARPACK, which is no longer supported values: [0, 1] | 0.0 |
| select | tokenizer | NLTK word tokenizer (see http://www.nltk.org/api/nltk.tokenize.html). Only considered when analyzer is set to ‘word’. If None, use the default scikit-learn vectorizer. If language is japanese, the ‘tiny-segmenter-jp’ tokenizer is used. values: [None, 'space', 'wordpunct', 'tweet', 'treebank', 'tiny-segmenter-jp', 'mecab' ] | sklearn\_tokenizer |
| 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 the ElasticNet model. values: [1.0, 2.0] | 1.5 |
| select | use\_bns | When True, enables bi-normal separation reweighting, which can enhance vector representations. Defaults to False. values: [False, True] | None |
| select | use\_char\_preprocessing | When True (the default), preprocess text by replacing non-ASCII characters with whitespace. values: [False, True] | False |
| select | use\_delta | When True, enables delta TF-IDF reweighting (values are weighted by calculating the difference of a word’s TF-IDF scores). Defaults to False. values: [False, True] | None |
| select | use\_hashing | When True, uses the hashing trick to find token string names for feature integer index mapping. Defaults to False. values: [False, True] | False |
| select | use\_idf | When True (the default), enables inverse-document-frequency reweighting. values: [False, True] | True |
| select | use\_marisa | When True (the default), uses the MARISA algorithm to store vocabulary. MARISA is a space-efficient trie data structure (see https://en.wikipedia.org/wiki/Trie). values: [False, True] | True |
| select | use\_term\_frequency | Controls whether to use DictVectorizer or CountVectorizer values: [False, True] | True |

5.3.3 Single Column Selector

Task that selects and filters for a single column from the dataset and enables other tasks in the pipeline to apply transformations only to this column.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| string | column\_name | Name of the column to select. | 6b5feab1b4ec84a4ec82ac28ec8b9ceab3b5ec82ac29 |

5.3.4 Bind branches

DataRobot splits up input data by data type, and implements separate data processing workflows for each data type. Therefore, numeric data, categorical data, text data, etc. all get preprocessing workflows optimized for that data type, for example imputing missing values for numerics, one-hot encoding categoricals, and ngrams + tf-idf weighting for text. DataRobot’s modeling tasks all know how to handle these types of data, and can take the output from the preprocessing directly as input.

However, DataRobot’s more advanced blueprints might implement several different ways of preprocessing the same data type, e.g. searching for Box-Cox transformations and searching for polynomial transformations for numeric variables, or calculating actuarial credibility ratios and one-hot encoding for categorical variables.

The bind tasks operates within a data type (numeric, categorical, text, etc.) and binds together the output of all the different preprocessing pipelines within that data type. The combined output is then ready to pass to the modeling task as a single input.

5.3.5 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] | True |
| 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] | 0.3 |
| 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 |
| select | loss | Loss function to be optimized. ‘ls’ refers to least squares regression. values: ['ls', 'labs', 'poisson', 'tweedie', 'gamma'] | gamma |
| 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] | 7 |
| 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)] | 0.0 |
| 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] | 3500 |
| 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. | 1.0 |
| 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'] | hist |
| float | tweedie\_p | Power parameter of the tweedie distribution. Only applicable when Tweedie loss is used. values: [1, 2] | 1.5 |

5.3.6 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 |

5.4 Literature Review and References

* Sidorov, Grigori, et al. “Syntactic n-grams as machine learning features for natural language processing.” Expert Systems with Applications 41.3 (2014): 853-860.
* Dunning, Ted. “Statistical identification of language.” Computing Research Laboratory, New Mexico State University, 1994. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.1958&rep=rep1&type=pdf>
* Franz, Alex, and Thorsten Brants. “All our n-gram are belong to you.” Google Machine Translation Team 20 (2006). <http://googleresearch.blogspot.pt/2006/08/all-our-n-gram-are-belong-to-you.html>
* 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
* [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

