WHITE PAPER

gathr

Build and operationalize machine learning models

Seamless training, testing, scoring, and model management

Introduction

Data science and machine learning (ML) have gained prominence in the enterprise digital transformation journey. ML automates decision-making and analytical reasoning capabilities. It leverages historical datasets from past experiences to give enterprises a competitive advantage for -

- Developing products
- Enhancing services
- Improving operational efficiency
- Adding certainty to decisions
- Predicting outcomes
- Optimizing ROI

Enterprises use ML for its varied advantages like:

1. High-speed data processing

As the machine learns and develops unique methods to solve a problem, it can process huge amounts of incoming data in seconds and detect data patterns that are not visible to the human eye.

2. Continuous improvement for real-time insights

Since the machine consistently learns from exposure to bigger datasets and scenarios, the accuracy of ML models improve with time. This continuous improvement helps enterprises gain real-time predictive and prescriptive insights. For example, AI-based conversational bots learn from users' questions, which improves their quality and accuracy with time.

3. Consistent decision making

ML-based decisions are consistent as they are based on training data. Machines take previous learnings, associated factors, and datasets into consideration before sharing insights and recommendations.

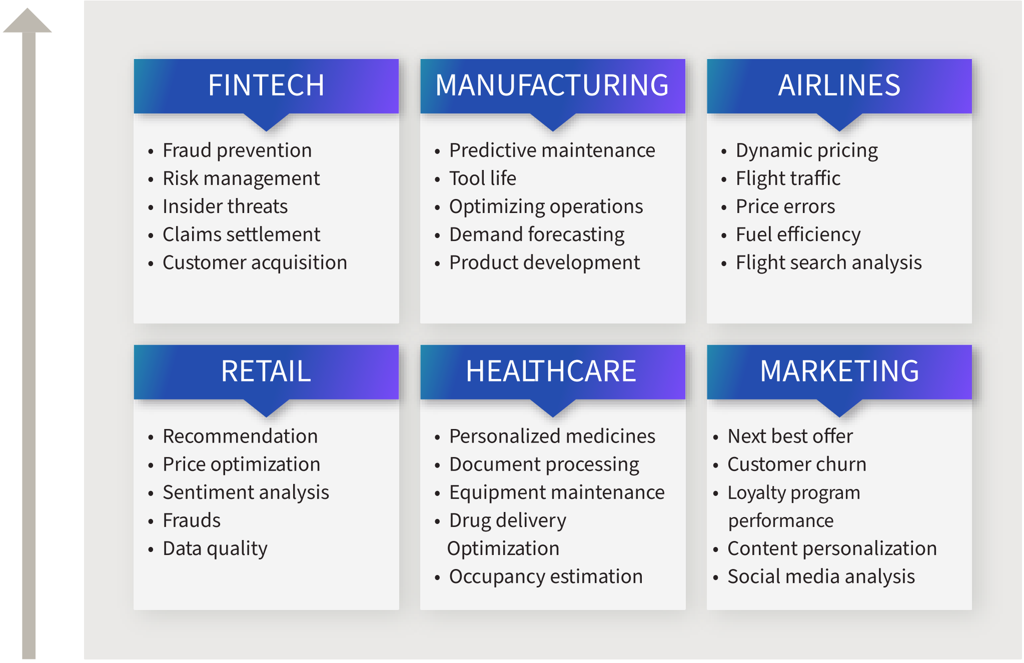
4. Reusability and extensibility

Once trained, ML models can be simultaneously used in an extensible way by multiple applications that need the same decisions.

5. Better risk management

ML models log all records and activities with underlying factors and cross-reference them with flagged variables. Therefore, in case of any error, users can retrace their steps for easy rectification, which helps enterprises manage risks.

