2nd National Conference on Engineering Applications of Emerging Technology in association with International Journal of Scientific Research in Science, Engineering and Technology | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
[https://ijsrset.com | doi : https://doi.org/10.32628/IJSRSET]

Robotic Process Automation for Machine Learning : A Comprehensive Review

Vaishnavi N, Sinchana K
Department of Computer Science and Engineering, Global Academy of Technology, Bangalore, Karnataka, India

ABSTRACT

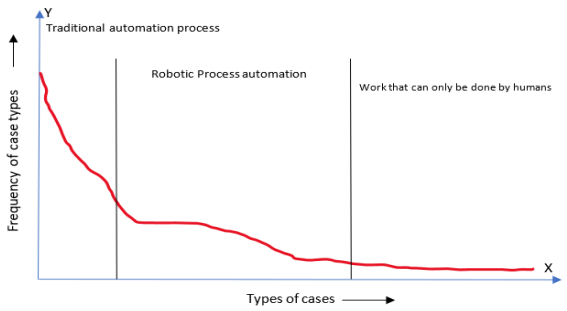
The tremendous growth in the field of RPA, AI and ML has been leading us back to the same question "what to be automated and why?" In this paper, we shift our focus towards how RPA is implemented for the Automation of Machine Learning (AutoML). AutoML is one of the advancements in ML which has reduced the tasks of Data scientists by automating the repetitive tasks like Data processing, Feature Engineering, Model Analysis and a few case studies of AUTO ML are discussed.

Keywords: AutoML, Pareto Distribution, feature Engineering, Hyper parameter optimization, Model interpretation.

I. INTRODUCTION

RPA (Robotic Process Automation) can be implemented in both hardware and software. Here, we'll shift our focus towards Software bots, specifically on the automation of Machine learning. RPA tools are a catch-all phrase for all the tools that interact with other systems just like humans. It aspires to replace humans with outside-in approach which is in contrast with the traditional "inside-out" approach. Here, we get an insight of automating the whole process of building models ML used for predictions. But to begin with a brief insight of why is ML used.

Machine learning (ML) is a branch of research concerned with understanding and developing strategies or methods for leveraging data to improve performance on a set of tasks. So when there's a new input the outcome is predicted based on the model built using the training data set. The main usage of ML is that it allows a computer to learn on itself to grow and change when exposed to new data. Also provides cheap and abundant computations on large datasets. We already know that harvesting huge data is feasible because of the usage of AI and ML in RPA even when there's a change in UI. We can notice a 'Pareto distribution' if we look at the relevance of RPA by taking different case types on the X-axis and the frequency of case types on the Y-axis



From the above graph, we can comprehend that around 80% of cases lie under 20% types of cases which is feasible and cheap for its automation. The remaining 20% of the cases are rare, less frequent, and needs human to act as glue by entering data and making decisions as automation of such cases are too expensive due to the involvement of proprietary systems. However, these cover around 80% of the case types and are time-consuming. Using RPA we can still automate such cases but can't be sure if economically viable or always possible.

II. EVOLUTION OF AUTOMATED MACHINE LEARNING

We can associate Automated Machine Learning with STP (Straight through Processing). STP was developed in the 1990s and became well-known for its fundamental characteristic of an automated system that employs only electronic transfers and needs no human contact in payments and securities trading (as well as other technical fields).This STP was one of the primary characteristics of WFM (Workflow Management) systems, but it was only relevant in a few circumstances, hence BPM (Business Process Management) solutions emerged. These BPMs were expensive. Though Automated ML is correlated to STP, it varies in two ways. The first is that RPA's "outside-in" method differs from STP's "inside-out" approach, therefore the current system stays untouched in RPA. The next aspect is that RPA aspired to be robust irrespective of the core system used. Between 1995 and 2015, ML tools and libraries exploded in popularity, with names like Weka (1990s), Scikit Learn (2007-10), H2O (2011), Spark MLlib (2013), and others. AutoML was developed by academics and practitioners, later startups focused on the Automation of ML. The very first attempt was made by Auto-Weka in the year 2013 from the Universities of British Columbia and Freiburg, followed by many others like Auto-sklearn (2014), H2O Driverless AI, TPOT, Darwin, Etc. All were built based on well-known Scikit learn. MLjar uses both scikit and tensor flow for its automation. Later large cloud providers and technology companies started providing AMLaaS (Automated Machine Learning as a Service) or standalone products, example of these are the popular Google cloud AutoML (2017) on Google cloud, Microsoft Azure-ML on Azure sales force's transmorgrif AI runs son spark ML and Uber's Ledwig on Horovod. While, Darwin, H2O driverless AI and Daterobot deals with time-series (The dataset the tracks a sample overtime) data on a view perspective. Whereas H2O-Driverless AI exports MOJO (Model object optimized) and POJO (Plain Old Java Optimized), for optimized data models to be easily deployed on any platforms supporting architecture of JAVA.

