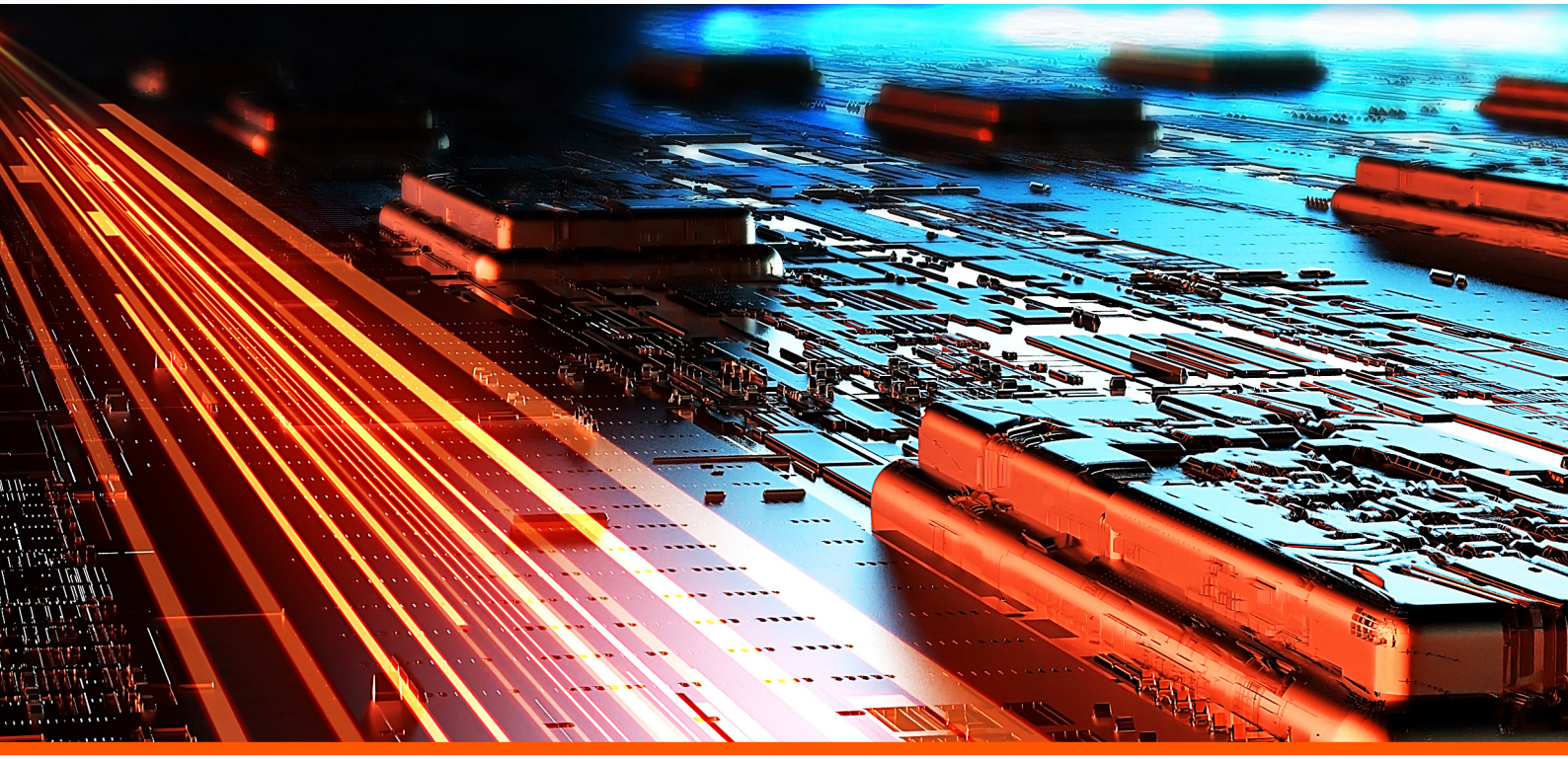


TECHNICAL HIGHLIGHTS 7.0 (MARCH 2021 GA FEATURES)

DataRobot's Automated Machine Learning

DataRobot

The document is a summary of feature highlights of the DataRobot Automated Machine Learning product. This list of features is not an exhaustive list of all the capabilities within the DataRobot Automated Machine Learning product as DataRobot continuously releases new features and functionality. For a comprehensive walkthrough of all capabilities and usage details, please see the DataRobot product documentation.