ML models are used across industries like:



FINTECH

- Fraud prevention
- Risk management
- Insider threats
- Claims settlement
- Customer acquisition

MANUFACTURING

- Predictive maintenance
- Tool life
- Optimizing operations
- Demand forecasting
- Product development

AIRLINES

- Dynamic pricing
- Flight traffic
- Price errors
- Fuel efficiency
- Flight search analysis

RETAIL

- Recommendation
- Price optimization
- Sentiment analysis
- Frauds
- Data quality

HEALTHCARE

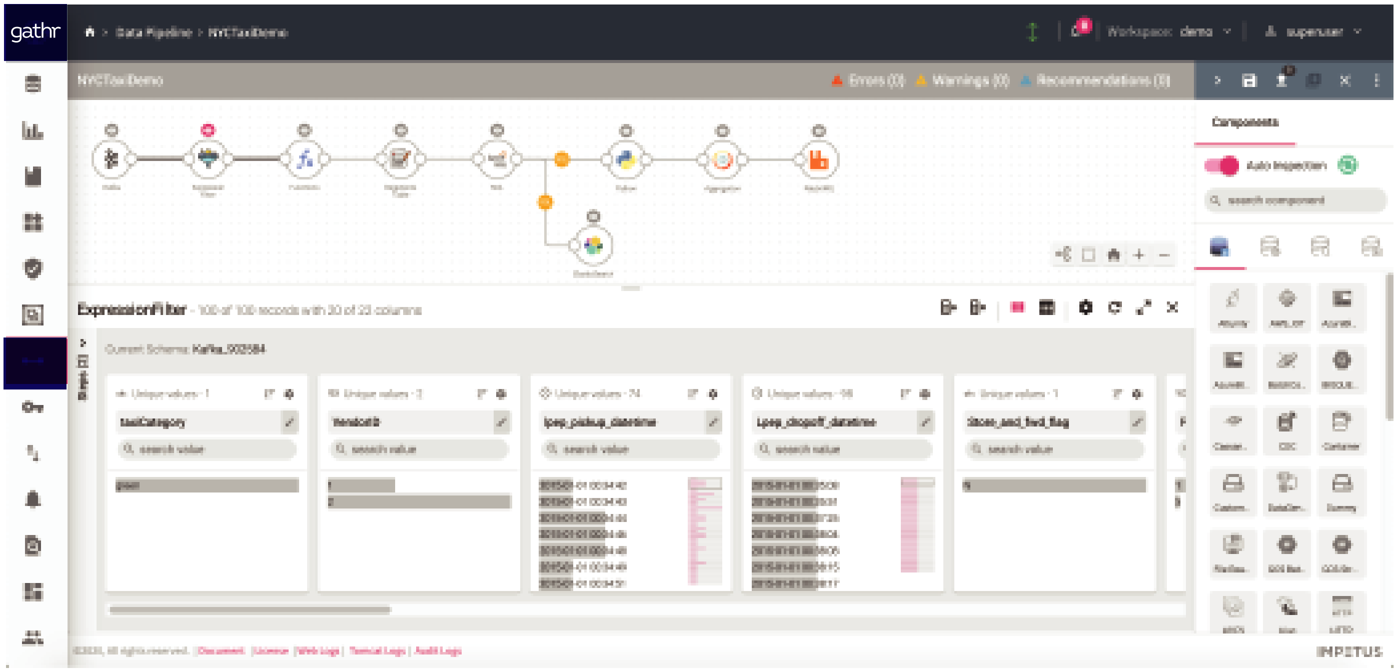
- Personalized medicines
- Document processing
- Equipment maintenance
- Drug delivery Optimization
- Occupancy estimation

MARKETING

- Next best offer
- Customer churn
- Loyalty program performance
- Content personalization
- Social media analysis

Build and operationalize ML models with a self-service data flow and analytics platform

You can build and operationalize production-grade robust applications with ML capabilities using Gathr. Gathr is a self-service ETL and analytics platform that helps create batch and streaming ETL pipelines using drag-and-drop operators on a visual IDE. It has a wide array of built-in operators for data sources, data preparation, data wrangling and transformation, and machine learning.