III. FUNCTIONALITIES OF AUTOML TOOLS

There are many automl tools both open source (auto-sklearn, auto weka, TPOT, azure-ml, auto-keras) and commercial (h2o-automl, data robot. The main task of data processing is the data type and schema detection that results in the production of proper feature engineering for the next step in the pipeline which isn't supported by most of the tools (data robot, h20-auotoml, azure-ml) as they depend on the data entered. Wherein TPOT, auto-keras require users to manually perform data pre-processing and only accept numerical feature matrices. These tools use different methods such as clustering, genetic algorithm, ANN, and imputation of missing values to select the best traits of each model and proceed to the next step. Auto-sklearn,

TPOT creates a ml pipeline by finding the dataset required using 'meta-learners' in order to reduce the time taken for data selection and hyper-parameter tuning. Whereas the h20-automl uses the naïve optimization algorithms to achieve best performance, particularly those problems where domain expertise typically biases the choice of optimization. And auto-keras uses network morphism to make Bayesian optimization more efficient. Mostly the commercialized tools are more provided with the functionality of prediction analysis resulting with the detailed representation through model dashboards, feature engineering and visualization methods.

IV. BASIC PIPELINE OF AUTOML TOOLS

The basic pipeline of AUTOML tools is shown in figure 1.

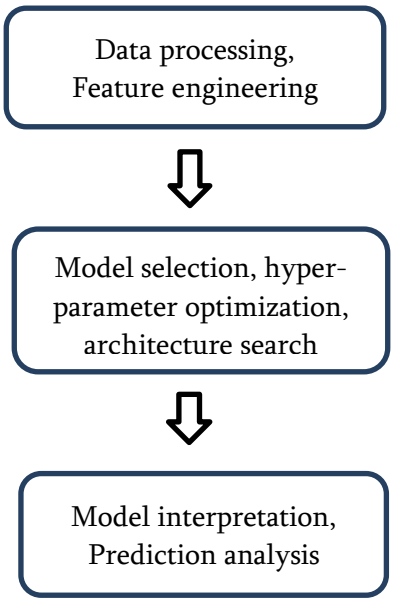


Fig.1 Data processing and feature engineering

In most ML pipelines, data processing is the initial step. At the moment, none of the AutoML tools are really good at handling this work, and it still takes a lot of human interaction. H2O-AutoML, H2O-driverlessAI, Data-Robot, MLjar, and Darwin, on the other hand, gain an advantage by detecting fundamental data kinds or schemas, which are currently confined to numerical, categorical, and time-series data. AutoML, Auto-sklearn, AzureML, and Ludwig, for example, may do feature engineering only based on user input criteria. Auto – sklearn necessitates the use of a label encoder to turn input into integers. TPOT and Auto-Keras are unable to process natural language input automatically; instead, the inputs must be encoded in integer language before data can be sent.

A. Model selection, hyper parameter optimization and architecture search

Different models with parameters such as hyper parameter optimization are created from the features gathered during the feature engineering process, and eventually the best model is chosen. Each tool aids in the development of models using established machine learning methods such as SVM, ANN, logistic regression, and tree-based algorithms. For supervised approaches, Data Robot, Auto-ml, H2O-Driverless AI, Auto-sklearn, MLjar, and TPOT all function in this way. Additional unsupervised approaches such as clustering and outlier identification are provided by Data Robot and H20-driverless AI. TPOT repeatedly selects the best features using genetic algorithms, whereas google cloud Automl and Auto-Keras employ neural architecture to pick the optimal model. Grid search, random search, and Bayesian search are all useful techniques for optimizing hyperparameters. SMAC (sequence model-based algorithm configuration) is used by auto-weka, whereas SMAC3, a reimplementation of SMAC, is used by auto-sklearn to execute Bayesian optimization effectively. On the parameter spaces, H20-Automl and MLjar employ random search, but H20-driverlessAI, Auto-ML, and auto Keras use both

