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 Adds new images to the dataset by transforming copies of existing images.
* 5.3.2 Grayscale Downscaled Image Featurizer
* 5.3.3 Image Features Post Processor
* 5.3.4 Logistic Regression with L1 or L2 penalty
* 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
* 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 Accuracy
* 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.

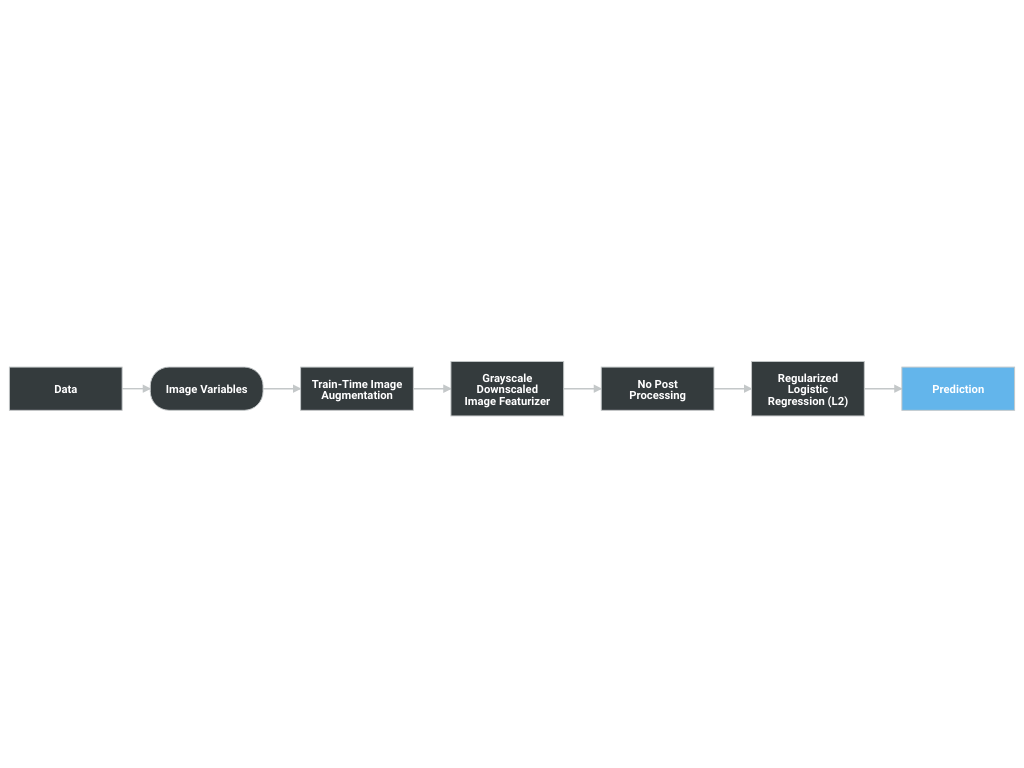
3.2 Model Development Purpose and Intended Use

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: Baseline Image Classifier. This model was developed in a project created with v7bfbfa5a7eba80ba of DataRobot. This model is denoted within DataRobot by the Project ID: 66b5df500edb94ae93e6b679 and the Model ID: 66b8266bbc42391209e6b7d9. The project was created on 2024-08-09 09:20:16.

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:

* Train-Time Image Augmentation
* Grayscale Downscaled Image Featurizer
* No Post Processing
* Regularized Logistic Regression (L2)

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 LogLoss and the project included a total of 1,570 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 (LogLoss) |
| cross\_validation | 1.6307 |
| holdout | 1.5527 |
| validation | 1.5759 |

