impact analytix

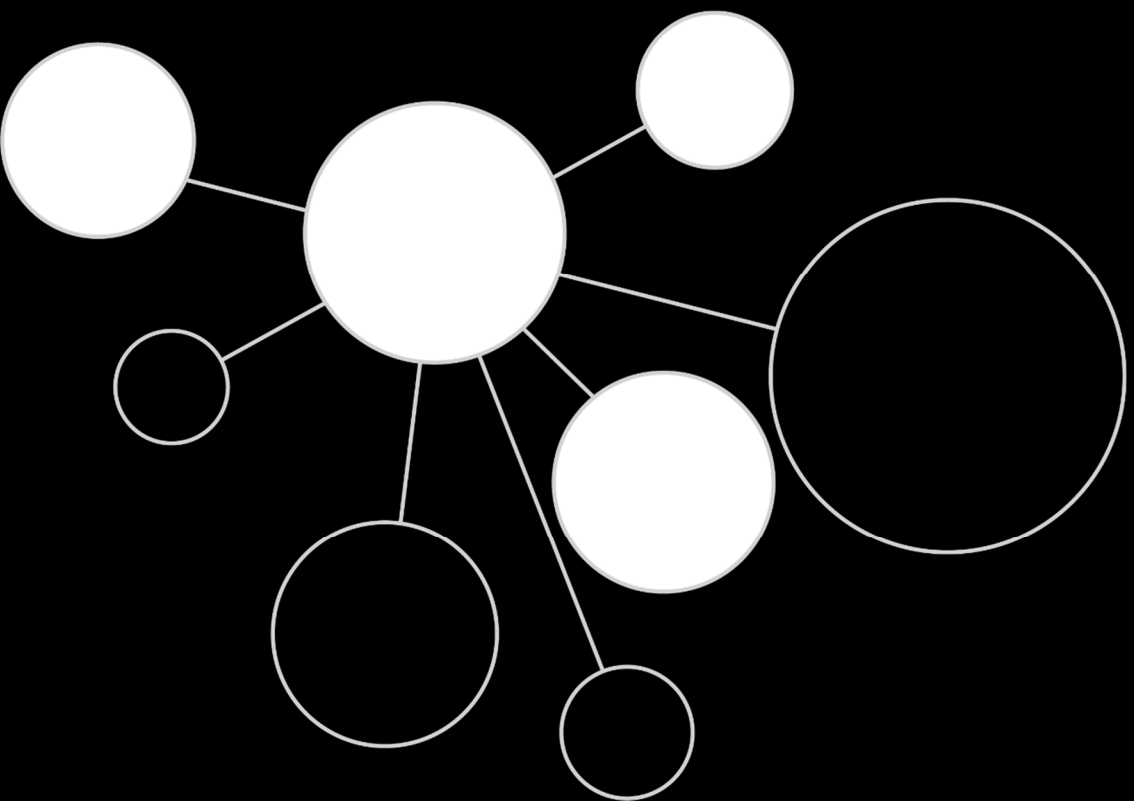
WHITE PAPER

MOVING FROM BI TO MACHINE LEARNING WITH AUTOMATION

BY JEN UNDERWOOD

TABLE OF CONTENTS

NOVEMBER 2017 – WHITE PAPER COMMISSIONED BY DATAROBOT



INTRODUCTION

Growing global demand for data scientists is far exceeding the expensive, scarce supply of qualified talent. Companies today use data scientists to find hidden insights and maximize the value of collected data by applying machine learning algorithms. To bridge the gap between demand and supply, automated machine learning is empowering a new class of data scientist – the citizen data scientist.

Gartner defines a citizen data scientist as a person who creates advanced diagnostic, predictive or prescriptive models whose primary role is outside the field of statistics or analytics.[1] Many data analysts and business intelligence professionals will likely progress into citizen data science roles. Automated machine learning is an ideal solution for them to take that leap.

With automated machine learning, you don't need to be a data scientist to get immediate predictive insights. If you use tools like Tableau, Qlik, TIBCO Spotfire or Excel, understand your business and have data, you can easily advance your career.

The next fundamental shift in the evolution of analytics is already happening. Next generation automated analytics technologies are disrupting current analytics approaches. Similar to the second wave of self-service analytics impacting the first wave of business intelligence, the third wave will change the game once again.

"Early adopters of machine learning automation tout unprecedented speed to insight and enhanced competitive advantage."

Early adopters of automated machine learning tout unprecedented speed to insight and enhanced competitive advantage. Gartner predicts in 2021, augmentation will create $2.9 trillion dollars of business value and $6.2 billion hours of worker productivity.[2]

Do not fear analytics automation. These solutions empower data savvy talent to combine the beauty of the human mind with the power of machine learning. Where business intelligence and visual analytics allow you to understand what happened in the past, automated machine learning is the key to successfully navigating the future.

BY 2020, 50% OF ANALYTICAL QUERIES WILL BE GENERATED VIA SEARCH, NATURAL-LANGUAGE OR WILL BE AUTOMATICALLY GENERATED[1].

