AutoML

Konstantinos Konstantinidis 2546
Nikolaos Stavrinos 2631

June 2021

1 Introduction

Automated machine learning, also referred to as automated ML or AutoML, is the process of automating the time consuming, iterative tasks of machine learning model development. It allows data scientists, analysts, and developers to build ML models with high scale, efficiency, and productivity all while sustaining model quality.

Traditional machine learning model development is resource-intensive, requiring significant domain knowledge and time to produce and compare dozens of models. With automated machine learning, you'll accelerate the time it takes to get production-ready ML models with great ease and efficiency.

2 Targets of AutoML

AutoML focuses on three targets.The first one is to accelerate human productivity while cutting costs.The second is to democratize machine learning for all irrespective of the level of expertise and the third one is to improve replicability of analyses, sharing of results, and facilitate collaborative analyses.

3 Advantages over traditional ML

The advantages of AutoML over traditional ML are a lot.The most important is that improves efficiency by automatically running repetitive tasks,which allows data scientists to focus more on problems instead of models.Another advantage is that automated ML pipelines also help avoid potential errors caused by manual work.Also AutoML is a big step toward the democratization of machine learning and allows everyone to use ML features.

4 Pipeline of AutoML

4.1 Data Ingestion

It is the process of obtaining data and importing data for use. Data can be sourced from multiple systems, such as Enterprise Resource Planning (ERP) software, Customer Relationship Management (CRM) software, and web applications. The data extraction can be in the real time or batches. Sometimes, acquiring the data is a tricky part and is one of the most challenging steps as we need to have a good business and data understanding abilities.

4.2 Data Preparation

There are several methods to preprocess the data to a suitable form for building models. Real-world data is often skewed—there is missing data, which is sometimes noisy. It is, therefore, necessary to preprocess the data to make it clean and transformed, so it's ready to be run through the ML algorithms.

4.3 ML model training

It involves the use of various ML techniques to understand essential features in the data, make predictions, or derive insights out of it. Often, the ML algorithms are already coded and available as API or programming interfaces. The most important responsibility we need to take is to tune the hyperparameters. The use of hyperparameters and optimizing them to create a best-fitting model are the most critical and complicated parts of the model training phase.

4.4 Model Evaluation

There are various criteria using which a model can be evaluated. It is a combination of statistical methods and business rules. In an AutoML pipeline, the evaluation is mostly based on various statistical and mathematical measures. If an AutoML system is developed for some specific business domain or use cases, then the business rules can also be embedded into the system to evaluate the correctness of a model.

4.5 Retraining

The first model that we create for a use case is not often the best model. It is considered as a baseline model, and we try to improve the model's accuracy by training it repetitively.

4.6 Deployment

The final step is to deploy the model that involves applying and migrating the model to business operations for their use. The deployment stage is highly dependent on the IT infrastructure and software capabilities an organization has.

5 AutoML frameworks

AutoML frameworks are getting better every day, and can provide high-performing ML pipelines, unique data insights, and ML explanations. No longer blackboxes, these powerful tools offer self-documenting capabilities and native Python notebook support.

5.1 Auto-SKLearn

Auto-SKLearn [**?**, **?**] is an automated machine learning software package built on scikit-learn. Auto-SKLearn frees a machine learning user from algorithm selection and hyper-parameter tuning. It includes feature engineering methods such as One-Hot, digital feature standardization, and PCA. The model uses SKLearn estimators to process classification and regression problems.Last Auto-SKLearn performs well in medium and small datasets, but it cannot produce modern deep learning systems with the most advanced performance in large datasets.

5.2 TPOT

TPOT is a tree-based pipeline optimization tool that uses genetic algorithms to optimize machine learning pipelines. TPOT is built on top of scikit-learn and uses its own regressor and classifier methods. TPOT explore thousands of possible pipelines and finds the one that best fit the data.

5.3 Auto-Keras

Auto-Keras is an open source software library ,which is built on top of the deep learning framework Keras, Auto-Keras provides functions to automatically search for architecture and hyper-parameters of deep learning models.In Auto-Keras, the trend is to simplify ML by using automatic Neural Architecture Search (NAS) algorithms. NAS basically uses a set of algorithms that automatically adjust models to replace deep learning engineers/practitioners.

5.4 H2O

H2O is an open source and distributed in-memory machine learning platform. H2O supports both R and Python. It includes an Automated Machine Learning module and uses its own algorithms to create pipelines. It uses exhaustive search for feature engineering methods and model hyper-parameters to optimize pipelines.

5.5 Auto-TS

Auto-TS is an open-source Python library with time series forecasting implementation. It can train multiple time series forecasting models including ARIMA, SARIMAX, FB Prophet, VAR, etc, in just one line of Python code, and then choose the best one out of it for predictions.

5.6 Google AutoML

Google AutoML is more adept at creating new models, but also is more resource intensive, as it requires the whole data set normally. It uses recurrent neural networks (RNN), convoluted neural networks (CNN) and long short-term memory (LSTM).

5.7 Microsoft Azure AutoML

Azure AutoML is a cloud-based service that can be used to automate building machine learning pipelines for classification, regression and forecasting tasks. Its goal is not only to tune hyper-parameters of a given model, but also to identify which model to use and how to pre-process the input dataset.

5.8 BigML's AutoML

BigML's AutoML is an Automated Machine Learning tool for BigML.The user needs to give it training and validation datasets and it will give back a Fusion with the best possible models using the least possible number of features. BigML's AutoML performs three main operations: Feature Generation, Feature Selection, and Model Selection.

6 Conclusion

Data scientists can accelerate ML development by using AutoML to implement really efficient machine learning. The essence of AutoML is to automate repetitive tasks such as pipeline creation and hyper-parameter tuning so that data scientists can spend more time on business problems on hand in practical scenarios. AutoML also allows everyone instead a small group of people to use the machine learning technology.The AutoML frameworks are getting better every day. The constant improvement is not only in their performance. The next generation of AutoML can provide insights and explanations about analyzed data, create new features, and generate documentation (with native Notebooks support).As a result AutoML will be an important part of machine learning in the future.

7 References

1. https://medium.datadriveninvestor.com/7-best-automatic-machine-learning-frameworks-202
2. https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml
3. https://neptune.ai/blog/a-quickstart-guide-to-auto-sklearn-automl-for-machine-learning
4. https://openml.github.io/automlbenchmark/automl_overview.html
5. https://medium.com/georgian-impact-blog/choosing-the-best-automl-framework-4f2a90cb182