5.5 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 Gamma Deviance. The model types considered during the model selection process included the following models, which are sorted by the Holdout score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Validation Score | Cross Validation Score | Holdout Score | Sample Percentage |
| Keras Slim Residual Neural Network Regressor using Adaptive Training Schedule (1 Layer: 64 Units) | 0.0057 | N/A | 0.0057 | 64.0 |
| Light Gradient Boosted Trees Regressor with Early Stopping (Gamma Loss) | 0.0058 | N/A | 0.006 | 64.0 |
| Light Gradient Boosting on ElasticNet Predictions (Gamma Loss) | 0.0064 | N/A | 0.0064 | 64.0 |
| Keras Deep Residual Neural Network Regressor using Training Schedule (2 Layers: 512, 512 Units) | 0.0072 | N/A | 0.0073 | 64.0 |
| Keras Deep Residual Neural Network Regressor using Training Schedule (3 Layers: 512, 64, 64 Units) | 0.0075 | N/A | 0.0075 | 64.0 |
| Keras Wide Residual Neural Network Regressor using Training Schedule (1 Layer: 1536 Units) | 0.0079 | N/A | 0.0079 | 64.0 |
| Generalized Additive2 Model (Gamma Loss) | 0.0144 | N/A | 0.0143 | 64.0 |
| Keras Slim Residual Neural Network Regressor using Training Schedule (1 Layer: 64 Units) | 0.0166 | N/A | 0.0167 | 64.0 |
| Keras Deep Self-Normalizing Residual Neural Network Regressor using Training Schedule (3 Layers: 256, 128, 64 Units) | 0.0201 | N/A | 0.0202 | 64.0 |
| Keras Residual AutoInt Regressor using Training Schedule (2 Attention Layers with 2 Heads, 2 Layers: 96, 96 Units) | 0.0242 | N/A | 0.0242 | 64.0 |
| Elastic-Net Regressor (mixing alpha=0.5 / Gamma Deviance) | 0.0327 | N/A | 0.033 | 64.0 |
| Elastic-Net Regressor (L2 / Gamma Deviance) | 0.0327 | N/A | 0.033 | 64.0 |
| Keras Residual Neural Factorization Machine Regressor using Training Schedule (2 Layers: 96, 96 Units) | 0.0709 | N/A | 0.0711 | 64.0 |
| Keras Residual Cross Network Regressor using Training Schedule (3 Cross Layers, 4 Layers: 96, 96, 72, 72 Units) | 0.4313 | N/A | 0.4358 | 64.0 |

5.6 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.
* Expert Judgment and Variable Selection: Summary of the expert judgment used during the variable selection process.
* Final Model Variables: Final feature list chosen.

5.6.1 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., Gamma Deviance) or the default assigned by DataRobot.

5.6.2 Expert Judgement and Variable Selection

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

5.6.3 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 (Gamma Loss) (Fast Feature Binning) 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).