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

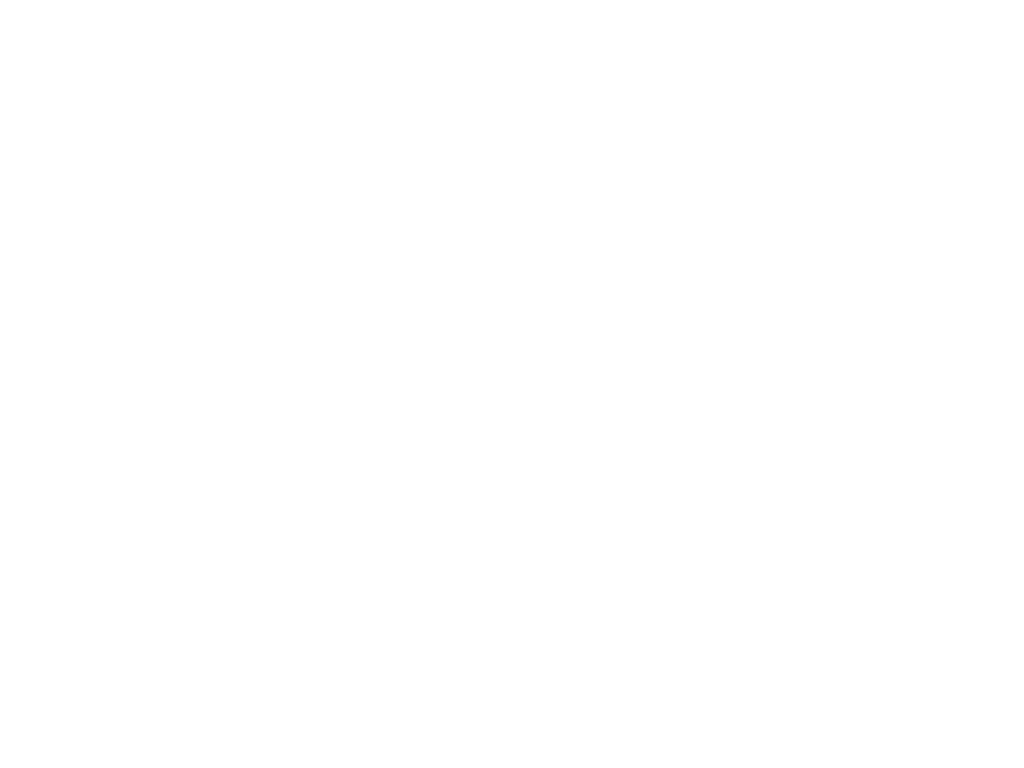
4.1 Feature Association

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:

* Train-Time Image Augmentation
* Grayscale Downscaled Image Featurizer
* No Post Processing
* Regularized Logistic Regression (L2)

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

5.3.1 Adds new images to the dataset by transforming copies of existing images.

This can boost predictions scores by helping to train a more generalizable model which has been exposed to more data. This technique is especially helpful for small datasets, where there is a higher risk of overfitting to the training data. As a tradeoff for seeing more data, this means that the models trained with augmentation will take longer. In addition, depending on the data set, some transformations can hurt prediction scores.

See “https://docs.datarobot.com/en/docs/modeling/special-workflows/visual-ai/tti-augment/index.html” documentation in DataRobot application for more information.

5.3.2 Grayscale Downscaled Image Featurizer

The grayscale downscaled image featurizer module converts an image into a numeric vector which is used as an input feature of a downstream model, such as a logistic regression or a random forest. This “featurization” process downsizes (to 20x20 pixels), grays, and flattens an image to single one-dimensional vector. Any model that uses this featurizer serves as a simple baseline model which will perform poorly most of the time relative to other models. However, if such a baseline model ever does well, then this is an indication that 1) one class of images is consistently darker than another, or 2) one class of images has a similar patch (like a watermark), or 3) there is little / no signal in the data to learn from Regardless, this is a sign that you should inspect your data for quality issues like ones mentioned above.

5.3.3 Image Features Post Processor

Handles all the post processing in one task specifically for image based features from pretrained CNN networks.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| select | pca\_whitening: | the type of pca whitening to use along with its input initialization type values: [`pca\_whitening`, `l2\_norm\_pca\_whitening`, `l1\_norm\_pca\_whitening`, `None`] | None |
| select | scaling\_method: | the scaling method to use before feeding it to further tasks values: [`Standardization`, `Robust\_Standardization`, `L1\_Normalization`, `L2\_Normalization`, `None`] | None |

5.3.4 Logistic Regression with L1 or L2 penalty

Logistic Regression (aka logit, MaxEnt) classifier.