This document does not cover the features available with DataRobot's Automated Time Series.

Contents:

Modeling Scenarios

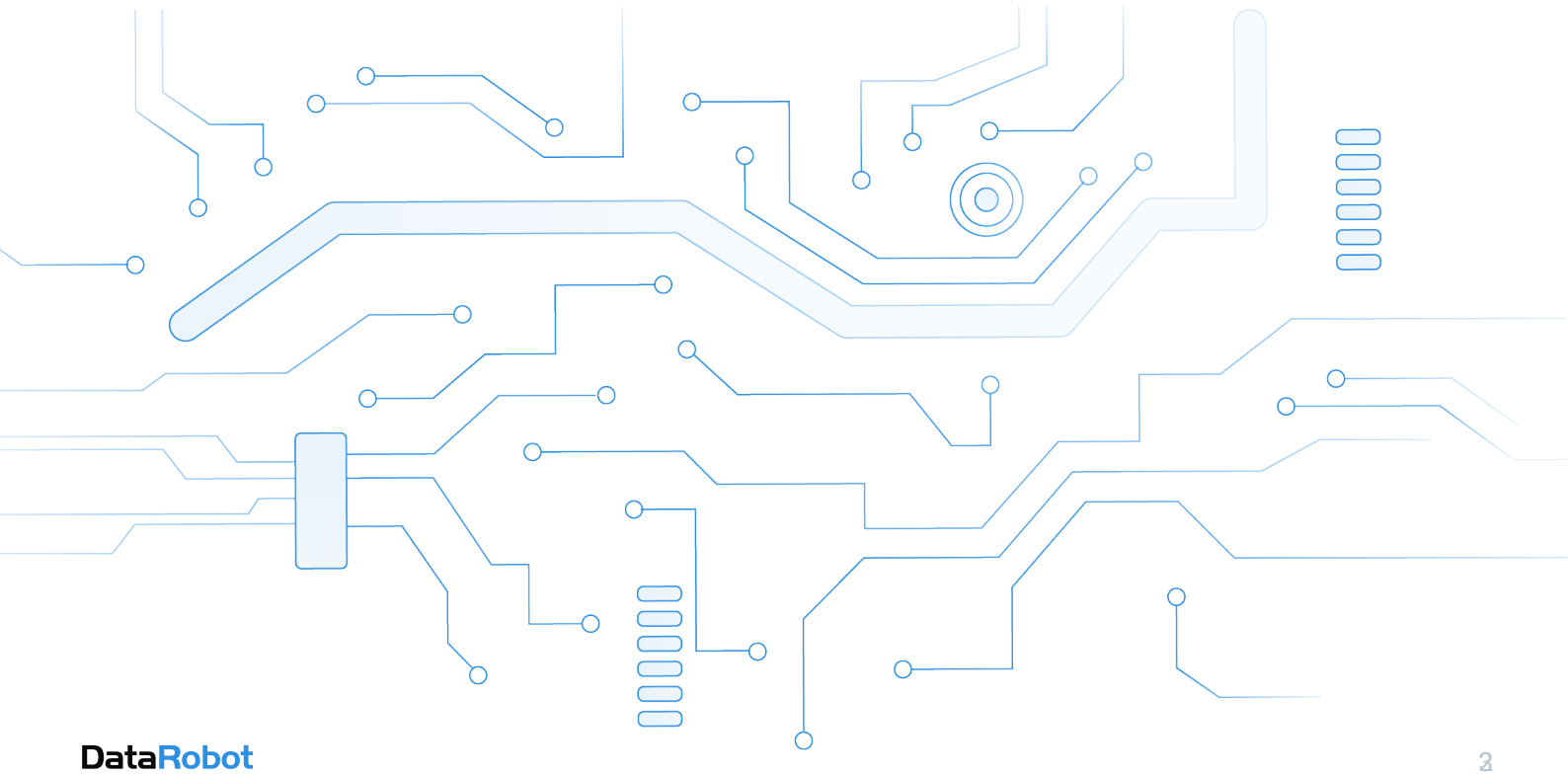
DataRobot supports a variety of supervised machine learning and unsupervised learning. Additionally, DataRobot is capable of multimodal modeling. It's possible to mix different feature types, such as numerical, categorical, text, geospatial, and images in the same model.