5.6.3.1 Model Features and Summary Statistics

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature Name | Var Type | Unique | Missing | Mean | Std Dev | Median | Min | Max | Target Leakage |
| 시군구 | Text | 338 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| 번지 | Categorical | 6555 | 184 | N/A | N/A | N/A | N/A | N/A | Low |
| 본번 | Numeric | 1522 | 62 | 564.95 | 515.67 | 470.0 | 0.0 | 4969.0 | Low |
| 부번 | Numeric | 328 | 62 | 5.96 | 46.86 | 0.0 | 0.0 | 2837.0 | Low |
| 아파트명 | Text | 6522 | 1712 | N/A | N/A | N/A | N/A | N/A | N/A |
| 전용면적(㎡) | Numeric | 14218 | 0 | 77.16 | 29.38 | 81.86 | 10.02 | 424.32 | Low |
| 계약년월 | Numeric | 198 | 0 | 201475.9 | 418.81 | 201507.0 | 200701.0 | 202306.0 | Low |
| 계약일 | Numeric | 31 | 0 | 15.8 | 8.72 | 16.0 | 1.0 | 31.0 | Low |
| 층 | Numeric | 73 | 0 | 8.87 | 5.98 | 8.0 | -4.0 | 69.0 | Low |
| 건축년도 | Numeric | 60 | 0 | 1998.76 | 9.33 | 2000.0 | 1961.0 | 2023.0 | Low |
| 도로명 | Text | 9195 | 973 | N/A | N/A | N/A | N/A | N/A | N/A |
| 해제사유발생일 | Date | 986 | 890216 | 2021-07-04 | 368.27 days | 2021-03-08 | 2020-02-21 | 2023-09-26 | Low |
| 등기신청일자 | Date | 181 | 883084 | 2023-06-16 | 57.27 days | 2023-06-22 | 2023-01-02 | 2023-09-22 | Low |
| 거래유형 | Categorical | 3 | 0 | N/A | N/A | N/A | N/A | N/A | Low |
| 중개사소재지 | Categorical | 575 | 0 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_단지분류(아파트,주상복합등등) | Categorical | 5 | 696619 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_전화번호 | Categorical | 1078 | 696280 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_팩스번호 | Categorical | 1356 | 698258 | N/A | N/A | N/A | N/A | N/A | Low |
| 단지소개기존clob | Numeric | 94 | 840380 | 542.005 | 752.98 | 144.0 | 1.0 | 2888.0 | Low |
| k\_세대타입(분양형태) | Categorical | 3 | 695698 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_관리방식 | Categorical | 3 | 695698 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_복도유형 | Categorical | 5 | 695962 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_난방방식 | Categorical | 4 | 695698 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_전체동수 | Numeric | 41 | 696577 | 14.79 | 17.69 | 10.0 | 1.0 | 124.0 | Low |
| k\_전체세대수 | Numeric | 520 | 695698 | 1184.24 | 1190.37 | 768.0 | 59.0 | 9510.0 | Low |
| k\_건설사(시공사) | Text | 344 | 696911 | N/A | N/A | N/A | N/A | N/A | N/A |
| k\_시행사 | Text | 553 | 697757 | N/A | N/A | N/A | N/A | N/A | N/A |
| k\_사용검사일\_사용승인일 | Date | 672 | 695814 | 2002-05-20T18:39:16.828813 | 3231.17 days | 2003-10-04T00:00:00 | 1976-07-09T00:00:00 | 2023-01-27T00:00:00 | Low |
| k\_연면적 | Numeric | 733 | 695698 | 161460.5 | 183831.36 | 101633.0 | 0.0 | 9591851.0 | Low |
| k\_주거전용면적 | Numeric | 738 | 695737 | 94169.65 | 101749.52 | 60280.0 | 2338.0 | 734781.0 | Low |
| k\_관리비부과면적 | Numeric | 734 | 695698 | 120667.46 | 128828.205 | 78125.0 | 0.0 | 969877.0 | Low |
| k\_전용면적별세대현황(60㎡이하) | Numeric | 347 | 695737 | 478.8 | 760.034 | 226.0 | 0.0 | 4975.0 | Low |
| k\_전용면적별세대현황(60㎡~85㎡이하) | Numeric | 386 | 695737 | 476.54 | 727.045 | 256.0 | 0.0 | 5132.0 | Low |
| k\_85㎡~135㎡이하 | Numeric | 244 | 695737 | 167.107 | 248.46 | 63.0 | 0.0 | 1500.0 | Low |
| k\_135㎡초과 | Numeric | 2 | 894794 | 70.0 | 0.0 | 70.0 | 70.0 | 70.0 | Low |
| k\_홈페이지 | Categorical | 220 | 804418 | N/A | N/A | N/A | N/A | N/A | Low |
| k\_등록일자 | Date | 125 | 886218 | 2018-09-28T18:58:16.806901 | 570.066 days | 2018-06-25T21:14:05 | 2017-02-01T10:49:21 | 2023-05-12T10:09:44 | Low |
| k\_수정일자 | Date | 742 | 695737 | 2023-09-03T09:00:50.819451 | 163.909 days | 2023-09-25T07:04:29 | 2020-02-17T04:28:42 | 2023-09-26T12:46:39 | Low |
| 고용보험관리번호 | Categorical | 526 | 730705 | N/A | N/A | N/A | N/A | N/A | Low |
| 경비비관리형태 | Categorical | 4 | 696847 | N/A | N/A | N/A | N/A | N/A | Low |
| 세대전기계약방법 | Categorical | 2 | 703084 | N/A | N/A | N/A | N/A | N/A | Low |
| 청소비관리형태 | Categorical | 4 | 696996 | N/A | N/A | N/A | N/A | N/A | Low |
| 건축면적 | Numeric | 454 | 695817 | 190762.0017 | 1736524.26 | 1623.8 | 0.0 | 31596200.0 | Low |
| 주차대수 | Numeric | 525 | 695817 | 1063.74 | 1235.21 | 683.0 | 0.0 | 12096.0 | Low |
| 기타/의무/임대/임의=1/2/3/4 | Categorical | 4 | 695698 | N/A | N/A | N/A | N/A | N/A | Low |
| 단지승인일 | Date | 734 | 696286 | 2015-05-17T09:25:31.234696 | 1330.404 days | 2015-10-16T10:47:13 | 1982-09-18T00:00:00 | 2023-08-04T15:39:34 | Low |
| 사용허가여부 | Categorical | 1 | 695698 | N/A | N/A | N/A | N/A | N/A | Low |
| 관리비 업로드 | Categorical | 2 | 695698 | N/A | N/A | N/A | N/A | N/A | Low |
| 좌표X | Numeric | 740 | 695785 | 127.0 | 0.091 | 127.015 | 126.8 | 127.18 | Low |
| 좌표Y | Numeric | 740 | 695785 | 37.55 | 0.053 | 37.54 | 37.45 | 37.69 | Low |
| 단지신청일 | Date | 258 | 695746 | 2013-08-02T02:03:17.947906 | 495.16 days | 2013-03-07T09:46:39 | 2013-03-07T09:46:12 | 2023-08-04T15:31:27 | Low |
| target | Numeric | 12875 | 0 | 57962.59 | 46347.64 | 44750.0 | 500.0 | 1450000.0 | 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.