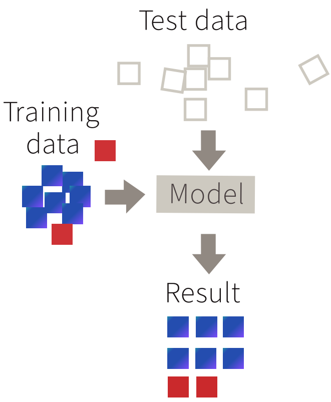


Gathr interface

Gathr helps you get the maximum value out of your data, maintain greater consistency within your data streams, and join streams with static data sources more efficiently.

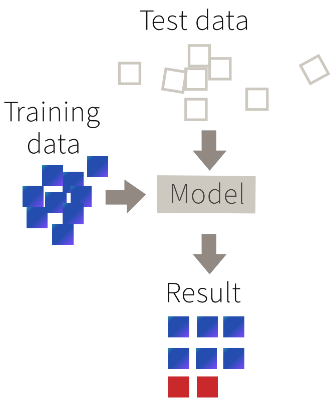
Gathr supports all major modeling techniques like:

Supervised learning



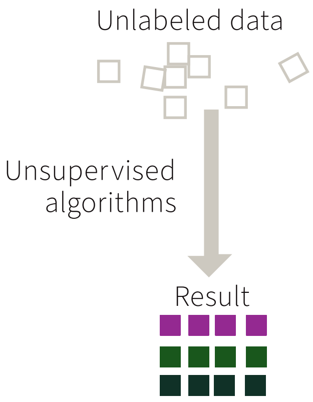
The model is trained with the help of a labeled dataset and learns the outcome based on the associated labels

Semi-supervised learning



The model is trained with a partially labeled dataset, which may be used to label the unlabeled data points

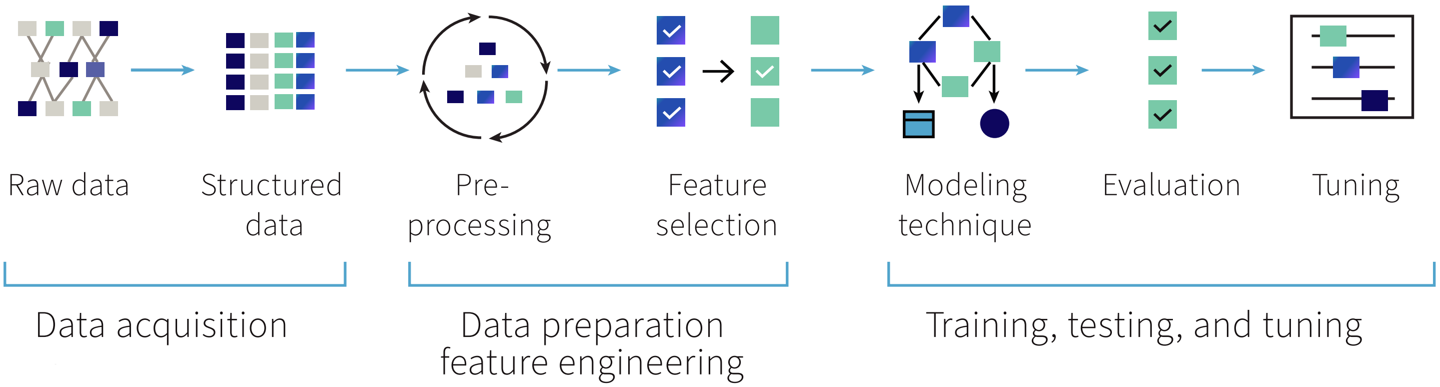
Unsupervised learning



The model is not provided with any labels and identifies patterns and groups them based on the similarity of features

Training and deployment of ML models

Gathr has a robust set of tools that enables powerful ML operations. It comes loaded with libraries and pre-built operators for effectively training ML models and deploying them on data pipelines.