The evolution of analytics cannot be ignored. You should begin exploring automated machine learning to deliver rapid, transformational value in minutes.

Gartner [1] https://www.gartner.com/newsroom/id/3570917
Gartner Symposium [2] October 2, 2017 event keynote

DATAROBOT FOR MACHINE LEARNING

DataRobot is the world's most advanced automated machine learning platform. DataRobot automates the machine learning process from data ingestion to deployment. It delivers immediate value and unmatched ease of use - no complicated math or scripting is required.

DataRobot automates feature engineering, finding key insights and hidden patterns. This invaluable technology expedites analytical investigation across millions of variable combinations that would be far too time consuming for manual human exploration.

"Domain knowledge and best practices used by the world's leading data scientists have been uniquely baked into DataRobot blueprints."

For optimal machine learning model performance, domain knowledge and best practices used by the world's leading data scientists have been uniquely baked into DataRobot blueprints. Users of all skill levels can safely apply machine learning with built-in optimizations and safeguards.

DataRobot supports popular advanced machine learning techniques and open source tools such as Apache Spark, H2O, Scala, Python, R, and TensorFlow. Using drag-and-drop, point-and-click guided menu options, DataRobot users can simply and quickly create predictive models with automated machine learning.

- Ingest data sources
- Select a target variable to predict
- Automatically generate features, extract balanced samples, build and iterate through 100s of machine learning models
- Visually explore top performing models and key findings
- Easily deploy and operationalize models

Machine learning development steps that used to take weeks or months of effort can now be completed in hours. By embedding DataRobot automated machine learning model intelligence into your reporting or business processes, you can quickly close the loop between insight to action.

In this white paper, we will discuss how to get started using DataRobot automated machine learning. We will introduce machine learning concepts, provide important data preparation tips, and walk you through the steps to apply DataRobot to your own data.

GETTING STARTED WITH MACHINE LEARNING

Before we dive into machine learning, let's briefly review the analytical maturity model and how it complements business intelligence and analytics.

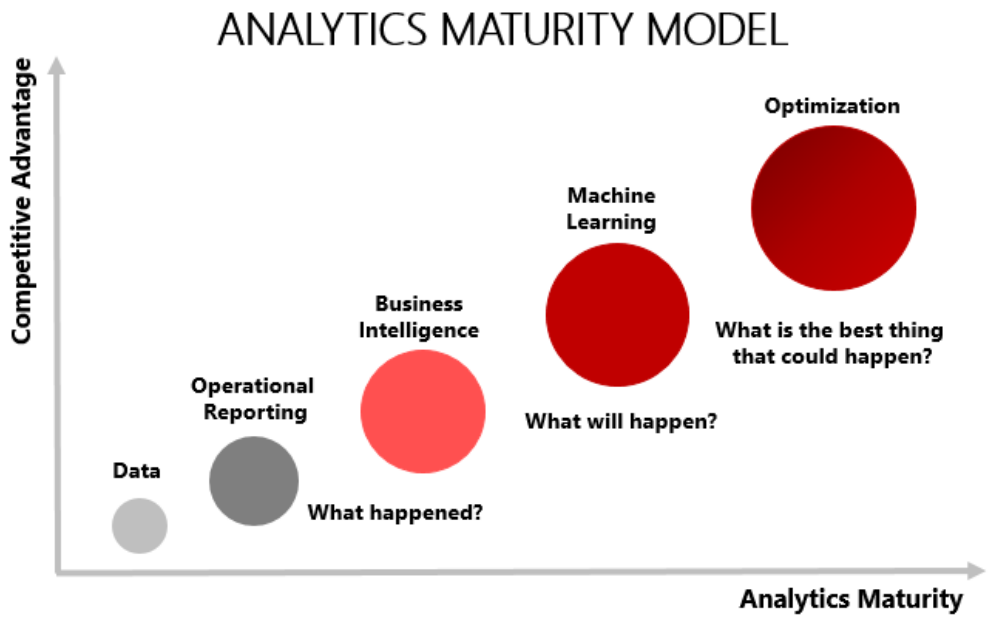


Figure 1 Analytics Maturity Model

Analytics maturity begins with the objective to transform raw data into operational reporting insight to reduce intuition-based decision-making. Data gets aggregated into data warehouses. At this level of analytics maturity, most of the time is spent querying, merging and preparing data. There is little time left for data analysis.

At the next phase, business intelligence is used to understand what happened in the past with analytical OLAP cubes, pivot table summaries, data visualization and dashboards. These essential data-driven decision tools typically require manual development processes to aggregate counts, sums and averages. Findings often reflect a single point in time and do not typically consider statistical significance.

Machine learning is used to answer what will happen next. An application of artificial intelligence (AI), machine learning provides systems with the ability to automatically learn from past experiences without being explicitly programmed. Machine learning captures statistically significant relationships between input variables and target variables. Automated machine learning expedites the process.