5.6.3.2 Data Quality Handling Report

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Name | Var Type | Missing Count | Missing Percentage | Imputation Name | Imputation Description |
| k\_135㎡초과 | Numeric | 1118495 | 100 | Missing Values Imputed | Imputed value: -9999 |
| 해제사유발생일 | Numeric | 1112839 | 99 | Missing Values Imputed | Imputed value: -9999 |
| k\_등록일자 (Day of Week) | Categorical | 1107832 | 99 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_등록일자 (Hour of Day) | Categorical | 1107832 | 99 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_등록일자 (Month) | Categorical | 1107832 | 99 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_등록일자 (Day of Month) | Numeric | 1107832 | 99 | Missing Values Imputed | Imputed value: -9999 |
| k\_등록일자 (Year) | Numeric | 1107832 | 99 | Missing Values Imputed | Imputed value: -9999 |
| 등기신청일자 | Numeric | 1103907 | 99 | Missing Values Imputed | Imputed value: -9999 |
| 단지소개기존clob | Numeric | 1050240 | 94 | Missing Values Imputed | Imputed value: -9999 |
| k\_홈페이지 | Categorical | 1005718 | 90 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 고용보험관리번호 | Categorical | 913304 | 82 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 세대전기계약방법 | Categorical | 878747 | 79 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_팩스번호 | Categorical | 872742 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 청소비관리형태 | Categorical | 871178 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 경비비관리형태 | Categorical | 870988 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_단지분류(아파트,주상복합등등) | Categorical | 870691 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_전체동수 | Numeric | 870630 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 단지승인일 (Day of Week) | Categorical | 870286 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 단지승인일 (Hour of Day) | Categorical | 870286 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 단지승인일 (Month) | Categorical | 870286 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 단지승인일 (Day of Month) | Numeric | 870286 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 단지승인일 (Year) | Numeric | 870286 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_전화번호 | Categorical | 870274 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_복도유형 | Categorical | 869890 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 건축면적 | Numeric | 869714 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 주차대수 | Numeric | 869714 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_사용검사일\_사용승인일 (Day of Week) | Categorical | 869696 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_사용검사일\_사용승인일 (Month) | Categorical | 869696 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_사용검사일\_사용승인일 (Day of Month) | Numeric | 869696 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_사용검사일\_사용승인일 (Year) | Numeric | 869696 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 좌표X | Numeric | 869670 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 좌표Y | Numeric | 869670 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 단지신청일 (Day of Week) | Categorical | 869625 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 단지신청일 (Hour of Day) | Categorical | 869625 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 단지신청일 (Month) | Categorical | 869625 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 단지신청일 (Day of Month) | Numeric | 869625 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 단지신청일 (Year) | Numeric | 869625 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_수정일자 (Day of Week) | Categorical | 869608 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_수정일자 (Hour of Day) | Categorical | 869608 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_수정일자 (Month) | Categorical | 869608 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_주거전용면적 | Numeric | 869608 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_전용면적별세대현황(60㎡이하) | Numeric | 869608 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_전용면적별세대현황(60㎡~85㎡이하) | Numeric | 869608 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_85㎡~135㎡이하 | Numeric | 869608 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_수정일자 (Day of Month) | Numeric | 869608 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_수정일자 (Year) | Numeric | 869608 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_세대타입(분양형태) | Categorical | 869563 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_관리방식 | Categorical | 869563 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_난방방식 | Categorical | 869563 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 기타/의무/임대/임의=1/2/3/4 | Categorical | 869563 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 사용허가여부 | Categorical | 869563 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 관리비 업로드 | Categorical | 869563 | 78 | Ordinal encoding of categorical variables | Imputed value: -2 |
| k\_전체세대수 | Numeric | 869563 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_연면적 | Numeric | 869563 | 78 | Missing Values Imputed | Imputed value: -9999 |
| k\_관리비부과면적 | Numeric | 869563 | 78 | Missing Values Imputed | Imputed value: -9999 |
| 번지 | Categorical | 225 | 0 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 본번 | Numeric | 75 | 0 | Missing Values Imputed | Imputed value: -9999 |
| 부번 | Numeric | 75 | 0 | Missing Values Imputed | Imputed value: -9999 |
| 거래유형 | Categorical | 0 | 0 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 중개사소재지 | Categorical | 0 | 0 | Ordinal encoding of categorical variables | Imputed value: -2 |
| 전용면적(㎡) | Numeric | 0 | 0 | Missing Values Imputed | Imputed value: 81.88 |
| 계약년월 | Numeric | 0 | 0 | Missing Values Imputed | Imputed value: 201507 |
| 계약일 | Numeric | 0 | 0 | Missing Values Imputed | Imputed value: 16 |
| 층 | Numeric | 0 | 0 | Missing Values Imputed | Imputed value: 8 |
| 건축년도 | Numeric | 0 | 0 | Missing Values Imputed | Imputed value: 2000 |

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.