Data acquisition

The quality of the acquired data determines the accuracy of an ML model. Therefore, acquiring the right data is crucial. Gathr lets you connect to a wide range of streaming and batch data sources for performing various data wrangling activities. Running data quality checks, identifying outliers, and enriching incoming data are common patterns that data scientists often spend their time on it. Gathr helps to increase productivity by providing out-of-the-box features and operators in a self-service model.

Data preparation and feature engineering

Data scientists prepare their data before models can consume it by:

- Scrubbing missing values
- Editing the columns with proper names
- Enriching the data by adding new sources
- Masking PII data
- Identifying data quality issues
- Eliminating outliers

Gathr has advanced data preparation capabilities with an intuitive UI to tackle data preparation challenges. Moreover, its wizard-based UI let you seamlessly handle feature selection and advanced pre-processing steps like imputation, binning, one-hot encoding, scaling, etc.

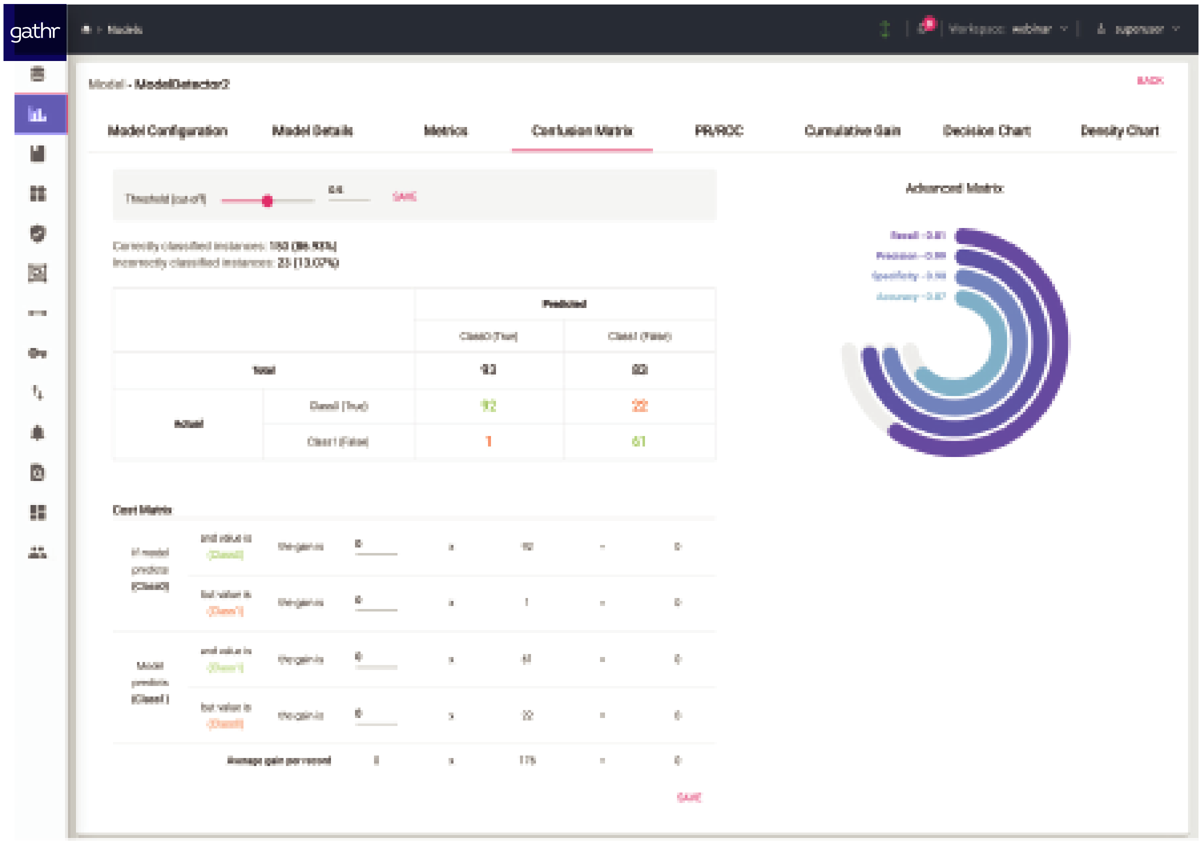
Training, testing, and tuning

A great performing model that strikes the right balance between complexity, performance, and accuracy needs multiple iterations. At first, you need to identify the type of algorithm required to run through the use case. Once identified, you can utilize the training capabilities of Gathr to train on your choice of technology.