Machine learning can efficiently analyze large volumes of data that may contain too many variables for manual business intelligence or traditional statistical analysis techniques. Time series, regression, neural networks, decision trees, classification, random forests, support vector machines and many other types of machine learning algorithms can automatically discover the signal in the noise. Hidden patterns and trends in data that a human mind would not feasibly be able to detect can be identified.

"Machine learning is ideal for efficiently analyzing large volumes of data that may contain too many variables for manual analysis techniques."

Common use cases for applying automated machine learning include financial credit scoring, preventing churn, optimizing inventory, direct marketing, conversion modeling, pricing, forecasting, detecting fraud and managing risks. There are countless other applications for applying this technology across banking, financial services, healthcare, insurance, government, energy, high tech and other industries.

The trained machine learning model output is usually combined with business intelligence and line of business application processes.

MACHINE LEARNING DOES NOT REPLACE BUSINESS INTELLIGENCE – IT SUPPLEMENTS AND IMPROVES IT.

Machine learning automation benefits include:

- Saves time by analyzing billions of data combinations instantly
- Finds the signal in the noise, reveals hidden patterns and trends that a human might never detect using manual approaches
- Improves accuracy by applying statistical significance, uncertainty and risk model estimates that manual approaches usually do not consider
- Helps minimize potential analytics bias since insights are purely data-driven versus manual analysis that is prone to human interpretation
- Closes the insight to action loop to optimize outcomes

Business intelligence skills easily transfer to automated machine learning, providing a solid foundation for success. Deep domain acumen, problem-solving capabilities and data are all you need to get started.

AN ITERATIVE PROCESS

Similar to Kimball dimensional modeling used in business intelligence projects, the process for developing high quality machine learning models is iterative in nature. Never start a project without first thinking about who should be involved, the workflow, and how your machine learning solution will be integrated into existing applications or business processes to make a positive impact. It is a best practice to start with small, quick win projects to gain momentum. Then leverage that momentum to continually achieve better results.

Success lies in identifying numerous opportunities and executing on as many as possible. You can get started by using data that is already available since data sources will never be complete or perfect. Modern data preparation and automated machine learning tools enable you to rapidly prototype ideas and determine whether they are worth pursuing. Once your first iteration is complete, you can proceed to other ideas, invest in additional data collection, further exploration, or start building a better version of your model.

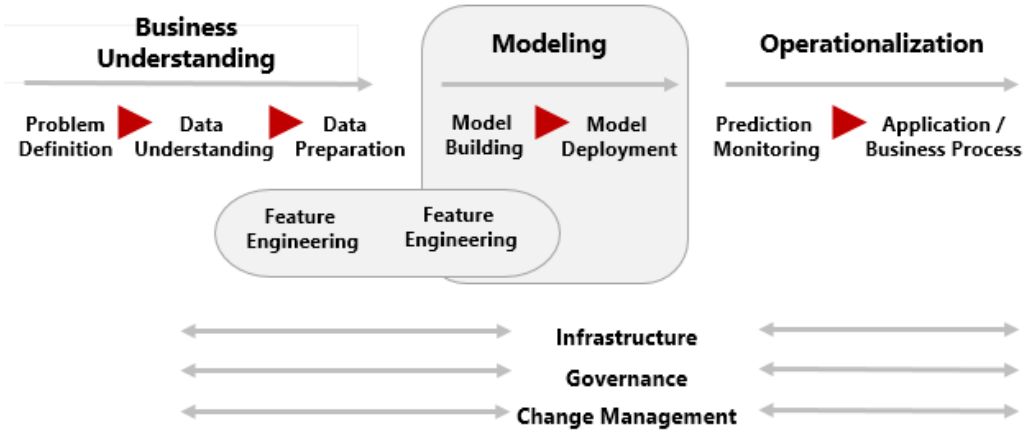


Figure 2 Iterative Machine Learning Process

The machine learning process begins with **Business Understanding**. This initial step focuses on defining the right problem to solve, recognizing the business objectives and requirements. After selecting a problem, you will collect and assess data. During the Data Understanding step, you will get familiar with available data sources, identify data quality problems, and perform exploratory analysis. Then in the Data Preparation step, you will cleanse the data, shape and transform it into a flattened format for loading into the machine learning automation platform.

During **Modeling,** you will select a metric to predict and let the machine learning platform automatically create new features, **Train** hundreds of predictive models, **Score** them and rank them by selected performance measures.