DataRobot uses standard modeling techniques to validate model performance and ensure that overfitting does not occur. DataRobot used a robust model k-fold cross-validation framework to test the out-of-sample stability of a model's performance. In addition to cross-validation 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 set aside 19.99996% of the training data as a holdout dataset. This dataset is used to verify that the final model performs well on data that has not been touched throughout the training process.
* For further model validation, the remainder of the data is divided into 5 cross validation partitions. To compensate for the overhead when working with large datasets, DataRobot first trains models on a smaller part of the data and uses only one cross-validation fold to evaluate model performance. Then, for the highest performing models, DataRobot increases the subset sizes. This results in only the best model being trained on the total cross-validation partition. For those models, DataRobot completes 5-fold cross-validation training and scoring. As a result, the mean score of complete model cross-validation is calculated across all folds. Those models that did not perform well will not have a cross-validation score. Instead, because they only had a "one-fold" validation, their score is reported in the Validation column.

The following figure summarizes the CV process used by DataRobot, where the blue denotes 80.00004% of the data available for training, which is then divided into 5-folds for cross-validation and and red denotes the holdout sample.



DataRobot calculates the Cross Validation scores for each of the training data partitions or folds. The project metric used to calculate the score is Gamma Deviance.

6.1.1 Cross Validation Scores

|  |  |
| --- | --- |
| Fold | Cross Validation Score (Gamma Deviance) |
| Fold 1 | 0.00633 |
| Fold 2 | 0.00635 |
| Fold 3 | 0.00654 |
| Fold 4 | 0.00626 |
| Fold 5 | 0.0064 |

6.1.2 Data Partitioning Methodology

Data partitions were selected by means of random sampling.

6.2 Model Performance (Sample Scores)

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

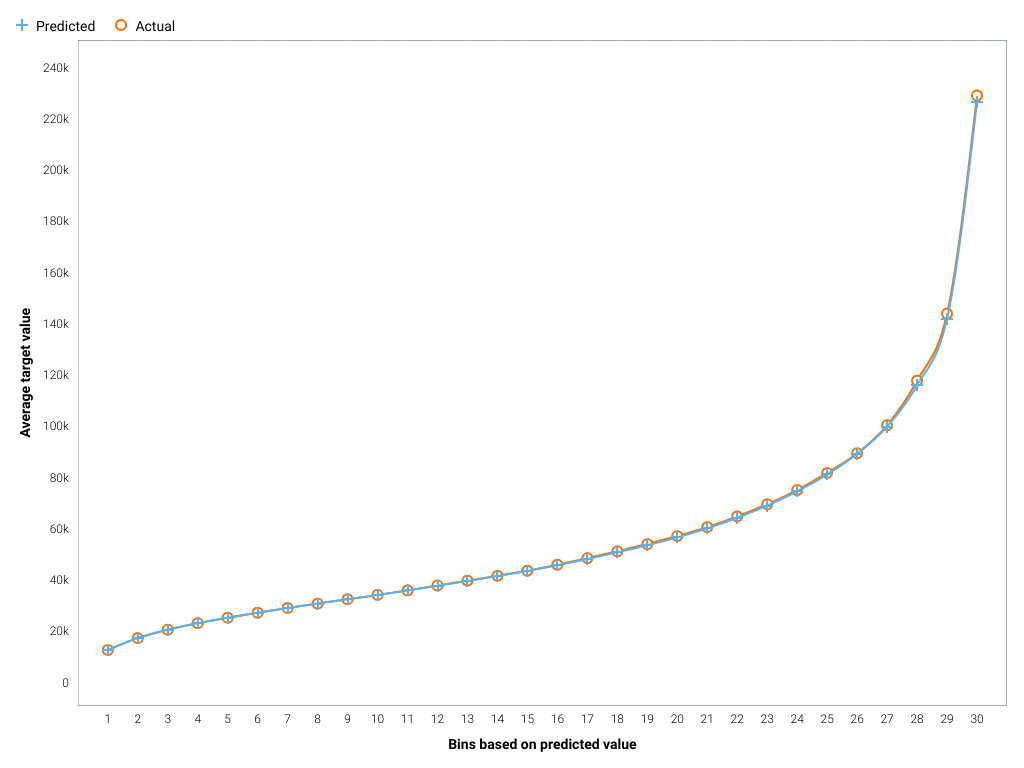
The performance metric used for this project was Gamma Deviance. The model performance results are presented below for in-sample testing:

|  |  |
| --- | --- |
| Scoring Type | Score (Gamma Deviance) |
| cross\_validation | 0.0064\* |
| holdout | 0.0064\* |
| validation | 0.0063\* |