random and Bayesian search. There are four main approaches to reducing model search and hyper parameter optimization time: a) The tools attempt to quickly find an initial parameter set; b) The tools use the relationship between model selection and hyper parameter optimization; and c) The tools use the relationship between model selection and hyper parameter optimization. c) Setting a maximum runtime for the tools to find the optimum model; d) Limiting the parameters can cause lethargic optimization.

B. Model selection and Prediction Analysis

The bulk of marketing tools, such as H2O-DriverlessAI, Data Robot, and Darwin, use this component, whereas non-commercialized solutions do not. In essence, it uses model dashboards, feature significance, and numerous visualisation tools such as the lift chart and prediction distribution to illustrate detailed findings. These tools also provide outlier data points where the best model fails to produce correct predictions, and they allow model interpretation approaches such as reason code, LIME, Shapley, partial dependence, and others.

V. SOME CASES STUDIES

Some of the case studies of robotic process automation is discussed in the below section

A. Datarobot

Date Robot, one of the well-known automated ML platform was launched in 2012, New England by Jeremy Achin and Thomas DeGodoy. It's modeling engine is a commercial product that supports parallel modeling applications, model construction, and optimizing tools. This engine can be accessed in different ways – cloud-based via internet and customer-specific in specific computing environments. It is used on a single data source for the prediction of a single target variable and optimization or fitting of a single parameter. Data-Robot provides an analysis of numerous machine learning algorithms to generate, employ, and build custom-made predictive models for every situation with its features like Hadoop cluster, plug-and-play, various database certifications, auto-ml, visualization tools, etc. Data - Robot also provides various methods for deploying finished predictive models like native and batch scoring, prediction APIs for scoring in real-time, and exportable prediction code

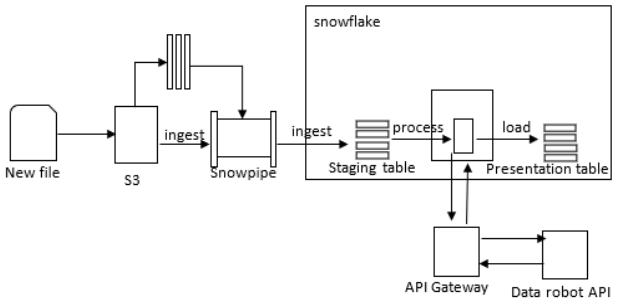


Fig.2 Pipeline in data robot

B. H2O Driverless AI

H20 is available for all major preferred languages(R, SCALA, and JAVA) and available through scripting in multiple languages and web GUI has an excellent model selection framework including forming stacked ensembles, good integration with big data platforms like Hadoop, spark, etc.Hence difficult to use with other frameworks as scikit learn.

H2O Driverless AI is a machine learning platform that uses artificial intelligence (AI) and automates the tough tasks of data science and machine learning. It aspires to reach the most accurate prediction, equivalent to that of professional data scientists but completes it in a fraction of minutes due to end-to-end automation.Automatic visuals and machine learning interpretability are also

available with driverless AI (MLI). Model openness and explanation are equally as crucial as prediction performance, especially in regulated businesses. Modeling pipelines (feature engineering and models) are exported as Python modules and Java standalone scoring artifacts (in full fidelity, without approximations).Commodity hardware is used to operate driverless AI. It was also built to make use of graphics processing units (GPUs), such as multi-GPU workstations and servers like the NVIDIA DGX-1, enabling training that is orders of magnitude quicker.