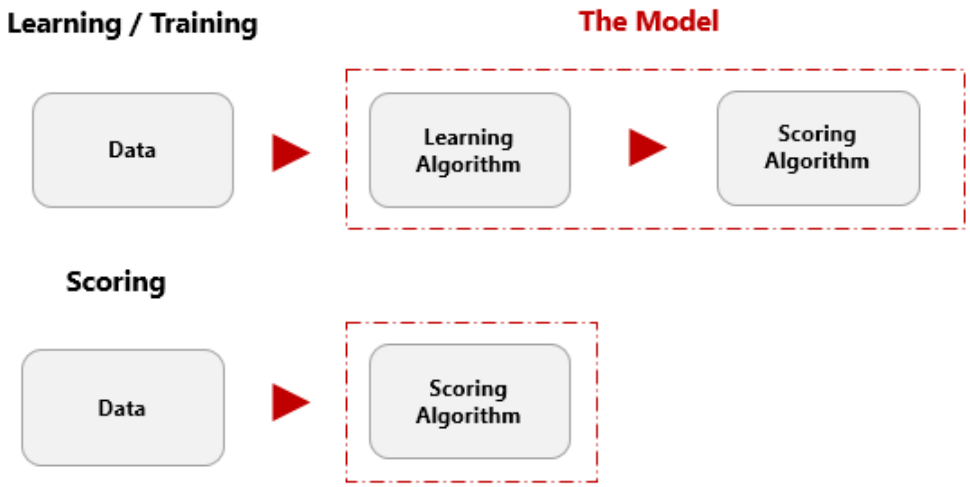


Figure 3 Two Phases of Automated Machine Learning

After reviewing ranked models, scores and estimated accuracy, you will select a model and then proceeding to **Operationalization.** In this step, you will deploy the model, continue to update it with new data and monitor performance. NOTE: Deployment of a machine learning model is generally not the end of your project. Machine learning models continuously evolve over time as new data is acquired, different questions arise and the environment changes.

WITH DATAROBOT AUTOMATED MACHINE LEARNING, MOST OF THE TIME CONSUMING, COMPLEX MODELING, EVALUATION AND DEPLOYMENT WORK IS AUTOMATICALLY DONE FOR YOU AND VASTLY SIMPLIFIED.

HUMAN ART OF AUTOMATED MACHINE LEARNING

Automated machine learning does not operate in isolation. The human art of automated machine learning requires business or domain expertise - selecting the right problems to solve, identifying crucial input data to collect and carefully thinking through the business process to prepare your data. After deployment, evaluation, tuning, and monitoring should continue to be led by human subject matter experts.

EXPLORATORY DATA ANALYSIS

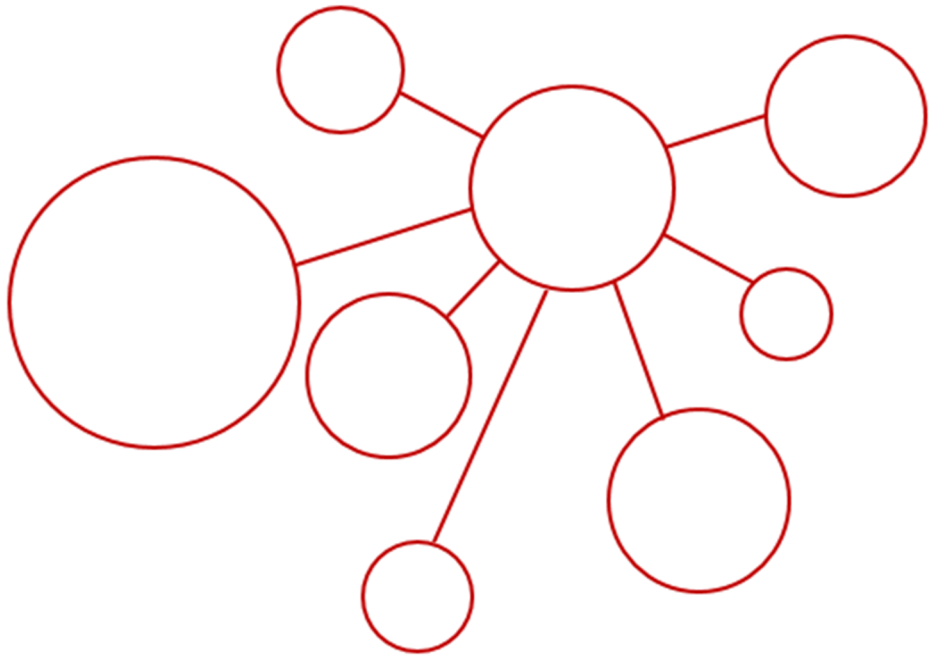
During this step, you'll visually explore variables, look for trends, extremes values, exceptions, skewed data, identify issues and review missing data. As you begin to explore and understand your data, DataRobot Informative Features can quickly help you determine what attributes provide the most information gain for building optimal machine learning models. Knowing what areas of your data influence outcomes is invaluable. This information alone helps the business understand where to focus limited time and resources.

"By looking at these findings, the business can immediately appreciate the variables that truly influence outcomes to get immediate value."

One of DataRobot's strengths is the visualization tools that identify influential features in relation to your desired model target variable. After you ingest data into DataRobot and select a metric to optimize, Feature Ranking and Feature Impact reports are available to you to interactively explore.