6.3 Sensitivity Testing and Analysis

6.3.1 Lift Chart

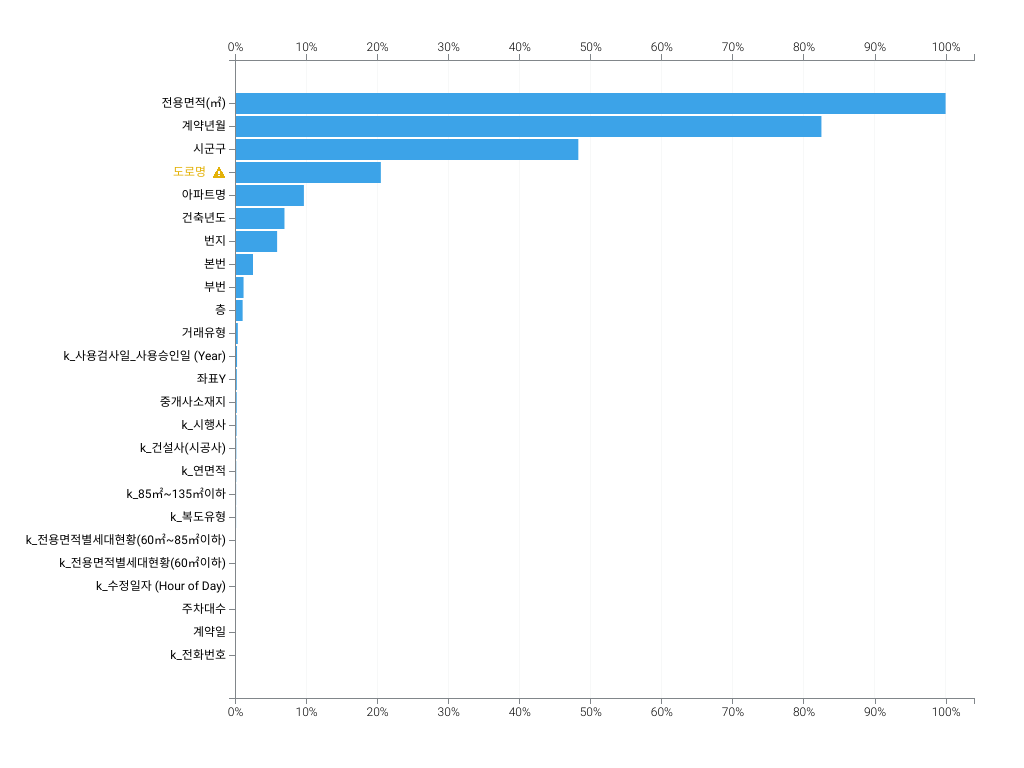
The Lift Chart sorts and groups numeric feature values into equal sized bins, depicting how well a model segments the target population and how capable it is of predicting the target. This helps the user to visualize model accuracy for each bin. The chart is sorted by predicted values -- lowest to highest 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.