Logistic is a class of generalized linear models that uses the binomial distribution to fit regression models to a binary (0/1) response variable. It is probably the most widely used binary classification models and is a good baseline by which to judge the performance of other classifiers. Logistic regression is notable in that it tends to produce well-calibrated classification probabilities without the need for post-processing.

This class implements regularized logistic regression using the liblinear library, newton-cg and lbfgs solvers. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted (and copied).

The newton-cg and lbfgs solvers support only L2 regularization with primal formulation. The liblinear solver supports both L1 and L2 regularization, with a dual formulation only for the L2 penalty.

Based on scikit-learn estimator sklearn.linear\_model.LogisticRegression

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Description | Best Searched |
| floatgrid | C | Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization. ‘auto’: grid of 50 numbers spaced evenly on a log10 scale from: If penalty is set to ‘l1’ then 1.0 to 1.0e+04 \* the lowest bound for C otherwise 1.0e-02 to 1.0e+02 values: {'floatgrid':[1e-10,1e10],'select':['auto']} | 0.03556480306223128 |
| select | class\_weight | Over-/undersamples the samples of each class according to the given weights. If not given, all classes are supposed to have weight one. The ‘auto’ mode selects weights inversely proportional to class frequencies in the training set. values: ['auto'] | None |
| bool | dual | Dual or primal formulation. Dual formulation is only implemented for l2 penalty with liblinear solver. Prefer dual=False when n\_samples > n\_features. values: [False,True] | False |
| bool | fit\_intercept | Specifies if a constant (a.k.a. bias or intercept) should be added the decision function. values': [False,True] | True |
| float | intercept\_scaling | Useful only if solver is liblinear. when self.fit\_intercept is True, instance vector x becomes [x, self.intercept\_scaling], i.e. a “synthetic” feature with constant value equals to intercept\_scaling is appended to the instance vector. The intercept becomes intercept\_scaling \* synthetic feature weight Note! the synthetic feature weight is subject to l1/l2 regularization as all other features. To lessen the effect of regularization on synthetic feature weight (and therefore on the intercept) intercept\_scaling has to be increased. values: [1e-10,1e10] | 1.00001 |
| select | penalty | Used to specify the norm used in the penalization. The newton-cg and lbfgs solvers support only l2 penalties. values: ['l1','l2'] | l2 |
| int | random\_state | The seed of the pseudo random number generator to use. values': [0, int(1e9)] | 1234 |
| float | tol | Tolerance for stopping criteria. values':[1e-15,1] | None |

5.4 Literature Review and References

* LIBLINEAR – A Library for Large Linear Classification http://www.csie.ntu.edu.tw/~cjlin/liblinear
* Hsiang-Fu Yu, Fang-Lan Huang, Chih-Jen Lin (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning 85(1-2):41-75. http://www.csie.ntu.edu.tw/~cjlin/papers/maxent\_dual.pdf
* Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.
* Harrell, Frank E. “Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.” Springer Science & Business Media, 2013. http://biostat.mc.vanderbilt.edu/tmp/course.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 LogLoss. 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 |
| Regularized Logistic Regression (L2) | 0.0509 | N/A | 0.0482 | 100.0 |
| Keras Slim Residual Neural Network Classifier using Training Schedule (1 Layer: 64 Units) | 0.3961 | 0.4507 | 0.5989 | 71.9746 |

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., LogLoss) 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 is a list of the final set of model feature/independent variables, as well as summary statistics for the Baseline Image Classifier model.

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 |
| ID | Image | 1413 | 0 | N/A | N/A | N/A | N/A | N/A | N/A |
| target | Numeric | 17 | 0 | 8.0 | 4.79 | 8.0 | 0.0 | 16.0 | 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

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 10.0% 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 90.0% 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 LogLoss.

6.1.1 Cross Validation Scores

|  |  |
| --- | --- |
| Fold | Cross Validation Score (LogLoss) |
| Fold 1 | 1.57588 |
| Fold 2 | 1.55334 |
| Fold 3 | 1.69921 |
| Fold 4 | 1.60853 |
| Fold 5 | 1.71644 |

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 out-of-sample records.