The Feature Ranking report measures how much each feature, by itself, is correlated with the target variable. A Feature Impact bar chart visualization displays measures of how much each feature contributes to the accuracy of a machine learning model. By looking at these findings, the business can immediately appreciate the variables that truly influence outcomes to get immediate value from a DataRobot project.

The intuitive DataRobot web-based interface allows anyone to interact with a very powerful platform, regardless of skill-level and machine learning experience. Users can drag-and-drop then let DataRobot do all the work or they can write their own models for evaluation by the platform. Built-in visualizations, such as Model X-Ray and Feature Impact, offer the deepest insights and a whole new understanding of your business.



impact analytix

UNDERSTANDING PROPER DATA COLLECTION AND PREPARATION

Data preparation is crucial for any analytics, business intelligence or machine learning effort. Don't overlook this important part of your machine learning project. Although DataRobot provides safeguards to prevent common mistakes and is robust enough to handle imperfect data, you'll still want to properly prepare your data to get the best possible results. Since each data set and business objective can be unique with varied data preparation challenges, the following basic tips are intended to help get you started.

Ideally, subject matter experts that understand the business domain, processes and the data source nuances would assist in collecting, cleansing and shaping data for DataRobot machine learning automation. Depending on your project, data preparation might be a one-time activity or a periodic one. As new insights are unveiled, it is common to further experiment by adding or changing DataRobot input data.

DATA PREPARATION BASICS

Unlike third-normal form transactional or dimensional patterns used in business intelligence, machine learning requires data input as a "flattened" table, view, or comma separated (.csv) flat file of rows and columns. Your view will need to contain an outcome metric, target variable, along with input predictor variables.

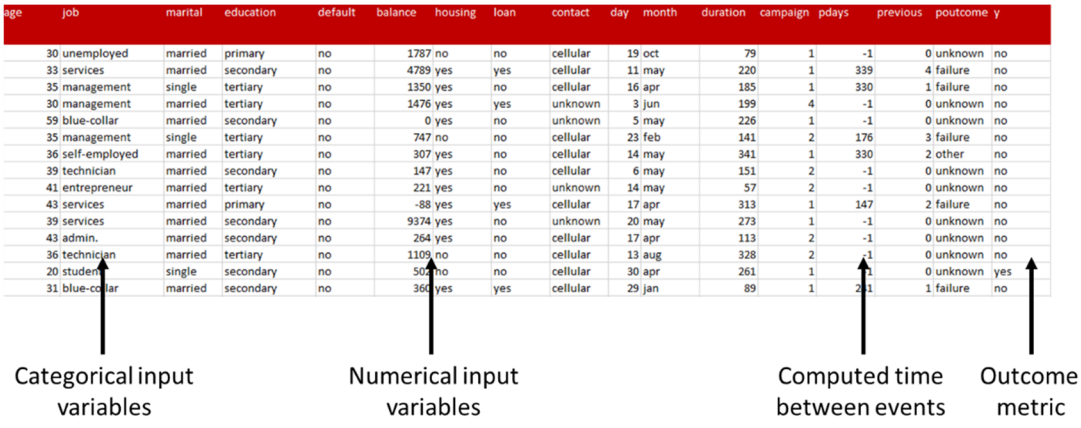


Figure 4 Example Prepared Data Set

If you have data stored in several tables in a data warehouse or relational database format, you will need to use record identifiers to join fields from multiple tables to create a single unified, flattened "view". For many target variables, input data is captured at various business process steps in multiple data sources. A sales process might have data in a CRM, email marketing program, Excel spreadsheet and accounting system. If that is the case, you will want to identify the fields in those systems that can relate, join or blend the different data sources together.

Prepared data should be collected at a level of analytical granularity that you can make decisions upon. Choose a granularity that is actionable, understandable, and useful in the event you incorporated the results into your existing business process or application. For example, if you wanted to make daily sales forecasts, you would need to input data at a day level versus week, month or year.

"Shaping data involves subject matter expert thought to creatively select, create and transform variables for maximum influence."

If you are trying to capture changes in data over a certain time period, check if your data source is only keeping the current state values of a record. Most data warehouses are designed to save different values of a record over time and do not overwrite historical data values with current data values. Transactional application data sources, Salesforce for instance, only contain current state value for a record. If you want to get a prior value, you would need to have a snapshot of the historical data stored or keep the prior value data in custom fields on the current record.

After identifying outcome metric business process data sources and fields, you will want to select any fields (input variables) that may directly affect the outcome. In doing so, ensure that the variable data is clean and consistent. The order and meaning of input variables should remain the same from record to record. Inconsistent data formats, "dirty data" and outliers can undermine the quality of analytical findings.

Then you will shape the data into analytical features with derived variables that you feel might describe or influence the outcome metric. If you ever saw the movie or read the book "Moneyball: The Art of Winning an Unfair Game" by Michael Lewis that discusses the revolution of baseball analysis with new performance metrics On-Base Percentage (OBP) and Slugging Percentage (SLG), essentially you will be using a similar approach. Shaping data involves subject matter expert thought to creatively select, create and transform variables for maximum influence. Please refer to DataRobot.com for additional resources on this important topic.