Binary Classification	Supports (0/1 or "Y"/"N") problems with 2 classes
Regression	Supports continuous numeric targets
Multiclass	Support up to 100 classes
Anomaly Detection	Supports looking for anomalous rows in a dataset

Modeling API

DataRobot offers an API for modeling. This allows the integration of DataRobot into production modeling pipelines.

Public API (Raw)	Documentation for the Public API is Available within the in-App Documentation.
R	A client package for R is maintained and published to CRAN.
Python	A client for python is maintained and available via pip and conda.



Automated Exploratory Data Analysis (EDA)

DataRobot's approach to analyzing datasets and summarizing their main characteristics.

Automatic Data Schema and Data Type	• Numeric (numerical statistics, mean, standard deviation, median, min, max) • Categorical • Boolean • Text • Special feature types date • Currency • Percentage • Length • Geospatial Points • Geospatial Lines or Polygons • Image
Data Visualization	• Histogram • Frequency distribution for top 50 items • Overtime • Column validity for modeling (non-empty, non-duplicate) • Average value • Outliers • Feature correlation to the target • Geospatial Map
Data Quality Checks	• Inliers • Outliers • Disguised missing values • Excess zeros • Target leakage • Missing Images • Duplicate Images
Feature Association Matrix	Supports numeric and categorical data with metrics: • Mutual Information • Cramer's V • Pearson • Spearman

Modeling Options

DataRobot offers a variety of options for tackling a variety of data science problems. Some of these include partitioning, monotonic constraints, and optimization metrics.

Partitioning Options	• Nested Cross Validation and Train-Validation-Holdout
	  • Random
	  • Stratified
	  • Number of folds and holdout size are configurable
		• Group Partitioning
		• Out of Time Validation/Backtesting
		• Select your own Partitioning

Monotonic Constraints	• Set monotonic constraints which ensure numeric featureseither steadily move up or down.	
Global Optimization Metrics (Weighted Versions are Available as Well)	• Gamma Deviance • Root Mean Squared Error • Mean Absolute Error • Gini Norm • R Squared • Fraction of Variance Explained Gamma • Fraction of Variance Explained Poisson • Fraction of Variance Explained Tweedie • Fraction of Variance Explained Binomial deviance • Fraction of Variance Explained Tweedie • Fraction of Variance Explained Multinomial deviance • Mean Absolute Percentage Error	• Poisson Deviance • Root Mean Log Squared Error • Symmetric Mean Absolute Percentage Error • Tweedie Deviance • Logarithmic Loss • Area Under the (ROC) Curve • Area Under Precision-Recall Curve • Kolmogorov-Smirnov • Maximum of Matthews Correlation Coefficient • Response rate in the top 10% highest predictions • Response rate in the top 5% highest predictions • Balanced Accuracy
Additional Model Specific Optimization Metrics	Keras Blueprint Also Offers: • Hinge • Square Hinge • Logcosh • Kullback_leibler_divergence • Binary_crossentropy • Quantile • Cosine Proximity	Light GBM Also Offers: • Huber Loss • Fair Loss XGBoost Also Offers: • Quantile Loss
Weighting	• Automated and Manual Downsampling • Weights	• Exposure • Offsets
Pairwise Interactions	• Set pairwise interactions to be included in the Generalized Additive Models.	
Target Leakage Detection	• Leaky features will automatically be removed from the modeling feature list.	