You can select the best ML models with Gathr by tuning hyperparameters. The platform performs various permutations and combinations on the provided hyperparameters' values to suggest the best model based on performance.

Model performance

Gathr lets you visualize key performance metrics to choose the best model that can deliver high accuracy. It also helps you consider various model dimensions with multiple views to understand performance effortlessly.



Model - ModelDetector2

BACK

Model Configuration | Model Details | Metrics | Confusion Matrix | PR/ROC | Cumulative Gain | Decision Chart | Density Chart

Threshold (cut-off) 0.6 SAVE

Correctly classified instances: 153 (86.93%)
Incorrectly classified instances: 23 (13.07%)

		Predicted	
		Class0 (True)	Class1 (False)
Total		93	83
Actual	Class0 (True)	92	22
	Class1 (False)	1	61

Advanced Matrix

Recall - 0.81
Precision - 0.99
Specificity - 0.98
Accuracy - 0.87

Cost Matrix

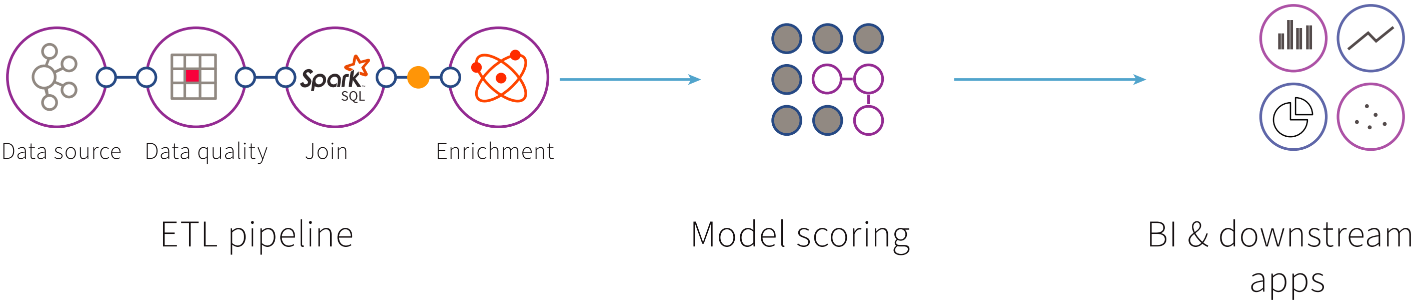
If model predicts (Class0)	and value is (Class0)	the gain is	0	x	92	=	0
	but value is (Class1)	the gain is	0	x	1	=	0
Model predicts (Class1)	and value is (Class1)	the gain is	0	x	61	=	0
	but value is (Class0)	the gain is	0	x	22	=	0
	Average gain per record		0	x	175	=	0