NOTE THAT DATAROBOT AUTOMATES FEATURE ENGINEERING DURING THE MACHINE LEARNING AUTOMATION PROCESS. IT IS A FANTASTIC FEATURE AND ONE OF THE TRUE BENEFITS OF AUTOMATION.

DELVING INTO MACHINE LEARNING MODELS

Since DataRobot automates the machine learning modeling process, you do not need to know all the algorithm types, how they work or the complex math behind them. Nor do you need to create different input data sets to accommodate algorithm differences. DataRobot has vastly simplified and expedited those historically tedious, time-consuming aspects of machine learning projects.

While automation and speed usually comes at the expense of model quality, that is not the case with DataRobot. It uniquely delivers on all fronts. DataRobot automatically searches through millions of combinations of algorithms, data preprocessing steps, transformations, features, and tuning parameters for the best machine learning models to test against your data. Each model is fine-tuned for the specific dataset and selected prediction target using blueprints designed by the world's leading data scientists.

"Currently DataRobot applies supervised machine learning techniques and ensembles."

TYPES OF MACHINE LEARNING MODELS

Currently DataRobot is optimized for supervised machine learning techniques and ensembles. Ensembles help improve machine learning results by combining predictions from several models. This approach provides better predictive performance than relying on a single model. In the future, unsupervised and reinforcement model types might be available.

SUPERVISED LEARNING

Supervised learning algorithms are mainly used in predictive modeling. Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features. The primary supervised learning algorithms include:

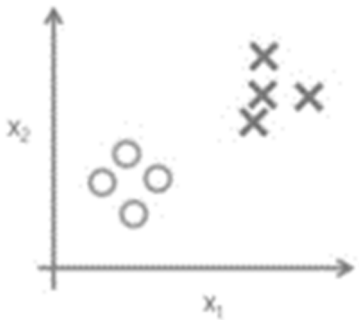


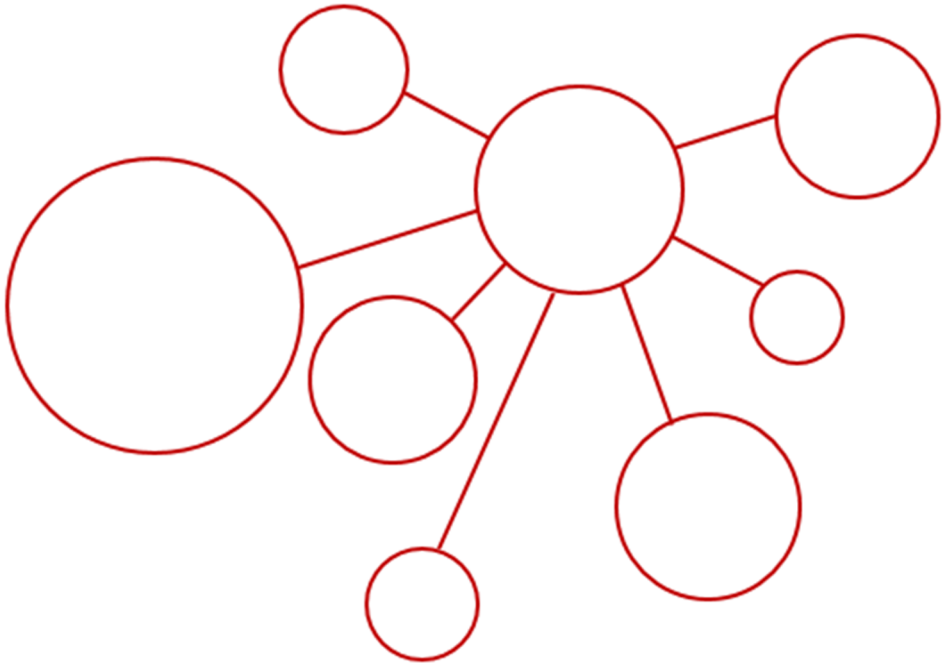
Figure 5 Supervised Learning

Classification: These algorithms build predictive models from training data which have features and class labels. They use features from training data on new, previously unseen data to predict class labels. Types of classification algorithms include but are not limited to decision trees, random forests, and support vector machines.

Regression: These algorithms are used to predict output values based on some input features obtained from the data. The output values in this case are continuous and not discrete. Types of regression algorithms include linear, multivariate, lasso and regression trees.

Ensembles: Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce more accurate solutions than a single predictive model.