Automated Feature Engineering for a Primary Dataset

DataRobot automatically creates and selects features in the modeling data. Preprocessing is a key part of the modeling process and includes dealing with messy data or performing model-specific operations that often increase that model's accuracy. For a given dataset, DataRobot's Autopilot will automatically create 100's-1000's of derived variables in order to enhance model accuracy even further.

DataRobot applies leakage detection safeguards when creating features, for example, not automatically creating a year feature when there is only two years present in the training data.

Numerics and Categoricals	• Missing Imputation (Median, Arbitrary) • Standardization • Search for ratios • Search for differences • Ridit Transform • DataRobot Smart Binning using a second model • Principal Components Analysis	• K-Means Clustering • One hot encoding • Ordinal encoding • Credibility intervals • Category counts • Variational Autoencoder
Date	• Month-of-year • Day of week • Day of year • Day of month	• Hour of day • Year • Month • Week
Text	Character / word ngram encoding Stopword removal Part of Speech Tagging / Removal TF-IDF scaling (optional sublinear scaling and binormal separation scaling) Hashing vectorizers for big data SVD pre-processing Cosine similarity between pairs of text columns (on datasets with 2+ text columns) Support for multiple languages, including English, Japanese, French, Korean, Spanish, Chinese, Portuguese, etc. Tokenizers: • None • Space • Wordpunt • Tweet • Treebank • Japtiny • Mecab • Language (stemmer and stopwords): • English • Danish • Dutch • Finnish	• French • German • Hungarian • Italian • Norwegian • Portuguese • Romanian • Russian • Spanish • Swedish • Japanese • Turkish • Stemmer • None • snowball • lancaster • porter • wordnet • Lemmatizer • None • Wordnet • Spacy

Images	Pretrained featurizers • Resnet50 (Pruned and unpruned) • Xception • Squeeznet • Efficientnet B0, B4 (Pruned and unpruned) • PreResnet10 • Darknet (Pruned and unpruned) • Mobilenetv3 (pruned) Feature pooling: • Average • Max • GEM • sum_with_center_prior • gem-max • avg-max	Feature scaling: • L1 • L2 Standardization / robust standardization All DataRobot image models can also model numeric, categorical, and text data. DataRobot also generates new features at from different layers from the pretrained networks.
Geospatial	• Polygon Area • Polygon Perimeter • Spatial Lags (Nearest Neighbors) • Local Indicators of Spatial Association (LISA) Featurizer • Spatial Kernel Featurizer	

Automated Feature Engineering Across Multiple Datasets

DataRobot can automatically aggregate and join datasets together. In the aggregation process, DataRobot will automatically engineer features from numerical, categorical, date, and text information.

Visual Lineage	• Full visual lineage of how all the features are generated.	
Feature Download	• Ability to download features that are automatically generated.	

Automated Numeric Feature Generation	• Search for
	  • Differences
	  • Ratios
	  • Equals
	• Count
	  • Min
	  • Max
	  • Avg counts
		• Min
		• Max
		• Sum
		• Avg
		• Missing

Automated Categorical Feature Generation	• Most frequent • Entropy • Summarized counts • Unique count • Missing count
Automated Date Features Generation	• Day of week • Day of month • Hour of day • Interval from previous • Time since las duration from creation date • Entropy of date difference • Pairwise data difference
Automated Text Features Generation	• Word/character counts • Summarized token counts

Modeling Algorithms

Very broadly, DataRobot includes a long list of open-source algorithms, including linear and tree-based models along with a few other types that don't fall neatly into those categories.