SAVE

Model performance with Gathr

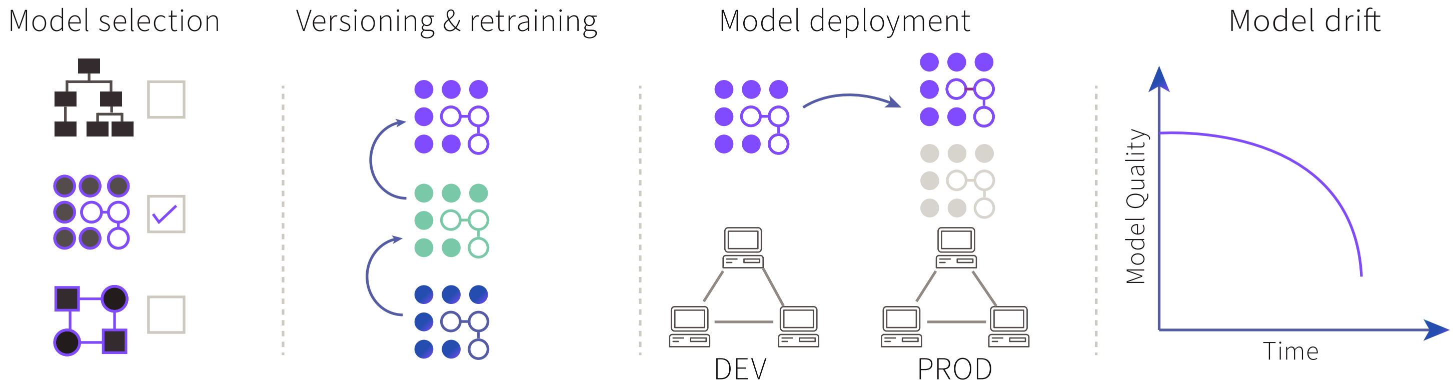
Model scoring

After choosing the best model, you can add models to scoring pipelines using Gathr's drag-and-drop feature, which lets you choose the algorithm and the model you want to deploy from a list of trained models.



Model management

Building production-grade models to predict outcomes consistently over a long time requires efficient model management capabilities. Model management involves the following steps:



Model selection

Selecting the right model is essential to solving unique and critical real-world problems. Model selection requires experimentation and the ability to consider various performance parameters of each trained model. Gathr streamlines this process by displaying performance metrics over various trained models, which can be utilized to choose the best fit model that can also be trained simultaneously.

Versioning and retraining

Typically, it takes multiple iterations to get a perfect production-ready model in place. To simplify version management, Gathr offers a powerful visual versioning and retraining mechanism that allows you to easily roll back to a previous version.

Model deployment

Model deployment on huge volumes of data in production is a tedious process, as each model needs to be deployed in a distributed manner. With Gathr, you can activate a new model in your scoring pipelines in a single click with no downtime.

Model drift

The prediction quality of models deteriorates with time, as data characteristics and parameters on which predictions are based change. Therefore, it is important to define criteria and monitor continously to detect any drift. Gathr comes bundled with powerful drift detection capabilities, which let you specify these criteria and get alerts in case of any drift.

Operationalizing models

Gathr can operationalize and manage the entire lifecycle of data science models – trained within the platform or using any other technologies like Python, R, KNIME, or RapidMiner, which cannot be executed in a distributed environment.

Model-as-a-service

Once built, multiple applications can use a model. However, in the absence of a robust deployment strategy, organizations often fail to reuse models. Gathr lets you migrate your stand-alone data science models to an application exposed as a REST service, which can be accessed and utilized by multiple teams across use cases.

Scaling models

Although models can handle certain loads, their scalability depends on the system's limit on which the models are deployed. Gathr can distribute your models on a Spark cluster for linear scalability. You can also use Kubernetes to scale models that you want to expose as a REST endpoint.

Auditing

Once you have a stable model in place, Gathr can capture all user actions inside a project and track changes.