Voting and averaging are two simple ensemble methods to understand. Voting is used for classification and averaging is used for regression. Stacking, also known as stacked generalization, is an ensemble method where predictive models are combined using another machine learning algorithm. Bagging, another technique, creates multiple models using the same algorithm using random dataset sub-samples. Lastly, boosting incrementally builds an ensemble by training each model with the same dataset but adjusts the weights according to the error of the last prediction. By automating ensemble methods, DataRobot generates high performing models.



impact analytix

EVALUATING AND DEPLOYING MACHINE LEARNING MODELS

DataRobot can turn an analyst or business intelligence specialist, who has domain knowledge about their business and an intimate knowledge of the data, into a pseudo data scientist who can quickly develop highly accurate machine learning models that are easy to deploy. Automated machine learning exponentially expedites the process of running through many model types, evaluating and finding the best models using industry standard cross-validation.

Cross-validation is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. In cross-validation, you split the input data into k subsets of data (also known as folds). You train your machine learning model on all but one (k-1) of the subsets, and then evaluate the model on the data that was not used for training. This k-fold cross-validation process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time.

"When assessing the performance of machine learning models, keep in mind what an error actually means."

MEASURING ACCURACY

The estimated performance of a model tells you how well it might perform on unseen data. Machine learning models within DataRobot can be evaluated using Estimation, Classification, Gain/Lift, ROI, ROC Curves and other types of accuracy estimation measurements. Different types of accuracy measurements are provided depending on the nature of the machine learning algorithm.

For example, with classification accuracy is simply the proportion of correctly classified instances. When evaluating a classification model, you'll review a confusion matrix showing the number of true positives, false negatives, false positives, and true negatives, as well as ROC, Precision/Recall, and Lift curves. The Precision of the model is the proportion of positives that are classified correctly. Recall is a measure of the true positive rate.

To review true positive rate versus false positive rate, a Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) value is used. The closer these curves are to the upper left corner, the better the classifier perform maximizing true positive rate while minimizing false positives.

When evaluating regression accuracy, you will use Mean Absolute Error, Root Mean

Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination. Errors here indicate the difference between the predicted value and the true value. Coefficient of determination, which is also known as R squared, is also a standard way of measuring how well the model fits the data

When assessing the performance machine learning models, keep in mind what an error means. In some cases, errors are less sensitive than in others (such as healthcare). If you can, visually explore errors with your favorite visual analytics or business intelligence tool. By finding where errors happen, you might be able to iteratively improve the input data for more accurate results.

MODEL TRANSPARENCY

These tools allow the user to evaluate "why" a particular model is performing as it does. And tools like Reason Codes go one step further, giving the Top X number of reason why a particular decision was made in the model, allowing the user to see at the individual decision level (oftentimes and individual person) why a decision was made for that individual (i.e. declined for a loan). This will help companies comply with new regulations around transparency for people when they are affected by machine learning algorithms, such as General Data Protection Regulation (GDPR) in the European Union.

DEPLOYING MACHINE LEARNING MODELS

The best predictive models have little to no organizational value unless they are rapidly operationalized. With DataRobot, deploying models for predictions can be done with a few mouse-clicks. After DataRobot is finished running your data, you can use selected, high performing models in a variety of ways.

- Upload new data for scoring in-app, and download the predictions from DataRobot in a .csv file
- Use DataRobot's REST API to integrate the chosen model into production systems to have DataRobot serve as the prediction server in enterprise environments (and allowing for rapid updates when new data is available or conditions change).
- Export the chosen model in Java, R, or other programming languages in environments that call for siloed predictions.

DataRobot deployment options allow easy integration of your chosen model into your production systems. You can deliver immediate value and unprecedented insights to the business in a matter of minutes.

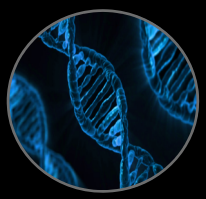
DATAROBOT IN ACTION

DataRobot is used in a wide variety of industries to solve some of today's most challenging problems. Countless applications and business processes across all industries are integrating machine learning to optimize outcomes. The following is an overview of the most popular use cases.

Financial Services
Wealth advisory companies can quickly develop, test, model, and deploy algorithms to assist with portfolio management services as well as help wealth managers prospect for new clients.

Banking and Investment
Banks are using machine learning to optimize all areas of their business from risk analysis and fraud detection to detecting churn, cross selling and marketing new offers.

Pharmaceutical
From drug discovery to clinical trial research and manufacturing, pharmaceutical companies are using machine learning to expedite and improve medicine and optimize operations.



Healthcare
Healthcare companies are using machine learning to improve outcomes, reduce expenses and improve efficiencies from readmission risk and occupancy rates to marketing.

Insurance
Insurance companies are using machine learning for underwriting, pricing, assessing risk, detecting fraud, direct marketing, and claim payment estimation.

Energy, Utilities and Telecom
To better serve customers, machine learning can be used to prevent churn, predict maintenance and schedule resources based on usage patterns and predicted demand.

Transportation and Logistics
Machine learning can be used to optimize fleet management operations, estimate delivery, forecast fuel usage, predict delays and improve scheduling.

Security, Law Enforcement and Military
Predicting and preventing terrorist attacks is a chief concern for intelligence and agencies, and predictive modeling based on historical data may help prevent them in the future.