Generalized Linear Models	• Penalty: L1 (Lasso), L2 (Ridge), ElasticNet, None • Distributions: Binomial, Gaussian, Poisson, Tweedie, Gamma, Huber, quantile • DR Frequency/Severity Models: 2 stage model (Binomial + Gaussian) for zero-inflated regression • Vowpal Wabbit (fast out-of-core linear models with optional polynomial feature search)	
Support Vector Machines	• Penalty: L1 (Lasso), L2 (Ridge), ElasticNet, None • Kernel: Linear, Nyström RFB, RBF • liblinear and libsvm	
Gaussian Processes	• RBF Kernel • Matérn kernel • Rational quadratic kernel	• Exp-Sine-Squared kernel • Dot-Product kernel
Tree Based Models	• Decision Tree (or CART) • Random Forest • ExtraTrees (or Extremely Randomized Forests) • Gradient Boosted Trees (Binomial, Gaussian, Poisson, Tweedie, Gamma, Huber, quantile)	• Extreme Gradient Boosted Trees - XGBoost (Binomial, Gaussian, Poisson) • LightGBM • AdaBoost • RuleFit

Deep Learning Models	• Feedforward Neural Networks • Neural Architecture Search • Deep Residual Networks • Self-Normalizing Neural Networks • Adaptive Learning Networks	• Attention-based text mining networks • State-of-the-art CNN architectures for images • Pre-trained CNN architectures for images
Anomaly Detection	• Isolation Forest • Local Outlier Factor • One Class SVM	• Double Median Absolute Deviation • Mahalanobis Distance • Anomaly Detection Blenders
Text Mining	• Linear n-gram models (character/word n-grams + tfidf + penalized linear/logistic regression) • SVD n-gram models (ngrams + tfidf + SVD) • Cosine similarity between pairs of text columns • Naive Bayes weighted SVM • Stochastic Gradient Descent • FastText • Word2Vec • Attention-based text mining networks	
Other	• K-Nearest Neighbors • DataRobot Eureqa: proprietary, patented genetic algorithm` • Generalized Additive Models (Rating Tables)	

Automated and Manual Model Ensembling

Automatically create ensembles from several models.

Averaging Methods	• Average • Median	
Stacked Modeling Ensembling	• Generalized Linear Model • Elastic Net • Partial Least Square	• Random Forest • Tensorflow • LightGBM

Automated and Manual Hyperparameter Tuning

Every model has different hyperparameters that need to be tuned for good model performance. DataRobot runs a DataRobot designed pattern search that is customized based on the characteristics of the dataset, optimization metric, and algorithm to select the hyperparameters of each algorithm. This method allows for accurate and fast hyperparameter tuning compared to traditional grid search.

The hyperparameter tuning results are available when you expand out a model and go under the 'Advanced Tuning'. In this screen, you can also test out whether other values beyond the grid search would improve the accuracy of a model. Selecting new hyperparameters will create a new model on the leaderboard with details on the hyperparameters selected.

Hyperparameters are available for all models. Every tuned model gets a unique immutable identifier to allow for easy experiment tracking.

Model Evaluation

Quickly and visually evaluate models, blueprints, feature impact, and more.

Learning Curves	See how model performance changes as sample sizes change.	
Speed Versus Accuracy	Visualize the tradeoff between runtime and predictive accuracy.	
Model Comparison	Evaluate models results on metrics like Dual Lift, Lift, ROC, and Profit/Loss Curve.	
Visual Aids for a Specific Model	• Lift Chart • Residuals • ROC Curve • Profit/Loss • Prediction Distribution • Predicted vs Actual	• Cumulative Lift • Cumulative Gain • Confusion Matrix • Accuracy over Time • Stability over Time • Accuracy over Space

Model Interpretability

Easily interpret and explain your models using visual tools and exportable information.

Visual Aids and Interpretability Tools	• Feature Impact (permutation based and Shap) • Conditional Permutation Importance for Location Features • Feature Effects (partial dependence) • Explanations (DataRobot's XEMP and Shap) • Coefficients • Rating Table • Rules (Hotspot/Rulefit) • WordCloud for Text • Tree-based Variable Importance • Embeddings for Images • Activation Maps for Images • Accuracy over Space

Model Assurance

DataRobot incorporates safeguards into all the models it creates. This ensures your models will provide predictions even during times of data drift or change.

