dotData

AUTOML PLATFORM EVALUATION GUIDE FOR BI PROFESSIONALS

A business-value focused evaluation guide to help you gauge how quickly your team is likely to become productive with AI and ML with a given AutoML platform.

ABOUT THIS GUIDE

This evaluation guide has been designed to support Business Intelligence, Analytics and Data Science organizations as they evaluate Automated Machine Learning (AutoML) Platforms in an effort to embrace, adopt and scale the data science discipline that lies at the heart of Machine Learning and Artificial Intelligence.

The rapid adoption of AI and Machine Learning in the enterprise requires businesses to expand their AI and ML efforts without negatively impacting the typically limited data science resources of the business. Whether your organization has a dedicated data science team or does not this guide has been designed to help you assess each platform's ability to automate as much of your data science process as is necessary to maximize your opportunity for success.

THE DATA SCIENCE WORKFLOW

The modern data science workflow lies at the heart of any Machine Learning or AI project. The process of leveraging available data to create ML/AI models is depicted below and is the foundation of this evaluation guide.

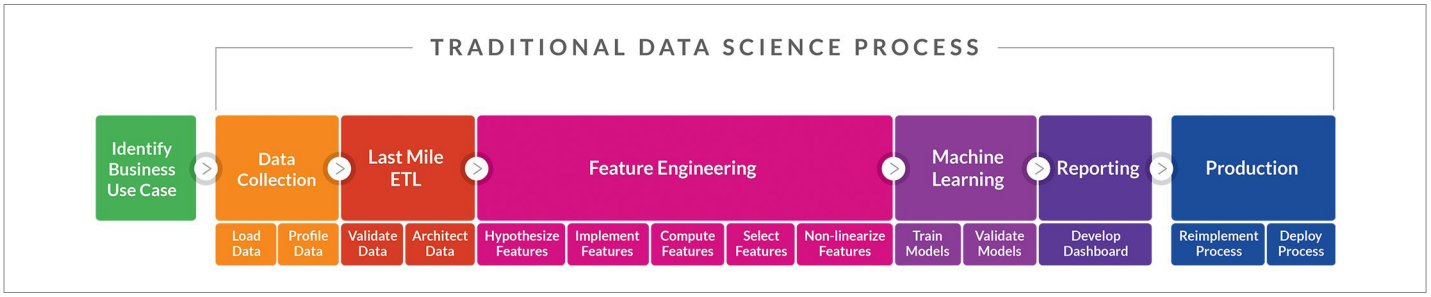

Figure 1: The Traditional Data Science Process

Any data science project is going to start with identifying business use cases and requirements. The process is also heavily dependent on the available resources of the business as well as the skill-set of the primary intended users. In order to make the best possible choice, organizations should start their evaluation by asking some fundamental questions:

- Who will be the primary intended users of the AutoML platform? The Data Science Team or the BI team?
- What are the skill-level and data science expertise of the primary user?
- Is the primary programming environment of the intended users Python?

This evaluation guide has been designed with the following assumptions in mind as responses to the above questions:

- Your BI & Analytics or Data Engineering team will be the primary users of any AutoML platform and will either be working in addition to or as a replacement for an already existing data science team
- The team has deep experience with data, data-wrangling, and manipulation, reporting and SQL, but does not have deep experience with statistical analysis, data science concepts or machine learning algorithms and techniques.
- Your analytics team is not a heavy user of Python and primarily uses SQL and BI tools

Key Takeaway:

The data science process that underlies development of AI/ML models is time-consuming and requires a lot of expertise when done manually. Even with typical AutoML tools, the feature engineering portion can take months.

EVALUATING A MODERN AUTOMATED MACHINE LEARNING (AUTOML) PLATFORM

This guide focuses on evaluating technical capabilities that are important when non-experienced data scientists or BI professionals use an AutoML platform. It is also important, however, that you consider the following non- functional attributes when evaluating AutoML platforms:

Platform Accessibility & Deployment

- Can all steps of the data science process be executed seamlessly within a single platform without the need for moving between systems and applications?
- Can all steps of the data science process be executed without involvement from specialists and/or data scientists?