C. Pipeline of H2O driverless AI

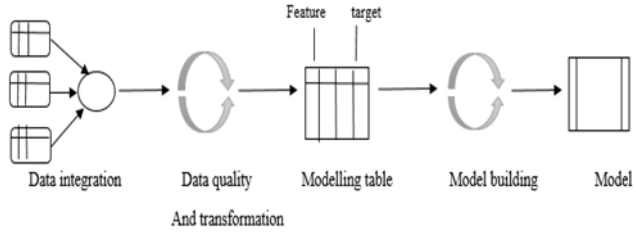


Fig.3Pipeline in H2O driverless AI

D. Azure AutoML

Azure AutoML is a cloud-based service that automates the creation of machine learning pipelines for activities like classification, regression, and forecasting. Its purpose is to determine which model to apply and how to pre-process the input dataset, as well as to tweak the hyper parameters of that model. Azure focuses on a lot of high-dimensional combinatorial optimization issues. In general, they approach these issues by developing probabilistic machine learning models to guide (automatic) experimental judgments, as well as meta-learning to minimize sample complexity and transfer information across related datasets or situations.Model explain ability techniques are available in H2O for both Auto-ML objects (groups of models) and individual models (eg: leader model). With a single function call, explanations may be created automatically, giving a convenient interface for exploring and explaining AutoML models.

E. Pipeline of azure autoML

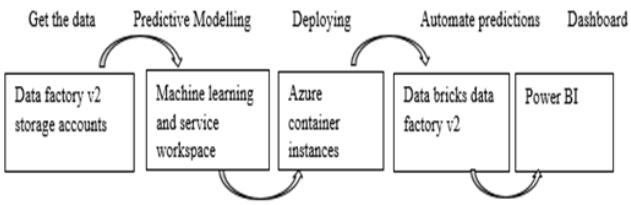


Fig.4Pipeline of AZURE automl

F. Google cloud autoML

Rather than starting from scratch when training models from the data we update Goolge cloud, Auto-ML implements automatic deep transfer learning and neural architecture search for language pair translation, natural language and image classification. They have different Auto-ML tools focusing on different applications few of them are listed below

G. Vertex AI

A unified platform will aid in the development, deployment, and scaling of additional AI models.

H. AutoMLtabular

Build and deploy cutting-edge machine learning models on structured data automatically. Handles a wide range of data primitives in tabular format.

I. AutoML Image

Using object detection and picture classification in the cloud or at the edge, you may gain insights.

- REST and RPC APIs are used.
- Locate things, as well as how many of them there are.
- Use custom labels to categorize photos.

J. AutoML Video

Creates compelling video experiences and effective content discovery.

- Using custom labels, annotate video
- Object identification and tracking

K. AutoML Text

Used to understand and comprehend the meaning of the text through the machine learning techniques

L. AutoML Translation

Used to translate between languages between dynamic

VI. CONCLUSION

The paper presented a study on evolution of automated machine learning for robotic process automationand auto machine learning tools for the robotic process automation is discussed and the hyper-parameters and optimization for RPA using auto ML is discussed and fewcase studies are discussed in which the importance of machine learning is highlighted.

VII. REFERENCES

[1]. K. Kharchenko, O. Beznosyk and V. Romanov, "Implementation of neural networks with help of a data flow virtual machine", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 407-410, 2018.

[2]. M. Lacity, L. P. Willcocks and A. Craig, Robotic process automation at telefonica o2, 2015.

[3]. M. C. Lacity and L. P. Willcocks, "A new approach to automating services", MIT Sloan Management Review, 2017.

[4]. S. Vishnu, V. Agochiya, R. Palkar et al., "Data-centered dependencies and opportunities for robotics process automation in banking", Journal of Financial Transformation, vol. 45, pp. 68-76, 2017.

[5]. Aguirre, Santiago & Rodriguez, Alejandro. (2017). Automation of a Business Process Using Robotic Process Automation (RPA): A Case Study. 65-71. DOI: 10.1007/978-3-319-66963-2_7.

[6]. van der Aalst, W. M., Bichler, M., & Heinzl, A. (2018). Robotic Process Automation. Bus Inf Syst Eng 60, pp.269–272.

[7]. Moffitt, K. C., Rozario, A. M., & Vasarhelyi, M. A. (2018). Robotic process automation for auditing. Journal of Emerging Technologies in Accounting, 15(1), 1-10.

[8]. Enríquez, J. G., Jiménez-Ramírez, A., Domínguez-Mayo, F. J., & García-García, J. A. (2020). Robotic Process Automation: A Scientific and Industrial Systematic Mapping Study. IEEE Access, 8, 39113-39129.