Missing Values Resilience	All DataRobot models can accommodate missing values, even if trained with no missing values.
New Levels Resilience	All DataRobot models can accommodate new levels not seen before in the training or testing data.
Automated Feature Selection	DataRobot automatically performs feature selection to evaluate the performance difference for a smaller feature list.
Automated Model Recommendation	DataRobot recommends the best model for your dataset after searching through many candidate models.
Automated Leakage Detection	Once provided a target, DataRobot will highlight and may remove features that are identified as potential target leakage.
Adaptive Training Schedules	DataRobot automatically identifies the optimal training schedule for deep learning models, ensuring more consistent and accurate deep learning model training.
Stacked Predictions	DataRobot creates stacked predictions on training data.
Synthetic AUC	Include the use of a synthetic AUC to rank the performance of anomaly models.

Model Documentation, Validation, and Compliance

DataRobot automates many critical compliance tasks associated with developing a model and, by doing so, decreases the time-to-deployment in highly regulated industries.

Blueprint Documentation	The preprocessing and algorithms for every blueprint are documented. Documentation includes hyperparameter descriptions, settings, open source code (if available), and literature references.
Neural Network Visualizer	The network layers and their sizes are offered in a visualization.
Automated Model Compliance Documentation	Automated documentation provides full insight into how a model is constructed, including the underlying assumptions, which are critical for regulated industries with compliance teams.

In the Datarobot 7.0 release, we also presented the Compliance report template builder. This allows you to create custom compliance reports that suit the specifics of your organizational standards and meet the needs of your regulators. Using the interactive template builder, you can add, remove, and arrange content in blocks to create the documentation structure you need.

DataRobot's Modeling Review and Automated Model Testing Process

DataRobot builds "blueprints", which represent the preprocessing and algorithm steps. Every blueprint is rigorously evaluated before being placed into production. Blueprints are assessed on accuracy across thousands of datasets, as well as other characteristics such as runtime, memory usage, and prediction consistency. For those we choose to incorporate, we have built out a suite of over 20,000 unique tests that run hundreds of times a day to ensure our code accuracy and reliability.

When using third party (and internal) modeling libraries, DataRobot ensures a careful vetting process, from open source license review, vulnerability scanning, build compatibility testing, and finally extensive modeling performance, backwards compatibility tests and validation testing before any changes are approved and added to the product. In this way, DataRobot is able to deliver a vast array of modern open source machine learning tools to diverse environments while guaranteeing consistent, reliable results, enterprise-grade security, and strict regulatory compliance.

Automatic Bias and Fairness Testing

Automatic bias and fairness testing is now generally available and includes a new cross-class accuracy insight. Bias and fairness testing allows you to flag protected features in your dataset and then actively guide you by selecting the best fairness metric to fit the specifics of your use case. Once your models are built, we then surface visual insights to illustrate the results of the selected bias and fairness test. If bias is identified, you can use the Cross-Class Data Disparity tool to perform root cause analysis, diagnosing the source of bias in your data and ultimately directing you towards mitigation steps in your data collection or processing.

Deployment Options

Easy and flexible model deployment based on your needs and ensure that models created can easily be put into production and deliver value.

REST API	Automated and containerized API deployment ensures the mobility and scalability of your model, easily slipstreaming in any of the heterogeneous systems your business depends upon.
Standalone Scoring Engine	Separate staging and production environments so that models can be tested and implemented in a stable, isolated environment. The Standalone Engine has the capability to run imported models without ever touching the development server from which they were exported.
Code Export (Exact)	Scoring code allows for the exporting a java jar file of a model.
Code Export (Approximate)	DataRobot Prime allows the creation of a java/python approximation mode which allows for a transparent exportable model.
Spark Scoring	Spark Scoring with DataRobot allows enterprises to score data for machine learning where it is located, eliminating the need to transfer and host that data on a central server.
Prediction Write Back or Export	Write back predictions into a database, or export predictions or feature impact, and feature effects.
Supported Integrations	Snowflake, Tableau, Microsoft SQL Server, Qlik

Model Monitoring and Management

Ensure the health of your models once in production

Model Health	Automatically quantify model stability over time to understand if/when performance may be changing.
Model Monitoring and Alerting	Monitor data drift and accuracy over time.
Prediction Intervals	Monitor if predictions fall outside of predetermined bounds.
Model Deployment Inventory	Comprehensive overview of all models in production.
Service Health	Track model-specific deployment latency, throughput, and error rate.