Model reproducibility

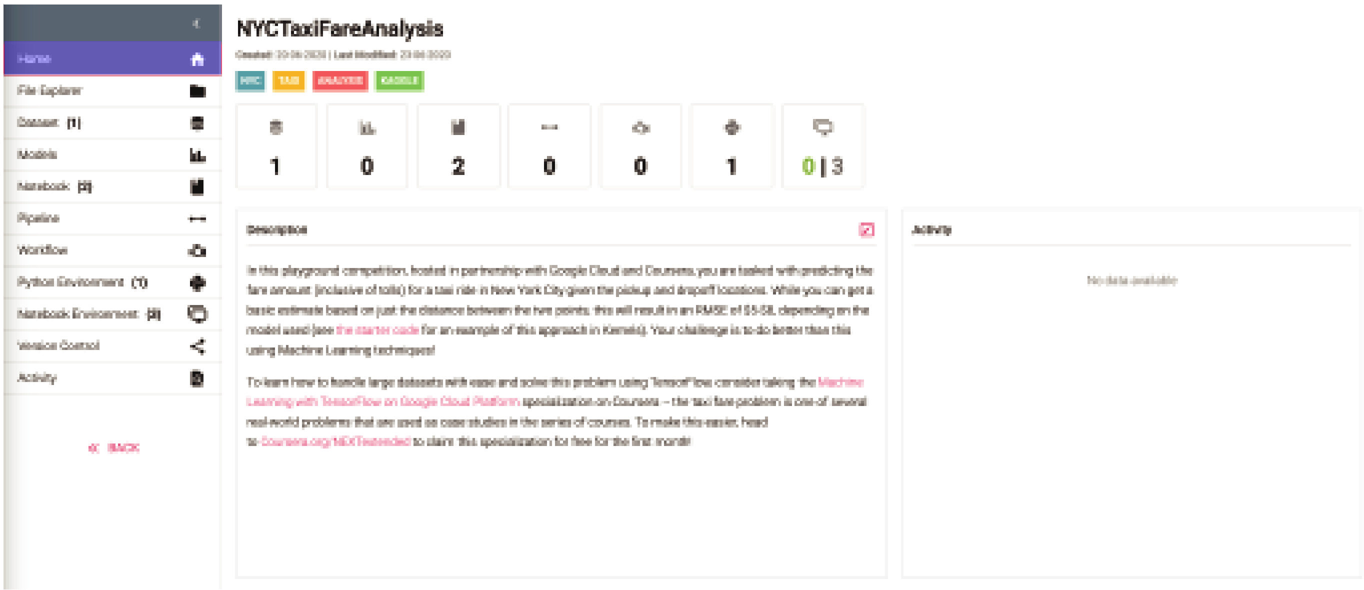
When new models are trained and refreshed on the production environment, Gathr can reproduce past training data and models. However, for compliance, you may need to completely recreate an environment from the past.

Additional platform capabilities

Gathr also has multiple features to help data scientists build models faster and manage them in an enterprise setup. Some of the capabilities are as follows:

User workspace

Each project can have its individual workspace, where all utilized data sources, code, pipelines, and clusters can be stored.



In this playground competition, hosted in partnership with Google Cloud and Coursera, you are tasked with predicting the fare amount (inclusive of tolls) for a taxi ride in New York City given the pickup and dropoff locations. While you can get a basic estimate based on just the distance between the two points, this will result in an RMSE of $5-$8, depending on the model used (see the starter code for an example of this approach in Kernels). Your challenge is to do better than this using Machine Learning techniques!

To learn how to handle large datasets with ease and solve this problem using TensorFlow, consider taking the Machine Learning with TensorFlow on Google Cloud Platform specialization on Coursera -- the taxi fare problem is one of several real-world problems that are used as case studies in the series of courses. To make this easier, head to Coursera.org/NEXTextended to claim this specialization for free for the first month!

Model performance with Gathr

Datasets

Data scientists need a data source to build models. Gathr lets you register your source once, which is then available to use in the models. You can update, prepare, and cleanse the original data source, which is automatically saved for future use.

Programming environment

Gathr allows you to create multiple self-sufficient stand-alone programming environments to execute codes. For example, you can run one environment on Python 2.7 and another on Python 3.6. All packages are managed and are specific to the environment in which they are installed.

Wizard-based training

Gathr supports a wide range of algorithms for classification, clustering, and regression analysis. Users can visually drag-and-drop an algorithm and train it by using an interactive UI.

In-built notebook

Developers who want to leverage notebooks for training their models can use the platform’s notebook interface. Models trained in these notebooks are automatically registered and can be further utilized in scoring pipelines.

Integration with GIT

All datasets, notebooks, models, workflows, tags, and versions of your work can be synced and managed with GIT.