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

|  |  |
| --- | --- |
| Scoring Type | Score (LogLoss) |
| cross\_validation | 1.6307 |
| holdout | 1.5527 |
| validation | 1.5759 |

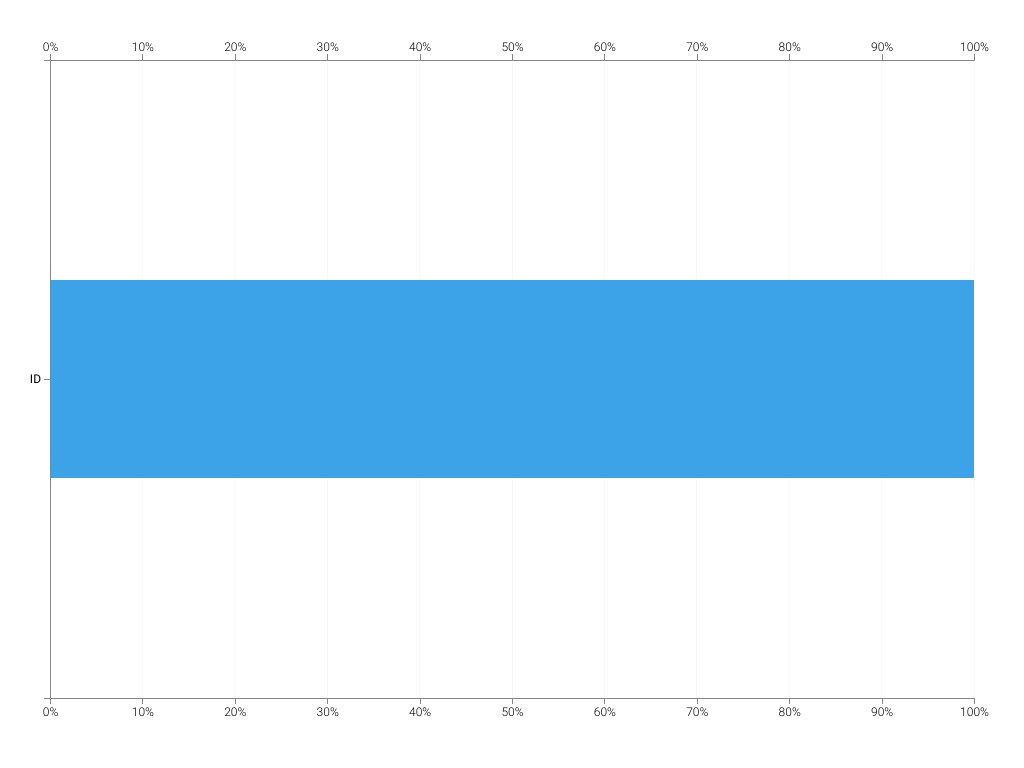
6.3 Sensitivity Testing and Analysis

6.3.1 Lift Chart

This section is not available when multiclass projects have more than 10 classes and are trained on 100% of the data.