Ease of Use & Automation

- How much manipulation of data must be performed before it is ready for ingestion by the AutoML platform?
- Is it easy for non-data scientists to understand the workflow of the application, the concepts, and steps necessary to proceed?
- How much of the process is truly automated? How much manipulation is needed at each step?

User Enablement

- What is the onboarding process that was outlined by the provider? Will the training encompass basic product usage? Or will it also provide data science basics to get the best use out of the platform?
- Will your users have ongoing access to data science professionals to help in their ongoing use of the product?

Deployment Flexibility

- Are there multiple deployment options (in-house, private cloud etc.) as well as support for multiple environments (AWS, Azure etc.)
- Is the platform scalable over time to account for increased user counts as well as growing data volumes?

Pricing and Packaging

- Are the available licensing options clear and transparent?
- Do the licensing options offer the right features and value for the price?
- Is the pricing model for the platform easy to understand?
- Is the pricing model for the platform flexible and scalable?
- Is the product priced in a way to provide all the necessary features for one price?

1.0 DATA INGESTION AND PREPARATION

Primary Goal of the Evaluation: How easily can I import and prepare my data?

When evaluating AutoML platforms, one of the most overlooked components of the process is data ingestion and preparation. In order for your Data Science process to work efficiently and quickly, having well prepared, data optimized for analytics is a critical first step. The evaluation criteria in this section are addressed from the perspective of the data engineering and BI team:

Evaluation Criteria:

Your data engineers and BI team should be able to:

- Upload data to the AutoML platform without having to write additional SQL code
- Convert entity relationships to analytical relationships with minimal work
- Add additional analytical relationships without having to write SQL code to provide new relationships that did not exist in the source tables
- Ingest either CSV files or raw database files without having to create normalized tables

Evaluation Considerations:

- ☐ Can the platform automatically ingest data from relational databases?
- ☐ Does the system automatically discover database relations without human intervention?
- ☐ Can users create custom relationships between database tables without using SQL code?
- ☐ Can the platform provide a base-level of data cleansing (identifying unique values, duplicates, illegal values etc.)

2.0 FEATURE ENGINEERING

Primary Goal of the Evaluation: How much manual work is involved in Feature Engineering?

Feature Engineering is the part of the data science process that takes the most amount of time. In fact, even among data science professionals, feature engineering is often the least-favorite activity and one that is wrought with trial and error. A typical feature engineering process requires a high degree of domain knowledge, but obtaining access to domain experts is time-consuming and problematic, slowing the whole process down. It's common for feature engineering to take months in many AI/ML projects. The evaluation criteria in this section are addressed from the perspective of your data science team, your BI team as well as from the perspective of domain experts that might have input into your platform selection:

Evaluation Criteria

Your AutoMl users should be able to:

- Work with any size data, even billions of records, to generate features quickly
- Discover features with no SQL code-writing and without having to manually process and analyze features
- Consume flat tables as well as relational data sets to generate new features
- Leverage varied data sets including geospatial and temporal data
- Be able to explain features to domain experts easily and quickly using system-generated blueprints

Evaluation Considerations:

- ☐ Can the system automatically explore all available database entity relationships and discover features based on available columns and relationships?
- ☐ Can the system generate features based on transactional tables without additional add-ons?
- ☐ Can the system generate geolocation-based features?
- ☐ Can the system explore text values and generate features based on text data?
- ☐ Can the platform create features using linear and non linear transformations ?
- ☐ Does the system explore previously profiled records to generate additional features?
- ☐ Can the platform prevent target leakage?
- ☐ Can the system evaluate millions of features and only expose the most relevant ones?
- ☐ Can the platform support feature engineering from databases with billions of records?
- ☐ Can the system create feature "blueprints" that are easily interpretable by domain experts?