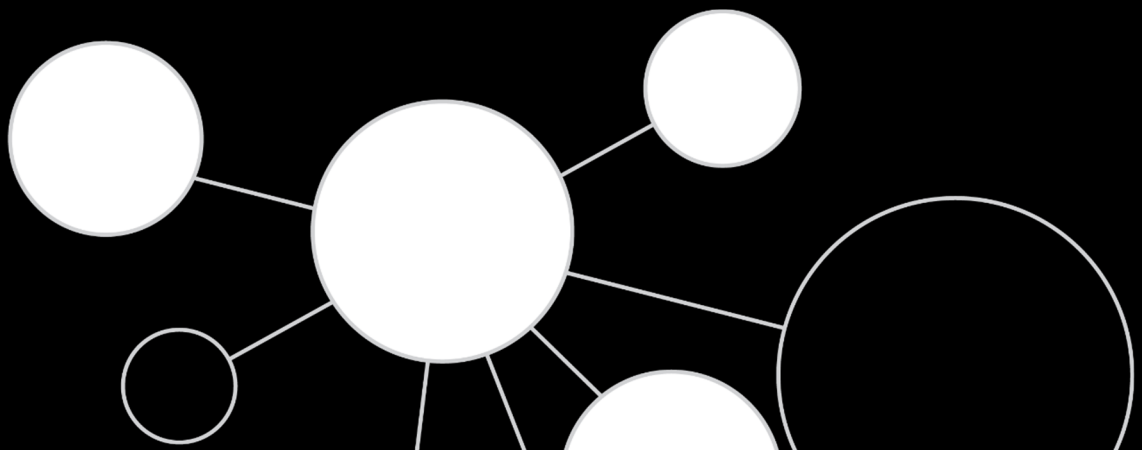
Sales, Marketing, and Customer Service
Machine learning is frequently used to score incoming leads, detect issues and provide dynamic pricing within ERPs, CRMs and other applications.

Information Technology
Cyber-security, insider threat and technology monitoring systems rely on machine learning to understand usage patterns, detect unusual behavior, and detect issues.

Retail and Ecommerce

Machine learning is often used to manage inventory, upsell, cross sell, provide dynamic pricing, detect fraud, and personalize offers in stores and online shopping applications**.**

Government and Public Agency Applications
Machine learning can be used to improve a variety of public services from applying for permits, paying taxes, detecting fraud, keeping track of pending legislation or reviewing spending.



CONCLUSION

"Organizations need to start investing and adopting next generation automation technologies now to survive and thrive."

No one wants to become the next extinct talent, business, solution or service that missed the early market signals and opportunities to evolve before it was too late. By relying only on historical reports, gut feel or intuition, instead of looking forward with predictions from automated machine learning, your career and business opportunities are likely to be underserved.

Automated machine learning is a significant competitive differentiator in a big data world, quickly transforming the digital business era. By 2020, information will be used to reinvent, digitalize, or eliminate 80% of business processes and products from a decade earlier.[3] Organizations need to move quickly, adopting next generation automation technologies now to survive and thrive.

Seize the opportunities now to experiment and adopt automated machine learning technologies to improve your decision-making processes. In this white paper, we explored how to get started with DataRobot automated machine learning. You don't need to be a data scientist to get immediate value from DataRobot. By combining the best of automation with your existing talent and technological investments, you can arm your organization with the right solutions to extract the most knowledge and value from your data assets.

RECOMMENDED NEXT STEPS

For additional information on machine learning automation, please contact an expert at DataRobot. It's easy to get started.

- DataRobot
  http://www.datarobot.com
- DataRobot Acceleration Packages
  https://www.datarobot.com/product/enterprise-ai/
- DataRobot Courses
  https://www.datarobot.com/education/all-courses/

Gartner Predicts [3] http://blogs.gartner.com/doug-laney/a-look-back-on-my-information-and-analytics-strategy-research-from-2014/

About DataRobot

DataRobot offers an enterprise machine learning platform that empowers users of all skill levels to make better predictions faster. Incorporating a library of hundreds of the most powerful open source machine learning algorithms, the DataRobot platform automates, trains and evaluates predictive models in parallel, delivering more accurate predictions at scale. DataRobot provides the fastest path to data science success for organizations of all sizes. For more information, visit www.datarobot.com.

About the Author

Jen Underwood, founder of Impact Analytix, LLC, is an analytics industry expert with a unique blend of product management, design and over 20 years of "hands-on" advanced analytics development. In addition to keeping a pulse on industry trends, she enjoys digging into oceans of data. Jen is honored to be an IBM Analytics Insider, SAS contributor, and former Tableau Zen Master. She also writes for InformationWeek, O'Reilly Media and other tech industry publications.

Jen has a Bachelor of Business Administration – Marketing, Cum Laude from the University of Wisconsin, Milwaukee and a post-graduate certificate in Computer Science – Data Mining from the University of California, San Diego. For more information, visit www.jenunderwood.com.

impact analytix