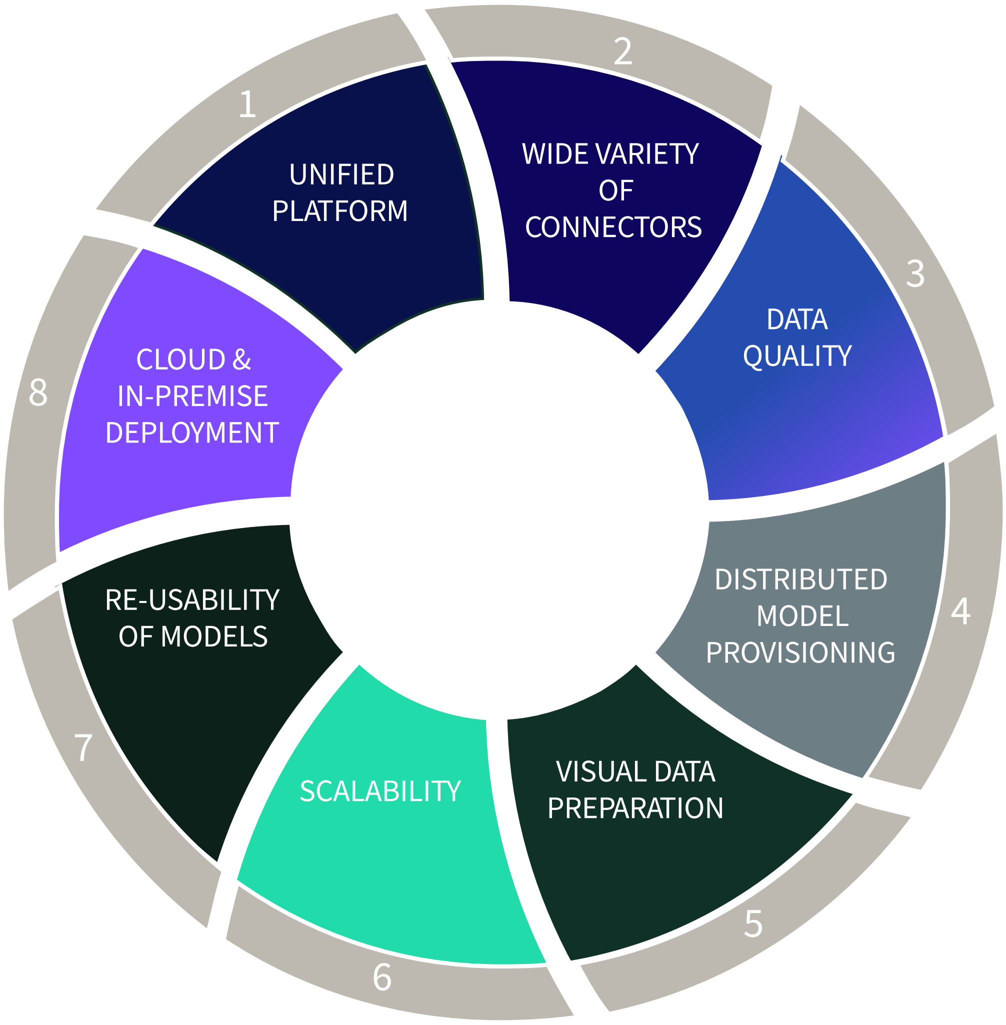
Training models using external technology

Many developers also prefer developing and training models using the technology of their choice, such as Python, R, H2O, etc. But many organizations face challenges when these models are taken to production for processing large datasets. With Gathr, you can import models trained externally and deploy them for scoring in a distributed environment.

Resource sharing

When multiple data scientists use the same infrastructure, fair distribution of resources across use cases often becomes a challenge. Gathr lets you launch clusters with predefined memory and cores to ensure proper resource sharing

Benefits of Gathr – an enterprise-grade ML and data analytics platform



www.gathr.one