Collaboration

Break down AI and ML silos, bring teams together to build high-impact ML and AI faster.

Sharing to Remove Silos	Users can promote knowledge sharing and collaboration across users, groups, or their entire organization--allowing them to generate more value from their AI projects using DataRobot. Maintain full control over what other collaborators can and cannot do with your shared object in DataRobot with built in roles and permissions on the shared object.
Notification Policies	Enforce policies for when specific users / tools need to be notified / alerted for DataRobot specified events. Know and fix an issue before your application users do.
Comments, Tags, and Notifications	Users can comment and tag other users on a particular Use Case, dataset in a Catalog, and model on the leaderboard to increase knowledge share and collaboration throughout their ML workflow.

Governance and Security

Single sign-on	Boost productivity of a workforce and IT department by providing seamless, secure access to and scalable management of apps, including DataRobot. **** Not available on our Managed AI Cloud at the moment.**
Multi-factor Authentication	Two-step verification process that helps safeguard access to DataRobot while maintaining simplicity for users. It provides additional security by requiring a second form of authentication and delivers strong authentication.
Built-in Roles for DataRobot	DataRobot has several built-in roles that you can assign to users, groups, or the entire organization. Role assignments are the way you control access to the different areas of the DataRobot platform.
Automated User Provisioning	DataRobot enables you to automate the creation, maintenance, and removal of user access by streamlining access control based on role, department, location, title, and other attributes. Changes in Active Directory are synchronized to tools like DataRobot within seconds.
Audit Logs	Help security teams maintain audit trails in DataRobot by providing visibility over user activity in DataRobot with our Activity Monitor. This maintains audit logs for administrative, modeling / deployment, and prediction activity.
Self-administer Users for Your Account	Reduce the time to manage DataRobot user accounts in your organization by leveraging DataRobot's self-administration capabilities.
Dynamic Configuration of SMTP Settings	Configuring SMTP settings for email notifications can be accomplished through the GUI rather than editing configuration files.
Approval Workflows	Ability to set priority levels for deployment and allow review of deployments.

Data Management: Data Ingest, AI Catalog, & Data Preparation

DataRobot' AI Catalog is a centralized collaboration hub for working with data and related assets. The AI Catalog can be connected to multiple data sources and allows you to seamlessly find, share, tag, and reuse data as well as execute simple data preparation. Data assets within AI Catalog can either be snapshots of tables/files or be queried dynamically from your data source. If the data is Snapshotted, those snapshots can be automatically refreshed periodically, and are also automatically versioned to preserve lineage and enhance the overall governance capabilities of DataRobot.

DataRobot has tested support for the following:

Data Connections	Databases & Data Lakes: HDFS*, Hive, Microsoft SQL Server, Elastic Search, Intersystems, KDB+, SAP HANA, MySQL, Oracle, PostgreSQL, AWS Athena, Amazon Redshif, Snowflake, Presto, TD-Hive, Google BigQuery, Google Cloud Storage, Azure Blob Storage, or Amazon S3. Generic Data Connectors: Import Data from URL HTTP, HTTPS, JDBC** *If deployed on Hadoop **If non-SaaS
File Types	.csv, .dsv, or .tsv, database tables, .xls, .xlsx; compressed formats .gz .bz2, .tar, .zip, .tar.gz/.tgz, .tar.bz2
Data Ingest Limitations	With the DataRobot managed cloud environment, dataset size is limited to 10 GB uncompressed. Larger ingest is available when installed within on-premise installations (physical hardware or VPC).
Data Prep	DataRobot Data Preparation SQL pre-processing Spark SQL blending