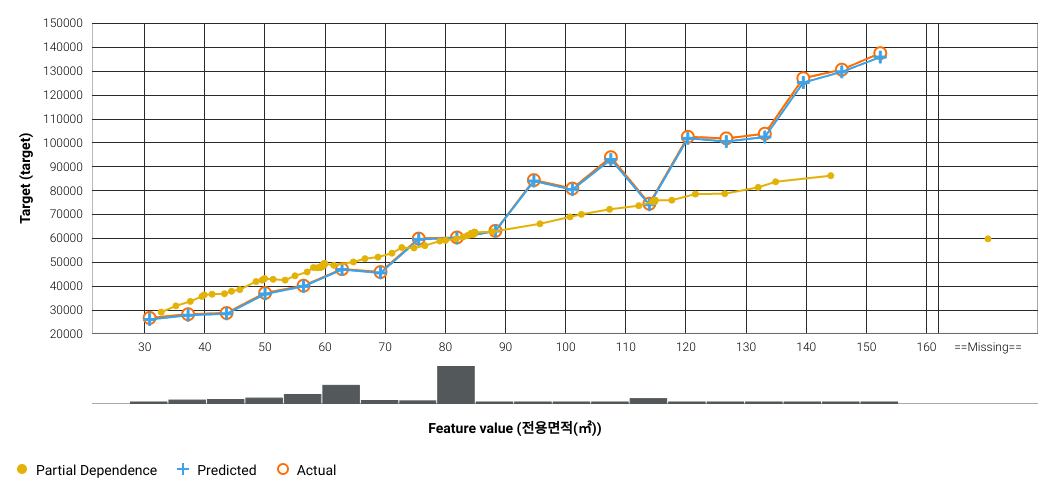


|  |  |  |
| --- | --- | --- |
| Feature Name | Impact Normalized | Impact Unnormalized |
| 전용면적(㎡) | 1.0 | 0.1861 |
| 계약년월 | 0.8252 | 0.1536 |
| 시군구 | 0.4832 | 0.0899 |
| 도로명 | 0.2052 | 0.0382 |
| 아파트명 | 0.0969 | 0.018 |
| 건축년도 | 0.0696 | 0.013 |
| 번지 | 0.0593 | 0.011 |
| 본번 | 0.0253 | 0.0047 |
| 부번 | 0.012 | 0.0022 |
| 층 | 0.0107 | 0.002 |
| 거래유형 | 0.004 | 0.0007 |
| k\_사용검사일\_사용승인일 (Year) | 0.0029 | 0.0005 |
| 좌표Y | 0.0025 | 0.0005 |
| 중개사소재지 | 0.0024 | 0.0004 |
| k\_시행사 | 0.0021 | 0.0004 |
| k\_건설사(시공사) | 0.0019 | 0.0003 |
| k\_연면적 | 0.0017 | 0.0003 |
| k\_85㎡~135㎡이하 | 0.0016 | 0.0003 |
| k\_복도유형 | 0.0015 | 0.0003 |
| k\_전용면적별세대현황(60㎡~85㎡이하) | 0.0014 | 0.0003 |
| k\_전용면적별세대현황(60㎡이하) | 0.0014 | 0.0003 |
| k\_수정일자 (Hour of Day) | 0.0012 | 0.0002 |
| 주차대수 | 0.001 | 0.0002 |
| 계약일 | 0.001 | 0.0002 |
| k\_전화번호 | 0.0009 | 0.0002 |
| k\_전체세대수 | 0.0009 | 0.0002 |
| k\_전체동수 | 0.0008 | 0.0002 |
| 좌표X | 0.0008 | 0.0001 |
| k\_주거전용면적 | 0.0008 | 0.0001 |
| 고용보험관리번호 | 0.0008 | 0.0001 |
| 건축면적 | 0.0005 | 0.0001 |
| 단지신청일 (Month) | 0.0005 | 0.0001 |
| k\_사용검사일\_사용승인일 (Month) | 0.0005 | 0.0001 |
| k\_난방방식 | 0.0005 | 0.0001 |
| k\_사용검사일\_사용승인일 (Day of Month) | 0.0005 | 0.0001 |
| 단지신청일 (Hour of Day) | 0.0004 | 0.0001 |
| k\_관리방식 | 0.0004 | 0.0001 |
| k\_단지분류(아파트,주상복합등등) | 0.0004 | 0.0001 |
| 등기신청일자 | 0.0004 | 0.0001 |
| 단지승인일 (Hour of Day) | 0.0004 | 0.0001 |
| 단지승인일 (Day of Week) | 0.0004 | 0.0001 |
| 단지승인일 (Month) | 0.0003 | 0.0001 |
| k\_팩스번호 | 0.0003 | 0.0001 |
| 단지승인일 (Day of Month) | 0.0003 | 0.0001 |
| k\_관리비부과면적 | 0.0003 | 0.0001 |
| k\_수정일자 (Day of Week) | 0.0003 | 0.0 |
| 관리비 업로드 | 0.0003 | 0.0 |
| k\_홈페이지 | 0.0002 | 0.0 |
| 단지신청일 (Year) | 0.0002 | 0.0 |
| 단지신청일 (Day of Week) | 0.0002 | 0.0 |
| 세대전기계약방법 | 0.0002 | 0.0 |
| 단지승인일 (Year) | 0.0002 | 0.0 |
| 단지소개기존clob | 0.0001 | 0.0 |
| k\_수정일자 (Day of Month) | 0.0001 | 0.0 |
| k\_사용검사일\_사용승인일 (Day of Week) | 0.0001 | 0.0 |
| 청소비관리형태 | 0.0001 | 0.0 |
| 경비비관리형태 | 0.0001 | 0.0 |
| k\_세대타입(분양형태) | 0.0001 | 0.0 |
| 해제사유발생일 | 0.0001 | 0.0 |
| 기타/의무/임대/임의=1/2/3/4 | 0.0001 | 0.0 |
| k\_등록일자 (Hour of Day) | 0.0001 | 0.0 |
| k\_등록일자 (Month) | 0.0 | 0.0 |
| 단지신청일 (Day of Month) | 0.0 | 0.0 |
| k\_등록일자 (Day of Week) | 0.0 | 0.0 |
| k\_수정일자 (Month) | 0.0 | 0.0 |
| k\_수정일자 (Year) | 0.0 | 0.0 |
| k\_등록일자 (Day of Month) | 0.0 | 0.0 |
| k\_등록일자 (Year) | -0.0 | -0.0 |
| k\_135㎡초과 | -0.0 | -0.0 |
| 사용허가여부 | -0.0 | -0.0 |