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 |
| ID | 1.0 | 2.6504 |

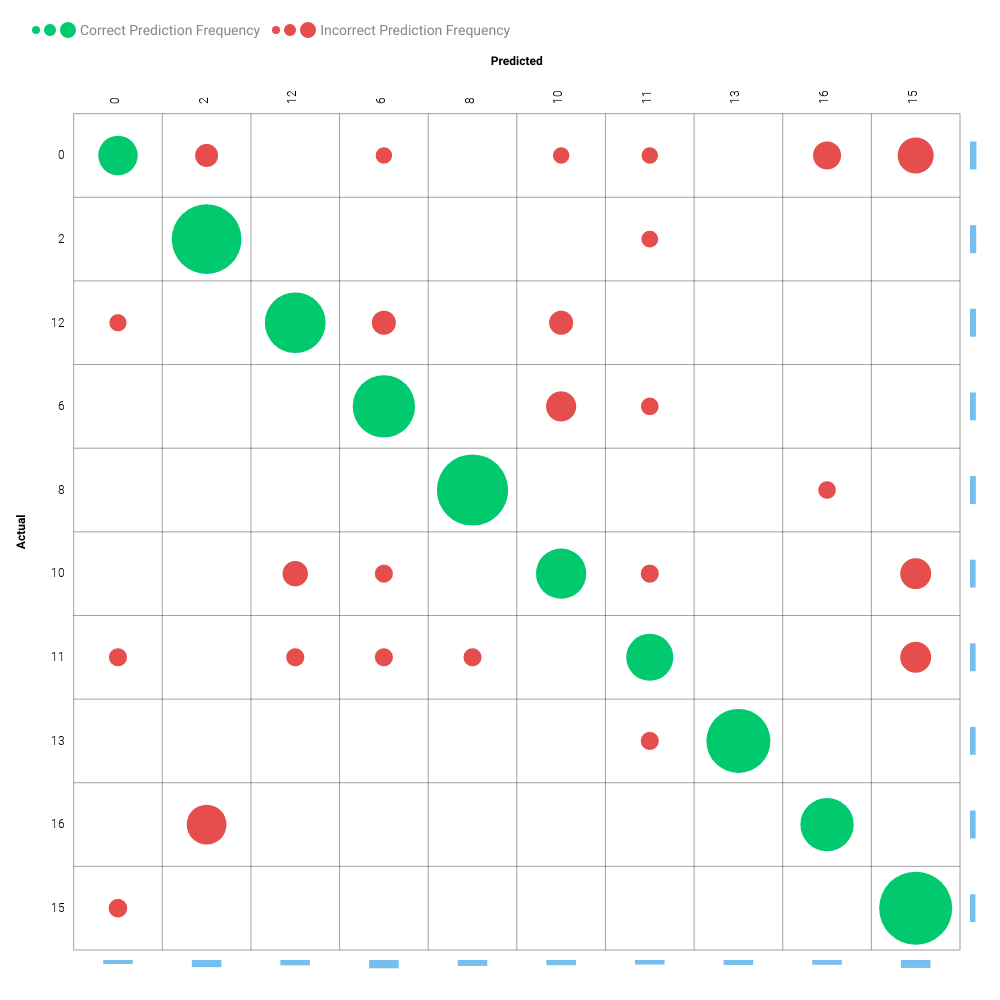
6.3.3 Accuracy

The table below lists, for each class, metrics to indicate how accurate the model was at predicting the class. The measurement is done by comparing the actual class assignment to the model's predicted assignment. Brief descriptions of each accuracy metric follow:

* F1 Score: A measure of the model's accuracy, based on the True Positive Rate and Positive Predictive Value.
* True Positive Rate: The ratio of true positives (correctly predicted as positive) to all actual positives. Also called sensitivity or recall.
* Positive Predictive Value: For all the positive predictions, the percentage of cases in which the model was correct. Also called precision.

|  |  |  |  |
| --- | --- | --- | --- |
| Class Name | F1 Score | True Positive Rate | Positive Predictive Value |
| 0 | 0.7692 | 1.0 | 0.625 |
| 1 | 0.25 | 0.1429 | 1.0 |
| 10 | 0.4 | 0.3 | 0.6 |
| 11 | 0.5556 | 0.625 | 0.5 |
| 12 | 0.5 | 0.5556 | 0.4545 |
| 13 | 0.3333 | 0.25 | 0.5 |
| 14 | 0.0 | 0.0 | 0.0 |
| 15 | 0.4762 | 1.0 | 0.3125 |
| 16 | 0.7 | 0.7 | 0.7 |
| 2 | 0.8696 | 0.9091 | 0.8333 |
| 3 | 0.3529 | 0.2143 | 1.0 |
| 4 | 0.4615 | 0.5455 | 0.4 |
| 5 | 0.8421 | 0.8889 | 0.8 |
| 6 | 0.5517 | 0.6667 | 0.4706 |
| 7 | 0.5 | 0.5714 | 0.4444 |
| 8 | 0.9091 | 0.9375 | 0.8824 |
| 9 | 0.9412 | 0.8889 | 1.0 |

For multiclass models, DataRobot provides a multiclass confusion matrix to help evaluate model performance. A confusion matrix is a visualization that helps you see which classes the model most often confuses with one another. The confusion matrix is the result of comparing the actual class against each of the possible predicted classes. The green points represent instances when the model predicted the correct class while the red points represent when the model was incorrect. The size of the points indicate the number of predictions made for a class. If your model has more than ten classes, only the first ten (based on actual frequency) are shown below, in descending order for the validation partition.



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.