3.0 MACHINE LEARNING

Primary Goal of the Evaluation: Can you evaluate features against the most relevant Machine Learning Algorithms?

Choosing the right machine learning model - and knowing how to optimize it for best performance is one of the hallmark features of any AutoML product. For the vast majority of businesses, however, having access to hundreds of ML algorithms is not very useful - that's because in the vast majority of business use-cases there are a handful of algorithms that tend to provide the best possible performance. When evaluating AutoML platforms, focusing too much time on the Machine Learning algorithm selection piece can slow the process and waste time and resources. The following evaluation criteria were created to provide the ideal features required in order to provide optimal results:

Evaluation Criteria

Your AutoMl users should be able to:

- Choose from state-of-the-art ML algorithms
- Have full automation of key steps like hyper-parameter optimization and search of feature preprocessing methods
- Be able to automatically choose ideal ML models based on automated methods
- Be able to provide transparency in selected models

Evaluation Considerations:

- ☐ Does the system support all state-of-the-art ML algorithms like scikit-learn, XGBoost, LightGBM, TensorFlow and PyTorch?
- ☐ Can you perform an automated hyper-parameter search of ML algorithms?
- ☐ Can the system automatically search for feature preprocessing methods?
- ☐ Can the system automatically avoid overfitting?
- ☐ Can ML model designers quickly and easily provide transparency for ML models?

4.0 PRODUCTION & OPERATIONALIZATION

Primary Goal of the Evaluation: How quickly and easily can you deploy ML models in production environment? Can you monitor models, discover data drift and quickly retrain them if production data changes over time?

Even when ML/AI models have been developed, many fail to become operationalized. The problem often is that data science teams run into roadblocks that are systemic in nature. IT departments might be wary of deploying AI/ML models to production environments. In addition, as production data changes over time, ML models need to be monitored, optimized and retrained, which can lead to production system downtimes if implementations were not done properly. The following evaluation criteria were created to provide the ideal features required in order to maximize the chances for success during operationalization. It is meant to be read by both your BI/Data Engineering as well as IT organizations:

Evaluation Criteria

Your AutoMl users should be able to:

- Deploy AI/ML models into production environments with just a few lines of code
- Continuously monitor model health and performance
- Dynamically retrain/update models as production data changes

Evaluation Considerations:

- ☐ Can you expose ML scoring pipelines using API calls?
- ☐ Can you expose feature pipelines for both flat-tables as well as relational table features?
- ☐ Can the system package ML/Feature pipelines and expose them via an API?
- ☐ Can the system monitor ongoing ML accuracy and degradation as production data changes?
- ☐ Can ML/AI models be retrained without downtime?

ABOUT DOTDATA

dotData Pioneered AutoML 2.0 to help business intelligence professionals add AI/ML models to their BI stacks and predictive analytics applications quickly and easily. Fortune 500 organizations around the world use dotData to accelerate their ML and AI development to drive higher business value. dotData's automated data science platform accelerates ROI and lowers the total cost of model development by automating the entire data science process that is at the heart of AI/ML. dotData ingests raw business data and uses an AI-based engine to automatically discover meaningful patterns and build ML-ready feature tables from relational, transactional, temporal, geo-locational, and text data.

dotData has been recognized as a leader by Forrester in the 2019 New Wave for AutoML platforms. dotData has also been recognized as the "best machine learning platform" for 2019 by the AI breakthrough awards, was named an "emerging vendor to watch" by CRN in the big data space and was named to CB Insights' Top 100 AI Startups in 2020. For more information, visit www.dotdata.com, and join the conversation on Twitter and LinkedIn.

Learn more about dotData:

dotData, Inc
2988 Campus Drive
Ste 100
San Mateo, CA 94403
https://dotdata.com

(415) 301-5600