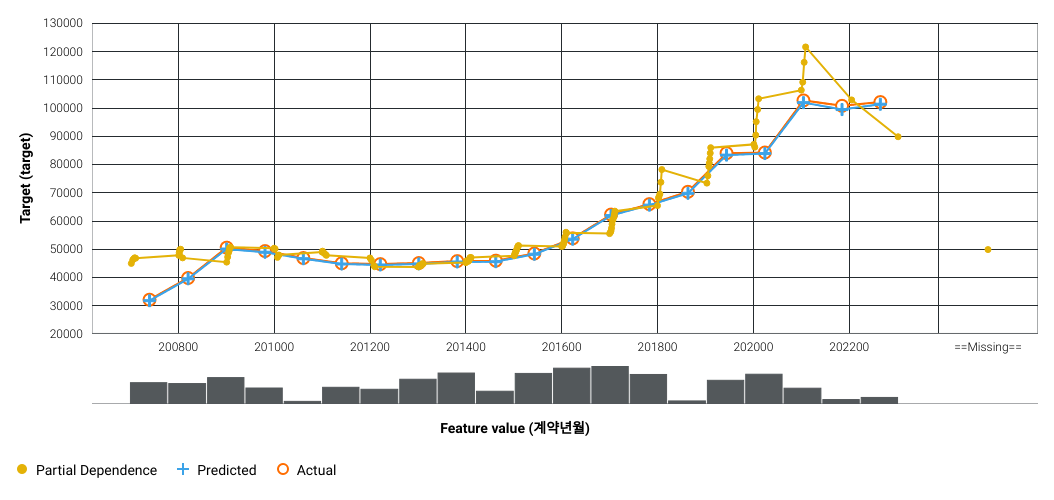
6.3.3 Sensitivity Analysis (Partial Dependence)

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.

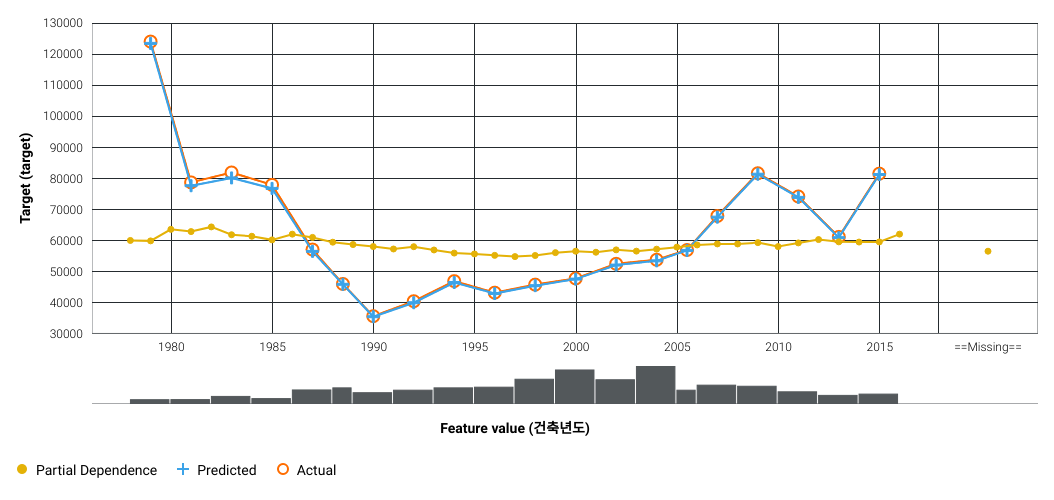
전용면적(㎡)



계약년월



건축년도



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.

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.